**Resumos MEI**

**Por que fazer experimentos?**

- Coletar evidências

- Validar hipótese

- Apoiar a definição, validação,

- Validar modelos

- Teorias de confirmação ...

**Propriedades:**

- Relevância (objetivos e resultados esperados são importantes para progresso (têm impacto?)

- Representatividade (é realista e representativo de cenários reais?)

- Repetibilidade (é possível repetir e obter resultados semelhantes?)

- Reprodutibilidade (informações suficientes que permita outros reproduzir experiência?)

- Análise e generalização de resultados

(análise dos resultados sólida? Generalização das conclusões é credível?)

- Custo (custo dos experimentos é compatível com os benefícios?)

**Tipos de experimentos:**

- Controlled experiments

- Case studies

- Pilot studies

- Benchmarks

[Field studies, Simulations, Surveys, Artifact/archive analysis, Rational reconstructions, Ethnographies, Quasi-experiments]

**Design Experiências** (Experimentos controlados /de laboratório):

**1.** Declaração do problema

**2.** Identifique variáveis

**3**. Gere hipóteses

**4.** Definir configuração/cenário

**5.** Desenvolver ferramentas/procedimentos

**6.** Executar experiência e coletar dados/medições

**7.** Realize análise de dados e teste hipóteses

**8.** Tirar conclusões (muitas vezes voltar início e reformule a definição do problema ou teste uma hipótese diferente)

De 1:5 é design of the experiment, restantes é medidas, análise e conclusões

**1-** Declaração do problema deve focada para permitir a identificação das variáveis do problema e tb devem ser abertas para permitir diferentes hipóteses para responder ao problema. Ex: Como x afeta y nas condições z?Para fazer boas declarações de problemas: conhecer a área (processo, sistema, técnica, produto), ser preciso e claro, ter certeza de que o problema é relevante

**2-** Variável dependente (variável de resposta): saída medida (ex: tempo de resposta, taxa de transferência, nº de bugs...). Variáveis independentes (fatores): input que pode ser mudado na experiencia (tamanho da memória, taxa de clock, tamanho do arquivo...)

Níveis:contínuos (tempo, tamanho bytes...) ou discreto (tipo de sistema, tipo algoritmo...)

Mudar fator de cada vez- cenário simples, a analise é simples, é fácil entender o efeito de um determinado fator na variável independente.

Fator completo-Alterar 2 ou + simultaneamente, + complexo. 2 vantagens: + eficazes (poupa tempo e esforço), permite estudos de interações entre fatores.

**Terminologia:**

- Baseline (golden run): conjunto valores de fatores (variáveis independentes) que representam cenário de linha de base.

-Repetition of golden run: estimar erro experimental (ruído) no sistema e identificar pequenos efeitos que possam variar no resultado.

- Randomization: minimizar potenciais biases incontroláveis, aleatoriamente atribuir fatores para “calcular” efeitos de possíveis terminologias estranhas.

- Bloqueio: experimento é dividido em segmentos homogéneos (sets de máquinas, utilizadores...) para melhorar precisão. Objetivo é controlar as variáveis de bloco para bloco.

- Confounding variable: extrair variáveis que influenciam as relações entre variáveis dependente e independentes.

**3-** Hipótese descreve relações provisórias entre fatores (v. independentes) e a variável resposta (dependente). Resposta provisória à declaração do problema. Pode ser direcional ou não direcional

Pode levar a um modelo que permite prever o que acontecerá em casos futuros. Muitas vezes objetivo dos experimentos é quantificar o relacionamento (ñ apenas confirmar que existe).

**4&5-** Complexidade do experimento, custo, ferramentas e estruturas disponíveis que podem ajudar, grau de automação.

**6-** Medições contínuas/discretas. Precisão e resolução. Medições básicas em computadores: contar, duração, tamanho, qualquer valor derivado da combinação de medidas básicas.

**7-** Análise exploratória de dados. Análise dados estatísticos: tabelas, gráficos, média, desvio padrão, lidando com erros de medição, intervalos de confiança, comparação estatística de alternativas, testes para verificar se os dados medidos se ajustam às distribuições conhecidas (qui-quadrado, testes KS...)

**8-** Relatório escrito da experiência é muitas vezes único resultado de meses/anos de trabalho.

Qualidade escrita essencial. Atributos relevantes:

- Ser claro (objetivos, abordagem, análise, conclusões)

- Credível (dados relatados, conclusão...)

- Autônomo

Nº erros encontrados testes depende do nº médio de horas de sono?

a) Dependente- nº de bugs

Independente- Horas de sono (media 2,4,6,8,10),

complexidade (alta, media, baixa), experiência programador, linguagem, duração tarefa.

c) Bugs relacionadas com o sono? (hipótese testada)

Avaliar complexidade: métricas de complexidade de software(sw), complexidade ciclomática(VG) – indentação.

**Resolução-** Resolução do instrumento de medição: menor diferença entre as medições fornecida por um dispositivo de medição. Ex: medir tempo de execução de um programa (milissegundos)

**Incerteza-** se repetirmos medição, resultados ligeiramente diferentes. Reflete a falta de precisão. 2 tipos:

**A diagram of a graph

Description automatically generatedA diagram with blue dots and lines

Description automatically generated**I. aleatórias- Ocorrem sem um padrão previsível.Podem ser reduzidos, mas nunca eliminados. Devem ser analisados estatisticamente e reportados no processo de medição.

I. sistemáticas-Desvios sistemáticos, - ou + que valor exato**.** Causas possíveis: ferramenta de medição imprecisas,calibração incorreta, tempo reação da ferramenta...Quando identificado, pode ser eliminado (uma das etapas do experimento).

**I.S.: casos especiais**

Warm-up: primeira medição pode ser diferente das subsequentes.

Ramp-up: é necessário um conjunto de medições para atingir valores estáveis

Hysteresis: resultado depende medições anteriores (histórico)

**Variabilidade -** 2 fontes diferentes:

- Limitações de precisão do instrumento medição: Mesmo que condições fossem totalmente estáveis, diferentes mediçõesmostrariam valores um pouco diferentes.

A diagram of a graph

Description automatically generated- Mudanças nas condições das medições: (ambiente, técnicas de manuseio...) Ex. alterações na carga do computador, estado do cache causam diferentes medições. Pequenas mudanças no ambiente são analisadas como incertezas aleatórias. Casos extremos levam a outliers (devem ser ignorados).

Uncertainty- deve ser analisado estatisticamente e relatado no processo de medição.

Inaccuracy- identificado e eliminado quando possível. Relatar erros sistemáticos vinculados.

Outlier- devem ser identificados e removidos da análise.

Resolution- certificar de que resolução é adequado aos objetivos do experimento e à análise necessária.

**Medir tempo em computador (ex.)**

Interrupções do temporizador- causa interrupções periódicas da CPU e corre o manipulador de interrupção do relógio que mantém a hora do sistema (legível por humanos). Razoavelmente preciso, resolução máxima é microssegundos.

Contador de carimbo (data/hora)- registro especial que conta os ciclos desde que a máquina foi inicializada. Depende da taxa CPU, pode mudar (economizar energia...). Pode alterar, dependendo da temperatura. Resolução- nanossegundos (precisa de muitos ciclos de processador para fazer uma leitura)

Servidor de horário e NTP (protocolo NetworkTime): Obtém hora de uma fonte padrão, para sincronizar relógio em rede. Pode levar a saltar no tempo, para frente/trás. Outros (específico do sistema...)

**Linux gettimeofday() (ex.)**

- Noção atualizada de tempo real de acordo com fonte externa

- Combina diferentes fontes de tempo

- Sincronizar contador de carimbo de data/hora

- Resultado do relatório em resolução de microssegundos

- Quando chamado, lê o contador data/hora atual e extrapola a partir da interrupção do relógio anterior

Medição simples- metas:

- Medição de atividade/operação do computador. Ex. classificação de um nº de itens

- Feito a partir do nível do usuário

- Sem ferramentas especializadas/externas

Alternativa 1:

t1 = gettimeofday();

<operação sendo medida>

t2 = gettimeofday();

print “tempo execução foi", t2 – t1, "\n";

Problemas:

- Imprecisão devido à sobrecarga de medição.

- Erro é altamente relevante se tempo de execução da “operação que está sendo medida” for de faixa semelhante ao tempo de execução de gettimeofday()

Alt 2: medições múltiplas + buffer:

for (i=0; i<N; i++) {

t1 = gettimeofday();

<operação sendo medida> t2 = gettimeofday();

tempo[i] = t2 - t1;}

print “t. médio é", avg(time[0.. N-1]), "\n”;

Prós e contras:

- Média é boa, evita a sobrecarga de impressão, normalmente é pesada

- Pode haver problemas de resolução se o tempo de execução for de faixa semelhante ao tempo de execução de gettimeofday(); +3

**Intervalos de confiança (básicos)**

- Quando realizamos varias medições da mesma coisa, podemos calcular intervalos de confiança

- Supor que medidas são amostras de uma distribuição (normal) (valor real + erro aleatório)

- Caracterizar a dispersão da distribuição

A graph of a graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidenceA graph with blue dots and red arrow pointing to the average

Description automatically generated- Encontre o intervalo que inclui as massas desejadas da densidade de probabilidade (ex. 90%)

- Supor que um conjunto de medições de uma distribuição normal (valor real + erro aleatório)

- Este conjunto é uma média, estimativa do valor real

- Se repetirmos isso com amostras diferentes, obteremos uma média ligeiramente diferente

- Múltiplos conjuntos de amostras induzem múltiplas amostras a partir da distribuição de médias

- Distribuição das médias é mais estreita que a distribuição básica, fornece uma estimativa mais precisa do valor real

A diagram of a normal distribution

Description automatically generatedA diagram of value and distribution of value

Description automatically generatedSuposição: médias refletem valor verdadeiro + algum erro/ruído aleatório (médias são distribuídas em torno do valor verdadeiro), dada a distribuição, podemos encontrar intervalo *h* que se espera que contenha ex.90% das médias

- Para 90% das médias, o valor verdadeiro está dentro de *h* ou média do intervalo ± *h* tem probabilidade 0,9 para incluir o valor real

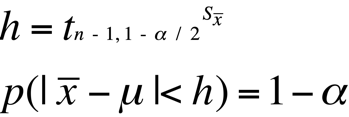
**Calcular intervalos de confiança:**

- *μ* é o meio real da distribuição baseada

A black letter x with a line

Description automatically generated- *x* é a média de *n* amostras

- distribuição for normal, então as médias terão distribuição t ou *Z* se amostra for grande (n ≥30)

- α é a incerteza aceitável (implica que o nível de confiança é 1 – α) e defina a meia largura

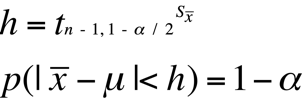
- ***t****n* - 1,1- *α* / 2, vem das tabelas (t para n<30)

*n* é nº de amostras; n-1 graus de liberdade (*t* Student) (n>30 utilizar z, distribuição normal)

A black letter x with a line

Description automatically generated- S é o desvio padrão das médias. Assumindo

as amostras de base são independentes, é calculado assim , *s* é o desvio-padrão das amostras.

