Testes de Significância Estatística para Avaliação de Algoritmos

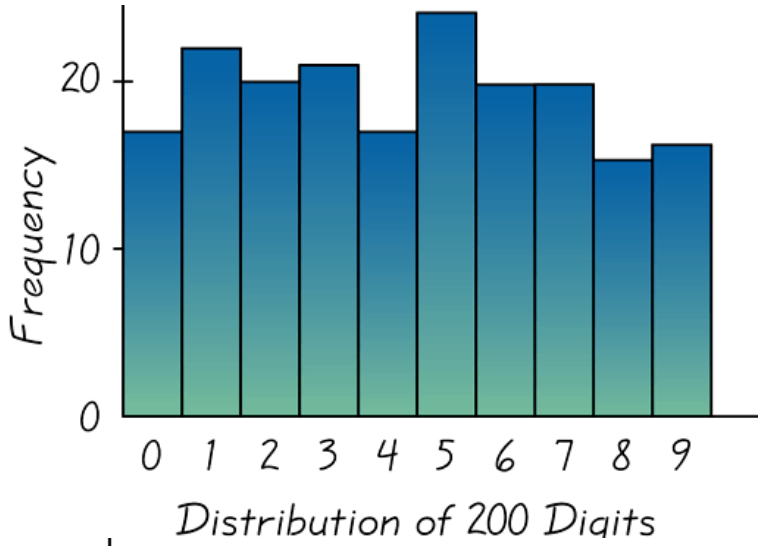
Prof. Eduardo R. Hruschka

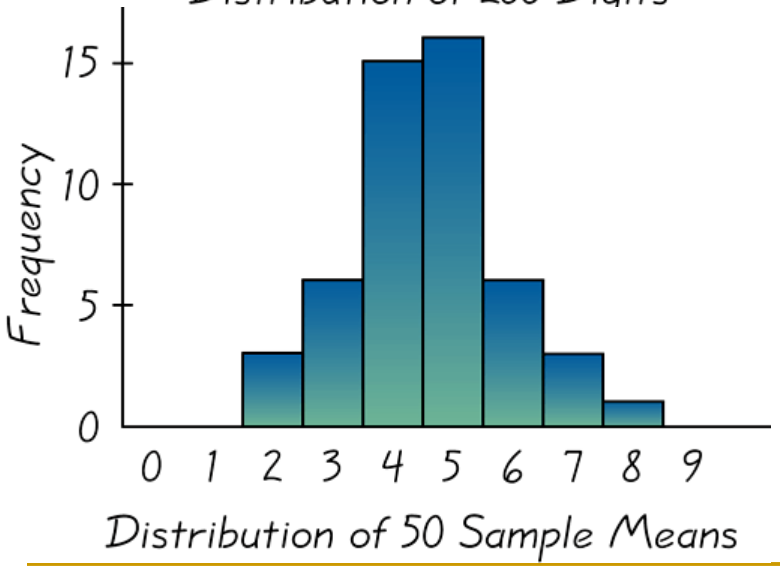
Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)

Agenda

1. Breve revisão sobre testes de significância;
2. Comparando dois algoritmos;

Aquecimento – TCL (intuição):





$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Table 5-6

SSN digits				$\bar{x}$
1	8	6	4	4.75
5	3	3	6	4.25
9	8	8	8	8.25
5	1	2	5	3.25
9	3	3	5	5.00
4	2	6	2	3.50
7	7	1	6	5.25
9	1	5	4	4.75
5	3	3	9	5.00
7	8	4	1	5.00
0	5	6	1	3.00
9	8	2	2	5.25
6	1	5	7	4.75
8	1	3	0	3.00
5	9	6	9	7.25
6	2	3	4	3.75
7	4	0		4.50
	2	8	6	
2	0	9	7	4.50
5	8	9	0	5.50
6	5	4	9	6.00
4	8	7	6	6.25
7	1	2	0	2.50
2	9	5	0	4.00
8	3	2	2	3.75
2	7	1	6	4.00
6	7	7	1	5.25
2	3	3	9	4.25
2	4	7	5	4.50
5	4	3	7	4.75
0	4	3	8	3.75
2	5	8	6	5.25
7	1	3	4	3.75
8	3	7	0	4.50
5	6	6	7	6.00

Triola, M.F., Introdução à Estatística, 1998.

