LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

Aula 7: *As diferenças de produtividade entre departamentos são estatísticamente significativas?*

Formulação e Teste de Hipóteses

Docente: Amílcar Moreira

Data & Hora: 10/11/2019, 20:30-22:30

Local: Edifício F2, Sala 111

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Na Aulas Anteriores**
  - **Exploramos as bases da estatísticas descritiva: análise univariada e bivariada (i.e. análise de relações entre variáveis)**
- **Objetivos da Aula**
  - **Parte Teórica**
    - **Perceber a diferença entre Estatística Descritiva e Estatística Inferencial**
    - **Perceber o que é uma amostra probabilística e que tipos de técnicas de amostragem existem**
    - **Perceber quais as características de uma Distribuição Normal**
    - **Perceber o papel da Teoria do Limite Central enquanto fundamento da Estatística Inferencial**
    - **Perceber o que é o Intervalo de Confiança, para que serve, e como é calculado**
  - **Parte Prática**
    - **Produzir o Intervalo de Confiança de um Média e de uma Proporção**

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Estatística Descritiva**
  - **Dá-nos as ferramentas para descrever dados de uma (ou mais variáveis) numa amostra**
    - **Medidas de tendência central (médias, modas, etc.)**
    - **Distribuição de frequências (proporções, percentagens, etc.)**
    - **Medidas de dispersão (variância, desvio padrão, etc.)**
  - **Dá-nos as ferramentas para descrever a relação entre variáveis dados de uma (ou mais variáveis) numa amostra**
    - **Medidas de Associação e Correlação**

- **Estatística Inferencial**
  - **Dá-nos as ferramentas para avaliarmos se a forma como os dados estão distribuídos, ou se a relação entre variáveis na amostra, podem ser inferidos para a população**
    - **Intervalos de Confiança**
    - **Testes de Hipóteses**

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **A possibilidade de inferir de uma amostra para uma população depende de duas condições fundamentais:**

  I. **Que amostra seja probabilística**

  II. **Que haja uma forma de demonstrar que a distribuição da amostra segue uma distribuição normal**

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Algumas definições importantes**
- **População**
  - **Conjunto de indivíduos, ou outras entidades, que pretendemos estudar.**
- **Base de Amostragem**
  - **Lista de todas as unidades da população de interesse a partir da qual a amostra será extraída (ex. lista de números de telefone).**
- **Amostra**
  - **Segmento da população de interesse que vai fazer parte do estudo.**
- **Amostra Probabilística**
  - **Amostra em que cada elemento da população tem igual probabilidade de ser seleccionado, e é seleccionado independentemente dos outros.**
- **Amostra Não-Probabilística**
  - **Amostra que não é escolhida segundo médodos probabilisticos.**

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Técnicas de Amostragem**

NÃO-PROBABILITICAS	
Por Conveniência	Os membros da amostra são selecionados em função dos interesses do investigador e da facilidade de acesso aos entrevistados.
Por Quotas	Os membros da amostra são selecionados (a partir da base amostral) de modo a que a amostra possa possa reflectir a composição da população de interesse por referência a um conjunto de categorias (género, idade, etc.).
Bola de Neve	Selecciona-se um conjunto de inquiridos de forma aleatória, a quem é depois pedido que indique alguém na população de interesse que possa responder. (O processo de selecção de entrevistados pára quando a adição de novos entrevistas não adiciona mais dados de relevo.)

- **Técnicas de Amostragem**

PROBABILITICAS	
Aleatória Simples	Os membros da amostra são selecionados de forma aleatória (ex. sorteio, Tabela de Números Aleatórios, data de nascimento, etc.) da base amostral.
Sistemática	Os membros da base amostral são ordenados de acordo com uma tabela de números aleatórios. É seleccionado, de forma aleatória, um membro da base amostral. Os restantes membros da amostra são escolhidos em função do seu número de identificação usando o seguinte critério (fracção da amostragem): Nº + i Em que i=N/n i, fração de amostragem N, total da população n, tamanho da amostra

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Técnicas de Amostragem**

PROBABILITICAS	
Estratificada	Primeiro divide-se a base amostral num conjunto de sub-grupos (estratos), mutuamente exclusivos (um membro da população só pode pertencer a um estrato) e exaustivos (nenhum membro da população é omitido). Exemplos de categorias de estratificação: características demográficas, tipo de empresa, tipo de sector económico, etc. Os membros de cada estrato são depois seleccionados de forma aleatória.
Por Clusters	A base de amostragem é divida em clusters (Unidades Primárias de Amostragem), formados em função dos interesses do investigador. O investigador pode optar por incluir todos os clusters (Amostragem por Clusters em Um Passo), ou apenas uma fracção, que é seleccionada de forma aleatória (Amostragem por Clusters em Dois Passos). Dentro de cada cluster, seleciona-se de forma aleatória os membros (Unidades Secundárias de Amostragem) a incluir na amostra.

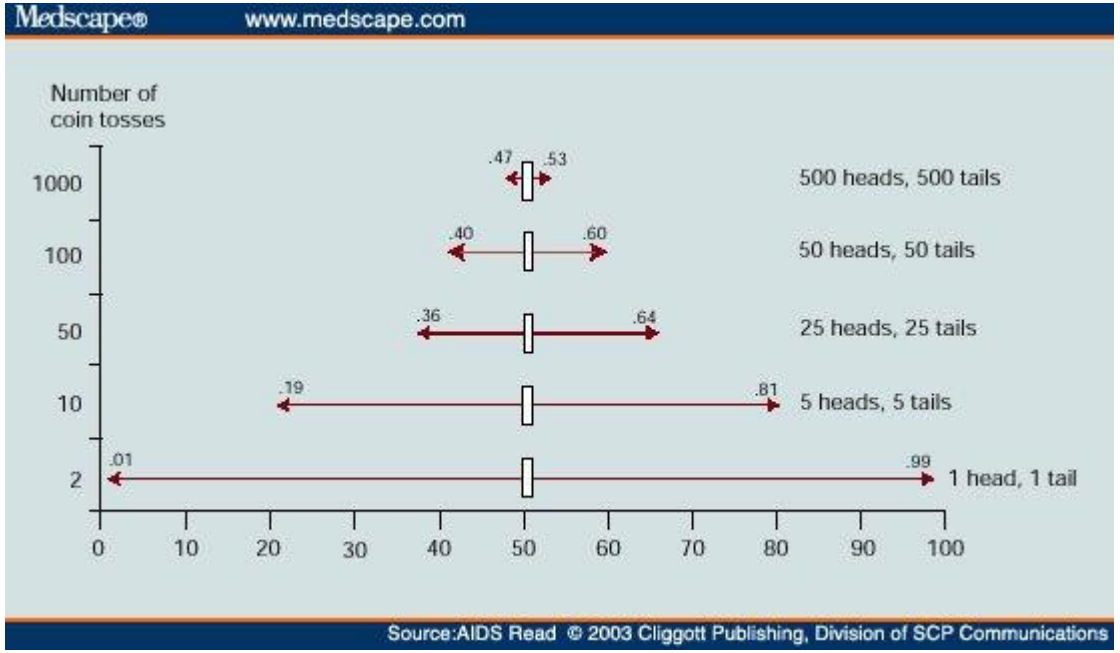
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Por que é que o tamanho da amostra é importante?**

  - **Quanto maior for o tamanho da amostra, Quanto maior a amostra, menor é a amplitude do intervalo de confiança – o que significa, maior precisão das nossas estimativas**

  - **Quanto maior for o tamanho da amostra, maior será a 'potência estatística' do estudo, que mede a probabilidade de encontrar um efeito estatístico que existe na realidade (evitando Erros de Tipo II)**



Fonte: http://gosu.talentrank.co/confidence-interval-and-sample-size/

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Como se calcula o tamanho da amostra?**

$$Tamanho\ da\ Amostra: \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$

Em que:

N: **População (Total)**

p: **Proporção da amostra (se desconhecida, assume-se 0.5)**

z **: z-score (se Intervalo de Confiança a 95% = 1.96; se a 99% = 2.57)**

e **: Margem de erro (se Intervalo de Confiança a 95% = 0.05; se a 99% = 0.01)**

Fonte: https://www.surveymonkey.com/mp/sample-size-calculator/

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **A possibilidade de inferir de uma amostra para uma população depende de duas condições fundamentais:**

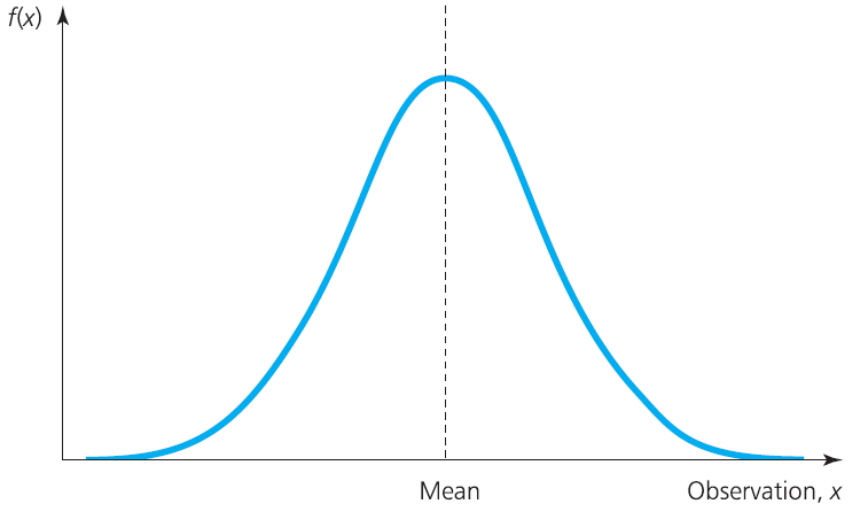
  I. **Que amostra seja probabilistica**

  II. **Que haja uma forma de demonstrar que a distribuição da amostra segue uma distribuição normal**

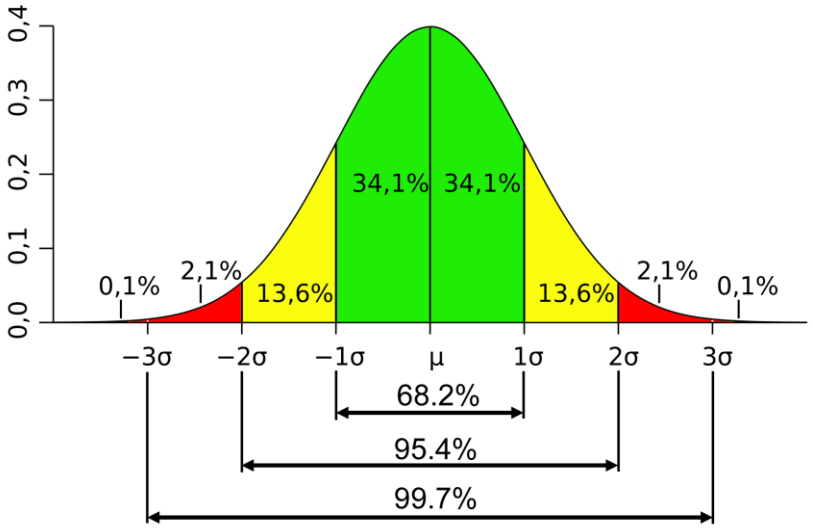
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **O que é uma Distribuição Normal (ou Curva de Gauss)?**
  - **Média = Mediana = Moda**
  - **Simétrica**
  - **Distribuição segue a regra dos 3 Sigmas**
    - **34,1% das observações da variável estão dentro de um desvio-padrão da média**
    - **68,2% das observações da variável estão dentro de (+ / -) um desvio-padrão da média**
    - **95,4% das observações da variável estão dentro de (+ / -) 2 desvio-padrão da média**
    - **99,7% das observações da variável estão dentro de (+ / -) 3 desvio-padrão da média**



