

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 experiência (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 homogêneos (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 (não 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?

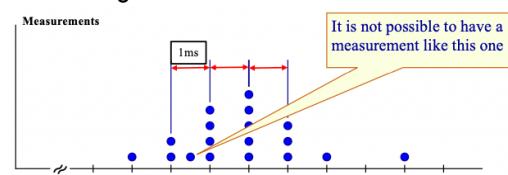
a) Dependente- nº de bugs

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)

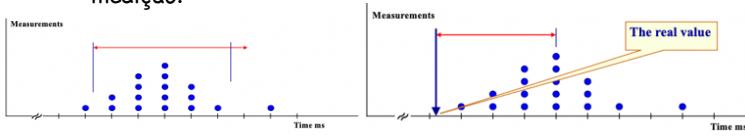
Avaliar complexidade: métricas de complexidade de software(sw), complexidade ciclomática(VG) - indentação.

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:

- 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 imprecisa, calibração incorreta, tempo reação da ferramenta... Quando identificado, pode ser eliminado (uma das etapas do experimento).

I.S.: casos especiais

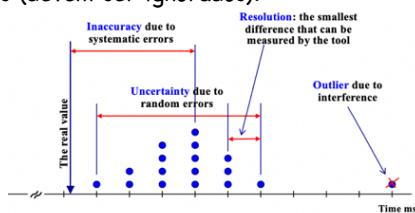
Warm-up: primeira medição pode ser diferente das subsequentes.

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 medições mostrariam valores um pouco diferentes.
- 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.

Resolução- 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:

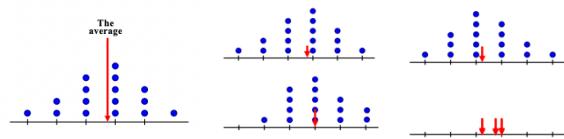
```
for (i=0; i<N; i++) {
t1 = gettimeofday();
<operação sendo medida> t2 = gettimeofday();
tempo[i] = t2 - t1;
print "t. médio é", avg(time[0.. N-1]), "\n";
```

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

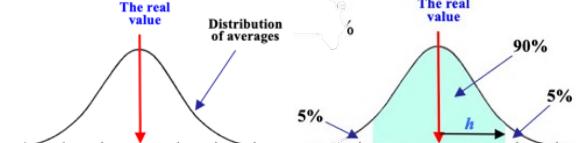
Intervalos de confiança (básicos)

- 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
- 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

Suposição: 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 $\pm h$ tem probabilidade 0,9 para incluir o valor real

Calcular intervalos de confiança:

- μ é o meio real da distribuição baseada
- \bar{x} é a média de n amostras
- distribuição for normal, então as médias terão distribuição t ou Z se amostra for grande ($n \geq 30$)
- α é a incerteza aceitável (implica que o nível de confiança é $1 - \alpha$) e defina a meia largura

$$h = t_{n-1, 1-\alpha/2} \frac{s_{\bar{x}}}{\sqrt{n}}$$

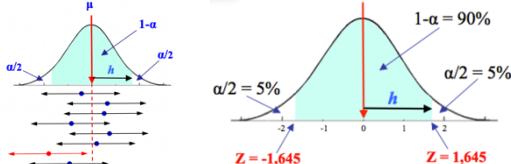
- $t_{n-1, 1-\alpha/2}$, vem das tabelas (t para $n < 30$)
 n é nº de amostras; $n-1$ graus de liberdade (t Student)
($n > 30$ utilizar z, distribuição normal)
- $s_{\bar{x}}$ é o desvio padrão das médias. Assumindo as amostras de base são independentes, é calculado assim s/\sqrt{n} , s é o desvio-padrão das amostras.

Intervalo de confiança:

$$p(|\bar{x} - \mu| < h) = 1 - \alpha$$

Com uma certeza de $1 - \alpha$, a distância entre uma amostra média \bar{x} e a verdadeira média μ é inferior à metade. Se repetirmos isso muitas vezes, e cada vez que desenharmos um segmento de $\pm h$ em torno de \bar{x} , logo em $1 - \alpha$ dos casos este o segmento incluirá μ .

Com uma certeza de $1 - \alpha$ a distância entre uma amostra média \bar{x} e o verdadeiro médio μ é inferior a h . Se repetir medição muitas vezes, e cada vez desenhar segmento de $\pm h$ em torno de \bar{x} , então em $1 - \alpha$ dos casos o segmento incluir μ



Assumindo que amostras são independentes, fórmula é:
 $\bar{x} \pm t^* s/\sqrt{n}$ - ($n \leq 30$, tabela t com df = $n-1$)
 $\bar{x} \pm Z^* s/\sqrt{n}$ - ($n \geq 30$, tabela Z para o normal padrão distribuição)

Onde:

s é desvio padrão das amostras n

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

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 ($n \leq 30$) 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 $\alpha = 5\%$? (bilateral)

- Subtrair α de 1: $1 - 0,05 = 0,95$
- Dividir resultado 2 (bilateral) $0,95/2 = 0,475$

