CI1027 – Introdução à Pesquisa em Ciência da Computação

Comparação de Experimentos

Alexandre Valadão Delazeri ⟨alexandre.delazeri@ufpr.br⟩

David Menotti Gomes ⟨web.inf.ufpr.br/menott⟩

Departamento de Informática (DInf) / Universidade Federal do Paraná (UFPR)

11 de maio de 2026

Outline

1. Motivação
2. O experimento como amostra
3. O intervalo de confiança
4. O tamanho de efeito
5. Comparando resultados
6. Em suma

Uma simples comparação

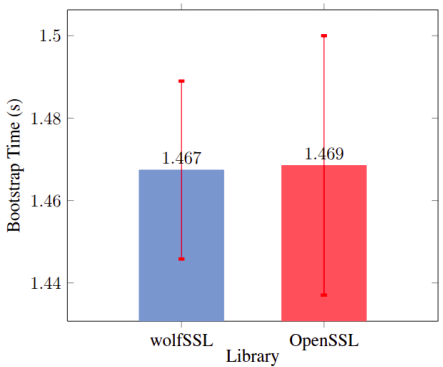


Figura: Tempo médio de execução de dois algoritmos com barra representando a variância. Imagem de [Ehlert et al., 2025].

Uma simples comparação

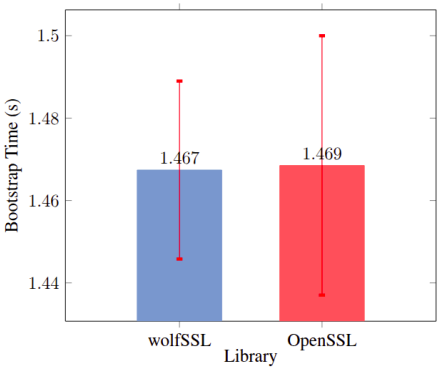


Figura: Tempo médio de execução de dois algoritmos com barra representando a variância. Imagem de [Ehlert et al., 2025].

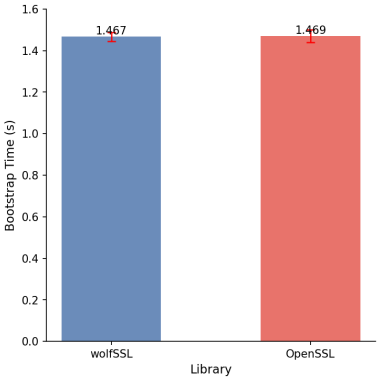


Figura: Mesmo gráfico sem truncagem no eixo, imagem própria.

Objetivo

Análise de resultados

- Fiz um experimento, obtive um resultado. Mas e agora?
- Como comparar resultados diferentes?
- Como comparar meu resultado com resultados de outros pesquisadores?

Dados dois resultados A e B, fundamentalmente existem 3 possibilidades:

- $A = B$
- $A > B$
- $B > A$

- Como distinguir?

Sua métrica não é um valor

- Obtivemos um resultado?
- "Ocorrências irrepetíveis não têm valor científico." – Popper, Karl. 1935.
- Mas se repetirmos o experimento, o resultado será o mesmo?
- O seu resultado é uma **Variável Aleatória**.

A variável aleatória

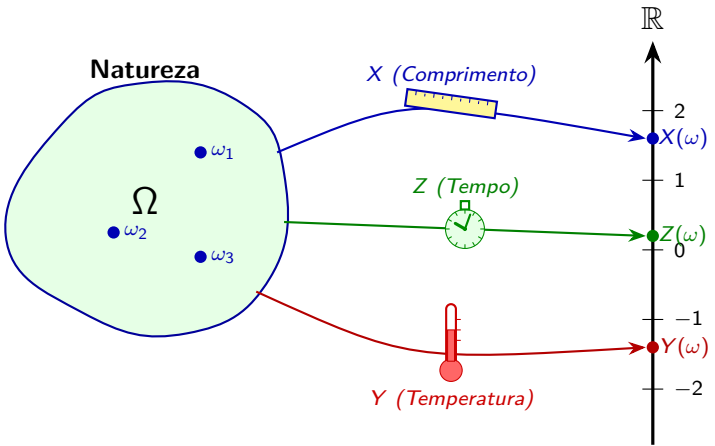


Figura: O conceito de variável aleatória.