Fonte: Waters, 2011: 354



Fonte:http://en.wikipedia.org/wiki/File: Standard_deviation_diagram.svg

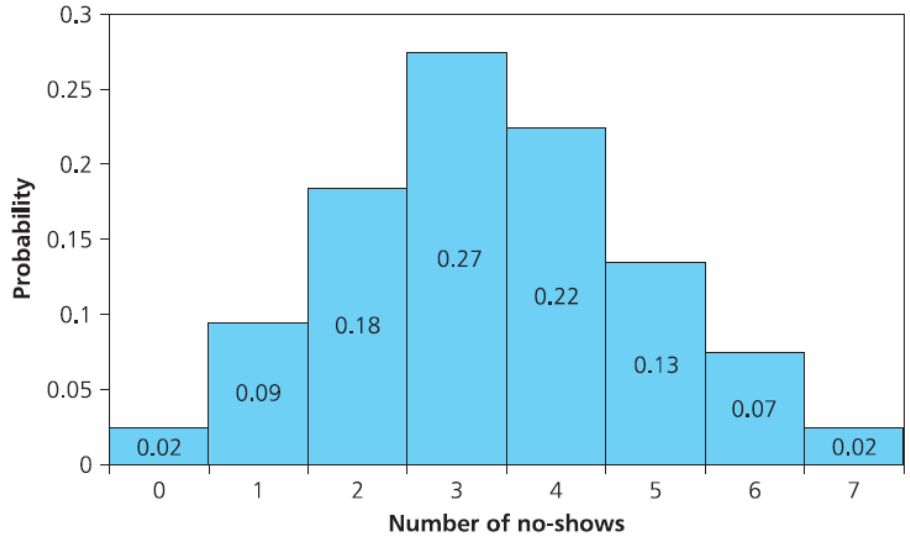
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **A curva normal exprime a distribuição de frequências numa variável…**

- **… mas também a distribuição de probabilidades ***

$$P = \frac{\textit{número de vezes em que re segistou um evento}}{\textit{número de observações}}$$



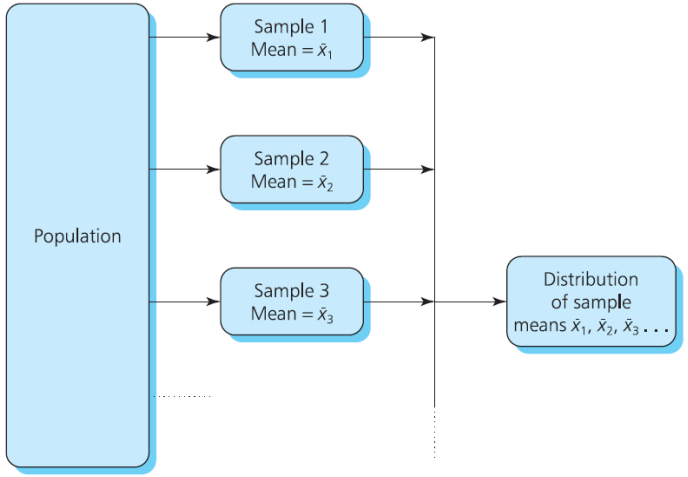
Fonte: Waters, 2011: 337

*** Esta ideia é importante para percebermos como são calculados os intervalos de confiança**

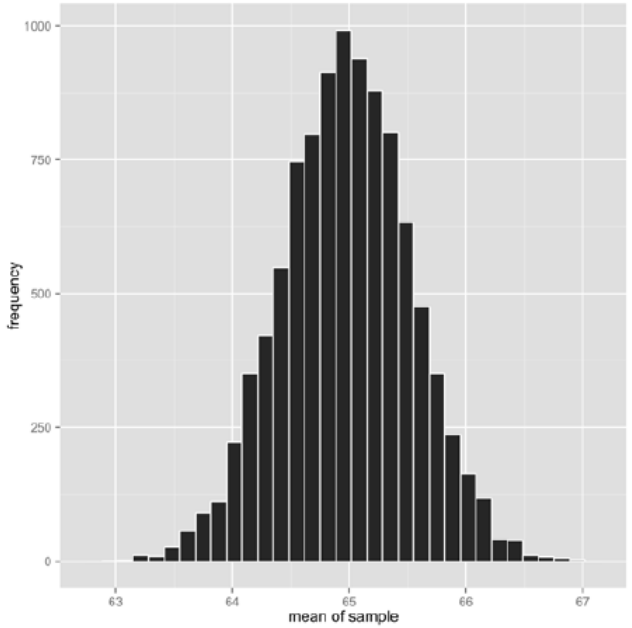
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **O Teorema do Limite Central sugere um conjunto de propriedades da Distribuição Amostral da Média que nos permite fazer a inferência estatística de uma amostra para uma população**



- **O que é a 'Distribuição Amostral Da Média'?**
  - **Uma dada população pode dar origem a um número de amostras**
  - **Cada amostra terá uma dada média (a chamada média amostral)**
  - **À forma como se distribuem as médias destas amostras chamamos 'Distribuição Amostral Da Média'**



- **O que diz o Teorema do Limite Central**
  - **Quando uma amostra é ≥ 30**
    - **a Distribuição Amostral da Média tende para a uma distribuição normal**
    - **O Desvio-Padrão da Distribuição Amostral da Média é o produto do seguinte rácio:**

$$\frac{\sigma}{\sqrt{N}}$$

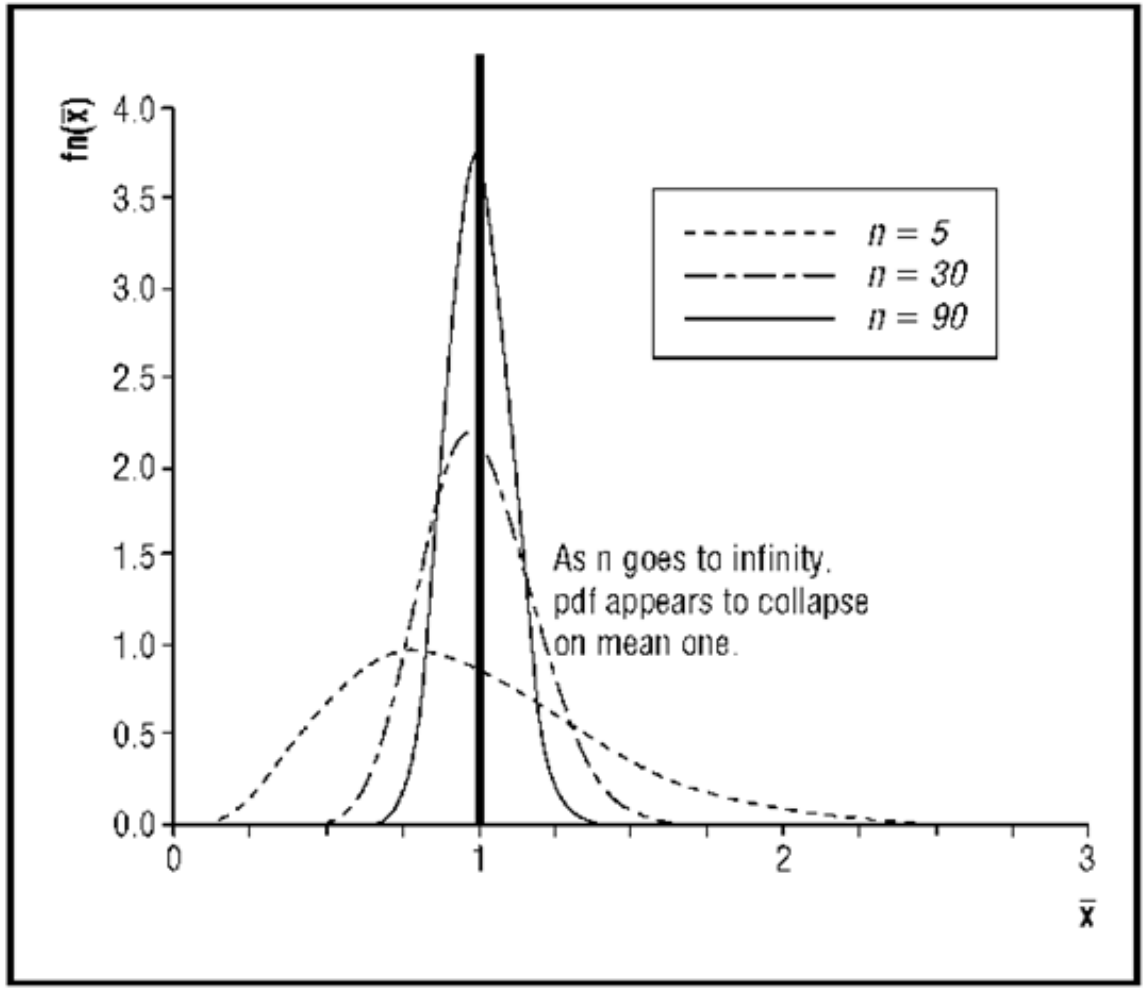
    - σ**: Desvio-Padrão da População**
    - $\sqrt{N}$**: Raiz quadrada do número de observações da amostra**

- **O que é que isto significa?**

  - **Quando uma amostra é ≥ 30, a média das médias amostrais tende para média populacional**

  - **A média da nossa amostra pode ser considerada uma aproximação da média da população**

  - **Quanto mais aumenta o tamanho da amostra, menor é o desvio-padrão da distribuição amostral da média, i.e. menor é a probabilidade de erro na amostra**



