CAPÍTULO 24

Experimento em sistemas colaborativos

Jacques Wainer

META

Apresentar conceitos sobre como realizar um experimento para avaliar um sistema colaborativo em comparação a alguma alternativa, por exemplo, a colaboração sem o sistema ou em comparação a outro sistema já em uso.

OBJETIVOS EDUCACIONAIS

Após o estudo desse capítulo, você deverá ser capaz de:

- Projetar desenhos experimentais para avaliação de sistemas colaborativos.
- Enumerar ameaças à validade de um experimento.
- Identificar o tipo de teste estatístico apropriado para analisar os dados coletados de um experimento.
- Analisar, a partir dos dados coletados no experimento, se o uso do sistema provocou uma diferença real (com diferença significativa estatisticamente).

RESUMO

Alguns sistemas colaborativos são adotados independente de uma avaliação que mostre benefício objetivo e mensurado. Correio eletrônico e outros sistemas apresentaram funcionalidades novas e facilidades tão óbvias que foram adotados independentemente de uma avaliação de impacto mais elaborada. Contudo, a maioria dos sistemas colaborativos não tem tal sorte. Na maioria dos casos é necessário demonstrar que o uso do sistema traz algum benefício em relação à alternativa: ou outro sistema já em uso ou nenhum sistema. Neste capítulo é discutido como realizar demonstrações por meio de avaliações quantitativas, e na maioria das vezes comparativas, de um sistema colaborativo. Neste capítulo não são enfocadas as fórmulas e os algoritmos associados aos testes estatísticos, pois estão disponíveis em livros-texto de estatística. O importante, do ponto de vista deste capítulo, é entender os conceitos centrais dos testes, por exemplo: erro de amostragem, inferência, p-valor, pré-requisitos de cada teste, dentre outros conceitos. Neste capítulo, a avaliação quantitativa de sistemas é discutida com o rigor apropriado para a realização de uma pesquisa científica.

24.1. Experimento

Um experimento é realizado para testar uma hipótese. Em pesquisas com sistemas colaborativos, a hipótese geralmente é que um sistema é melhor do que outro, ou então que se obtêm melhores resultados da colaboração com o uso de um dado sistema do que sem o uso desse mesmo sistema. Um experimento é realizado quando é necessário convencer sobre os benefícios do uso de um sistema a partir de uma avaliação quantitativa e com validade estatística, o que é amplamente aceito pela comunidade científica.

Experimento é caracterizado pela manipulação de algumas variáveis em situações laboratoriais, isto é, artificiais ou semiartificiais. Diferentemente das ciências naturais (Física, por exemplo) em que os resultados são obtidos diretamente do uso de um equipamento sem envolver seres humanos, em Ciências da Computação um experimento necessariamente envolve um grupo de pessoas e várias medidas sobre as pessoas. Experimento também envolve controle – o experimentador decide que grupo de pessoas fará o quê e quem participará de quais grupos. Isto difere de pesquisas observacionais nas quais o pesquisador não tem controle sobre o grupo de pessoas. As seguintes atividades devem ser realizadas num experimento (Figura 24.1):

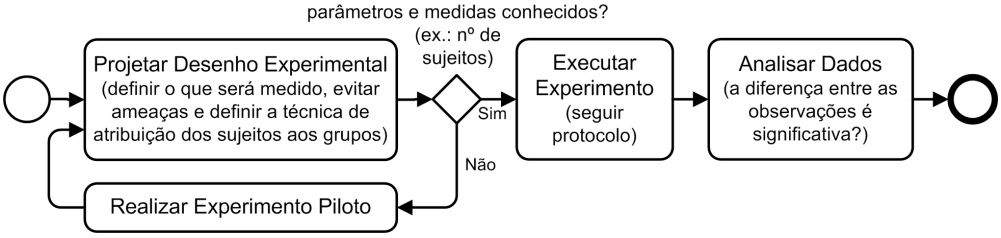


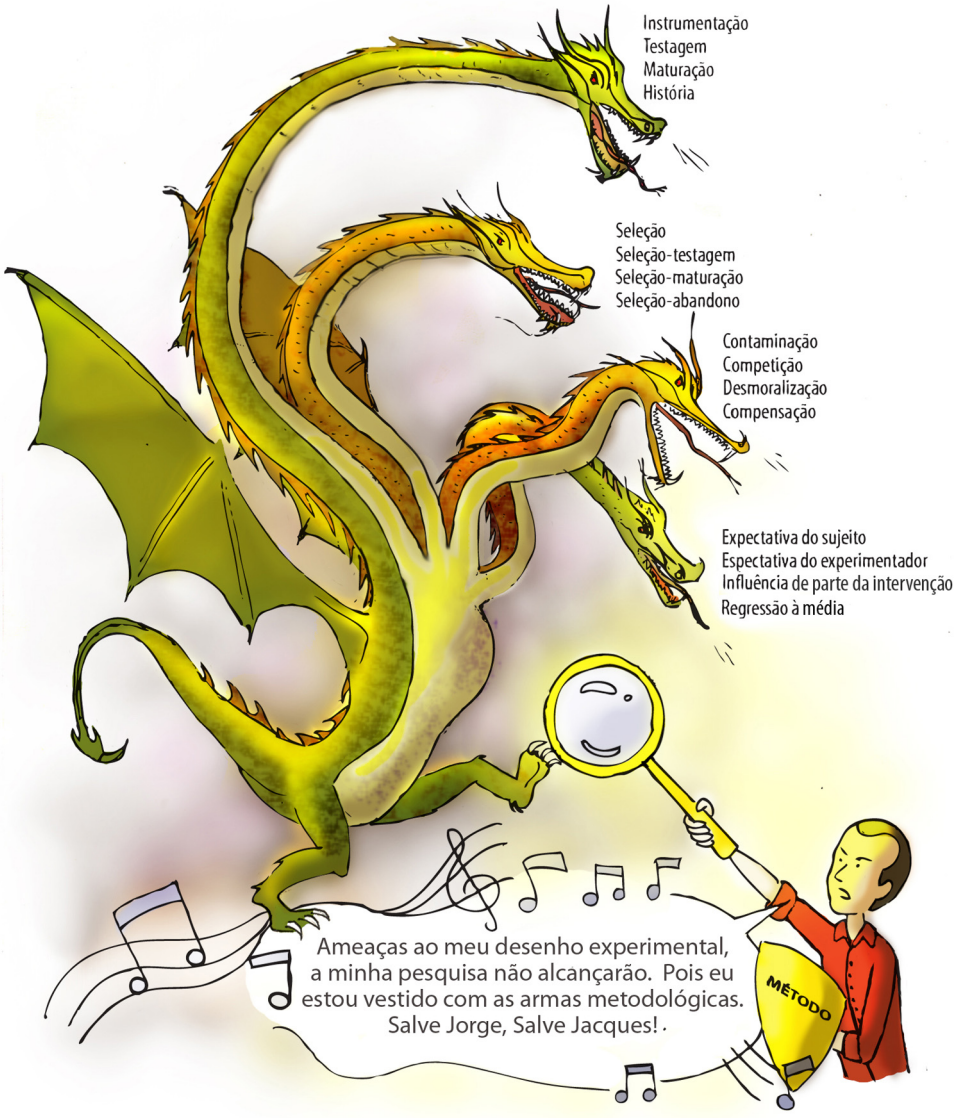
Figura 24.1 – Processo de um experimento

Experimento requer conhecimento de estatística. Neste capítulo, o enfoque é nos conceitos e nos procedimentos para a realização adequada de testes estatísticos e não serão discutidas as fórmulas, pois o leitor poderá consultá-las em livros-textos específicos de Estatística. Aconselha-se que o leitor também adote algum pacote estatístico para a realização de testes

por computador, como as Funções de Estatística em Excel, ou sistemas específicos como SPSS, SAS, e R que é um ótimo pacote estatístico gratuito que roda em todos os sistemas operacionais.

24.2. Ameaças à validade do experimento

Se um experimento for projetado inadequadamente, pode levar o pesquisador a obter uma conclusão errada sobre a hipótese. O desenho experimental[1] deve ser projetado para evitar algumas ameaças à validade da conclusão sobre o experimento.



1 Experimental Design. Há várias alternativas para a tradução dessa expressão em português, por exemplo, "delineamentos experimentais", ou "projetos experimentais". Neste capítulo optamos por "desenho experimental" [Hochman 2005].

Validade interna é a confiança que se tem de que o efeito observado é realmente devido à manipulação feita, e não em função de outros fatores. Validade externa é a confiança que se tem de que o efeito observável é generalizável, ou seja, mesmo acreditando que o efeito é decorrente da manipulação realizada num grupo, tal efeito se repetirá em outros grupos?

Começaremos com o problema da validade interna. Vamos supor um experimento para verificar se o uso de um sistema de apoio a reuniões diminui o tempo para tomar uma decisão.

EXPERIMENTO E1

Oferecemos aos participantes do experimento um problema P1, esperamos que eles resolvam o problema e medimos o tempo decorrido até a resolução. Instalamos um sistema, definimos outro problema P2 e exigimos que os participantes usem o sistema para resolver o problema, e medimos o tempo até a resolução de P2.

Do ponto de vista de desenho experimental, esse experimento tem o desenho: um grupo com pré-teste e pós-teste, representado por:

$$O_1 \quad X \quad O_2$$

onde O_1 representa o pré-teste (ou uma Observação), O_2 representa o pós-teste, e X representa a introdução do sistema de apoio a reuniões. Em desenhos experimentais, o X é denominado intervenção.

Mesmo que o tempo para a solução de P2 seja muito menor do que o tempo para a solução de P1, pode-se afirmar que a diminuição do tempo foi causada pelo uso do sistema de apoio a reunião? Ou, de outro ponto de vista, quais são as explicações rivais para a diminuição do tempo de resolução do problema P1 para o P2? Essas explicações alternativas são as ameaças à validade interna do experimento.

