SCC 5933 - METODOLOGIA DE PESQUISA CIENTÍFICA EM COMPUTAÇÃO

MÉTODOS DE PESQUISA QUANTITATIVA E QUALITATIVA PARA A CIÊNCIA DA COMPUTAÇÃO

Profa. Sandra M Aluisio

Avaliação

- Prova (30/6) e
- Avaliação do Site para a Pesquisa do Aluno (30/6)

- Disponibilizar o **link do site num arquivo**, via escaninho

- **Aprovação: C**[5 , 7) **B**[7 , 8.0) **A**[8.0 , 10]
- (**Frequência ≥ 75%**).

Elementos do Site

- Título da Pesquisa
- Tema
- Lacuna/problema
- Hipóteses e Objetivo
- Justificativa/motivação
- Resumo com **Estruturação Explícita das partes componentes**
- Metodologia de Desenvolvimento dos métodos (ou dos sistemas)
- Metodologia de Avaliação
- **Abas:**
  - Equipe, Contato, Duração da Pesquisa
  - Publicações e Slides ou Pôsteres, Demos/Pilotos
  - Monografia de Qualificação, Dissertação/Tese
  - Dados/Recursos ou Benchmarks criados, Links Interessantes

- **Métodos de pesquisa quantitativa e qualitativa para a Ciência da Computação**

  Jacques Wainer

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

  - pesquisa analítica ➔ provas matemáticas, análise assintótica
  - pesquisa quantitativa } EMPÍRICA ou Experimental
  - pesquisa qualitativa
  - pesquisa bibliográfica ➔ revisão sistemática

Métodos qualitativos

- Observação cuidadosa:
  - dos ambientes onde o sistema proposto é ou será usado
  - das perspectivas dos usuários ou potenciais usuários
- Métodos: estudos qualitativos observacionais, pesquisa-ação (intervencionistas)

- Em HCI
  - Método de Inspeção Semiótica (MIS) [de Souza et al. 2010]

    de Souza, C. S., Leitão, C. F., Prates, R. O., Bim, S.A., da Silva, E.J. (2010). Can inspection methods generate valid new knowledge in HCI? The case of semiotic inspection. In: International Journal of Human-Computer Studies, pp. 22-40 (2010).

    http://www.repositorio.ufop.br/bitstream/123456789/4410/1/ARTIGO_CanInspectionMethods.pdf

Os métodos quantitativos

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”.

- Medida numérica de poucas variáveis objetivas, enfatizando a comparação de resultados e uso intensivo de métodos estatísticos
- Métodos:

(1) uso de dados sintéticos* : *benchmarks*, simulações e competições ➔ shared tasks da CONLL usa ML para PLN (http://ifarm.nl/signll/conll/) – dados são reais

Três classes de benchmarks:

- para avaliar o tempo de execução do programa.
- para avaliar se um programa consegue obter um resultado e resultam num conjunto de medidas binárias (resolveu ou não resolveu o problema).
  - especificam não só exemplos de problemas, mas também sua solução.
- para avaliar a qualidade da resposta do programa (resposta binária ou medida de erro)

(2) técnicas estatísticas para a comparação de conjuntos de medidas

(3) uso de questionários (*surveys*)

(4) desenhos experimentais

* devem, em princípio, representar a possível diversidade de dados reais

EMNLP 2015 Workshop on Discourse in Machine Translation

- http://www.idiap.ch/workshop/DiscoMT/shared-task
- Important dates
- February 2015 **Training data** release
- 4 May 2015 Release of **test data** for pronoun translation task (não tem a classe)
- 10 May 2015 Submission deadline for pronoun translation task
- 11 May 2015 Release of **test data** for cross-lingual pronoun prediction task (não tem a classe)
- 18 May 2015 Submission deadline for cross-lingual pronoun prediction task
- 28 June 2015 System **paper submission deadline**
- 21 July 2015 Notification of acceptance
- 11 August 2015 Camera-ready papers due
- September 2015 DiscoMT 2015 workshop in Lisbon (in conjunction with EMNLP)
- ➔ Métricas de avaliação são comuns a todos os competidores e assim a competição avança o estado da arte de tarefas. Baselines são oferecidos; ideal os sistemas superarem esses.
- **Evaluation**
- The **classification** results will be evaluated against the **gold standard translations** from the test set. For the **pronoun-focused translation task**, the submissions will be **scored manually**.

