



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

Formulação e Teste de Hipóteses

Docente: Amílcar Moreira

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

Local: Edifício F2, Sala 111

- **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

- **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

- **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 métodos probabilísticos.



ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

TÉCNICAS DE AMOSTRAGEM

NÃO-PROBABILÍSTICAS

Conveniência

Por Quotas

Bola de Deve

PROBABILÍSTICAS

Aleatória Simples

Sistemática

Estratificada

Por Clusters



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 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 entrevistados 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	<p>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):</p> $N^{\circ} + i$ <p>Em que $i = N/n$</p> <p>i, fracção de amostragem N, total da população n, tamanho da amostra</p>

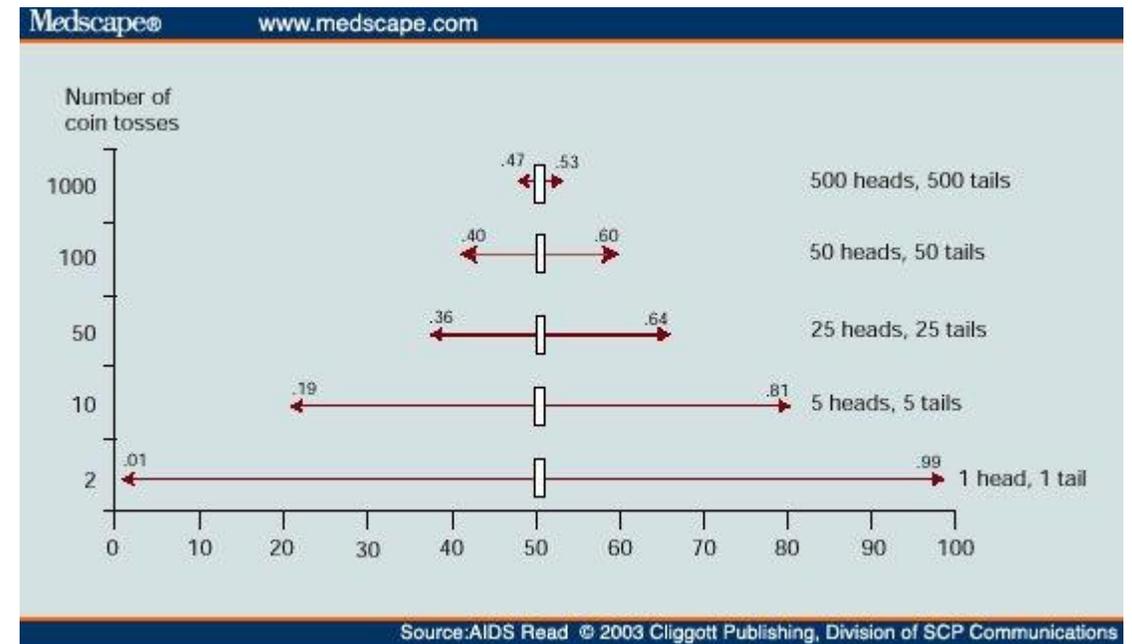
- **Técnicas de Amostragem**

PROBABILITICAS	
Estratificada	<p>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).</p> <p>Exemplos de categorias de estratificação: características demográficas, tipo de empresa, tipo de sector económico, etc.</p> <p>Os membros de cada estrato são depois seleccionados de forma aleatória.</p>
Por Clusters	<p>A base de amostragem é dividida em clusters (Unidades Primárias de Amostragem), formados em função dos interesses do investigador.</p> <p>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).</p> <p>Dentro de cada cluster, selecciona-se de forma aleatória os membros (Unidades Secundárias de Amostragem) a incluir na amostra.</p>

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.talentrunk.co/confidence-interval-and-sample-size/>

- Como se calcula o tamanho da amostra?

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

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)

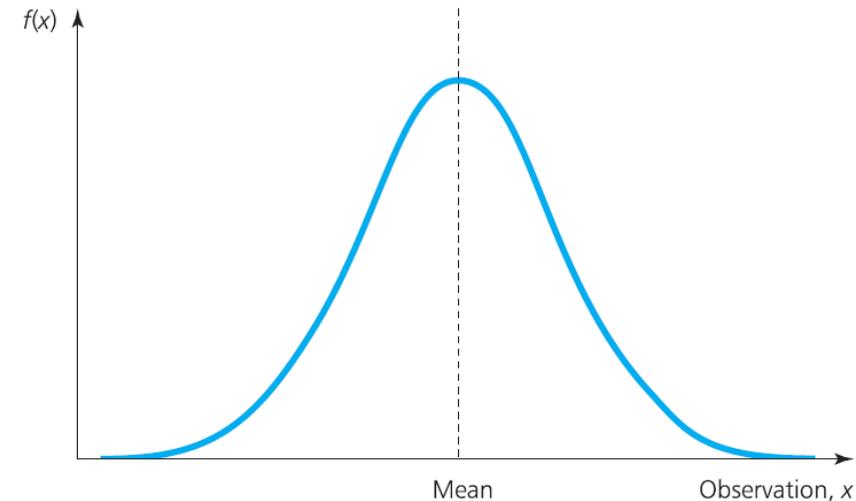
Fonte: <https://www.surveymonkey.com/mp/sample-size-calculator/>