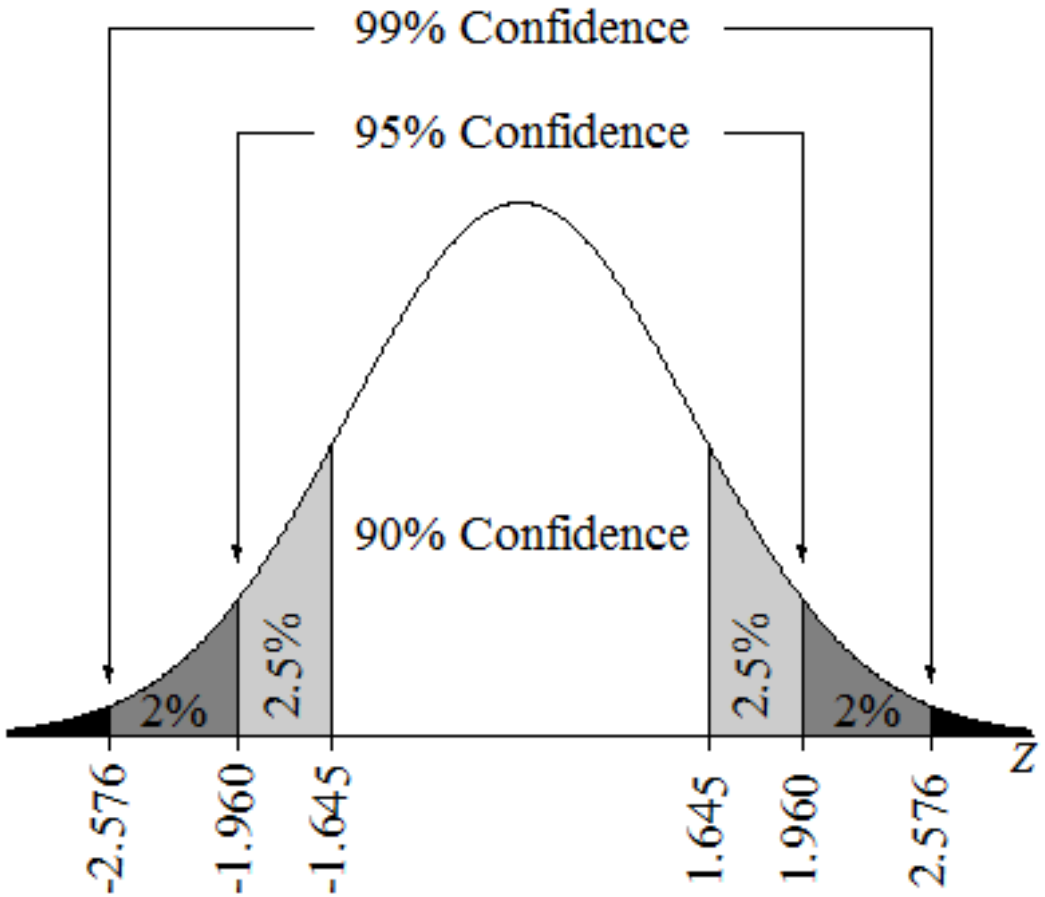
Fonte: http://what-when-how.com/social-sciences/law-of-large-numbers-social-science/

Com base no Teorema do Limite Central , podemos calcular uma estatística que nos permite aferir até que ponto a média da nossa amostra é uma boa estimativa da média da população:

Intervalo de Confiança

- **O que é o Intervalo de Confiança?**
  - **É o Intervalo de Valores (*CI*) dentro do qual se estima que a média se situe na população**

- **O que é o Grau de Confiança?**
  - **Probabilidade de o intervalo de confiança capturar o parâmetro (neste caso a média) da população**
  - **Por norma, adota-se um Grau de Confiança de 95%**
    - **Se quisermos, podemos adotar um Grau de Confiança maior (99%)…**
    - **ou menor (90%)**
  - **Interpretação:**
    - ***Ex*: Intervalo de Confiança com um Grau de Confiança a 95%**
    - **Se fizéssemos 100 inquéritos, em 95% dos casos o intervalo de confiança iria conter a média da população**



Fonte: https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/estimate-the-difference-between-population-proportions-2-of-3/

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **O que é o Grau de Confiança?**

  - **De notar que, associado a um determinado Grau de Confiança, temos sempre um determinado valor critico (*z*), baseado no Erro-Padrão**

  - **Estes valores são usados para calcular a amplitude do Intervalo de Confiança…**

Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

Fonte: http://www.biochemia-medica.com/en/journal/18/2/10.11613/BM.2008.015

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Como se calcula o Intervalo de Confiança?**

MARGEM DE ERRO

$$CI_x^{99} = \bar{x} \pm \{z^{99} \times SE\}$$

CI_x^{99} **Intervalo, com um Grau de Confiança a 99%**

$\bar{x}$ **Média da variável x**

z^{99} **Valor crítico para um Grau de Confiança a 99%**

SE **Erro-Padrão**

- **Como se calcula o Intervalo de Confiança?**

<u>Isto é equivalente a</u>

MARGEM DE ERRO

$$CI_x^{99} = \bar{x} \pm \left\{ z^{99} \times \frac{SD}{\sqrt{n}} \right\}$$

SD **Desvio-Padrão para um Grau de Confiança a 99%**

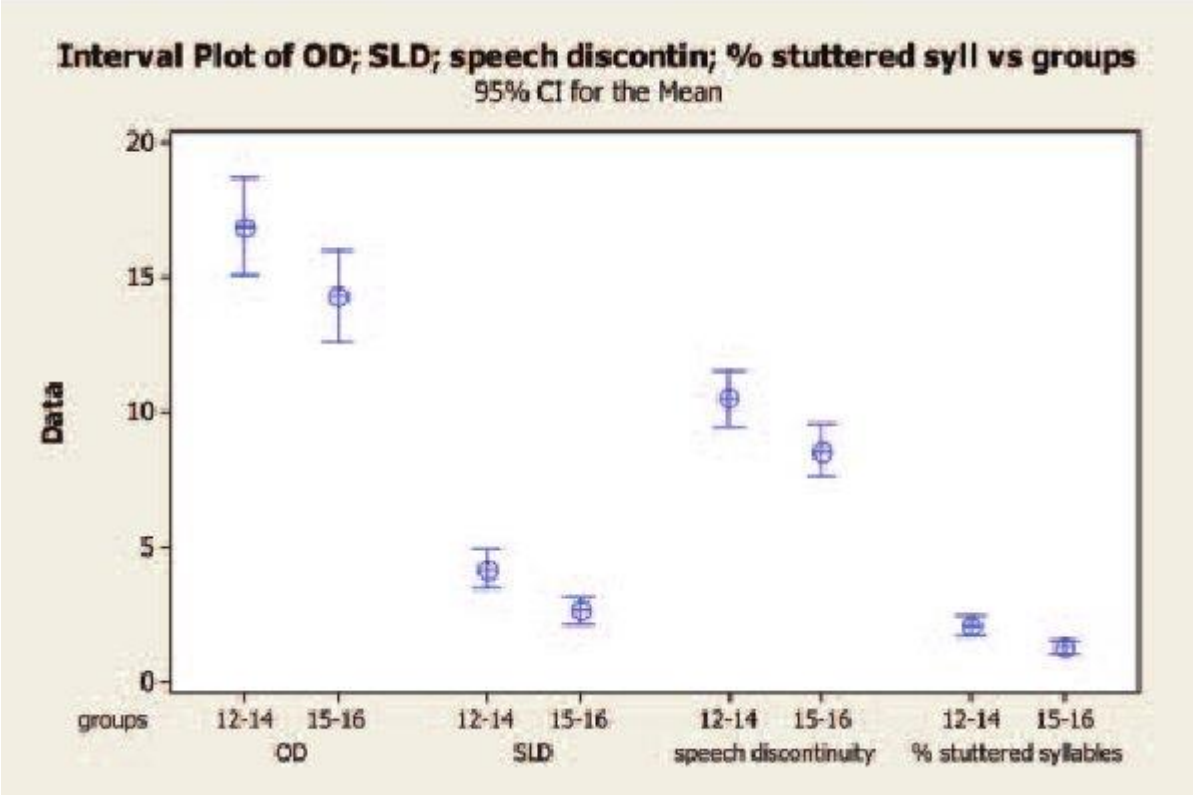
CI_x^{99} **Intervalo, com um Grau de Confiança a 99%**

$\bar{x}$ **Média da amostra**

z^{99} **Valor crítico para um Grau de Confiança a 99%**

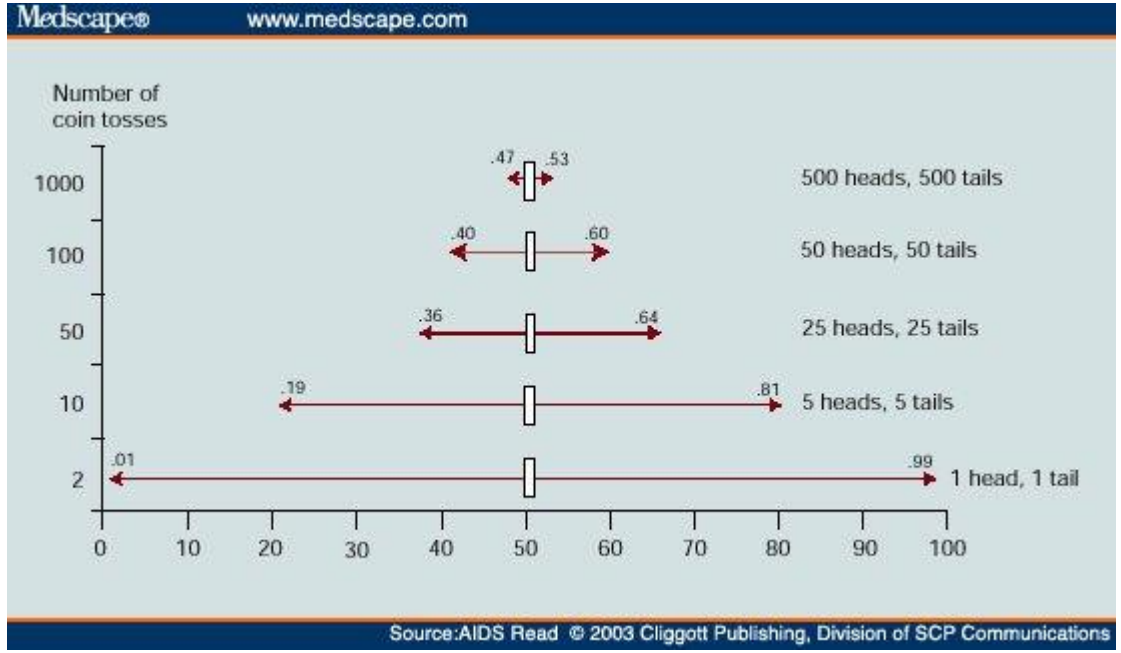
$\sqrt{n}$ **Raiz-quadrada do número de observações da amostra**

- **O que nos diz o Intervalo de Confiança?**
  - **Grau de precisão da média**
    - **Quanto maior a amplitude do Intervalo de Confiança, menor o grau de precisão**



Fonte: https://www.researchgate.net/publication/5988752_Fluency_variation_in_adolescents/figures?lo=1

- **O que nos diz o Intervalo de Confiança?**
  - **Grau de precisão da média**
    - **Quanto maior a amplitude do Intervalo de Confiança, menor o grau de precisão**
    - **O que afeta a amplitude?**
      - **Quanto maior a amostra, menor é a amplitude**
      - **Quanto maior é o Erro-Padrão, maior é a amplitude**



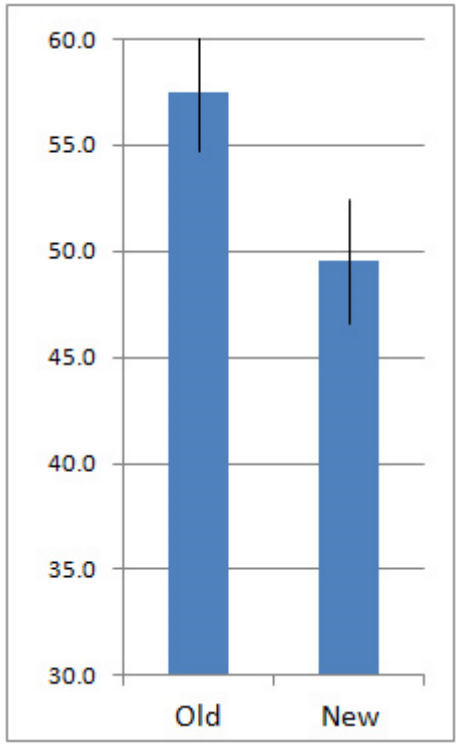
Fonte: http://gosu.talentrank.co/confidence-interval-and-sample-size/

