Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação

Jacques Wainer

Instituto de Computação – UNICAMP
wainer@ic.unicamp.br

Abstract

This course will discuss some of the scientific research methods relevant to Computer Science. We will discuss many aspects of both quantitative and qualitative methods. These methods are relevant to evaluate systems, to verify theories, and to understand working practices that may illuminate the specification of systems. Among the quantitative methods we will discuss the evaluation of programs using synthetic data, significance tests, surveys, and design of experiments. Among the qualitative methods we will discuss observational research (case studies, ethnography, and so on) and action research. Whenever possible we point out published research that used these methods.

Resumo

Este minicurso abordará métodos de pesquisa científica apropriados para a Ciência da Computação. Abordaremos vários aspectos de métodos quantitativos e qualitativos. Estes métodos são apropriados para a avaliação de sistemas, para a verificação de teorias, e para o entendimento de práticas de trabalho que podem iluminar a elaboração de requisitos de sistemas. Dentre os métodos quantitativos abordaremos avaliação de programas usando dados artificiais, testes de significância estatística e o uso de questionários e experimentos. Dentre os métodos qualitativos, abordaremos variações em pesquisa observacional (estudo de caso, etnografia, etc.) e pesquisa-ação. Sempre que possível apontamos pesquisas publicadas onde os métodos são utilizados.

5.1. Metodologias de pesquisa científica em Ciência da Computação

Pesquisa em Ciência da Computação (abreviado como CC, doravante) envolve na maioria dos casos a construção de um programa, de um modelo, de um algoritmo ou de um sistema *novo*. Novidade é considerada como algo fundamental da pesquisa em CC. De vez em quando (dependendo da subárea e da sorte) apenas a apresentação do programa/modelo/sistema novo é considerado como a pesquisa em si, e há vários exemplos de artigos em revistas e

conferências onde um sistema/modelo/algoritmo novo é apresentado e comparado informalmente com as alternativas já publicadas.

Este texto usará o termo *programa* para indicar um software que resolve um problema específico, e cujo usuário é normalmente o próprio autor. Um *sistema* é um software que interage com usuários que não são o desenvolvedor do sistema, e que faz parte das atividades do usuário.

O autor acredita que cada vez mais as revistas de qualidade em Ciência da Computação vão exigir avaliação dos sistemas e programas criados, e que esta avaliação deverá ser cada vez mais rigorosa. Em áreas onde alguma forma de avaliação já é um requisito para a publicação, por exemplo algoritmos de mineração de dados, revisões estão cada vez mais exigentes com o rigor da avaliação, na experiência pessoal do autor.

Nos casos onde a simples criação de um programa/sistema novo não é suficiente como pesquisa é necessário encontrar algum "conhecimento"[1] sobre o programa/sistema obtido de forma mais metodológica. Ou dito de outra forma, é preciso **avaliar** o programa/sistema.

Há várias definições de avaliação, principalmente porque avaliação é algo de importância em muitas áreas, por exemplo, educação, administração, políticas publicas, além da computação. Uma definição de avaliação, adaptada de [Guba and Lincon 1981], é:

> Avaliação é o processo de julgar o mérito e valor de um sistema de informação.

Avaliação deve ser contrastada com dois outros conceitos, verificação e validação. **Verificação** é o processo de julgar a aderência de um sistema de informação com a sua especificação, e **validação** é o processo de julgar quão bem um sistema de informação resolve o problema para o qual ele foi concebido. E avaliação de uma forma ampla é o processo de verificar para que serve e quanto serve um sistema. A diferença entre os termos surge porque, por exemlplo, o problema para qual o sistema foi concebido não é o "problema real", ou porque resolvendo este problema surgem outros problemas, ou porque a especificação pode ser incompleta ou errada frente ao problema para o qual o sistema foi concebido, etc.

Neste curso, o valor e o mérito de um programa é exatamente a sua capacidade de resolver o problema para o qual ele foi concebido, e portanto a avaliação de programa equivale à validação do programa. Já sobre sistemas, exatamente porque ele vai ser usado em "lugares reais" por "usuários reais" o mérito do sistema pode não ter nada a ver com o problema para o qual ele foi concebido.

Métodos de pesquisa científica discutidos neste curso são usados comumente para avaliar programas e sistemas. Mas avaliação não é o único uso

[1] Aspas são usadas em termos que podem numa primeira aproximação ser lidos no seu sentido usual, mas que numa segunda aproximação são tópicos que suscintam maior debate.

destes métodos. Outras aplicações dos métodos de pesquisa científica são: entendimento de requisitos e práticas de trabalho e demonstrações e refutações de teorias sobre sistemas de informação.

Em alguns domínios da CC, usa-se métodos de pesquisa científica para entender as práticas e o contexto de trabalho em uma organização, em um grupo, ou em comunidades, como subsídio para especificar sistemas de informação para esta organização, grupo ou comunidade. Nós chamaremos isso de "entender" os requisitos de um sistema, em vez de usar o termo mais comum em engenharia de software, "análise de requisito," pois o resultado não é necessariamente uma especificação dos requisitos de um sistema futuro. O autor está mais familiarizado com o uso de técnicas de pesquisa qualitativa para entender práticas de trabalho colaborativo, que podem ou não guiar o desenvolvimento de um sistema de informação. Usando exemplos clássicos em CSCW (Computer supported collaborative work), [Bentley et al. 1992] usa uma técnica qualitativa que discutiremos abaixo, para especificar requisitos importantes e não importantes para um sistema de auxilio ao controle de trafico aéreo. Já [Watts et al. 1996] discute como canais de voz hierárquicos são usados para controlar missões da NASA, apenas para ilustrar como uma colaboração intensa, envolvendo centenas de pessoas, pode ser coordenada através de um canal de comunicação de baixo volume.

Finalmente, métodos científicos são usados em CC para "fazer ciência", como é entendido em outras áreas científicas. Sem entrar em muitos detalhes do que é ciência (que é o assunto da Epistemologia, uma subárea da Filosofia), a maioria das ciências naturais (que estudam fenômenos naturais) trabalha com teorias ou "leis" genéricas, que explicam as "observações" e as "descobertas". Um caso clássico em Física, que é o exemplo mais paradigmático de ciência natural, é a teoria da gravitação e as três leis do movimento de Newton, que explica não só as leis de Keppler para os movimentos planetários, mas a queda de corpos na Terra, etc. Para que Newton pudesse criar as suas leis, incontáveis astrônomos tiveram que coletar dados quantitativos sobre posição de planetas, etc., e Kepler teve que criar um conjunto de leis que resumia/explicava o movimento planetário. Mas as leis de Newton não explicam o pequeno efeito da presseção da órbita de Mércurio, e este, entre outros fenômenos, levou à substituição das leis de Newton pela teoria geral da relatividade de Einstein.

De forma análoga, a coleta de dados (preferencialmente quantitativos) sobre custos de desenvolvimento de software, ou sobre defeitos de software, ou sobre resistência na adoção de um novo sistema, etc., pode levar à criação de leis gerais ou teorias em diferentes áreas da computação. E a descoberta de um fenômeno que não pode ser explicado por essas teorias, pode levar à sua substituição por outra teoria. Por outro lado, ciência da computação, se for uma ciência[2], tem características peculiares - não só há poucas "leis" da

[2] Os artigos [Denning 2005, Tichy 1998, Newell and Simon 1976, Brooks 1996] entre ou-

computação, mas as leis que existem:

- têm um caráter estatístico - isto é, elas são leis válidas para grandes quantidades de exemplos, mas não necessariamente para um exemplo
- são rasas[3] - isto é, elas se parecem mais com "descobertas," e não com leis ou teorias gerais que permitem a derivação de várias conclusões (que podem ser posteriormente verificadas ou não)
- a maioria delas são datadas - isto é, as leis são válidas para um particular período, e não se espera que elas continuem válidas indefinidamente

Vejamos cada item. É difícil pensar em leis ou teorias da computação, isto é, formulações genéricas que são empiricamente verificáveis, mas não são necessariamente verdadeiras. Dizer que o problema do caixeiro viajante é NP-completo não é uma lei empírica, é uma conclusão matemática da definição do problema, e da definição de NP completo. Um exemplo de lei da computação, ou da engenharia de software, é a que afirma que a manutenção de software consome pelo menos 60% do custo total do software ([Huff 1990], por exemplo).

Esta lei é claramente estatística: encontrar um projeto, cujo custo de manutenção foi de 40% do custo total, não invalida a lei. A lei também não permite derivar muitas conclusões além do que ela diz. E finalmente, a proporção 60% reflete um particular momento e situação, que depende da longevidade média dos sistemas, do tipo de sistema, das práticas de desenvolvimento e manutenção de sistemas, etc. Por exemplo, [Koskinen 2007] contém uma tabela com os diferentes resultados de proporção do custo de manutenção em diferentes artigos em diferentes momentos.

De qualquer forma, o uso de métodos quantitativos é necessário para coletar dados para definir e atualizar as leis/teorias da computação. E se estamos tratando de teorias determinísticas, métodos quantitativos e qualitativos podem descobrir exemplos que não são explicados por uma teoria determinística, e que portanto requerem a elaboração de uma nova teoria que suplanta a anterior.

Conhecimento em ciência da computação é obtido usando as seguintes grandes metodologias:

- pesquisa analítica
- pesquisa quantitativa
- pesquisa qualitativa
- pesquisa bibliográfica

A pesquisa quantitativa e a qualitativa são coletivamente chamadas de pesquisa **empírica**.

tros, discutem se ciência da computação é ou não uma ciência.

[3] Rasa (*shallow*) é o termo usado em [Kitchenham et al. 2002].

A pesquisa analítica, que não será objeto desse minicurso, é o método mais comum de gerar/obter conhecimento sobre programas e algoritmos. A pesquisa analítica faz algumas pressuposições sobre os dados do programa ou sobre a máquina onde o programa vai ser executado, e *prova* matematicamente que o programa tem algumas propriedades interessantes. A análise de complexidade assintótica de algoritmos é uma forma de pesquisa analítica - há um conjunto de pressuposições sobre a infra-estrutura computacional onde o algoritmo "roda," e prova-se propriedades sobre tempo de execução, uso de memória, etc. Análise de algoritmos e programas onde se modela a distribuição de probabilidade dos dados (por exemplo [Mezard et al. 2002]) são também exemplos de pesquisa analítica.

A pesquisa quantitativa é baseada na medida (normalmente numérica) de poucas variáveis objetivas, na ênfase em comparação de resultados e no uso intensivo de técnicas estatísticas. Os métodos quantitativos a ser abordados neste minicurso são:

- uso de dados sintéticos: *benchmarks*, simulações e competições
- técnicas estatísticas para a comparação de conjuntos de medidas
- uso de questionários (*surveys*)
- desenhos experimentais

Há mais um conjunto de técnicas usual em pesquisa quantitativa, mas que não será abordado neste curso, a pesquisa **correlacional** ou **modelagem estatística**. A idéia da pesquisa correlacional é verificar a existência de alguma relação estatística entre poucas variáveis em um grande conjunto de dados. No caso mais comum, busca-se uma correlação significativa entre duas variáveis. Em estudos mais elaborados, usa-se modelos mais complexos que relacionam as várias variáveis (a correlação mede a aderência dos dados a um modelo linear que relaciona duas variáveis). Em CC há poucos exemplos de estudos correlacionais, mas uma área relacionada onde se usa modelagem estatística é o estudo de impactos da informática na produtividade de empresas e países [Wainer 2003, Brynjolfsson and Hitt 1998, Dewan and Kraemer 1998].

A pesquisa qualitativa baseia-se na observação cuidadosa dos ambientes onde o sistema está sendo usado ou onde será usado, do entendimento das várias perspectivas dos usuários ou potenciais usuários do sistema, etc. Os métodos qualitativos a ser apresentados nesse minicurso são:

- estudos qualitativos observacionais
- pesquisa-ação (ou estudos qualitativos intervencionistas)
- outras formas de avaliação qualitativa

A pesquisa bibliográfica, que não será abordada neste curso, não é apenas coletar e resumir *alguns* artigos relevantes à pesquisa, como tem sido feito em artigos e teses em computação. Por pesquisa científica bibliográfica nos referimos a duas práticas comuns nas Ciências da Saúde: **revisão**

sistemática e **meta-análises**. Em ambas, o objetivo é coletar *todos* os artigos publicados que reportam a algum experimento quantitativo pelo qual estamos interessados e resumir os vários resultados. A revisão sistemática termina em uma avaliação qualitativa e quantitativa desses vários resultados (por exemplo distribuição e homogeneidade dos resultados), enquanto que meta-análise usa técnicas estatísticas avançadas para agregar os vários resultados experimentais num único número. O site [Castro 2007] apresenta um curso virtual sobre meta-análise. O artigo [Chen and Rada 1996] é uma revisão sistemática que sumariza 23 experimentos sobre usabilidade de hipertextos; [Hundhausen et al. 2002] é uma revisão sistemática sobre efetividade educacional de visualização de algoritmos.

5.2. Métodos quantitativos

A pesquisa quantitativa vem da tradição das ciências naturais, onde as variáveis observadas são poucas, objetivas e medidas em escalas numéricas. Filosoficamente, a pesquisa quantitativa baseia-se numa visão dita **positivista** onde:

- as variáveis a serem observadas são consideradas objetivas, isto é, diferentes observadores obterão os mesmos resultados em observações distintas
- não há desacordo do que é *melhor* e o que é *pior* para os valores dessas variáveis objetivas
- medições numéricas são consideradas mais ricas que descrições verbais, pois elas se adequam à manipulação estatística

A essência da pesquisa quantitativa em ciência da computação é verificar o quão "melhor" é usar um programa/sistema novo frente à(s) alternativa(s).