A primeira ameaça é que talvez P2 seja mais fácil do que P1 e, portanto, mais rápido de ser resolvido. Essa ameaça é chamada de instrumentação, e se baseia na ideia de que a diferença entre O_1 e O_2 talvez seja decorrente de um erro na medição. Por exemplo, os testes de O_1 e O_2 são instrumentos suficientemente diferentes, ou as observações são feitas por pessoas diferentes, etc.

Vamos analisar outra ameaça. Após a experiência em resolver P1, os membros do grupo ficaram mais cientes de como os participantes agem nas reuniões e quais são os comportamentos mais benéficos para o grupo. Ao resolver P2, e sabendo que o importante é o tempo para a resolução do problema, pela experiência anterior, os participantes agirão da forma que eles já sabem que leva à resolução do problema mais rapidamente. Essa ameaça é chamada de testagem, decorre do fato de que passar por O_1 prepara os participantes para um melhor desempenho em O_2. Um exemplo típico da ameaça de testagem é aplicar a mesma prova no pré e no pós-teste numa intervenção educacional – uma possível melhora pode ter sido causada pelo fato dos participantes já terem feito a prova antes e não necessariamente causada pela intervenção.

Vamos supor que o experimento E1 tenha sido realizado num único dia: o primeiro problema foi resolvido durante a manhã, depois os participantes foram treinados a usar o

sistema no começo da tarde, e no fim da tarde resolveram o segundo problema. Digamos que o tempo para resolução de P2 tenha sido maior. Podemos afirmar que o sistema piorou a eficiência do grupo? Uma explicação alternativa é que os participantes estavam cansados. Essa ameaça é chamada de maturação, os sujeitos dos experimentos podem se tornar menos capazes com o tempo, independentemente da intervenção. Agora analisemos um exemplo de maturação em que os sujeitos se tornam mais capazes com o tempo. Uma empresa mede o número de erros de programação durante um ano, instala uma ferramenta CASE, e no ano seguinte mede de novo os erros. Programadores tendem a se tornar melhores com a prática, e depois de um ano de projetos, os programadores da empresa devem estar mais competentes, o que também explicaria a diminuição de erros por projeto. Assim, se houve uma diminuição do número de erros, não é possível afirmar se foi por causa da ferramenta CASE ou por maturação dos programadores.

Vamos supor que as duas medições de E1 tenham sido feitas separadas por meses. Uma ameaça é que talvez a empresa tenha promovido alguma iniciativa para que o grupo funcionasse melhor entre as resoluções de P1 e P2. Por exemplo, a empresa pode ter contratado consultores para melhorar as relações interpessoais e a dinâmica de grupo. Ou vários outros problemas semelhantes tenham sido resolvidos pelo grupo entre P1 e P2 e o grupo aprendeu a resolver melhor os problemas. Essa ameaça é chamada de história, que é a possibilidade de que tenha ocorrido um ou mais eventos externos ao experimento que sejam a real causa da melhoria dos resultados.

Nem todas essas ameaças podem ser relevantes para um dado experimento, mas se o pesquisador projetar o desenho E1, precisará explicar quais das ameaças não se aplicam e por quê. Para evitar essas ameaças, o pesquisador deve projetar um desenho experimental com dois grupos.

Vamos analisar o experimento a seguir:

EXPERIMENTO E2

Definimos dois grupos de pessoas. Para o grupo 1 damos um sistema e medimos o resultado da resolução do problema P1. Para o grupo 2 não damos o sistema e medimos o resultado para o mesmo problema P1.

O desenho experimental de E2 é:

X	*O*	*grupo 1*
	O'	*grupo 2*

O grupo 1, que sofre a intervenção, é chamado de grupo experimental, e o outro, que não sofre a intervenção, é chamado de grupo de controle. Neste caso, o efeito de *X* é verificado a partir da comparação entre *O* e *O'*, isto é, se o resultado do grupo experimental é melhor do que o resultado do grupo de controle.

Para este desenho experimental, as ameaças de instrumentação, testagem, maturação e história não são relevantes (reflita no porquê). Mas isso não significa necessariamente que *X* seja a causa de possíveis resultados melhores do grupo experimental. Uma explicação alternativa é que, por um acaso, o grupo 1 já era melhor do que o grupo 2, e a diferença observada nos resultados apenas reflete a diferença pré-existente entre os grupos. Essa ameaça chama-se seleção: o fato dos grupos serem diferentes influencia a medida obtida com o teste.

Já que a seleção pode influenciar os resultados, então talvez valha a pena medir os dois grupos antes da intervenção. Analisemos o seguinte experimento:

EXPERIMENTO E3

Definimos dois grupos e medimos o tempo decorrido até a resolução de P1 em cada grupo. O grupo experimental aprende a usar o sistema. Oferecemos o problema P2 para ambos os grupos e medimos o tempo para a resolução.

O desenho experimental de E3 é:

O_1	*X*	O_2	*grupo 1*
O'_1		O'_2	*grupo 2*

O efeito de X é verificado a partir da comparação entre O_2 - O_1 e O'_2 - O'_1, isto é, se a redução de tempo do grupo experimental é maior do que a redução de tempo do grupo de controle. Com isso, mesmo que os grupos sejam diferentes, a seleção não é mais uma ameaça, pois estamos falando da redução em cada grupo e não apenas do valor do pós-teste. A comparação é em termos relativos em vez de absolutos.

Há outras ameaças, similares à seleção, que derivam das diferenças entre os grupos:

- **Interação seleção-testagem.** É possível que o grupo experimental tenha a capacidade de aprender mais rápido do que o grupo de controle, e o grupo pode ser mais eficiente na solução do segundo problema independente do uso do sistema.

- **Interação seleção-maturação.** É possível que o grupo experimental tenha uma taxa de maturação diferente da taxa do grupo de controle. Se o experimento for executado num mesmo dia, é possível que o grupo experimental tenha mais resistência física e tenha ficado menos cansado do que o grupo de controle.
- **Interação seleção-abandono (ou mortalidade seletiva).** Se o experimento for executado ao longo de várias semanas ou meses, é possível que nem todos os membros que fizeram o pré-teste estejam disponíveis para o pós-teste. A mortalidade seletiva aparecerá se houver uma dupla tendência: se os indivíduos de somente um dos grupos tiverem probabilidade maior de abandonar o experimento, ou se os melhores ou os piores indivíduos tiverem probabilidade maior de abandonar o experimento. Se os melhores do grupo de controle abandonarem o experimento, então a segunda observação tende a ser mais baixa já que os melhores saíram, e isso poderá ser uma explicação alternativa para que o ganho do grupo experimental seja maior do que o ganho do grupo de controle. Outra explicação alternativa é se os piores do grupo experimental abandonarem o experimento.

As ameaças foram caracterizadas até agora como explicações alternativas para o efeito da intervenção, ou seja, quando o ganho observado num grupo não é de fato decorrente da intervenção. Ameaças também podem provocar o efeito inverso, constituir-se numa explicação alternativa para um efeito nulo, isto é, mesmo observando que o ganho do grupo experimental não tenha sido maior do que o ganho do grupo de controle, é possível que a intervenção tenha efeito positivo, mas a ameaça tende a tornar os resultados mais homogêneos. Não há uma nomenclatura estabelecida, mas vamos chamar uma ameaça de positiva se amplifica as diferenças dos resultados podendo mascarar o fato da intervenção não ter funcionado, e negativa quando tende a homogeneizar os resultados podendo mascarar o fato de a intervenção ter um efeito positivo.

Algumas ameaças são intrinsecamente positivas ou negativas. Testagem é sempre positiva. Instrumentação pode ser positiva: P2 é mais fácil do que P1 e, consequentemente, a medida do pós-teste é melhor do que a medida do pré-teste independentemente do uso do sistema. Mas instrumentação também pode ser negativa: P2 é mais difícil do que P1 e, consequentemente, a medida do pós-teste é pior do que a medida do pré-teste mesmo que o sistema tenha apoiado a resolução do problema.

Vejamos algumas ameaças quando um grupo sabe da existência do outro grupo, e em alguns casos os membros entre os grupos até se comunicam. As ameaças a seguir são intrinsecamente negativas, com a exceção da ameaça de desmoralização, que é sempre positiva.

- **Contaminação**. Membros do grupo experimental podem ensinar aos membros do grupo de controle algumas das técnicas às quais estão sendo submetidos. A contaminação tem grande potencial de ocorrer em educação quando os alunos do grupo experimental passam os conteúdos que receberam para os alunos do grupo de controle. Em computação também pode ocorrer contaminação.
- **Competição**. Membros do grupo de controle podem se sentir preteridos frente aos do grupo experimental, porque é atribuído um valor positivo à intervenção – por exemplo, o grupo experimental passa a usar um "sistema legal". Os membros do grupo de controle podem se sentir motivados a provar que são tão bons quanto os do grupo experimental e que mereceriam também usar aquele sistema.

- **Desmoralização**. Membros do grupo de controle podem se sentir preteridos, mas ao contrário do que ocorre no comportamento competitivo, podem se sentir desmotivados e assim ter um desempenho pior.
- **Compensação**. Pode ser que alguma autoridade crie medidas compensatórias para favorecer o grupo de controle ao sentir que o grupo foi preterido por não receber a intervenção.

Existem também as ameaças que são causadas pelas expectativas dos envolvidos na pesquisa:

- **Efeito da expectativa do sujeito (efeitos placebo e Hawthorne)**. O efeito placebo é muito conhecido na medicina: mesmo quando é dado um remédio inócuo (por exemplo, uma pílula contendo farinha), o paciente ainda pode achar que os sintomas melhoraram. A expectativa que o paciente tem de melhorar, por ter tomado o que ele acha que é um remédio, causou uma melhora! Um efeito similar ao placebo, e talvez mais relevante para experimentos em Ciências da Computação, é o efeito Hawthorne: ocorre um efeito positivo apenas pelo fato de os sujeitos saberem que estão sendo observados. Em experimentos computacionais, o efeito Hawthorne precisa ser levado em consideração. Exemplos: engenheiros de software que sabem que estão sendo observados melhoram a produtividade ou a qualidade do software gerado, usuários melhoram o desempenho, alunos melhoram o aprendizado etc.