1. Testes de significância (revisão)

→ Inferência estatística: métodos para tirar conclusões a partir de dados.

- Probabilidades expressam a *força* das conclusões;
- Testes de significância se baseiam em distribuições amostrais de estatísticas;
  - Probabilidades para afirmar o que aconteceria se utilizássemos o método de inferência muitas vezes.
- Ter em mente a importância de se realizar *experimentos controlados*.

- Objetivo de um Teste de Significância (TS):
  - Avaliar a evidência oferecida pelos dados em favor de uma afirmação sobre a população.
- Raciocínio subjacente:
  - O que aconteceria se repetíssemos muitas vezes a amostra (experimento)?
- Exemplo 1 (Moore, 2000):
  - Tomates - embalagens de 227g.

Moore, D., A Estatística Básica e Sua Prática, LTC, 2000.

- Supor que uma amostra revele que o peso médio das embalagens é de 225g (< 227g na embalagem);
- Como não poderíamos esperar que todas as embalagens pesariam exatamente 227g:
  - A diferença se deve simplesmente *ao acaso*?

    OU
  - Máquina empacotadora está com problemas?

- Um TS testa uma hipótese específica, usando dados amostrais para decidir sobre sua validade;
- Queremos saber se $\mu = 227$ g.
- Hipótese nula (H_0): é a afirmação sendo testada;
  - Proposição sobre ausência de diferença.
  - TS avaliará a força da evidência contra H_0.

- A hipótese alternativa (H_1) é a afirmação para a qual procuramos evidência. Por exemplo:
  - H_0: $\mu = 227$ g (μ é o peso médio das embalagens).
  - H_1: $\mu \neq 227$ g.
- A evidência que usaremos para decidir entre H_0 e H_1 é a média da nossa amostra de dados.

- *Lógica* de um TS:
  - Assumir H_0 verdadeira (embora ela possa ser falsa);
  - Qual é a probabilidade de obter dados *tão extremos* quanto aqueles de que dispomos se H_0 é verdadeira?
  - Improvável:
    - Tendemos a *duvidar* de H_0 ;
  - Provável:
    - Tendemos a *acreditar* em H_0 .
- O que significa obter dados *tão extremos* quanto aqueles de que dispomos?

- Máquina de empacotamento necessita de revisão?

$$H_0: \mu = 227g \quad X \quad H_1: \mu \neq 227g$$

- Qual é a probabilidade de extrair uma amostra aleatória tal como a nossa se H_0 é verdadeira?

Média amostral = 225g, *n* = 4;

Desvio padrão populacional conhecido = 5g.

- Em termos mais precisos:
  - *Se assumimos* H_0 verdadeira, qual é a distribuição amostral para as médias das amostras de tamanho 4?

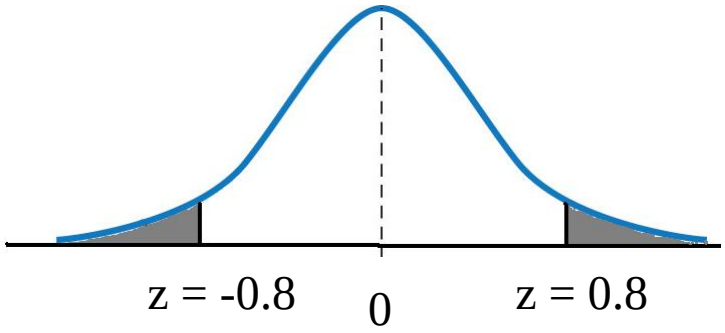
$$\bar{x} \sim N\left(\mu(\bar{x}) = 227, \sigma(\bar{x}) = \frac{5}{\sqrt{4}} = 2.5\right)$$

  - Qual é o *escore z* da média amostral de que dispomos? Este valor é aqui denominado de ***estatística de teste***:

$$z = \frac{\bar{x} - \mu(\bar{x})}{\sigma(\bar{x})} = \frac{225 - 227}{2.5} = -0.8$$

Observação: Estamos assumindo, por enquanto, que *X* possui distribuição normal, mas essa restrição pode ser relaxada ...