População vs. Amostra

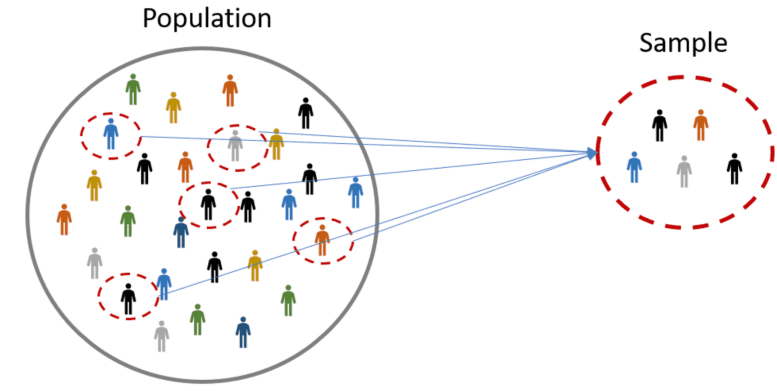


Figura: População e Amostra. Fonte: Omniconvert

O conceito de experimento

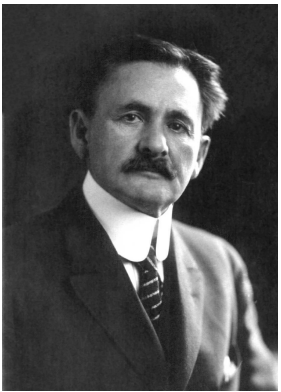
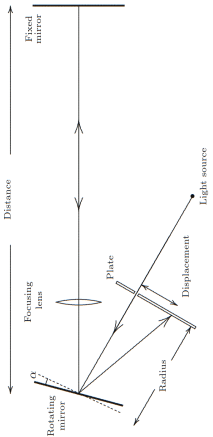


Figura: Albert A. Michelson



O conceito de experimento começou a ser formalizado no fim do século 19.

- Cada medida é entendida como realização de uma variável aleatória.
- Cada medição deve ter a mesma distribuição.
- Obtidas sob mesmas condições.

Figura: esquemática do experimento, fonte: [Dekking et al.]. Foram realizadas 100 medidas, com valor máximo 300,070 km/s, mínimo 299,620 km/s, média 299,852.4 km/s e desvio padrão 28 km/s.

Independentes e Identicamente Distribuídos?

Supomos que:

- Independentes: Cada processo de medida não interfere nos outros.
- Identicamente distribuídos: As condições são idênticas e as medições todas seguem a mesma distribuição de probabilidade.
- (i.i.d.)

O que pode dar errado?

Independência

- Ex: dependência entre medições (ex: *caching*, *throttling*, *carry-over effects*).
- *Benchmarks* em CPU com *thermal throttling*.
- Experimentos sequenciais onde o resultado anterior influencia o próximo.

Distribuição Idêntica

- Ex: *drift* de população.
- Realizamos uma sequência de medidas sem perceber que o processo muda com o tempo.
- Pense em um servidor com degradação de performance devido a *uptime* alto.

Ok, mas e agora?

- Suponha então que nossas medidas são i.i.d.
- Ainda assim, nosso experimento é apenas uma amostra.
- Se repetirmos o experimento, obteremos um novo resultado.
- E se ao invés de estimar um valor pontual pudéssemos determinar um intervalo de valores plausíveis?

Intervalo de confiança

- O intervalo de confiança fornece uma faixa de valores plausíveis para o valor verdadeiro, com base em nossos dados.

Interpretação gráfica

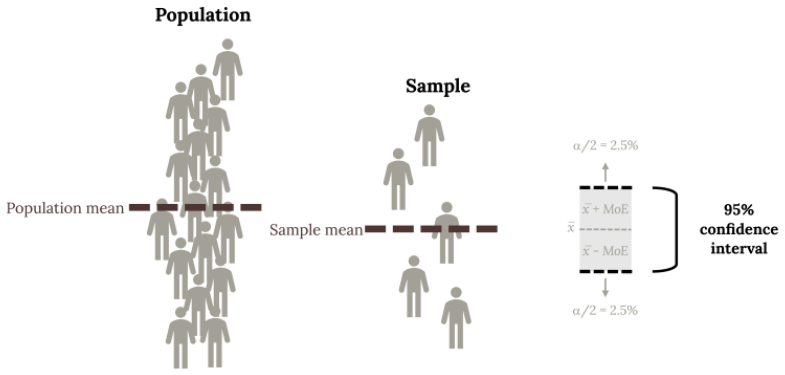


Figura: Visualização de um intervalo de confiança. Imagem adaptada de [Russell et al]