Foco: métodos quantitativos

- Usam intensivamente métodos estatísticos.
  - Essência: verificar quão melhor é nosso método proposto frente a alternativas, usando métricas como P, R, F-measure, acurácia, tempo.
  - Bastante usados na CC
- Essa é a razão das hipóteses virem já associadas com as medidas de avaliação.

O que fazer com isso?

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

- Pesquisa quantitativa mede as variáveis de interesse objetivamente (medidas ou observadas).
  - Tempo de Execução (dado real)
  - Programa acertou ou não a resposta (dado binário)
  - Em NLP, por exemplo,
    - desejamos saber se um novo método proposto tem desempenho significativamente melhor, com relação a métricas padrão da área (P, R, F-measure), que abordagens anteriores.
  - Temos classes de medidas: **categóricas (ou nominais), ordinais, intervalares, medidas de razão**

Classes de Medidas

- IMPORTANTE: define que tipo de teste estatístico usar para verificar se 2 conjuntos são significativamente diferentes ou não.

- Medidas Categóricas (ou Nominais): sexo, estado, país, diagnóstico médico.
  - Não há operações com esse dado.
- Medidas ordinais: classe socioeconômicas, série escolar, gravidade.
  - Operação de ordenação
- Medidas intervalares: medida em célsius.
  - O intervalo pode ser comparado. **o valor nulo não corresponde à ausência da característica medida**
- Medidas de razão: massa corporal, idade, tempo, pressão arterial ou temperatura Kelvin.
  - Razões entre 2 medidas fazem sentido. **o zero corresponde à ausência da característica medida.**

Significância estatística – testes usados (2 conjuntos)

- Chi-quadrado:

  medidas categóricas e algumas vezes com ordinais (número de observações é no mínimo 5)

- Fischer Exact test:

  versão mais elaborada que chi-quadrado. Usado quando o tamanho das amostras são pequenas

- Teste t-student:

  medidas intervalares e de razão, distribuídas de forma normal

- Teste T pareado:

  usado para conjuntos correspondentes (notas da P1 e notas da P2); mesmas condições que teste t-student

- Teste Wilcoxon rank-sum test (ou teste de Wilcoxon-Mann-Whitney)

  quando as condições do test t-student não são verdadeiras (não-normalidade ou variâncias muito diferentes). Usado para medidas ordinais, intervalares e de razão.

Avaliações em PLN

- Melhorar o estado da arte de tecnologias de língua
  - Avaliação da melhoria de um sistema versus outro é a variável de interesse.

- Métricas para comparar sistemas não são geralmente distribuídas normalmente: não se usa teste t-student e sim Wilcoxon.

- What's in a p-value in NLP?
  - Proceedings of the Eighteenth Conference on Computational Language Learning, pages 1–10, Baltimore, Maryland USA, June 26-27 2014.
  - http://www.aclweb.org/anthology/W14-1601

Significância Estatística

- Testes Estatísticos, hipótese nula, p-value, significância do teste, pressuposições do teste, variáveis dependentes e independentes...
- Variáveis:

Independentes

Todas aquelas que são manipuladas ou controladas

Dependentes

São aquelas que queremos estudar para ver os efeitos das mudanças nas variáveis independentes

Normalmente temos apenas uma variável dependente

Variáveis dependentes e independentes

Variáveis independentes:

São aquelas que *podemos controlar e mudar*

Escolher as variáveis não é fácil e, normalmente, exige *conhecimento do domínio*

Possuem um certo *efeito sobre as variáveis dependentes*

Variáveis dependentes:

Mede o *efeito dos tratamentos (Saúde)*

Normalmente, é *definida somente 1 variável dependente* derivada diretamente das hipóteses

Na maioria das vezes *não é diretamente mensurável*

Variáveis dependentes e independentes

Exemplo de Variáveis em CC:

estudar os efeitos de um novo método de desenvolvimento de software com relação à produtividade dos desenvolvedores.

Considerando que um método OO será introduzido no lugar de um método baseado em funções (procedimental)

Variável dependente:

produtividade

Variáveis independentes:

Método de desenvolvimento

Experiência do pessoal

Suporte de ferramentas

Ambiente de trabalho

Hipóteses de Pesquisa

Uma hipótese deve ser declarada formalmente.

e os dados coletados durante a execução experimental deverão ser usados para, se possível, rejeitar a hipótese.