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

Valor + próximo do coeficiente Z está na interseção da linha 1,9 coluna 0,06. Somando 2 valores é $Z = 1,96$ para $\alpha = 5\%$

z	Confidence Level
0.0	0.0000
0.01	0.0040
0.02	0.0080
0.03	0.0120
0.04	0.0160
0.05	0.0190
0.06	0.0220
0.07	0.0250
0.08	0.0279
0.09	0.0308
0.10	0.0337
0.11	0.0366
0.12	0.0395
0.13	0.0423
0.14	0.0452
0.15	0.0481
0.16	0.0509
0.17	0.0538
0.18	0.0566
0.19	0.0594
0.20	0.0622
0.21	0.0650
0.22	0.0678
0.23	0.0706
0.24	0.0733
0.25	0.0760
0.26	0.0787
0.27	0.0814
0.28	0.0840
0.29	0.0866
0.30	0.0891
0.31	0.0916
0.32	0.0940
0.33	0.0964
0.34	0.0987
0.35	0.1010
0.36	0.1032
0.37	0.1054
0.38	0.1075
0.39	0.1096
0.40	0.1116
0.41	0.1136
0.42	0.1155
0.43	0.1174
0.44	0.1192
0.45	0.1210
0.46	0.1227
0.47	0.1244
0.48	0.1261
0.49	0.1277
0.50	0.1293
0.51	0.1308
0.52	0.1323
0.53	0.1337
0.54	0.1351
0.55	0.1364
0.56	0.1377
0.57	0.1390
0.58	0.1402
0.59	0.1414
0.60	0.1426
0.61	0.1437
0.62	0.1448
0.63	0.1458
0.64	0.1468
0.65	0.1478
0.66	0.1487
0.67	0.1496
0.68	0.1504
0.69	0.1512
0.70	0.1520
0.71	0.1527
0.72	0.1534
0.73	0.1540
0.74	0.1546
0.75	0.1551
0.76	0.1556
0.77	0.1560
0.78	0.1564
0.79	0.1568
0.80	0.1571
0.81	0.1574
0.82	0.1577
0.83	0.1580
0.84	0.1582
0.85	0.1584
0.86	0.1586
0.87	0.1588
0.88	0.1590
0.89	0.1591
0.90	0.1592
0.91	0.1593
0.92	0.1594
0.93	0.1595
0.94	0.1596
0.95	0.1597
0.96	0.1598
0.97	0.1599
0.98	0.1600
0.99	0.1601

Confidence Level	z
0.90	1.645
0.91	1.70
0.92	1.75
0.93	1.81
0.94	1.88
0.95	1.96
0.96	2.05
0.97	2.17
0.98	2.33
0.99	2.575

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

	90%	99%
n of samples	32	32
Z	1.65	2.575
S (std dev)	330.51	330.51
average	3130.31	3130.31
Confidence interval	96.11	150.45
Exec. time minimum	3034.20	2979.86
Exec. time maximum	3226.42	3280.76

- 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 \tilde{n} , Vulnerabilidade detectada ou \tilde{n}

Modelo binomial

Propriedades variável binomial:

- É binária (assume apenas 1 de 2 valores 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 \tilde{n} afeta o resultado de outro ensaio.

Distribuição por proporção da amostra:

Proporção da amostra: $\hat{p} = \frac{x}{n}$

(amostra = Conjunto de ensaios)

\hat{p} é proporção da amostra c/ resultado de interesse. É estimativa proporção da população p .

\hat{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 \tilde{n} sucessos ser maior que 10 ($np \geq 10$ e $n(1-p) \geq 10$)
 - 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

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Package	exec. time
74	
66	
88	
68	
...	
87	
79	
78	
72	
86	
85	
86	

32

times

avg = 78.15

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 nenhouver certeza de que dados seguem distribuição normal, devemos testá-la para normalidade.

Conf. A	Conf. B
exec. time	exec. time
74	69
66	71
88	80
68	88
79	64
68	65
87	74
79	76
78	89
72	68
86	67
85	72
86	72

Avg A = 78.15

n = 13

Avg B = 73.58

n = 12

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 sensu*)
- 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.

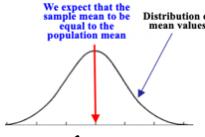
Hip. são rejeitadas ou aceites c/ algum grau de certeza

Hipótese: colocada em perspectiva

- 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 às 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, com 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 (H_0) é uma afirmação sobre o parâmetro populacional (ex. média populacional) que é considerado verdadeiro, (resposta provisória ao problema em estudo)

Ex.: H_0 - Nova configuração não afeta tempo de execução.

- Hipótese alternativa (H_1) é afirmação que contradiz diretamente a nula ao afirmar que o valor real da população não é igual ao valor da nula, (isso é o que achamos que está errado na hipótese nula).

Ex.: H_1 - Tempo execução é diferente na nova configuração (menor ou maior).

Decisão tomada no teste centra-se na hipótese H_0 (nula)

- Ideia é mostrar evidências de que H_0 é improvável, a fim de a rejeitar. Se não o fizer, hipótese nula é mantida.
- O contrário é não fazer nada, encargo é colocado sobre o pesquisador para demonstrar que H_0 não é provável que seja verdade. [Experiências definidas para coletar dados para mostrar que H_0 , não é verdade]

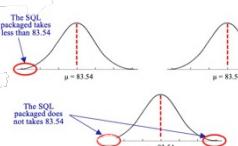
Step 2- Select the criteria for a decision

- 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 - Compute 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. Medir distância entre a média da amostra e a média da população em H_0 . Quanto maior o valor de $|Z_c|$, mais indicará que H_0 não é verdadeiro.