Calculando um intervalo de confiança

Um exemplo, intervalo de confiança de 95%

- Em 20 amostras: $\bar{x} = 1.467$, $S = 0.0216$
- A teoria diz que Margem de Erro $\propto$ SE.
- Casos simples: $\text{SE} = \frac{\sigma}{\sqrt{N}}$. Diferenças, quocientes $\mapsto$ expressões mais complicadas.
- Para confiança de 95% o *score* é 2.09.
- Existe a fórmula $\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{N}}$.
- Substituindo: $1.467 \pm 2.09 \cdot 0.0216/\sqrt{20} = 1.467 \pm 0.0101 = [1.457, 1.477]$

Atenção:

- Todo processo de estimação introduz hipóteses sobre os dados.
- Mesmo em casos em que não conhecemos a distribuição exata, ainda assim podemos estimar o intervalo (ex.: outras aproximações, simulação, bootstrap).

Um intervalo de confiança para nossos dados

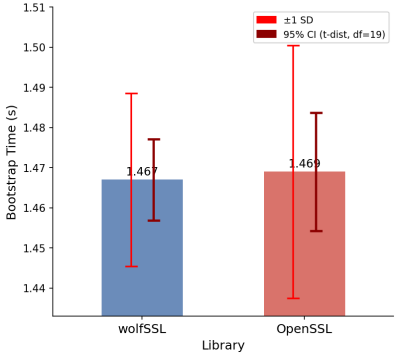


Figura: Tempo médio de execução de dois algoritmos agora com intervalo de confiança.

- O mesmo gráfico de antes, agora com intervalo de confiança de 95%.
- Observe que o intervalo de confiança é muito menor do que o desvio padrão reportado no artigo.
- Intuitivamente, há evidências de que valores plausíveis para as médias estão dentro dos intervalos $[1.457, 1.476]$ e $[1.455, 1.483]$, respectivamente.

Como interpretar o intervalo de confiança

Mas o que $X\%$ de confiança significa?

- Isso significa que o intervalo foi construído por um método que, se repetido muitas vezes, conteria o valor verdadeiro em aproximadamente 95% dos casos.
- **Importante:** isso não significa que haja 95% de probabilidade de o valor verdadeiro estar nesse intervalo. A média é um valor específico fixo. É a média da amostra que depende da amostra.

Regra prática #1:

Sempre que viável, reporte um intervalo de confiança

- Um único valor, sem estimativa de incerteza, não é muito significativo.

Tamanho de efeito:

Significância

Mesmo se algo ocorre, é algo importante?

Pense na comparação de implementações do processo de autenticação:

- Diferença observada: 2 milissegundos. Isso importa?
- Talvez
  - Servidor: 100000 autenticações/dia $\mapsto$ 3.3 minutos economizados.
  - Dispositivo IoT: 1 autenticação/dia $\mapsto$ irrelevante.
- Vale a pena investigar?

Aspectos práticos

O tamanho de efeito como medida de importância

- O tamanho de efeito tenta medir o impacto do resultado.
- Existem medidas livres de escala para tamanho de efeito (como *Cohen's d* e *Hedges' g*).
- Cada uma se adapta a diferentes cenários e hipóteses sobre os dados.
- Quanto menor o tamanho do efeito, maior a quantidade de dados necessários para detectá-lo.
- Quanto mais dados, mais provável que efeitos irrisórios sejam detectados como significantes no teste.
- Em geral, se o efeito é muito menor que o desvio padrão é necessário cautela.