EXPERIÊNCIA NUMA FÁBRICA DO BAIRRO HAWTHORNE

O caso que gerou a teoria do efeito Hawthorne decorre de um estudo da produtividade dos trabalhadores numa fábrica localizada no bairro Hawthorne (Chicago, EUA). O objetivo era investigar a relação entre a luminosidade e a produtividade. Constatou-se que a produtividade aumentava independentemente da intensidade de iluminação. A teoria é que a produtividade dos trabalhadores aumentou porque eles sabiam que estavam sendo observados.

- **Efeito da expectativa do experimentador**. O efeito da expectativa do experimentador acontece em alguns exemplos onde o pesquisador interage intensamente com o sujeito, e as crenças do experimentador causam um efeito no sujeito, ou ao menos nos testes realizados pelo sujeito. Por exemplo, quando o pós-teste requer alguma avaliação subjetiva, o pesquisador pode "melhorar" as notas do pós-teste, ainda que inconscientemente, se estiver esperando que a intervenção seja positiva e souber quais são os testes advindos da intervenção. Outro fenômeno é o efeito Pigmaleão ou Rosenthal em educação: quando os professores receberam a informação falsa de que na classe tinham alunos mais inteligentes do que a média, os alunos obtiveram resultados melhores do que os alunos pertencentes a classes em que os professores não receberam a informação falsa. A teoria propõe que o experimentador passa sinais inconscientes que acabam influenciando os sujeitos. O efeito da expectativa do experimentador também é relevante para experimentos em computação. Se o experimentador é o criador de um sistema, pode passar aos usuários sinais que indicam sua expectativa de que aquele sistema é útil. A ameaça também está presente quando a avaliação do pós-teste tiver algum aspecto subjetivo a ser avaliado pelo criador (por exemplo, se for preciso classificar os erros de software em sérios ou não).

Por fim, também são conhecidas as seguintes ameaças à validade interna de um experimento:

- **Influência de apenas parte da intervenção**. Este efeito não tem um nome padrão, aparece em diferentes domínios com diferentes nomes. A ideia é que o efeito não decorre da intervenção como um todo, mas somente de parte da intervenção. Por exemplo, a partir de um desenho experimental apropriado, foi mostrado que houve ganhos de aprendizado dos alunos quando usado um sistema de aprendizagem a distância contendo vários sistemas colaborativos como wiki, blog, fórum, bate-papo, dentre outros. Nem todos os sistemas e todas as funcionalidades são necessários para a obtenção do mesmo efeito, e o pesquisador não saberá indicar o que, exatamente, é responsável pelo efeito.
- **Regressão à média**. É um efeito decorrente do uso de pré-teste para selecionar o grupo experimental. Espera-se que os sujeitos que tiraram os piores resultados no pré-teste melhorem esse resultado, independentemente da intervenção. Para explicar esse fenômeno, vamos assumir que o resultado do pré e do pós-teste é um número aleatório entre 0 e 100 independentemente das características intrínsecas do sujeito. Sendo assim, espera-se que os sujeitos que tiraram os piores 10% resultados no pré-teste tenham no pós teste uma média de 50 pontos (daí o termo regressão para a média), que daria a impressão que esse grupo melhorou do pré-teste para o pós-teste. É claro que nenhum teste é totalmente aleatório, mas quase nenhum teste é totalmente não aleatório – sempre existe um componente de "sorte ou azar" num teste: "estudei toda a matéria menos o assunto daquela questão"; ou "eu estava com dor de cabeça no dia que escrevemos o programa". O componente aleatório de cada teste tende a fazer as pessoas que tiraram as piores notas no pré-teste terem notas melhores, pelo menos um pouco. Isto também é verdade para os que tiraram as melhores notas no pré-teste, que tenderão a ter notas menores no pós-teste.

As ameaças à validade externa são mais sutis. O objetivo de um experimento é obter um conhecimento que possa ser generalizado. Se o experimento mostra que o uso de um sistema provocou uma diminuição significativa do tempo de resolução dos problemas, então espera-se que esse conhecimento possa ser generalizado para a afirmação: "o uso do sistema reduz o tempo de resolução de problemas". Há duas generalizações nessa afirmativa:

- que o resultado vai valer para qualquer pessoa, em qualquer lugar, em qualquer ambiente, em qualquer tempo; e
- que o resultado vai valer para situações reais (não experimentais, não artificiais).

A primeira generalização é a mais forte, está relacionada ao quanto especial foi o grupo de pessoas escolhidas para fazer parte do experimento, ao quanto especial também foram o local, o momento e o ambiente onde foi feito o experimento. Essa capacidade de generalização está relacionada ao problema de amostragem (ver Seção 24.4).

A segunda generalização é um caso particular da primeira no que se refere ao ambiente: generalização dos resultados de ambientes mais artificiais (o ambiente de laboratório usado na experimentação) para ambientes mais naturais (usuários usando o sistema no cotidiano). Essa segunda generalização causa algumas confusões de nomenclatura. Alguns autores consideram que os efeitos de expectativa do sujeito e do experimentador são ame-

aças à validade externa e não à validade interna, porque esses efeitos impedem a segunda generalização – são efeitos que aparecem somente porque os resultados foram obtidos em situações artificiais.

24.3. Desenhos experimentais relevantes para Sistemas Colaborativos

Nesta seção são discutidos os desenhos experimentais mais relevantes para experimentação em computação. Assumiremos que:

- O resultado dos testes é um número, e que podemos tirar a média destes números (é o que chamaremos de medida intervalar na seção seguinte).
- Que valores maiores são "melhor" – assumiremos que os testes medem coisas como "notas" que quanto maior melhor, e não coisas como "tempo de execução" que quanto menor melhor.

Indicaremos abreviadamente que o resultado do pós-teste é melhor do que o resultado do pré-teste, por:

$$\mu \text{ (pós-teste)} >> \mu \text{ (pré-teste)}$$

Se essa comparação dos resultados é verdadeira e as ameaças foram tratadas, então diremos que a intervenção teve um efeito positivo.

Um grupo, pós-teste

O desenho experimental é:

$$X \quad O$$

Aplica-se a intervenção e mede-se o resultado. Este desenho sofre de todas ou quase todas as ameaças à validade interna. Não é possível saber se: o resultado obtido O já valia antes da intervenção ou se foi efeito de algum outro fenômeno que não seja X. Apesar das ameaças, há uma situação em que esse desenho pode ser usado – quando o resultado O é tão obviamente espetacular que é de senso comum que ele não valia antes, e que nenhum outro efeito conhecido pode ter levado a O. Por exemplo, este desenho experimental é suficiente para justificar o uso do paraquedas: poucas pessoas que saltaram de um avião usando um paraquedas nos últimos 50 anos morreram: isso não valia antes do uso do paraquedas, e não se conhece outra explicação que poderia gerar esse efeito! Não está claro, para o autor deste capítulo, que resultados deste tipo são possíveis em computação. Recomenda-se fortemente que este desenho não seja projetado para um experimento.

Um grupo, pré e pós-teste (ou experimento antes-depois)

Este desenho é muito comum em experimentos que envolvem sistemas. Mede-se o ambiente antes da adoção do sistema (O_1) e depois (O_2):

$$O_1 \quad X \quad O_2$$

Se o grupo obtém resultado "melhor", então a intervenção tem um efeito positivo:

$$\mu(O_2) >> \mu(O_1)$$

Como foi discutido na seção anterior, quando o pesquisador projeta um experimento seguindo esse desenho, precisa estar ciente e discutir as ameaças de instrumentação, testagem, maturação e história com relação ao experimento realizado.

Um grupo, pré e pós-teste com remoção da intervenção

O desenho experimental é representado por:

$$O_1 \quad X \quad O_2 \quad X^{-1} \quad O_3$$

O objetivo é aumentar a confiança de que X é realmente a causa da mudança na variável de teste, pois quando X é removido (X^{-1}) a variável de teste piora. Deve-se mostrar que:

$$\mu(O_2) >> \mu(O_1) \quad \text{e} \quad \mu(O_2) >> \mu(O_3)$$

Este desenho só é aplicável se a intervenção puder ser totalmente removida. Se a intervenção leva os sujeitos a aprenderem algo (um assunto ou uma forma de agir), remover a intervenção não levará as pessoas a esquecerem o que aprenderam, e portanto a intervenção não pode ser removida totalmente.

Como este desenho tem um aspecto de "confirmação" que a intervenção melhorou e sua remoção piorou o resultado dos testes, pode apresentar resultados impossíveis de serem interpretados. Por exemplo, vamos dizer que:

$$\mu(O_2) >> \mu(O_1) \quad \text{e} \quad \mu(O_2) \approx \mu(O_3)$$

Neste caso a melhora de O_2 indica que a intervenção foi positiva, mas a sua remoção indica que ela é irrelevante. Como interpretar esse resultado? Veja que não podemos usar o argumento que os sujeitos aprenderam com a intervenção e usaram isso quando a intervenção foi removida, pois só podemos usar este desenho quando sabemos que não há "aprendizado"!

Este desenho pode ser estendido, e reintroduzimos a intervenção após O_3. Esta extensão é chamada de intervenção repetida:

$$O_1 \quad X \quad O_2 \quad X^{-1} \quad O_3 \quad X \quad O_4$$

Neste novo desenho há ainda mais um nível de confirmação, mas de novo, isto abre mais possibilidades de não ser possível interpretar o resultado.