Intervalo de confiança:

A black letter x with a line

Description automatically generatedCom uma certeza de 1-a, a distância entre uma amostra média x e a verdadeira média *μ* é inferior à metade

A black letter x with a line

Description automatically generatedSe repetirmos isso muitas vezes, e cada vez que desenharmos um segmento de ±*h* em torno de x, logo em 1-α dos casos este o segmento incluirá *μ.*

A black letter x with a line

Description automatically generatedA black letter x with a line

Description automatically generatedA diagram of a function

Description automatically generatedA diagram of a function

Description automatically generatedCom uma certeza de 1-α a distância entre uma amostra média e o verdadeiro médio *μ* é inferior a *h*. Se repetir medição muitas vezes, e cada vez desenhar segmento de ±*h* em torno de x, então em 1-α dos casos o segmento incluir *μ*

Assumindo que amostras são independentes, fórmula é:

A black letter x with a line

Description automatically generated ± ± *t* \*s/√*n* – (n≤30, tabela *t* com df = n-1)

A black letter x with a line

Description automatically generated± ± *Z* \*s/√*n* - (n≥30, tabela *Z* para o normal padrão distribuição)

Onde:

*s* é desvio padrão das amostras *n*

Ex. α = 0,1 o valor z = 1.645. Representa o ponto no eixo em que a área sob a curva normal padrão é 1- α (ex. 90% para α = 0,1)

Premissas:

- Amostras básicas vêm de uma distribuição normal. [Caso contrário, mas tiver uma variância finita, as médias ainda serão normais, mas isso exigirá um *n* maior.].

- As amostras base são independentes [Caso contrário, talvez o uso de lotes maiores reduza a correlação entre eles].

- Se o número de amostras for pequeno (n ≤ 30) assumimos uma distribuição *t* Student.

Antes de calcular intervalos de confiança:

- Limpar os dados primeiro

- Remover outliers que indiquem interferência ou espúrios medições. (medições superior/inferior; analisar os dados e decidir sobre os outliers a eliminar)

- Remover os efeitos de aquecimento e histórico

**Ex: qual é o grau de confiança Z para α = 5%?** (bilateral)

1. Subtrair α de 1: 1 – 0,05 = 0,95

2. Dividir resultado 2 (bilateral) \_0,95/2 = 0,475

3. Observar tab. *Z* e localizar resultado (0,475).

A table of numbers with text

Description automatically generatedA screenshot of a data

Description automatically generatedValor + próximo do coeficiente *Z* está na interseção da linha 1,9 coluna 0,06. Somando 2 valores é *Z* = 1,96 para α = 5% Confianças Comuns

**A screenshot of a paper with numbers and text

Description automatically generated**A table of numbers with numbers

Description automatically generatedEx. medir a execução tempo de um programa. Repetir execução do programa c/ diferentes cargas e momentos diferentes, no mesmo computador.

- Tempo execução (95%) = 3130,3196,11

- Tempo execução ( 99%) = 3130,31150. 45

Normalmente testar 2 resultados possíveis [variável dependente é binária (2 resultados mutuamente exclusivos). Supor que uma distribuição binomial seja uma boa aproximação para esses casos]

Ex.: Erro detectado ou ñ, Vulnerabilidade detectada ou ñ

**Modelo binomial**

Propriedades variável binomial:

- É binária (assume apenas 1 de 2 valore possíveis.

- É observada um nº conhecido de vezes *(n):*

[Cada observação é chamada de tentativa.

Nº de vezes que o resultado de interesse (ex. detecção de erro) é observado é *x*, chamado de nº de “sucessos”.]

- Probabilidade de ocorrência do resultado de interesse é a mesma para cada tentativa.

- Ensaios são independentes e o resultado de um ensaio ñ afeta o resultado de outro ensaio.

A black and white math equation

Description automatically generated with medium confidenceDistribuição por proporção da amostra:

Proporção da amostra:

A black and white math equation

Description automatically generated with medium confidence (amostra = Conjunto de ensaios)

A black and white math equation

Description automatically generated with medium confidenceP é proporção da amostra c/ resultado de interesse. É estimativa proporção da população *p.*

P varia de amostra para amostra de forma aleatória.

Grandes *n* de amostras, distribuição pode ser considerada como normal. Mas o grande nº de amostras deve:

- nº de sucessos e ñ sucessos ser maior que 10 (*np* ≥ 10 e *n(1-p)* ≥ 10)

- ser pelo menos 20 vezes menor do que a população (população deve ser muito maior)

Assim, assumimos que média da distribuição da amostra é aproximadamente igual à proporção real da população *p*.

Intervalos de confiança para a proporção da população

Para amostras maiores, a amostra da distribuição

da amostra da proporção é aproximadamente normal:

- Erro standard (SE) da amostra da proporção é dado por:

- Intervalo de confiança para proporção da população é

A square root of a mathematical equation

Description automatically generatedA square root of a mathematical equation

Description automatically generatedA square root of a mathematical equation

Description automatically generated

A diagram of a number

Description automatically generated

**Cenário de teste de hipótese 1** - Supor administrador de banco de dados de um grande sistema de informação, está insatisfeito c/ tempo execução de um pacote SQL.

A partir de dados históricos (milhares de execuções de pacotes anteriores), sabe que tempo médio de execução do pacote é de 83,54 seg. c/ desvio padrão de 16,36.

Alterar o ajuste do banco de dados e executar o pacote várias vezes para verificar o efeito.

Questões: Nova configuração teve algum efeito? Nova é melhor ou pior?

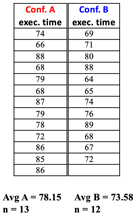
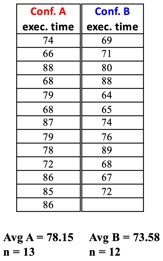
Resp: Usamos a distribuição de dados normal porque:

- Cada execução é independente das anteriores;

- Variabilidade das medições resultam de alterações aleatórias na condição de execução.

Se ñ houver certeza de que dados seguem distribuição normal, devemos testá-la para normalidade.

**Cenário de teste de hipótese 2** - Supor administrador de banco de dados de um grande sistema de informação. Banco de dados acabou de ser instalado e está a tentar 2 configurações de afinação: Conf. A e Conf. B.



Usa um pacote SQL para testar o tempo de execução para cada configuração. Depois de executar várias vezes o pacote SQL em ambos configurações, tomar decisão.

Pergunta: Qual é a melhor?

**Hipótese** – é a tentativa de resposta

Qual é uma hipótese?

– Proposta de explicação para um determinado fenômeno

– Suposição sobre a eficiência de um componente/sistema – Declaração sobre parâmetros de uma população (visão . estatística)

Escopo (finalidade):

– Abstrato: sobre o mundo (*lato senso*)

– Concreto: sobre determinado projeto/aparelho

Tipos:

– Explicativo: explica o fenômeno, identifica relações/ causalidades entre variáveis/elementos do fenômeno

– Preditivo: prevê a observação de um fenômeno, antecipa o resultado

Hipótese tem de ser avaliada para ser considerada verdadeira. Pode ser rejeitada ou confirmada.

Hipótese verdadeira significa que a probabilidade de ser correta é 'alta' e de ser incorreta é 'baixa'.

Estatísticas são necessárias para quantificar o significado de "alto" e "baixo" e decidir sobre validade da hipótese.

Hip. são rejeitadas ou aceites c/ algum grau de certeza

**Hipótese: colocada em perspetiva**

- Tema: Assunto (área focada) de interesse, onde a lacuna ou dificuldade de ser resolvido está incluído. Essencial para contextualizar hipótese. Ex. qualidade código (s/ bugs)

- Problema: Objeto do estudo. Pressupõe questões claras e explícitas que formulam o problema a ser resolvido. Ex. metode de desenvolvimento relacionado com nº bugs?

- Hipótese: resposta provisória as perguntas. Se hipótese for confirmada, a resposta é considerada correta (com determinado grau de certeza). Ex. H0- soft. desenvolvido por CMMi5 tem mesmo bugs do desenvolvido por Scrum.

H1- software desenvolvido por CMMi5 não tem mesmos bugs do que desenvolvido por Scrum.

**Estatística inferencial e teste de hipóteses**

- Permite avaliar comportamento em amostras para saber + sobre comportamento de toda a população

- Muitas vezes, toda a população é muito grande (ou mesmo infinita) ou não é acessível

- A partir do teorema do limite central, sabemos que a probabilidade de selecionar qualquer outro valor médio da amostra desta população é normalmente distribuída.

Teste de hipótese- forma sistemática de testar alegações ou ideias sobre um grupo/população,

A diagram of distribution of a normal distribution

Description automatically generatedcom base em amostras selecionadas das tais populações.