Test statistic:

$$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}}$$

Mean of the sample Mean of the population
Standard deviation of the population Standard error
Number of elements in the sample

Step 4 - Make a Decision

- Valor do teste estatístico (Z_c) 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 (não alcançará significância)

Cenário 1* - Non-directional (Two Tail)

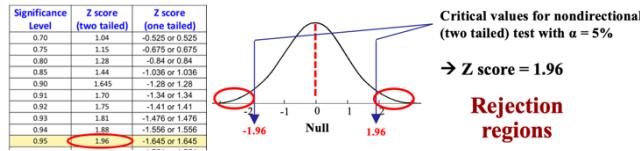
Step 1 - State the hypothesis

- H_0 - nova configuração não afeta sobre tempo de execução.
 H_1 - tempo execução é diferente na nova configuração (menor ou maior)
- Testar se hipótese nula H_0 é verdadeira.

Step 2 - Select the criteria for a decision

Considerar nível de significância 5% [$\alpha=0,05$ $1-\alpha = 0,95$]
Localizar pontuação Z (tabela normal padrão) que representa os valores críticos

Valor 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.



Critical values for non-directional (two tailed) test with $\alpha = 5\%$
 $Z_c = 1.96$
Rejection regions

Step 3 - Compute the test statistics

- Test statistic:
- $$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}} = \frac{78.15 - 83.54}{16.36 / \sqrt{32}} = -1.86$$

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 $Z_c = -1.86$ é 0,0314. Como é bilateral $P = 0,0314 \times 2 = 0,0628 \rightarrow P = 6,28\%$

Significa que probabilidade de obter uma média de 78,15 se H_0 for verdadeiro é 6,28%



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

H_0 - nova configuração não afeta sobre tempo de execução.

(tempo médio execução é 83,54)

H_1 - tempo execução é diferente na nova configuração

(menor ou maior)

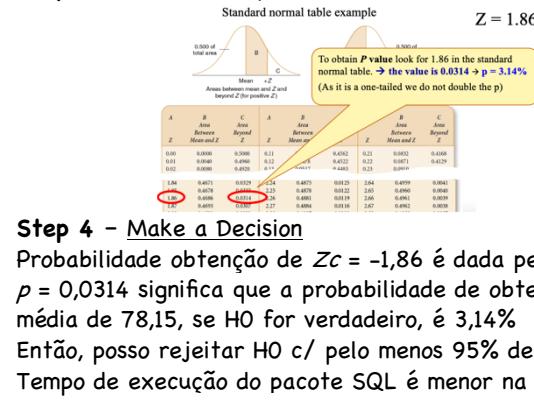
- Testar se hipótese nula H_0 é verdadeira.

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 - Compute the test statistics

- Test statistic:
- $$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}} = \frac{78.15 - 83.54}{16.36 / \sqrt{32}} = -1.86$$

Step 3 - Obtain the p value



Típos de erro

Conclusão pode estar errada, pois estamos a analisar uma amostra com número limitado de elementos n . [Falso positivo - homem gravido, Falso negativo - gravida não gravida]

Decision	
Truth in the population	
True	Correct $1 - \alpha$
False	Type II error β
	Correct $1 - \beta$ (Power)

False negative

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

T-test

- Teste t segue a distribuição T Student (se hipótese nula for verdadeiro)
- T-test 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 não 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)
- Step 2: Compute the test statistics:

$$t_c = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Mean of the sample Mean of the population (assumed, often stated as a target)
 Standard deviation of the sample Number of elements of the sample

Ex3. hipótese 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 não 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-

$H_0: \mu = 70$ (classe sabe programar em C# c/ competência equivalente a 70)

$H_1: \mu_1 > 70$ (classe é melhor em programação C# do que a pontuação de 70)

Step2-

Média da amostra: 79,17

Desvio-padrão amostra: 13,17

Step3-



TABLE D t distribution critical values									
df	Upper-tail probability p								
	.25	.20	.15	.10	.05	.025	.01	.005	.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.207	4.781	6.999	9.249
3	0.727	0.941	1.282	1.645	2.351	3.182	3.587	5.893	8.281
4	0.741	0.943	1.190	1.353	2.131	2.776	3.078	4.604	6.223
5	0.708	0.906	1.143	1.282	1.980	2.447	2.778	4.025	5.783
6	0.718	0.908	1.139	1.240	1.943	2.421	2.707	3.899	5.591
7	0.717	0.906	1.133	1.232	1.925	2.398	2.621	3.726	5.353
8	0.703	0.883	1.100	1.283	1.883	2.262	2.500	3.559	5.233
9	0.699	0.874	1.088	1.263	1.851	2.201	2.428	3.478	5.144
10	0.697	0.874	1.083	1.263	1.839	2.178	2.398	3.407	5.054
11	0.695	0.874	1.085	1.263	1.839	2.171	2.391	3.399	4.964
12	0.695	0.874	1.085	1.263	1.839	2.171	2.389	3.393	4.874

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 H_0 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 não 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
- Na prática, valor p tb dá ideia do tamanho do efeito.