5.2.1. Dados sintéticos

Algumas áreas da computação usam conjuntos de dados ou conjuntos de exemplos definidos pela comunidade como forma de avaliar os programas. Tais conjuntos de dados, ou *benchmarks*, devem em princípio representar a possível diversidade dos "dados reais". Em algumas áreas da computação é possível selecionar um subconjunto do *benchmark* para avaliar os programas, em outras, todo o *benchmark* deve ser usado. Nos casos mais completos, deve-se executar o programa novo P_n e um ou mais programas competitivos $P_1, P_2, \ldots P_k$ nos mesmos dados do *benchmark*, e usar técnicas estatísticas de comparação de conjuntos de medidas para determinar se existem diferenças significativas entre os resultados do programa novo P_n e os resultados dos programas competitivos P_1 até P_k.

Outra alternativa (menos desejada) é que os autores dos programas P_1, P_2, etc. publiquem, não só o desempenho dos seus programas, mas também dados como o intervalo de confiança (ver abaixo) dessas medidas, de forma que o autor do programa novo P_n pode compará-lo com alguma sofisticação estatística com os outros P_i.

A maioria dos *benchmarks* podem ser agrupados em três classes principais. No primeiro grupo, *benchmarks* que são usados para avaliar o tempo de execução do programa. O segundo grupo são os *benchmarks* que são usados para avaliar se um programa consegue obter um resultado (dentro talvez de certas restrições de tempo de execução), e resultam num conjunto de medidas binárias (resolveu ou não resolveu o problema). Esta segunda classe de *benchmarks* normalmente especifica não só exemplos de problemas, mas também sua solução. Finalmente, a terceira classe de *benchmarks* é usada para avaliar a qualidade da resposta do programa. Programas que usam heurísticas, que fazem aproximações, etc., nem sempre retornam a resposta "correta", e *benchmarks* dessa familia são usados para avaliar a qualidade da resposta - que pode ser tanto binária (acertou/errou), quanto uma medida de erro.

Criar, manter, atualizar, contribuir e analisar as características de *benchmarks* é uma atividade que deve ser considerada como cientificamente nobre e importante. Por exemplo, [Holte 1993] mostra que um *benchmark* muito usado em aprendizado de máquina (o conjunto de dados do UCI) tinha um sério viés - os exemplos eram muito fáceis, e, portanto, quase todos as técnicas usadas tinham resultados muito bons nesses exemplos. Esse fenômeno é chamado de **efeito de teto**, quando os programas atingem perto do máximo possível no benchmark. Isso dificulta a análise do desempenho do programa - se todos os programas têm diferenças de desempenho muito pequenas, é possível que elas desapareçam na análise de significância. O efeito contrário, **efeito-chão**, acontece quando os programas de *benchmark* são muito difíceis e o desempenho dos programas é todo muito baixo.

Algumas áreas da CC permitem que os programas sejam testados com dados gerados artificialmente por meio de simuladores. De novo, no melhor caso, os exemplos gerados por simulação seriam testados, tanto no programa P_n, quanto nos programas alternativos P_1 até P_k.

É importante perceber que dados gerados através de uma simulação têm um viés, já que o gerador cria exemplos segundo uma distribuição de probabilidade que pode não corresponder aos "dados reais". Assim, algum cuidado deve ser tomado quando se compara resultados de programas diferentes com dados gerados por simuladores diferentes, pois é provável que cada simulador gere exemplos com diferentes distribuições.

Finalmente, algumas áreas da CC possuem a tradição de competições, ou desafios. Os dados que serão usados para avaliar os programas não são sabidos de antemão. Só durante a competição os dados são apresentados, e a avaliação dos programas se faz naquele momento. Em áreas como Inteligência Artificial, especialmente nos subdomínios onde os programas são competitivos, como jogos, algumas formas de robótica, etc., a competição é normalmente um conjunto de "partidas" entre os programas/sistemas competidores, e a avaliação do programa se dá por quantos pontos ele acumula durante a competição.

5.2.1.1. O que fazer com isso?

O pesquisador deve informar-se da disponibilidade de *benchmarks*, simuladores e competições na sua área de pesquisa.

5.2.2. Significância estatística

5.2.2.1. Tipos de medida

Em pesquisa quantitativa, assume-se que as variáveis de interesse podem ser medidas objetivamente. Mas o que é medido pode variar. Por exemplo, no caso de *benchmarks* descritos acima, uma medida possível é o tempo de execução, outra medida possível é se o programa acertou ou não a resposta. A primeira medida é um real, a segunda um dado binário.

Genericamente, classifica-se as medidas nas seguintes classes:

medidas categóricas ou nominais Medidas categóricas indicam apenas a classe do dado, e a única operação possível é verificar se o dado tem um ou outro valor. Uma medida categórica clássica é sexo: masculino ou feminino. Não há nenhuma operação que faça sentido com esses dois valores: ordená-los, fazer operações matemáticas, etc. Mesmo que se codifique 1 para o sexo masculino e 2 para o feminino, não faz nenhum sentido "somar" o sexo de um grupo de pessoas, ou tirar a média do sexo, etc. Variáveis categóricas podem ter mais de dois valores, por exemplo, estado ou país de origem, diagnóstico médico, etc. Nestes casos, a codificação das categorias em números deve ser entendida com cuidado. Se atribuímos 1 para Acre, 2 para Alagoas, 3 para Amazonas, e assim por diante, não faz nenhum sentido somar esses números, subtrair um do outro, dizer que Piauí é maior que Para só porque seu código é um número maior, etc.

medidas ordinais Medidas ordinais também atribuem classes aos dados, mas é possível ordená-las de maior para menor. Um exemplo clássico é classe socioeconômica. Normalmente usa-se as classes A,B,C e D, e existe uma ordem entre elas: de A para D nessa ordem. Outras variáveis ordinais comuns são dificuldade de um projeto de software, nível (e não anos) de educação de um profissional, etc. Se os valores ordinais são codificados com números, por exemplo, classe socioeconômica A = 4, B = 3, C = 2, etc., então a ordem dos números reflete a ordem dos valores, mas é preciso ter em mente que a diferença dos números/código não faz nenhum sentido - a distância entre a classe A e B não é a mesma que a distância entre B e C mas a diferença dos códigos é idêntica.

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 medida de temperatura em Célsius. Medidas intervalares garantem que as diferenças entre duas medidas (o intervalo) é algo que pode ser comparado: quando a temperatura é de $20^{\circ}C$ não se pode dizer que

está duas vezes mais quente que quando a temperatura está $10°C$, mas pode-se dizer que o ganho de temperatura quando se vai de $10°C$ para $20°C$ é duas vezes maior do que quando se vai de $20°C$ para $25°C$.

medidas de razão Medidas de razão atribuem ao dado um número real onde o zero é absoluto, e portanto razões entre duas medidas fazem sentido. O exemplo clássico é a medida de temperatura em Kelvin, ou medida de distância em metros, etc.

Para a computação provavelmente não há muitos exemplos onde a diferença entre medidas de razão e intervalares é muito relevante, e, portanto, neste texto nós falaremos em medidas *pelo menos intervalares*, para referirnos a intervalares ou de razão. Então, em computação é preciso ter em mente pelo menos as diferenças entre medidas categóricas, ordinais e medidas pelo menos intervalares.

Esta classificação de medidas é importante por duas razões: o tipo de medida define que tipo de estatística é possível usar para sumarizar os dados, e que tipo de teste estatístico deverá ser usado para verificar se dois conjuntos de dados são significativamente diferentes ou não. Vamos nos concentrar nas estatísticas para sumarização de um conjunto de dados nesta seção, e na próxima falaremos de comparações e de significância estatística.

Dado um conjunto de medidas categóricas, o único jeito de descrever os dados é apresentar a *distribuição de freqüências*: 2% dos produtos vieram do Acre, 14% de Alagoas, 13% do Amazonas, e assim por diante. E para sumarizar os dados, o máximo que se pode fazer é apresentar a *moda* ou o valor com maior freqüência. Para medidas ordinais, a medida sumarizadora mais comum é a mediana, o valor que divide o conjunto de dados em duas metades. A forma de descrever os dados pode ser também através da distribuição de freqüências para cada um dos valores.

Medidas pelo menos intervalares são sumarizadas através da média, e do desvio padrão. Medidas de razão também permitem coisas como média geométrica e média harmonia, que não faz sentido para medidas intervalares.

Um dado muito comum em computação é se o programa produz o resultado correto ou não para um particular dado de entrada. Para resguardar-se de erros, o pesquisador deve considerar essa medida como uma medida categórica! Mesmo usando a codificação tradicional de 0 para falha e 1 para sucesso, o pesquisador não deve enganar-se e pensar que está trabalhando com uma medida de razão, embora contas como a média parecem fazer sentido. Se o programa acertou 30 exemplos e errou 12, dizer que o programa acerta $30/(30+12) = 71.4\%$ parece fazer sentido. O que está acontecendo é que o número 71.4% é a descrição da distribuição dos valores - certo 71.4% das vezes e errado 28.6% das vezes. Se, neste caso, a conta da média acabou resultando em uma proporção, ou seja, um número que descreve a distribuição dos dados levar essa analogia mais a fundo é um erro. Para comparar dois programas quanto à sua corretude, não se usa o teste t (ver abaixo), que seria

apropriado para uma medida pelo menos intervalar, e sim o teste chi-quadrado (ver abaixo), que é usado para medidas categóricas.

5.2.2.2. Significância estatística

Nota: Esta seção é apenas introdutória para o assunto de significância estatística. Nós não explicaremos a mecânica e as pressuposições dos testes estatísticos, nem mesmo as fórmulas que resumem os testes. O objetivo desta seção é apenas associar o nome do teste a um problema específico (por exemplo, verificar se a média de dois conjuntos de medidas são significativamente diferentes). As fórmulas para o teste devem ser obtidas de outras fontes. Finalmente, o autor não é especialista em estatística, e as regras colocadas aqui talvez não sejam as mais modernas - se o leitor tem uma forte opinião de que outro teste é mais apropriado para o seu problema, deve seguir sua intuição e conhecimento.

Testes estatísticos são procedimentos que fazem uma particular pressuposição sobre os dados, a chamada **hipótese nula**, e mais uma serie de outras pressuposições (que chamaremos de **condições do teste**) sobre os dados, e calculam a probabilidade que alguma **propriedade relacionada** aos dados seja verdadeira, dadas as pressuposições. Essa probabilidade calculada é chamada **valor p** ou (*p-value*). Se as condições do teste são verdadeiras, e o p-value é suficientemente baixo, então o pesquisador pode assumir que a hipótese nula é falsa, ou como é normalmente fraseado, "há evidencias suficientes para rejeitar a hipótese nula". O valor do p-value abaixo do qual se assume que a hipótese é falsa é usualmente 0.05 ou 0.01. Se o valor de corte é 0.05, então diz-se que a **significância** do teste é (1-0.05), ou 95%. Se o valor de corte é 0.01, diz-se que a significância do teste é 99%. Se p-value calculado é maior que o valor de corte, então a conclusão que se tira é que "não há evidências para rejeitar a hipótese nula".

Por exemplo, o teste T, que será discutido em mais detalhes abaixo, tem as seguintes características:

- a hipótese nula é que os dois conjuntos de dados são duas amostras independentes de uma mesma população
- as condições do teste é que os dados dos dois conjuntos são pelo menos intervalares, têm distribuições normais e os dois conjuntos têm a mesma variância
- a propriedade relacionada é que a diferença das médias de dois conjuntos de dados tenha um valor igual ou maior que o obtido nos dados

Na maioria das vezes os testes estatísticos são usados para comparar dois ou mais conjuntos de medidas. Por exemplo, tem-se os tempos de execução de dois programas diferentes para dados gerados por um simulador, e quer-se verificar se a diferença no tempo de execução médio é significativa. Assim, temos o conjunto D_1 de medidas de tempos de execução do programa 1, e o

conjunto D_2 de medidas de tempos de execução do programa 2. A idéia por traz da maioria dos testes estatísticos é assumir que existe uma única fonte de dados D, e que tanto D_1 como D_2 são amostragens dessa mesma fonte D. Se isso é verdade então D_1 e D_2 não são realmente diferentes entre si, e, portanto, os programas 1 e 2 também não são muito diferentes!

A hipótese nula é exatamente a idéia de que D_1 e D_2 são amostragens da mesma fonte D. Mas D_1 e D_2 têm algumas propriedades diferentes, por exemplo, a média. Se as medidas são ordinais, então talvez a mediana seja diferente, se são categóricas, talvez a distribuição dos valores seja diferente, etc. Esta característica dos dois conjuntos que o teste vai explorar é o que chamamos de propriedade relacionada aos dados. Finalmente, assumindo a hipótese nula, e usando os dados D_1 e D_2 para calcular algumas propriedades da fonte D, e fazendo mais algumas pressuposições (as condições do teste), pode-se calcular a probabilidade que em duas amostragens de D, a tal propriedade relacionada, seja tão grande quanto se encontrou em D_1 e D_2. Se essa probabilidade for muito baixa, tradicionalmente menor que 0.05, assume-se que D_1 e D_2 **não** são amostragens da mesma fonte D, e que são *significativamente* diferentes.