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

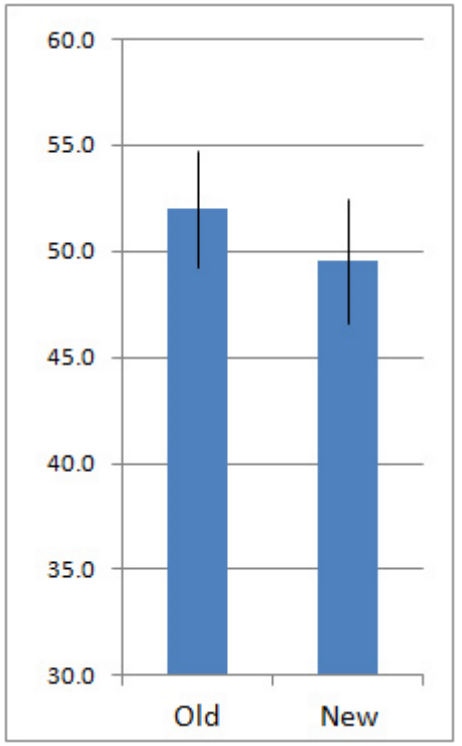
- **O que nos diz o Intervalo de Confiança?**
  - **Significância estatística**

Os intervalos de confiança não se sobrepõem



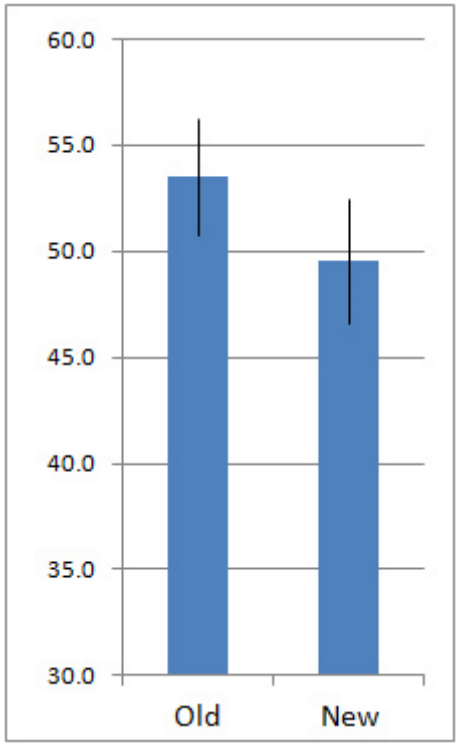
A diferença é estatisticamente significativa

Há uma grande sobreposição entre os intervalos de confiança



A diferença não é estatisticamente significativa

Há alguma sobreposição entre os intervalos de confiança



Mais vale aplicar um teste estatístico

Fonte: https://measuringu.com/ci-10things/

ANÁLISE DE DADOS EM GRH
Aula 7: *Podemos confiar nas sondagens? (Parte III)*

- **Mas, o Intervalo de Confiança só é útil para estimar a média populacional?**

Não…

…. Podemos calcular o Intervalo de Confiança para uma série de estatísticas (proporções, medianas, coeficientes de beta, etc.)

- **Ex: Intervalo de Confiança de uma proporção**

MARGEM DE ERRO

$$CI_p^{99} = \hat{p} \pm \left(z^{99} \times \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}} \right)$$

CI_x^{99}

Intervalo, com um Grau de Confiança a 99%

$\hat{p}$

Proporção da categoria, na amostra

z^{99}

Valor crítico para um Grau de Confiança a 99%

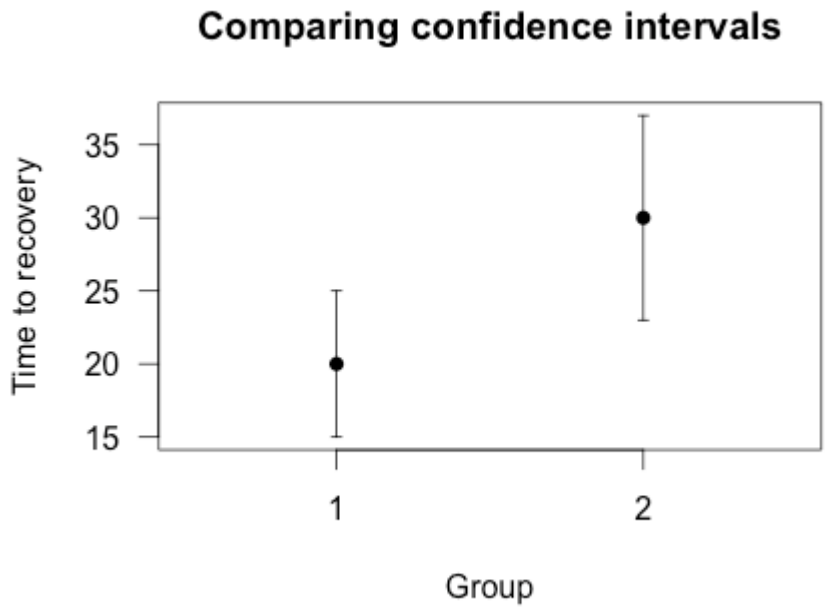
n

Número de observações da amostra

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Como devemos representar graficamente um Intervalo de Confiança?**
  - **Se se tratar de uma média**
    - **Gráfico 'Alto-Baixo'**

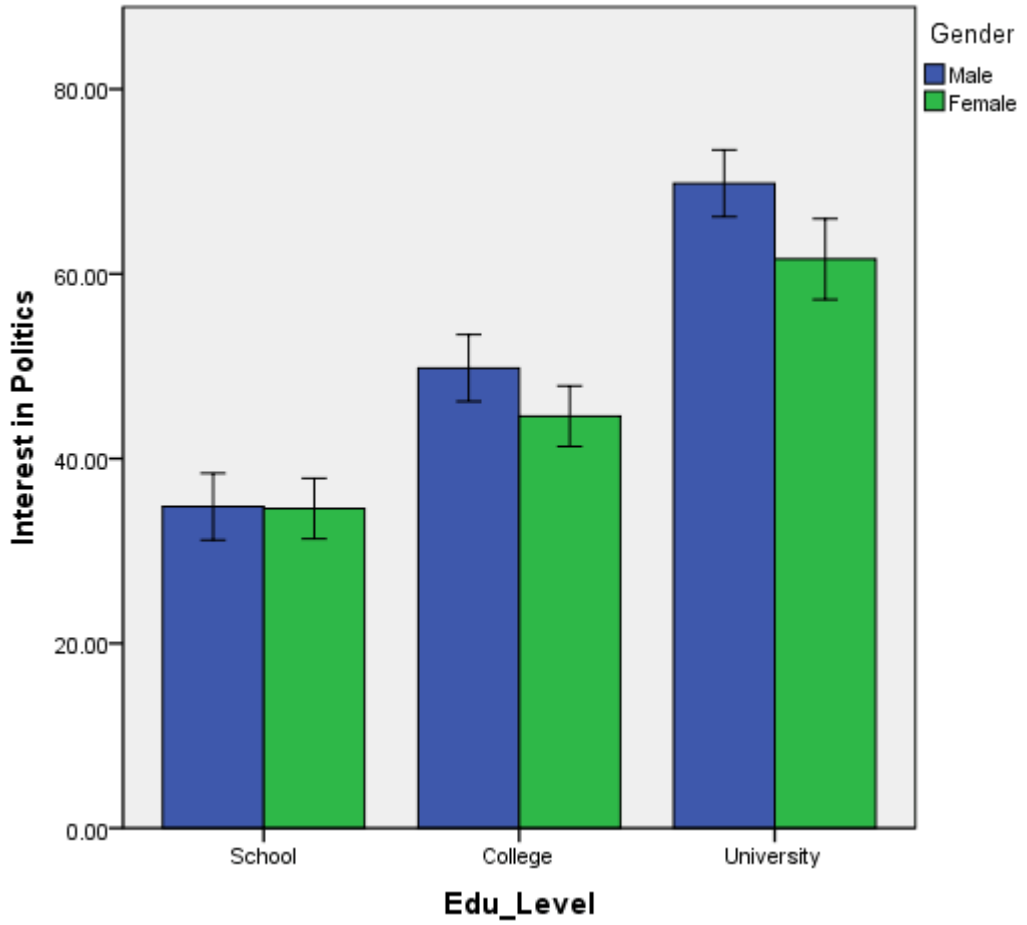


Fonte: https://www.statisticsdonewrong.com/significant-differences.html

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Como devemos representar graficamente um Intervalo de Confiança?**
  - **Se se tratar de uma média**
    - **Gráfico 'Alto-Baixo'**
  - **Se se tratar de uma proporção**
    - **Gráfico de Barras com Intervalo de Confiança**

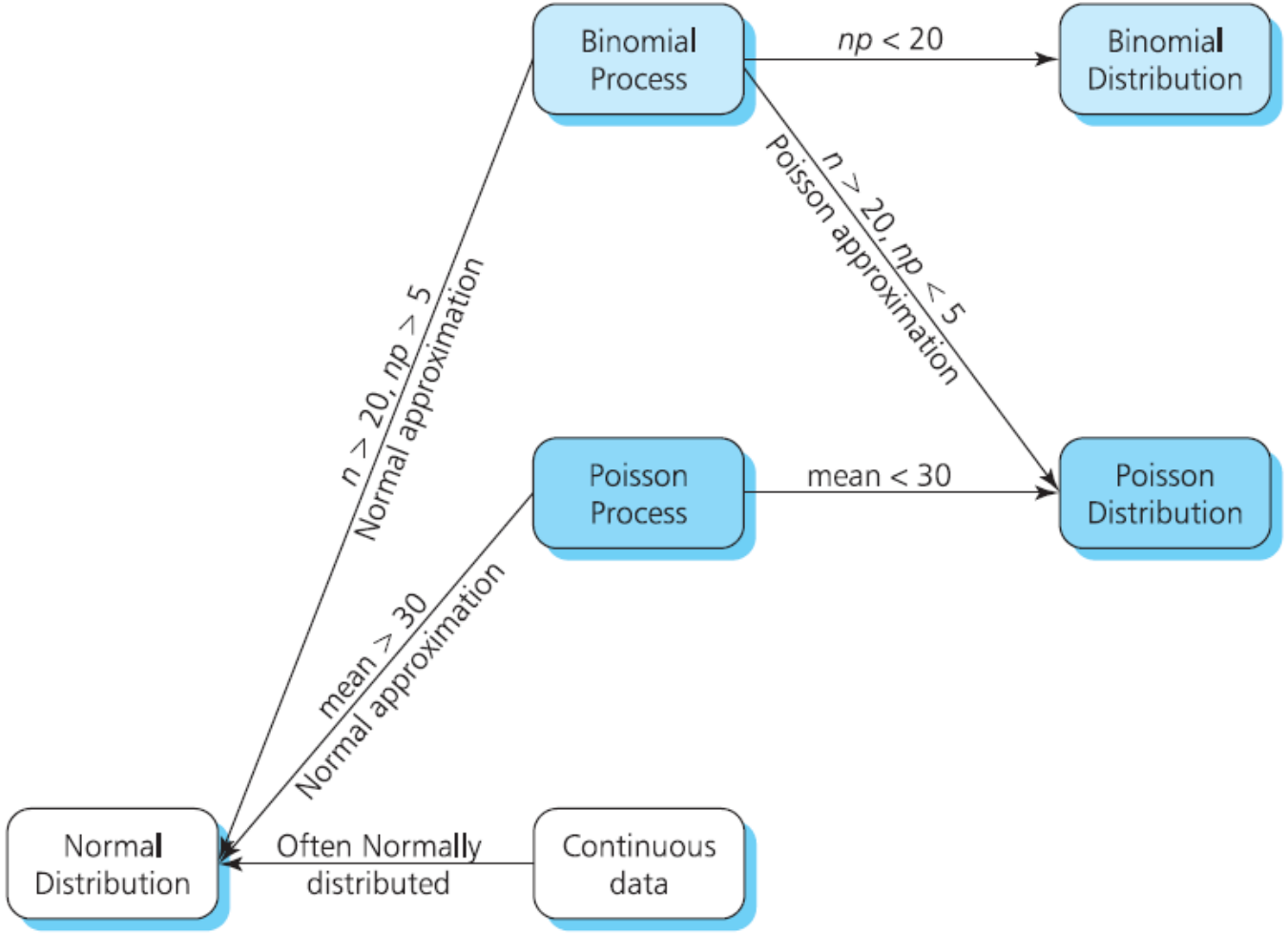


Fonte: https://statistics.laerd.com/spss-tutorials/clustered-bar-chart-using-spss-statistics-2.php