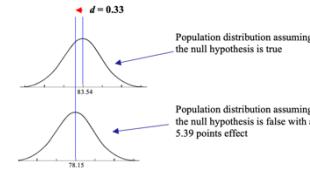
Description of Effect	Effect Size (d)
Small	$ d < 0.2$
Medium	$0.2 < d < 0.8$
Large	$ d > 0.8$

$Cohen's d = \frac{M - \mu}{\sigma}$

Mean of the sample Mean of the population
 Standard deviation of the sample Standard deviation of the population

Cenário 1* Cohen's d formula

$$Cohen's d = \frac{M - \mu}{\sigma} = \frac{78.15 - 83.54}{16.36} = -0.33$$



Teste de hipóteses usando T-Test (uma amostra)

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 $n-1$ e $n-2$)
- Teste estatísticos são agora 2 amostra

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Means of the two samples Hypothesized difference between the population means (0 if testing for equal means)
 Standard deviation of the two samples Number of elements of the two samples

Cenário 2*

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

$\mu_1 = 78.15$ $s_1 = 7.94$ $n = 13$: A

$\mu_2 = 73.58$ $s_2 = 8.33$ $n = 12$: B

Step 1 -

- $H_0: \mu_1 = \mu_2$ (configurações A e B são iguais em tempo de execução)

- $H_1: \mu_1 > \mu_2$ (configuração B é + rápida que A (tempo de execução é maior na A))

Step 2 -

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{78.15 - 73.58 - 0}{\sqrt{\frac{7.94^2}{13} + \frac{8.33^2}{12}}} = 1.402$$

Step 3 - Dado que as dimensões das amostras são $n = 13$ e $n = 12$, grau de liberdade é o menor $n-1 \Rightarrow 11$

TABLE D t distribution critical values									
df	Upper-tail probability p								
	.25	.20	.15	.10	.05	.025	.01	.005	.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.207	4.781	6.999	9.249
3	0.727	0.941	1.190	1.353	2.131	2.776	3.078	4.604	6.223
4	0.741	0.943	1.139	1.282	1.980	2.447	2.707	3.899	5.591
5	0.718	0.906	1.143	1.240	1.943	2.347	2.612	3.726	5.353
6	0.717	0.906	1.139	1.232	1.935	2.345	2.598	3.699	5.233
7	0.703	0.883	1.100	1.283	1.883	2.262	2.500	3.559	5.144
8	0.699	0.874	1.088	1.263	1.839	2.178	2.398	3.478	5.054
9	0.697	0.874	1.083	1.263	1.839	2.171	2.389	3.399	4.964
10	0.695	0.874	1.085	1.263	1.839	2.171	2.389	3.393	4.874
11	0.695	0.874	1.085	1.263	1.839	2.171	2.389	3.393	4.874

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 H_0 for verdadeira é de 9,42%. Não 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.

$$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}}$$

Se dimensão amostras for grande ($n \geq 30$) - Teste Z 2 amostras para comparar médias de 2 amostras grandes independentes.

$$Z_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Se 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.

$$t_c = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Se tamanho amostra for pequeno ($n < 30$) - Teste T de 2 amostras para comparar médias de duas amostras independentes.

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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 em sites.

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: $p \geq 0,10$ - diferença observada é "não significativa" $0,05 \leq p > 0,10$ - "marginalmente significativa" $0,01 \leq p > 0,05$ - 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 não, 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!

Teste estatística:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Step 2-

Large samples	Small samples
$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$t = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$

Tamanho do efeito:

$$d = \frac{|\hat{p} - p|}{p(1-p)}$$

$$d = \frac{|\hat{p} - p|}{p(1-p)}$$

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: $p \leq 0,75$ (proporção vulnerabilidades detetadas não é superior a 0,75 (0,75 é alvo declarado pelo CTO))

H1: $p > 0,75$ (proporção vulnerabilidades detetadas é superior 0,75 (ferramenta é melhor do que o limite indicado pelo CTO))

Step 2- $z = 1.088$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.7806 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{237}}} = 1.088$$

Step 3- Olhando para tabela Z para $z = 1,088$ (uma cauda), descobrimos que p está entre 0,1401 e 0,1379. Usando calculadora $p = 0,1383 = 13,83\%$

Z	A Area Between Mean and Z	B Area Beyond Z	C Area Beyond Mean
0.71	0.2611	0.2389	
0.72	0.2642	0.2358	
0.73	0.2673	0.2327	

Except from the Z table

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

Medir os intervalos de confiança:

Proporção de vulnerabilidades detetadas - cobertura da ferramenta

$$\hat{p} = \frac{x}{n} = \frac{185}{237} = 0.7806$$

CI para um nível de confiança de 95%

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7806 \pm 1.96 \times \sqrt{\frac{0.7806(1-0.7806)}{237}} = 0.7806 \pm 0.0527$$

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

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

$$\hat{p} = \frac{x}{n} = \frac{38}{223} = 0.1704$$

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

CI para um nível de confiança de 95%

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1704 \pm 1.96 \times \sqrt{\frac{0.1704(1-0.1704)}{223}} = 0.1704 \pm 0.0493$$

Proporção falsos positivos (95% conf.) = $0,1704 \pm 0,0493$

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

Null hypothesis	Alternate hypothesis	Number of tails
P1 = P2	P1 ≠ P2	2
P1 ≥ P2	P1 < P2	1
P1 ≤ P2	P1 > P2	1

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

n2 é o tamanho da amostra 2

$$p^{\wedge} = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2} \quad SE(p^{\wedge}) = \sqrt{p^{\wedge}(1-p^{\wedge}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{p_1 - p_2}{\sqrt{p^{\wedge}(1-p^{\wedge}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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 não 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.

$$E(X_{1i} - X_{2i}) = \text{diferença entre entradas para par de dados} \quad \bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

$n = \text{nº de pares}$

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

$$S_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}}$$

Usar média d das diferenças entre as entradas de dados emparelhadas nas amostras dependentes e o desvio padrão S_d 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 $n \geq 30$ ou por uma distribuição T com $n-1$ graus de liberdade se $n < 30$.