A mesma idéia de uma fonte D pode ser usada para apenas um conjunto de medidas D_1, que teria sido amostrada de D. Dados D e outras pressuposições, qual a probabilidade de que a propriedade relacionada a D_1 tenha sido tão grande quanto a observada? Isto pode ser usado, por exemplo, para julgar se D_1 tem uma distribuição normal, que é uma das condições de teste comuns a ser verificada. Numa outra conta, usualmente para conjuntos de dados pelo menos intervalares, podemos calcular um parâmetro p de D (usualmente a média), de tal forma que a probabilidade que uma amostragem de $D(p)$ tenha a média que D_1 tem seja exatamente 0.05. Há dois valores possíveis para esse parâmetro, quando ele é maior que a média de D_1 e quando ele é menor. Este intervalo é chamado de **o intervalo de 95% de confiança** para a média de D_1. O intervalo de 95% de confiança de um parâmetro de D_1 é o intervalo correspondente ao maior e menor valores para esse parâmetro da fonte de dados D, que garantem que o p-value da propriedade relacionada de D_1 é igual a 0.05. Assim, se o intervalo de 95% de confiança para a média de D_1 dados é 4 ± 2.3 então:

- a média de D_1 é 4
- 1.7 é o menor valor da média de D para o qual o p-value será 0.05 ou mais, e
- 6.3 é o maior valor da média de D para o qual o p-value será 0.05 ou mais

Testes estatísticos para a comparação de dois conjuntos de dados e os intervalos de confiança dos dados são relacionados. Se os intervalos de confiança dos dois conjuntos têm alguma intersecção, então o teste dirá que as diferenças não são significativas. Assim, se o pesquisador publicou o intervalo

de confiança para o tempo médio de execução do programa P1, então para compará-lo com o programa P2, só é preciso gerar o intervalo de confiança para os tempos de execução de P2, e verificar se há alguma intersecção. Isso só funciona para a comparação de *dois* conjuntos de dados, e normalmente apenas para medidas pelo menos intervalares.

Dados dois testes T1 e T2, diremos que T1 **é mais forte** que T2, se para os mesmos dados o p-value calculado por T1 é menor que o de T2. Um teste é mais forte que outro, normalmente, 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 permite tal conclusão. Um teste é dito **não-paramétrico**, se entre suas condições de teste não há qualquer pressuposição que os dados têm alguma distribuição fixa. Exatamente porque assumem menos pressuposições nos dados, testes não-paramétricos são mais fracos que seus correspondentes paramétricos.

Os testes mais comuns para calcular o intervalo de confiança são:

teste T é usado para calcular intervalo de confiança para a média de medidas pelo menos intervalares distribuídas de forma normal

intervalo de Wald é usado o intervalo de confiança de uma proporção p, desde que np e $n(1-p)$ sejam > 5, onde n é o número total de dados

Vejamos agora os testes mais comuns para a comparação de apenas 2 conjuntos de dados.

teste T é usado quando se for verificar se a média de um conjunto de medidas pelo menos intervalares é maior que a média de outro conjunto. As condições do teste são:

- as variâncias dos dois conjuntos são iguais (que pode ser verificado usando o teste F ou o teste de Levene)
- que os dados dos dois conjuntos estão distribuídos segundo uma normal (que pode ser verificado usando o teste de Shapiro-Wilk)

teste T pareado usa-se nas mesmas condições do teste T, mas quando cada medida de um dos conjuntos pode ser colocada em correspondência com uma do outro conjunto. Por exemplo, o primeiro conjunto indica as notas dos alunos de uma classe na primeira prova e o segundo, a nota dos mesmos alunos na segunda prova. Ou o primeiro conjunto são os tempos de execução do programa P1 num conjunto de exemplos, e o segundo conjunto, os tempos de execução do programa P2 nos mesmos exemplos. O teste T pareado é mais forte que o teste T

teste U de Mann-Whitney ou Wilcoxon rank-sum test se as condições do teste T (não pareado) não são verdadeiras (não-normalidade ou variâncias muito diferentes), então o teste U deve ser usado. Este é um teste não-paramétrico, usado para medidas pelo menos ordinais

Wilcoxon signed-rank test é a alternativa não-paramétrica do teste T pareado

Teste do Chi-quadrado (ou qui-quadrado) é usado para verificar se duas distribuições são significativamente diferentes ou não (a hipótese nula é que elas são iguais). Como o chi-quadrado trabalha com distribuições, ele é usado sempre que as medidas são categóricas e algumas vezes com medidas ordinais. O chi-quadrado é usado também para verificar independência de duas categorias. O teste chi-quadrado só pode ser usado se o número de observações em cada categoria é no mínimo 5

Fisher exact test é uma versão mais elaborada do Chi-quadrado e útil quando nem todos os x_i são maiores que 5

Os testes acima só são apropriados para a comparação de dois conjuntos. Comparar mais que dois grupos ou conjuntos de dados é chamado **comparações múltiplas**. O problema de usar os testes acima em comparações múltiplas é que a significância diminui com o número de comparações. Se considerarmos que o grau de significância de 95% significa que a "probabilidade que a conclusão esteja certa" é de 0.95, então se compararmos 4 conjuntos de dados (D_1 a D_4), que resulta em $4 \times 3/2 = 6$ comparações, e cada comparação é feita com nível de significância de 95%, e se chegarmos à conclusão que

> **Afirmação 1:** a média de D_1 é maior que a média de D_2 que por sua vez é maior que a média de D_3 e D_4 que não são diferentes entre si

então, numa primeira aproximação, a probabilidade da afirmação 1 acima ser correta é de 73,5%. Se cada uma das comparações é independente das outras (e elas não são!), então a probabilidade de todas as comparações serem corretas (que resultou na afirmação 1) é de $0.95^6 = 0.735$.

Portanto, utilizar várias vezes um teste para dois conjuntos reduz a significância da conclusão. De uma forma intuitiva (e incorreta!), a significância de 95% indica a chance de chegar à conclusão errada uma vez a cada 20 vezes, logo, se o teste for usado 20 vezes, pelo menos uma das conclusões estará errada!

Os testes mais comuns para múltiplas comparações de variáveis pelo menos intervalares são:

one way ANOVA o teste ANOVA é usado para testar a hipótese que mais de 2 conjuntos de medidas pelo menos intervalares "não têm todos a mesma média". Isto é, a hipótese nula do ANOVA é que todos os conjuntos têm a mesma média, e, portanto, a rejeição da hipótese nula é que "pelo menos um dos conjuntos não tem a mesma média dos outros," mas o ANOVA *não diz* quais conjuntos têm ou não a mesma média

Kruskal-Wallis o correspondente não-paramétrico ao one-way ANOVA. Usado para medidas pelo menos ordinais

comparações múltiplas para verificar qual de mais de 2 conjuntos de medidas são diferentes entre si, há vários métodos. Alguns deles são conhecidos como ANOVA post-hoc tests, pois só devem ser aplicados depois que o ANOVA confirmou que nem todas as médias são iguais. Outros testes e técnicas são: Bonferroni, Tukey HSD, Scheffe, técnicas baseadas no Studentized range, etc. [Toothaker 1993]

Para finalizar, deve-se notar que na maioria das vezes queremos mostrar que o nosso programa P_n é (significantemente) melhor que a alternativa P_0. Isto é, a hipotese nula (que os dois conjuntos vieram da mesma população) é o que *não* desejamos. Assim, quando o p-value é menor que 0.05, e rejeitamos a hipotese nula, concluimos o que queríamos, que P_n é melhor que a alternativa. Mas de vez em quando queremos mostrar que P_n é igual ou equivalente a P_0, por exemplo, que o nosso programa P_n tem a mesma taxa de acerto que a alternativa P_0, mas, por exemplo roda em menos tempo, ou utiliza menos recursos. Nesse caso queremos "provar" a hipotese nula. É *errado* dizer que se o p-value é maior que 0.05 então não rejeitamos a hipotese nula e, portanto, provamos que os dois conjuntos de dados são equivalentes. Por exemplo, [Lew 2006] discute esse tipo de análise, chamada de **experimentos reversos**.

5.2.2.3. Desenhos fatoriais

Normalmente os programas P_i que estão analisados possuem várias variações ou alternativas. Vamos dizer que o programa P_N tem 2 parâmetros, onde o primeiro pode assumir 2 valores e o segundo 3. Por exemplo, o primeiro parâmetro define um tipo de busca em grafos - largura ou profundidade - enquanto o segundo define o tamanho de um buffer intermediário no programa que pode assumir os valores 1000, 5000 ou 10000. O objetivo do pesquisador é tentar entender a relação entre os valores dos dois parâmetros.

Em estatística o que chamamos de parâmetro é chamado de **fator**, e, portanto, o problema acima têm 2 fatores. O número de valores possível para cada valor é chamado de **nível** (*level*). Portanto, o primeiro fator tem 2 níveis e o segundo 3 níveis. A escolha de que combinações de níveis devemos explorar é conhecido como desenho experimental. Aqui estamos usando o termo "desenho experimental" num sentido diferente do que usaremos na seção 5.2.4, onde falaremos da construção de experimentos envolvendo pessoas.

O desenho é um desenho fatorial (completo) (*(full) factorial design*) se todas as combinações de níveis para todos os fatores são testados. Se nem todas as combinações são testadas o desenho é chamado **fatorial parcial**. Obviamente, o central de um desenho parcial é a escolha de que combinações usar. Livros mais avançados de estatística (por exemplo [Box et al. 1978]) normalmente cobrem os desenhos parciais e a análise estatística apropriada para esses desenhos.

Um desenho parcial que é muito usado e que *não* é muito interessante é o um-fator-por-vez. Se houver 3 ou mais fatores, cada um com vários níveis, esse

desenho assume um nível para cada fator como sendo o padrão; mantendo-se os outros fatores no padrão, altera-se os níveis de um só fator a cada vez. Esse desenho óbvio deve ser evitado [Czitrom 1999].

5.2.2.4. O que fazer com isso?

Esta seção deve ser suficiente para que o pesquisador saiba numa primeira aproximação que teste usar e quando para comparar conjuntos de dados. Todos os testes mencionados estão implementados em uma variedade de pacotes estatísticos, inclusive pacotes gratuitos como o R[4].

5.2.3. Questionários

As principais referências para esta seção são [Pfleeger and Kitchenham 2001] e os outros artigos da série que começa com [Kitchenham and Pfleeger 2002a].

Questionários são uma forma rápida e simples para avaliar as opiniões, objetivos, anseios, preferências, crenças, etc. de pessoas. Mas por ser uma forma simples, se malconcebida, pode levar a um viés considerável. Para este curso, questionários são um conjunto de perguntas com respostas predefinidas ou perguntas de **resposta fechada** (*closed questions*), que são respondidas, ou pelos próprios sujeitos da pesquisa (questionários **auto-aplicados**), ou por observadores que estão avaliando os sujeitos. Se um questionário é auto-aplicado, diremos que os sujeitos da pesquisa são os **respondentes**.

O uso de questionários envolve as seguintes fases:

- elaboração das perguntas e respostas
- amostragem da população
- avaliação das respostas
- análise dos resultados

Mas antes de entrarmos nessas fases, é importante notar que um questionário, na maioria das vezes, é um instrumento que através das perguntas tenta avaliar ou medir uma variável invisível ou **latente**. Por exemplo, se o pesquisador tem uma teoria sobre "estilos de programação", boa parte do questionário será sobre questões que avaliam os diferentes aspectos de cada um dos "estilos". Este questionário deve ter várias propriedades. Entre elas

- confiabilidade (*reliability*)
- validade (*validity*)
- não desencorajar o usuário a respondê-lo

Confiabilidade é a propriedade que diferentes aplicações do questionário (para o mesmo respondente) devem dar resultados próximos. Dessa forma um instrumento é confiável se sua utilização não causa "muito ruído". Validade é a

[4] www.r-project.org

propriedade que aquilo que é medido pelo instrumento é uma boa aproximação da variável latente que se quer medir.

A confiabilidade de um instrumento pode ser avaliada de várias formas, embora [Kitchenham and Pfleeger 2002b] afirme que ainda não existe consenso sobre o que é uma confiabilidade baixa, e quais as implicações disso para as conclusões tiradas do questionário. Uma forma de avaliar a confiabilidade é aplicar o questionário duas vezes para um mesmo grupo de pessoas, tomando cuidado para que o tempo entre as aplicações não seja tão grande que as respostas teriam mudado, nem tão curto que as pessoas se lembrem do que responderam na primeira aplicação. Se as diferenças entre as duas aplicações é muito grande, a confiabilidade do questionário é pequena. O artigo [Kitchenham and Pfleeger 2002b] propõe que se use a estatística alfa de Cronbach ou a estatística tau de Kendall para calcular o grau de correspondência entre as respostas antes e depois. Essas estatísticas são medidas de correlação não-paramétricas, já que as respostas a uma questão, mesmo que sejam codificadas com números, são variáveis categóricas, ou no máximo ordinais, para as quais testes paramétricos não são apropriados.

