# PPT 2 / T / MEI

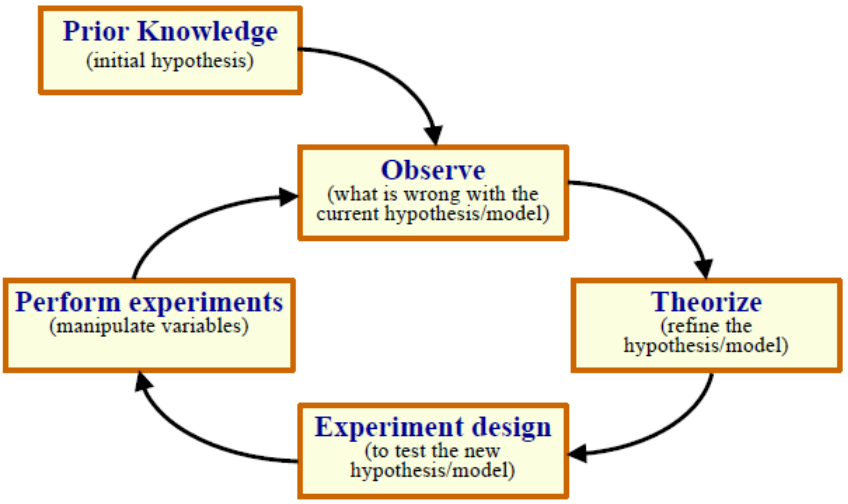
| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 2 |

## Experiências em informática:

> 💡 A ideia fundamental em MEI é confrontar a realidade com a experiência fundamentada com dados.

Os engenheiros querem afinar sistemas, otimizar, verificar, requisitos → testando um produto/software, e validar /verificar mecanismos e/ou soluções, medir e avaliar eficiência de mecanismos de sistemas; acessar a eficácia de processos em software development, processes, etc.

### A metodologia cientifica é:




- Utilizar conhecimento prévio e observações sobre um certo fenómeno
- Formular hipóteses
- Construir um modelo do fenómeno
- Usar uma hipótese para prever os resultados
- Testar as hipóteses usando experiências
- Analisar os resultados da experiência
- Voltar ao principio, se necessário e redefinir a experiência

## Propriedades Chave:

- **Relevância**: Será que vale mesmo a pena, é de relevante responder ao problema em questão tendo em conta o contexto.
- **Representatividade**: será que isto representa o que vai acontecer na realidade? Será que a experiência é só valida para o meu laboratório? *Queremos por indução provar que X em qualquer caso irá funcionar como X.*

- **Repetividade**: as experiêcnias e os seus resultados não podem só ser repetíveis uma vez, tal como os seus resultados.
- **Reprodutível**: outra identidade tem de conseguir repetir a experiência.
- **Custo**

## Tipos de Experiências:

As seguintes irão ser as mais importantes/abordadas durante o curso:

- **Controlled Experiments:** investigation of a testable hypotesis, in which conditions are set up to isolate the variables of interest (”independent variables”) and test how they affect certain measurable outcomes.
  - **São utilizadas para:** análise quantitativa de benefícios de certa ferramenta; estabelecer relações de causa-efeito
  - **Não funcionam quando:** é difícil aplicar se não conseguimos simular as condições certas no laboratório; confiança limitada no que toca à simulação vs cenário real; ignora certos fatores externos.
- **Case Studies:** technique for detailed exploratory investigations, that attempt to understand and explain a phenomenon or test theories.
  - **São utilizadas para:** respostas detalhadas do *porquê.* Ganhar *deep insights* de causa-efeito. Testar teorias em situações complexas, onde temos pouco controlo da variável.
  - **Não funcionam porque:** difícil encontrar casos de estudo. Difícil de quantificar os resultados.
- **Pilot Studies:** introdução controlada de uma ferramenta/técnica num projeto real onde o investigador não consegue controlar o contexto mas consegue visualizar o efeito contra uma baseline ou experiência prévia.
  - **São utilizadas para:** medir os benefícios num contexto real
  - **Não funcionam porque:** difícil a sua execução num contexto real empresarial, de uma ideia que ainda não foi provada.
- **Benchmarks:** um teste de comparação entre diferentes alternativas de ferramentas ou técnicas.

As menos importantes são:

- **Field Studies:** exploratory study used where little is currently known about a problem or where we wish to check that our research goals are grounded in real-life settings
- **Simulations:** An executable model of the software development process, developed from detailed data collected from past projects, used to test the effect of process innovations
- **Surveys:** A comprehensive system for collecting information to describe, compare or explain knowledge, attitudes and behavior over large populations

## Como desenhar uma experiencia controlada:

1. **Problema.** Precisamos de o definir de forma concisa e científica
2. **Definição de variáveis**
   a. Dependentes, Independentes e os níveis das mesmas.
   b. Podemos alterar um fator de cada vez ou *full factorial* (mudar mais que um fator ao mesmo tempo).
   c. Conceito de Baseline e repetition of golden run.
3. **Hipótese** → possível resposta ao ponto 1.
4. **Definir o setup/cenário**
5. **Desenvolver ferramentas/software** que por vezes é necessário para a obtenção de dados
6. **Correr as experiências**
7. **Fazer o data analysis** e **testar as hipóteses** consoante esses dados analisados.

8. **Ver a conclusão e analisar**. Se nenhuma das hipóteses se verificar, voltar ao ponto 1.

> 💡 O problema não pode ser ultra focado, nem ultra geral. Para formular um bom problema temos de ter algum *background* na àrea em que a experiência vai ser feita.

***Nota:*** *o aspeto mais importante é separar as variáveis dependentes e independentes, e classificar os seus níveis:*

- **Dependente**: output que iremos medir *(response time, throughput, no. bugs).*
- **Independente**: variáveis de input que podem ser alteráveis durante o decorrer da experiência*(memory size, clock rate, channel bandwidth, etc. etc. )* → o que se quer testar

Os níveis, referem-se aos valores que foram utilizados. Vamos fazer um nível com pouca potência, um nível superior, com mais alguma potencia e um nível com muita potencia. Assim temos níveis de potencia diferente pelo qual vamos ter resultados diferentes.

> 💡 O importante é saber qual os diferentes níveis que devemos aplicar, neste caso de potência a ser utilizado, de forma a podermos ter resultados que sejam relevantes e significativos para a experiência.

> *Normalmente deve-se só mudar um valor ao mesmo tempo. Mas mesmo assim existe o* ***full factorial,*** *em que se altera todos os valores e consegue-se obter resultados na mesma.*

## Terminologia:

- **Baseline - golden run:** *set of factor values (independente variables) that represent a baseline*
- **Repetition of golden runs:** repetição
- **Randomization:** tenta mudar aleatoriamente as variáveis da experiência. Isto para tentar minimizar fatores externos. *"Vou-te dar um codigo ao acaso, mas não sempre codigos dificeis de forma a influenciar o tempo que o demoras a resolver."*

## Hipóteses:

As hipóteses podem ser:

- **Direcionais**: o valor de uma variável sobe enquanto que a outra desce.
- **Não direcional:** nao tem nada a ver uma com a outra
- *H0* e *H1*: *H0* → o meu software é pior que X. Ao combater esta ficamos com a *H1* → o meu software é melhor que X.

> → isto é direcional, no sentido em que a hipótese é formulada de forma a ver a relação entre duas variáveis.

# Exercício:

Exemplo: experiência sobre a relação do número de bugs com as horas de sono do programador. Os programadores são diferentes tais como a complexidade do codigo a desenvolver, as linguagens, etc.

**a) Variáveis dependentes:** Número médio de Bugs.

**b) Variáveis independentes** (por ordem de impacto): horas de sono, (2h, 6,h, 10h), complexidade do programa*** (neste caso usamos, baixa, média, ou alta) feito, experiência do programador (begginer, medium, senior), linguagem de programação, duração da tarefa*** (curta, média, alta)

** Métricas de Mcabbe, ou complexidade ciclomática (por gráfos de caminho independente).*

- Vg0-4 → muito simples; Vg5-10 → médio; Vg11- 15 → difícil.
- *** → preciso de ter cuidado com variáveis que se relacionam. Quando acontece este caso queimamos uma.

**c)**

- **Hipótese H0:** o numero de horas de sono não influencia o número de bugs.
- **H1**: o número de horas de sono influencia o número de bugs
  (não direcionais)

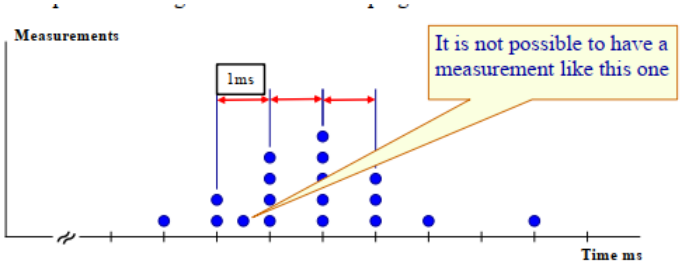
# PPT 3 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 3 |

## Medidas e Intervalos de confiança

### Resolução:

Refere-se à resolução do instrumento de medida → **smallest difference between measurements provided by a measuring device.**




Todos os instrumentos de medida também se caracterizam por uma resolução, por uma **precisão** e uma **exatidão**.

- **Precisão**: as medidas sucessivas terem resultados similares ou não muito distantes
- **Exatidão**: as medidas serem similares à medida real.

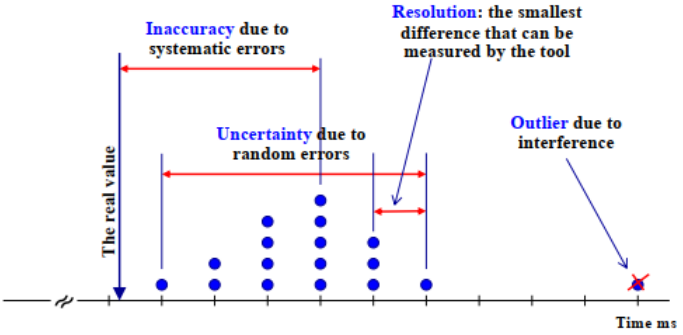
### Incerteza:

Refere-se à incerteza da medida. Se medirmos a mesma coisa duas vezes não será similar → isto refere-se à **incerteza de medição.** Há dois tipos:

- **Aleatórias:** variações na medida que ocorre sem um padrão previsível.
  - Ocorrem sem um padrão
  - Podem ser reduzidas mas nunca eliminadas
  - Devem ser analisadas estatisticamente
- **Sistemáticas:** variações na medida que fazem com que a medida seja maior ou menor sistematicamente.
  - Desvios sistemáticos do valor real
  - Acontecem graças a diversos motivos → instrumento de medida mal calibrado, delay, etc.
  - Quando identificado, pode ser eliminado totalmente.
- **Casos especiais:** este tipo de incerteza pode-se dever a certos casos específicos da experiência como p.e.: **warm-up, ramp-up, hysteresis** → o outcome depender de medidas prévias.

### Variabilidade:

- Limite da precisão do instrumento: apesar do *setup* ser igual para a experiência, o ambiente pode ligeiramente mudar.
- Mudanças no ambiente da experiência: pequenas mudanças podem ocorrer durante a experiência → temperatura, bandwidth, cache state, etc.
  - Em certos casos extremos, pode levar à ocorrência de *outliers.*



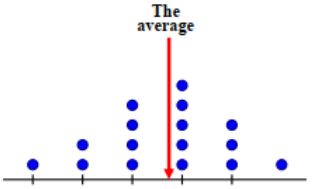

*Os outliers e as incertezas devem ser sempre analisadas estatisticamente e reportadas.*

## Intervalos de Confiança:

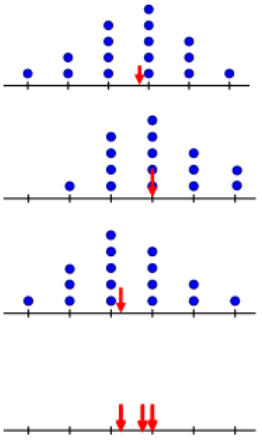
Quando medimos a mesma coisa, diversas vezes, devemos calcular intervalos de confiança.

- O que pretendemos com isto é dizer qual é a percentagem de que eu se fizer duas medidas seguidas, a segunda tem de estar 0,% dentro dessa medida anterior.

- Assumimos que um set de medidas têm uma distribuição normal.
- Assim este set de medidas têm uma média, que é uma estimativa do valor real.
- Se repetirmos isto com diferentes medidas, iremos ter uma média ligeiramente diferente.