Um grupo, pré e pós-teste com variável dependente não equivalente

O desenho experimental é:

$$(O_1, V_1) \quad X \quad (O_2, V_2)$$

O que torna esse desenho diferente é que duas variáveis são medidas tanto no pré quanto no pós-teste. A variável *O* é a tradicional variável que se espera medir o efeito da intervenção. A segunda variável, *V*, é chamada de dependente porque mede um fenômeno que se quer observar, e não equivalente porque deve ser insensível à intervenção. A variável *V* não equivalente deve ser sensível à história e à maturação da variável *O*, mas não à intervenção. Assim, a variação sofrida por *V* mede as duas ameaças e seu efeito é "retirado" do ganho entre V_2 e V_1.

A maioria dos experimentos que usam variável dependente não equivalente faz a seguinte análise: mostra-se que houve ganho entre o pré e o pós-teste e mostra-se que a variável não equivalente "não mudou muito":

$$\mu(O_2) >> \mu(O_1) \quad \text{e} \quad \mu(V_2) \approx \mu(V_1)$$

Vejamos um exemplo. Vamos supor que um sistema de workflow é usado para rotear tarefas entre os funcionários de uma empresa, mas não dá suporte à execução das tarefas em si. Mesmo assim, graças ao roteamento automático e ao controle promovidos pelo sistema, espera-se que o tempo de execução de um processo diminua. Esta é a medida *O*. Mas uma diminuição de O_1 para O_2 pode ser devido à maturação – os funcionários ficam cada vez mais eficientes no trabalho que fazem. Assim deve-se medir também o tempo de execução das tarefas, que é a variável *V* dependente e não equivalente. Se houve maturação, a medida *V* deve diminuir entre antes e depois. Se não houve maturação, a medida *V* não deve ser modificada pelo uso do sistema de workflow.

Dois grupos, apenas pós-teste

O desenho mais simples com dois grupos é:

$$\begin{array}{ll} X & O \\ & O' \end{array}$$

A intervenção tem resultados positivos se:

$$\mu(O) >> \mu(O')$$

Como discutido na seção anterior, desenhos com dois grupos reduzem as ameaças de história e maturação, entre outras, mas são ameaçados pela seleção (combatida com técnicas de atribuição dos sujeitos aos grupos conforme discutido na próxima seção) e também pela competição, contaminação, entre outras ameaças. Por outro lado, este desenho de dois grupos com apenas pós-teste é um dos únicos que evita as ameaças de testagem e instrumentação, já que não há pré-teste.

Dois grupos, pré e pós-teste

$$\begin{array}{lll} O_1 & X & O_2 \\ O'_1 & & O'_2 \end{array}$$

A intervenção tem um efeito positivo se o grupo experimental tiver um ganho maior do que o ganho do grupo de controle:

$$\mu(O_2 - O_1) >> \mu(O'_2 - O'_1)$$

Este desenho, embora reintroduza os problemas de instrumentação e testagem, é insensível à ameaça de seleção. Já que estamos medindo o ganho nos testes (pós-teste menos o pré-teste), mesmo que os grupos sejam diferentes antes da intervenção, seremos capazes de verificar se a intervenção tem um efeito positivo.

24.4. Atribuição dos participantes aos grupos

Experimentos com dois grupos requerem a atividade de atribuição: decidir quem vai participar do grupo de controle e do experimental.

CATÁLOGO DE DESENHOS EXPERIMENTAIS

Há outros desenhos possíveis, discutidos em livros-texto como o do Shadish (2002). Há também a possibilidade de combinar ideias de vários desenhos, por exemplo, acrescentar uma variável dependente não equivalente em quase todos os desenhos apresentados que possuam um pré-teste.

Atribuição por conveniência

O caso mais comum em experimentos de computação é a atribuição por conveniência ou não aleatória, o que ocorre quando os grupos de controle e experimental estão definidos por critérios que o experimentador não tem controle. Por exemplo, os programadores do time A já existente usam um sistema colaborativo e os do time B não usam. Ou os alunos da turma A usam um sistema de aprendizagem colaborativa, e os da turma B não usam. Os grupos já estavam pré-formados, o experimentador não teve controle sobre quem está no grupo A ou B.

Atribuição por conveniência sofre da ameaça da seleção e da sobreposição de outras ameaças: seleção-maturação, seleção-testagem, e seleção-abandono. O time A pode já ter mais tempo de prática em comum; o time B pode ter uma dinâmica de grupo particularmente tensa; a turma A pode ter sido formada com os melhores alunos do ano anterior; a turma B pode ter aulas numa sala particularmente barulhenta etc.

Todo desenho de dois grupos que não usa a atribuição aleatória é chamado na literatura de quasi-experimento. Em algumas ciências, a ameaça de seleção é considerada um risco sério, e essas ciências recomendam evitar a atribuição por conveniência. Em Ciências da Computação, o risco da ameaça de seleção não deve dissuadir o experimentador de usar este tipo de atribuição, e se a seleção e suas interações forem uma ameaça relevante, o pesquisador deve adotar um projeto de desenho experimental que evite essas ameaças.

Atribuição aleatória

Na atribuição aleatória, os sujeitos são atribuídos aos grupos de forma randômica. A aleatoriedade na atribuição "mistura" os dois grupos, de tal forma que eles são provavelmente homogêneos e equivalentes, e portanto não há a ameaça de seleção.

A atribuição aleatória de fato não garante que os grupos sejam equivalentes – pode acontecer de apenas os "bons participantes" caírem num dos grupos. A atribuição aleatória garante que tais situações são improváveis somente se o número de pessoas for grande. Se o número de pessoas for pequeno, não há garantias de equivalência entre os grupos mesmo com atribuição aleatória. Por exemplo, se há seis pessoas e elas forem aleatoriamente distribuídas em dois grupos do mesmo tamanho, é alta a chance das duas melhores pessoas caírem no mesmo grupo.

Casamento (matching)

Casamento é o processo de cuidadosamente selecionar quem participará de cada um dos grupos de tal forma que os grupos sejam o mais iguais possível. Por exemplo, se o desempenho no pré-teste for a melhor medida para prever o desempenho no pós-teste, então deve-se construir os grupos de tal forma que a distribuição dos valores do pré-teste seja o mais parecida nos dois grupos. Por exemplo, o sujeito com o melhor pré-teste vai para o grupo 1 e o segundo melhor para o grupo 2, o terceiro melhor vai para o 2 e o quarto melhor para o 1, e assim por diante.

Casamento também pode ser feito com o uso de outras variáveis diferentes do pré-teste, como anos de experiência, idade, condição física, ou qualquer outra variável relevante para o resultado da intervenção.

QUEM SÃO OS SUJEITOS DE UM EXPERIMENTO: INDIVÍDUO, GRUPO OU ORGANIZAÇÃO?

Um sistema colaborativo pode trazer impactos para o individuo, como a melhora de satisfação ou a melhora de produtividade; ou pode trazer ganhos para o grupo que usa o sistema, geralmente em termos de produtividade; ou para a organização que fomenta ou impõe o uso destes sistemas, também na forma de ganhos em produtividade. Por exemplo, em sistemas colaborativos como correio eletrônico, editor cooperativo de texto, ambiente de aprendizagem colaborativa etc., os ganhos são mais evidentes com relação aos indivíduos, pois se tornam mais eficientes em algum aspecto ou se sentem mais satisfeitos em fazer o trabalho. Sistemas como apoio a reuniões, suporte ao desenvolvimento colaborativo de software, entre outros, trazem ganhos mais óbvios para o grupo; por exemplo, aumenta a eficiência em resolver problemas ou em desenvolver softwares. Finalmente, sistemas como workflow e gerenciamento de conhecimento, parecem trazer ganhos mais claramente definidos para a organização.

A definição de quem são os sujeitos de um experimento depende do nível em que o benefício pode ser melhor evidenciado. Se você estiver pesquisando ganhos no nível individual, cada sujeito será uma pessoa. Se o ganho a ser medido estiver no nível de grupo, então os sujeitos são os grupos experimental e de controle. E se o ganho é no nível da organização, então várias organizações irão compor tanto o grupo de controle quanto o grupo experimental.

24.5. Execução de experimentos

O conjunto de procedimentos para a execução de um experimento é denominado protocolo. Algumas ameaças não são tratadas no desenho do experimento, mas sim na sua execução.

Por exemplo, ameaças como contaminação podem ser resolvidas mantendo os grupos experimentais e de controle longe um do outro. Desta forma os grupos não se encontram e, portanto, não trocam informação. Competição e desmoralização podem ser combatidas omitindo a informação de que o grupo de controle está fazendo parte de um experimento, bastando manter os participantes fazendo as coisas como sempre fazem.

Uma técnica de execução de experimento, importante em outras áreas mas praticamente impossível na computação, é não informar aos sujeitos se eles estão no grupo de controle ou experimental – chamado de experimento cego. Isto evita a competição, desmoralização e a expectativa do sujeito, mas a utilidade em computação é claramente limitada: como evitar que o sujeito perceba que ele está usando um sistema novo? Um protocolo ainda mais restritivo, chamado duplo cego, é manter o experimentador também ignorante sobre quem são os grupos de controle e o experimental. Isto evitaria que a expectativa do experimentador fosse uma ameaça. Finalmente, o protocolo triplamente cego impõe que quem faz as análises do experimento também não saiba quais eram os dois grupos para evitar que o analista se torne "criativo" com as técnicas estatísticas. Embora "cegar" o sujeito possa ser difícil em experimentos computacionais, nessa área é uma boa prática fazer o experimentador e o analista não identificarem os grupos de controle e o experimental.

24.6. Análise dos resultados de um experimento

Para analisar os resultados de um experimento, considere as questões:

- A diferença entre o pós e o pré-teste é real?
- A diferença entre o pós e o pré-teste é importante na prática?
- A diferença entre o pós e o pré-teste vale o custo da intervenção?