A validade do questionário é bem mais difícil de medir, e na maioria das vezes só pode ser avaliado com experimentos que comparam os resultados de usar um questionário com os resultados de usar outros métodos que avaliam a mesma variável latente. Normalmente distingue-se as seguintes variações de validade para questionários:

validade de conteúdo é uma avaliação subjetiva por especialistas que o questionário contempla todos os aspectos da variável latente que se espera medir.

validade de critério é uma avaliação de quanto os resultados do questionário são correlacionados com outros instrumentos e medidas da mesma variável latente. Por exemplo, se um questionário novo está sendo desenvolvido para avaliar a qualidade do desenvolvimento de software, então os resultados do questionário não podem ser muito diferentes ou contrários aos resultados de questionários que medem a mesma variável, tais como os questionários do CMM ou do SPICE.

validade de construto é uma avaliação a longo prazo se diferentes formas de coletar dados (principalmente para questionários que não são auto-aplicados) produz resultados similares.

As questões de validade e confiabilidade do instrumento indicam que é melhor usar questionários já elaborados (e validados) que desenvolver um questionário novo. Em outras áreas, as várias validações de questionários são por si só uma forma válida e importante de pesquisa científica. Em ciência da computação há poucos exemplos de validação de questionários.

5.2.3.1. Desenvolvimento de um questionário

Se o pesquisador tiver que desenvolver um questionário novo, então deve ter cuidado na elaboração das questões e das respostas. Quanto às questões, normalmente recomenda-se:

- que as questões sejam fraseadas de forma simples, concisa, e direta
- que as questões sejam neutras
- que as questões não sejam fraseadas na negativa ("você acha que uma ferramenta CASE não é importante para ...")
- que as questões não contenham mais de uma pergunta ou conceito ("A ferramenta CASE ajudou você ou a sua equipe na sua habilidade de especificar projetos complexos e a entender as especificações de outras equipes?")

A elaboração das respostas predefinidas também é muito importante. O princípio que deve nortear essas respostas é o **balanço** - os extremos opostos das alternativas devem ser de igual intensidade e em direções opostas, e as alternativas intermediárias devem ser em igual número nas duas "direções" possíveis. Por exemplo, as respostas para a questão "Você achou que as funções de ajuda do sistema?"

1. foram super ótimas!!!
2. ajudaram muito
3. ajudaram em algumas ocasiões
4. não usei a função de ajuda

não são balanceadas pois não permitem que o respondente expresse que a função de ajuda o atrapalhou.

Uma prática comum para conseguir o balanço é usar a escala de Likert: as questões são fraseadas como afirmativas e o respondente deve escolher uma das alternativas

1. discordo totalmente
2. discordo
3. não discordo nem concordo
4. concordo
5. concordo totalmente

Há opiniões que a escala de Likert deve ser usada sem uma alternativa neutra.

5.2.3.2. Amostragem e não-resposta

Dois problemas comuns em questionários, especialmente questionários auto-aplicáveis, é o viés de amostragem e a não-resposta. Por exemplo, uma pequisa sobre práticas de desenvolvimento de software envia 200 questionários para os diretores de projetos de várias empresas. Destes, 30 retornam, alguns sem resposta para todas as questões.

O problema de viés de amostragem, também chamado de viés de **cobertura**, é a escolha inicial das 200 empresas. Elas são todas as empresas do ramo que se quer estudar? Elas são representativas das empresas que se quer estudar? Não há algum viés importante na seleção dessa amostra, por exemplo, empresas que se registraram na última conferência sobre técnicas de programação para tempo real, e, portanto, são empresas de projetos de sistemas de tempo real?

O segundo problema é que só 30 questionários voltaram respondidos, dos 200 enviados. Questionários enviados e não respondidos são chamados de não-resposta de unidade (*unit nonresponse*). A taxa de resposta do questionário é de 30/200, ou 15%. Taxas de resposta baixa são um problema comum em questionários enviados por correio, e algo impossível de calcular para questionários disponibilizados pela Internet. Mas mais sério que a taxa de resposta é um possível viés de resposta. Não há muitos problemas se quem responde o questionário pode ser considerado como uma amostragem aleatória das empresas para as quais o questionário foi enviado. Mas talvez as não-respostas tenham um viés, por exemplo, empresas pequenas têm menor probabilidade de responder o questionário que empresas grandes. Dessa forma, os questionários respondidos não são representativos do universo de empresas que se quer avaliar.

O terceiro problema é chamado de não-resposta ao item (*item nonresponse*) - nem todas as questões do questionário foram respondidas. Há duas técnicas-padrão para tratar não-resposta ao item: ponderação e imputação, que não serão abordadas neste texto.

5.2.3.3. Análise dos resultados

Existem várias maneiras de utilizar os dados de um questionário. Vamos discutir algumas dessas formas de utilização dos dados e alguns problemas associados.

A forma mais comum de utilização de dados de questionários é apenas reportá-los. Normalmente descreve-se a distribuição das respostas de cada questão (ou das questões mais relevantes). Neste caso, o "mais correto" é lembrar-se que as respostas são medidas no máximo ordinais (por exemplo, na escala de Likert), e que as estatísticas descritivas usadas devem ser apropriadas para o tipo de medida. Para medidas categóricas, deve-se descrever a distribuição de frequencia das respostas, ou, se deseja-se resumir o resultado,

reporta-se a resposta mais frequente. Para medidas ordinais, se deseja-se resumir o resultado, reporta-se a mediana das respostas, e talvez os primeiro e terceiro quartis.

A segunda forma de utilização de um questionário é agregar as várias respostas (talvez ponderadas por pesos) num só índice que reflete uma medida de interesse. Questionários de usabilidade de software, por exemplo SUMI [Kirakowski 2007] e QUIS [Shneiderman 1992], resumem as respostas de várias questões em poucos números que medem a qualidade do software nas dimensões de usabilidade. Questionários cujas respostas são agregadas em poucos valores-resumos devem ter sido validados pelo menos até o nível de validade de critério (ver acima). Nestes casos, embora não haja guias de conduta explícitos, o autor acredita que o valor-resumo de cada questionário pode ser considerado uma medida pelo menos intervalar, e portanto as técnicas de significância estatística para estas medidas podem ser aplicadas.

A terceira forma de utilização dos dados de questionários é buscar correlações entre as respostas de diferentes questões. Isto deve ser feito com cuidado. O problema central é relacionado com o problema de comparações múltiplas - fazendo-se várias correlações entre questões é provável que algumas correlações sejam erradamente avaliadas como significantes, quando elas não o são. [Kitchenham and Pfleeger 2002c] discute brevemente esse problema. Se o pesquisador sabe de antemão que questões ele vai tentar correlacionar entre si, então essas correlações espúrias não são tão importantes, mas se o pesquisador está "caçando" coisas interessantes nos dados do questionário, isso deverá ser levando em consideração.

5.2.3.4. O que fazer com isso?

Esta seção discutiu vários aspectos da elaboração e poucos aspectos da análise de questionários. A primeira lição é usar um questionário já feito e analisado, se possível. Se o pesquisador precisar desenvolver um questionário novo, deve tomar cuidado na elaboração das perguntas e respostas. Por outro lado a comunidade de CC como um todo (com exceção da área de usabilidade) não parece ainda ter desenvolvido critérios para avaliar os questionários em si, ou o resultado de questionários, e, portanto, parece ainda haver espaço para um certo "amadorismo" no uso de questionários.

5.2.4. Experimentos

As fontes principais para essa seção são [Marczyk et al. 2005, Trochim 2006].

Experimentos são atividades caracterizadas pela manipulação de algumas variáveis, e a observação de outras, em situações artificiais ou semi-artificiais. No caso de experimento em CC estamos sempre nos referindo a experimentos que envolvem seres humanos, e vários deles. Diferente das ciências naturais (física por exemplo), onde um "experimento" pode ser apenas uma medida feita em um equipamento especial e complexo, em ciências da computação di-

remos que um experimento necessariamente envolve um grupo (talvez grande) de pessoas, e várias medidas relativas a essas pessoas. Experimentos também envolve controle - o experimentador pode decidir que grupo de pessoas fará o quê, e em alguns casos, pode decidir quem participará de quais grupos. Isto difere de pesquisas observacionais (que veremos a seguir), onde o pesquisador não tem controle sobre o grupo de pessoas.

Esta seção cobrirá os conceitos de validade interna e externa de um experimento, e discutirá os vários desenhos experimentais[5], e quais são os tipos de ameaças à validade interna que eles apresentam.

Validade interna é a confiança que se tem de que o efeito observado é realmente devido à manipulação feita, e não a outros fatores. Uma ameaça à validade interna é uma outra possível explicação para o efeito observado que não a ação ou a manipulação feita. **Validade externa** é a confiança que se tem que o efeito observável é generalizável, ou seja, mesmo acreditando que para *esse grupo* o efeito é devido à manipulação, tal se repetirá para outros grupos?

Começaremos com o problema da validade interna. Vamos supor um experimento onde se pretende verificar se o uso de uma ferramenta CASE diminui o número de erros de um grupo de programadores. Vejamos dois possíveis experimentos:

- **experimento E1:** oferecemos aos programadores um projeto artificial (P1), esperamos que eles completem o projeto, contamos os erros de P1, instalamos a ferramenta CASE, atribuímos um outro projeto (P2) e exigimos que os programadores usem a ferramenta, e contamos os erros de P2.
- **experimento E2:** contamos os erros de um projeto anterior à instalação (P3), instalamos o CASE, e contamos os erros de um projeto (P4) posterior à instalação do CASE.

Do ponto de vista de desenho experimental, tanto E1 como E2 têm o desenho um-grupo/pré-teste/pós-teste, representado por

$$O_1 \quad X \quad O_2$$

onde O_1 representa o pré-teste (ou uma Observação), O_2 o pós-teste, e X a introdução do CASE. Em desenhos experimentais, chama-se o X de **intervenção**.

Mesmo que os erros de P2 sejam significativamente menores que os de P1, pode-se afirmar que a ferramenta CASE é a causa? Ou, de um outro ponto de vista, quais são as explicações rivais para a diminuição do número de erros de P1 para P2? Essas explicações alternativas são as ameaças à validade interna do experimento.

[5] Em inglês *experimental design*. Há várias alternativas em português para essa expressão, por exemplo, “delineamentos experimentais”, ou “projetos experimentais”. Neste texto optamos por “desenho experimental” [Hochman et al. 2005].

A primeira ameaça é que talvez P2 seja mais fácil que P1, e, portanto, teria menos erros. Essa ameaça é chamada de **instrumentação**, e se baseia na idéia que talvez a diferença ente O_1 e O_2 é devido a um erro na medição. Por exemplo, os testes de O_1 e de O_2 são suficientemente diferentes, ou as observações são feitas por pessoas diferentes, etc.

Uma outra ameaça, no caso de E1, é que talvez por causa do P1 (e dos erros do P1 que foram mostrados a eles) os programadores ficaram mais cientes dos seus erros e, portanto, não cometeram os mesmos erros em P2. Essa ameaça é chamada de **testagem**, ou seja, que o fato de passar por O_1 prepara os programadores para ir melhor no O_2.

Uma outra ameaça, no caso de E2, é que talvez os programadores ficaram melhores com o tempo, principalmente se muito tempo se passou entre P3 e P4. Programadores ficam melhores com a prática, assim como alunos ficam mais "sabidos" com o tempo, etc. Essa ameaça é chamada de **maturação**, ou seja, que os sujeitos dos experimentos podem tornar-se mais capazes com o tempo, independentemente da intervenção.

Uma outra ameaça é que talvez houve alguma iniciativa da empresa para diminuir o número de erros entre P1 e P2 (ou mais provavelmente entre P3 e P4). Essa ameaça é chamada de **história**, ou seja, a possibilidade de que haja um evento externo ao experimento que causou a melhora dos resultados.

Nem todas essas ameaças acima podem ser relevantes para o experimento em questão, mas se o pesquisador adotar o desenho acima, ele precisará explicar quais das ameaças não se aplicam e porquê. As outras ameaças são relevantes em desenhos experimentais onde há dois grupos. Vamos considerar o experimento a seguir:

- **experimento E3:** escolhemos dois grupos de programadores (talvez de divisões diferentes). Para o grupo 1, seguimos o experimento E1, e para o grupo 2, seguimos E1, mas sem a ferramenta CASE.

O desenho de E3 é:

$$\begin{array}{lll} O_1 & X & O_2 \\ O_3 & & O_4 \end{array}$$

e o efeito de X é em principio verificado se $O_2 - O_1 > O_4 - O_3$. 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**. Este desenho neutraliza as ameaças de maturação e história, mas por outro lado introduz as seguintes ameaças:

seleção é possível que o grupo 1, que usou o CASE, acabou sendo escolhido porque a sua divisão é mais dinâmica e estaria mais apta e disposta a aceitar a introdução da ferramenta CASE. E talvez esse dinamismo e entusiasmo é que causou a melhora da diferença do pós-teste menos o pré-teste.

mortalidade seletiva pode ser que os membros do grupo 2 pertençam a uma divisão menos dinâmica e tendam a sair da empresa com mais freqüência, principalmente os mais capazes. Assim os resultados do teste O_4 são piores que os de O_2 porque os melhores programadores do segundo grupo têm maiores probabilidades de saírem do experimento no meio.

contaminação pode ser que os membros do grupo experimental ensinem aos membros do grupo de controle algumas das técnicas às quais eles estão sendo submetidos. Isso é obviamente muito fácil em educação, onde membros do grupo experimental passam o material e/ou lições que receberam para membros do grupo de controle. Em computação, em alguns casos pode haver contaminação. O capítulo 7 de [Collins and Pinch 1998] discute o caso de contaminação em testes clínicos de remédios para AIDS - os pacientes dividiam as suas doses em 2 e trocavam com outros pacientes, para diminuir a chance que tivessem recebido o placebo (ver abaixo).

comportamento competitivo pode ser que os membros do grupo de controle se sintam preteridos frente aos do grupo experimental, e podem se mostrar motivados a competir com o grupo experimental ("para mostrar que nós somos bons mesmo que não nos tenham dado o CASE").

comportamento compensatório pode ser que alguma autoridade sinta que o grupo de controle foi preterido e crie medidas compensatórias para o grupo.