Se a hipótese pode ser rejeitada/aceita então

conclusões podem ser feitas, com base no **teste de hipótese** levando em consideração alguns riscos

Hipótese nula vs Hipóteses alternativas

A definição de um experimento é formalizada por meio de hipóteses.

Duas hipóteses devem ser formuladas.

Hipótese nula (H0) – declara que não existem condições de tendência ou padrões em um experimento.

Hipótese nula vs Hipóteses alternativas

É a hipótese que queremos REJEITAR com a maior significância (certeza) possível.

Exemplo: “Uma técnica nova de inspeção encontra, na média (μ), o mesmo número de falhas (#F) que a técnica antiga”

$$H_0: \mu_{\#F\_antiga} = \mu_{\#F\_nova}$$

Hipótese nula vs Hipóteses alternativas

Hipótese Alternativa (H1) – é declarada a favor do que rejeita a hipótese nula.

Exemplo: “Uma técnica nova de inspeção encontra, na média (μ), mais falhas (#F) que a técnica antiga”.

$$H_1: \mu_{\#F\_antiga} < \mu_{\#F\_nova}$$

Hipótese nula vs Hipóteses alternativas

A hipótese nula H0 representa *a circunstância que está sendo testada, e o objetivo dos testes de hipóteses é sempre tentar rejeitar a hipótese nula.*

A hipótese alternativa H1 representa o que se deseja provar ou estabelecer, sendo formulada para contradizer a hipótese nula.

Erros: type I and type II

Existem vários testes estatísticos de hipótese.

Todos estão baseados na ideia de que as hipóteses são formuladas antes dos testes estatísticos serem escolhidos e realizados.

O teste de hipóteses envolve diferentes tipos de riscos:

Ou o teste rejeita uma hipótese verdadeira

Ou o teste não rejeita uma hipótese falsa

Erros: type I and type II

Repare que, ao testarmos uma hipótese nula, chegamos a uma conclusão:

rejeitá-la, ou não rejeitá-la

Entretanto, devemos lembrar que tais conclusões ora são corretas, ora são incorretas **(mesmo quando fazemos tudo corretamente!).**

Este é o preço a ser pago por estarmos trabalhando em uma situação onde a variabilidade é inerente !!!

Erros: type I and type II

Type-I-error

Ocorre quando um teste estatístico **indica** um padrão/relacionamento mesmo que não exista um padrão/relacionamento real

A probabilidade de cometer um erro desse tipo pode ser expressa como:

$$P(\text{type-I-error}) = P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira})$$

No exemplo de hipóteses apresentado, type-I-error é a probabilidade de rejeitar H0 **mesmo que as 2 técnicas**, na média (μ), **encontrem o mesmo número de falhas** (#F)

Erros: type I and type II

Type-II-error

Ocorre quando um teste estatístico **não indica** um padrão mesmo se tal padrão/relacionamento existir

A probabilidade de cometer um erro desse tipo pode ser expressa como:

$$P(\text{type-II-error}) = P(\text{não rejeitar } H_0 \mid H_0 \text{ é falsa})$$

No exemplo de hipóteses apresentado, type-II-error é a probabilidade de não rejeitar H0 **mesmo que as 2 técnicas, na média, possuam médias** (μ) **do número de falhas** (#F) **encontradas diferentes**

Erros: type I and type II

		O Verdadeiro Estado da Natureza	
		A hipótese nula é verdadeira.	A hipótese nula é falsa.
Decisão	Decidimos rejeitar a hipótese nula.	**Erro tipo I** (rejeição de uma hipótese nula verdadeira)	Decisão correta
	Não rejeitamos a hipótese nula.	Decisão correta	**Erro tipo II** (Não rejeição de uma hipótese nula falsa)

Controle de riscos: type I and type II

O tamanho do erro depende de diferentes fatores

Um exemplo é a habilidade do teste estatístico revelar um padrão/relacionamento verdadeiro em dados coletados

Conhecido como o Poder do Teste (*P)*

Controle de riscos: type I and type II

O poder de um teste estatístico é a probabilidade do teste revelar um padrão verdadeiro se H0 for falsa.