**Etapas de teste de hipóteses**

1. Declarar hipótese/afirmação a ser testada

2. Selecionar critérios para uma decisão

3. Calcular a estatística de teste

4. Tomar decisão

**Cenário de teste de hipótese 1\***

**Step 1-** State the hypothesis

- Hipótese nula (H0) é uma afirmação sobre o parâmetro populacional (ex. média populacional) que é considerado verdadeiro, (resposta provisória ao problema em estudo) Ex.: H0- Nova configuração ñ afeta tempo de execução.

- Hipótese alternativa (H1) é afirmação que contradiz diretamente a nula ao afirmar que o valor real da população ñ é igual ao valor da nula, (isso é o que achamos que está errado na hipótese nula).

Ex.: H1- Tempo execução é diferente na nova configuração (menor ou maior).

Decisão tomada no teste centra-se na hipótese H0 (nula)

- Ideia é mostrar evidências de que H0 é improvável, a fim de a rejeitar. Se ñ o fizer, hipótese nula é mantida.

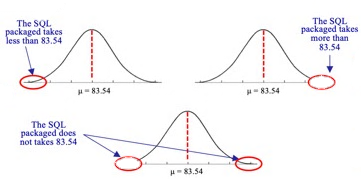
- O contrário é ñ fazer nada, encargo é colocado sobre o pesquisador para demonstrar que H0 ñ é provável que seja verdade. [Experiências definidas para coletar dados para mostrar que H0, ñ é verdade]

**Step 2–** Select the criteria for a decision

- Definir critério significa indicar o nível de significância.

- Nível significância refere-se ao critério de julgamento sobre qual decisão é feita em relação ao valor declarado na hipótese nula.

- Nível de significância típico é 5%. Significa que quando a probabilidade de obter uma determinada amostra média é inferior 5%, supondo que hipótese nula seja verdadeira, concluímos que a amostra utilizada para calcular a média é muito improvável e rejeitamos a hipótese nula. [Escolher onde colocar grau de confiança - 1 Tail, 2 Tail]

Média amostra = à média da população, se hipótese nula for verdadeira - valores possíveis de média da amostra é normalmente distribuída. Pelo menos 95% da média amostra serão abrangidas 2 SD da média da população (nível confiança 5%)

**Step 3 –** Cumpute the test statistics

- Selecionar amostra aleatória da população e medir média amostra. Ex: executar *n* vezes, medir média =78,15

- Para tomar decisão, preciso avaliar quão provável é esse resultado da amostra, se média populacional declarada pela hipótese nula (83,54) for verdadeira.

- Estatística de teste é fórmula para determinar probabilidade de obter resultados amostrais se hipótese nula for verdadeira. Valor da estatística de teste é usado para tomar uma decisão sobre a hipótese nula.

A diagram of a sample

Description automatically generatedMedir distância entre a média da amostra e a média da população em H0. Quanto maior o valor de |Zc|, mais indicará que H0 ñ é verdadeiro.

**Step 4** – Make a Decision

- Valor do teste estatística (Zc) é chave para tomar uma decisão sobre a hipótese nula. Decisão é baseada na probabilidade de obtenção de uma amostra média, dado que o valor declarado na hipótese nula é verdadeiro.

- *P* é a probabilidade de obter um resultado amostra, dado que o valor declarado na hipótese nula é verdadeiro.

Ex: *P* < 5% rejeita hipótese nula (alcançar significância),

*P* > 5% manterá hipótese nula (ñ alcançará significância)

**Cenário 1\* - Non-directional (Two Tail)**

**Step 1-** State the hypothesis

**H0- nova configuração ñ afeta sobre tempo de execução.**

**H1 - tempo execução é diferente na nova configuração (menor ou maior)**

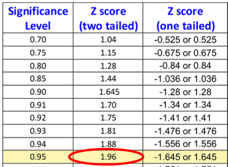
**- Testar se hipótese nula H0 é verdadeira.**

**Step 2–** Select the criteria for a decision

Considerar nível de significância 5% [a=0,05 1-a = 0,95]

Localizar pontuação Z (tabela normal padrão) que representa os valores críticos

A diagram of a graph

Description automatically generatedValor crítico é um valor de corte que define os limites além dos quais menos de 5% das médias da amostra podem ser obtidas se a hipótese nula for verdadeira.

**Step 3 –** Cumpute the test statistics

A black numbers and a white background

Description automatically generated- Test statistic:

**Step 4** – Make a Decision

Probabilidade obtenção de *Z* = -1,86 é dado pelo valor *P*. Para obter P procurar 1,86 na tabela. Valor de *p* para *Zc* = -1,86 é 0,0314

Valor de *p* para *Z* = -1,86 é 0,0314. Como é bilateral P = 0,0314 x 2 = 0,0628 - > P = 6,28%

A diagram of a function

Description automatically generatedSignifica que probabilidade de obter uma média de 78,15 se H0 for verdadeiro é 6,28%

A screenshot of a math test

Description automatically generated

**Abordagem + pragmática:**

1. Declarar hipótese/afirmação a ser testada

2. Calcular a teste estatístico

3. Obter valor de *p*

4. Tomar decisão

**Cenário 1\* Pragmatic**

**Step 1-** State the hypothesis

**H0- nova configuração ñ afeta sobre tempo de execução.**

(tempo médio execução é 83,54)

H1 - tempo de execução do é menor na nova configuração

**- Testar se hipótese nula H0 é verdadeira.**

Apenas hipótese alternativa mudou. Ensaios direcionais ou unilaterais são ensaios de hipóteses em que a hipótese alternativa é declarada como maior (>) ou menor (<) que o valor indicado na hipótese nula.

**Step 2 –** Cumpute the test statistics

A black numbers and a white background

Description automatically generated- Test statistic:

A diagram with a yellow text

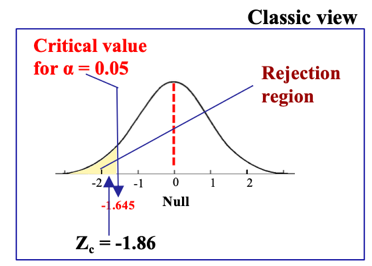
Description automatically generated with medium confidence**Step 3 –** Obtain the p value

**Step 4** – Make a Decision

Probabilidade obtenção de *Zc* = -1,86 é dada pelo valor *p*.

*p* = 0,0314 significa que a probabilidade de obter uma média de 78,15, se H0 for verdadeiro, é 3,14%

Então, posso rejeitar H0 c/ pelo menos 95% de confiança

Tempo de execução do pacote SQL é menor na nova configuração

**Tipos de erro**

A diagram of a choice

Description automatically generated with medium confidenceConclusão pode estar errada, pois estamos a analisar uma amostra c/ nº limitado de elementos *n.* [Falso positivo- homem gravido, Falso negativo– gravida ñ gravida]

E.II - Decisão errada consiste em manter uma falsa hipótese nula. Isto é, não fazer nada. Podemos fazer + experiências e testar novamente a hipótese nula. Não é assim mau…

E.I - Decisão errada é rejeitar um verdadeiro

hipótese nula. Significa rejeitar noções anteriores de verdade que são de facto verdadeiro (equivalente julgar uma pessoa inocente culpada). Podemos controlar diretamente a probabilidade de um erro do Tipo I declarando um nível alfa.

Correct Power- Conclusão forte: esta é a decisão que procuramos quando testamos a hipótese. Se o testarmos, significa temos dúvidas sobre essa hipótese.

**Cenário 1\* + questões**

E se ñ houver um nº relativamente grande de amostras? E se ñ soubermos desvio-padrão da população?

Usar *t Test*

**T-test**

- Teste t segue a distribuição T Student (se hipótese nula for verdadeiro)

- T-teste deve ser usado quando: tamanho da amostra é pequeno (n < 30), desconhece-se o desvio-padrão das populações

- Quando o nº de amostras é grande, o t teste e Z dão resultados semelhantes

2 tipos:

- 1 amostra: usado para comparar média da amostra com média da população conhecida.

- 2 amostras: utilizados para comparar duas amostras.

(Amostras independentes: grupos separados ñ relacionados)

**Teste de hipóteses usando T-Test (uma amostra)**

- Segue os mesmos passos que ensaio Z

- Teste estatístico é agora o teste t (fórmula igual *p*)

- Valor crítico vem da tabela T (considerando n-1 graus de liberdade)

A mathematical equation with arrows pointing to the sides

Description automatically generated- Step 2: Cumpute the test statistics:

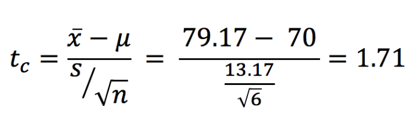
**Ex3. hipotese c/ T-test (1 tail)**

Professor quer saber se alunos são bons em C#. Ele quer que turma consiga pontuar acima de 70 (0-100), mas ñ quer examinar todos alunos. Seleciona aleatoriamente 6 alunos e dá-lhes um teste de C#. 6 alunos recebem notas de 62, 92, 75, 68, 83, e 95. Ele pode ter 90% de confiança de que pontuação média seria acima de 70?

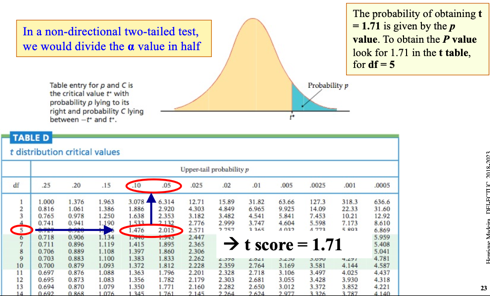
Step1-

H0: *μ* = 70 (classe sabe programar em C# c/ competência equivalente a 70)

H1: *μ*1 > 70 (classe é melhor em programação C# do que a pontuação de 70)

Step2-

Média da amostra: 79,17

Desvio-padrão amostra: 13,17

Step3-

Step4- Probabilidade de obtenção t = 1,71 é dada pelo valor de *p*. Para obter valor *p* procurar 1,71 na tabela t, para df = 5

Valor de *p* situa-se entre 5% e 10% (*p* = 7,4%)

- p < 10% : rejeita hipótese nula (alcance significado)

Significa que probabilidade de obter uma pontuação média de 79,17, se H0 for verdadeira, é de 7,4%

Conclusão: turma é melhor em programação C# do que a pontuação de 70

**Medir o tamanho de um efeito**

- Decisão de rejeitar a hipótese nula significa que o efeito é significativo. Teste hipóteses ñ informa sobre o tamanho do efeito.