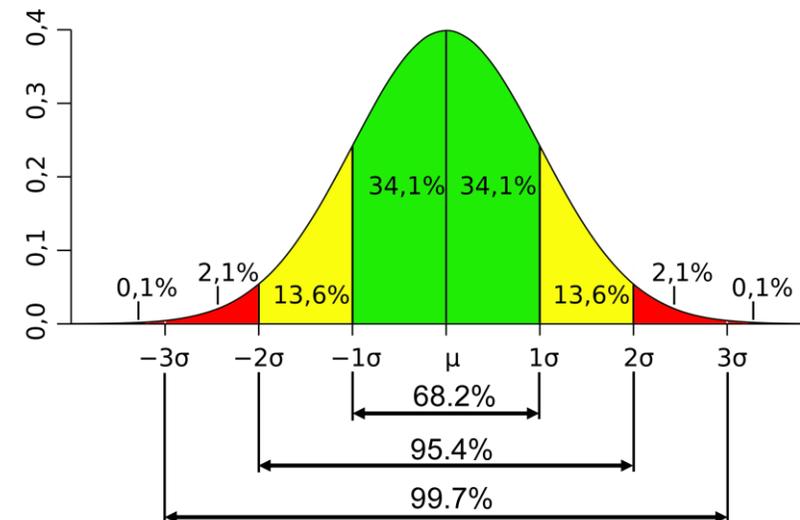
- **A possibilidade de inferir de uma amostra para uma população depende de duas condições fundamentais:**
 - I. **Que amostra seja probabilística**
 - II. **Que haja uma forma de demonstrar que a distribuição da amostra segue uma distribuição normal**

- 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

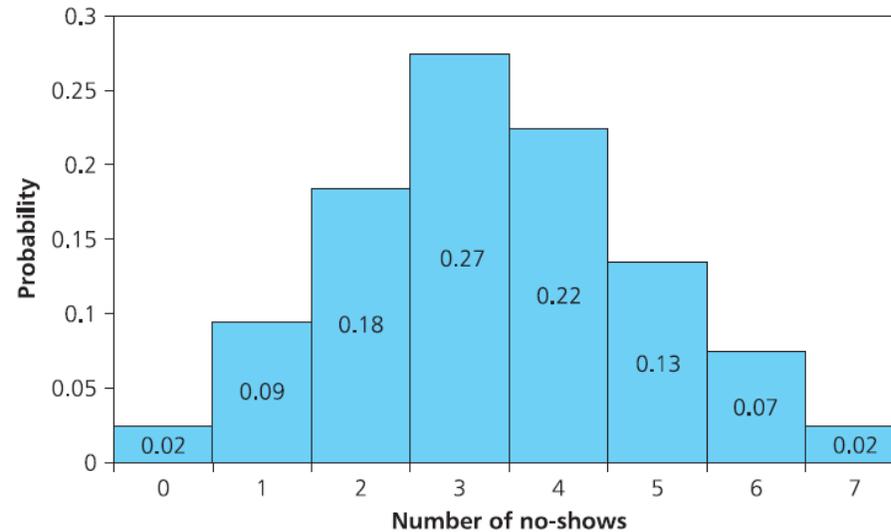
ANÁLISE DE DADOS EM GRH

Aula 7: Formulação e Teste de Hipóteses

- A curva normal exprime a distribuição de frequências numa variável...
- ... mas também a distribuição de probabilidades *

$$P = \frac{\text{número de vezes em que se registou um evento}}{\text{número de observações}}$$

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

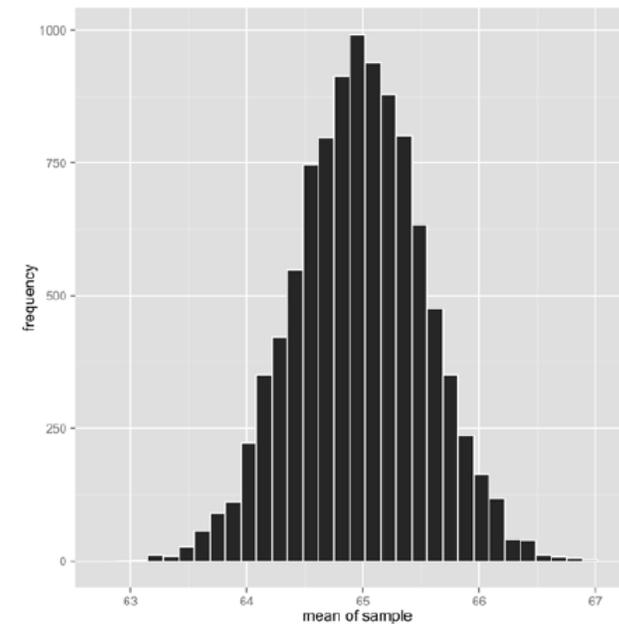
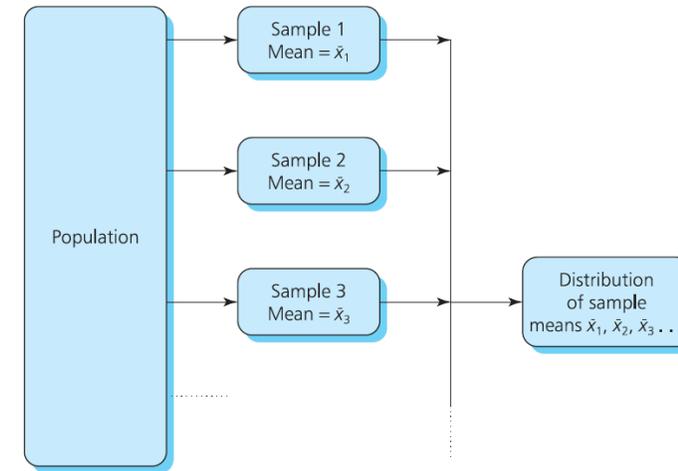