Finalmente, existem outras ameaças não facilmente ilustradas pelo exemplo. Elas são:

regressão à média A regressão à média é um efeito difícil de explicar e se baseia em usar o pré-teste para selecionar o grupo experimental. Vamos usar um exemplo de educação e não de computação para ilustrar esse efeito. Se você fizer duas provas, e olhar apenas para os 30% piores alunos na primeria prova, a média deles na segunda prova será melhor que a média na primeira - independentemente do que você fizer no meio. Esta é a regressão para a média, e ela é explicada pela idéia que as provas são medidas com erro do "verdadeiro conhecimento" do aluno. Assim as duas medidas serão diferentes por uma quantidade aleatória, mas como você selecionou os 30% piores na primeria prova, e para alguns deles a nota não pode ser mais baixa (pois eles tiram zero, por exemplo), então mais alunos do grupo terão as notas aleatoriamente maiores que menores. Dessa forma, se os alunos foram escolhidos para serem do grupo experimental por causa da sua nota no pré-teste, então o experimentador não saberá dizer se o efeito de melhora da nota é devido à intervenção ou à regressão a média.

efeito de expectativa do sujeito - efeitos placebo e hawthorne O efeito placebo é muito conhecido na medicina, e diz que se você der um remédio

inócuo (por exemplo uma pílula feita de farinha) para um paciente, ele vai dizer que seus sintomas melhoraram. A expectativa que o paciente tem de melhorar (porque ele tomou o que ele acha que é um remédio) causa a melhora! Um efeito similar ao placebo e talvez mais relevante para a experimentos em CC é o efeito Hawthorne, que diz que pode haver um efeito positivo apenas pelo fato dos sujeitos saberem que estão sendo estudados/observados. O caso que gerou a teoria do efeito foi numa fábrica chamada Hawthorne, onde se estudou o efeito de níveis de iluminação na produtividade dos trabalhadores, e descobriu-se que a produtividade dos trabalhadores aumentava independentemente de mudanças no nível de luminosidade - a teoria é que sabendo que estavam sendo observados melhorou a produtividade dos trabalhadores. Em experimentos computacionais, o efeito Hawthorne pode ser relevante e precisa ser levado em consideração - engenheiros de software que sabem que estão sendo estudados ou observados podem melhorar sua produtividade ou a qualidade do software gerado, alunos podem melhorar seu aprendizado, usuários podem melhorar seu desempenho, etc.

efeito de expectativa do experimentador O efeito de 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 pelo menos nos testes realizados pelo sujeito). Um exemplo claro desse efeito acontece quando o pós-teste requer alguma avaliação subjetiva do pesquisador - se o pesquisador sabe de quem são os testes submetidos à intervenção, e ele espera que a intervenção seja positiva, então ele pode "melhorar" as notas do pós-teste, mesmo inconscientemente. Um outro exemplo do fenômeno, mesmo quando não há o componente subjetivo no pós-teste, é o efeito Pigmaleão ou Rosenthal em educação - quando professores foram (falsamente) informados que suas classes tinham alunos mais inteligentes que a média, os alunos tiveram resultados muito melhores que alunos similares, mas cujos professores não receberam a falsa informação. A teoria propõe que o experimentador passa sinais inconscientes que acabam influenciando os sujeitos. O efeito de expectativa do experimentador também pode ser muito relevante para experimentos em CC. Se o experimentador é o criador de um sistema, ele pode passar aos sujeitos sinais que indicam sua expectativa que seu sistema é útil e bom, e, é claro, se a avaliação do pós-teste tiver algum aspecto subjetivo (por exemplo, se for preciso classificar os erros de software em sérios ou não).

influência de parte da intervenção Este efeito não tem um nome-padrão, mas aparece em diferentes domínios com diferentes nomes. A idéia é que o efeito observado não é devido à intervenção como um todo, mas devido a apenas parte dela. Vamos ver um exemplo na área de sistemas de apoio à decisão. Tais sistemas fazem várias perguntas a seu usuário, e propõe uma resposta/solução, mas cabe ao usuário aceitá-la ou levá-la

em consideração quando tomar a sua decisão. Um exemplo típico são os sistemas para apoio ao diagnóstico - o sistema pede ao médico várias informações e retorna uma ou mais alternativas de diagnóstico, mas cabe ao médico fazer o diagnóstico final. Descobriu-se que parte do efeito desses sistemas é o chamado efeito de *checklist* [Wyatt 1998] porque o sistema faz todas as perguntas que considera potencialmente importantes. Isso força o médico a pensar nas alternativas que normalmente ele não pensaria quando está fazendo o diagnóstico.

Na literatura, desenhos de experimentos são ditos **verdadeiramente experimentais** se a seleção dos membros dos grupos de controle e do grupo experimental é feita de forma aleatória, o que é indicado por um A antes de cada grupo. A escolha dos grupos de forma aleatória elimina as ameaças de seleção e mortalidade seletiva.

O desenho experimental abaixo é considerado um dos mais completos, pois elimina quase todas as ameaças, com a exceção de contaminação, comportamentos competitivos e o efeito expectativa para sujeitos e experimentadores. Esse desenho é chamado de dois-grupos, apenas pós-teste, seleção aleatória. A análise estatística apropriada é alguma comparação entre o pós-teste do grupo experimental e do grupo de controle - por exemplo o teste t se as medidas são pelo menos intervalares.

$$\begin{array}{ccc} A & X & O \\ A & & O \end{array}$$

Em medicina, onde os efeitos de expectativa são muito importantes, usa-se o experimento **duplo cego**, com placebo, cujo desenho é:

$$\begin{array}{ccc} A & X & O \\ A & X_0 & O \end{array}$$

onde X_0 é uma intervenção inócua, o placebo, com a mesma apresentação que X. Como X e X_0 têm a mesma apresentação, os sujeitos não sabem se estão recebendo X ou X_0 e, portanto, isto elimina as ameaças de contaminação, comportamento competitivo e expectativa do sujeito. O pesquisador que está administrando a intervenção também não sabe se o sujeito está recebendo X ou X_0, e isso elimina a ameaça de expectativa do experimentador. Em uma primeira análise, não parece possível desenhar experimentos duplo cego em computação - como criar uma intervenção placebo X_0 com a mesma apresentação que a intervenção "correta" X? Mas há artigos que discutem como controlar alguns efeitos de expectativa em experimentos de engenharia de software [Silva and Travassos 2004]. Além do desenho de dois grupos, pós-teste apenas, há outros desenhos verdadeiramente experimentais, como o desenho de 4 grupos de Solomon.

Nem sempre é possível escolher os membros do grupo experimental e de controle de forma aleatória. Por exemplo, em pesquisa educacional, pode-se

escolher classes aleatoriamente, mas não alunos - todos os alunos de uma mesma classe recebem a mesma intervenção. Da mesma forma, programadores que trabalham num mesmo projeto não podem ser aleatoriamente atribuídos para usar ou não usar uma ferramenta CASE. Desenhos experimentais onde a seleção dos grupos não é aleatória são chamados de **quase experimentais**. Desenhos quase experimentais de uma forma ou de outra envolvem a idéia dos grupos de controle e experimentais, mas a seleção dos membros de cada grupo não é aleatória. Desenhos experimentais onde não há o grupo de controle, que é comparado com o grupo experimental, são chamados de **pré-experimentais** (alguns autores chamam esses desenhos de **não-experimentais**).

O desenho pré-experimental mais comum é o discutido acima, de um só grupo, com pré e pós-testes, que sofre de vários problemas de validade interna.

$$O \quad X \quad O$$

Outro desenho que é historicamente classificado como pré-experimental, embora possua grupo de controle, é o 2-grupos, pós-teste apenas, representado por:

$$\begin{array}{ccc} N & X & O \\ N & & O \end{array}$$

onde o N é apenas indicativo que a seleção não é aleatória. Numa interpretação ingênua, se O do grupo experimental é significativamente superior ao do grupo de controle, então se confirma que a intervenção foi eficaz. Mas como deve ficar claro, este desenho não controla as ameaças de seleção e mortalidade seletiva, e como qualquer desenho com grupo de controle, pode sofrer de contaminação, e de comportamentos competitivos e compensatórios.

De uma forma geral, desenhos pré-experimentais são considerados muito fracos e devem ser evitados. Os desenhos quase-experimentais são um equilíbrio interessante entre factibilidade e força do experimento, e provavelmente são os desenhos experimentais mais frequentes em computação.

O desenho quase-experimental mais comum é o pré-teste/pós-teste para grupos não-equivalentes, cuja representação é:

$$\begin{array}{cccc} N & O & X & O \\ N & O & & O \end{array}$$

A análise estatística deste desenho não é simples; o leitor deve consultar [Trochim 2006] e [Reichardt 1979].

Um desenho quase experimental curioso e representativo é a série temporal interrompida, representada abaixo. O curioso deste desenho é que embora haja apenas o grupo experimental, ele é usado como seu próprio controle, através das observações repetidas, tanto antes, como depois da intervenção.

$$O \quad O \quad O \quad X \quad O \quad O \quad O$$

A idéia da série temporal interrompida é que se as várias observações antes da intervenção e depois da intervenção têm um padrão claro, e são diferentes entre si, então a diferença pode ser atribuída à intervenção X. Os exemplos à esquerda na figura 5.1 ilustram a situação onde se pode assumir que X é a razão das diferenças medidas antes e depois; os exemplos no painel à direita ilustram situações onde não é possível dizer que X causou algum efeito mensurável. Outros desenhos na mesma linha da série temporal interrom-

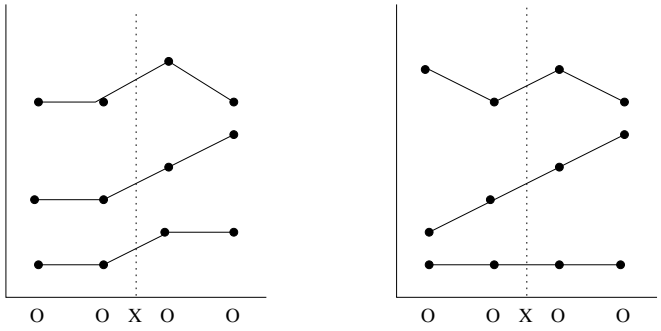

Figura 5.1. Dois tipos de resultados de um experimento de série temporal interrompida

pida são desenhos reversos, onde, por exemplo, depois de usar a ferramenta CASE e medir os erros, remove-se a ferramenta no próximo projeto. Há vários outros desenhos quase-experimentais e desenhos que combinam componentes quase-experimentais com seleção aleatória. O leitor deve consultar [Trochim 2006], por exemplo.

Todos os desenhos descritos acima, talvez com a exceção dos quase-experimentais baseados em séries temporais, são desenhos **transversais** (*cross sectional*), isto é, eles fazem um corte no tempo, e fazem as medidas nesse corte. Esse desenho pode não ser apropriado para responder questões que envolvam a evolução dos sujeitos. Desenhos experimentais que fazem várias medidas através do tempo em grupos de sujeitos são chamados de desenhos **longitudinais**. As ameaças à validade interna e a análise estatística de desenhos longitudinais não serão abordadas aqui.

As ameaças à validade externa são mais sutis, pois o conceito de validade externa não é tão bem definido. O objetivo final de um experimento é gerar um conhecimento que pode ser generalizado. Se o experimento mostra que houve uma diminuição estatisticamente significativa dos erros *pelo uso do CASE*, então espera-se que esse conhecimento possa ser generalizado para

"O uso daquela ferramenta CASE reduz o número de erros". O problema é que há duas generalizações na expressão acima:

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

A primeira generalização é a mais forte, e a mais comum quando se considera as ameaças à validade externa. A segunda generalização é mais fraca; pode-se mesmo pensar que ela é um caso particular da primeira no que se refere ao ambiente - generalização dos resultados de ambientes mais artificiais (o ambiente da experimentação) para ambientes mais naturais.

Vamos rapidamente tratar da segunda generalização, porque ela causa algumas confusões de nomenclatura. Por exemplo, alguns autores consideram que os efeitos de expectativa do sujeito e do observador são ameaças à validade externa e não à validade interna. Isso é justificado porque esses efeitos impedem a segunda generalização - são efeitos que aparecem porque foram obtidos em situações experimentais/artificiais.

A primeira generalização será menos "certa," quão mais especial for o grupo de pessoas que foram escolhidas para fazer parte do experimento, quanto mais especial for o local, o momento e o ambiente onde foram feitos os experimentos. Esses problemas são normalmente classificados como problemas de amostragem - como selecionar uma amostra de uma população, já discutido na seção de questionários.