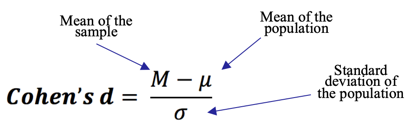
- Dimensão do efeito é uma medida estatística da dimensão de um efeito numa população. Faz sentido quando a hipótese nula é rejeitada.

- Cohen´s d mede nº de desvios-padrão, efeito mudado para acima ou abaixo da média da população declarada pela hipótese nula.

**Cohen´s d formula de medida**

- Convenções de tamanho de efeito de Cohen são usadas para interpretar o tamanho do efeito

- Se valores *d* forem negativos, efeito mudou para abaixo da média da população

****A table with text and numbers

Description automatically generated- Na prática, valor *p* tb dá ideia do tamanho do efeito.

A graph of a normal distribution

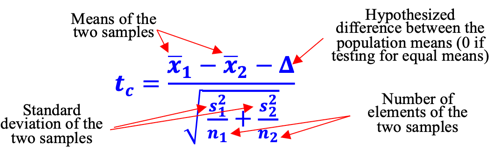
Description automatically generatedA yellow background with black numbers

Description automatically generated**Cenário 1\* Cohen´s d formula**

**Teste de hipóteses usando T-Test (uma amostra)**

- Segue mesmos passos que Z

- Valor crítico provém da tabela *t* (graus de liberdade são os menores n1-1 e n2-1)

- Teste estatísticos são agora 2 amostra

**Cenário 2\***

Importante: consideramos que as amostras de medição obtidas com cada configuração são independentes.

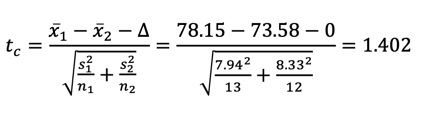
**μ1 =** 78.15 **s1 =** 7.94 **n =** 13 : **A**

**μ2 =** 73.58 **s2 =** 8.33 **n =** 12 : **B**

Step 1 –

- H0: **μ**1 = **μ**2 (configurações A e B são iguais em tempo de execução)

- H1: **μ**1 > **μ**2 (configuração B é + rápida que A (tempo de execução é maior na A)

Step 2 –

A table with numbers and words

Description automatically generated

A screenshot of a computer

Description automatically generatedStep 3 – Dado que as dimensões das amostras são *n* = 13 e n = 12, grau de liberdade é o menor *n*-1 => 11

Step 4 – Valor *p* para *t* = 1,402 e *df* = 11 está entre 5% e 10% (tabela T), valor *p* exato é 0,0942 (p=9,42%) (calculadora online). Significa que probabilidade de obter uma pontuação média de 73,58 se H0 for verdadeira é de 9,42%. Ñ foi possível provar que configuração B é mais rápida do que A c/ 95% de confiança.

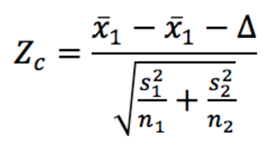
**Quando utilizar**

Step 2 - Compute the test statistic:

Se dimensão da amostra for grande (n>30) e se conhecer a população - Teste Z para comparar média da amostra c/ média população.

A mathematical equation with black text

Description automatically generated

Se dimensão amostras for grande (n≥30) – Teste Z 2 amostras para comparar médias de 2 amostras grandes independentes.

A mathematical equation with black text

Description automatically generated with medium confidenceSe tamanho amostra é pequeno (n < 30) e μ da população não é conhecida (é um alvo) - Teste T de 1 amostra para comparar uma média da amostra com a média da população.

A mathematical equation with black text

Description automatically generatedSe tamanho amostras for pequeno (n < 30) - Teste T de 2 amostras para comparar médias de duas amostras independentes.

**Step 3 – Obtain p value**

Teste estatístico é convertido em uma condição de probabilidade, o valor *p*. Pode ser obtido usando tabelas t ou usando cálculo do valor *p* emsites.

Valor *p* responde à pergunta "se hipótese nula é verdadeira, qual é probabilidade de observando os dados medidos?"

**Step 4 – Obtain *p* value**

Pequenos valores de *p* fornecem evidências contra a hipótese nula, pois significa que os dados observados são improváveis quando hipótese nula é verdade. Convenções:

p ≥ 0,10 - diferença observada é "ñ significativa"

0,05 ≤ p > 0,10 - "marginalmente significativa"

0,01 ≤ p > 0,05 - diferença é "significativa"

p < 0,01 - diferença é "altamente significativa"

**Inferências de proporção**

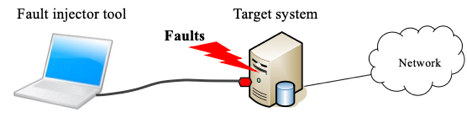
Em experiências de computador, a variável dependente tem apenas 2 resultados possíveis. Ex.: sistema travou ou ñ, caso de teste bem-sucedido ou teste falhado.

**Cenário tradicional de injeção de falhas (ex.)**

- Injetou 1000 falhas e 756 o sistema detetou erros

- Injetou 1000 falhas e em 89 falhas o sistema caiu

- Injetou 1000 falhas e em 56 falhas o sistema produziu saída errada s/ aviso (corrupção silenciosa dos dados).

Variável dependente é binária (2 resultados mutuamente exclusivos). Supor que distribuição binomial seja uma boa aproximação para esses casos

**Estatística de ensaio e dimensão do efeito**

Considerando que, para amostras maiores, a distribuição da amostra da proporção da amostra é aproximadamente normal. Para amostra pequena de população normal, usar

estatística t. Grau de liberdade é n-1. Cuidado c/ amostras muito pequenas!

A mathematical equation with black letters

Description automatically generatedA black and white math equation

Description automatically generatedA diagram of a number of objects

Description automatically generated with medium confidenceA mathematical equation with black text

Description automatically generatedTeste estatística: Tamanho do efeito:

A close-up of a math problem

Description automatically generatedStep 2-

Step 3- Usando tabela Z padrão ou calculadoras online.

(tabela t/calculadoras com df = n-1 para amostras pequenas)

**Ex.5: Teste hipótese para proporção**

- Desenvolveu nova ferramenta de verificação de fraquesas no código para utilizar no desenvolvimento de aplicações web. CTO da empresa decidiu que empresa pode comercializar a ferramenta se for capaz de detetar pelo menos 75% vulnerabilidades do código c/ taxa de falsos positivos inferior a 15%.

- Para testar, utilizou um benchmark composto semeadas com vulnerabilidades representativas. Total 237 vulnerabilidades foram injetadas no código. Ferramenta detetou 185 dessas vulnerabilidades, mas tb indicou 38 vulnerabilidades erradas (falsos positivos). Falsos positivos foram confirmados por inspeção manual. Pode reportar ao CTO c/ 95% confiança de que ferramenta pode detetar + de 75% da vulnerabilidades e c/ menos de 15% de falsos positivos?

Resp- Como problema tem 2 alegações independentes (proporção de vulnerabilidades e proporção de falsos positivos), fazemos 2 testes de hipóteses separados.

Primeiro testar hipótese de vulnerabilidades detetadas.

Step 1-

H0: p≤0,75 (proporção vulnerabilidades detetadas ñ é superior a 0,75 (0,75 é alvo declarado pelo CTO)

A mathematical equation with numbers

Description automatically generatedH1: p>0,75 (proporção vulnerabilidades detetadas é superior 0,75 (ferramenta é melhor do que o limite indicado pelo CTO)

A screenshot of a table

Description automatically generatedStep 2- *z*= 1.088

Step 3- Olhando para tabela Z para z = 1,088 (uma cauda), descobrimos que *p* está entre 0,1401 e 0,1379.

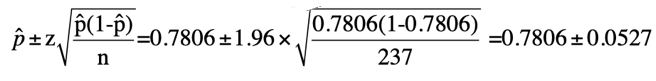
Usando calculadora *p* = 0,1383 = 13,83%

Step 3- Como *p* = 13,83% ñ podemos rejeitar hipótese nula. Isso significa que ferramenta ñ é melhor do que 0,75 com 95% de confiança. Como ferramenta falha no 1º teste, ñ preciso testar falsos positivos. Conclusão: continuar a tentar melhorar a ferramenta!

**Medir os intervalos de confiança:**

A black numbers and a line

Description automatically generated with medium confidenceProporção de vulnerabilidades detetadas - cobertura da ferramenta



CI para um nível de confiança de 95%

Z = 1,96 - tabela Z para 95% de confiança, 2 caudas

Proporção de vulnerabilidades detetadas (95% de confiança) = 0,7806 ± 0,0527 Cobertura do instrumento entre 72,79% e 83,33% com 95% de confiança

A number with a carrot

Description automatically generated with medium confidenceProporção de falsos positivos

223 - falsos positivos, devemos considerar todos *n* 223 vulnerabilidades detectadas: 185 (correcto) + 38 (incorretamente indicado como vulnerabilidades)

CI para um nível de confiança de 95%

Proporção falsos positivos (95% conf.) = 0,1704 ± 0,0493

Detecção falsos positivos da ferramenta situa-se entre 12,11% e 21,98%, com 95% de confiança.

**Teste Z, 2 proporções - teste de hipótese para a diferença entre proporções**

Teste de hipótese para determinar se a diferença entre 2 proporções é significativa.