Fonte: Waters, 2011: 337

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’

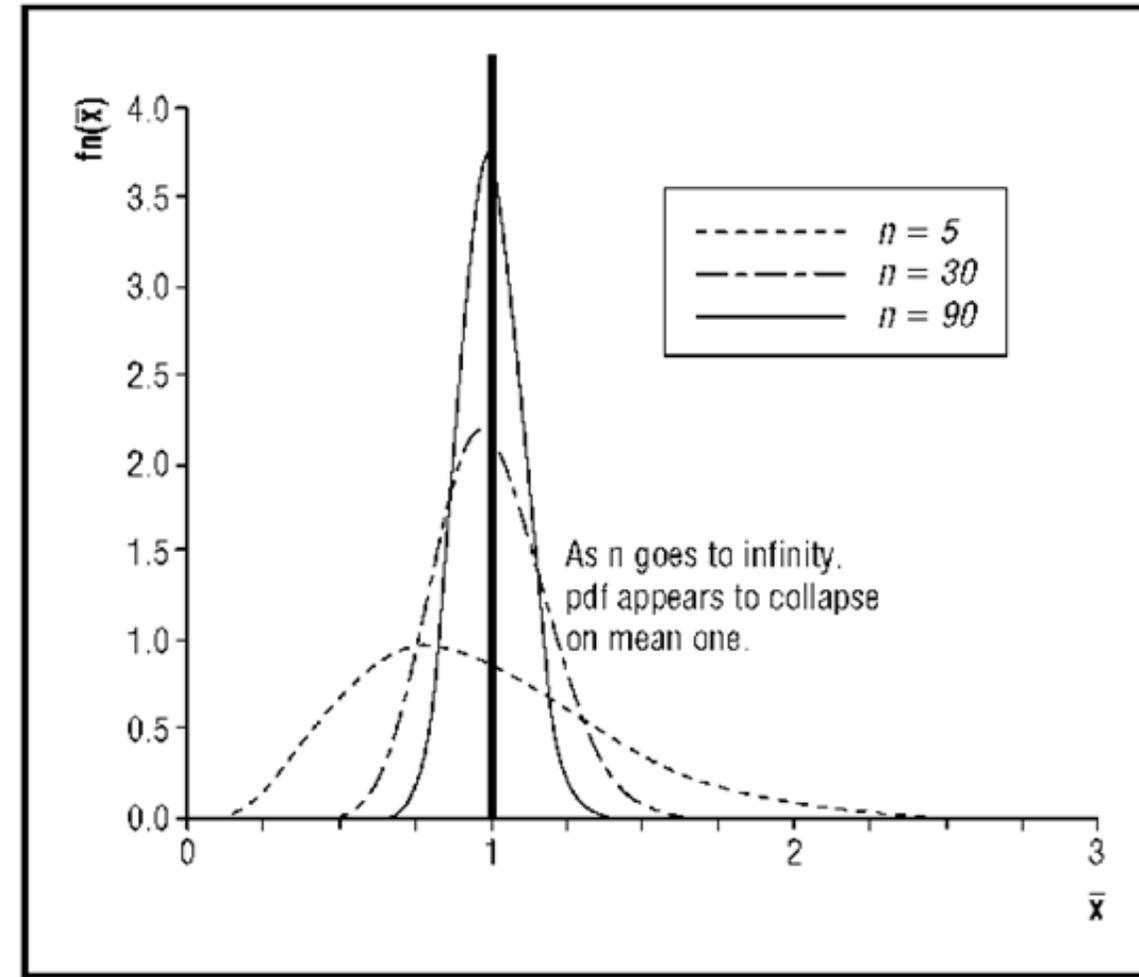


- **O que diz o Teorema do Limite Central**
 - **Quando uma amostra é ≥ 30**
 - **a Distribuição Amostral da Média tende para a uma distribuição normal**
 - **O Desvio-Padrão da Distribuição Amostral da Média é o produto do seguinte rácio:**

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

- **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



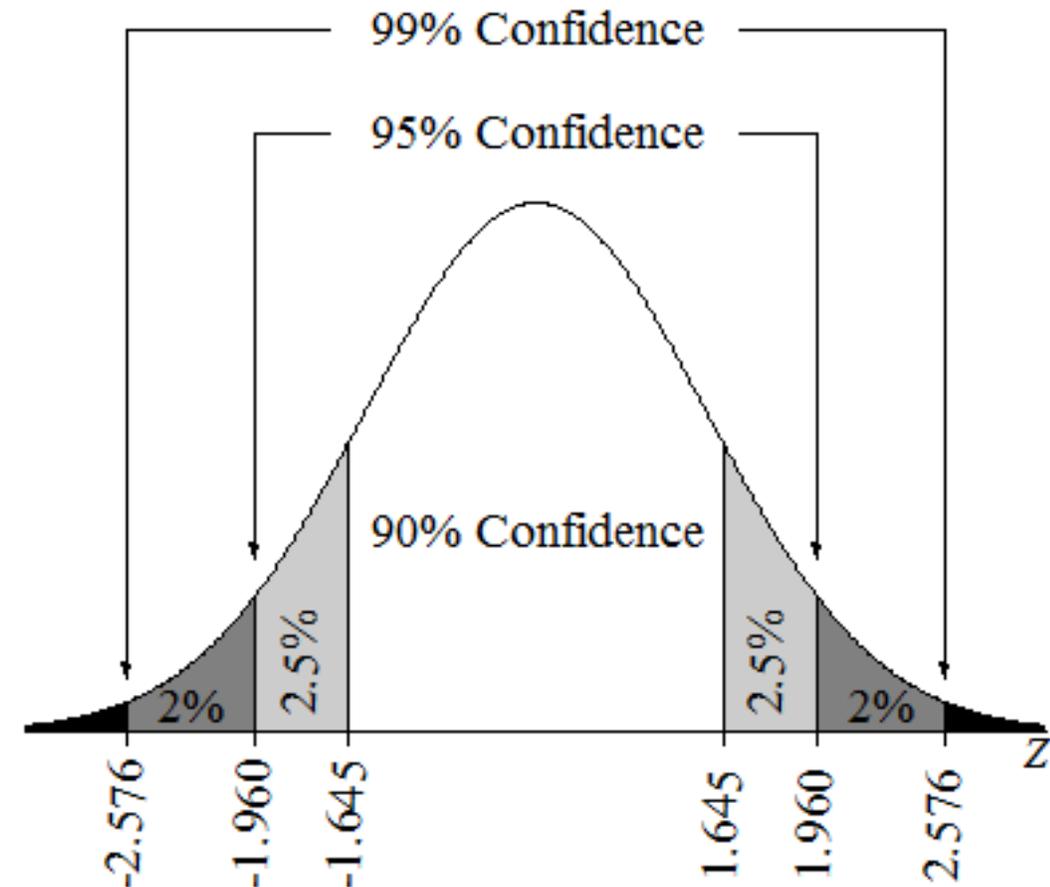
Fonte: <http://what-when-how.com/social-sciences/law-of-large-numbers-social-science/>