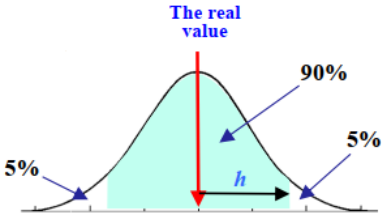



- Múltiplas sets de medidas, e o cálculo das suas médias, levam à formação de múltiplas amostras da distribuição das médias.
- A distribuição das médias é *narrower* do que a distribuição base. O que leva a que haja uma estimativa mais exata.



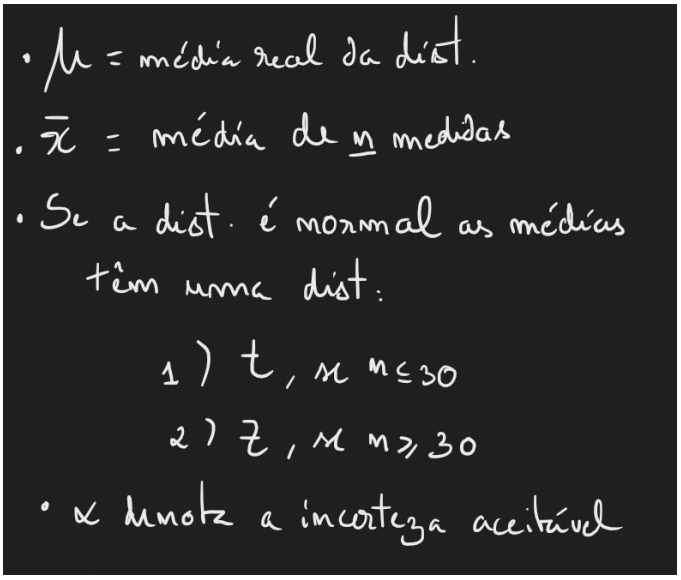
**Assim supomos que:**

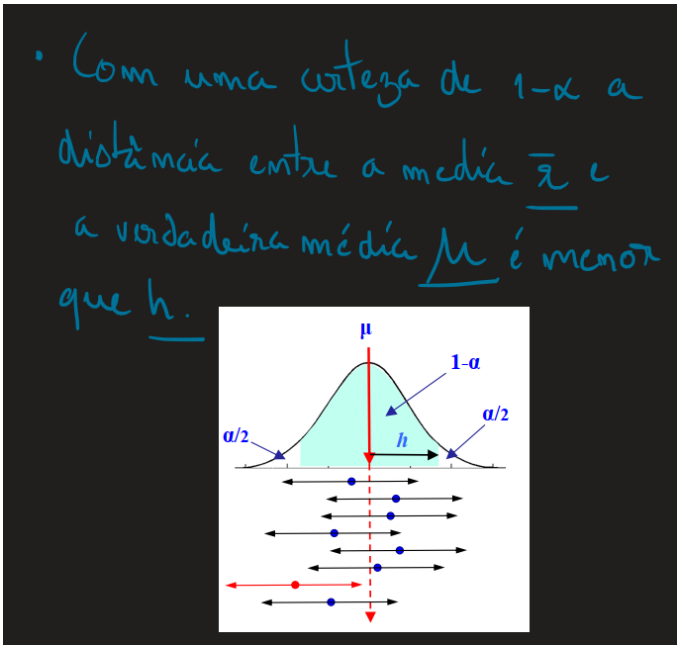
- A média reflete um valor real, mais algum erro/ruído.
- Logo, as médias são **distribuídas** à volta do valor real.
- Dada a distribuição, conseguimos encontrar o range (*h*) em que é esperado que contenha 90% das médias → *90% é só um exemplo.*




Para 90% das médias, o valor real está contido entre *h*.

**Calcular os intervalos de confiança:**



Assim assumimos que:

- $\bar{x} \pm t \times s/\sqrt{n}$, para $n \leq 30$

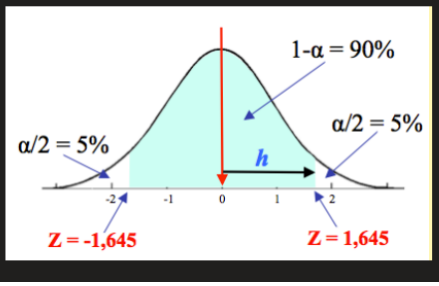
ou

- $\bar{x} \pm z \times s/\sqrt{n}$, para $n \geq 30$

onde:

$s \rightarrow$ standard deviation of the $n$ samples

Ex:



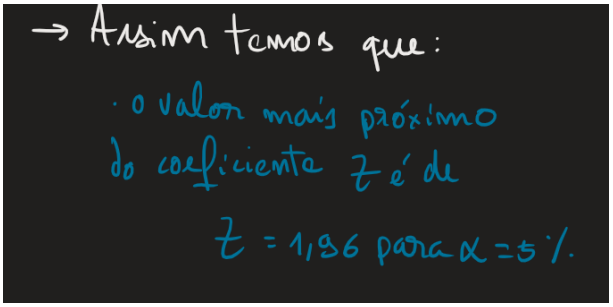

**Exemplo prático:**

**Example: what is the confidence coefficient Z for α = 5%? (two-tailed test)**

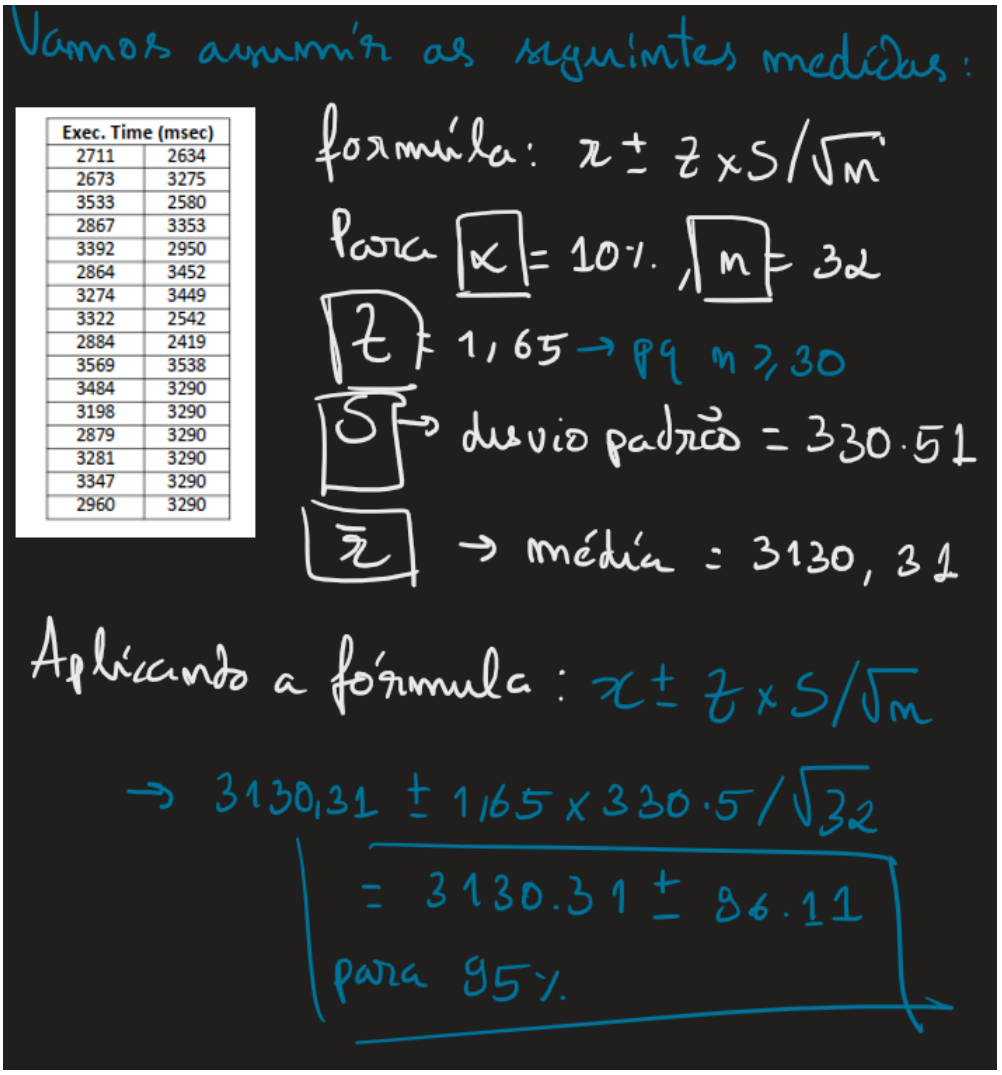
$1 - 0,05 = 0,95$

$0,95 / 2 = 0,475$

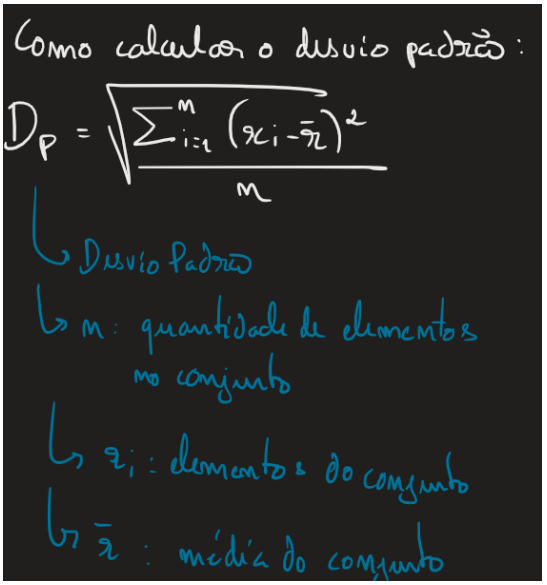

Agora temos de localizar 0,475 na tabela da dist. Z.




**Exemplo Prático n2:**




**Extra:**
**1) Cálculo do Desvio Padrão:**




### 2) Valores importantes dos níveis de confiança e valores de Z:

| Confidence Level | Z |
|---|---|
| 0.70 | 1.04 |
| 0.75 | 1.15 |
| 0.80 | 1.28 |
| 0.85 | 1.44 |
| 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 |

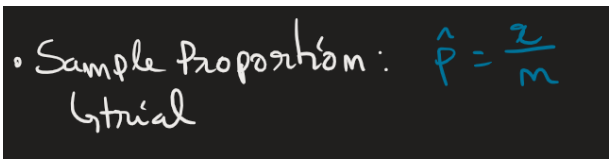
---

### Inferences for proportion - Binomial:

- Em computação, acontece que muita das vezes, as variáveis de estudo só têm um de dois possíveis resultados.

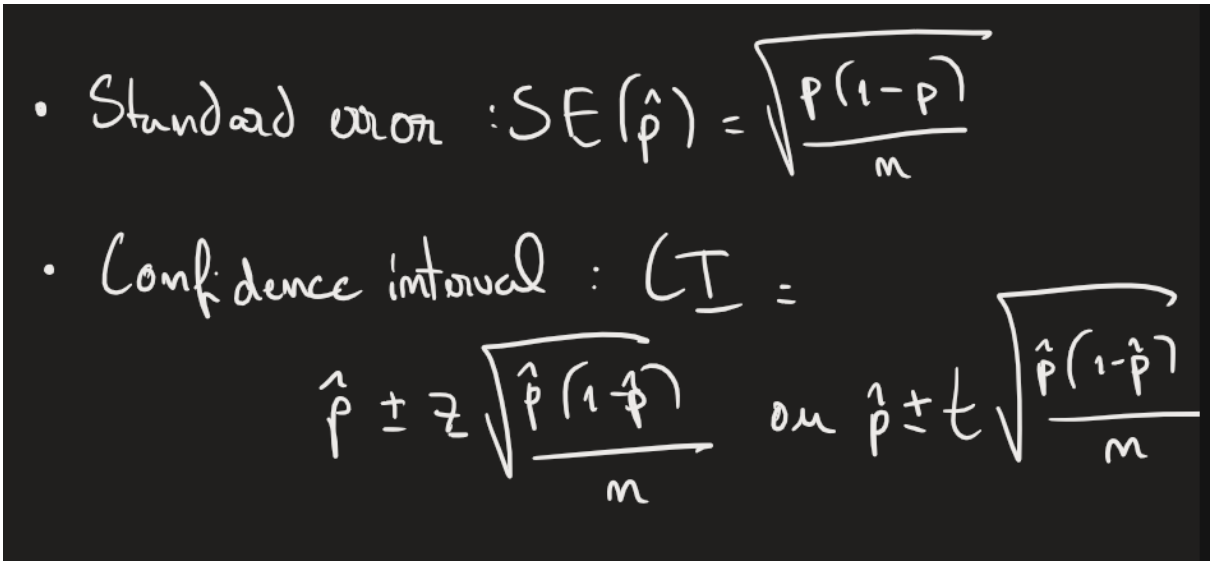
Neste caso, o modo que melhor se aplica é o **Binomial:**

1. Cada observação é chamada de “trial”.
2. O número de vezes que o resultado de interesse acontece é *x.* Normalmente chama-se até de *números de sucessos.*
3. A variável é observada um número de vezes → *n* vezes.
4. A percentagem de que o resultado de interesse aconteça é sempre o mesmo em todas as *trials.*
5. Os *trials* são independentes e o resultado de cada um não afeta o output do *trial* sucessivo.