Para ex. anterior, este teste pode ser útil se pretender comparar proporções de vulnerabilidades detetadas (cobertura) por 2 ferramentas concorrentes, T1 e T2.

Medições obtidas c/ ferramenta são independentes das medições obtidas c/ a outra (medições ñ relacionadas em amostras independentes).

Mesmos passos que teste hipóteses para proporção, mas requer estatística de teste ligeiramente diferente.

- Supor que temos 2 proporções populacionais, P1 E P2

- Hipótese pode ser testada sobre a diferença entre as 2 proporções populacionais

A table with text and symbols

Description automatically generated- Dependendo do experimento e objetivos do pesquisador, 1 ou + dessas hipóteses podem ser relevantes para serem testadas.

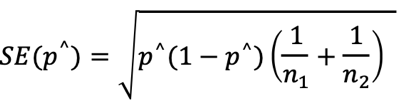
Para calcular erro padrão (SE), precisamos de proporção amostral agrupada (assumir que é semelhante a toda a população)

p1 é proporção da amostra da população 1,

p2 é a proporção da amostra da população 2,

n1 é o tamanho da amostra 1

A red oval with black text

Description automatically generatedA math equation with black letters

Description automatically generated with medium confidencen2 é o tamanho da amostra 2

**Amostras dependentes**

Amostras

Até agora, estudámos testes hipóteses em diferentes cenários de amostragem c/ distribuições (aproximada) conhecidas (Testes Paramétricos):

Tipo de amostras: amostras grandes (≥30): distribuição Z (normal); amostras pequenas (≤30): T distribuição dos alunos

Nº de amostras:

- Amostra única: apenas um grupo de observações; ensaio contra uma média hipotética; Z grandes e T pequenas.

- 2 amostras: 2 grupos de observações; testar a diferença entre médias; Z grandes e T pequenas.

-3 ou + amostras: vários grupos; testar variância (anova)

Natureza das amostras

- Amostras independentes: grupos ñ estão relacionados e observações são verdadeiramente independentes

- Amostras dependentes: quando uma observação num grupo está relacionada c/ observação num outro grupo. Tb chamado de pares combinados, amostras combinadas...

**Ex. Amostras dependentes**

Ex.1: Amostra 1: Downloads por dia das aplicações comercializadas pela empresa

Amostra 2: Downloads por dia das mesmas aplicações após campanha publicitária da empresa na Google

Ex.2: A 1: nº vulnerabilidades de segurança de código encontradas em inspeções de código por 10 engenheiros

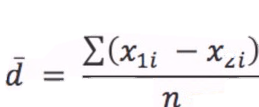
A 2: nº vulnerabilidades de segurança encontradas nas inspeções de código pelos mesmos 10 engenheiros após formação em programação de segurança.

- Amostras são dependentes porque medições podem ser emparelhadas em relação a cada aplicação (exemplo 1) ou a cada engenheiro (exemplo 2), mas existem diferenças sutis entre esses 2 exemplos…

**Teste diferença entre médias, 2 amostras dependentes**

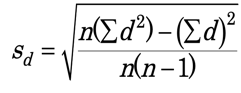
Teste hipótese de 2 amostras c/ amostras dependentes é

com base na média *d* das diferenças entre os dados pareados entradas nas amostras dependentes.

E(X1i- X2i) = diferença entre

entradas para par de dados

n= nº de pares

Desvio padrão Sd das diferenças entre as entradas de dados emparelhadas nas amostras dependentes

Usar média *d* das diferenças entre as entradas de dados emparelhadas nas amostras dependentes e o desvio padrão *Sd* somente se estas condições forem atendidas:

- Amostras devem ser selecionadas aleatoriamente.

- Amostras devem ser dependentes (emparelhadas).

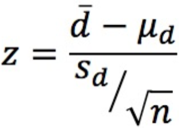
- Ambas populações devem estar normalmente distribuídas.

Se condições forem satisfeitas, então distribuição de amostragem para *d* é aproximada por um valor normal

distribuição para n≥30 ou por uma distribuição T com n-1 graus de liberdade se n < 30.

**Teste estatístico 2 amostras dependentes**

Grandes conjuntos amostras emparelhadas (n ≥ 30) *Z*

*- d:* média diferença entre entradas emparelhadas amostras dependentes

- ***μ****d:* diferença assumida entre médias (geralmente 0 pq H0 é conservadora)

A mathematical equation with black text

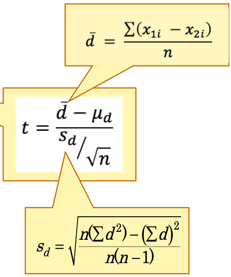
Description automatically generated- *Sd*: desvio padrão das diferenças entre entradas de dados emparelhadas nas amostras dependentes

- √n: nº amostras emparelhadas.

Pequenos conjuntos amostras emparelhadas (n <30) *t*

-√n: grau de liberdade é n-1.

A yellow rectangular boxes with black and white text

Description automatically generated with medium confidenceSetp 2- (amostras grandes) (amostras pequenas)

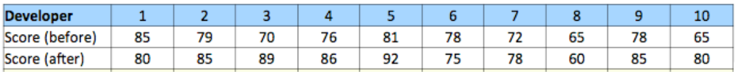
Setp 3- Usar tabela Z ou calculadora online ou tabelas/calculadoras T com df = n-1 para pequenas

**Ex.** **Teste 2 amostras dependentes**

Empresa insatisfeita c/ qualidade código produzido pelos desenvolvedores de aplicações, pois nº de vulnerabilidades de segurança, como injeção de SQL é bastante alto.

Desenvolvedores fizeram teste escrito que consiste em escrever código e registar notas teste.

Desenvolvedores passaram por treinamento específico sobre escrever código seguro e após treinamento repetir teste escrito (ñ o mesmo, mas similar).

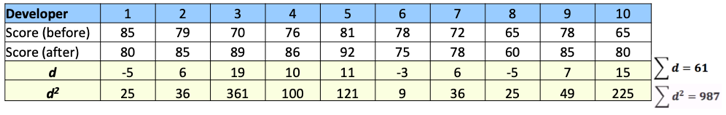
Relatar c/ 95% confiança que treinamento melhorou as habilidades?

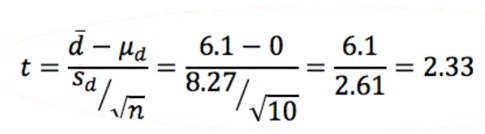
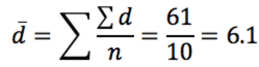
Setp 1-

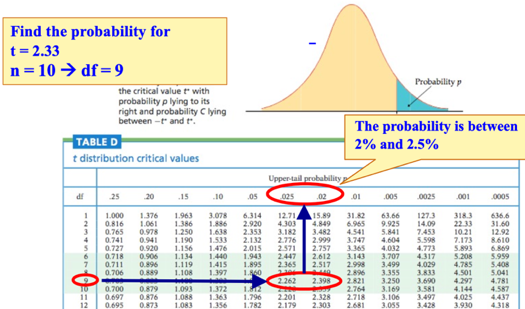
H0: *μd* ≤0 (pontuações testes após treinamento ñ são melhores (treinamento ñ melhorou habilidades)

H1: *μd* >0 (treinamento melhorou habilidades e resultados testes após treinamento são melhores (Claim).

Setp 2- adicionar cálculos intermediários à tabela

A black and red line with numbers

Description automatically generatedTeste analise dentro quadro

Setp 3-

Setp 4-

Valores pequenos de *p* fornecem evidências contra a hipótese nula, pois significam que dados observados são improváveis quando a hipótese nula é verdadeira.

Como *p* está entre 2% e 2,5% o efeito é “significativo”. Podemos rejeitar H0 c/ 95% confiança.

Treinamento melhorou habilidades dos desenvolvedores!

**Resumo amostras dependentes**

Usar quando:

- Medidas repetidas para o mesmo indivíduo/sistema...

- Estudos c/ pares correspondentes de membros da família

Vantagens:

- Fontes conhecidas de possível bias são controladas

- Desvio padrão do teste estatístico é geralmente menor, tornando o poder do teste maior do que um teste Z ou T

Desvantagens:

- Algum caso é difícil encontrar mesmos objetos/ participantes

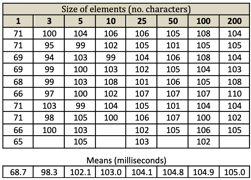
- Quando hipótese nula é rejeitada, muitas vezes é difícil argumentar que a diferença se deve a eventos globais e ñ ao teste-reteste dos mesmos indivíduos.

**ANOVA- Análise De Variância**

**Cenário 2\*-** Ñ pode utiliza teste T de 2 amostras (pelo menos diretamente)

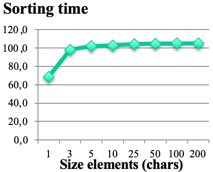
Ex.: Qual é impacto do tamanho dos elementos no tempo de classificação? - Como tamanho dos elementos em uma matriz (cada elemento é uma sequência de caracteres) afeta tempo necessário para classificar todos elementos da matriz?

Depois de realizar algumas experiências com Quicksort (software de ordenamento memoria e tempo) e uma matriz de 10000 elementos, obteve seguintes resultados:

A table with numbers and numbers

Description automatically generatedA graph with blue and green diamonds

Description automatically generatedTamanho dos elementos importa?

Depois observar resultados, decidiu fazer + experimentos:

Mas problema precisa de + do que apenas olhar para os meios utilizando uma abordagem informal…

Objectivo é verificar se a diferença entre as médias das amostras múltiplas é significativa. Usar ANOVA!

**Abordagem Informal**

Para saber se a diferença entre múltiplas médias de amostras é significativa:

- Utilizar abordagens gráficas informais (análise exploratória de dados): gráficos, parcelas de caixa lado a lado, histogramas múltiplos.