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/>

- **O que é o Grau de Confiança?**
 - De notar que, associado a um determinado Grau de Confiança, temos sempre um determinado valor crítico (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>

- 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?

Isto é equivalente a

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

MARGEM DE ERRO

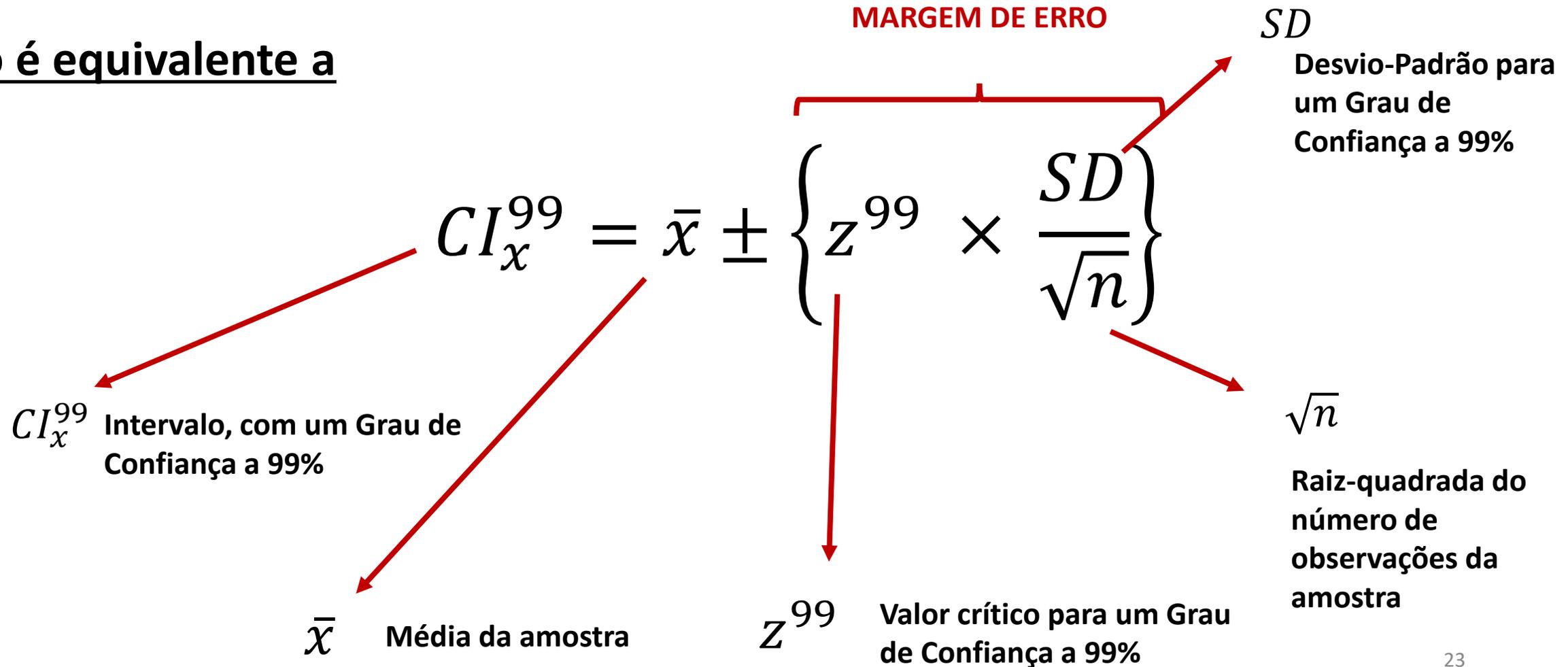
SD
Desvio-Padrão para um Grau de Confiança a 99%