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **E se a população não segue uma distribuição normal?**
- **Em alguns casos, podemos fazer aproximações… isto é tratar certas distribuições como se tratassem de distribuições normais.**
- **Mas isso não é uma preocupação por agora!**



Calcular o Intervalo de Confiança

De uma Média
De uma Proporção

Calcular o Intervalo de Confiança

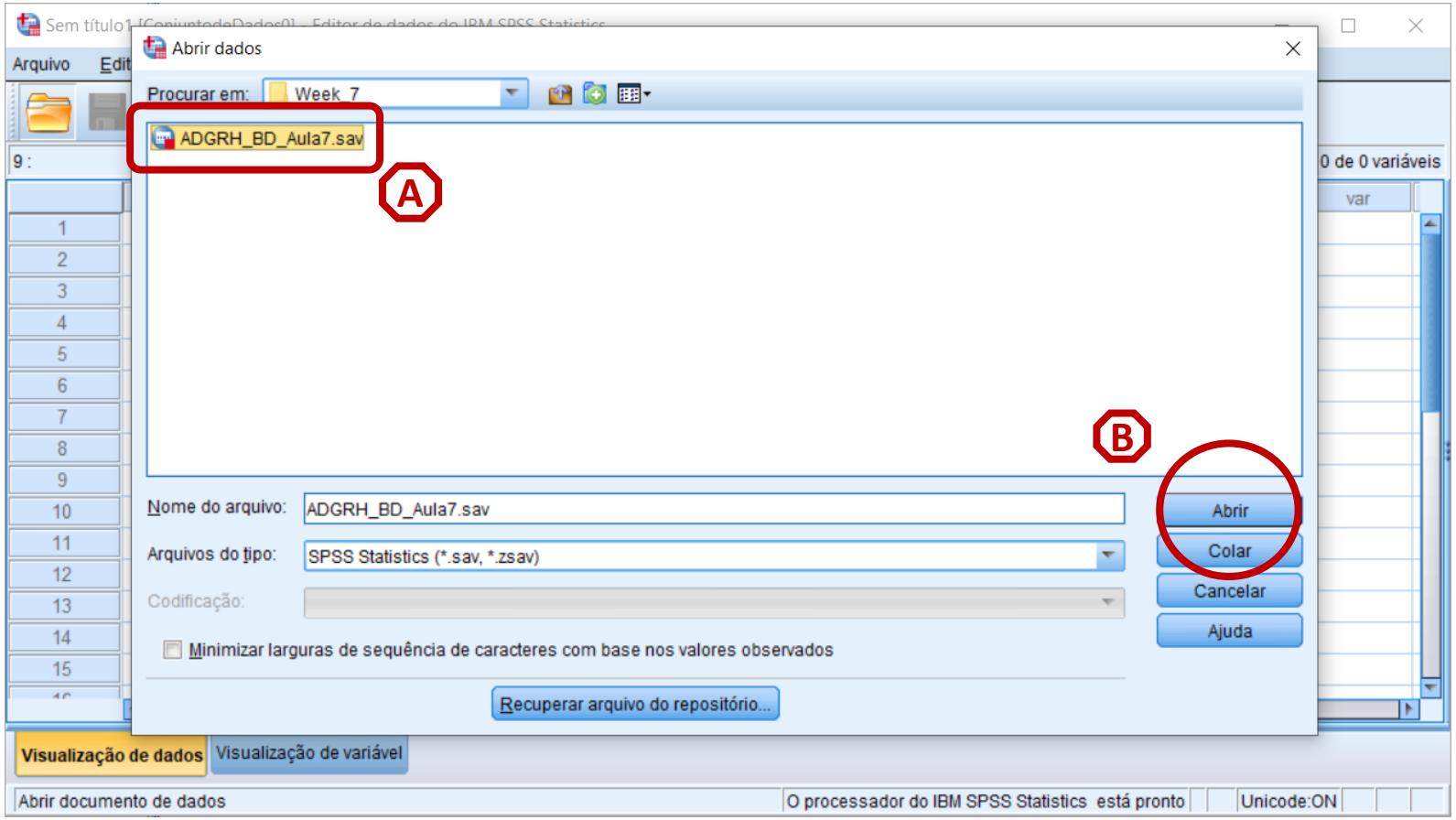
De uma Média

Intervalo de Confiança: Média

- **Objectivo:**
  - **Qual é o intervalo de confiança da variável que mede os salários na empresa (y_wage2)?**

Intervalo de Confiança: Média

- **Selecionar ficheiro 'Database_Tutorial_7.sav'**
- **Selecionar 'Abrir'**

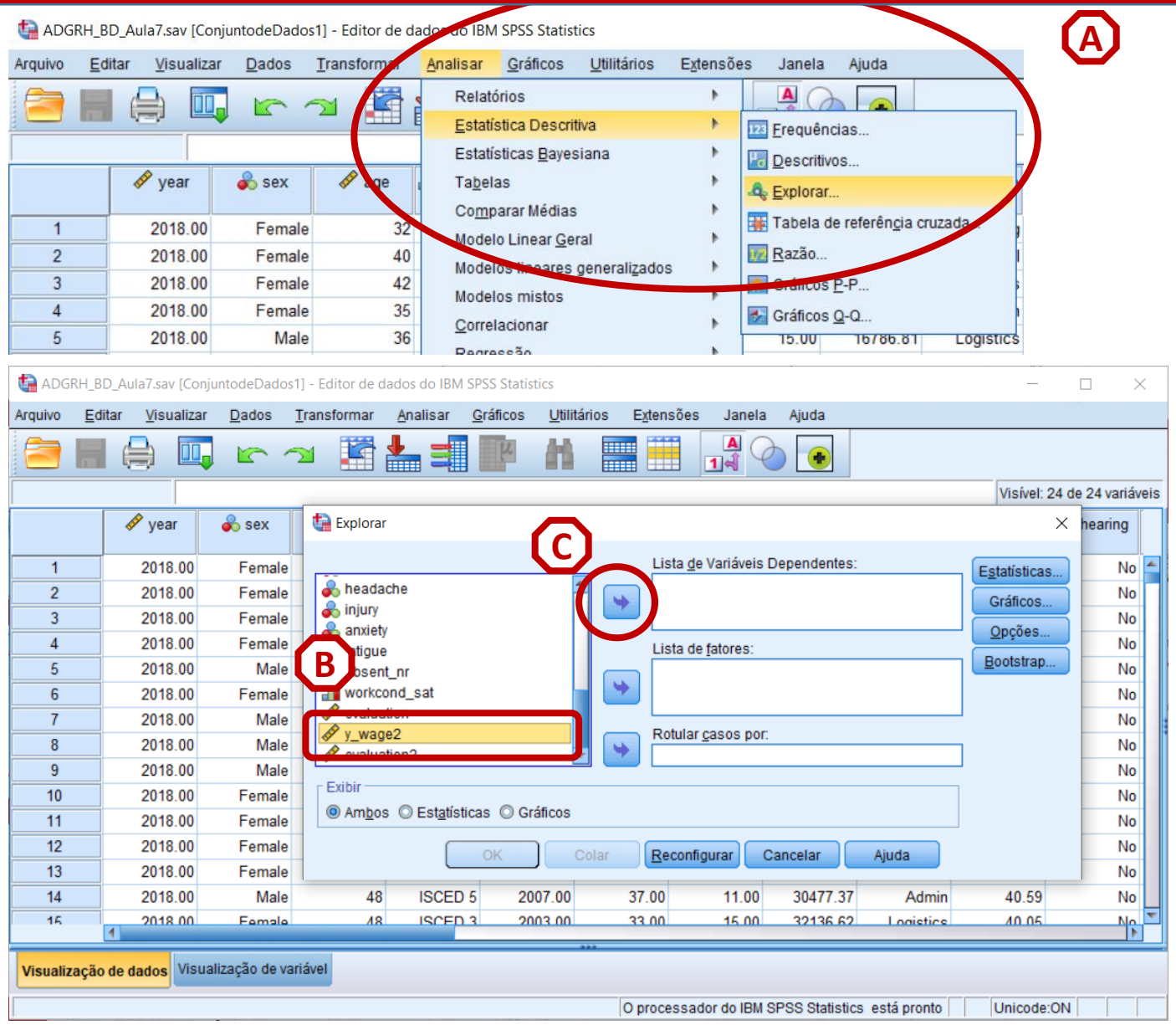


ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

Intervalo de Confiança: Média

- **Selecionar 'Analisar' / 'Estatisticas Descritivas' / 'Explorar'** (A)
- **Selecionar a variável 'y_wage2'** (B)
- **Colocar na caixa 'Lista de Variáveis Dependentes'** (C)