- Quais são os outros escores *z* tão extremos quanto o disponível (na *direção* de H_1)?



- Quão provável é obter uma média amostral tão extrema quanto a nossa?
  - *Valor P* do TS.
  - $P = 0{,}4237$.

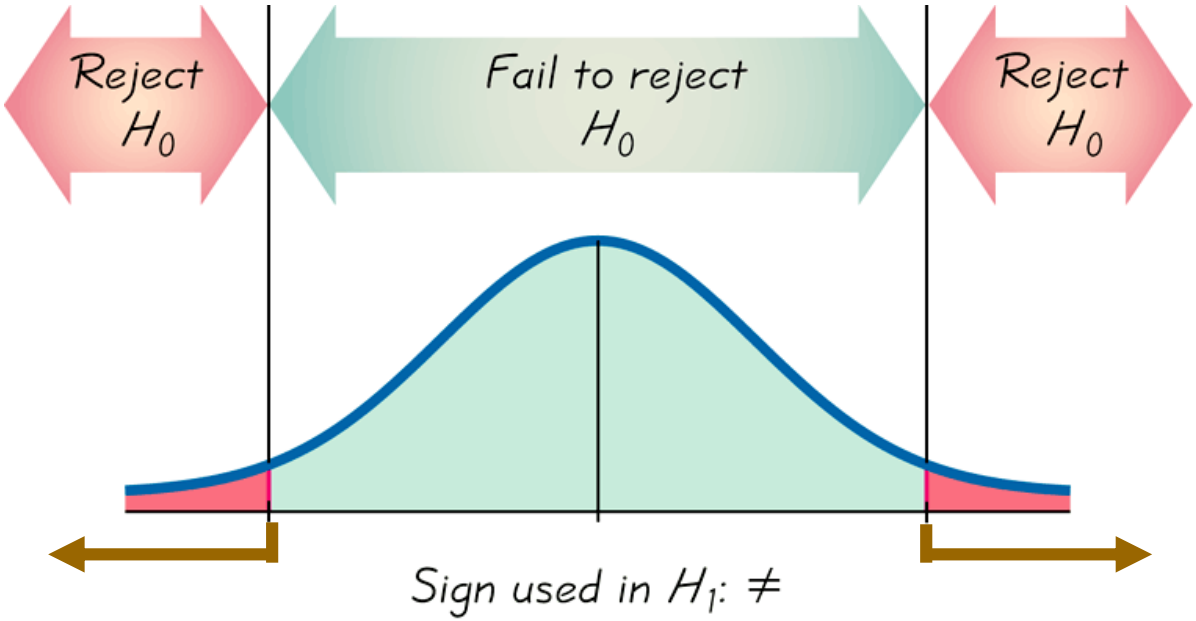
$\rightarrow$ *P* é a probabilidade de que somente uma variação aleatória (do processo de amostragem) é responsável pela diferença observada.

- Um valor pequeno de *P* implica que tal variação aleatória provavelmente NÃO é a única responsável pela diferença observada;
- Rejeitar H_0, i.e., possuímos evidência de que a verdadeira propriedade da população é significativamente diferente de H_0.
- Mas o que pode ser considerado um valor pequeno de *P* ?

- Nível de significância (α) é o maior valor tolerado para *P* a fim de rejeitarmos H_0, refletindo quanta evidência necessitamos contra H_0;
  - Se $P \leq \alpha$: rejeitamos H_0.
  - Se $P > \alpha$: falhamos em rejeitar H_0 .