Este é uma estimativa da proporção da população *p.* Ele varia de sample para sample numa forma aleatória.

> 💡 Para um ***n*** grande, podemos assumir uma distribuição normal. Mas também precisamos de um *n* de sucessos e insucessos maior ou igual a 10.



# PPT 4 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| Aula | T |
| # Aula Nº | 4 |

## Teste de Hipóteses

### Cenário 1:

Estamos a fazer um teste a um pacote de SQL, para saber se é mais rápido que o que temos atualmente. Temos milhares de tempos de execução do pacote antigo, e fizemos 32 testes para o novo pacote SQL.

**Comentários:**

O tempo de execução é bastante grande. Corremos 32 vezes o teste. Será que a confiança dos resultados são bons o suficiente? Para isto fazemos o teste de confiança.

### Cenário 3:

Estamos a fazer uma nova base de dados SQL. Queremos saber se um pacote A, é mais rápido do que um pacote B. Fizemos 12 testes para o pacote B, e 13 para o pacote A. Qual será o pacote mais rápido?

**Comentários:**

No cenário 1 tínhamos milhares de testes. Neste cenário nem temos um desvio padrão dos dados.

## O que é uma hipótese?

Uma hipótese é sempre uma resposta concreta a um problema, normalmente de caráter muito geral. **Normalmente refere-se à aferição de algo, a partir de uma população de amostras.** Tenta explicar o porquê de X funcionar de forma Y.

Quando a responder a uma hipótese, podemos ser **explicatórios,** ou **preditivos.** Em MEI, seremos, maior parte das vezes preditivos.

> 💡 Queremos prever o comportamento de algo, a partir de uma amostra de dados, retirados do fenómeno que é alvo de estudo.

As hipóteses são validadas com uma probabilidade de estar correta ou não.

> *Esta hipótese é muito difícil de ser correta.*

Isto significa que só conseguimos verificar se uma hipótese tem uma probabilidade muito grande de ser correta, ou se a probabilidade de ser incorreta é pequena.

Para isto normalmente temos de obedecer a um **grau de certeza → degree of certainty -** que nos irá dizer se a hipótese pode ser rejeitada ou aceite.

### Topic, Problem and Hypotesis:

A hipótese é uma resposta concreta a um problema (uma das opções de resolução do problema, visto que normalmente esta vem aos pares, *H0 → H1).* O tópico é a área de foco, seguido do problema e objeto de estudo.

Exemplos de questões:

## Research-study questions

- **Exploratory**
  Understand a phenomenon (subject of study) and clarify its features
- **Base-rate**
  Characterize the occurrence patterns of the phenomenon
- **Relational**
  Identify possible relations of the phenomenon under study with other phenomenon
- **Causal**
  Identify cause and effect related to the phenomenon under study

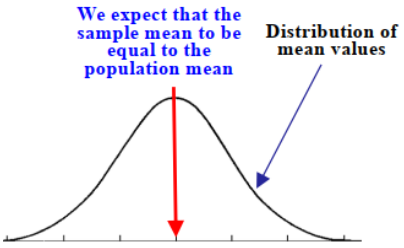
## Engineering questions

- **Design and architecture**
  Define the best engineering processes and the best architecture for products
- **Measure and optimization**
  Measure and evaluate figures of merit correctly and use the measurements to optimize products and processes
- **Benchmark and choose**
  Measure to compare and choose among alternatives (components, systems, processes)
- **Verification and validation**
  Confirms that a given implementations works as specified (verification) and solves the intended problem as expected (validation)

# Hypothesis testing and infering statistics

- Permite-nos testar o comportamento em *samples* para aprendermos mais sobre o comportamento da população total. Normalmente, isto acontece porque a população pode ser infinita ou muito grande.
- A partir do teorema do limite central, sabemos que:

> 💡 **Independentemente da forma da distribuição original de uma população, a distribuição das médias amostrais será aproximadamente normal se o tamanho da amostra for suficientemente grande.**




# Passos para testar uma hipótese:

## 1. Constatar a hipótese a ser testada

Isto passa por gerar a hipótese, e fazer o estudo dos dados. Também devemos analisar o facto da distribuição ser normal ou não.
Para a distribuição ser normal, temos de verificar se cada execução da experiência é independente das execuções antigas. Para além disso, a variabilidade das medições devem resultar de mudanças aleatórias dentro da experiência.

Depois disto, iremos definir *H0,* e *H1.* A ideia passa por mostrar que *H0 (a negação de H1)* é provavelmente rejeitável, de forma a podermos passar para a *H1* que é a hipótese realmente a ser testada.

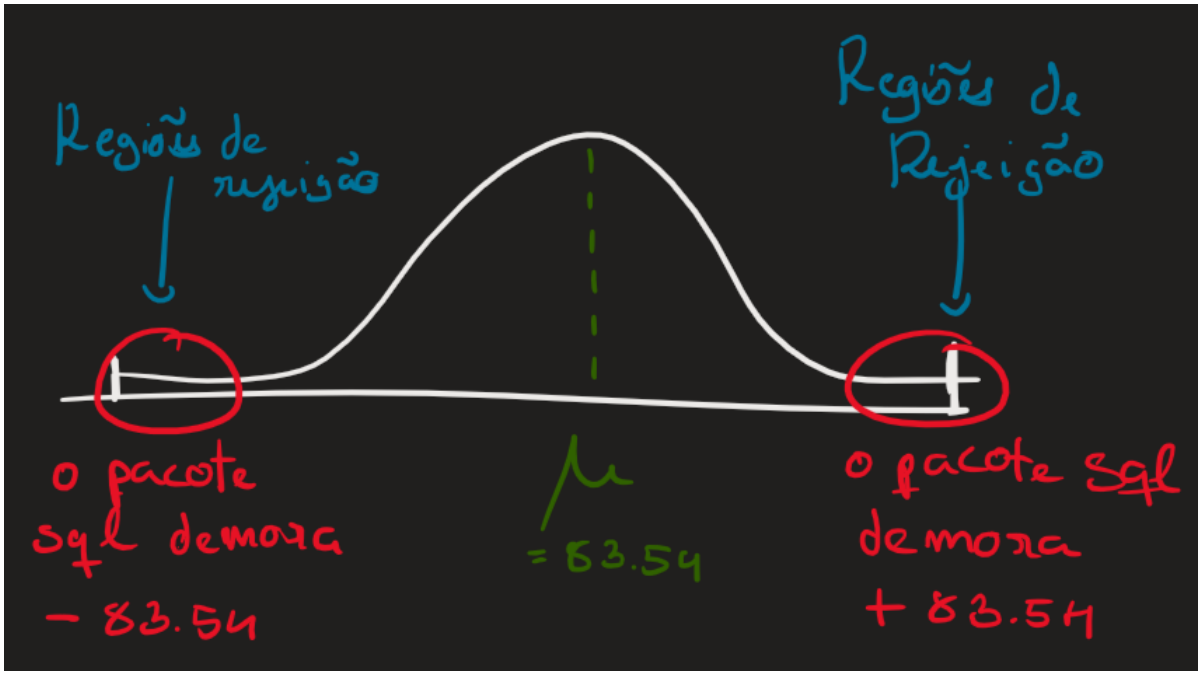
## 2. Selecionar o critério p/a decisão

Queremos ter um nível de significância, ou seja de *alfa (a),* seja 95%, 99%. O valor de significância refere-se ao critério pelo qual uma decisão é feita, tendo em conta o valor constatado na hipótese nula. O valor comum é de 5%. Isto significa que a probabilidade de obtermos uma certa média de uma sample é menos do que 5%, tendo em conta que a hipótese nula é verdadeira.

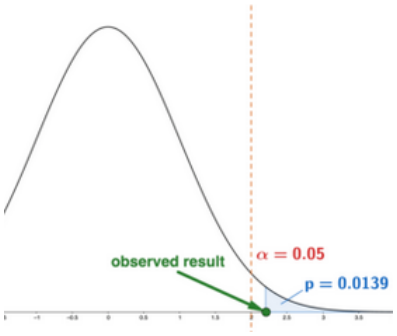
> 💡 **A ideia depois é concluir que a sample utilizada para calcular a média é demasiado improvável de acontecer, logo rejeitamos a hipótese.**

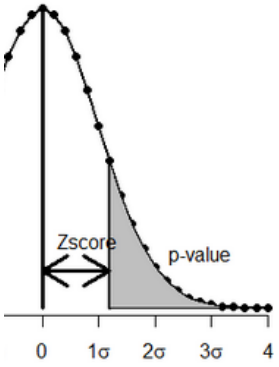
A hipótese alternativa estabelece onde colocar o nível de confiança.

*One Tailed (Uma hipótese refere-se ao mais, e outra hipótese refere-se ao menos.*

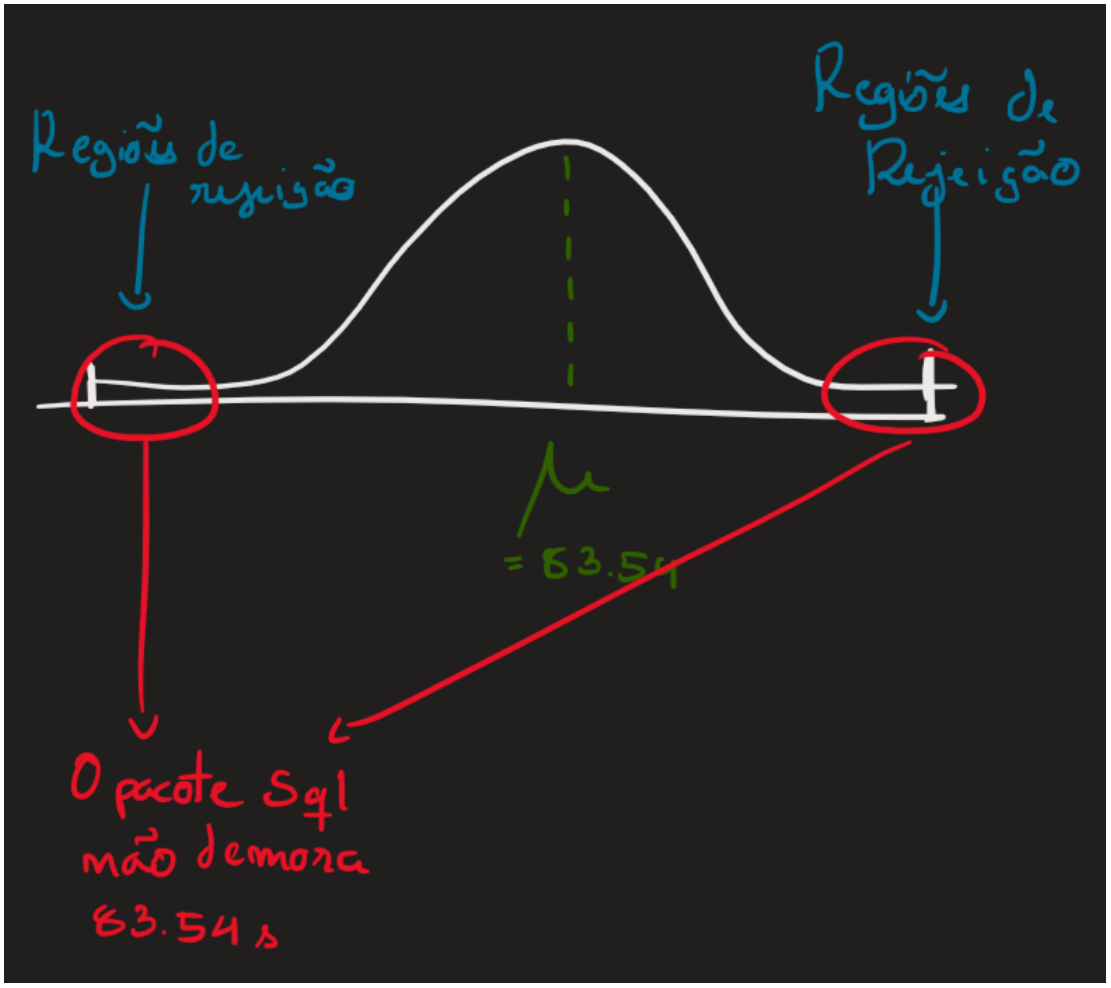



*Nota: o P-value corresponde à área delimitada pelo limite imposto pelo Z-Score ou T-Score, ou seja, o p-value é a área da distribuição, e o Z-value é a distância ao 0 (centro do gráfico).*




*Two-Tailed:*

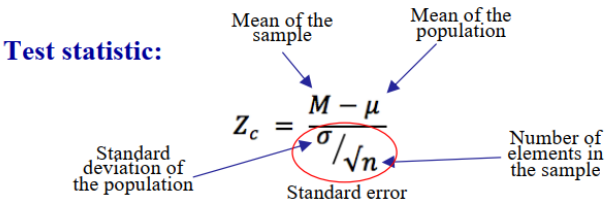



O teste de cauda dupla é para quando queremos verificar se por exemplo a média de um grupo é diferente de um valor específico. O teste de cauda único tem direção (esquerda ou direita), e estamos interessados em detetar uma diferença para uma direção específica.