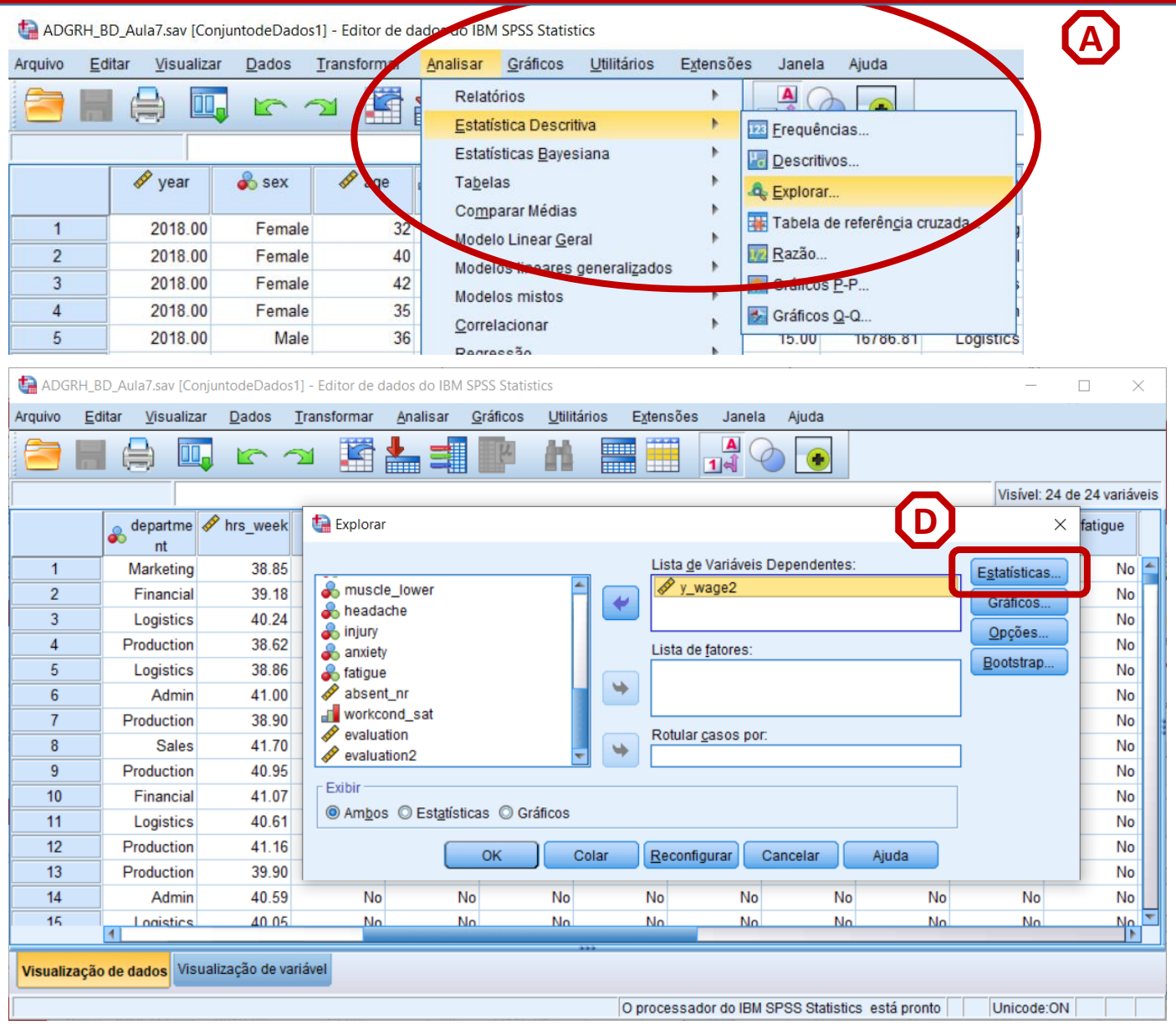


ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

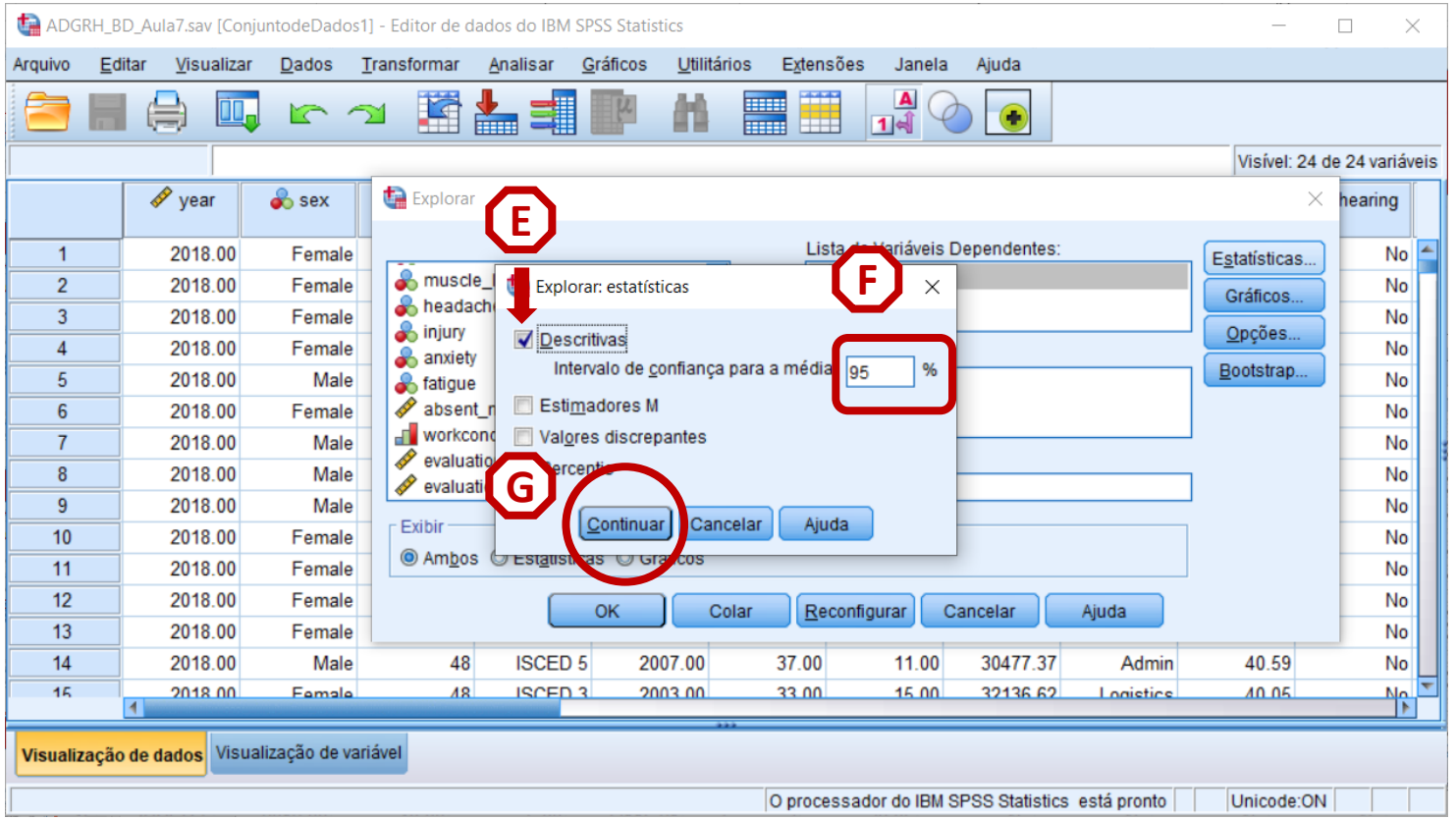
Intervalo de Confiança: Média

- **Selecionar 'Analisar' / 'Estatísticas Descritivas' / 'Explorar'** (A)
- **Selecionar a variável 'y_wage2'** (B)
- **Colocar na caixa 'Lista de Variáveis Dependentes'** (C)
- **Selecionar 'Estatísticas'** (D)



Intervalo de Confiança: Média

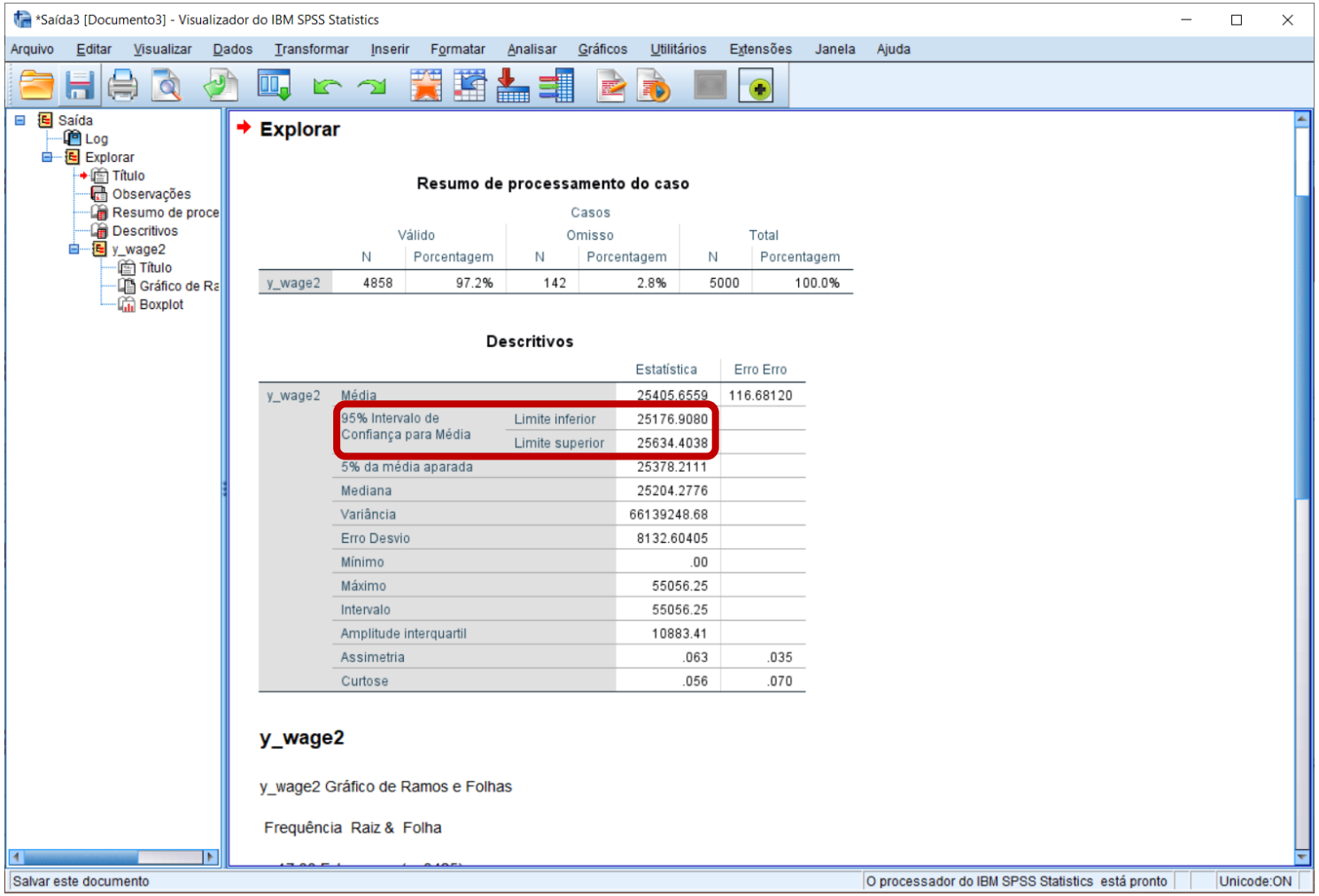
- Selecionar 'Analisar' / 'Estatísticas Descritivas' / 'Explorar' — A
- Selecionar a variável 'y_wage2' — B
- Colocar na caixa 'Lista de Variáveis Dependentes' — C
- Selecionar 'Estatísticas' — D
- Selecionar 'Descritivos' — E
- Definir um Grau de Confiança de '95%' — F
- Selecionar 'Continuar'/OK — G



Intervalo de Confiança: Média

- **O resultado é publicado no 'Visualizador de Resultados'**

PODEMOS DIZER, COM 95% DE CONFIANÇA, QUE O VALOR DO SALÁRIO MÉDIO ANUAL NA POPULAÇÃO ESTÁ ENTRE €25.177 E €25.634.