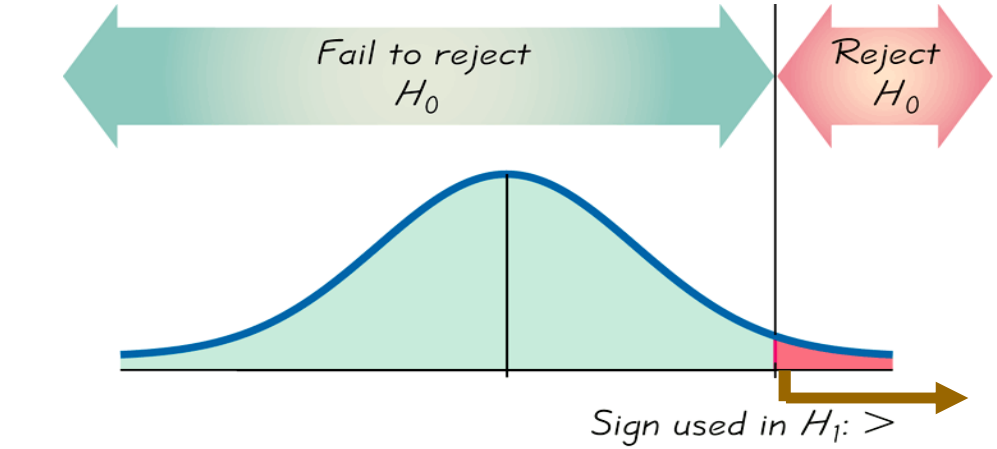
- Valores comuns para α: 10%, 5% e 1%.
  - No nosso exemplo, *P*=42,37%.
  - O que concluímos?

- TS unilaterais (unicaudais) e bilateral (bicaudal):
  - TS bilateral: H_0 contém o sinal de igualdade (=).



Triola, M.F., Introdução à Estatística, 1998.

- TS unilaterais direito e esquerdo:



Diretrizes

- $n \geq 30$: escore z, estimando $\sigma = s$;
- $n < 30$ e histograma essencialmente *não normal*: métodos não paramétricos;
- $n < 30$ e histograma *normal*, com σ conhecido: usar escore z;
- $n < 30$ e histograma *normal*, com σ desconhecido: usar estatística t;

2. Comparando dois algoritmos

- Cenários típicos:
  - Reportar desempenho de um novo algoritmo;
  - Comparar diversos algoritmos num problema particular (e.g., classificação de clientes);

$\rightarrow$ Foco em alguma(s) medida(s) que capturem a capacidade do algoritmo em solucionar o problema.

- Testes usualmente empregados:
  - Teste *t*;
  - Teste de Wilcoxon.
- Para situações específicas de sua pesquisa, estudar literatura de sua respectiva área de atuação...
- Para efeito de ilustração, veremos um exemplo da área de aprendizado de máquina;

- Assumamos que c_i^j é um escore de desempenho do *j*-ésimo algoritmo na *i*-ésima base de dados.

- Para os valores c_i^j obtidos: diferenças de desempenho são estatisticamente significativas?
  - Preferivelmente os algoritmos devem ser rodados nas mesmas amostras: planejar bem os experimentos.

Exemplo 2 (Demsar, 2006):

	Alg. A	Alg. B
Adult	0.763	0.768
Breast	0.599	0.591
Wisconsin	0.954	0.971
Cmc	0.628	0.661
Ionosphere	0.882	0.888
Iris	0.936	0.931
Liver	0.661	0.668
Lung	0.583	0.583
Lymph...	0.775	0.838
Mushroom	1.000	1.000
Tumor	0.940	0.962
Rheum	0.619	0.666
Voting	0.972	0.981
Wine	0.957	0.978

Teste *t* (revisão):

→ Verificar se a diferença média (B-A) é significativamente diferente de zero;

→ n<30, histograma *normal* (?), σ desconhecido.