### 3. Computar o teste estatístico:

Selecione uma sample aleatória da população e meça a média da sample. Para tomarmos uma decisão, temos de avaliar o quão provável o resultado desta sample é, tendo em conta que a hipótese nula (83.54) é verdadeira. Se for abaixo de 5%, então a sample é muito difícil de ocorrer, e rejeitamos a hipótese nula.

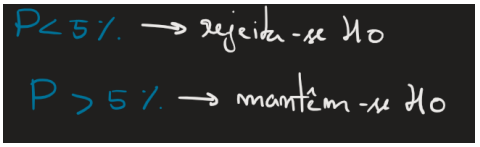
---




O desvio padrão indica o quanto os valores num conjunto de dados tendem-se a afastar da média desse conjunto. Noutras palavras quantifica a dispersão ou a variabilidade dos dados.

### Tomar a decisão:

O **P value,** é a probabilidade de obtermos uma sample, sendo que o valor constatado na hipótese nula é verdadeiro (83.54). O P-Value é o menor valor de significância (*a*) para o qual *H0* é rejeitada, pois o P-value contêm mais informação sobre o eventual desvio relativamente a H0.




Quanto menor for o *P-value* maior é a certeza na rejeição de *H0.*

## Voltando ao Exemplo

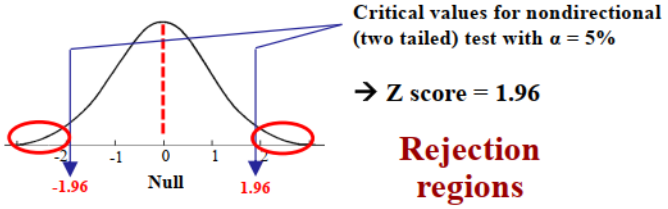
*H0 → a nova configuração não tem efeito no tempo de execução*

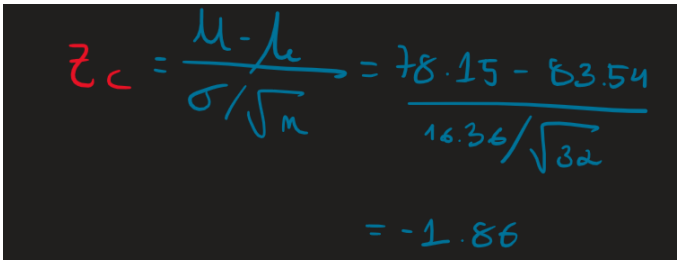
*H1 → a nova configuração faz diferença no tempo de execução (pode ser maior ou menor).*

Considerando o nível de significância a 5% → *a = 0.05.* → *1-a = 0.95*

Temos de localizar o Z na tabela que representa os **valores críticos.** Os **valores críticos** são um corte que estabelece a fronteira pelo qual menos de 5% das probabilidades podem ser obtidas se a hipótese nula for verdadeira.

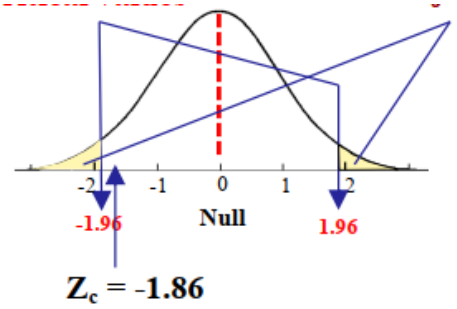
| Significance Level | Z score (two tailed) | Z score (one tailed) |
|---|---|---|
| 0.70 | 1.04 | -0.525 or 0.525 |
| 0.75 | 1.15 | -0.675 or 0.675 |
| 0.80 | 1.28 | -0.84 or 0.84 |
| 0.85 | 1.44 | -1.036 or 1.036 |
| 0.90 | 1.645 | -1.28 or 1.28 |
| 0.91 | 1.70 | -1.34 or 1.34 |
| 0.92 | 1.75 | -1.41 or 1.41 |
| 0.93 | 1.81 | -1.476 or 1.476 |
| 0.94 | 1.88 | -1.556 or 1.556 |
| 0.95 | 1.96 | -1.645 or 1.645 |
| 0.96 | 2.05 | -1.751 or 1.751 |
| 0.97 | 2.17 | -1.881 or 1.881 |
| 0.98 | 2.33 | -2.054 or 2.054 |
| 0.99 | 2.575 | -2.326 or 2.326 |




A probabilidade de obtermos Z = -1.86 é dada pelo **P-value.** Para obtermos o **P-Value,** devemos procurar por 1.86 na tabela da distribuição normal. Neste caso ele é igual a 0.0314. Neste caso, como falamos de um teste *two-tailed, multiplicamos esse valor por x2* ⇒ ***0.0628 ⇒ p = 6.28%***

> 💡 Assim temos que a probabilidade de obtermos uma média de 78.15 se *H0* é verdadeiro é de 6.28%. Ou seja, temos de reter a hipótese nula, visto que definimos 5% como o nosso grau de confiança.




$Z_c = -1.86$

# PPT 5 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 5 |

## Cenário 1 - Two-Tailed:

Corremos um pacote de SQL, para verificar se a configuração é mais rápida do que a configuração anterior. Este pacote novo foi corrido 32x. Queremos saber se este novo pacote teve algum efeito, e se a configuração é melhor.

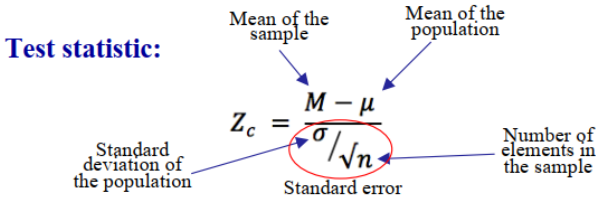
Para confiar nestes valores de média do tempo, devemos fazer um teste → Vamos também fazer um teste sem nenhuma direção: **O tempo é melhor ou é pior do que a configuração anterior?**

**Temos 4 passos** *(de forma mais prática)***:**

1. **Constatar hipótese a ser testada:** a nova configuração não tem efeito no tempo de execução do pacote SQL *vs.* o tempo de execução do novo pacote SQL é menor.
2. **Selecionar o critério para a decisão:** neste caso consideramos o **nível de significância** *a = 0.05* → O **nível de confiança** é *1-a = 0.95.*

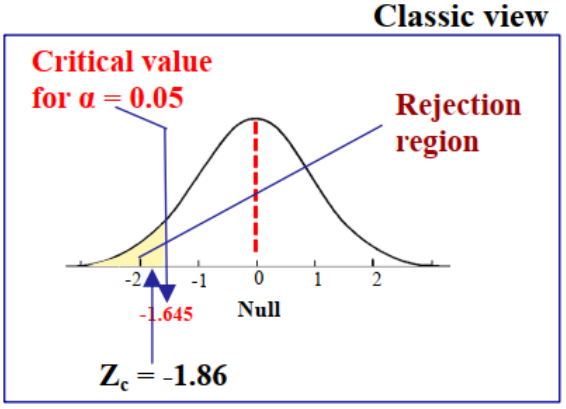
   Se escolhemos o método clássico vamos definir uma região crítica. Se o resultado que obtivermos estiver dentro da região crítica então temos de rejeitar a hipótese *H0.*

   **Se fizermos a experiência e o p-value estiver dentro da região crítica, temos de rejeitar *H0.***
3. **Calcular o teste estatístico:** neste caso é através da formula.




   Ao obtermos o Z-Score através desta forma, conseguimos obter o *p-value* consultando a tabela Z. Através do *P-value,* podemos obter a probabilidade para verificarmos se é maior ou não que o **nível de confiança.**
4. Tomar uma decisão: Se o Z não estiver dentro da região de rejeição, então não vamos rejeitar a nossa hipótese.

   Temos o utilizar o Z-Score para fazer uma probabilidade. A probabilidade é de 3.14, no exemplo dado, mas como temos uma figura com 1 lado → porque estamos com uma experiência *one-tailed,* então não precisamos de multiplicar por 2. Assim **temos de rejeitar a hipótese nula.**



**Nota:** vamos assumir que a hipótese nula é verdade → não há diferença entre as duas configurações. Então a probabilidade de executarmos 32x a experiência, calcularmos o valor da média, e obtivermos uma média maior é mais alto que o valor de confiança que estipulamos → logo temos que afirmar que a experiência não é representativa o suficiente do problema que temos.

**Na prática, o que temos ao fazer uma destas experiências é:**

1. Definir a hipótese;
2. Calcular o *p-value* e efetuar o teste estatístico.
3. Tomar uma decisão.

## Cenário 2 - Two-Tailed:

Aqui o cenário não é direcional, e two-tailed. Queremos saber se a nova configuração é diferente. *H0* mantêm-se e *H1* é que se altera.

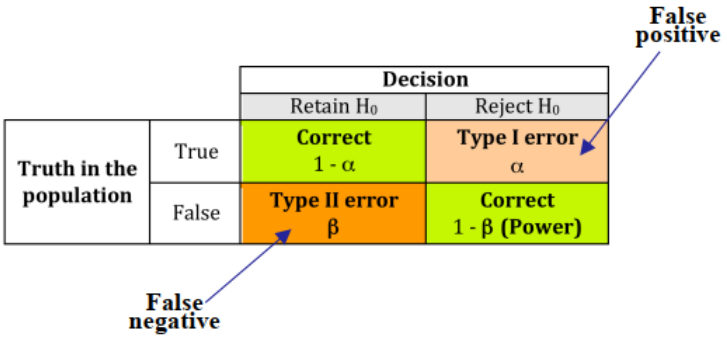
- *H0:* a nova configuração não faz diferença nenhuma.
- *H1:* a nova configuração faz diferença.

Assim vamos tomar os mesmos passos da última experiência, com todos os dados iguais, mas só temos a diferença de no passo 4, multiplicarmos o valor que foi obtido, por 2, visto que o gráfico só tem duas região critica.

A conclusão agora também vai ser diferente, apesar dos dados serem todos iguais.

**A probabilidade de termos uma média de 78.15 *se H0 é verdadeira* é de pelo menos aproximadamente 5% → isto verifica-se mais frequente do que devia de acontecer.** Assim não podemos rejeitar H0, com pelo menos de 95% de confiança.

## Tipos de erros:




**Há erros com consequências mais graves que outros** → Se tivermos um erro do tipo II, é o equivalente a fazer nada na configuração SQL p.e. Se **fosse do tipo I podíamos ter colocado uma configuração que seria pior em termos de tempo.**

Resumidamente temos:

- Tipo I: rejeitar uma hipótese nula, sendo ela verdadeira - **falso positivo**
- Tipo II: não rejeitar uma hipótese nula, sendo ela falsa - **falso negativo.**

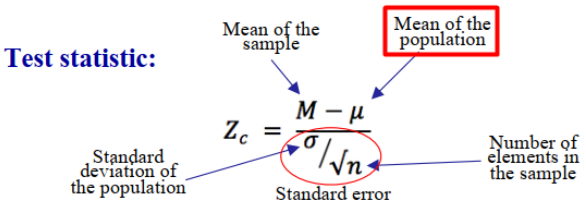
## Outras questões - Cenário Alternativo:

E se ocorrer a situação de que não temos uma quantidade de testes apropriados → <32x?

**No seguinte exemplo só temos 13,** porém para este tipo de casos existe uma distribuição, **t-student** que permite efetuar o teste estatístico como se tratasse de uma distribuição normal. Para isso, o que precisamos, é de <32x amostras e que as amostras sejam tendencialmente normais, ou seja, que não exista algo que nos permita afirmar que a distribuição apresentada não seja normal.