Teste estatístico 2 amostras dependentes

Grandes conjuntos amostras emparelhadas ($n \geq 30$) Z

$$Z = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$$

- d : média diferença entre entradas emparelhadas amostras dependentes

- μ_d : diferença assumida entre médias (geralmente 0)

H_0 é conservadora

- S_d : desvio padrão das diferenças entre entradas de dados emparelhadas nas amostras dependentes

- \sqrt{n} : nº amostras emparelhadas.

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

- \sqrt{n} : grau de liberdade é $n-1$.

Setp 2- (amostras grandes)

(amostras pequenas)

$$\bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

$$t = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$$

$$S_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}}$$

$$S_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}}$$

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 (\tilde{n} o mesmo, mas similar).

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

Developer	1	2	3	4	5	6	7	8	9	10
Score (before)	85	79	70	76	81	78	72	65	78	65
Score (after)	80	85	89	86	92	75	78	60	85	80

Setp 1-

H0: $\mu_d \leq 0$ (pontuações testes após treinamento \tilde{n} são melhores (treinamento \tilde{n} melhorou habilidades)

H1: $\mu_d > 0$ (treinamento melhorou habilidades e resultados testes após treinamento são melhores (Claim)).

Setp 2- adicionar cálculos intermediários à tabela

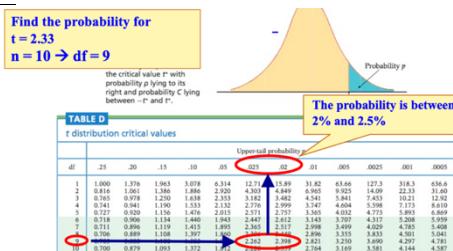
Teste analise dentro quadro

Developer	1	2	3	4	5	6	7	8	9	10
Score (before)	85	79	70	76	81	78	72	65	78	65
Score (after)	80	85	89	86	92	75	78	60	85	80
d	-5	6	19	10	11	-3	6	-5	7	15
d^2	25	36	361	100	121	9	36	25	49	225

$$\bar{d} = \frac{\sum d}{n} = \frac{61}{10} = 6.1 \quad S_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}} = \sqrt{\frac{10(987) - 3721}{10(10-1)}} = \sqrt{68.32} = 8.27$$

$$t = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}} = \frac{6.1 - 0}{8.27 / \sqrt{10}} = \frac{6.1}{2.61} = 2.33$$

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:

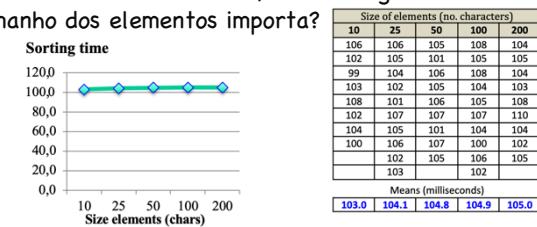
- Algum caso é difícil encontrar mesmos objetos/participantes
- Quando hipótese nula é rejeitada, muitas vezes é difícil argumentar que a diferença se deve a eventos globais e não ao teste-reteste dos mesmos indivíduos.

ANOVA- Análise De Variância

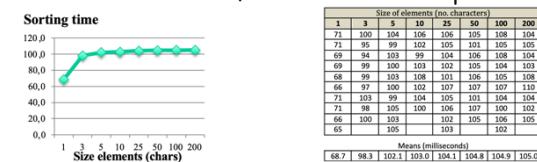
Cenário 2*- \tilde{N} pode utilizar teste T de 2 amostras (pelo menos diretamente)

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 memória e tempo) e uma matriz de 10000 elementos, obteve seguintes resultados: Tamanho 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 dependentes: 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: $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ (médias de todos os grupos são iguais)
H1: nem todas médias são iguais

- \tilde{N} 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 \tilde{N} com discrepâncias graves

Size of elements (no. characters)	10	25	50	100	200
103	108	108	110	108	108
108	107	109	108	112	108
105	106	111	111	108	108
105	106	111	111	108	108

- 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: $H_0: \mu_1 = \mu_2 = \mu_3$.
- 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.

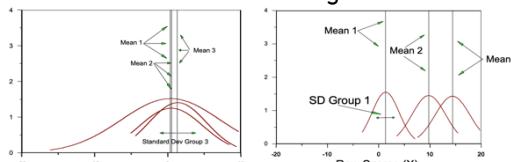
Resultados: 3 amostras da mesma população

Médias dos 3 grupos não 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

Médias amostrais estão distantes umas das outras, então a variância entre as médias será grande



Suponha que façamos estudo e encontremos os seguintes resultados (qualquer gráfico). Como saberíamos se existe um efeito real ou não? Para decidir, podemos comparar a variância observada nas médias com o que esperaríamos obter, se H_0 for verdadeira (não há diferença nas médias).

Definições de termos na ANOVA

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

\bar{X}_G - grande média, assumindo todas as observações.

\bar{X}_A - média de qualquer nível do IV (grupo).

\bar{X}_{A_i} - média de um nível específico (1 neste caso).

X_i - a observação ou dados brutos para i^{th} medição.