Problemas:

- Diferenças para as bases de dados são "comensuráveis"?
- Diferenças entre as 2 variáveis aleatórias são *Normais*?
- Médias afetadas por *outliers?*

	Alg. A	Alg. B
Adult	0.763	0.768
Breast	0.599	0.591
Wisconsin	0.954	0.971
Cmc	0.628	0.661
Ionosphere	0.882	0.888
Iris	0.936	0.931
Liver	0.661	0.668
Lung	0.583	0.583
Lymph...	0.775	0.838
Mushroom	1.000	1.000
Tumor	0.940	0.962
Rheum	0.619	0.666
Voting	0.972	0.981
Wine	0.957	0.978

Computando a estatística de teste:

$$t = \frac{\bar{x} - \mu(\bar{x})}{s/\sqrt{n}} = \frac{0{,}0155 - 0{,}0000}{0{,}02/3{,}74} = 2{,}89$$

- *t_crítico* para α=1% (unilateral) é 2,65.
- O que concluímos?

- Teste de Wilcoxon:
  - Baseado nos *ranks* das diferenças;
  - Assume comensurabilidade qualitativa (*ranks / postos*):
    - Grandes diferenças contam mais;
    - Magnitudes das diferenças não são levadas em conta.
  - Não assume distribuições *Normais*;
  - Efeito dos *outliers* é atenuado.

Voltemos ao *Exemplo 2*...

	Alg. A	Alg. B	Diferença	*Rank/Posto*
Adult	0.763	0.768	+0.005	3.5
Breast	0.599	0.591	-0.008	7
Wisconsin	0.954	0.971	+0.017	9
Cmc	0.628	0.661	+0.033	12
Ionosphere	0.882	0.888	+0.006	5
Iris	0.936	0.931	-0.005	3.5
Liver	0.661	0.668	+0.007	6
Lung	0.583	0.583	0.000	1.5
Lymph...	0.775	0.838	+0.063	14
Mushroom	1.000	1.000	0.000	1.5
Tumor	0.940	0.962	+0.022	11
Rheum	0.619	0.666	+0.047	13
Voting	0.972	0.981	+0.009	8
Wine	0.957	0.978	+0.021	10

Analisando intuitivamente o sumário obtido...

	(B – A)	*Rank*
Adult	+0.005	3.5
Breast	-0.008	7
Wisconsin	+0.017	9
Cmc	+0.033	12
Ionosphere	+0.006	5
Iris	-0.005	3.5
Liver	+0.007	6
Lung	0.000	1.5
Lymph...	+0.063	14
Mushroom	0.000	1.5
Tumor	+0.022	11
Rheum	+0.047	13
Voting	+0.009	8
Wine	+0.021	10

Ranks favoráveis ao *Algoritmo B*:
$\mathbf{R}^+ = 3.5 + 9 + 12 + 5 + 6 + 14 + 11 + 13 + 8 + 10 + 1.5 =$ **93**

Ranks favoráveis ao *Algoritmo A*:
$\mathbf{R}^- = 7 + 3.5 + 1.5 =$ **12**.

Dados sugerem que *B* é melhor do que *A*...
→ Há significância estatística?

Estatística de teste:
$T = \min(R^+, R^-)$
→ Valores críticos para *T* disponíveis em tabelas;
→ Neste caso, rejeita-se H_0 para α=5% (T_{max}=17).

	(B – A)	*Rank*
Adult	+0.005	3.5
Breast	-0.008	7
Wisconsin	+0.017	9
Cmc	+0.033	12
Ionosphere	+0.006	5
Iris	-0.005	3.5
Liver	+0.007	6
Lung	0.000	1.5
Lymph...	+0.063	14
Mushroom	0.000	1.5
Tumor	+0.022	11
Rheum	+0.047	13
Voting	+0.009	8
Wine	+0.021	10

Observações finais:

- Muito debate sobre o uso de TS;
- Vimos duas abordagens razoavelmente bem aceitas;
- Resultados de TS devem ser usados com cuidado e sobriedade;
- TS proporcionam uma certa evidência da validade sobre os resultados obtidos;
- Significância estatística X significância prática?