5.2.4.1. O que fazer com isso?

Esta seção abordou alguns aspectos importantes de desenhos experimentais. A conclusão óbvia é que se deve usar experimentos duplo cego de dois grupos, com pós-teste apenas, já que esse desenho não parece sofrer de quase nenhuma ameaça à validade interna. Mas como vimos, nem sempre é possível usar esse desenho. O que se deve fazer é usar o mais forte desenho experimental possível (preferencialmente da familia dos quase-experimentais), e ter claro quais são as ameaças à validade interna e externa desse desenho, e se possível argumentar que algumas dessas ameaças não são relevantes para a situação em questão, e pensar em mecanismos para neutralizar as outras ameaças que podem ser relevantes.

5.3. Métodos Qualitativos

As fontes principais para essa seção são [Myers 1997, Yin 2005] e [Mays and Pope 1995a].

Numa primeira definição, métodos qualitativos diferem de métodos quantitativos porque se ocupam de variáveis que não podem ser *medidas*, apenas *observadas*. Essa é uma dicotomia muito simplista. Métodos qualitativos vêm das ciências sociais, em oposição aos métodos quantitativos que derivam das

ciências naturais. Essa diferença na origem já é suficiente para que visões diferentes sobre o que é ciência, e como se faz ciência, tornem definições suscintas sobre o que é um ou outro método muito difícil. De um modo geral, métodos qualitativos em CC são métodos que se caracterizam por ser um estudo aprofundado de um sistema no ambiente onde ele está sendo usado, ou, em alguns casos, onde se espera que o sistema seja usado. Métodos qualitativos sempre envolvem pessoas, e na maioria das vezes sistemas.

Contrário a fontes como [Myers 1997], que classifica a pesquisa qualitativa em 4 grupos, eu acho a divisão em apenas dois grupos mais produtiva: a **pesquisa observacional** e a **pesquisa-ação** (*action research*). A pesquisa observacional tem como objetivo observar o ambiente, mas não modificá-lo; já o objetivo central da pesquisa-ação é modificar o ambiente. É claro que só a presença do pesquisador causa alguma modificação no ambiente, mas essa modificação não é o objetivo da pesquisa observacional, e algumas variantes da pesquisa observacional tentam eliminar esse efeito.

Uma nota, antes da próxima seção: na discussão de métodos qualitativos usaremos alguns exemplos da área de "sistemas de informação" ou "gerência de sistemas de informação" (em inglês *information systems* ou *management information systems*). Essa área normalmente não se inclui nas áreas de pesquisa dos departamentos de ciência da computação no Brasil. A área tem um caráter mais aplicado, e estuda desde como resolver problemas tecnológicos práticos no uso e desenvolvimento de sistemas de informação, até os impactos econômico/financeiros e sociais desses sistemas nas organizações, e até os problemas na adoção ou desenvolvimento de novos sistemas. No Brasil, pesquisadores nessa área (que chamarei de sistemas de informação) se concentram em alguns departamentos de administração ou de engenharia de produção. Mas a área de sistemas de informação tem interfaces importantes com algumas subáreas tradicionais da CC, em particular as subáreas de sistemas colaborativos, engenharia de software, interfaces humano-computador e informática médica. Embora não seja uma subárea tradicional de CC, a literatura sobre métodos qualitativos em sistemas de informação é muito rica e merece ser lida.

5.3.1. Estudos observacionais

As principais referências para esta seção são [Mays and Pope 1995a], [Yin 2005], [Klein and Myers 1999] e [Dube and Pare 2003].

Segundo vários autores (por exemplo [Orlikowski and Baroudi 1991]) a pesquisa qualitativa onservacional pode ser dividida segundo a perspectiva filosófica ou epistemológica que a embasa em:

positivista

interpretativista

crítica

A perspectiva positivista tenta seguir de perto os fundamentos da pesquisa quantitativa, ou seja, que existem variáveis objetivas "no mundo," que embora

não possam ser medidas, podem ser observadas. Pesquisa com viés positivista tenta falar em teorias, em provar e desprovar essa teorias, etc. Um artigo que discute profundamente a pesquisa qualitativa positivista em sistemas de informação é [Dube and Pare 2003].

A perspectiva interpretativista propõe que não há variáveis objetivas, e que tudo que é observado depende de interpretação de um observador, e que diferentes pessoas não só observarão o mesmo "fato" de forma diferente, mas atribuirão valor a esse fato de forma diferenciada. A pesquisa em CC de fundo interpretativista tenta iluminar e elucidar as diferentes perspectivas/valores/interpretações das pessoas envolvidas com o sistema. Um artigo que discute profundamente a pesquisa qualitativa interpretativista é [Klein and Myers 1999].

A perspectiva crítica entende o mundo como a construção histórica e social de relações de poder e dominação. Nesta visão sistemas de informação provavelmente herdam da sociedade relações de poder, alienação e dominação, e revelar essas heranças é o objetivo central da pesquisa qualitativa de fundo crítico. [Myers and Young. 1997] é um bom exemplo de pesquisa qualitativa de fundo crítico em CC.

[Dube and Pare 2003] analisou 210 artigos de pesquisa observacional na área de sistemas de informação de 1990 a 2000, e descobriu que 87% deles seguem a perspectiva positivista, 12% a prespectiva interpretativista e apenas 1% a perspectiva crítica. Embora esses resultados sejam da decada de 1990, não acredito que a predominância da visão positivista tenha se alterado recentemente. Portando, o resto desta seção abordará essencialmente a perspectiva positivista.

A pesquisa observacional positivista é na sua maioria chamada de **descritiva** ou **exploratória**, isto é, busca descrever de forma "objetiva" e direta "eventos" e "fatos" de interesse. A pesquisa sobre canal de voz hierárquico na NASA [Watts et al. 1996] é um exemplo de pesquisa com o objetivo meramente descritivo. A pesquisa exploratória, além de descrever o fenômeno, faz propostas para novas teorias, ou novas observações, novas métricas para medir o fenômeno, etc. Finalmente, a pesquisa é **explanatória** se ela busca provar ou desprovar uma teoria particular. [Markus 1983] é considerado como uma das mais claras pesquisas explanatórias. O artigo confronta três teorias sobre resistência à implantação de um novo sistema computacional; uma delas diz que as pessoas serão contra um sistema se ele não corresponder com a visão que a pessoa tem da tarefa, de estilos cognitivos, etc. Entre os casos observados na pesquisa, há o de uma pessoa que era primeiramente favorável a um novo sistema, mas quando ela foi promovida passou a ser contra. Este exemplo contradiz com a teoria (se ela for uma teoria determinística), e é importante para ilustrar o que se chama de **controles naturais** na pesquisa explanatória - a busca de exemplos que se aproximam muito do poder de um experimento, mas não são artificialmente impostos.

5.3.1.1. Estilos da pesquisa observacional

Existem dois *estilos*[6] extremos de pesquisa observacional, com diferentes métodos e objetivos. Os estilos são chamados na literatura de:

- estudo de caso
- etnografia

O que mais caracteriza a separação em dois estilos é o nível de envolvimento entre o pesquisador e os sujeitos da pesquisa observacional. No estudo de caso, o pesquisador interage com os sujeitos de uma forma semi-formal, enquanto que em etnografia o pesquisador "vive e trabalha" junto com os sujeitos. Além do grau de envolvimento, normalmente os dois estilos usam de fontes de dados diferentes e têm objetivos diferentes, mas não existe uma fronteira clara entre eles.

O que caracteriza o estudo de caso é que o pesquisador interage com os sujeitos geralmente de uma forma semi-formal, através de entrevistas e conversas programadas, e normalmente tem acesso a documentos, dados e outros materiais formais da organização. O objetivo da pesquisa de estudo de caso é descobrir "o que as pessoas escrevem" e "o que as pessoas dizem", ou, em outras palavras, descobrir as práticas formais da organização e os valores, opiniões e atitudes dos sujeitos. O estudo de caso é um exemplo de "observador como participante" na classificação de [Gold 1958] sobre o grau de envolvimento do pesquisador e dos sujeitos. No grau "observador como participante" a interação é curta e semi-formal, e não há a formação de relacionamentos entre o pesquisador e os sujeitos.

Através da análise de documentos e de dados o pesquisador tem acesso a procedimentos formais da organização, aos tempos de duração de cada atividade, etc. que é muito relevante para a pesquisa em engenharia de software, por exemplo. Através de entrevistas com os sujeitos, o pesquisador pode descobrir aspectos importantes desses participantes, coisas que eles "querem e podem falar", ou seja, suas opiniões, seu valores, etc.

A etnografia é historicamente uma técnica usada na antropologia para entender culturas, geralmente primitivas. O pesquisador passa meses nessas culturas, vivendo com e como os sujeitos, para entender os valores e as práticas dessa cultura. Como ferramenta de pesquisa qualitativa em computação, a etnografia tem sido usada para descobrir "o que as pessoas fazem". Como na antropologia, o pesquisador passa um bom tempo com os sujeitos, e tenta conscientemente "fazer parte" do grupo. Em alguns casos o pesquisador aprende a fazer o que os sujeitos fazem, mas na maioria das vezes ele apenas observa os sujeitos no seu dia-a-dia (no trabalho). O etnógrafo pode ter acesso a documentos da organização, mas o que é central é que ele observe as pessoas trabalhando, e que interaja com elas, não só para entender o que elas estão fazendo, mas para criar um relacionamento de confiança e descontração

[6] A idéia de estilos não é mencionada na literatura em pesquisa qualitativa.

entre o pesquisador e os sujeitos. Na classificação de [Gold 1958], o método etnográfico é "participante como observador".

O objetivo central do estudo etnográfico é entender como as pessoas trabalham, e normalmente esse entendimento não é possível apenas entrevistando ou mesmo conversando com elas. Muitas das práticas de trabalho são tácitas, isto é, não são conscientes o suficiente para que as pessoas falem sobre elas, e se forçadas a fazê-lo elas acabam reinterpretando e filtrando muitas dessas práticas. Um exemplo exagerado é como explicar como andar de bicicleta.

É claro que muito raramente numa pesquisa de estilo etnográfico é preciso "viver e trabalhar" com os sujeitos por vários meses. [Hughes et al. 1994] fala em uma etnografia "rápida e suja" (*quick and dirty*) e outros falam em etnografia de curta duração (*short term*) para se referir à pesquisa etnográfica de dias ou semanas. A duração da etnografia deve ser suficiente para que dois objetivos tenham sido atingidos. O primeiro é que os sujeitos passem a estar confortáveis com a presença do pesquisador e voltem a fazer o que normalmente fazem - o efeito perturbador da presença do pesquisador na rotina dos sujeitos se dissipa quando eles se acostumam com o pesquisador. O segundo objetivo é que o pesquisador tenha observado um conjunto amplo o suficiente de comportamentos e práticas. Usa-se o termo **saturação** para indicar o momento quando tudo o que o etnógrafo observa já foi observado antes, e nesse momento ele tem alguma confiança (mas nunca a certeza) que já observou os comportamentos mais comuns.

Áreas da computação que estão interessadas em processos, decisões e opiniões optam por um estilo de estudo de caso. As áreas de engenharia de software, de sistemas de informação, de informática médica são exemplos onde os aspectos formais da organização são importantes, e, portanto, onde um estilo de estudo de caso é interessante. As áreas de sistemas colaborativos (CSCW) e interfaces (HCI) são exemplos onde é mais interessante saber o que as pessoas "realmente" fazem e como elas fazem isso, e não o que elas "dizem que fazem" ou o que o organograma da empresa "diz que elas fazem". Estas áreas usam mais o estilo etnográfico de pesquisa observacional.

5.3.1.2. Técnicas de pesquisa observacional

De modo geral, o problema central da pesquisa qualitativa é o rigor: como garantir que os dados e as conclusões obtidas são confiáveis, válidas e generalizáveis. Dentro da validade, a questão central é como controlar a **subjetividade** ou o **viés** do pesquisador: como saber se as conclusões realmente vêm dos dados ou se vêm de posições pré-experimentais do pesquisador.

O resto desta seção discutirá alguns métodos que tentam controlar algumas ameaças à confiabilidade e a validade da pesquisa qualitativa observacional. As técnicas abaixo são em principio relevantes tanto para estudos de caso como para etnografias.

amostragem fundamentada em teoria ou direcionada (*purposive* ou

***theorical sampling*)** A seleção das amostras em pesquisa qualitativa não é aleatória, mas busca especificamente casos extremos. A própria definição de quem será o próximo ambiente a ser observado pode ser determinado durante a pesquisa. Isto garante que as fontes mais diversas serão estudadas, e que a pesquisa cobre o espectro das possibilidades (mas sem nenhuma preocupação especial com o típico ou o representativo).

separação de observação e de teorização A coleta de dados e a teorização devem ser feitas em momentos independentes (embora seja permitido que aconteçam em ciclos de observação seguidos de teorização). O pesquisador deve anotar "tudo" que acontece na observação de campo, em cadernos que devem depois ser reanalisados.

teoria fundamentada em dados (*grounded theory*) É uma forma de análise de dados qualitativos (textos escritos, fala e entrevista dos participantes, etc.) que busca extrair dos próprios dados e de padrões repetitivos dos dados as teorias que explicam tais dados.

triangulação Na sua primeira acepção, triangulação é a utilização de várias fontes para o mesmo fato. Numa visão mais moderna, a triangulação consiste em buscar pelo menos duas formas/fontes para cada dado e análise da pesquisa. Pode-se usar mais de uma técnica de coleta de dados, por exemplo, análise de documentos e entrevistas, ou observação e questionários, ou pode-se usar mais de um pesquisador observando o ambiente. O uso de múltiplos pesquisadores é também chamado de **codificação múltipla**.

parceiro neutro Utilização de um pesquisador experiente não envolvido diretamente na pesquisa para validar e/ou criticar as conclusões do pesquisador principal. É similar à idéia de triangulação, mas centrado na análise e planejamento e não nas observações (como a codificação múltipla).

validação pelos sujeitos A validação pelos sujeitos consiste em mostrar os dados coletados e/ou a análise dos mesmos a alguns dos sujeitos da pesquisa, respeitando-se as questões éticas previamente combinadas. Além de permitir ao pesquisador verificar se suas anotações e conclusões são coerentes com o pensamento dos sujeitos, pode-se utilizar a técnica de retorno para os sujeitos como parte do processo de coleta, utilizando dados previamente coletados como fonte de discussão.