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

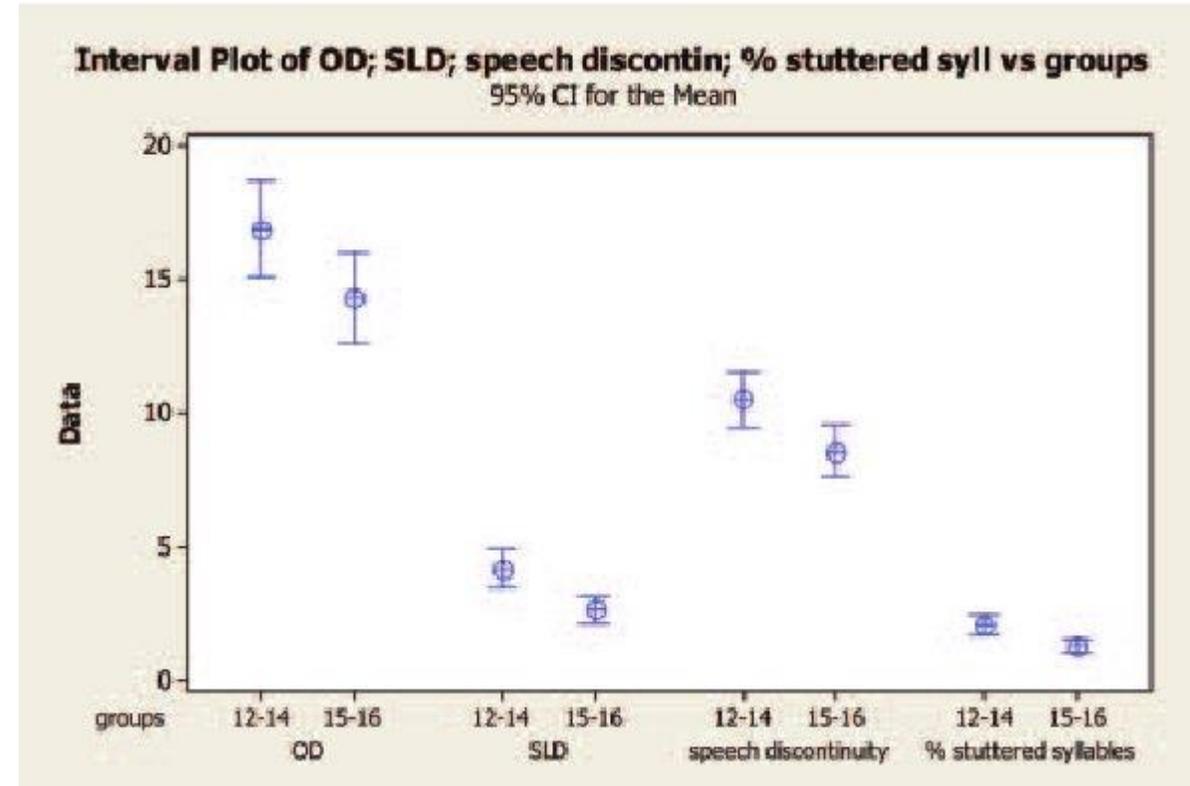
CI_x^{99} Intervalo, com um Grau de Confiança a 99%

\bar{x} Média da amostra

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



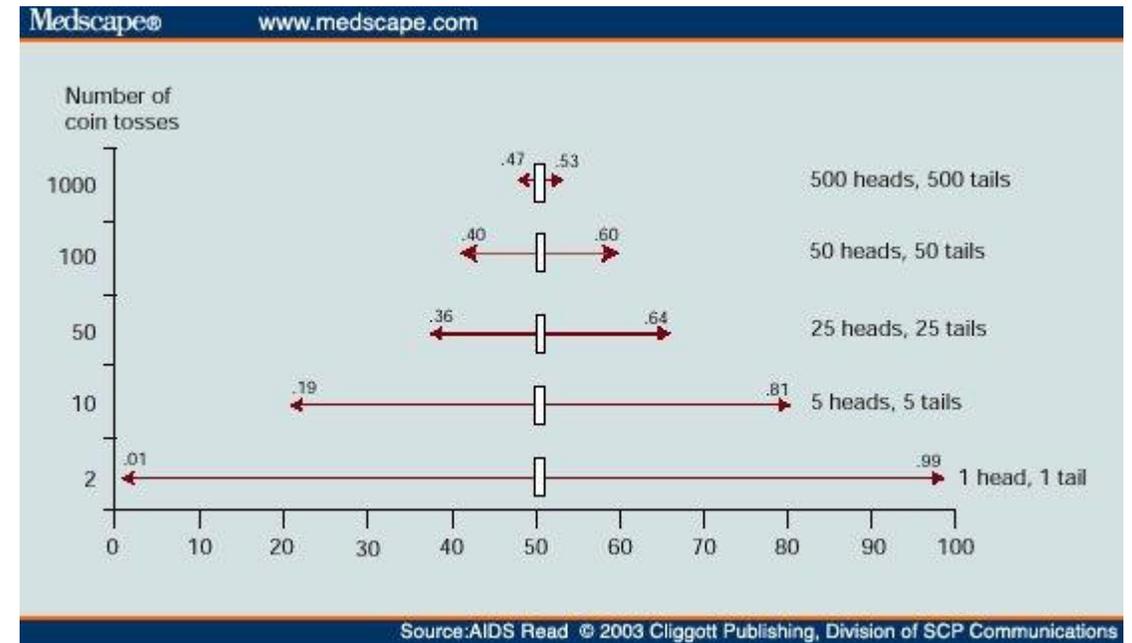
- O que nos diz o Intervalo de Confiança?
 - Grau de precisão da média
 - Quanto maior a amplitude do Intervalo de Confiança, menor o grau de precisão



Fonte:

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

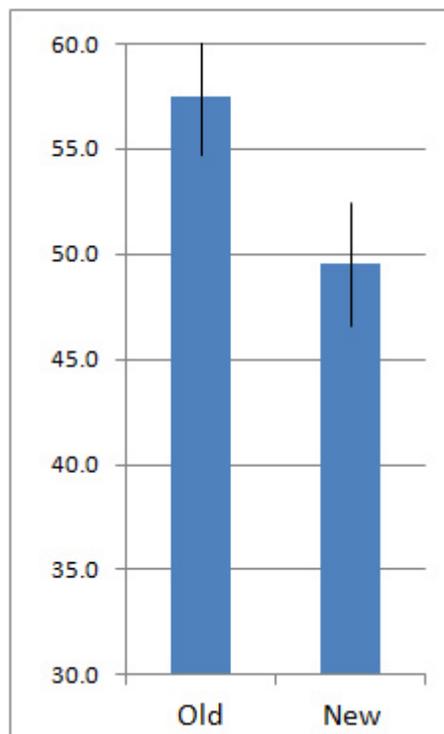
- O que nos diz o Intervalo de Confiança?
 - Grau de precisão da média
 - Quanto maior a amplitude do Intervalo de Confiança, menor o grau de precisão
 - O que afeta a amplitude?
 - Quanto maior a amostra, menor é a amplitude
 - Quanto maior é o Erro-Padrão, maior é a amplitude



Fonte: <http://gosu.talentrunk.co/confidence-interval-and-sample-size/>

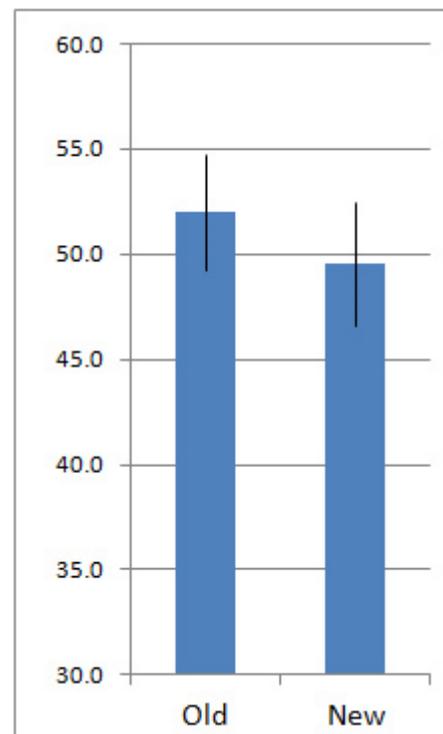
- O que nos diz o Intervalo de Confiança?
 - Significância estatística

Os intervalos de confiança não se sobrepõem



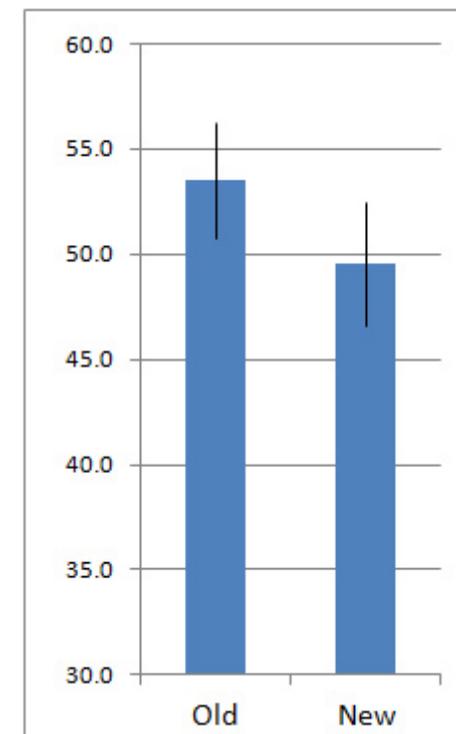
A diferença é estatisticamente significativa

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

- **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

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

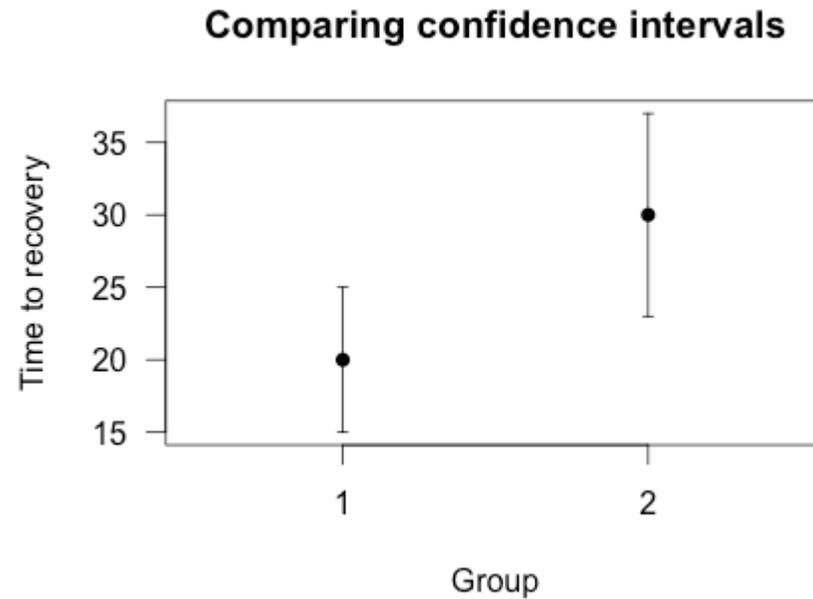
CI_x^{99} Intervalo, com um Grau de Confiança a 99%