Calcular o Intervalo de Confiança

De uma Proporção

- **Podemos calcular o Intervalo de Confiança de uma proporção/percentagem?**
- **O SPSS é particularmente limitado**
  - **Regra geral: não é possível calcular (diretamente) o Intervalo de Confiança de uma proporção/percentagem**
  - **Exceções**
    - **Quando fazermos testes de significância estatística**
    - **Para variáveis binomiais (com valores 0 e 1)**
      - **<u>Pressuposto: a média dessa variável representa o proporção de observações com valor 1</u>**

ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Podemos calcular o Intervalo de Confiança de uma proporção/percentagem?**

Em alternativa, o Intervalo de Confiança pode ser calculado à mão…

https://select-statistics.co.uk/calculators/confidence-interval-calculator-population-proportion/

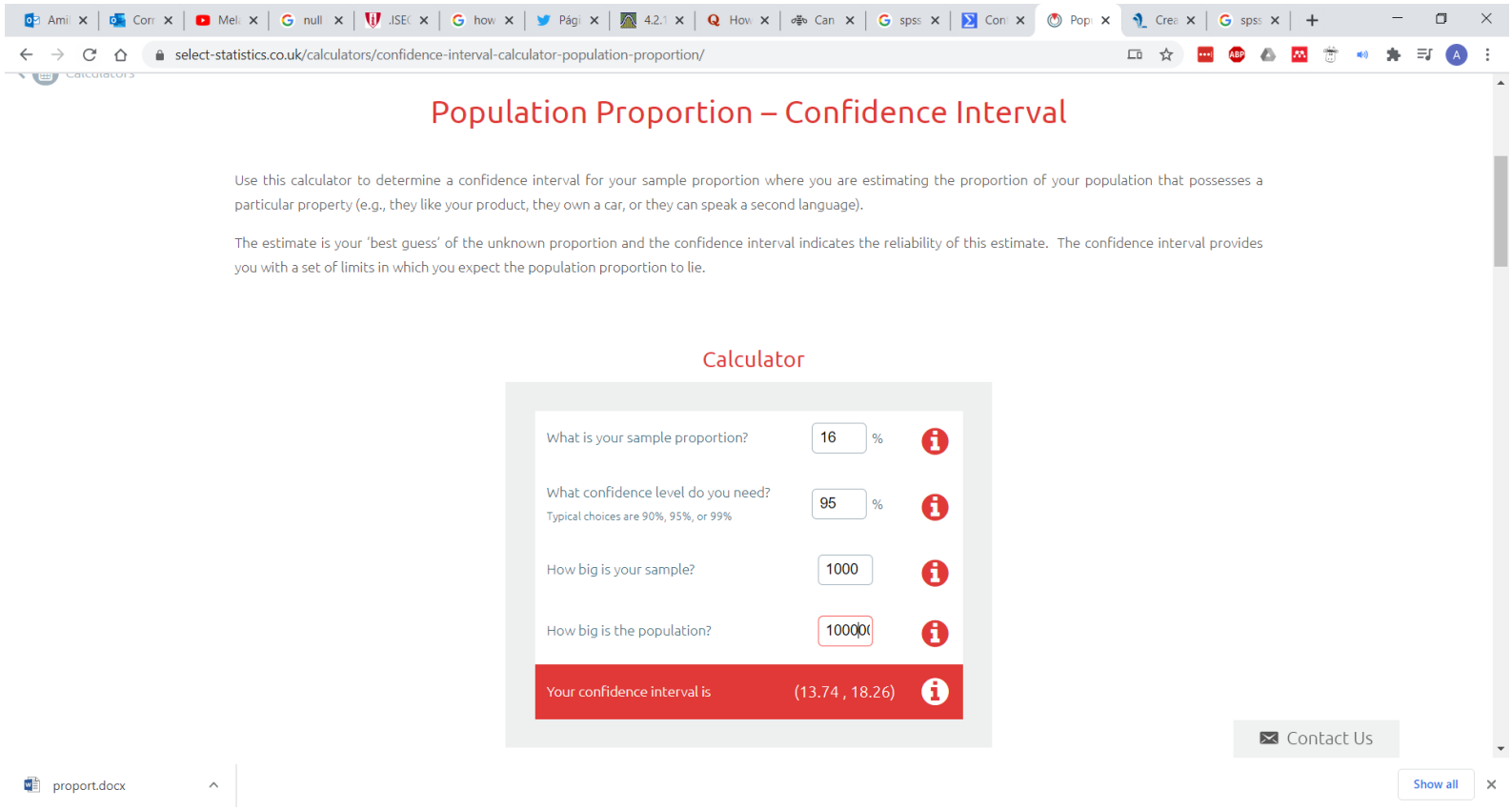
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- **Podemos calcular o Intervalo de Confiança de uma proporção/ percentagem?**

Interpretação:

PODEMOS DIZER, COM 95% DE CONFIANÇA, QUE O A PERCENTAGEM DE 'X' ESTARÁ ENTRE 13.7% E 18,3%..



- **Podemos calcular o Intervalo de Confiança de uma proporção/percentagem?**

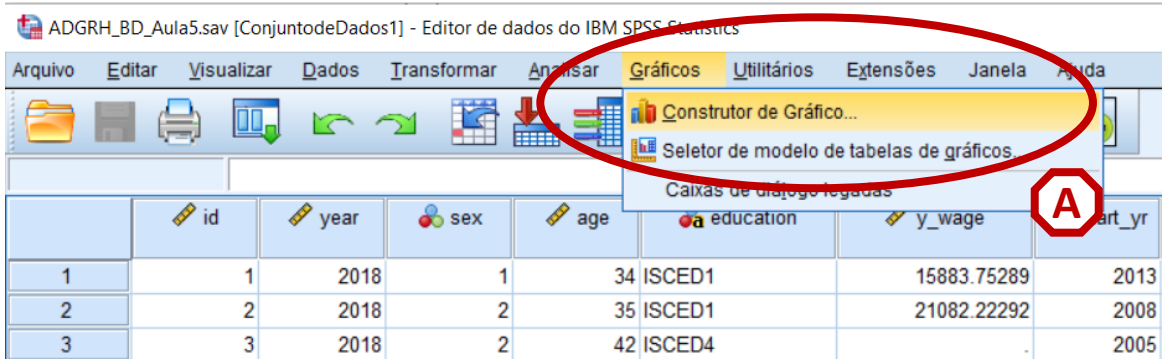
A outra alternativa é representar os intervalos de confiança visualmente

ANÁLISE DE DADOS EM GRH

Aula 5: As Mulheres Recebem o Mesmo que os Homens?

Gráfico de Barras com Erros

- Selecionar 'Gráficos' / 'Construtor de Gráfico' (A)
- Selecionar 'Não mostrar este diálogo novamente' (B)
- Selecionar 'OK' (C)



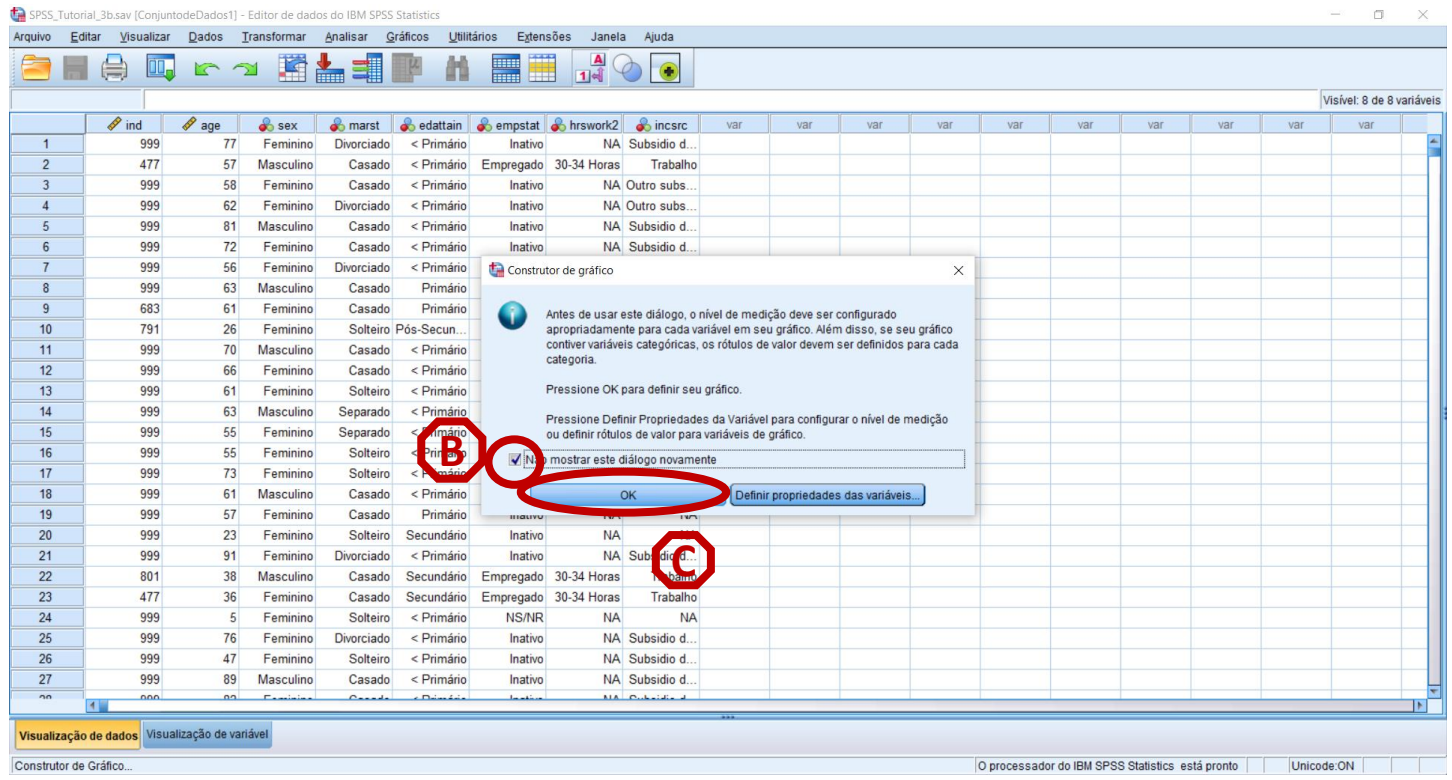


Gráfico de Barras com Erros

- **Selecione 'Barras'** (D)
- **Selecionar (com duplo-clique) o Gráfico de Barras (simples)** (E)