[Wilson 2006] mostra o uso de triagulação em pesquisa em interfaces; [de Souza et al. 2005] usa teoria fundamentada em dados para entender práticas de desenvolvimento de software.

5.3.2. Pesquisa-ação

A bibliografia para essa seção é [Avison et al. 1999, Baskerville 1999].

Pesquisa-ação (*action research*) é uma forma de pesquisa qualitativa que busca modificar o ambiente que está sendo estudado através da ação do pesquisador. O resultado da pesquisa-ação em computação é a descrição de um caso de tentativa (bem-sucedida ou não) de modificação de uma organização ou grupo através do desenvolvimento (opcional) e a implantação de um sistema (por exemplo [Lindgren et al. 2004]).

A pesquisa-ação é uma idéia desenvolvida na Psicologia, e depois adotada por várias ciências sociais aplicadas, onde deixa-se a postura do cientista que observa e aprende observando pela do cientista que atua, modifica e aprende dessa ação. A pesquisa-ação carrega um forte componente ideológico, já que quase sempre atrás da ação de modificação está uma visão de como "tornar as coisas melhores". Palavras como "participativo" "democrático" e "justo" são quase sempre associadas a pesquisa-ação em áreas como educação, problemas sociais, etc.

De uma forma geral, na pesquisa-ação espera-se que o pesquisador interaja com os sujeitos (ou a organização). Da interação surge uma definição de quais são os problemas que devem ser resolvidos. Num segundo momento, tanto o pesquisador, como os sujeitos trazem diferentes formas de teorias e conhecimentos para a criação da solução. Essa solução é posta em prática e analisada/avaliada - deu certo ou não? resolveu o que tinha sido definido como o problema a ser resolvido? criou outros problemas?, etc. Da análise dos resultados, os participantes (pesquisador e sujeitos) devem reavaliar suas teorias e conhecimentos, que pode gerar um novo ciclo. O novo ciclo pode começar de uma nova definição do problema ou de uma nova definição da solução.

[Baskerville 1999] define as seguintes etapas de uma pesquisa-ação:

infra-estrutura cliente-sistema a definição de um acordo/contrato entre o pesquisador e a organização (ou comunidade) sobre o escopo da pesquisa, os papéis que cada um assumirá, etc.

diagnóstico a definição conjunta e colaborativa do que é o problema a ser resolvido.

planejamento da ação construção da solução que espera-se resolverá o problema.

tomada da ação a implantação da solução.

avaliação a análise/avaliação dos resultados da ação.

aprendizado a adaptação das teorias que foram usadas para formular a solução, tendo em vista a avaliação.

Descrita dessa forma, a pesquisa-ação não parece diferir de outras duas atividades que envolvem cientistas da computação: desenvolvimento de sistemas e consultorias. Uma forma de desenvolvimento de sistemas mais participativa também envolve a definição junto com os sujeitos dos requisitos do sistema (a solução), a implementação do sistema e avaliação dos resultados. As diferenças centrais parecem ser, tanto de enfoque, como de postura:

- desenvolvimento de sistemas normalmente não começam com a fase de diagnóstico, muito menos diagnóstico participativo - o problema a ser resolvido normalmente é definido antes que o processo de desenvolvimento de sistemas comece
- desenvolvimento de sistemas necessariamente acredita que a solução é desenvolver um sistema novo, enquanto que a ação em pesquisa-ação pode não necessariamente ser um sistema novo
- desenvolvimento de sistemas encara a avaliação como um teste - o sistema consegue ou falha em resolver o problema (e portanto existe um viés para que a avaliação seja positiva!). Em pesquisa-ação a avaliação deve ser tão neutra quanto possível - em princípio o que está sendo avaliado não é a ação mas as teorias que levaram ao planejamento daquela e não outras ações
- não há muita preocupação com o aprendizado participativo no desenvolvimento de sistemas - os desenvolvedores provavelmente aprenderão muito com o desenvolvimento de sistemas (e em alguns casos o aprendizado é formalizado com reuniões de finalização de projeto, coleta de estatísticas, etc.) mas há pouco retorno para os sujeitos do sistema.

Mesmo com as diferenças em postura e ênfase, artigos que discutem "lições aprendidas na prática" do desenvolvimento de um sistema, mesmo que ele não tenha o caráter participativo desejável, podem ser considerados como uma forma de pesquisa-ação, segundo [Avison et al. 1999], por exemplo.

Consultorias feitas por pesquisadores em ciência da computação se aproximam mais da pesquisa-ação no que se refere ao diagnóstico, que normalmente é parte do processo de consultoria, mas se afasta no que se refere ao caráter participativo do planejamento da ação. Normalmente em consultorias as organizações esperam que os consultores proponham a solução baseada na sua experiência prévia, em vez de construir a solução com os membros da organização.

5.3.3. Outras formas qualitativas de avaliação

A principal referência para esta seção é [Wyatt and Friedman 1997]

Quando se fala de avaliação de um sistema, outros métodos qualitativos são possíveis e já foram usados em publicações científicas. Mencionaremos brevemente os seguintes métodos de avaliação, retirados de [House 1980]:

- avaliação como crítica artística
- avaliação por comitê de especialistas (*professional review*)

A idéia da avaliação como crítica artística é chamar um especialista na área e fazer com que esse especialista use o programa ou o sistema e expresse sua opinião sobre a experiência de usá-lo. Essa forma de avaliação assemelha-se à crítica artística - o crítico que é um especialista na área e que tem uma percepção refinada e experiente avalia, segundo sua visão, o que é usar o sistema,

da mesma forma como um crítico artístico avalia o que foi ler o livro, ou ver o filme. A perspectiva filosófica é essencialmente interpretativista: as pessoas têm opiniões e valores sobre o que é bom e ruim sobre o programa/sistema, e aceita-se que diferentes pessoas tenham diferentes visões, assim como dois críticos podem divergir na avaliação de um filme. O uso de avaliação como crítica artística é mais comum em áreas como interface humano-computador; em particular [Bertelsen and Pold 2004] argumenta que a área deveria fazer mais uso desse método.

A avaliação por especialistas se aproxima mais da pesquisa qualitativa observacional como descrita acima, mas em vez de usar um pesquisador qualitativo experiente, que usa várias técnicas para manter sua subjetividade sob controle, usa-se um grupo de especialistas. Espera-se que por ser um grupo, as divergências entre eles limitem o efeito das várias subjetividades, e que por serem especialistas, os avaliadores não se prendam apenas ao que é visível e óbvio.

5.3.4. Publicação de pesquisa qualitativa

O pesquisador que usa métodos qualitativos tem que ter alguns cuidados na hora de publicar seus resultados. O primeiro problema de publicar os resultados de uma pesquisa qualitativa é defender o mérito dos métodos qualitativos, se a publicação é em uma área da CC que valoriza a pesquisa quantitativa. [Mays and Pope 1995b] e [Pope and Mays 1995] são exemplos de dois artigos que fazem essa defesa no campo da pesquisa em saúde, que é uma das mais rigorosas quanto a seus métodos. O problema parece ser maior se o pesquisador pretende publicar artigos de pesquisa-ação - é preciso mostrar que pesquisa-ação não é apenas o desenvolvimento de um sistema, ou o resultado de uma consultoria, que normalmente não são publicados como artigos científicos.

Mesmo que a área de CC já aceite bem pesquisa qualitativa, o segundo cuidado é quase sempre necessário. Na pesquisa quantitativa, a competência do pesquisador é central na escolha do desenho experimental e na escolha do teste estatístico, mas "fazer" pesquisa é essencialmente uma tarefa mecânica. E a qualidade da pesquisa pode ser julgada apenas pela escolha do desenho experimental e dos testes estatísticos. Mas em pesquisa qualitativa, a competência do pesquisador é necessária durante o "fazer" da pesquisa. O pesquisador qualitativo tem que manter suas idéias pré-pesquisa controlados, tem que separar a coleta de dados da teorização, tem que estar aberto e aproveitar as oportunidades que aparecem durante a pesquisa, tem que ter habilidade de entrevistar os participantes, etc. Se o pesquisador não é competente, fica difícil acreditar nos resultados da pesquisa, *sejam eles quais forem*.

O pesquisador qualitativo precisa então mostrar evidências no artigo que é um pesquisador competente! O texto da publicação tem que ser rico o suficiente para dar confiança ao revisor e aos leitores que a pequisa foi feita por alguém competente. Parte desse processo de convencer os revisores e leito-

res é mostrar que o pesquisador está ciente da literatura em metodologia de pesquisa qualitativa. Em alguns casos o pesquisador precisa não só declarar o enfoque filosófico da pesquisa, mas também ligar a sua pesquisa com as teorias fundamentais da pesquisa qualitativa (por exemplo [Butler 2000] centra uma pesquisa sobre CASE na teoria hermenêutica). Teorias como fenomenologia, etnometodologia, interacionismo simbólico, construtivismo e hermenêutica (ver [Flick et al. 2004] para capítulos curtos sobre essas teorias) são consideradas como teorias que fundamentam a pesquisa qualitativa, pois de uma forma ou outra elas definem o que é o "pessoal" e o "social," e qual a fronteira entre eles. O autor não entende o suficiente sobre essas teorias para ser mais claro.

Artigos como [Dube and Pare 2003] e [Klein and Myers 1999] discutem, não só como executar uma pequisa qualitativa rigorosa, mas como *relatá-la* de forma rigorosa.

5.4. Ética em pesquisa em computação

Pesquisa qualitativa, questionários e experimentos necessariamente envolvem seres humanos. Em outras áreas, em particular nas ciências da saúde, há um conjunto quase consensual de quais são os padrões éticos para pesquisas que envolvem seres humanos. O autor não conheçe nenhuma discussão particular para as questões éticas relativas à pesquisa em Ciência da Computação, então, em vez de listar qual é o comportamento considerado correto, iremos apenas listar as questões que parecem pouco claras. Enquanto a área não definir padrões de comportamento ético em pesquisa, cada pesquisador terá que refletir e agir segundo suas convicções nessas e outras questões[7].

Quanto à participação em experimentos:

- O sujeito de um experimento em ciência da computação deve ser informado que ele participa de um experimento ou isso não é necessário?
- Se ele tiver que ser informado, é preciso que ele o seja antes e concorde em participar do experimento, ou só é preciso que ele aprove, após o experimento, que os dados sejam utilizados na pesquisa, desde que certas salvaguardas sejam tomadas?

Em ciências da saúde, exige-se que o sujeito seja informado e concorde *antes* de participar do experimento. Mais do que isso, exige-se que o sujeito assine um termo de consentimento informado, onde deve haver salvaguardas sobre a possibilidade do sujeito decidir sair do experimento no meio dele, sobre as responsabilidades, riscos e custos de participar do experimento, etc. O princípio por trás desta exigência é que um experimento em saúde pode de várias

[7] Infelizmente talvez a Ciência da Computação não tenha tempo de definir seus próprios padrões de ética em pesquisa. Agencias financiadoras de pesquisa como a FAPESP de São Paulo exigem que pesquisas que envolvam seres humanos tenham sido aprovadas pelas comissões de ética em pesquisa das respectivas universidades. Mas essas comissões têm uma tradição das ciências da saúde, e que, como veremos, têm padrões éticos que não necessariamente se aplicam a pesquisa em ciência da computação.

formas “fazer mal” ao paciente: uma droga que não tem o efeito esperado e, portanto, não cura o paciente, efeitos colaterais ainda desconhecidos de uma droga experimental, sofrimento físico devido a tratamentos e exames, etc. Por outro lado, em documentários, normalmente não se informa ao sujeito que ele está participando de um documentário. Apenas depois das filmagens o sujeito é informado e pede-se que ele consente na divulgação de sua imagem.

A questão então é se um experimento de computação se parece mais com um experimento em saúde ou com um documentário. Pode um experimento em computação “fazer mal” ao sujeito?

Pesquisas em saúde exigem que os resultados da pesquisa não tenham nenhuma forma de identificação dos pacientes. A divulgação de informação sobre a saúde de uma pessoa pode ser prejudicial a ela - pode dificultar obter seguro-saúde, ou pode fazê-la perder o emprego, ou ser ostracisada por seu grupo social, etc. Pesquisadores em saúde normalmente dão garantias de anonimato ao paciente nos resultados publicados da pesquisa e pedem o consentimento para o uso dos dados. Em computação, a divulgação de informações é também prejudicial ao sujeito, de tal forma que tanto garantias de anonimato como permissão de uso são necessários?