\hat{p} Proporção da categoria, na amostra

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

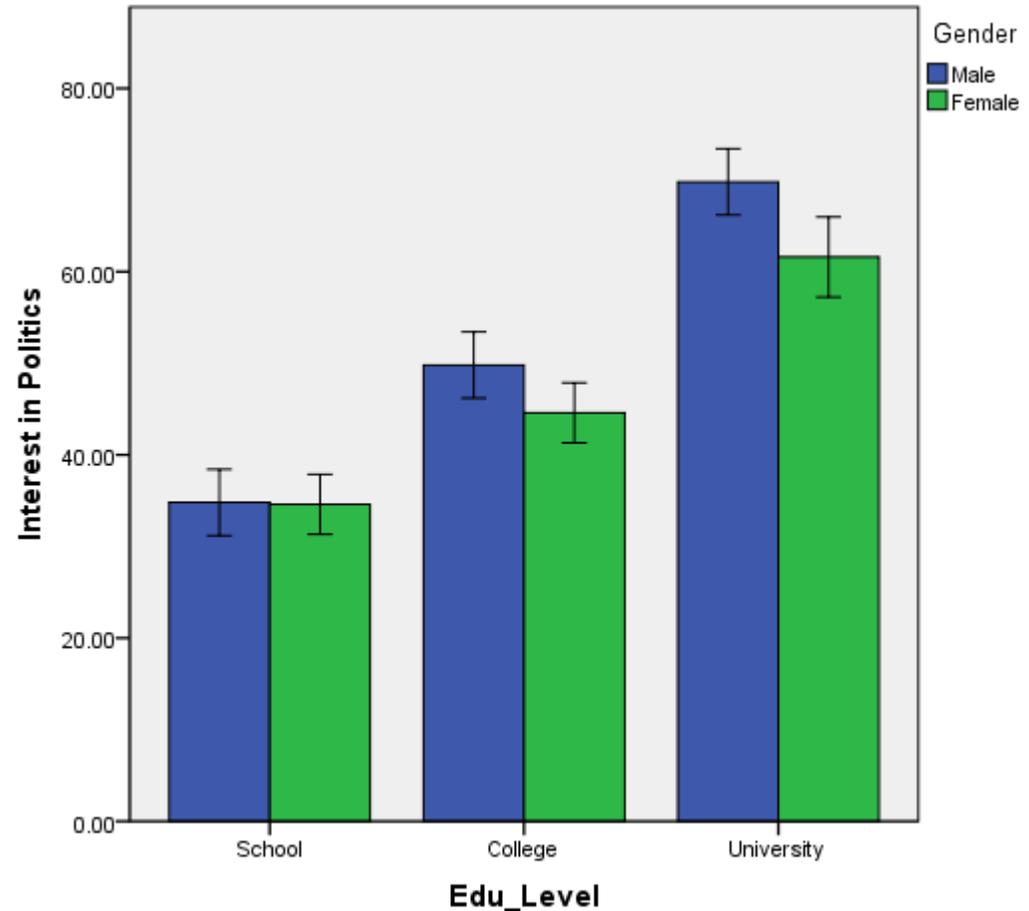
n Número de observações da amostra

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



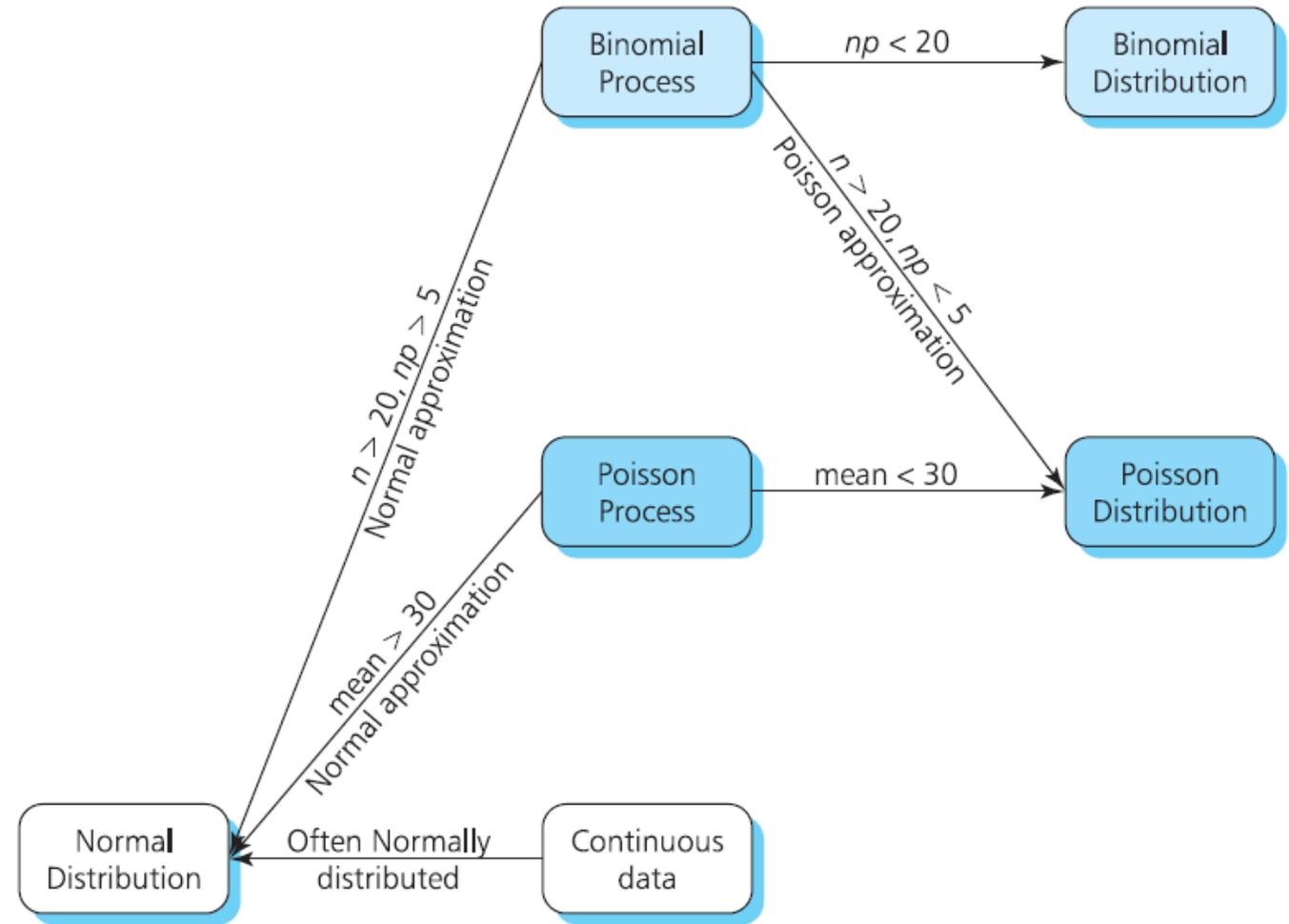
Fonte: <https://www.statisticsonewrong.com/significant-differences.html>