Assim, a única coisa que mudamos é o **teste estatístico.** Em vez de irmos buscar o valor do teste a uma tabela Z, vamos fazer o mesmo só que a uma tabela T.

Para além disso temos outra diferença:



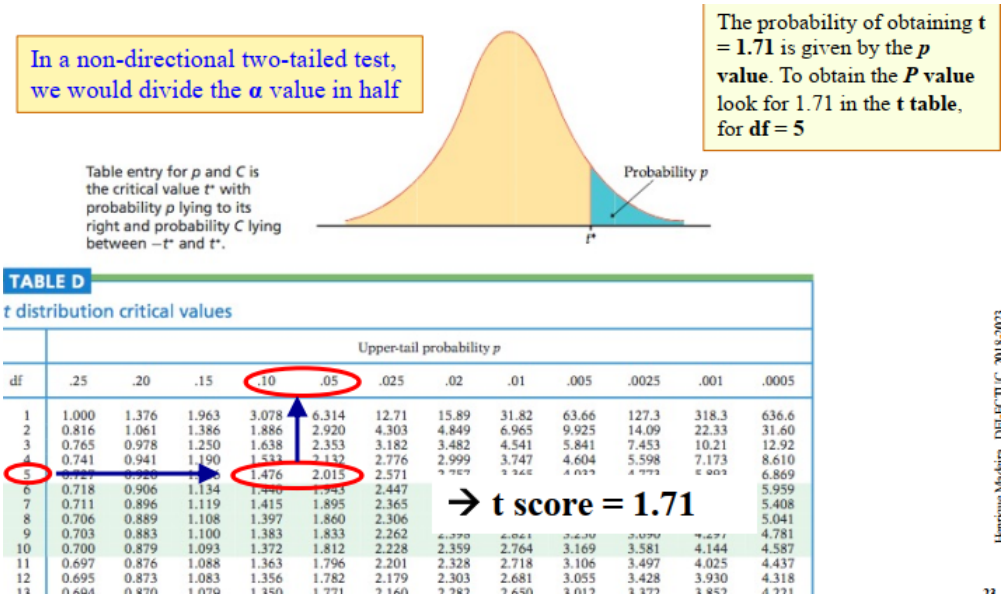

Aqui não utilizamos a média da população, mas sim o objetivo da experiência. Se afirmarmos: **queremos testar se o pacote é mais rápido do que 70 segundos,** então o objetivo assim será 70 segundos e funcionará como a média da população.

- Average of the sample: 79.17
- Standard deviation of the sample: 13.17

**Test statistic:**

$$t_c = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = 1.71$$

Ao consultar a tabela, o **grau de liberdade** é dado por N-1, ou seja, o N ⇒ número de vezes que fizemos o teste.




Como conseguimos verificar neste caso, a probabilidade de obter t=1.71, para um grau de liberdade (df) de 5, equivale a 7.4% (está entre 5% e 10%). **Logo como a p-value é menor que 10% temos de rejeitar a hipótese nula.**

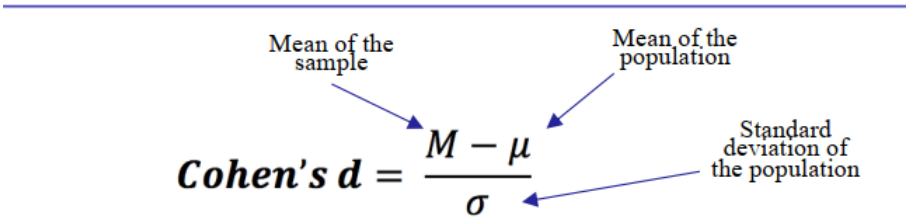
# PPT 6 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 6 |

## Medição do tamanho de um efeito:

A decisão de rejeitar a hipótese nula, normalmente significa que um certo **efeito é significativo.** Ao testarmos uma hipótese, não temos obtemos informação de quão influente é esse efeito. O tamanho do efeito ou **effect size,** não informa o quão grande este efeito é.

Para calcularmos isto, podemos utilizar o **Cohen's d,** medindo assim o número de **desvios padrões, que um certo efeito teve acima ou abaixo da média da população, afirmada na hipótese nula.**




**Cohen's effect size conventions** are often used to interpreter the effect size

If values of *d* are negative, the effect shifted below the population mean

| Description of Effect | Effect Size (*d*) |
|---|---|
| Small | $\|d\| < 0.2$ |
| Medium | $0.2 < \|d\| < 0.8$ |
| Large | $\|d\| > 0.8$ |

Na prática, também verificamos o efeito através do valor de ***p.***

## Exemplo:

- Voltamos ao mesmo exemplo dos pacotes SQL. Queremos saber se um pacote SQL é mais rápido do que o original. Fizemos a experiência 32x. Sabemos que a média da população (o pacote antigo) é de 83.54 segundos com um desvio de 16.36.
- **Será que o novo pacote teve algum efeito?** Para isto iremos então utilizar a fórmula de Cohen:

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

Os valores da medida do efeito de Cohen têm o seguinte significado:

**Cohen's effect size conventions** are often used to interpreter the effect size

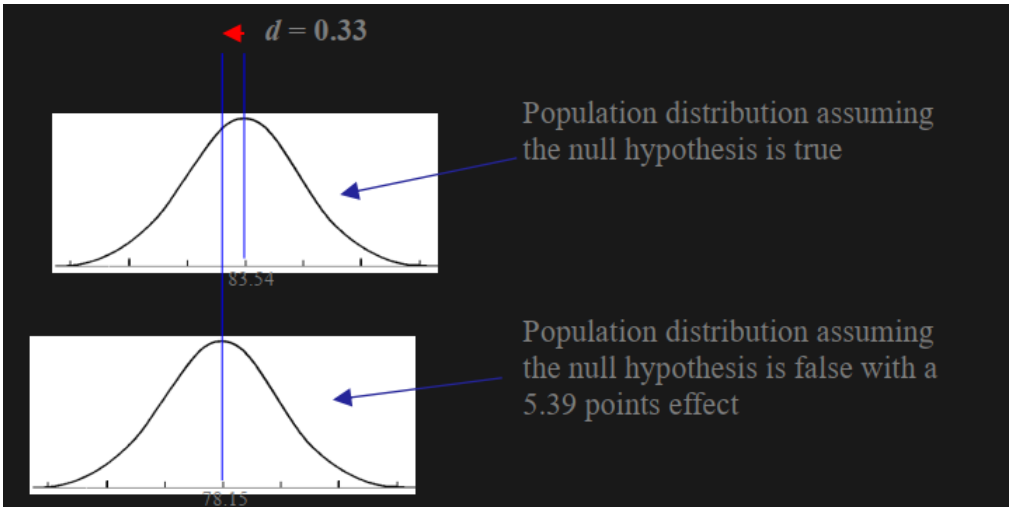
If values of *d* are negative, the effect shifted below the population mean

| Description of Effect | Effect Size (*d*) |
|---|---|
| Small | $\|d\| < 0.2$ |
| Medium | $0.2 < \|d\| < 0.8$ |
| Large | $\|d\| > 0.8$ |

podemos verificar que portanto, o *p-value* em si, também dá uma ideia (se bem que com pouco rigor) do tamanho do efeito.

**Conclusão:** *the observed effect shifted 0.33 standard deviations below the mean.*

A representação deste valor pode ser vista também da seguinte forma:

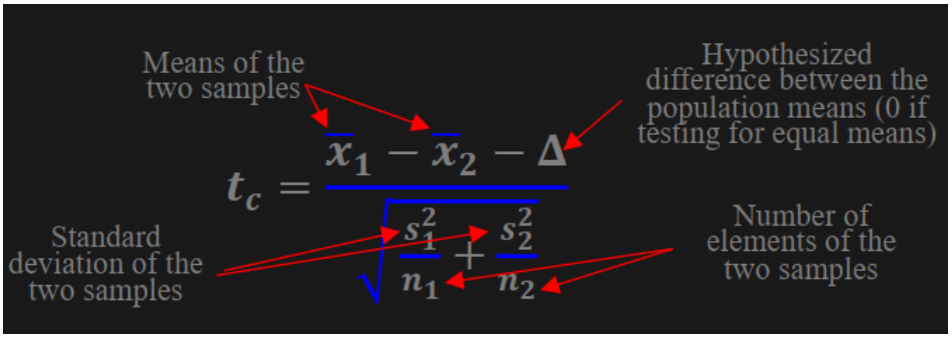



# Two Sample T-Test:

Neste tipo de teste iremos comparar uma **sample A** com uma **sample B.** Temos certos aspetos que são importantes antes de começar:

1. Temos amostras independentes ou dependentes?
2. Ambas foram testadas poucas vezes, e vezes diferentes
3. Conhecemos o desvio padrão das amostras? Neste caso não

Assim vamos ter o mesmo procedimento só que vamos consultar a tabela T na mesma. A formula do teste estatístico vai alterar para termos em consideração o desvio padrão das duas amostras e o desvio para as duas amostras.




**Neste exemplo, temos que:**

- Temos duas configurações diferentes, e pretendemos testar se uma é mais rápida que a outra.
- Vamos então definir hipóteses:
  - **H0:** a configuração A, e a configuração B são iguais.

$$H_0: \mu_1 = \mu_2$$

  - **H1**: a configuração B é a melhor. → *nota: neste caso nos vamos assumir que uma delas é a melhor, normalmente assumimos que a melhor é a que nos interessa.*

$$H_1: \mu_1 > \mu_2$$

**Agora vamos calcular o nosso T.**

| Sample | Configuration | n | $x$ | s |
|---|---|---|---|---|
| 1 | A | 13 | 78.15 | 7.94 |
| 2 | B | 12 | 73.53 | 8.33 |

**Test statistic:**

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

Para encontrar o valor na tabela utilizamos o nosso t, e calculamos o grau de liberdade, que é dado pelo número de amostras tiradas -1. Como neste caso temos dois números de amostras tiradas vamos utilizar na formula o mais pequeno número de amostras retiradas → 12-1 = 11.

Verificamos na tabela t que, o nosso valor se encontra entre 0.05 e 0.1, entre portanto 10% e 5%.

- **Sabemos que a probabilidade representa: a probabilidade de fazermos uma medida outra vez e calhar que as duas probabilidades são iguais é à volta de 9%, ou seja verificar o mesmo resultado. Não podemos assim rejeitar H0, porque temos que p > 5%.**

*Nota: Se fosse pedido o mesmo para um valor de confiança de 90%, então poderíamos ter aceitado a hipótese.*

## Resumo:

💡 **Quando temos mais de 30 samples para cada variável, deve-se utilizar o Two-Samples Z Test, como normalmente:**

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

💡 **Quando temos um sample size pequeno ( n < 30 ), queremos estudar só uma variável, e não sabemos a média da população, ou seja, é um target, então devemos utilizar o One Sample T-Test:**

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

💡 **Quando temos uma sample size pequeno, mas temos mais do que uma variável a estudar, devemos utilizar o Two-Samples T Test:**

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

💡 **Um P-value prova que, contra H0, os resultados observados e retirados da experiência são difíceis de obter, quando a hipótese 0 (nula) é verdade.**

- $p \geq 0.10$ → the observed difference is “not significant”
- $0.05 \leq p < 0.10$ → the observed difference is “marginally significant”
- $0.01 \leq p < 0.05$ → the observed difference is “significant”
- $p < 0.01$ → the observed difference is “highly significant”

# PPT 7 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 7 |

# Testes de Proporção:

Em vários casos, as variáveis dependentes das experiências, só têm **dois possíveis resultados:** *um erro detetado ou não detetado; uma vulnerabilidade detetada ou não; o sistema crashou ou não.* **Dizemos que a variável dependente é binária.**

- Normalmente utilizamos este tipo de testes para saber se dados estão corrompidos, se um teste crashou, se foram detectadas intrusões, etc. **Resumidamente, queremos saber se um fenómeno acontece ou não.**

### Propriedades:

- A variável sendo binária só pode ter 2 estados possíveis
- A variável é observada um *n* número de vezes.
    - A cada observação chama-se de **trial**. O número de vezes que o resultado que queremos que aconteça, se verifica, chama-se de **sucessos**.
- A probabilidade do resultado de interesse ocorrer para cada trial é sempre igual.
- Os trials são independentes se um trial não depende do resultado de outro trial.

*Nota: a isto se chama "modo binomial".*

## Proporção da Sample e Distribuição da Sample

**Proporção da sample (set de trials) é dada pela seguinte formula:**

$$\hat{p} = \frac{x}{n}$$

- Aqui ^*p* é a proporção da sample, *with the outcome of interest.* É uma estimativa da proporção da população total, *p*. O ^*p* varia de sample para sample de uma forma aleatória.