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

Soma dos Quadrados (SS) é frequentemente seguida por uma variável entre parênteses, como $SS_{(W)}$ que indica qual soma dos quadrados se refere:

$$SS_{(T)} = \sum (X_i - \bar{X}_G)^2$$

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

$$GM = \frac{\sum n_i \bar{x}_i}{\sum n_i}$$

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

$$SS_{(T)} = \sum (X_i - \bar{X}_G)^2$$

Soma 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.

$$SS_{(W)} = \sum df s^2$$

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

$$SS_{(B)} = \sum n_A (\bar{X}_A - \bar{X}_G)^2$$

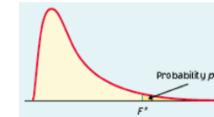
ANOVA: estatística F

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: ($F = \text{teste estatístico}$)

Variation Among Sample Means
Variation Among Individuals In Each Sample

$$F = \frac{SS_{(B)}}{SS_{(W)}} / \frac{(k-1)}{(n-k)}$$



Obtendo o crítico da tabela F

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

Tab. F para $\alpha = 0,05$, procurar valor crítico para $F(3,9)$

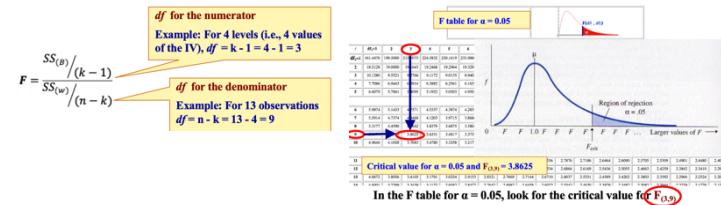


Tabela ANOVA

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

	SS	df	MS	F	P	
Between	$SS_{(B)}$	$k-1$	$\frac{SS_{(B)}}{k-1}$	$\frac{MS_{(B)}}{MS_{(W)}}$	Tail area above F	$SS_{(B)} = \sum n_i (\bar{X}_i - \bar{X}_G)^2$
Within	$SS_{(W)}$	$n-k$	$\frac{SS_{(W)}}{n-k}$			$SS_{(W)} = \sum (X_i - \bar{X}_i)^2$
Total	$SS_{(W)} + SS_{(B)}$	$n-1$				$SS_{(T)} = \sum df s^2$

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.

Size of elements (no. characters)		
3	5	10
100	104	106
95	99	102
94	103	99
99	100	103
99	103	108
97	100	102
103	99	104
98	105	100
100	103	
	105	

Step 1-

$H_0: \mu_3 = \mu_5 = \mu_{10}$ - Todas amostras têm médias iguais

$H_1: \mu_3 \neq \mu_5 \neq \mu_{10}$ - Nem todas médias são iguais

- Não diz como ou quais diferem

- Pode acompanhar múltiplas comparações

Step 2- Determinar as características das amostras em comparação. $n=27$, $k=3$

Cálculo:

	3 chars	5 chars	10 chars
Sample size	9	10	8
Mean	98.3	102.1	103.0
Std. Dev.	2.74	2.38	2.98
Variance	7.50	5.66	8.86

- Média ponderada das médias amostrais individuais

$$GM = \frac{\sum n_i \bar{x}_i}{\sum n_i} = \frac{9(98.3) + 10(102.1) + 8(103.0)}{9 + 10 + 8} = 101.10$$

Variação entre grupos

- Variação entre cada média da amostra e a média geral
- Cada variação do grupo é ponderada pelo tamanho da amostra

$$SS_{(B)} = \sum n_i (\bar{X}_i - \bar{X}_G)^2 = 9(98.3 - 101.10)^2 + 10(102.1 - 101.10)^2 + 8(103.0 - 101.10)^2 = 107.77$$

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

$$SS_{(W)} = \sum df s^2 = 8(2.74)^2 + 9(2.38)^2 + 7(2.98)^2 = 172.90$$

Agora podemos preencher tab. ANOVA unidirecional

	SS	df	MS	F	P
Between	SS(B)	k-1	$\frac{SS(B)}{k-1}$	$\frac{MS(B)}{MS(W)}$	Tail area above F
Within	SS(W)	n-k	$\frac{SS(W)}{n-k}$		
Total	SS(W) + SS(B)	n-1			

Step 3- Valor P para $F(2,24) = 7.48$ é 0,002986 (usando calculadora online)

Step 4- $H_0: \mu_3 = \mu_5 = \mu_{10}$ - Todas amostras têm médias iguais
 $P \approx 0,003$ à H_0 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 semelhantes)
- Independência de erros (amostras aleatórias independentes)

Two-way ANOVA: Hipótese nula

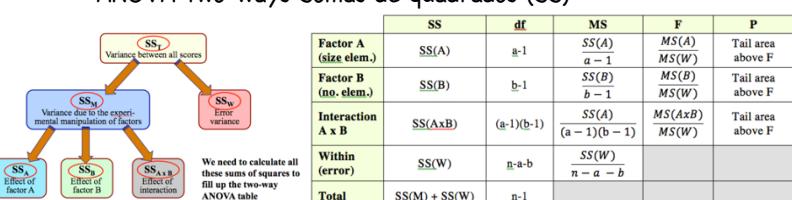
Testa 3 hipóteses simultaneamente:

- Nenhuma diferença nas médias devido ao fator A
 $H_0: \mu_1 = \mu_2 = \dots = \mu_a$.
- Nenhuma diferença nas médias devido ao fator B
 $H_0: \mu_1 = \mu_2 = \dots = \mu_b$
- S/ interação dos fatores A e B - $H_0: AB_{ij} = 0$

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).