- Mas saber se a diferenças entre os grupos (factores) são significativamente dependes: diferença entre os meios; desvios-padrão cada grupo; tamanhos amostras.

- Usar ANOVA para determinar o valor P (a partir da estatística F - outra distribuição e teste como T ou Z)

**One-way ANOVA**

- é usada para testar afirmação de que 3 ou + médias populacionais são iguais

- é extensão dos 2 testes de amostras independentes

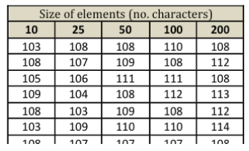
- Testa as seguintes hipóteses:

H0:μ1=μ2=μ3...=μk(médias de todos os grupos são iguais) H1: nem todas médias são iguais

- Ñ diz como ou quais diferem. Precisa acompanhar múltiplas comparações (+ testes).

- Variável dependente: variável que está ser comparada

- Variável independente: variável fatorial usada para definir amostras (grupos)

- Níveis: valores da variável independente selecionada. Cada nível originará uma amostra (grupo)

-Ex: Variável dependente: tempo de classificação

Variável independente: tamanho dos elementos

Níveis: 5 níveis [10 caracteres, 25c, 50c, 100c, 200c]

**Suposições da ANOVA**

- Cada grupo é aproximadamente normal (pode ser verificado informalmente observando o histograma dos dados ou usar teste de normalidade)

- Consegue lidar com alguma anormalidade, mas ñ com discrepâncias graves

- Desvios padrão de cada grupo são aproximadamente iguais. Regra prática: proporção entre o maior e o menor desvio padrão da amostra deve ser menor que 2:1.

**Rationale for ANOVA (justificativa?)**

- Ter pelo menos 3 médias para testar (cada média é de uma amostra). Ex: H0:μ1=μ2=μ3.

- Poderíamos usar teste t de 2 amostras para testar, 2 de cada vez. Mas ANOVA testaremos todos de uma vez.

[Nº de comparações (testes) aumenta quando se utilizam teste t. Erro tipo I (falso positivo) tb aumenta muito. Ex:

Nº amostras- 2, 3, 4, 5. Nº teste- 1, 3, 6, 10]

- Em vez de usar diferença média, ANOVA usa variância das médias do grupo em relação à média geral de todos os grupos.

- Lógica é a mesma do teste t/z: comparar variância observada entre as médias (diferença observada nas médias no teste t/z) c/ o que esperaríamos obter

- Supor temos 3 amostras da mesma população. Resultados: 3 amostras da mesma população

Médias dos 3 grupos ñ são exatamente iguais, mas são próximas, portanto variação entre médias será pequena.

- Supor temos 3 amostras de diferentes populações. Resultados: 3 amostras de 3 populações diferentes

A diagram of a graph

Description automatically generatedA diagram of a normal distribution

Description automatically generatedMédias amostrais estão distantes umas das outras, então a variância entre as médias será grande

Supor que façamos estudo e encontremos os seguintes resultados (qualquer gráfico). Como saberíamos se existe um efeito real ou ñ? Para decidir, podemos comparar a variância observada nas médias com o que esperaríamos obter, se H0 for verdadeira (ñ há diferença nas médias).

**Definições de termos na ANOVA**

Dividir análise da variância em partes significativas que correspondam: efeito variável independente (IV) e erros.

A black letter with a square and a square

Description automatically generated - grande média, assumindo todas as observações.

 - média de qualquer nível do IV (grupo).

 - média de um nível específico (1 neste caso).

 - a observação ou dados brutos para medição.

Variação é soma dos quadrados dos desvios entre um valor e a média da amostra (grupo)

A black letter on a white background

Description automatically generatedSoma dos Quadrados (SS) é frequentemente seguida por uma variável entre parênteses, como que indica qual soma dos quadrados se refere:

**Fontes de variações**

- Somas dos quadrados medem 3 fontes de variação:

-Grupos (variação entre médias de grupo)

-Erro (variação dentro dos grupos)

-Total (SST = SSG + SSE)

- Graus liberdade (*n* observações, *k* amostra significado)

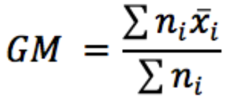
-df(total)=n–1

-df(dentro)=n-k

-df(entre)=k-1

df (total) = df (entre) + df (dentro)

**ANOVA cálculos**

Média geral é média ponderada das médias da amostra individual

A black symbols on a white background

Description automatically generatedSoma total dos quadrados vem da distância de todas as pontuações até a média geral. Este é o grande total.

A black symbol with a white background

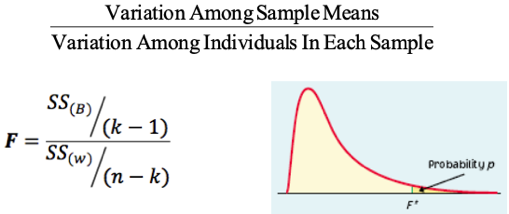
Description automatically generatedSoma dos quadrados dentro do grupo vem da distância das pontuações às médias amostrais. Graus liberdade são iguais à soma *df* individuais cada amostra. Isso indica erro.

A black letters on a white background

Description automatically generatedSoma dos quadrados entre grupos representa a distância das médias amostrais da média geral. Isso indica efeitos IV.

**ANOVA: estatística F**

Teste estatístico para ANOVA é chamada a partir da estatística F, que obtemos do teste F.

Estatística F determina se variação entre médias das amostras é significativa: (F= teste estatistico)

**Obtendo o crítico da tabela F**

Como obter o valor crítico das tabelas F para um determinado α? Ex. α = 0,05?

A graph of a function

Description automatically generatedA person standing next to a number of equations

Description automatically generatedTab. F para α = 0,05, procurar valor crítico para F(3,9)

**Tabela ANOVA**

A table with text and symbols

Description automatically generated with medium confidenceA screenshot of a math equation

Description automatically generated- tab. é resumo de todos os elementos necessários para o cálculo do valor P

A table of numbers with numbers

Description automatically generated**Ex. One-way ANOVA**

Como tamanho elementos de array afeta tempo para classificar todos os elementos do array?

Consideraremos 3 níveis da variável independente, tamanho dos elementos: 3, 5 e 10 caracteres.

Tabela mostra tempos de classificação em

milissegundos obtidos c/ Quicksort e um array de

10.000 elementos.

Step 1-

H0: μ3=μ5=μ10 - Todas amostras têm médias iguais

H1: μ3≠μ5≠μ10 - Nem todas médias são iguais

– Ñ diz como ou quais diferem

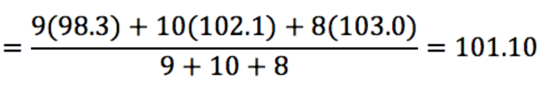
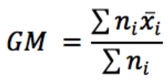
– Pode acompanhar múltiplas comparações

A table with numbers and a number of objects

Description automatically generated with medium confidenceStep 2- Determinar as características das amostras em comparação. *n* =27, *k* =3

Cálculo da Grande Média

– Média ponderada das médias amostrais individuais



Variação entre grupos

– Variação entre cada média da amostra e a média geral

– Cada variação do grupo é ponderada pelo tamanho da amostra



Dentro da variação do grupo

– É o total ponderado das variações individuais

– Ponderação é feita com graus de liberdade. Para cada amostra, *df* é um a menos que o tamanho da amostra



A table with numbers and letters

Description automatically generatedA table with text and numbers

Description automatically generatedAgora podemos preencher tab. ANOVA unidirecional

Step 3- Valor P para F(2,24) = 7,48 é 0,002986 (usando calculadora online)

Step 4- H0:μ3=μ5=μ10- Todas amostras têm médias iguais

P ≈ 0,003 à H0 é rejeitado!

**Two-way ANOVA**

- Testa igualdade de 2 ou + médias populacionais quando são utilizadas 2 variáveis independentes: fator A e B (+ de 2 fatores: ANOVA multidirecional).

- Cada variável independente (fatores) pode ter qualquer nº de níveis.

- Mesmos resultados da one-way ANOVA separada em cada variável. Mas interação pode ser testada.

- Economiza tempo e esforço, em comparação com testes ANOVA unilaterais consecutivos.

**Suposições da Two-way ANOVA**

- Normalidade (populações são normalmente distribuídas)

- Homogeneidade (populações têm variâncias semelhante)

- Independência de erros (amostras aleatórias independentes)

**Two-way ANOVA: Hipotese nula**

Testa 3 hipóteses simultaneamente:

- Nenhuma diferença nas médias devido ao fator A H0:μ1.=μ2.=...=μa.

- Nenhuma diferença nas médias devido ao fator B H0:μ.1=μ.2=...=μ.b

A black and white logo

Description automatically generated- S/ interação dos fatores A e B – H0: = 0

**One-way VS. Two-way**

- Teste segue os mesmos passos em ambos os casos.

- Two-way é semelhante (mas tem + linhas do que a tabela one-way, pois há + somas de quadrados e + graus de liberdade para calcular).

- Fórmulas são iguais.

- Métrica F é a mesma.

- Interpretação dos resultados (two-way é + complexa, mas ainda bastante direta).

A diagram of anova analysis

Description automatically generatedANOVA two-ways Somas de quadrados (SS)

A table with numbers and letters

Description automatically generated

A table with numbers and numbers

Description automatically generated**Ex. Two-ways ANOVA**

Como nº elementos em uma matriz e tamanho

de cada elemento afetam tempo de classificação?

Considerar 3 níveis da variável independente tamanho do

elemento (10, 50 e 100 caracteres) e 2 níveis da variável independente nº de elementos (10.000 e 50.000).

Tab. mostra tempos de classificação em milissegundos obtidos c/ implementação Quicksort.