**Importante:** para um número de *n* grande, podemos assumir uma distribuição normal, sendo que:

- Haja um número de sucessos e de não sucessos ≥ 10
- Que a sample seja pelo menos 20x mais pequena do que a população

*Quando tivermos o caso de uma sample pequena, poodemos utilizar, mais uma vez, a tabela de t-student, mas devemos ter cuidado particular ao abordar situações como estas e nos certificarmos que os dados são representativos o suficiente.*

## Intervalos de Confiança para proporções:

- The standard error (SE) of sample proportion is given by

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

- The confidence interval (CI) for population proportion is

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

## Teste estatístico e tamanho do efeito:

*Para samples grandes (maiores que trinta):*

- Test statistic

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

Proportion observed in the sample

Proportion expected/ assumed for the population

Number of observations

- Effect size

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

*Para samples pequenas (menores que trinta):*

- Test statistic

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

**But care should be taken with quite small samples!**

- Effect size

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

## Exemplo:

Desenvolveu-se uma ferramenta que scaneia vulnerabilidades de código no desenvolvimento de aplicações. A ferramenta só pode ir para o publico se detetar pelo menos 75% das vulnerabilidades com uma rate de falsos positivos abaixo de 15%. Utilizei uma ferramenta para introduzir vulnerabilidades no código: de um total de 237 injetadas, foram detectadas 185, com 38 detectadas erradamente.

- Aqui, a proporção refere-se à quantidade de vezes, do total realizadas, que foi detetado algo. Se fizemos 100 vezes uma injeção e foi detetada 60, a proporção é de 60%.
- Para o exemplo vamos ter:
  - **H0:** a proporção de vulnerabilidades detetada não é melhor do que 0.75. p ≤ 0.75
  - **H1:** a proporção de vulnerabilidades detetada pela ferramenta é melhor do que 0.75. p > 0.75.

  *Nota: como temos diferentes problemas (queremos saber tanto a taxa de falsos positivos como de positivos), necessitamos de diferentes hipóteses.*
- Aplicando a fórmula:

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

**Z = 1.088**

- Agora vamos à tabela Z, para que *z = 1.088,* visto que o teste é *one-tailed.* Vemos assim que o Z está entre 0.1401 e 0.1379. Ao utilizarmos uma calculadora vemos que *p=0.1383.*
- Visto que o *p* está entre 14% e 13%, não podemos rejeitar a hipótese nula. Como a ferramenta falhou neste teste, nem sequer precisamos de testar o caso dos falsos positivos.
- **Conclusão: a ferramenta tem de ser melhor a detetar as vulnerabilidades.** Utilizando o que foi dado na última aula, podemos adicionalmente verificar o *effect size:*

Small ***p*** values provide evidence against the null hypothesis, as it means that the observed data are unlikely when the null hypothesis is true.

**Conventions:**

- $p \geq 0.10$ → the observed difference is **"not significant"**
- $0.05 \leq p < 0.10$ → the observed difference is **"marginally significant"**
- $0.01 \leq p < 0.05$ → the observed difference is **"significant"**
- $p < 0.01$ → the observed difference is **"highly significant"**

## Medição dos Intervalos de confiança:

### Proporção de vulnerabilidades detetadas:

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

Z = 1.96
from Z table for 95% confidence, two tailed

CI for a confidence level of 95%

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

**Proportion of vulnerabilities detected (95% confidence) = 0.7806 ± 0.0527**

Coverage of the tool between 72.79% and 83.33% with 95% confidence

### Proporção dos falsos positivos:

Para calcular os falsos positivos devemos considerar todas as vulnerabilidades detetadas, ao calcular a proporção:

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

For the false positives we should consider all the vulnerabilities detected: 185 (correct) + 38 (incorrectly indicated as vulnerabilities)

CI for a confidence level of 95%

Neste caso então já não temos que colocar todos os casos, mas sim todos os casos positivos.

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

**Proportion of false positives (95% confidence) = 0.1704 ± 0.0493**

False positive detection of the tool is between 12.11% and 21.98% with 95% confidence

---

# Caso Extra:

## Two-Proportion Z-Test:

Quando queremos verificar se a diferença entre duas proporções é significativa. Digamos, utilizando o exemplo anterior, se quisermos comparar as proporções de vulnerabilidades detetadas entre duas ferramentas, P1 e P2.

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

Em caso de termos uma experiência, uma destas hipóteses pode ser relevante para ser testada.

Adicionalmente para conseguirmos computar o teste, precisamos de uma **pooled sample proportion -** proporção de amostras combinadas, ou seja, juntar as diferentes amostras assumindo que estas amostras passam agora a representar a população total.

$$\hat{p} = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2}$$

p1 is the sample proportion from population 1,
p2 is the sample proportion from population 2,
n1 is the size of sample 1
n2 is the size of sample 2.

The standard error (SE) is:

$$SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

**The test statistic**

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

Depois disto, obedecendo aos passos de todos os testes até agora, é obter o valor p, ou uma estimativa dele, e tomar uma decisão.

# PPT 8 / T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| ⁝≡ Aula | T |
| # Aula Nº | 8 |

## Resumo das últimas aulas:

So far we have studied hypothesis testing in different sampling scenarios with (approximately) known distributions (parametric tests):

- Type of samples
  - Large (≥30) samples: Z (normal) distribution
  - Small (<30) samples: T Student distribution
- Number of samples
  - **Single sample**: only one group of observations; test against a hypothetical **mean** or **proportion**; Z test for large samples and T test for small ones.
  - **Two samples**: two groups of observations; test the difference between **means** or **proportions**; Z test for large samples and T test for small ones.
  - **Three or more samples**: several groups; test the variance (**ANOVA**)
- Nature of the samples
  - **Independent samples**: groups are not related and observations are truly independent
  - **Dependent samples**: when one observation/measurement in a group is related to one observation in a another group. Also called matched pairs, matched samples, etc.

Agora vamos analisar o caso de termos *samples* dependentes e 3 ou mais.

**Exemplo 1:**

- Número de downloads de um grupo de aplicações, depois de uma campanha de publicidade

**Exemplo 2:**

- Número de vulnerabilidades encontrados em code inspections pelos 10 mesmos engenheiros depois de um treino de segurança.

## Duas samples dependentes:

São samples dependentes porque as medidas podem ser relacionadas, ou com a aplicação em si e a sua campanha, e com o engenheiro em questão, no segundo exemplo.

O teste estatístico que iremos utilizar é baseado na média das diferenças entre as entradas de pares de dados nas samples dependentes.

entries in the dependent samples.

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

Difference between entries for a data pair

Number of pairs

The standard deviation $S_d$ of the differences between the paired data entries in the dependent samples

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

Podemos utilizar esta metodologia se e só se:

- As samples são selecionadas aleatoriamente

- As samples são dependentes (pares)
- Ambas as populações forem normalmente distribuídas

Podemos dizer que há uma distribuição normal neste caso n ≥ 30, ou uma distribuição *T* caso n < 30. Vamos ter testes estatísticos diferentes para sets de pares de dados maiores ou menores:

$$z = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

Test statistics for **large sets** of paired samples (n ≥ 30)

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

Test statistics for **small sets** of paired samples (n < 30). The degree of freedom is n-1.

Depois disto, os mesmos passos para o cálculo de hipóteses estatísticas são utilizados, lembrando sempre como calcular tanto o *d* como o *Sd*.

---

## Exemplo

A qualidade do desenvolvimento de uma Web App, foi medida antes e depois dos engenheiros terem sido alvos de um treino. Temos resultados da qualidade antes da formação, e depois da formação:

| 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 |

Com 95% de certeza, conseguimos dizer que o treino melhorou a capacidade dos engenheiros em desenvolver aplicações Web?

**As hipóteses são:**

- ***H0:*** os scores depois do treino não são melhores do que antes do treino ⇒ *u ≤ 0.*
- ***H1:*** os scores depois do treino são melhores do que os scores anteriores ao treino ⇒ *u > 0.*

**A questão é: conseguimos verificar com 95% de confiança que o treino melhorou as capacidades dos developers?**

**Para este teste temos de fazer alguns cálculos intermédios:**

| 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 | $\sum d = 61$ |
| $d^2$ | 25 | 36 | 361 | 100 | 121 | 9 | 36 | 25 | 49 | 225 | $\sum d^2 = 987$ |

**Depois de fazermos o cálculo do *d* e do *d^2, de forma a termos o seu somatório, temos que:***

$$\bar{d} = \sum \frac{\sum d}{n} = \frac{61}{10} = 6.1$$

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

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

**Agora temos de consultar a tabela, para *t=2.33* e *n = 10* ou *df = 9***

Temos que, através da tabela, a probabilidade está entre 2% e 2.5%. Assim podemos dizer que o **efeito é significativo**, logo somos obrigados a rejeitar H0 com 95% de confiança. **O treino conseguiu melhorar as capacidades dos developers.**

---

## Resumo:

**Usamos este tipo de teste quando:**

- Temos medidas repetidas para os mesmos indivíduos, sistemas ou componentes.
- Estudos com pares de membros familiares

**Vantagens:**

- Potencias fontes de enviesamento são conhecidas e controladas.
- O desvio padrão do teste estatístico irá ser menor, tornado este teste mais preciso do que um teste Z ou T

**Desvantagens:**

- Em certos casos é complicado encontrar os mesmos objetos/participantes
- Quando a hipótese nula é rejeitada, é difícil argumentar que a diferença se deve a eventos globais e não em relação ao teste e reteste dos mesmos indivíduos.

# PPT 9 /T / MEI

| | |
|---|---|
| ↗ Course | MEI |
| Aula | T |
| # Aula Nº | 9 |

## Análise da Variância - ANOVA

Vamos abordar o mesmo problema das configurações SQL. Neste caso vamos ter 3 configurações diferentes, e queremo-las comparar, para saber qual é a melhor, ou se são todas iguais ou ainda indiferentes.

Também podemos utilizar outro exemplo. Queremos verificar se o tamanho dos elementos num array, afetam o tempo necessário para dar *sort* a todos os elementos do array. A pergunta é: **O tamanho dos elementos contidos em cada array faz diferença?**

Temos várias variáveis que temos de ter em conta. O problema precisa mais do que só olhar para médias usando uma abordagem informal - O objetivo é testar se a diferença entre várias médias de samples é significante. Para isto precisamos de **ANOVA → Análise da Variância.**

| Size of elements (no. characters) | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1** | **3** | **5** | **10** | **25** | **50** | **100** | **200** |
| 71 | 100 | 104 | 106 | 106 | 105 | 108 | 104 |
| 71 | 95 | 99 | 102 | 105 | 101 | 105 | 105 |
| 69 | 94 | 103 | 99 | 104 | 106 | 108 | 104 |
| 69 | 99 | 100 | 103 | 102 | 105 | 104 | 103 |
| 68 | 99 | 103 | 108 | 101 | 106 | 105 | 108 |
| 66 | 97 | 100 | 102 | 107 | 107 | 107 | 110 |
| 71 | 103 | 99 | 104 | 105 | 101 | 104 | 104 |
| 71 | 98 | 105 | 100 | 106 | 107 | 100 | 102 |
| 66 | 100 | 103 | | 102 | 105 | 106 | 105 |
| 65 | | 105 | | 103 | | 102 | |
| Means (milliseconds) | | | | | | | |
| 68.7 | 98.3 | 102.1 | 103.0 | 104.1 | 104.8 | 104.9 | 105.0 |

Informalmente, para compararmos médias de diferentes samples, podemos utilizar gráficos. Mas para saber se os fatores que utilizamos para realizar a experiência são ou não significativos, necessitamos de analisar:

- Diferença das médias
- O desvio padrão de cada grupo
- O tamanho das samples.

## One-Way ANOVA:

O One-Way anova é usado para testar se a média de tres ou mais populações são iguais. Podemos considerar que a ANOVA é uma extensão do *two-independent sample tests,* que vimos na aula 7.

**Através do ANOVA testamos as seguintes hipóteses:**

- H0: *u1 = u2 = u3 … = uk* → a média de todos os grupos são iguais
- H1: Nem todas as médias são iguais

**Então temos que, tendo em conta o último exemplo:**

- **Variável Dependente:** a variável que estamos a comparar - sorting time
- **Variável Independente:** o fator variável que estamos a usar para definir os grupos/samples → tamanho dos elementos
- **Níveis:** neste caso temos 5 níveis (10 chars, 25, 50, 100 e 200).