Para responder a primeira questão, é preciso verificar se a diferença realmente existe ou se é apenas um efeito da sorte. Não há uma resposta booliana, no máximo é possível responder em termos de probabilidade: a diferença é provavelmente real, ou então, provavelmente não há uma diferença real entre o pós e o pré-teste. Por exemplo, suponha que num experimento antes-depois, a média dos pré-testes deu 95,3 (unidade arbitrária) e a média dos pós-testes deu 97,1. Podemos dizer que a diferença entre estas médias é real? Não necessariamente. Precisamos verificar se a diferença entre o antes e o depois é estatisticamente significativa.

Se a diferença é provavelmente real, então poderemos tentar responder se a diferença é importante. Por exemplo, o uso de um sistema de apoio a reunião reduz 20 minutos de uma reunião que normalmente dura 2 horas, e esta diferença é real (no sentido da primeira questão). Parece que a diferença de 20 minutos "vale a pena", já que a duração da reunião foi reduzida em 1/6. Mas se a diferença for de apenas 2 minutos, ainda que seja uma diferença real, esta diferença não parece ser útil na prática. Não há um padrão muito estabelecido para dizer se a diferença é "útil na prática", mas a boa notícia é que em pesquisa científica geralmente não é preciso determinar o grau da utilidade prática! A grande maioria das pesquisas científicas, com exceção de algumas pesquisas em Medicina, geralmente é finalizada com a obtenção da resposta para a primeira questão.

A terceira questão é se os ganhos da intervenção compensam os custos. Esta é uma análise de custo/benefício e geralmente é o que motiva a realização da pesquisa. Também não abordaremos essa análise.

O objetivo desta seção é discutir os conceitos necessários para responder a primeira questão. Se o pesquisador for inexperiente com o uso dos procedimentos estatísticos descritos nas próximas subseções, deve procurar uma pessoa experiente para apoiá-lo na análise dos dados coletados no experimento.

24.6.1. Tipos de medida

Em pesquisa quantitativa, assume-se que as variáveis de interesse são medidas objetivas: tempo de execução, respostas corretas, quantidade de mensagens enviadas etc. As medidas são classificadas nos seguintes tipos:

- Medida categórica ou nominal. Um dado é classificado em uma categoria. Uma medida categórica clássica é sexo: masculino ou feminino. A única operação possível é verificar se o dado tem um ou outro valor; não é possível ordená-los nem fazer operações matemáticas. Mesmo que se codifique 1 para o sexo masculino e 2 para o feminino, não faz sentido "somar" o sexo de um grupo de pessoas, ou tirar a média do sexo. A codificação das categorias em números deve ser entendida com cuidado. Por exemplo, com relação aos estados brasileiros, se atribuirmos 1 para Acre, 2 para Alagoas, 3 para Amazonas, e assim por diante, não faz sentido somarmos esses números, subtrair um do outro, ou dizer que Piauí é maior do que Pará só porque seu código é um número maior.

- Medidas ordinais. Medidas ordinais também atribuem classes aos dados, mas é possível ordená-los. Um exemplo típico é classe socioeconômica. Geralmente são definidas as classes A, B, C e D, sendo a classe A com mais poder aquisitivo e a classe D com menos poder aquisitivo. Outras variáveis ordinais são: nível educacional, dificuldade de um projeto de software, nível de maturidade dos processos de software de uma organização, entre outros exemplos. Se os valores ordinais são codificados com números, como no exemplo de classe socioeconômica A = 4, B = 3, C = 2 e D = 1, então a ordem dos números reflete a ordem dos valores, mas a diferença entre os números-dos-códigos não faz sentido: a distância entre a classe A e B não é a mesma que a distância entre as classes B e C.

- Medidas intervalares. Medidas intervalares atribuem ao dado um número real, mas o zero da escala é arbitrário. O exemplo clássico de medida intervalar é a temperatura em graus Celsius. Medidas intervalares garantem que as diferenças entre duas medidas (o intervalo) é algo que pode ser comparado: pode-se dizer que o ganho de temperatura de 20ºC para 30ºC é duas vezes maior do que quando de 10ºC para 15ºC. Contudo, não se pode dizer que um dado é múltiplo do outro, por exemplo, é errado dizer que a 20ºC está duas vezes mais quente do que em 10ºC.

- Medidas de razão. Medidas de razão atribuem ao dado um número real onde o zero é absoluto e, portanto, a razão entre duas medidas faz sentido. Exemplos: tempo transcorrido, distância, idade, altura etc. Nestes casos, pode-se dizer que 20 é o dobro de 10 (unidades arbitrárias).

Em computação não há muitos exemplos onde a diferença entre medidas de razão e intervalares sejam relevantes. O importante é compreender as diferenças entre as medidas categórica, ordinal e as demais (intervalar e de razão). Como detalhado nas próximas subseções, a classificação das medidas é importante porque define: que tipo de estatística deve ser usada para sumarizar os dados, e que tipo de teste estatístico deverá ser usado para verificar se dois conjuntos de dados são significativamente diferentes.

24.6.2. Sumarização dos dados (estatística descritiva)

Dado um conjunto de medidas categóricas, os dados são descritos por meio da distribuição de frequências: 2% dos produtos vieram do Acre, 14% de Alagoas, 13% do Amazonas, e assim por diante. Para sumarizar os dados, apresenta-se a moda, isto é, o valor mais frequente.

Para medidas ordinais, a medida sumarizadora mais empregada é a mediana, o valor que divide o conjunto de dados em duas metades. Para a descrição dos dados, também deve ser apresentada a frequência de cada um dos valores.

As escalas intervalares e de razão são sumarizadas através da média e do desvio padrão. Medidas de razão também permitem média geométrica e média harmônica que não fazem sentido para medidas intervalares.

Um procedimento comum em computação é verificar se um programa produz o resultado correto para alguns dados de entrada. O pesquisador deve considerar essa medida como categórica. Mesmo usando a codificação tradicional de 0 para falha e 1 para sucesso, o pesquisador não deve pensar que está trabalhando com uma medida de razão. A origem do engano é porque falsamente parece possível realizar contas como a média: se o programa acertou 30 e errou 12, dizer que o programa acerta 30/(30+12) = 71,4% até parece fazer sentido. O que de fato está acontecendo é que o número 71,4% é a descrição da distribuição dos valores: certo em 71,4% das vezes, e errado em 28,6% das vezes. Já que essa medida é categórica, para comparar dois programas quanto à corretude, conforme será visto nas próximas subseções, deve-se usar o teste Chi-quadrado porque é o teste indicado para medidas categóricas, e não se deve usar o teste t que é para medidas intervalar e de razão.

24.6.3. Da amostra para a população (estatística inferencial)

Enquanto a estatística descritiva abrange métodos para descrever os dados, a estatística inferencial, que nos interessa nesta seção, abrange os métodos para fazer afirmações sobre a população (conjunto de dados) a partir de medidas sobre uma amostra (subconjunto de dados tirado da população). Na grande maioria das vezes, a população é um conjunto infinito, abstrato, e desconhecido; já a amostra é um conjunto finito de dados concretamente coletados. Considere, por exemplo, que o objetivo é descobrir a média da população dada a média e outras informações de uma amostra (o que requer que a medida seja intervalar ou de razão). A média ou outra medida da população é chamada de parâmetro, enquanto qualquer medida da amostra é chamada de estimativa. A estatística inferencial faz previsões sobre os parâmetros da população a partir das estimativas da amostra.

Os métodos da estatística inferencial exigem que a amostra seja aleatoriamente escolhida da população, ou mais precisamente, que cada elemento da amostra tenha sido escolhido

aleatoriamente da população e que a escolha de um elemento da amostra não influencia ou é influenciada pela escolha do próximo elemento.

O principal fenômeno a ser entendido é o erro amostral, o fato de que o valor de uma estimativa na amostra quase nunca vai ser exatamente igual ao valor do parâmetro correspondente. Por exemplo, suponha que um programa estatístico gere uma população de 1000 valores aleatórios entre 0 e 100. Este é um exemplo bastante incomum porque a população é finita e conhecida. Neste caso, podemos calcular a média da população somando-se todos os valores e dividindo por 1000 – o resultado deverá ser próximo de 50. Uma amostra pode ser obtida selecionando um subconjunto dessa população. Digamos que nossa amostra seja formada pelos elementos entre as posições 40 a 45, inclusive. A média dessa amostra de apenas 6 elementos pode ser bastante diferente da média da população. Essa diferença é o erro amostral.

O erro amostral para amostras de tamanho n tem uma distribuição conhecida: uma gaussiana (distribuição normal) de média zero e desvio padrão igual a $\sigma / \sqrt{n}$, onde σ é o desvio padrão da população. Vamos ilustrar essa distribuição. Usando um programa estatístico, a partir de uma população de 1000 valores aleatórios entre 0 e 100, geramos um conjunto de 300 amostras, cada amostra com 8 elementos aleatórios da população. Foi calculada a média de cada amostra e calculado o erro amostral. Na Figura 24.2 é apresentado o histograma do erro amostral do nosso conjunto de amostras, e a gaussiana que melhor se ajusta ao histograma.