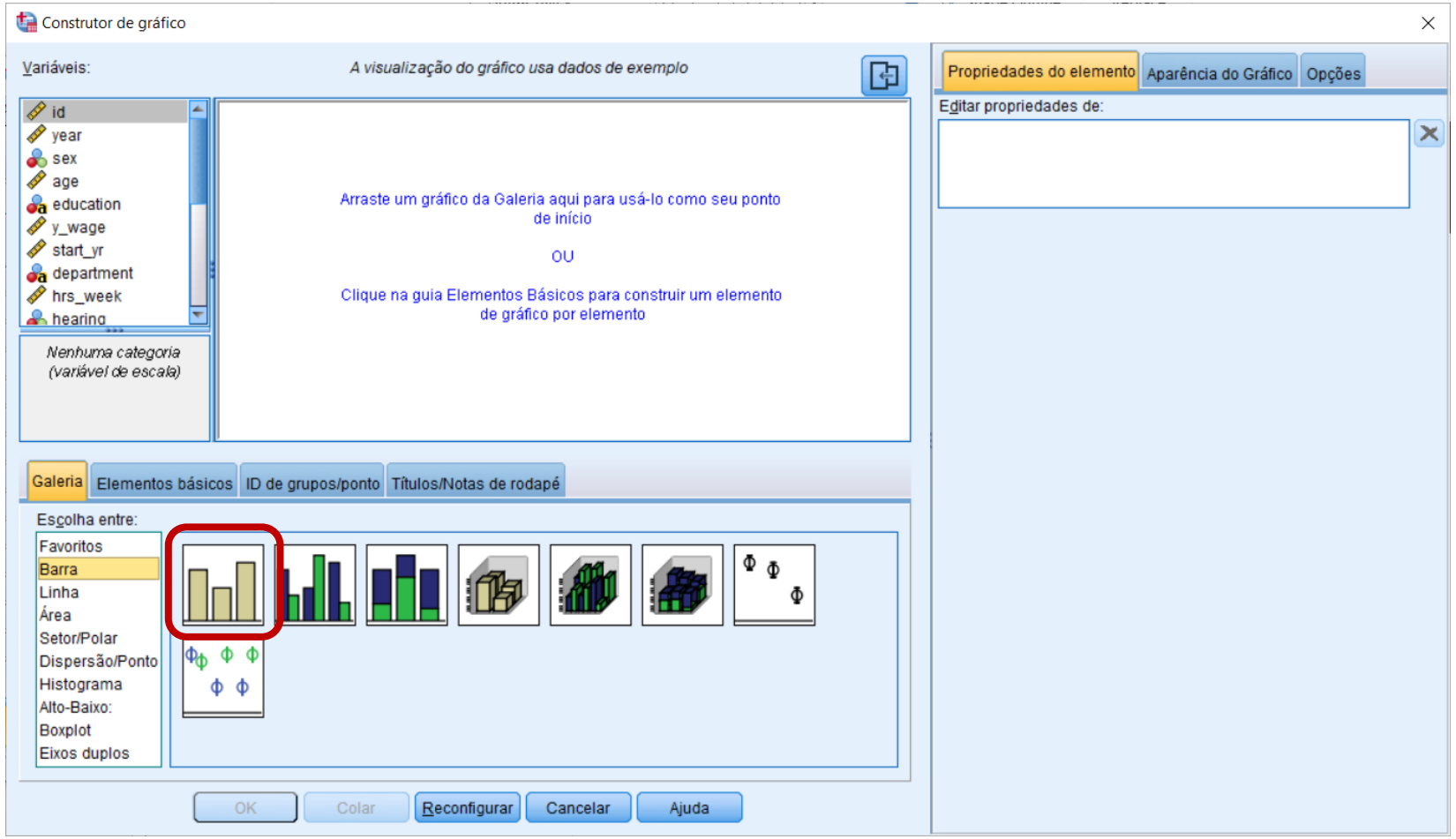


Gráfico de Barras com Erros

- Selecione ‘Barras’ (D)
- Selecionar (com duplo-clique) o Gráfico de Barras (simples) (E)
- Selecionar a variável ‘department’ (F)
- Colocar a variável ‘department’ no ‘Eixo X’ (G)

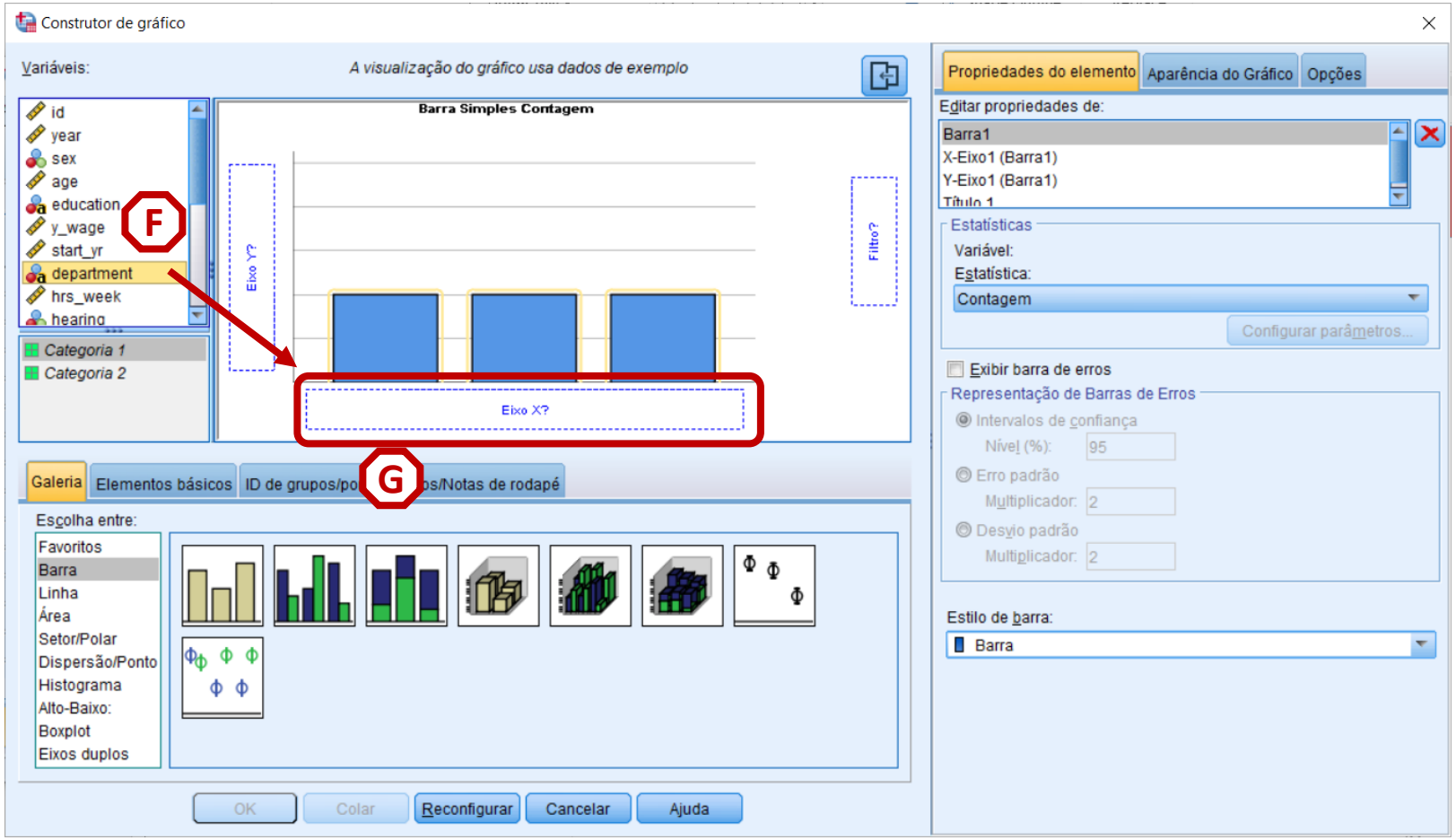


Gráfico de Barras com Erros

- Selecione ‘Barras’
- Selecionar (com duplo-clique) o Gráfico de Barras (simples)
- Selecionar a variável ‘department’
- Colocar a variável ‘department’ no ‘Eixo X’
- Seleccionar ‘Exibir Barra de Erros’
- Selecionar ‘OK’

D E F G H I

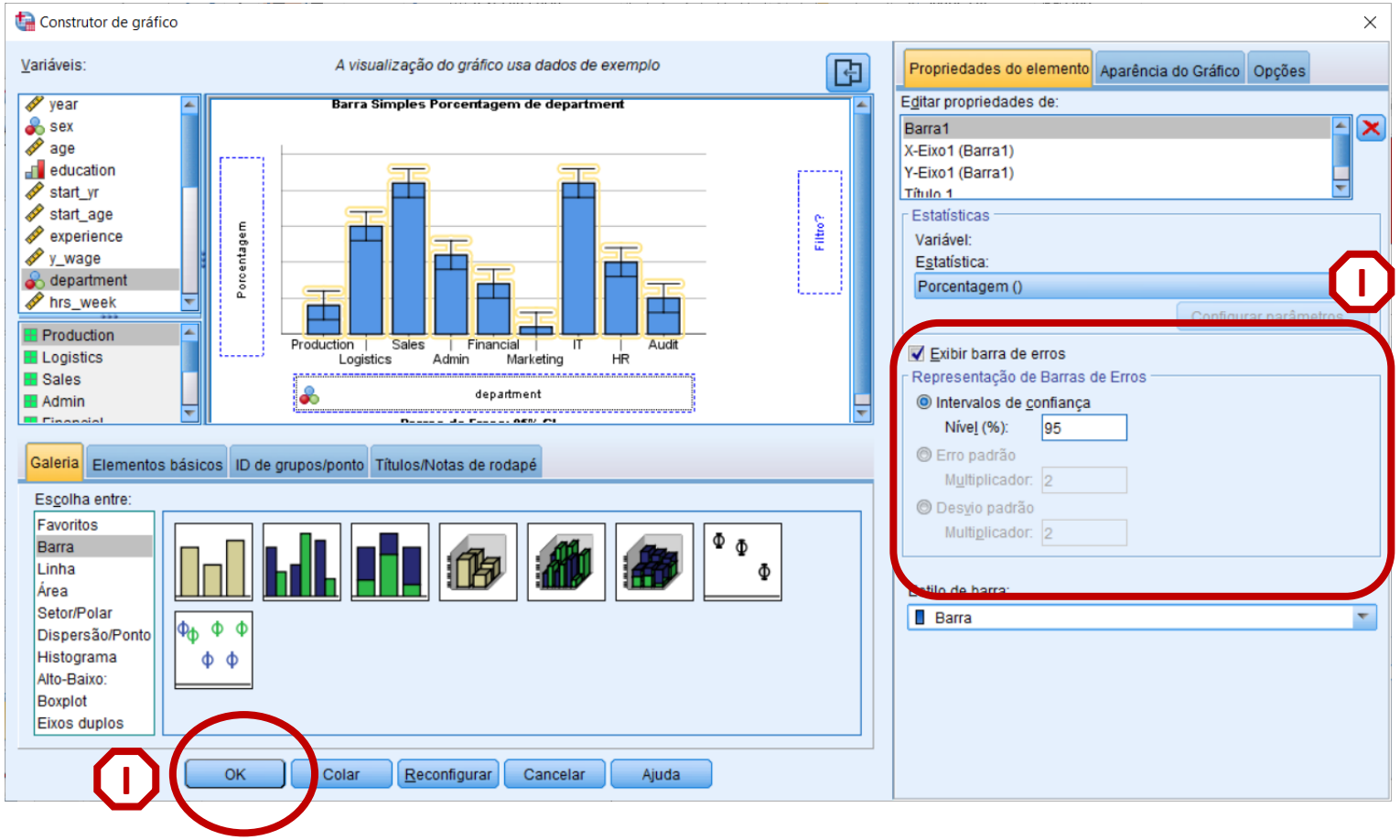


Gráfico de Barras com Erros

- **O gráfico é publicado no 'Visualizador de Resultados'**