**Para podermos utilizar o teste da ANOVA temos de assumir que:**

- Cada grupo é “aproximadamente” normal
- O desvio padrão de cada grupo é aproximadamente igual

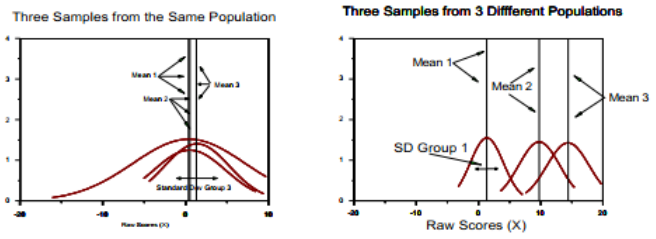
Para o exemplo, temos pelo menos 3 médias para testar. Podemos usar dois T-test para testar 2 de cada vez, mas o ANOVA consegue-os analisar aos 3 de uma só vez. Para além disso é um processo logicamente muito mais demorado:

The number of comparisons (tests) increase when using two sample t-tests to test successively pairs of samples at a time. Type I error (false positive) also increase dramatically.

| Number of samples | Number of tests |
|---|---|
| 2 | 1 |
| 3 | 3 |
| 4 | 6 |
| 5 | 10 |

Assim em vez de usar a diferença média, usamos a variância das médias do grupo em relação à média geral de todos os grupos. A lógica continua a ser a mesma do Z-test ou do T-test → comparar a variância observada entre as médias, com o que esperaríamos obter (o nosso objetivo).

Para melhor perceber isto graficamente:




Supondo que tínhamos médias de 3 diferentes populações. Os resultados seriam como na figura superior. Quando as médias são mais separadas, a variância irá ser maior. Como é que aqui conseguimos verificar se há algum efeito ou não?

> 💡 **Para decidir, temos de comparar a variância observada entre as médias, e se H0 for verdade, quer dizer que não existe diferença entre as médias.**

### Voltando ao Exemplo…

- O mesmo problema das configurações, desta vez vamos ter 3 configurações SQL. Quero saber se de uma vez, consigo comparar os 3 testes, comparando-as aos pares. A com B, B com C, A com C.
- Também temos o exemplo do processamento de uma array, cada elemento é uma string de chars, afetam o tempo necessário pra dar sort a todos os elementos da Array.
  - O tamanho dos elementos interessa? Nos níveis que temos presente (10, 25, 50, 100) , não se nota muito diferença. mas…
- Se aumentarmos os níveis, vemos que faz bastante diferença, dos níveis iniciais para os últimos principalmente (1, 3, 5)

Para determinar a variância que existe entre as médias de cada população iremos utilizar ANOVA, que por consequência diz-nos se existe diferença nos resultados entre cada uma das médias dessas populações → **ou seja, o efeito é real.**

## Definições dos elementos da ANOVA:

$\bar{X}_G$ The Grand Mean, taken over all observations.

$\bar{X}_A$ The mean of any level of the IV (group).

$\bar{X}_{A_1}$ The mean of a specific level (1 in this case).

$X_i$ The observation or raw data for the $i^{th}$ measurement.

**Variation** is the sum of the squares of the deviations between a value and the mean of the sample (group)

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

**Sum of Squares** (SS) is often followed by a variable in parentheses such as $SS_{(B)}$ or $SS_{(W)}$ that indicates which sum of squares is referred to:

- **O Sum of Squares,** mede 3 fontes possíves de variação: Grupos (entre a média dos grupos), o erro (variação dentro dos grupos), e o Total. Para cada um destes também teremos diferentes graus de liberdade:

df (total) = $n - 1$
df (within) = $n - k$
df (between) = $k - 1$

df (total) = df (between) + df (within)

## Cálculos ANOVA:

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

The grand mean is the weighted average of the individual sample means

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

The total sum of squares comes from the distance of all the scores to the grand mean. This is the big total.

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

The within-group sum of squares comes from the distance of the scores to the sample means. The degrees of freedom are equal to the sum of the individual *df* for each sample. This indicates error.

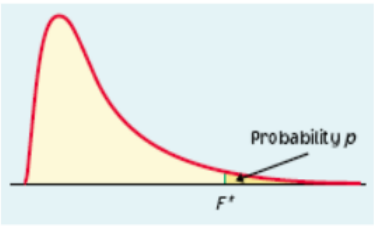
$$SS_{(B)} = \sum n_A (\bar{X}_A - \bar{X}_G)^2$$

The between-groups sum of squares represents the distance of the sample means from the grand mean. This indicates IV effects.

Quando a calcular um teste estatístico ANOVA, iremos utilizar a **F-Statistic - da tabela F-Test** este determina se a variação entre as médias das samples é significante:

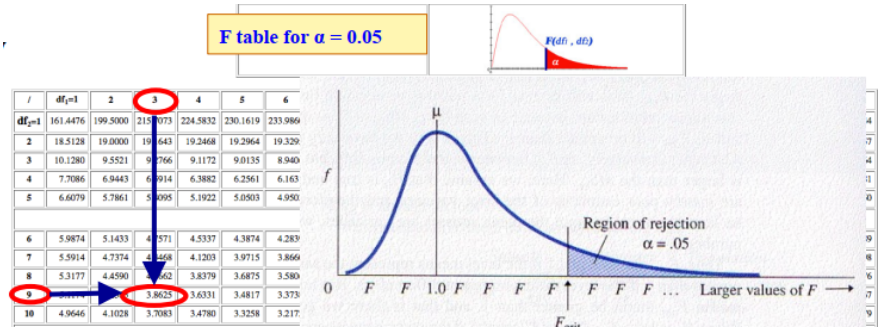
$$\frac{\text{Variation Among Sample Means}}{\text{Variation Among Individuals In Each Sample}}$$

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




Conseguimos encontrar este valor (o valor critico), na tabela F, para um certo *a* e um certo grau de liberdade.

Na tabela, para procuramos o valor de F(3,9) por exemplo - *o 3 refere-se ao número de níveis - 1, e o 9 refere-se a n - k (observações menos o número de níveis).*




### Tabela ANOVA:

Todos os cálculos que necessitamos de efetuar para chegar ao *p-value,* podem ser encontrados na tabela ANOVA.

| | 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 | | | |

O F-Value é dado por:

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

## Exemplo:

Os passos para realizar o teste estatístico são iguais. Quando chegamos à parte dos cálculos é que podemos ter algum trabalho extra. Voltando ao exemplo anterior, *como é que os elementos de uma array afetam o tempo necessário para dar sort a todos os elementos do array?*

Definimos só 3 níveis da variável independente.

Definimos as hipóteses:

- H0: todas as samples têm médias iguais
- H1: nem todas as médias são iguais

> 💡 é importante perceber que não descobrirmos assim quais é que diferem de quais, simplesmente sabemos se o tamanho tem efeito no tempo ou não.

Depois temos de preencher a tabela ANOVA para n=27, e k=3:

| 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 | |

| | 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 |

Para isso, iremos calcular:

- **Grand mean Calculation:** *weighted average of the individual sample means*

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

| | 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 |

$$GM = \frac{9(98.3) + 10(102.1) + 8(103.0)}{9 + 10 + 8} = 101.10$$

- **Between Group Variation:** *variation between each sample and the grand mean*

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

$$SS(B) = 9(98.3 - 101.10)^2 + 10(102.1 - 101.10)^2 + 8(103.0 - 101.10)^2 = 107.77$$

$$GM = 101.10$$

| | 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 |

- **Within Group Variation:** *weighted total of the individual variations*

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

| | 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 |

$$SS(W) = 8(2.74)^2 + 9(2.38)^2 + 7(2.98)^2 = 172.90$$

Assim, ficamos com a tabela preenchida:

| | SS | df | MS | F | P |
|---|---|---|---|---|---|
| **Between** | 107.77 | 2 | 53.88 | 7.48 | ≈ 0.003 |
| **Within** | 172.90 | 24 | 7.20 | | |
| **Total** | 280.67 | 26 | | | |

Assim temos que o F = 7.48, com MS(B)\MS/W). Conseguimos chegar ao p-value através de uma calculadora online → 0.003. Depois disto se realizar, só resta é tomar a decisão, para verificar (ou não) se a variação entre as médias das samples é significante.

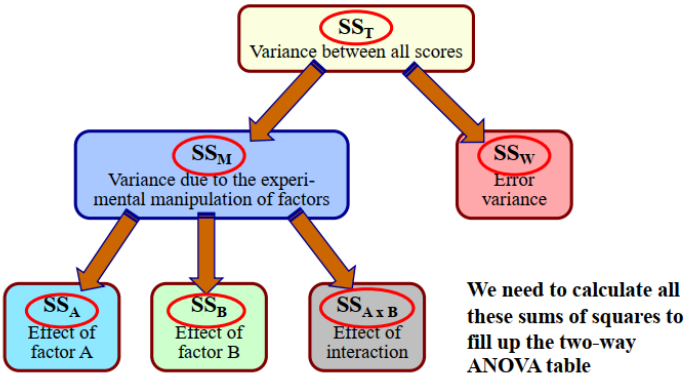
# Two-Way ANOVA:

Serve para testarmos a equidade entre 2 ou mais médias de população, quando utilizamos 2 variáveis independentes - dai também poder ser chamado de “multi-way Anova”. Aqui teremos os mesmos resultados que obtemos em testes One-Way ANOVA consecutivos ou seja, **por variável.** Mesmo assim, ao contrário do ANOVA One-Way, temos que a interação da variável independente pode ser testada.

O procedimento é o mesmo de sempre, mas há algumas diferenças na definição das hipóteses:

- H0: Não há diferença entre as médias devido ao fator X
  - *U1 = U2 = … Un*
- H0: Não há diferença entre as médias devido ao fator Y
  - *U1 = U2 = … Un*
- H0: Não há diferença entre as médias devido ao fator X ou fator Y
  - *ABij = 0.*

Para o cálculo do Sum of Squares, temos que calcular as seguintes:




O Two-Way ANOVA passa a ter mais colunas e linhas:

| | SS | df | MS | F | P |
|---|---|---|---|---|---|
| **Factor A** | $SS_{(A)}$ | a-1 | $\frac{SS(A)}{a-1}$ | $\frac{MS(A)}{MS(W)}$ | Tail area above F |
| **Factor B** | $SS_{(B)}$ | b-1 | $\frac{SS(B)}{b-1}$ | $\frac{MS(B)}{MS(W)}$ | Tail area above F |
| **Interaction A x B** | $SS_{(AxB)}$ | (a-1)(b-1) | $\frac{SS(A)}{(a-1)(b-1)}$ | $\frac{MS(AxB)}{MS(W)}$ | Tail area above F |
| **Within (error)** | $SS_{(W)}$ | n-a-b | $\frac{SS(W)}{n-a-b}$ | | |
| **Total** | $SS_{(M)} + SS_{(W)}$ | n-1 | | | |

*Voltando ao nosso exemplo….*

> Como é que o número de elementos numa array, e o tamanho de cada elemento afetam o *sorting time?*

Para economizar tempo, iremos considerar só 3 níveis da variável independente **tamanho de cada elemento,** e 2 níveis para a variável independente **número de elementos.**

| 10 chars | | 50 chars | | 100 chars | |
|---|---|---|---|---|---|
| **10K** | **50K** | **10K** | **50K** | **10K** | **50K** |
| 106 | 313 | 105 | 308 | 108 | 312 |
| 102 | 307 | 101 | 309 | 105 | 307 |
| 103 | 308 | 106 | 307 | 108 | 307 |
| 103 | 309 | 105 | 307 | 104 | 311 |
| 108 | 306 | 106 | 311 | 105 | 312 |
| 102 | 308 | 107 | 310 | 107 | 308 |
| 104 | 310 | 101 | 311 | 104 | 308 |
| 105 | 312 | 107 | 311 | 100 | 309 |
| 108 | | 105 | | 106 | 312 |
| | | | | 102 | |

Definição de hipóteses:

**Hypothesis**

- **$H_0$:** $\mu_{a.10} = \mu_{a.50} = \mu_{a.100}$ → The size of elements to be sorted is not relevant for the sorting time (means are equal)
- **$H_0$:** $\mu_{b.10k} = \mu_{b.100K}$ → The number of elements to be sorted is not relevant for the sorting time (means are equal)
- **$H_0$:** $AB_{ij} = 0$ → There are no interaction between the size and the number of elements
- $H_1$: → Not all the means are equal and there is interaction (only right tail is possible).