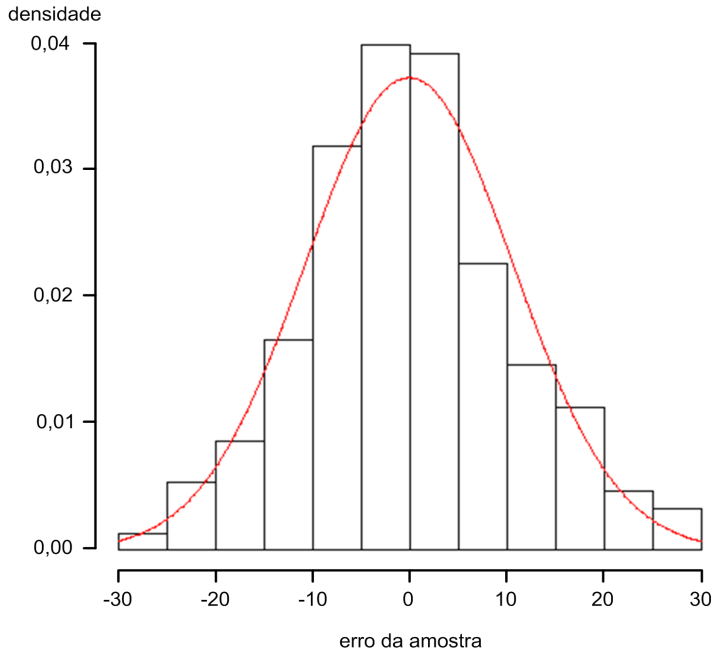


Figura 24.2. Histograma do erro amostral para amostras com 8 elementos da população descrita no texto

Dado a média da amostra, podemos começar a calcular a média da população. O erro amostral mais provável é 0, que é a média da distribuição do erro amostral. Portanto, o valor mais provável para a média da população é o próprio valor da média da amostra. Mas temos quase certeza de que a média da população não vai ser exatamente igual a essa estimativa, mas se fi-

xarmos alguma probabilidade, digamos 90%, podemos dizer que com 90% de probabilidade a média da população vai ser a média da amostra mais ou menos um certo valor *x*. Esse valor *x* é o valor na distribuição do erro amostral tal que a área da gaussiana de 0 até *x* é igual a 0,45 (a metade de 0,9). Infelizmente, para terminar essa conta precisamos do desvio padrão da população, que nós não temos, mas também podemos aproximá-lo usando o desvio padrão da amostra. A conta final fica mais complexa, pois também há um erro na nossa estimativa do desvio padrão e, portanto, a distribuição resultante não mais é uma gaussiana, mas sim uma distribuição parecida com a normal, chamada t de Student.

Essa probabilidade 90% é chamada de grau de confiança. Dado um grau de confiança de 90% podemos calcular um intervalo em torno da média da amostra de tal forma que a média da população estará dentro do intervalo com aquela probabilidade. Esse intervalo é conhecido como intervalo de confiança. O número 90% é arbitrário, poderíamos escolher qualquer grau de confiança, mas tradicionalmente as ciências escolhem apenas 90%, 95%, e 99% como os graus de confiança usados, e o grau de confiança de 95% é o mais frequente.

Dado um grau de confiança g, o número $\alpha=1-g$ é conhecido como erro do tipo I, também denominado Falso Positivo, que consiste em chegar a um resultado significativo estatisticamente quando de fato aconteceu por acaso. Por isso α deve ser baixo, para que se tenha baixa probabilidade de concluir erradamente sobre a hipótese testada no experimento.

OUTRAS MEDIDAS ALÉM DA MÉDIA DA POPULAÇÃO A PARTIR DE UMA AMOSTRA

A ideia do intervalo de confiança é aplicada a várias medidas mais complexas derivadas de amostras. Por exemplo, uma amostra que contenha pares de pontos, calcula-se o intervalo de confiança para o parâmetro correlação. Mas há medidas para as quais o cálculo exato do intervalo de confiança é complexo ou impossível, por exemplo, a mediana.

Existe uma técnica chamada de bootstrap (Efron, 2003) que é capaz de gerar aproximações de intervalos de confiança para quase qualquer medida de uma amostra. Se o pesquisador precisa gerar intervalos de confiança para medidas complexas de uma amostra deve considerar utilizar essa técnica.

MEDIDAS DA POPULAÇÃO A PARTIR DE DADOS CATEGÓRICOS OU ORDINAIS

Da mesma forma com que calculamos um intervalo de confiança para a média, podemos calcular intervalos de confiança para a proporção de um valor categórico ou ordinal na amostra. Se a amostra é de medidas categóricas, e 32% delas tem valor "A", pode-se calcular um intervalo de confiança para essa proporção. O método exato para calcular o intervalo chama-se o método de Clopper-Pearson. Mas há métodos aproximados que funcionam de forma diferente para diferentes condições. O método mais popular é o método da normal (ou de Wald), que funciona se nenhuma das proporções for perto de 0 ou 1 (e com outras condições). Talvez o método mais recomendado seja o método de Agresti-Coull, também chamado de método de Wald modificado.

24.6.4. Testes de diferença

Agora que conhecemos o conceito de erro amostral, podemos falar de testes de significância para a diferença. Vamos supor 2 conjuntos de dados, e a média do primeiro é maior do que a média do segundo. Esses dois conjuntos são amostras, e queremos fazer afirmações sobre as médias das respectivas populações. Mas mesmo que as médias dos dois conjuntos sejam diferentes, as amostras podem ser da mesma população e a diferença é devido apenas ao erro amostral. Testes de diferença fazem essa verificação. Testes estatísticos calculam a probabilidade de que as diferenças encontradas nas médias de dois conjuntos sejam devido apenas ao erro amostral. Se esta probabilidade é maior do que um valor mínimo (que é 1 menos o grau de confiança), então diremos que a diferença entre os conjuntos não é estatisticamente significativa.

Testes de diferença (esse nome não é padrão) é um de vários tipos de testes estatísticos. Um teste estatístico tira conclusões sobre a população a partir das amostras. A conclusão é baseada na aceitação ou rejeição de uma hipótese. Devem ser definidas duas hipóteses: a hipótese nula, denominada H_0, e a hipótese alternativa, denominada H_1. A hipótese nula é a hipótese do "não efeito", é formulada com o objetivo de ser rejeitada, é a negação do que se está tentando afirmar (H_1). Para o teste de diferença significativa, são definidas as hipóteses:

- H_0: as duas amostras são de uma mesma população e as diferenças entre os dois conjuntos de medidas é apenas devido ao erro de amostragem.
- H_1: As amostras não são da mesma população, há uma diferença significativa global entre as observações das amostras.

A decisão sobre aceitar ou rejeitar a hipótese alternativa é baseada na probabilidade da hipótese nula ser verdadeira. Deve-se definir o nível de significância α, que é o valor máximo de probabilidade aceitável para rejeitar a hipótese nula (quanto menor o valor α, maior a certeza da decisão). Além da hipótese nula, também são feitas outras pressuposições sobre os dados que chamaremos de condições do teste. Pode-se calcular a probabilidade de que uma propriedade relacionada aos dados seja verdadeira, dadas as suas pressuposições. Essa probabilidade calculada é chamada valor p ou p-valor.

A hipótese nula para o teste de diferença é que os dados de dois conjuntos *C1* e *C2* vieram de uma mesma população e portanto a diferença entre a média de *C1* e a de *C2* é apenas devido ao erro de amostragem. É possível calcular qual a probabilidade que o erro de amostragem seja igual ou maior que a diferença das médias de *C1* e *C2*. Esta probabilidade é o p-valor. Se o p-valor é suficientemente baixo, então é muito improvável que a diferença seja apenas devido ao erro de amostragem, e portanto é improvável que os dois conjuntos sejam amostras de uma mesma população. Se o grau de confiança desejado é de 95%, então qualquer p-valor menor que 0,05 (α) é suficiente para rejeitar a hipótese nula, e afirma-se que a diferença entre as médias é estatisticamente significativa. Por outro lado, valores de p-valor acima do α implicam que a diferença não é estatisticamente significativa, e que a hipótese nula não pode ser descartada.

24.6.5. Seleção do teste adequado

Existem diferentes testes estatísticos para a comparação de dois conjuntos de dados. Cada teste estatístico estima o p-valor baseado em algumas pressuposições sobre a população (condições do teste), principalmente, sobre os conjuntos *C1* e *C2*.

Dados dois testes estatísticos, T1 e T2, diremos que T1 é mais forte do que T2 se para os mesmos dados o p-valor calculado por T1 é menor do que o de T2. Em geral, um teste é mais forte do que outro se faz mais pressuposições sobre os dados, e em alguns casos, um teste mais forte pode definir que a hipótese nula é falsa enquanto um teste mais fraco não possibilita tal conclusão.

Um teste é dito não paramétrico se, dentre suas condições de teste, não há qualquer pressuposição de que os dados têm alguma distribuição fixa. Exatamente por assumir menos pressuposições sobre os dados, testes não paramétricos são mais fracos do que seus correspondentes paramétricos – Tabela 24.1. Os testes Z e T são paramétricos. Se os dados satisfazem as condições dos testes paramétricos, então estes testes devem ser os escolhidos, caso contrário é preciso usar os testes não paramétricos. Para dados ordinais, o teste não paramétrico é o teste U, e para dados categóricos, os testes Chi-quadrado e exato de Fisher.

Tabela 24.1. Testes para comparação de 2 conjuntos de dados

	TESTES NÃO PAREADOS	TESTES PAREADOS	FORÇA
Testes paramétricos	teste Z	teste Z pareado	+ forte
	teste T	teste T pareado	
Testes não paramétricos	teste U de Mann-Whitney ou teste Wilcoxon rank-sum	teste Wilcoxon signed-rank	
	teste do Chi-quadrado	–	
	teste exato de Fischer	–	+ fraco

Outro componente importante na escolha do teste é se os dados nos dois conjuntos são pareados. Dados são pareados se cada medida de um dos conjuntos pode ser colocada em correspondência com uma medida do outro conjunto. Por exemplo, um primeiro conjunto que indica as notas dos alunos de uma classe na primeira prova é pareável com um segundo que indica a nota dos mesmos alunos na segunda prova. Outro caso: o primeiro conjunto são os tempos de execução do programa P1 para certas entradas, e o segundo conjunto são os tempos de execução do programa P2 para as mesmas entradas. Testes pareados são sempre mais fortes do que seus correspondentes não pareados. Os testes do Chi-quadrado e o teste exato de Fischer não tem versões pareadas.

O componente final na escolha do teste estatístico é se o teste deve ser usado na sua versão bicaudal ou unicaudal. Nós temos usado o termo teste de diferença para verificar se a diferença entre a média de dois conjuntos é significativamente diferente entre si. Este é o teste bicaudal, porque a média de C_1 pode ser tanto maior como menor do que a média de C_2 e, portanto, é preciso testar as duas hipóteses. Se por alguma razão sabe-se que a média de C_1

não pode ser menor que a média de C_2, então o teste de diferença pode apenas verificar se a média de C_1 é significativamente maior que a de C_2 sem testar a hipótese dela ser menor. Este é o teste unicaudal – só testa-se uma das alternativas possíveis. Testes unicaudais são sempre mais fortes que suas versões correspondentes bicaudais.