Step 1-

Fator A – Tamanho (car.) elementos a serem classificados

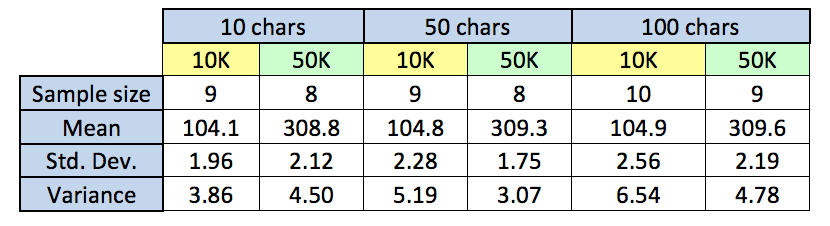
Fator B – Nº elementos a serem classificados

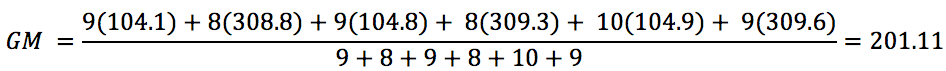
- H0:μa.10=μa50=μa.100 - tamanho elementos a serem classificados ñ é relevante para tempo de classificação (médias são iguais)

- H0:μb.10k=μb.100K - nº elementos a serem ordenados ñ é relevante para tempo ordenação (médias são iguais)

- H0:ABij=0 - Ñ há interação entre tamanho e nº de elementos

- H1: Nem todas médias são iguais e há interação (só é possível na cauda direita).

Step 2- Determinar características das amostras em comparação.



Média ponderada das médias da amostra individual:

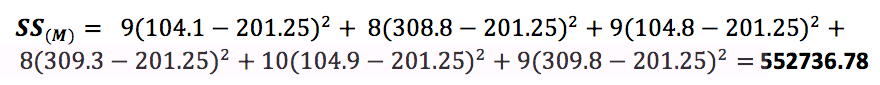
SS(T)- soma total dos quadrados vem da distância de todas as pontuações da média geral. Este é grande total.

A diagram of a measurement

Description automatically generated with medium confidenceRealizar cálculo: SS(T)=552982 (este valor ñ é necessário nos cálculos)

A diagram of a number of numbers

Description automatically generated with medium confidenceSS(M)– soma dos quadrados que dá a variância devido à manipulação experimental de fatores (todos fatores são considerados aqui). Cálcular:



SS(A)– soma dos quadrados do fator A (efeito do fator A, tamanho de cada elemento a ser classificado). Organizar dados de acordo c/ observação do fator A.



A table with numbers and numbers

Description automatically generatedA black letters and numbers on a white background

Description automatically generatedA table with numbers and numbers

Description automatically generatedA black letter and numbers

Description automatically generated with medium confidence SS(B)– soma dos quadrados para o fator B (efeito fator B, nº elementos a serem classificados) Organizar dados de acordo c/ observações do fator B. Calcular médias



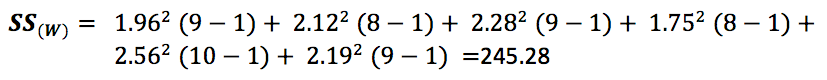
SS(AxB)– soma dos quadrados para a interação entre o fator A e o fator B. Realizar cálculo:



Dentro da variação do grupo

– É o total ponderado das variações individuais (erro)

– Ponderação é feita c/ graus de liberdade. Para cada amostra, df é um a menos que o tamanho da amostra



A table with numbers and letters

Description automatically generatedSumario Tabela:

A graph on a sheet of paper

Description automatically generatedStep 3- Encontrar p

Step 4-

H0: μb.10=μb.50=μb.100 - tamanho elementos a classificar ñ é relevante para tempo de classificação (médias são iguais)

H0: μ10K.a=μ50K. – nº elementos a ordenar ñ é relevante o tempo de classificação (médias são iguais)

A black text on a white background

Description automatically generatedH0: AB1=0 - Ñ há interação entre tamanho/nº elementos

H1: Nem todos as media são iguais e há interação (apenas a cauda direita é possível).

A math equations on a white background

Description automatically generated with medium confidenceP=36% para fator A, dimensão de elementos a ordenar- H0 é mantido para este fator: tamanho elementos ñ é relevante para o período de classificação

P≈0 para fator B, nº elementos a ser classificado- H0 é rejeitado: nº elementos a serem classificados é significativo e determina tempo de classificação.

P=63% para interações entre fator A e B- H0 é mantido: ñ existe interações entre tamanho e nº elementos.

H0 é rejeitado

**Inferência estatística não paramétrica**

Testes hipóteses ñ paramétricos - existem situações para as quais ñ é possível aplicar estatísticas paramétricas: dados têm classificação, mas ñ têm interpretação numérica clara, como as preferências do utilizador; parâmetro da população cuja distribuição é desconhecida (medianas, variâncias ...).

Pós: métodos ñ paramétricos fazem menos suposições do que os paramétricos; são livres de distribuição

Contras: nos casos em que um teste paramétrico seria adequado, os testes ñ paramétricos têm menos potência.

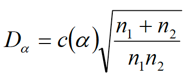
**A black text on a white background

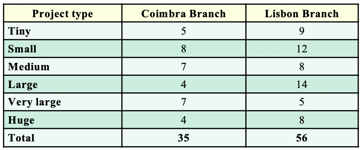
Description automatically generatedTeste Kolmogorov-Smirnov (2 amostras)**- compara 2 funções de distribuição empírica. Teste estatístico é:

A graph of a number of fractions

Description automatically generatedF1,n1 e F2,n2 são funções de distribuição empírica e "supx" é função suprema. Valores da estatística de ensaio são tabulados. Teste K-S avalia a significância da divergência máxima *D* entre duas curvas de frequência cumulativa. Se *D* for maior do que um valor crítico para um determinado ⍺, então as diferenças entre as duas funções são significativas.

**Valores críticos para teste K-S 2 amostras (2 lados)**

Quadro apresenta valores d críticos para valor de 0,05 (v. superior) e 0,01 (valor inferior) para vários tamanhos de amostra. \* significa ñ pode rejeitar H0 independente dos dados observados. Para amostras maiores, o valor crítico aproximado *D⍺* é dado pela equação:

A number of numbers on a white background

Description automatically generatedA graph with numbers and a number

Description automatically generatedCoeficiente *c(⍺)* para os valores típicos de décimos é:

Ex.: Empresa está analisar falhas nos projectos SW, objetivo é avaliar falhas dos projectos em Coimbra e Lisboa são diferentes (significância 95%) ou ñ. Se houver diferença, empresa reavaliará as práticas c/ maior taxa de insucesso.

Step 1- H0: Distrib1 = Distrib2 (são semelhantes, taxa insucesso é mesma nos 2 ramos)

H1: Distribuição1 ≠ Distribuição2 (são diferentes, taxa de insucesso é diferente, filial c/ maior percentagem de insucesso é pior)

A math equations and numbers

Description automatically generated with medium confidenceStep 2- significância 95%*→* ⍺=0,05. Valor crítico K-S para valores grandes de n1 e n2 é (aproximadamente) dada pela fórmula:

1.36*→* coeficiente c(⍺) é ⍺ = 0,05

0.29*→* Valor crítico, se diferença entre 2 distribuições cumulativas for superior a este *→* rejeitar H0)

A black text on a white background

Description automatically generatedStep 3- teste estatístico é:

A table with numbers and a number

Description automatically generatedEsta é diferença máxima no valor absoluto (classes do projeto são ordenadas)

Step 4- Valor crítico: D⍺(crític.)=0.29

Teste estatístico: D(máx.)=0,082

D(max) é inferior D(critic.). Conclusão: ñ rejeitar H0

Ñ podemos dizer que distribuições sejam diferentes, c/ 95% confiança.

Teste pressuposto da normalidade: normalidade dos dados é necessária para teste-T, pode ser necessário testar os dados se seguem distribuição normal. Métodos gráficos para testar normalidade incluem histograma e qq-gráfico.

Tb podem ser utilizados testes ñ paramétricos como K-S e Shapiro-Wilk.

**Bootstrappin-** estima precisão estatísticas das amostras, desenhando aleatoriamente c/ substituição do conjunto de pontos de dados. É sado para estimar desvio padrão e circuito de computação.

Procedimento: 1) Reamostrar os dados c/ substituição, de modo que tamanho reamostra seja igual ao tamanho do conjunto de dados original. 2) Calcular estatística a partir da reamostra da 1ª etapa. 3) Repitir etapas 1 e 2 muitas vezes (1000 a 10000) para obter estimativa + precisa da distribuição.

Distribuição bootstrap aproxima-se da distribuição amostral da estatística. Esta estatística pode ser diferente da média (mediana, variância, correlação...); geralmente têm a mesma forma e distribuição que a distribuição de amostragem, mas ñ estão centradas na estatística (da amostra original).Tb é conhecido como método estatístico intensivo em computador.

Ex.: Aproximar distribuição da proporção da amostra para nº de cabeças encontrado de 100 lançamentos de moedas.

A graph of a number of columns

Description automatically generated with medium confidenceGerar amostra tamanho 100 (x1,..., x100). Reamostrar aleatoriamente observações c/ substituição e calcular a proporção de cabeças para cada amostra. Reamostrar várias vezes (Monte Carlo) para obter distribuição de bootstrap empírica da média da amostra.

Ensaios de randomização- ensaio estatístico em que a distribuição estatística de ensaio em H0 é obtida através do cálculo de todos os valores possíveis da estatística de ensaio sob rearranjos dos rótulos nos pontos de dados observados. Muito utilizado como teste estatístico de 2 amostras. Procedimento: semelhante ao bootstrap, mas amostra é construída s/ substituição. Pressuposto: rótulos podem ser trocados em H0. Condição suficiente para a permutabilidade: variáveis são i.i.d.

A graph of a number of numbers and a line

Description automatically generated with medium confidence**Ex. Teste hipóteses usando testes de randomização (2 amostras independentes) Cenário 2-** Reatribuir aleatoriamente as observações entre os 2 grupos e calcular média. Reamostrar e calcular média várias vezes (Monte Carlo) para obter distribuição empírica da média da amostra. Calcular *P* para dado valor de significância c/ base na distribuição empírica da média da amostra.