Como reportar o tamanho de efeito

Exemplos de métricas

- Informais:
  - Diferença: $B - A$ (útil principalmente quando a diferença é grande).
  - Diferença relativa: $\frac{B-A}{A}$ (bom para comparar).
  - Proporção: $\frac{\text{Métrica Nova}}{\text{Métrica Antiga}}$ (comum para métricas de performance).
- Métricas padronizadas:
  - d de Cohen é a diferença em unidades de desvio padrão: $d = \frac{A-B}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}$.
  - Existem muitas outras, adaptadas para diversas situações.

O tamanho de efeito e o tamanho amostral

d	Tamanho amostral mínimo
0.01	~ 156977
0.074 (wolfSSL)	~ 1570
0.2 (pequeno)	~ 394
0.5 (médio)	~ 64
0.8 (grande)	~ 26

Tabela: Tamanho amostral mínimo para valores específicos do d de Cohen com poder 80% e 5% de significância. Um d de 0.074 representa uma diferença entre medias de aproximadamente $1/13$ do desvio padrão.

Regra prática #2:

Sempre que viável, reporte tamanhos de efeito

- Reporte o tamanho de efeito com intervalo de confiança.
- Compare o efeito observado com o efeito esperado definido antes do estudo.
- Discuta o impacto prático do efeito no contexto do problema.

Essa diferença é real?

- Observamos duas medidas diferentes, A e B.
- Observamos que talvez exista uma diferença entre elas que vale a pena estudar.
- Fundamentalmente existem duas possíveis razões para isso:
  - Existe uma diferença real entre as duas populações.
  - Variação natural da medida devido ao processo amostral.
- Como distinguir entre esses casos?
- A inferência estatística é uma ferramenta que visa ajudar a responder essa pergunta.

Hipóteses

A hipótese H_0

- Assuma, por hipótese, que não há diferença entre as amostras. Queremos saber o quão surpreendentes são os nossos dados.
- Intuitivamente esse é o pior caso. Se quero mostrar que algo existe, parta do pressuposto que não existe.

A hipótese alternativa H_a

- O efeito que acreditamos que exista.

Erros

Tipos de erros		**Hipótese nula (H_0) é**	
		Verdadeira	**Falsa**
Minha decisão sobre H_0	**Não rejeita**	Inferência correta (Probabilidade $= 1 - \alpha$)	Erro de Tipo II (Probabilidade $= \beta$)
	Rejeitar	Erro de Tipo I (Probabilidade $= \alpha$)	Inferência correta (Probabilidade $= 1 - \beta$)

Figura: Adaptado de [Wikipédia].

Um mundo “ideal”

Etapas

- Enunciar uma hipótese
- Planejar o estudo
  - Determinar tamanho de efeito esperado
  - Determinar estatística de teste
  - Determinar nível de significância
  - Determinar tamanho de amostra
- Coletar dados
- Realize o teste especificado
- Computar efeito observado

O que é o p-valor?

O p-valor é uma **medida de compatibilidade** entre nossos dados **e a hipótese nula**

- Dados pouco compatíveis com a hipótese H_0 fornecem evidência de que ela possivelmente seja falsa.

Definição

- Probabilidade de que, dado que a hipótese H_0 seja verdadeira, sejam observados dados tão (ou mais) extremos quanto os nossos.
- Em símbolos (ex. para diferença de médias):

$$p = \mathbb{P}\Big(|\bar{X}_1 - \bar{X}_2| \geq |\bar{x}_1 - \bar{x}_2| \mid H_0\Big)$$

O que significa ter “evidencia estatística”

A evidência é um gradiente, não uma linha de chegada

- “Evidência” é o grau de surpresa ao observar nossos dados, assumindo que a hipótese H_0 seja a realidade.

- Em um mundo onde H_0 seja verdadeira:
  - **p-valor alto ($p > 0.10$):** Os dados não são surpreendentes.
  - **p-valor pequeno ($p < 0.05$):** Evidências de que os dados são incomuns.
  - **p-valor minúsculo ($p < 0.001$):** Os dados seriam extremamente bizarros.

Exemplo de cálculo de p-valor

- Suponha que tenho alguma razão para desconfiar que uma dada moeda é viciada.
- Tomo H_0: a moeda é justa, H_a a moeda tem uma chance maior de resultar em caras.
- Meu experimento: Jogo a moeda 30 vezes e conto o número de caras, C.
- Digamos que observei 20 caras, isto é 2/3s dos lançamentos resultaram em caras. $C_{obs} = 20$.
- Qual o p-valor?
- $p = \mathbb{P}(C \geq C_{obs} \mid H_0) = \mathbb{P}(C \geq 20 \mid H_0) = \sum_{k=20}^{30} \binom{30}{k} (0.5)^{30} \approx 0.0494$.
- Significante?