Como regra geral, deve-se sempre usar o teste mais forte possível, desde que suas condições sejam satisfeitas. A exceção é que se deve usar os testes bicaudais em vez dos unicaudais, a não ser em casos onde já se sabe, por razões teóricas, que uma das médias não pode ser menor do que a outra.

Vejamos em mais detalhes o teste T, que é um dos testes mais usados na prática. O teste T tem as seguintes pressuposições para calcular o p-valor:

- os dados de cada um dos conjuntos são independentes;
- os dados dos dois conjuntos são medidas intervalares ou de razão;
- os dois conjuntos têm distribuições normais; e
- os dois conjuntos têm o mesmo desvio padrão.

Nem sempre é simples ou mesmo recomendado verificar todas essas condições. É impossível verificar estatisticamente o primeiro critério, que os dados dos conjuntos são amostrados independentemente. Esta condição só é verificada pelo desenho experimental e pelos procedimentos de coleta. Então, na prática, este critério nem é discutido, embora seja o mais importante. O segundo critério, que os dados sejam pelo menos intervalares, é facilmente verificável. O terceiro critério, que os dados dos conjuntos estejam distribuídos segundo uma normal, é provavelmente o menos relevante. Há testes para verificar se um conjunto de dados segue a distribuição normal. Dois dos mais famosos são o teste de Shapiro-Wilk e o de Kolmogorov-Smirnov, e destes, o primeiro deve ser usado. Estudos recentes indicam que se o número de dados das amostras é pequeno, tais testes não são muito precisos para detectar a não normalidade, e por outro lado, se o número de dados for grande, o próprio requisito de normalidade dos dados para o teste T não é importante. Portanto, como regra geral, verificar este critério não é necessário e se reduz a verificar o número de dados em cada amostra. Se ambas as amostras têm mais do que 20 dados (segundo alguns guias de pacotes estatísticos), este critério não precisa ser verificado. Finalmente o quarto critério, igual variança entre os dois conjuntos, pode ser desconsiderado se em vez de usar a versão tradicional do teste T, for usada a modificação de Welch do teste T. A modificação de Welch torna o teste T um pouco mais fraco.

24.6.6. Teste de equivalência

Em tecnologia, de vez em quando temos que mostrar que duas soluções são equivalentes em algum aspecto central, e provavelmente diferentes em algum aspecto menos central. Por exemplo, ao desenvolvedor um sistema de videoconferência, deseja-se mostrar que os resultados de uma reunião presencial e de uma reunião a distancia são equivalentes no que diz respeito à qualidade ou ao tempo da reunião (aspecto central), mas a reunião a distancia é mais conveniente ou barata (aspecto menos central). Em geral, usa-se um desenho com dois grupos para verificar a equivalência.

Um exemplo clássico, não em computação, são os medicamentos genéricos. Deve-se mostrar que uma droga tem o mesmo efeito das drogas originais, mas assume-se que elas serão mais baratas. Testes de equivalência são tão importantes em farmacologia e áreas afins que a maioria dos textos que explicam tais testes são desta área, e algumas vezes os testes são até chamados de testes de bioequivalência.

TESTE PARA DEMONSTRAR MÉDIAS IGUAIS

Para mostrar que dois conjuntos são equivalentes, deve-se usar o TOST (two one-sided test). Informações sobre o teste TOST podem ser consultadas em Graphpad (2010).

Mostrar que dois conjuntos têm a mesma média não é a mesma coisa que mostrar que a diferença não é estatisticamente significativa! Esse é um erro muito comum, e é discutido na literatura sob diferentes nomes: "provar a hipótese nula", ou "ausência de evidência não é o mesmo que evidência de ausência (de diferença)".

24.6.7. Análise dos resultados em função do desenho experimental

Agora podemos discutir a análise estatística apropriada para cada desenho experimental. A primeira questão apresentada no início desta seção – se a diferença entre os resultados é real – consiste em verificar se a diferenças entre dois conjuntos de medidas é significativamente diferente.

Para os desenhos experimentais envolvendo um grupo – "antes-depois", "um grupo pré e pós-teste com remoção da intervenção" e "um grupo, intervenção repetida" –, a verificação se $\mu(O_2) >> \mu(O_1)$ ou equivalente, é a verificação se a diferença entre os conjuntos de medidas O_2 e O_1 é estatisticamente significativa usando os testes apropriados da Tabela 24.1, dependendo do tipo de medida dos testes e se satisfazem os requisitos apropriados de cada teste. Em particular, se o mesmo grupo de pessoas fez o pré e o pós-teste, então devem ser usados os testes pareados.

No caso do desenho experimental envolvendo dois grupos só com pós-teste, a análise se $\mu(O) >> \mu(O')$ também é feita usando os testes da Tabela 1.

Para os desenhos com "variável dependente não equivalente" e "dois grupos, pré e pós-testes" requer mais elaboração. Para o desenho com variáveis não equivalentes, uma primeira análise apenas mostraria que a diferença entre os valores pré e pós da variável não equivalente não é significativa. Isto é, numa abordagem menos rigorosa mas ainda aceitável, mostrar que $\mu(V_2) \approx \mu(V_1)$ é apenas mostrar que não é verdade que $\mu(V_2) >> \mu(V_1)$ – lembre-se da discussão sobre testes de equivalência da seção anterior: mostrar que a diferença não é significativa não é a

ANÁLISE DA VARIÁVEL DEPENDENTE NÃO EQUIVALENTE

Uma análise mais elaborada da variável não equivalente pode ser encontrada em Reynolds (1987). Resumidamente: as variáveis O_1 e V_1 são convertidas em variáveis estandardizadas (média = 0 e desvio padrão = 1). A conversão também é aplicada às variáveis O_2 e V_2, mas subtraem a média de O_1 e V_1, respectivamente, e dividem pelos desvios padrão das variáveis do pré-teste. Após as conversões, os ganhos na variável V e O podem ser comparados entre si, e os autores mostram que o ganho em O é significativamente maior do que o ganho em V.

mesma coisa que mostrar que não há diferença. Numa abordagem mais rigorosa, deve-se usar o TOST para mostrar que $\mu (V_2) \approx \mu (V_1)$.

Para o desenho "dois grupos, pré e pós-testes", que é o desenho mais complexo na nossa lista, a literatura recomenda três formas de análise:

- análise do ganho
- análise de covariância (ANCOVA)
- análise de covariância com correção de confiabilidade (reliability)

A análise de ganho assume que o pré-teste e o pós-teste são da mesma natureza, e é feita a subtração dos dois valores obtidos para cada sujeito. Se o sujeito *i* obteve o_{1i} no pré-teste e o_{2i} no pós-teste, então o ganho da pessoa *i* é $\delta i = o_{2i} - o_{1i}$. O conjunto de ganhos para o grupo experimental Δ deve ser comparado aos ganhos do grupo de controle Δ' e diremos que a intervenção foi positiva se:

$$\mu (\Delta) >> \mu (\Delta')\Delta$$

TAMANHO DO EFEITO PARA ANALISAR A IMPORTÂNCIA PRÁTICA

Uma vez determinado que a diferença mesurada é real, precisamos determinar se a diferença é importante na prática. Por exemplo, suponha que num desenho experimental "dois grupos, só pós-teste" a diferença do grupo de controle e do experimental seja estatisticamente significativa e na média seja igual a 5 minutos de diferença para a resolução de um problema com e sem um sistema colaborativo - isto é relevante na prática? Para começar a determinar essa relevância, desenvolveu-se uma medida chamada de tamanho do efeito (effect size).

É importante notar que a relevância da redução de 5 minutos depende do tempo total médio para resolver o problema. Uma diferença de 5 minutos num problema cuja solução demora 20 minutos parece muito mais importante do que a mesma diferença de 5 minutos se o problema demora 5 horas para ser resolvido. Assim, o ganho relativo (e não o ganho absoluto) é uma medida melhor da importância do efeito. Mas uma medida ainda mais importante é o ganho dividido por alguma medida similar ao desvio padrão dos tempos. Se há muita variação nos tempos, por exemplo, por volta de 50 minutos, então o ganho de 5 minutos é pequeno. Se por outro lado só há uma variação de 10 minutos entre as medidas, o ganho de 5 minutos já é muito mais interessante.

Há várias formas de calcular o tamanho do efeito, dependendo principalmente de como se calcula o denominador (a medida similar ao desvio padrão das medidas), tais como o tamanho de efeito definido por Cohen (chamado de d), por Hedges (g) e a medida delta de Glass. Cohen também definiu o tamanho de efeito para que o ganho seja considerado importante. A definição é arbitraria, mas amplamente aceita. Segundo Cohen, um tamanho de efeito até 0,2 é considerado baixo, de 0,2 a 0,8 é considerado médio, e acima de 0,8, grande. Na falta de melhores métodos para avaliar a importância do efeito da intervenção, na prática pode-se usar estas definições de Cohen.

Note que para fazer esta análise é preciso que os mesmos sujeitos tenham feito tanto o pré-teste quanto o pós-teste em cada um dos grupos, pois é preciso fazer a subtração para cada sujeito.

Por outro lado, pode ser que o pré-teste e o pós-teste não sejam de mesma natureza. Por exemplo, pode-se achar que a experiência prévia de pessoas em projetos é determinante na velocidade de resolver um problema usando um sistema colaborativo. Assim, a medida do pré-teste pode ser anos de trabalho ou número de projetos já executados, e a medida do pós-teste pode ser tempo para resolver o problema. Diz-se que é uma medida substituta de pré-teste e obviamente não faz sentido fazer a subtração com o pós-teste. Nestes casos utiliza-se a análise por ANCOVA, que pode ser consultada em Trochim (2006), onde também se encontra discutida a análise da correção de confiabilidade no teste ANCOVA.