Para tanto, ao realizar um experimento devemos escolher um **teste com o maior *P possível***.

$$P = (\text{rejeitar } H_0 \mid H_0 \text{ é falsa}) = 1 - P(\text{type-II-error})$$

P-value

- *Se as condições do teste são verdadeiras e o p-value é baixo então o pesquisador pode assumir que a hipótese nula é falsa (há evidências para rejeitar a hipótese nula).*
- *O valor do p-value abaixo do qual se assume que a hipótese nula é falsa é 0.05 ou 0.01*

- *A significância do teste é 1 – p-value, ou seja, deve ser 95% ou 99%*
- *Se o p-value calculado é maior que o valor de corte então: não há evidências para rejeitar a hipótese nula.*
- ➔ *Mostrem sempre o p-value calculado. Para PLN usa-se o valor de corte de 0.0025*

Seleção dos participantes

A seleção dos participantes está diretamente relacionada à generalização dos resultados de um experimento

Para tanto, a seleção deve ser representativa para a população
Seleção de participantes = amostra de uma população
A amostragem pode ser probabilística ou não-probabilística

Amostragem probabilística: a *probabilidade da seleção de cada participante é conhecida*

Amostragem não-probabilística: a *probabilidade da seleção de cada participante não é conhecida.*

Servem para sondagens sem propósitos inferenciais, nestes casos, os processos que envolvem comparações estatísticas que usem cálculos científicos não são válidos.

Amostragens probabilísticas

- **Amostragem aleatória simples**
  - É aquela em que toda amostra possível de mesmo tamanho tem a mesma chance de ser selecionada a partir da população.
- **Amostragem sistemática**
  - Consiste em um elemento aleatório, por exemplo, **um nome a cada dez de uma lista**, a décima peça produzida em uma linha de produção etc. Sua principal vantagem é sua simplicidade e flexibilidade, sendo mais fácil de instruir os trabalhadores de campo.
- **Amostragem estratificada**
  - Consiste em dividir ou estratificar a população em um certo número de subpopulações que não se sobrepõem e então extrair uma amostra de cada estrato.

Amostragens não probabilísticas

- **Amostragem de voluntários**
  - É quando os próprios componentes da população se voluntariam para participar de uma pesquisa.
- **Amostragem por bola de neve**
  - escolhem-se voluntários e estes indicam "conhecidos" com o mesmo perfil para responder entrevistas ou questionário e assim sucessivamente. Formam-se redes de referência.
- **Amostragem por cotas**
  - Consiste em buscar repetir a proporção de elementos de cada estrato da população, na amostragem por cotas os elementos da amostra não são selecionados através de sorteio.
- **Amostragem por escolha racional**
  - É quando o pesquisador busca na população uma parte dela que interessa, ou seja, os participantes são escolhidos por terem uma ou mais características específicas.

Seleção dos participantes

O **tamanho da amostra tem impacto sobre** a generalização dos resultados de um experimento

Quanto maior a amostra, menor é a chance de errar ao generalizar os resultados, pois tendemos ao universo.

Princípios gerais para escolher o tamanho da amostra:

Se existir uma ampla variabilidade na população, uma amostra de tamanho maior é necessária;

A análise dos dados pode influenciar a escolha do tamanho da amostra.

Limitações da Pesquisa Experimental

As pesquisas experimentais constituem o mais valioso procedimento disponível aos cientistas para testar hipóteses que estabelecem relações de causa e efeito entre as variáveis.

Em virtude de suas possibilidades de controle, os experimentos oferecem garantia muito maior do que qualquer outro delineamento de que a variável independente causa efeitos na variável dependente.

A despeito, porém, de suas vantagens, a pesquisa experimental apresenta várias limitações.

Primeiramente, existem muitas variáveis, cuja manipulação experimental se torna difícil ou mesmo impossível.

Limitações da Pesquisa Experimental

Uma série de características humanas, tais como idade, sexo ou histórico familiar, não podem ser conferidas às pessoas de forma aleatória.

Outra limitação consiste no fato de que muitas variáveis que poderiam ser tecnicamente manipuladas estão sujeitas as considerações de ordem ética que proíbem sua manipulação.

Não se pode, por exemplo, submeter pessoas a atividades estressantes com vistas a verificar alterações em sua saúde física ou mental.

Agradecimentos

- Parte dos slides vieram da apresentação de

- Pesquisa Experimental
  - Nemesio Freitas Duarte Filho
  - Kleberson Junio do Amaral Serique
  - Prof. Dra. Renata Pontin