Quanto a pesquisas qualitativas, normalmente as organizações autorizam tais pesquisas depois de certas negociações. Nessas negociações o pesquisador terá de enfrentar questões como:

- que garantias de anonimato da organização na publicação final dos resultados são apropriadas? Pesquisa qualitativa deve ser muito explícita na descrição do ambiente que foi estudado e de certa forma isso conflita com as demandas de anonimato da organização
- a organização tem poder de veto na publicação dos resultados? Pesquisa qualitativa pode tanto revelar problemas quanto vantagens competitivas das organizações, e a organização pode temer a revelação de tais informações, mesmo com as garantias de anonimato acordadas
- se a organização já autorizou a pesquisa, é preciso pedir consentimento de cada um dos sujeitos estudados? Isto é, se a organização autorizou o pesquisador a entrevistar os seus funcionários, é preciso separadamente pedir o consentimento de cada um dos funcionários entrevistados?
- um problema real em computação é que sistemas acabam por substituir pessoas. Se a pesquisa qualitativa é feita com a intenção de especificar um sistema, o pesquisador é eticamente obrigado a informar aos sujeitos da pesquisa que talvez seus empregos estejam ameaçados?

5.5. Considerações finais

Este capítulo apresentou vários conceitos e técnicas associadas a pesquisa quantitativa e qualitativa que podem e devem ser aplicados em Ciência da Computação. O autor acredita que este texto cobre em largura e em profundidade questões metodológicas que não são discutidas em computação, mas

que o texto não é auto-suficiente. Se o pesquisador, por exemplo, decide usar pesquisa-ação como forma de pesquisa, ele deve buscar nas fontes citadas um maior aprofundamento no assunto. Isso vale para todas as técnicas apresentadas aqui.

O autor acredita que rigor científico é algo definido essencialmente pela comunidade científica. Um pesquisador deve no mínimo seguir métodos definidos como padrão pela sua comunidade científica, mas é quase sempre vantajoso usar métodos mais rigorosos que os da comunidade. Isso não só melhora as chances de aceitação do trabalho, mas educa e "melhora" a própria comunidade. Se a sua subárea dentro da computação não usa comumente significância estatística, então introduzi-la nos seus artigos tem o duplo benefício de melhorar o seu trabalho e melhorar a comunidade. Por outro lado, um pesquisador terá mais dificuldades em ter trabalhos aceitos se ele é menos rigoroso que a sua comunidade como um todo.

Finalmente, este capítulo reflete a experiência/história do autor. Em particular, o autor está ciente que não há referências suficientes para artigos nas áreas de desempenho (*performance*), que têm uma tradição de uso de métodos quantitativos e de uso de técnicas estatísticas, e na área de interface humano-computador, que é uma área que usa muito experimentos, questionários, pesquisa qualitativa observacional, etc., além de ter uma tradição de discussão metodológica (discussão sobre os próprios métodos de pesquisa) talvez única na computação. Infelizmente o autor não tem um conhecimento dessas áreas que lhe permitisse escolher mais (ou melhores) referências bibliográficas.

5.5.1. Agradecimentos

Cada uma das seções deste capítulo tem a colaboração de alunos da disciplina MO901 "Questões epistêmicas e metodológicas em Ciência da Computação" oferecida no 2o semestre de 2006, no Instituto de Computação da UNICAMP. A seção 5.2.1 contou com a colaboração de Danilo Lacerda, a seção 5.2.2 com a de Leandro Rodrigues Magalhães de Marco, a seção 5.2.3 com a de Fábio Bezerra, a seção 5.2.4 com a de Patricia Rocha de Toro, e a seção 5.3.1 com a de Claudia Galindo Brasotini e Vania Paula de Almeida Neris.

O autor gostaria também de agradecer aos alunos André Covic Bastos, Denis Neves de Arruda Santos e Paulo Lopes, cujas contribuições não foram incluídas neste texto.

Referências bibliográficas

[Avison et al. 1999] Avison, D. E., Lau, F., Myers, M. D., and Nielsen, P. A. (1999). Action research. *Communications of the ACM*, 42(1):94–97.

[Baskerville 1999] Baskerville, R. L. (1999). Investigating information systems with action research. In *Communications of the Association for*

Information Systems, volume 2. Association for Information Systems. http://cis.gsu.edu/ rbaskerv/CAIS_2_19/CAIS_2_19.html.

[Bentley et al. 1992] Bentley, R., Hughes, J. A., Randall, D., Rodden, T., Sawyer, P., Shapiro, D., and Sommerville, I. (1992). Ethnographically-informed systems design for air traffic control. In *CSCW '92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 123–129, New York, NY, USA. ACM Press.

[Bertelsen and Pold 2004] Bertelsen, O. W. and Pold, S. (2004). Criticism as an approach to interface aesthetics. In *NordiCHI '04: Proceedings of the third Nordic conference on Human-computer interaction*, pages 23–32, New York, NY, USA. ACM Press.

[Box et al. 1978] Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley, New York.

[Brynjolfsson and Hitt 1998] Brynjolfsson, E. and Hitt, L. M. (1998). Beyond the productivity paradox. *Commun. ACM*, 41(8):49–55.

[Butler 2000] Butler, T. (2000). Transforming information systems development through computer-aided systems engineering (case): lessons from practice. *Information Systems Journal*, 10(3):167–193.

[Castro 2007] Castro, A. A. Curso de revisão sistemática e metanálise. http://www.virtual.epm.br/cursos/metanalise. Acessado em 1/2007.

[Chen and Rada 1996] Chen, C. and Rada, R. (1996). Interacting with hypertext: A meta-analysis of experimental studies. *Human-Computer Interaction*, 11(2):125–156.

[Collins and Pinch 1998] Collins, H. and Pinch, T. (1998). *The Golem at large: what you should know about technology*. Cambridge University Press.

[Czitrom 1999] Czitrom, V. (1999). One-factor-at-a-time versus designed experiments. *American Statistician*, 53(2):126–131.

[de Souza et al. 2005] de Souza, C., Froehlich, J., and Dourish, P. (2005). Seeking the source: software source code as a social and technical artifact. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 197–206, New York, NY, USA. ACM Press.

[Denning 2005] Denning, P. J. (2005). Is computer science science? *Commun. ACM*, 48(4):27–31.

[Dewan and Kraemer 1998] Dewan, S. and Kraemer, K. L. (1998). International dimensions of the productivity paradox. *Commun. ACM*, 41(8):56–62.

[Dube and Pare 2003] Dube, L. and Pare, G. (2003). Rigor in information system positivist case research: current practices, trends and recommendations. *MIS Quarlerly*, 27(4):597–635.

[Flick et al. 2004] Flick, U., von Kardoff, E., and Steike, I., editors (2004). *A*

Companion to Qualitative Research. Sage.

[Brooks 1996] Brooks, F. P. (1996). The computer scientist as toolsmith II. *Commun. ACM*, 39(3):61–68.

[Gold 1958] Gold, R. (1958). Roles in sociological field investigation. *Social Forces*, 36:217–223.

[Guba and Lincon 1981] Guba, E. G. and Lincon, Y. S. (1981). *Effective evaluation*. Jossey-Bass.

[Hochman et al. 2005] Hochman, B., Nahas, F. X., Oliveira, R. S., and Ferreira, L. M. (2005). Desenho de pesquisa. *Acta Cirurgica Brasileira*, 20(2).

[Holte 1993] Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90.

[House 1980] House, E. R. (1980). *Evaluating with validity*. Sage.

[Huff 1990] Huff, S. (1990). Information systems maintenance. *The Business Quarterly*, 55:30–32.

[Hughes et al. 1994] Hughes, J., King, V., Rodden, T., and Andersen, H. (1994). Moving out from the control room: ethnography in system design. In *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 429–439, New York, NY, USA. ACM Press.

[Hundhausen et al. 2002] Hundhausen, C. D., Douglas, S. A., and Stasko, J. T. (2002). A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing*, 13(3):259–290.

[Kirakowski 2007] Kirakowski, J. Sumi: Software usability measurement inventory. http://sumi.ucc.ie/. Acessado em 3/2007.

[Kitchenham and Pfleeger 2002a] Kitchenham, B. and Pfleeger, S. (2002a). Principles of survey research: part 2: designing a survey. *ACM SIGSOFT Software Engineering Notes*, 27(1):44–45.

[Kitchenham and Pfleeger 2002b] Kitchenham, B. and Pfleeger, S. (2002b). Principles of survey research: part 4: questionnaire evaluation. *ACM SIGSOFT Software Engineering Notes*, 27(3):44–45.

[Kitchenham and Pfleeger 2002c] Kitchenham, B. and Pfleeger, S. (2002c). Principles of survey research: part 6: data analysis. *ACM SIGSOFT Software Engineering Notes*, 28(2):24–27.

[Kitchenham et al. 2002] Kitchenham, B., Pfleeger, S. L., Pickard, L., Jones, P., Hoaglin, D., Emam, K. E., and Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering,*, 28(8):721–734.

[Klein and Myers 1999] Klein, H. K. and Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, 23(1):67–93.

[Koskinen 2007] Koskinen, J. Software maintenance costs. http://www.cs.jyu.fi/ koskinen/smcosts.htm. Acessado em 1/2007.

[Lew 2006] Lew, M. J. (2006). Principles: When there should be no difference - how to fail to reject the null hypothesis. *Trends in Pharmacological Sciences*, 27(5):274–278.

[Lindgren et al. 2004] Lindgren, R., Henfridsson, O., and Schultze, U. (2004). Design principles for competence management systems: A synthesis of an action research study. *MIS Quarterly*, 28(3):435–472.

[Marczyk et al. 2005] Marczyk, G., DeMatteo, D., and Festinger, D. (2005). *Essentials of Research Design and Methodology*. John Wiley and Sons.

[Markus 1983] Markus, M. L. (1983). Power, politics, and mis implementation. *Commun. ACM*, 26(6):430–444.

[Mays and Pope 1995a] Mays, N. and Pope, C. (1995a). Qualitative research: Rigour and qualitative research. *British Medical Journal*, 311:109–112.

[Mays and Pope 1995b] Mays, N. and Pope, C. (1995b). Qualitative Research: Rigour and qualitative research. *BMJ*, 311(6997):109–112.

[Mezard et al. 2002] Mezard, M., Parisi, G., and Zecchina, R. (2002). Analytic and Algorithmic Solution of Random Satisfiability Problems. *Science*, 297(5582):812–815.

[Myers 1997] Myers, M. D. (1997). Qualitative research in information systems. In *MISQ Discovery,*, volume 2. MIS Quarterly. http://www.qual.auckland.ac.nz/.

[Myers and Young. 1997] Myers, M. D. and Young., L. W. (1997). Hidden agendas, power, and managerial assumptions in information systems development: An ethnographic study. *Information Technology & People*, 10(3):224–240.

[Newell and Simon 1976] Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19(3):113–126.

[Orlikowski and Baroudi 1991] Orlikowski, W. and Baroudi, J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information Systems Research*, 2(1):1–28.

[Pfleeger and Kitchenham 2001] Pfleeger, S. and Kitchenham, B. (2001). Principles of survey research: part 1: turning lemons into lemonade. *ACM SIGSOFT Software Engineering Notes*, 26(6):44–45.

[Pope and Mays 1995] Pope, C. and Mays, N. (1995). Qualitative Research: Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *BMJ*, 311(6996):42–45.

[Reichardt 1979] Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group design. In Cook, T. D. and Campbell, D. T., editors,

Quasi experimentation: design and analysis issues for field studies. Rand McNally.

[Shneiderman 1992] Shneiderman, B. (1992). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley, 2nd edition.

[Silva and Travassos 2004] Silva, L. and Travassos, G. (2004). Tool-supported unobtrusive evaluation of software engineering process conformance. In *Proceedings. 2004 International Symposium on Empirical Software Engineering, 2004. ISESE '04.*, pages 127 – 135.

[Tichy 1998] Tichy, W. F. (1998). Should computer scientists experiment more? *Computer*, 31(5):32–40.

[Toothaker 1993] Toothaker, L. E. (1993). *Multiple Comparison Procedures*. Sage Publications.

[Trochim 2006] Trochim, W. M. Research methods knowledge base. http://www.socialresearchmethods.net/kb/. Acessado em Jan 2007.

[Wainer 2003] Wainer, J. (2003). O paradoxo da produtividade. In Ruben, G., Wainer, J., and Dwyer, T., editors, *Informática, Organizações e Sociedade no Brasil*, pages 7–55. Cortez.

[Watts et al. 1996] Watts, J. C., Woods, D. D., Corban, J. M., Patterson, E. S., Kerr, R. L., and Hicks, L. C. (1996). Voice loops as cooperative aids in space shuttle mission control. In *CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pages 48–56, New York, NY, USA. ACM Press.

[Wilson 2006] Wilson, C. E. (2006). Triangulation: the explicit use of multiple methods, measures, and approaches for determining core issues in product development. *interactions*, 13(6):46–ff.

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

[Wyatt and Friedman 1997] Wyatt, J. and Friedman, C. P. (1997). *Evaluation Methods in Medical Informatics*. Springer.

[Yin 2005] Yin, R. K. (2005). *Estudo de Caso: Planejamento e Metodos*. Bookman, 3a edição edition.