ANOVA two-ways Somas de quadrados (SS)



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

- $H_0: \mu_{10} = \mu_{50} = \mu_{100}$ - tamanho elementos a serem classificados não é relevante para tempo de classificação (médias são iguais)

- $H_0: \mu_{10K} = \mu_{50K}$ - nº elementos a serem ordenados não é relevante para tempo ordenação (médias são iguais)

- $H_0: AB_{ij} = 0$ - Não há interação entre tamanho e nº de elementos

- H_1 : Nem todas médias são iguais e há interação (só é possível na cauda direita).

Step 2- Determinar características das amostras em comparação.

Sample size	10 chars		50 chars		100 chars	
	10K	50K	10K	50K	10K	50K
Mean	104.1	308.8	104.8	309.3	104.9	309.6
Std. Dev.	1.96	2.12	2.28	1.75	2.56	2.19
Variance	3.86	4.50	5.19	3.07	6.54	4.78

$$GM = \frac{\sum n_i \bar{x}_i}{\sum n_i} = \frac{9(104.1) + 8(308.8) + 9(104.8) + 8(309.3) + 10(104.9) + 9(309.6)}{9 + 8 + 9 + 8 + 10 + 9} = 201.11$$

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.

Realizar cálculo: $SS(T) = 552982$ (este valor não é necessário nos cálculos)

\bar{X}_G is the GM (Grand Mean), taken over all the observations/scores.

The observation or raw data for the i^{th} measurement/score

SS(M)- soma dos quadrados que dá a variância devida à manipulação experimental de fatores (todos fatores são considerados aqui). Cálcular:

$$SS_{(M)} = \sum n_i (\bar{X}_i - \bar{X}_G)^2$$

Number of samples of level i Average of the scores of level i

$$SS_{(M)} = 9(104.1 - 201.25)^2 + 8(308.8 - 201.25)^2 + 9(104.8 - 201.25)^2 + 8(309.3 - 201.25)^2 + 10(104.9 - 201.25)^2 + 9(309.6 - 201.25)^2 = 552736.78$$

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.

10 chars	
10K	50K
106	313
102	307
103	308
103	309
108	306
102	308
104	310
105	312
108	305

50 chars	
10K	50K
105	308
101	309
106	307
105	307
104	311
105	312
107	310
101	311
107	311
104	308
100	309
106	312
102	305

100 chars	
10K	50K
106	313
102	307
103	309
108	305
105	307
108	307
104	311
105	312
107	308
101	311
104	308
100	309
106	312
102	305

$$SS_{(A)} = \sum n_i (\bar{X}_i - \bar{X}_G)^2$$

$$SS_{(A)} = 17(200.82 - 201.25)^2 + 17(201.08 - 201.25)^2 + 19(201.84 - 201.25)^2 = 10.81$$

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

10K	
106	105
102	101
103	106
103	105
108	106
102	107
104	101
105	107
108	105

50K	
313	308
307	309
308	307
309	307
306	311
310	311
312	308
311	309
313	311

$$SS_{(B)} = \sum n_i (\bar{X}_i - \bar{X}_G)^2$$

$$SS_{(B)} = 28(104.75 - 201.25)^2 + 25(309.32 - 201.25)^2 = 552721.12$$

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

$$SS_{(AxB)} = SS_M - SS_A - SS_B = 552736.78 - 10.81 - 552721.12 = 4.84$$

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
- $$SS_{(w)} = \sum df s^2 = 1.96^2 (9 - 1) + 2.12^2 (8 - 1) + 2.28^2 (9 - 1) + 1.75^2 (8 - 1) + 2.56^2 (10 - 1) + 2.19^2 (9 - 1) = 245.28$$

Sumário Tabela:

	SS	df	MS	F	P
Factor A (size elem.)	10.81	2	5.41	1.06	Tail area above F
Factor B (no. elem.)	552721.12	1	552721.12	108166.81	Tail area above F
Interaction A x B	4.84	2	2.42	0.47	Tail area above F
Within (error)	245.28	48	5.11		
Total	552982.05	52			

Step 3- Encontrar p

	SS	df	MS	F	P
Factor A (size elem.)	10.81	2	5.41	1.06	Tail area above F
Factor B (no. elem.)	552721.12	1	552721.12	108166.81	Tail area above F
Interaction A x B	4.84	2	2.42	0.47	Tail area above F
Within (error)	245.28	48	5.11		
Total	552982.05	52			

Step 4-

H0: $\mu_B.10 = \mu_B.50 = \mu_B.100$ - tamanho elementos a classificar não é relevante para tempo de classificação (médias são iguais)

H0: $\mu_{10K} = \mu_{50K}$. - nº elementos a ordenar não é relevante o tempo de classificação (médias são iguais)

H0: $AB_{ij}=0$ - Não há interação entre tamanho/nº elementos

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

P=36% para fator A, dimensão de elementos a ordenar- H0 é mantido para este fator: tamanho elementos não 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: não existe interações entre tamanho e nº elementos.
H0 é rejeitado

Inferência estatística não paramétrica

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

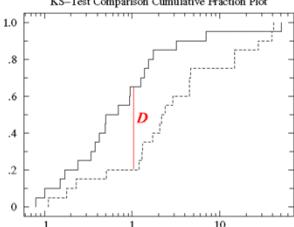
Pós: métodos não paramétricos fazem menos suposições do que os paramétricos; são livres de distribuição

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

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

$$D_{(\max)} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|$$

F1,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 α , então as diferenças entre as duas funções são significativas.