Interpretando o exemplo

- Possivelmente.
- Por um lado, 2/3s dos lançamentos serem iguais parece inusitado.
- Por outro lado, isso ocorre aproximadamente 1 em cada 21 vezes.
- No fim, vivemos neste mundo, onde coisas inusitadas podem acontecer de vez em quando, não no "mundo médio" onde apenas coisas tipicas ocorrem.
- Mas o teste pode servir de razão para buscarmos mais evidencia.

Regra Pratica #3

Guarde os dados tão completos quanto possível

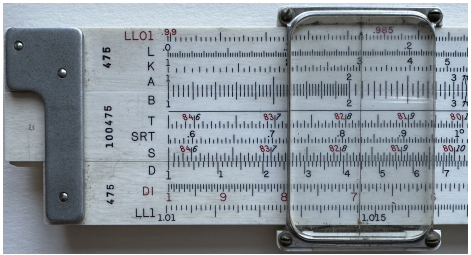
- Não reporte apenas as medidas de interesse.
- Guarde os dados brutos e a ordem em que eles foram observados.
- Essas informações podem ser úteis em análises futuras.
- Sempre que possível reporte o *dataset* coletado e não apenas as estatísticas de resumo.

Por que 0.05?

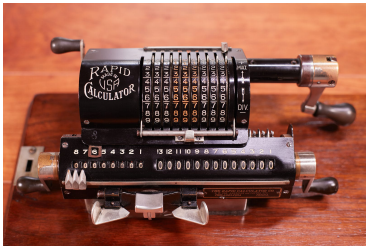
O valor para o qual $P = 0.05$, ou 1 em 20, é 1.96 ou quase 2; é conveniente então tomar esse valor como o limite para julgar se um desvio deve ser considerado significativo ou não. Desvios excedendo duas vezes o desvio padrão são então formalmente considerados como significantes. Usando esse critério, devemos atribuir significância a um resultado negativo apenas uma vez a cada 22 testes.
— Fisher. R, 1925, Statistical Methods for Research Workers[1]

[1] Tradução própria.

A conveniência de Fisher



(a) Régua de cálculo. Fonte [Wikipedia].



(b) Calculadora de 1925. Fonte [Zalewski]

Figura: O estado da arte das ferramentas para fazer contas em 1925.

- Note então que $p = 0.05$ começou por conveniência, em particular conveniência de cálculo. Em 1925 as ferramentas disponíveis para calculo eram essas ou mesmo ainda mais rudimentares.

O que **NÃO** é o p-valor?

O p-valor **NÃO** é:

- Probabilidade que H_0 é verdadeira
- Probabilidade de que o resultado seja devido ao acaso
- Proporcional à magnitude do efeito
- Medida de importância do resultado
- Evidência direta em favor da hipótese alternativa
- Resposta definitiva se sim ou se não

Recomendações práticas

1. **Realize repetições suficientes.** Um teste, ou mesmo alguns poucos, não são dados.
2. **Reporte intervalos de confiança.** Um valor, sem estimativa de incerteza, é pouco informativo.
3. **Reporte a variância**, não apenas as médias.
4. **Reporte o tamanho do efeito**, não apenas o p-valor.
5. **Decida o que testar antes de realizar o experimento.**
   Não ajuste a análise só para obter significância depois de ver os resultados.
6. **Verifique suas hipóteses.**
   Hipóteses são aproximações da realidade, quanto menos próximas, maior o erro.

Leitura Complementar

- M. N. Magalhães e A. C. P. de Lima, *Noções de probabilidade e estatística*.
- F.M. Dekking et al., *A Modern Introduction to Probability and Statistics*
- G. E. Noether, *Introduction to Statistics: The Nonparametric Way*
- B. Efron, *An Introduction to the Bootstrap*
- The American Statistician, Volume 73, Issue sup1 (2019), *Statistical Inference in the 21st Century: A World Beyond $p < 0.05$*