Agora vamos tentar preencher a tabela ANOVA:

| | 10 chars | | 50 chars | | 100 chars | |
|---|---|---|---|---|---|---|
| | 10K | 50K | 10K | 50K | 10K | 50K |
| Sample size | 9 | 8 | 9 | 8 | 10 | 9 |
| 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 |

$GM = \frac{\sum n_i \bar{x}_i}{\sum n_i}$ Weighted average of the individual sample means

E vamos calcular todos os restantes parâmetros para a ANOVA, preenchendo no final a tabela:

| | SS | df | MS | F | P |
|---|---|---|---|---|---|
| **Factor A (size elem.)** | SS(A) | a-1 | $\frac{SS(A)}{a-1}$ | $\frac{MS(A)}{MS(W)}$ | Tail area above F |
| **Factor B (no. elem.)** | SS(B) | b-1 | $\frac{SS(B)}{b-1}$ | $\frac{MS(B)}{MS(W)}$ | Tail area above F |
| **Interaction A x B** | SS(AxB) | (a-1)(b-1) | $\frac{SS(A)}{(a-1)(b-1)}$ | $\frac{MS(AxB)}{MS(W)}$ | Tail area above F |
| **Within (error)** | SS(W) | n-a-b | $\frac{SS(W)}{n-a-b}$ | | |
| **Total** | SS(M) + SS(W) | n-1 | | | |

- **SS(t):** total sum of squares comes from the distance of all the scores from the grand mean (this is the big total).

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

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

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

Performing the calculation: $SS_{(T)} = 552982$

- **SS(m):** sum of squares that gives the variance due to the experimental manipulation of factors (all factors are considered here):

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

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

Number of samples of level *i*

Average of the scores of level *i*

Performing the calculation:

$$SS_{(M)} = 9(104.1 - 201.25)^2 + 8(308.8 - 201.25)^2 + 9(104.8 - 201.25)^2 + 8(309.3 - 201.25)^2 + 10(104.9 - 201.25)^2 + 9(309.8 - 201.25)^2 = \mathbf{552736.78}$$

- **SS(a):** sum of squares for factor A (effect of factor A, size of each element to be sorted)

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

$$SS_{(A)} = 17(200.82 - 201.25)^2 + 17(201.00 - 201.25)^2 + 19(201.84 - 201.25)^2 = \mathbf{10.81}$$

- **SS(b):** The sum of squares for factor B (effect of factor B, number of elements to be sorted)

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

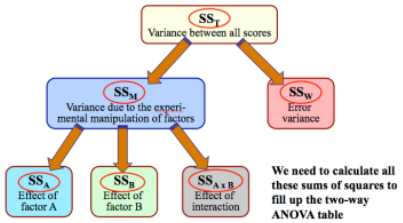
$$SS_{(B)} = 28(104.75 - 201.25)^2 + 25(309.32 - 201.25)^2 = 552721.12$$

- **SS(a*x*b):** The sum of squares for the interaction between factor A and factor B

$$SS_{(A\times B)} = SS_M - SS_A - SS_B$$

Performing the calculatic

$$SS_{(AxB)} = 552736.78 \cdot$$




- **Within Group Variation:** Is the weighted total of the individual variations (error). The weight is made with the degrees of freedom ( -1 !):

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

$$SS_{(W)} = 1.96^2\,(9-1) + 2.12^2\,(8-1) + 2.28^2\,(9-1) + 1.75^2\,(8-1) + 2.56^2\,(10-1) + 2.19^2\,(9-1) = 245.28$$

Preenchendo a 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 | | | |

Ao calcularmos o *p-value,* temos que:

1. Para F(2, 48) = 1.06, temos que $p = 0.36 \Rightarrow$ 36%. H0 é mantido.
2. Para F = 1081, temos que $p = 0 \Rightarrow$ Logo, H0 é rejeitado.
3. Para F(2, 48) = 0.47, temos que $p = 63 \Rightarrow$ 63%. Logo, temos que H0 é mantido.

**H0, no seu todo, é rejeitado, visto que tem influência, no segundo ponto.**

# PPT 10 /T /MEI

| | |
|---|---|
| ↗ Course | MEI |
| Aula | T |
| # Aula Nº | 10 |

## Testes não paramétricos:

Existem certas situação que requerem testes não paramétricos. Os dados podem não ter qualquer interpretação, ou o interesse num certo parâmetro da população ser desconhecida (médias, variâncias, etc).

> 💡 É para os casos onde os dados não seguem uma distribuição paramétrica (à partida) que iremos utilizar estes métodos:

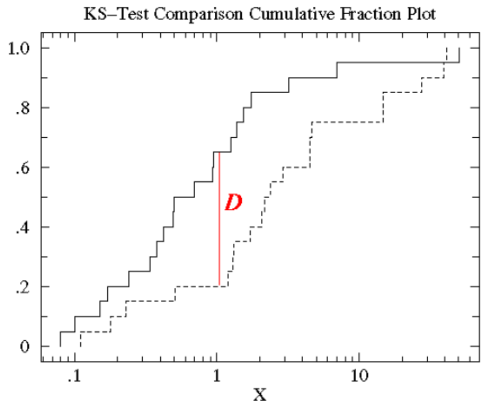
**Pros:** têm menos suposições do que testes paramétricos.

**Contras:** os testes são normalmente menos rigorosos.

> *Existe uma panóplia de testes não paramétricos, mas iremos focar no teste:*

## Kolmogorov-Smirnov Test (two sample):

Na prática o que este teste faz é avaliar a significância da **divergência máxima (D) entre duas curvas cumulativas**. Se o D for maior que os valores críticos, para um certo grau de confiança, então a diferença é significativa.




Cumulativa porque as samples não irão ser medidas individualmente, como iremos verificar.

A tabela dos valores críticos (D-Values) podem estar assinalados com um *, o que significa que não podemos rejeitar H0, **independentemente do observado D.**

| $n_2 \backslash n_1$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | * * | * * | * * | * * | * * | * * | * * | * * | * * | * * |
| 2 | * * | * * | * * | * * | * * | 16/16 * | 18/18 * | 20/20 * | 22/22 * | 24/24 * |
| 3 | * * | * * | 15/15 * | 18/18 * | 21/21 * | 21/24 24/24 | 24/27 27/27 | 27/30 30/30 | 30/33 33/33 | 30/36 36/36 |
| 4 | | 16/16 * | 20/20 * | 20/24 24/24 | 24/28 28/28 | 28/32 32/32 | 28/36 32/36 | 30/40 36/40 | 33/44 40/44 | 36/48 44/48 |
| 5 | | | * * | 24/30 30/30 | 30/35 35/35 | 30/40 35/40 | 35/45 40/45 | 40/50 45/50 | 39/55 45/55 | 43/60 50/60 |
| 6 | | | | 30/36 36/36 | 30/42 36/42 | 34/48 40/48 | 39/54 45/54 | 40/60 48/60 | 43/66 54/66 | 48/72 60/72 |
| 7 | | | | | 42/49 42/49 | 40/56 48/56 | 42/63 49/63 | 46/70 53/70 | 48/77 59/77 | 53/84 60/84 |
| 8 | | | | | | 48/64 56/64 | 46/72 55/72 | 48/80 60/80 | 53/88 64/88 | 60/96 68/96 |
| 9 | | | | | | | 54/81 63/81 | 53/90 70/90 | 59/99 70/99 | 63/108 75/108 |
| 10 | | | | | | | | 70/100 80/100 | 60/110 77/110 | 66/120 80/120 |
| 11 | | | | | | | | | 77/121 88/121 | 72/132 86/132 |
| 12 | | | | | | | | | | 96/144 84/144 |

For relatively larger sample sizes, the approximate **critical value** $D_\alpha$ is given by the equation

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

The coefficient ***c(α)*** for typical values of **α** is:

| α | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|---|
| c(α) | 1.22 | 1.36 | 1.48 | 1.63 | 1.73 | 1.95 |

## Exemplo de um teste K-S:

Uma empresa de SW está a analisar falhas em projetos durante os 3 últimos anos. O objetivo é avaliar as falhas em 2 sucursais da empresa (Coimbra e Lisboa) e se elas têm diferença, com 95% de confiança:

| 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** |

As hipóteses para o problema são:

- **H0:** as duas frequências acumulativas são semelhantes, ou seja, a taxa de falha entre as duas sucursais é semelhante.
- **H1:** as duas frequências são diferentes, ou seja, a taxa de falha é diferente (existe uma com maior taxa de falha do que a outra).

Ora para uma significância de 95%, os valores críticos são dados pelas seguintes fórmulas:

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

Coefficient ***c(α)*** for $\alpha = 0.05$

**Critical value**
(If the difference between the two cumulative distributions is higher than this → reject $H_0$)

O teste estatístico é dado pela seguinte fórmula:

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

Para o valor crítico de 0.29, e o valor da estatística de teste de 0.082, a conclusão é que não podemos rejeitar H0, ou seja, não podemos dizer que a taxa de projetos falhados têm alguma diferença.

## Teste de Normalidade:

O teste de KS também tem a particularidade de poder testar a normalidade de uma sample de dados (especialmente quando nos referimos ao T-Test, em que necessitamos obrigatoriamente de a testar).

*Adicionalmente, temos outros tipos de testes…*

## Bootstrapping:

Técnica estatística de reamostragem que é utilizada para estimar a precisão de estatísticas amostrais, como médias, medianas, desvios padrão e intervalos de confiança, através da geração de múltiplos conjuntos de dados a partir da amostra original.

A ideia é simular a aleatoriedade do processo de amostragem gerando múltiplas amostras a partir da amostra original.

A reamostragem é feita "com reposição", o que significa que um mesmo ponto de dados pode ser escolhido mais de uma vez em uma única amostra simulada. Esse processo cria múltiplas amostras simuladas, cada uma com variações leves em relação à amostra original.

## Testes de Permutação:

Num teste de permutação, a distribuição da estatística de teste sob a hipótese nula (H0) é obtida considerando todas as possíveis combinações ou permutações das etiquetas nos data points observados. Vamos baralhar todos os dados, sem qualquer consideração prévia, e calcular para cada alteração a estatística de teste de interesse, vendo se uma qualquer medida irá fazer diferença. Depois fazemos uma distribuição com esses dados (de todas as estatísticas de teste). Essa distribuição representa a distribuição da estatística de teste sob a suposição de que a hipótese nula é verdadeira. Comparamos por fim, a estatística de teste observada nos dados originais com a distribuição obtida a partir das permutações.

# PPT 11/T/ MEI

| | |
|---|---|
| ↗ Course | MEI |
| Aula | T |
| # Aula Nº | 11 |

## Experiências com pessoas

Quais são os usos das experiências em que as pessoas sejam intervenientes?

- Engenharia de Software Empírica
  - Como as pessoas projetam sistemas?
  - Quais são os procedimentos de engenharia de software eficazes?
- Testes de Usabilidade em Interação Humano-Computador (HCI)
- Perceção do Utilizador sobre a Utilização e Desempenho do Sistema
- Avaliação de Segurança
- Avaliação do Mercado de Produtos de Software/Computador
- Qual é a técnica que melhor funciona?
- Variação entre a experiência de diferentes grupos.

## Técnicas experimentais:

- Observação e análise de dados: como é que os utilizadores se comportam sozinhos (consiste em monitorização de grandes quantidade de dados).
- Experiências controladas: como é que os utilizadores realizam certas tarefas.
- Entrevistas e questionários: tentar perceber como é que os utilizadores percecionam algo

### Observação comportamental do utilizador:

Observação detalhada de um pequeno número de pessoas, durante o seu trabalho. Tentar perceber quais são os seus problemas, necessidades, criadores de valor, etc.

Para melhor perceber isto podemos utilizar experiências que estejam relacionadas com o comportamento do utilizador já que o nosso objetivo é tentar verificar o que é que lhes interessa num sistema, a sua performance durante a utilização do mesmo e por último, o seu comportamento.

Um fator importante é correlacionar certas métricas que podemos ir retirando durante a experiência (keystrokes, performance, tempo) com o seu comportamento.

### Experiências controladas com utilizadores:

Fazer um estudo experimental formal, definindo o sistema e os fatores de avaliação, criar tarefas, definir medidas e observações a retirar, arranjar quem seja objeto de estudo, vê-los a tentar completar as tarefas, etc.

A ideia é medir a sua performance em diferentes tarefas, analisar em vídeo numa fase mais tardia a sua performance, e coletar (de uma forma muito geral) a experiência subjetiva do utilizador (questões sobre a experiência, utilizando escalas de Likert ou Diferencial Semântica).

Ter sempre em conta que o que é importante numa experiência deste tipo é o bom senso: não fazer demasiadas tarefas, não desgastar o utilizador, fazer uma avaliação imparcial, testar antes as tarefas e a experiência várias vezes, etc.

Adicionalmente, também teremos de escolher os candidatos, utilizando processos de recrutamento screening.

### Entrevistas e Questionários:

Não há grande factos a adicionar, mas devemos ter em atenção o bias com que fazemos as entrevistas, de forma a que tenhamos utilizadores que sejam representativos dos diferentes níveis que possamos ter definido.