24.7. Planejamento de experimentos

Um experimento que envolve seres humanos necessita de um planejamento cuidadoso dado o investimento de tempo e talvez de dinheiro para a realização. O planejamento de um experimento envolve:

- definir os instrumentos de medida dos pré e pós-teste;
- definir o desenho experimental;
- definir o número de sujeitos.

O primeiro passo de um experimento é definir qual é o efeito que se quer medir e como medi-lo. Se é esperado que o sistema melhore a velocidade de resolução de problemas, então deve-se medir esse tempo. Se é esperado que o sistema melhore a satisfação dos usuários, então deve-se elaborar questionários que meçam a satisfação. Com a definição da medida também se obtém uma primeira avaliação da duração do experimento. Por exemplo, se o pré e o pós-testes são o tempo de resolução de um problema que demora semanas para ser resolvido, então o experimento como um todo vai demorar pelo menos o dobro deste tempo mais o tempo para treinar os usuários no uso do sistema.

Uma vez definido o instrumento de medida, deve-se definir o desenho experimental. Uma interação importante do instrumento com o desenho é se há as ameaças de testagem e instrumentação. Se essas ameaças são importantes, então deve-se escolher um desenho sem pré-teste. A segunda interação importante é o tempo. Se o experimento for longo, então história, maturação, contaminação e outras ameaças são mais relevantes do que se o experimento for curto, e portanto o desenho deverá evitar essas ameaças. Uma regra genérica é fazer o desenho o mais rigoroso possível, onde rigoroso significa que o desenho, por si só, evita o maior número de ameaças. Contudo, nem sempre é possível usar um desenho rigoroso, e nem todas as ameaças potencialmente presentes em um desenho experimental são importantes em todas as pesquisas. O pesquisador deve listar as ameaças relevantes ao experimento, e usar um desenho que as evite.

O próximo passo é definir o protocolo de execução do experimento e o método de atribuição dos participantes aos grupos (se tiver sido projetado um experimento com grupo de controle). Com estas duas definições devem ser eliminadas as ameaças importantes que não tiverem sido evitadas pelo desenho experimental.

O último passo é a determinação do número de sujeitos. O objetivo é ter o menor número de sujeitos que garanta a significância estatística das diferenças encontradas. Não há muita teoria para a determinação do tamanho da amostra (sample size) para a maioria dos desenhos apresentados. O que é conhecido é uma formula para calcular o número de elementos em cada conjunto para que a diferença das médias seja estatisticamente significativa, dado:

- a diferença das médias dos dois conjuntos (δ);
- o desvio padrão (suposto igual) dos dois conjuntos (σ);
- o grau de confiança;
- e uma medida similar ao grau de confiança chamada poder (power).

A fórmula para calcular o número de sujeitos em cada grupo (que assumimos ser igual) é complicada, mas existe uma simplificação da fórmula para os valores mais comuns do grau de confiança e da medida de power. Esta formula é conhecida como a equação de Lehr:

$$n = 16 / \Delta \quad \text{onde} \quad \Delta = \delta / \sigma \Delta$$

Note que Δ é o tamanho do efeito (apresentado num quadro da seção anterior). Contudo, a diferença da média e o desvio padrão são obtidos somente após a realização do experimento, então como usá-los no planejamento? Há duas soluções e uma terceira solução parcial para esse problema.

A primeira solução real é obter tanto a diferença das médias e o desvio padrão de dados da literatura. Se alguém já fez um experimento similar, os resultados daquele experimento podem ser usados como dados para o cálculo do tamanho da amostra. Mesmo que o experimento da literatura use medidas diferentes das que você planeja no seu experimento, o cálculo do tamanho do efeito é considerado uma medida que pode ser transferida de um experimento para outro. Portanto, pode-se usar o tamanho do efeito dos experimentos da literatura para calcular o tamanho da amostra para o novo experimento.

A segunda solução é fazer um experimento piloto cujo objetivo não é obter certeza estatística na diferença das médias, mas sim obter os valores para essa diferença e para o desvio padrão. Com os dados do experimento piloto, pode-se calcular o número de sujeitos do experimento "pra valer".

A solução parcial é assumir algum valor para o tamanho do efeito no seu experimento. A literatura afirma que tamanhos de efeito menores que 0.2 ou 0.3 não são relevantes na prática. Então use 0.3 nos cálculos para o número de sujeitos – se o tamanho de efeito obtido no experimento for maior, então o número de sujeitos será provavelmente suficiente para garantir a significância estatística da diferença, e se for menor, provavelmente a diferença obtida não é importante o suficiente. Assim, o número mínimo de sujeitos em cada grupo deve ser 16/0.3 = 54. Se este número parece excessivo, lembre-se que ele é o número de sujeitos por grupo necessário para detectar um fenômeno muito sutil, no limite do que seria algo que "vale a pena na prática".

Se ao final do experimento não for encontrada uma diferença estatisticamente significativa que indique que a intervenção tem o efeito desejado, ainda assim será importante publicar o resultado, pois a publicação do tamanho do efeito (ou da diferença das médias e do desvio padrão) é um dado importante para a comunidade planejar o próximo experimento.

Esperamos que este capítulo seja útil para apoiá-lo a realizar experimentos com sistemas colaborativos. O objetivo não é esgotar o assunto, mas sim apresentar uma visão geral dos conceitos necessários para realizar um experimento nessa área. Para aprofundar o seu conhecimento, estude textos específicos como os listados nas Leituras Recomendadas.

EXERCÍCIOS

24.1. A presença de uma ameaça à validade interna não depende apenas do desenho, mas também das condições particulares da execução do experimento. Por exemplo, suponha um experimento conduzido com um grupo durante um só dia, com os sujeitos dentro de um laboratório fechado. Quais as ameaças que essas particularidades eliminam? E que ameaças são mais relevantes?

E se o experimento for conduzido:

- com 2 grupos, experimental e controle, em laboratórios diferentes, sem que os sujeitos possam sair do laboratório durante o experimento?
- em uma empresa real, conduzindo seus negócios habituais, durante uma semana?
- em uma empresa real, conduzindo seus negócios habituais, durante vários meses?

24.2. Para cada um dos desenhos experimentais listados na Seção 24.3, indique quais são as possíveis ameaças à validade interna. Assuma o pior caso em que as condições particulares do experimento não eliminam nenhuma das ameaças.

24.3. Se 6 pessoas vão ser divididas em dois grupos iguais, qual a probabilidade de duas pessoas específicas (por exemplo "as melhores") caírem no mesmo grupo? E a probabilidade delas caírem no grupo experimental?

24.4. Dado o conjunto de dados

{99, 75, 31, 85, 7, 78, 56, 83, 61, 11, 78, 94, 57, 20, 6, 17, 8, 98, 62, 82}

Usando algum pacote estatístico, ou usando as formulas disponíveis em um livro texto de estatística, calcule o intervalo de 95% de confiança para a média.

24.5. Suponha um conjunto de execuções de um programa com diferentes dados: 35 deram certo e 12 deram erro. Discuta qual é a amostra e qual é a população. Qual é o intervalo de 95% de confiança para a taxa de acerto do programa, usando tanto Wald como o Wald modificado.

24.6. Suponha os dados da questão 4 e o seguinte conjunto de medidas:

{55, 109, 118, 113, 79, 35, 94, 54, 87, 58, 33, 115, 68, 58, 45, 37, 110, 44, 59, 63}

Usando algum pacote estatístico, ou usando as fórmulas disponíveis em um livro texto de estatística, calcule o p-valor para:

- teste t
- teste t unicaudal
- teste t pareado

- teste de Wilcoxon rank sum
- teste de Wilcoxon signed rank (pareado)

Qual a ordem parcial de força dos testes? O p-valor calculado é compatível com essa ordem?

LEITURAS RECOMENDADAS

- Métodos de pesquisa quantitativa e qualitativa para a ciência computação (Wainer, 2007). O presente capítulo se baseou num texto que escrevi para um curso sobre pesquisas em computação. Sempre que me perguntam sobre um bom livro de estatística, digo que ainda não descobri o meu favorito. A maioria dos livros não cobre testes estatísticos como eu acho que deva ser apresentado.
- Experimental and quasi-experimental designs for generalized causal inference (Shadish *et al.*, 2002). É um bom livro sobre desenhos experimentais.
- Research Methods Knowledge Base <http://www.socialresearchmethods.net/kb> É um excelente site sobre desenhos experimentais.
- Intuitive Biostatistics: Choosing a statistical test <http://www.graphpad.com/www/book/choose.htm> Site que discute qual teste estatístico usar em cada situação.

REFERÊNCIAS

EFRON, B., TIBSHIRANI, R. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

GRAPHPAD, http://www.graphpad.com/library/BiostatsSpecial/article_182.htm (acessado em 9/2010)

HOCHMAN, B., NAHAS, F. X., OLIVEIRA, R. S., FERREIRA, L. M. Desenho de pesquisa. *Acta Cirurgica Brasileira*, 20(2), 2005.

REYNOLDS, K. D., WEST, S. G. A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11(6):691–714, 1987.

SHADISH, W. R., COOK, T. D., CAMPBELL, D. T. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Co, 2002.

TROCHIM, W. M.K http://www.socialresearchmethods.net/kb/statnegd.php 2006 (acessado em 9/2010).

WAINER, J. Métodos de pesquisa quantitativa e qualitativa para a ciência computação. In: *Atualização em Informática*, KIWALTOWSKI, Tomasz, BREITMAN, Karin. (Org.), Sociedade Brasileira de Computação e Editora PUC-Rio, 2007.

WYATT, J. Quantitative evaluation of clinical software, exemplified by decision support systems. *International Journal of Medical Informatics*, 47(3):165–173, 1998.