**Teste de Mann-Whitney-** assume apenas um nível ordinal de medição, baseia se na classificação das pontuações. Testa 2 amostras da mesma população (equivalente teste-t ñ pareado de 2 amostras).

Classifica conjunto de pontuações n1 e n2 do + baixo (classificação 1) ao mais alto. Seja R1 a soma das classificações da amostra + pequena (a do tamanho n1)

A mathematical equation with numbers and symbols

Description automatically generatedTeste estatístico:

Valores para U são tabulados para n<20. Aproxima-se da distribuição normal para tamanhos maiores.

**Wilcoxon signed-ranks test -** pressupõe um nível de medição de intervalo. Testa 2 amostras pareadas da mesma população (equivalente teste-t ñ pareado de 2 amostras).1)Calcular diferença entre cada par e classificar. 2) Cada classificação recebe o sinal da diferença a que corresponde. 3) Somar classificações positivas e somar classificações negativas. Teste estatístico é a menor soma. Valores são tabulados para n<20. Aproxima-se da distribuição normal para tamanhos maiores.

**Sign Test -** teste tem apenas em conta o sinal das diferenças entre os pares (equivalente teste-t ñ pareado de 2 amostras e menos potente do que os postos sinalizados de Wilcoxon). É aplicado quando ñ há informações de intervalo.1) Calcular sinal das diferenças entre cada par e ignorar aqueles que ñ têm diferença (reduzir o tamanho da amostra de acordo)2)Teste estatístico é nº de pares c/ sinal menos frequente.

Valores são tabulados para n<25. Aproxima-se da distribuição normal para tamanhos maiores (uma vez que está relacionada c/ distribuição binomial).

**Experiências c/ Pessoas**

Sistemas/programas informáticos são desenvolvidos, mantidos e utilizados por pessoas, estes grupos ñ são as mesmas pessoas e é difícil antecipar o que as outras pensam e fazem. Necessidade de experimentar para descobrir e melhorar.

Áreas de Aplicação: Engenharia software - Como pessoas projetam sistemas? Quais são bons procedimentos?

Testes usabilidade IHC - Como é que utilizam sistemas?

Quais são boas orientações de design de interface?

Perceção usuário sobre uso/desempenho do sistema – C/ o que os usuários se preocupam?

Avaliação segurança- perceção dos utilizadores sobre a segurança do sistema e o seu impacto na usabilidade

Avaliar mercado de produtos Informáticos

E redes sociais, obviamente...

Ex. perguntas: Que processos/técnicas funcionam melhor? Resp: teste vs. inspeção de código; desenho detalhado vs. programação ágil; menus amplos vs. menus profundos

Variação entre programadores/utilizadores experientes e novatos? Resp. como tornar + fácil de aprender para iniciantes? Como tornar eficiente para experientes? Ambos podem utilizar os mesmos mecanismos? ...

Técnicas experimentais

1) Observação e análise de dados (ver como utilizadores se comportam por conta própria enquanto utilizam aplicações, através monitorização e análise de dados

2) Experiências controladas (ver se utilizadores realizam tarefas pré-definidas e como comportamento muda quando parâmetro do sistema específico é alterado)

3) Entrevistas e inquéritos (entender por que usuários se comportam daquela maneira, preferências, necessidades... Impressão subjetiva do participante

**1) Observar comportamento utilizador:** inquérito contextual, observação pormenorizada de pequeno nº de pessoas durante o seu trabalho normal. Quais são os verdadeiros problemas? Necessidades reais? Onde pode trazer valor real?

Nível profundo de exigência: utilizadores ñ sabem articular aquilo que precisam. O que querem nem sempre é o que precisam. Que precisam pode ñ estar relacionado c/ computador... observação e análise para descobrir.

Sistema é alvo da avaliação: ñ os utilizadores.

Abordagem Experimental: olhar comportamento utilizador

Objetivos: Compreender o que interessa aos utilizadores no sistema, o seu desempenho e comportamento. Utilizar registos da atividade, gravação teclas/ecrã, rato... para reduzir ao máximo a intrusão

Analisar atividade cada utilizador separadamente; nº participantes é pequeno ("observação" de grande nº de utilizadores na Internet através do registo automático pode ser feita através de abordagens de IA)

Para cada tarefa, mede o desempenho dos utilizadores.

Ex. conclusão da tarefa (Sim/Não), nº de erros, tempo...

Correlacionar desempenho c/ comportamento.

**2) Experiências controladas c/ utilizadores -** etapas:

1. Definir objetivos do sistema ou módulo em Avaliação - Que serviços/funcionalidades oferece?

2. Criar conjunto tarefas que são executadas para atingir esses objetivos

3. Definir medidas/observações: desempenho (ex. nº erros, conclusão da tarefa, tempo...), opinião subjetiva dos participantes (questionários)

4. Obter pessoas representativas utilizadores do sistema

5. Observar (registar) tentando executar as tarefas

Aplicável aos clientes de sítio web, programadores de uma nova aplicação.

**Aspetos experimentais**

Medir desempenho nas diferentes tarefas - média dos utilizadores de cada tarefa (nº médio de erros...)

Gravar vídeo (captura ecrã) para análise posterior-

(comportamentos, expressões e erros)

Recolher experiência subjetiva: formular declaração ou pergunta sobre sistema. Opinião c/ nº de posições (para evitar a temida tendência pontual)

Ex: Discordo totalmente, D, C, Concordo totalmente

Perguntas devem ser apresentadas sob a forma de declarações, respondida utilizando escala proposta. Ex.

"O sistema é fácil de utilizar"

**Definição tarefas (ex. estudos usabilidade)**

Demasiadas tarefas, ñ é possível testar todos, fazer lista de tarefas e classificá-las por importância, escala de 1-6, classificar pelo grau de dúvida (ou feedback proprietário ou usuário informal) têm sobre eles, escala 1-6. Multiplicar duas classificações e classificar resultado. Testar tarefas topo: importantes e + requerem intervenção utilizador, definir objetivos primeiro, ñ o procedimento. Objetivo é saber qual procedimento os utilizadores irão utilizar (ser específico e claro sobre o que usuários façam), criar sequência razoável, evitar uso de palavras que aparecem na interface. Juntos, ñ devem demorar muito tempo, estimar quanto tempo levará (especialista que conhece o sistema), multiplicar por 3 a 10, dependendo do perfil do testador.

**Seleção de pessoas-** recrutamento: encontrar pessoas

c/ base em dados demográficos gerais e características diversas– idade, nível rendimento, uso computador...

Triagem: encontrar pessoas certas, filtrar aqueles que correspondem à demografia, mas provavelmente ñ são úteis, por muitas razões..., testadores devem estar interessados (mas ñ predispostos) no sistema, talvez usar sistema semelhante e estar disponíveis datas previstas para testes, ñ devem trabalhar na indústria ou para concorrentes

**Realizar sessão de utilizador**

Explicar que utilizador está ajudar a testar o sistema, ñ é sistema que está a testar o utilizador. Ñ há respostas erradas. Se ñ compreenderem ou tiverem dificuldades, tudo bem, objetivo é saber sobre isso. Usuário deve dizer o que ele está a tentar fazer e porque, usuários ñ deve ter vergonha (registar processo, organizador do estudo está apenas em segundo plano)

**Plano Experimental -** Melhor ter vários pequenos do que um estudo enorme, nº indivíduos pode ser baixo 5-6, suficiente para sentir resultados, ñ necessariamente boas estatísticas.Realizar sessão piloto: descobrir adequação para diferentes dados demográficos do usuário, verificar se tarefas são razoáveis e se descrição do sistema e tarefas é compreensível.

**Entrevistas e inquéritos-** População entrevistada:

bias amostragem (pretende utilizadores representativos),

dimensão amostra: quanto maior melhor, mas experiências c/ pessoas são dispendiosas

Formulação perguntas: perguntas neutras para ñ afetar resultados, ordem perguntas tb é importante. Pré-testar perguntas numa pequena amostra para detetar e corrigir problemas. Análise estatística dos resultados

**Tipos entrevistas-** Entrevista ñ estruturada (intercâmbio e recolha de informações totalmente livres, utilizado como instrumento exploratório nas fases iniciais estudo, quando investigador ainda ñ sabe muito)

Entrevista semiestruturada (seguir esboço pré-definido de perguntas, permitir que usuário expanda vários tópicos, tb questionário em linha em que as perguntas dependem de respostas anteriores.

Entrevista estruturada (preenchimento questionário pré-definido)

**Estrutura questionário**- Título, Breve introdução (do que se trata), Questões demográficas (quem responde), Começar c/perguntas fáceis, Deixar questões sensíveis fim. Seja claro e nítido e pré-teste todas perguntas

Tipos: escolha múltipla ou escala, numérico (quantas vezes por dia utiliza?), texto aberto (o que faria na pág. seguinte?), ao dar opções, incluir "N/A", "Outro"; fornecer explicações em texto, além da escala.

Escala desejável é discutível: nº de pontos deve situar-se entre 4 e 8, nº pontos deve ser uniforme para evitar resposta intermediária indecisa.

**Considerações éticas-** são essenciais, devem ser tomadas medidas explícitas para evitar problema, mas ingrediente chave para experiências c/ pessoas é confiança.

Estudos incluem tratamento de informações confidenciais numa organização. Isto deve ser tido em conta.

Principais fatores éticos: consentimento informado, aprovação do comité de revisão, confidencialidade, manipulação de resultados sensíveis-incentivos, feedback.

A table of numbers with yellow and green lines

Description automatically generated with medium confidenceA table of numbers and percentages

Description automatically generatedTabelas *Z* e *t* :

Usar calculadora online one-way ANOVA - https://www.danielsoper.com/statcalc/calculator.aspx?id=43

Fazer exescicios

Passar aquilo que me explicaram