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

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

$$D_\alpha = c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Coeficiente $c(\alpha)$ para os valores típicos de décimos é:

n1	3	4	5	6	7	8	9	10	11	12
1	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*	*	*
8	*	*	*	*	*	*	*	*	*	*
9	*	*	*	*	*	*	*	*	*	*
10	*	*	*	*	*	*	*	*	*	*
11	*	*	*	*	*	*	*	*	*	*
12	*	*	*	*	*	*	*	*	*	*

Project type	Coimbra Branch	Lisbon Branch
Tiny	5	9
Small	8	12
Medium	7	8
Large	4	14
Very large	7	5
Huge	4	8
Total	35	56

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

Step 1- H0: Distribuição1 = Distribuição2 (são semelhantes, taxa de insucesso é mesma nos 2 ramos)

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

Step 2- significância 95% → $\alpha=0,05$. Valor crítico K-S para valores grandes de n1 e n2 é (aproximadamente)

$$D_\alpha = c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36 \sqrt{\frac{35 + 56}{35 \times 56}} = 0.29$$

dada pela fórmula:

1.36 → coeficiente $c(\alpha)$ é $\alpha = 0,05$

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

Step 3- teste estatístico é:

$$D_{(\max)} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|$$

Esta é diferença máxima no valor absoluto (classes do projeto são ordenadas)

Projects type	Coimbra Branch	Lisbon Branch	Cum% Coimbra	Cum% Lisbon	Diff abs. value
Tiny	5	9	0,143	0,161	0,018
Small	8	12	0,371	0,375	0,004
Medium	7	8	0,571	0,518	0,054
Large	4	14	0,686	0,768	0,082
Very large	7	5	0,886	0,857	0,029
Huge	4	8	1,000	1,000	0,000
Total	35	56			

Step 4- Valor crítico: $D_{(\text{crític.})}=0.29$

Teste estatístico: $D_{(\text{máx.})}=0,082$

$D_{(\text{máx.})}$ é inferior $D_{(\text{critic.})}$. Conclusão: não rejeitar H0. Não 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. Também podem ser utilizados testes não paramétricos como K-S e Shapiro-Wilk.

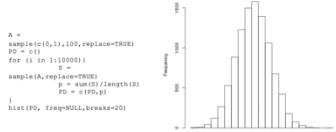
Bootstrapping- estima precisão estatísticas das amostras, desenhando aleatoriamente c/ substituição do conjunto de pontos de dados. É usado para estimar desvio padrão e circuito de computação.

Procedimento: 1) Reamostrar os dados c/ substituição, de modo que reamostra seja igual ao tamanho do conjunto de dados original. 2) Calcular estatística a partir da reamostra da 1ª etapa. 3) Repetir etapas 1 e 2

muitas vezes (1000 a 10000) para obter estimativa + precisa da distribuição.

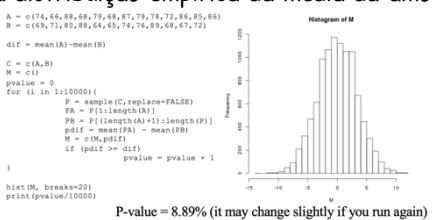
Distribuição bootstrap aproxima-se da distribuição amostral da estatística. Esta estatística pode ser diferente da média (mediana, variância, correlação...); geralmente têm a mesma forma e distribuição que a distribuição de amostragem, mas não estão centradas na estatística (da amostra original). Tb 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. Gerar amostra tamanho 100 (x_1, \dots, x_{100}). 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.



Ensaio de randomização- ensaio estatístico em que a distribuição estatística de ensaio em H_0 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 H_0 . Condição suficiente para a permutabilidade: variáveis são i.i.d.

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 não pareado de 2 amostras).

Classifica conjunto de pontuações n_1 e n_2 do + baixo (classificação 1) ao mais alto. Seja R_1 a soma das classificações da amostra + pequena (a do tamanho n_1)

Teste estatístico:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

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 não 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 não pareado de 2 amostras e menos potente do que os postos sinalizados de Wilcoxon). É aplicado quando não há informações de intervalo. 1) Calcular sinal das diferenças entre cada par e ignorar aqueles que não 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 não 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- percepção 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 não sabem articular aquilo que precisam. O que querem nem sempre é o que precisam. Que precisam pode não estar relacionado c/ computador... observação e análise para descobrir. Sistema é alvo da avaliação: não 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

Analizar 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.

Aspectos 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, não 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, não 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, não 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 não são úteis, por muitas razões..., testadores devem estar interessados (mas não predispostos) no sistema, talvez usar sistema semelhante e estar disponíveis datas previstas para testes, não devem trabalhar na indústria ou para concorrentes

Realizar sessão de utilizador

Explicar que utilizador está ajudar a testar o sistema, não é sistema que está a testar o utilizador. Não há respostas erradas. Se não compreenderem ou tiverem dificuldades,

tudo bem, objetivo é saber sobre isso. Usuário deve dizer o que ele está a tentar fazer e porque, usuários não 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, não 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 comprehensível.

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 não 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 não estruturada (intercâmbio e recolha de informações totalmente livres, utilizado como instrumento exploratório nas fases iniciais estudo, quando investigador ainda não 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.

Tabelas Z e t:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4895	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4993	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997

t distribution critical values

df	Upper-tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.900	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.894	2.365	2.517	2.994	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.794	2.201	2.328	2.718	3.105	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
<i>z</i> [*]	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%