- 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>

- 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

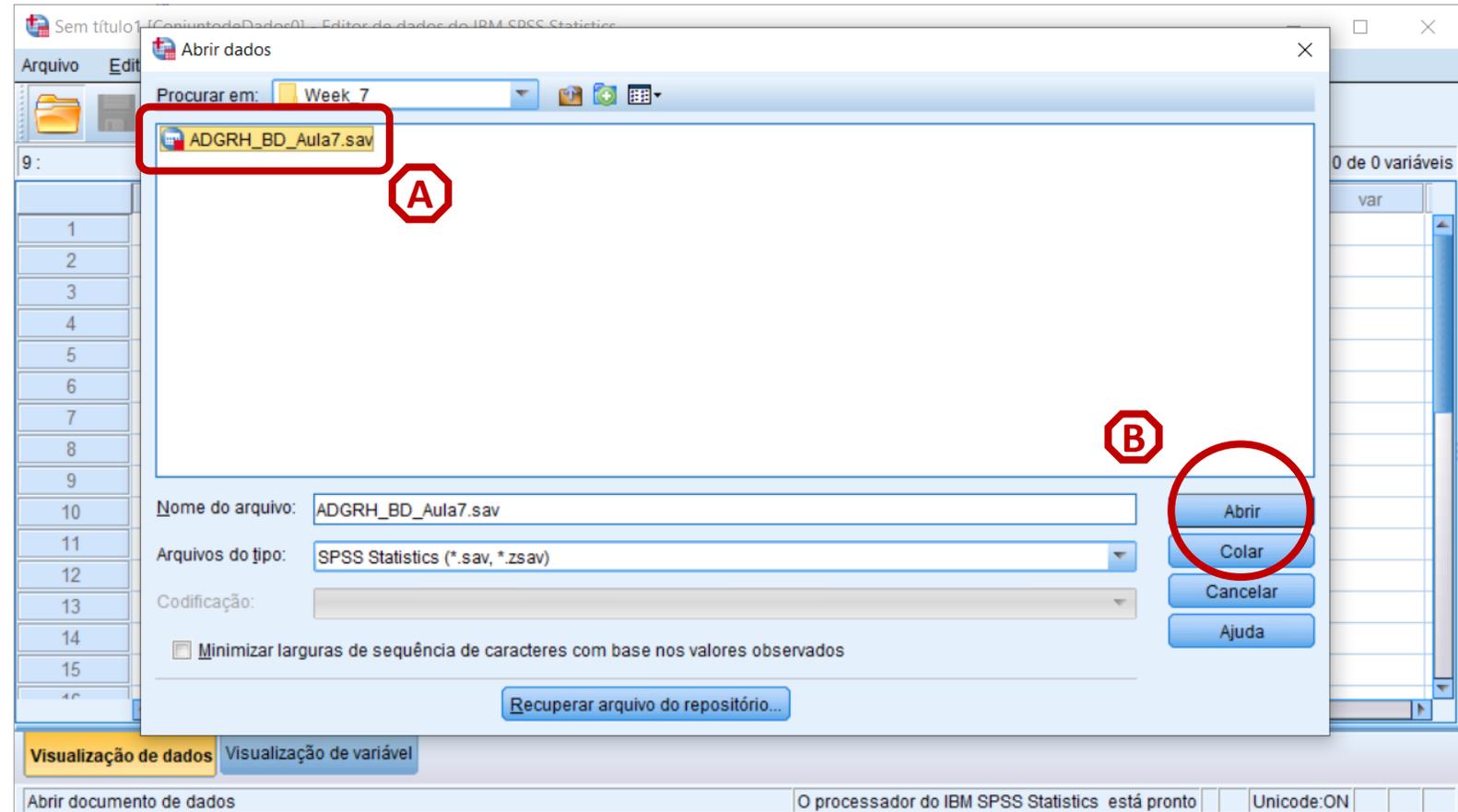
De uma Média

Intervalo de Confiança: Média

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

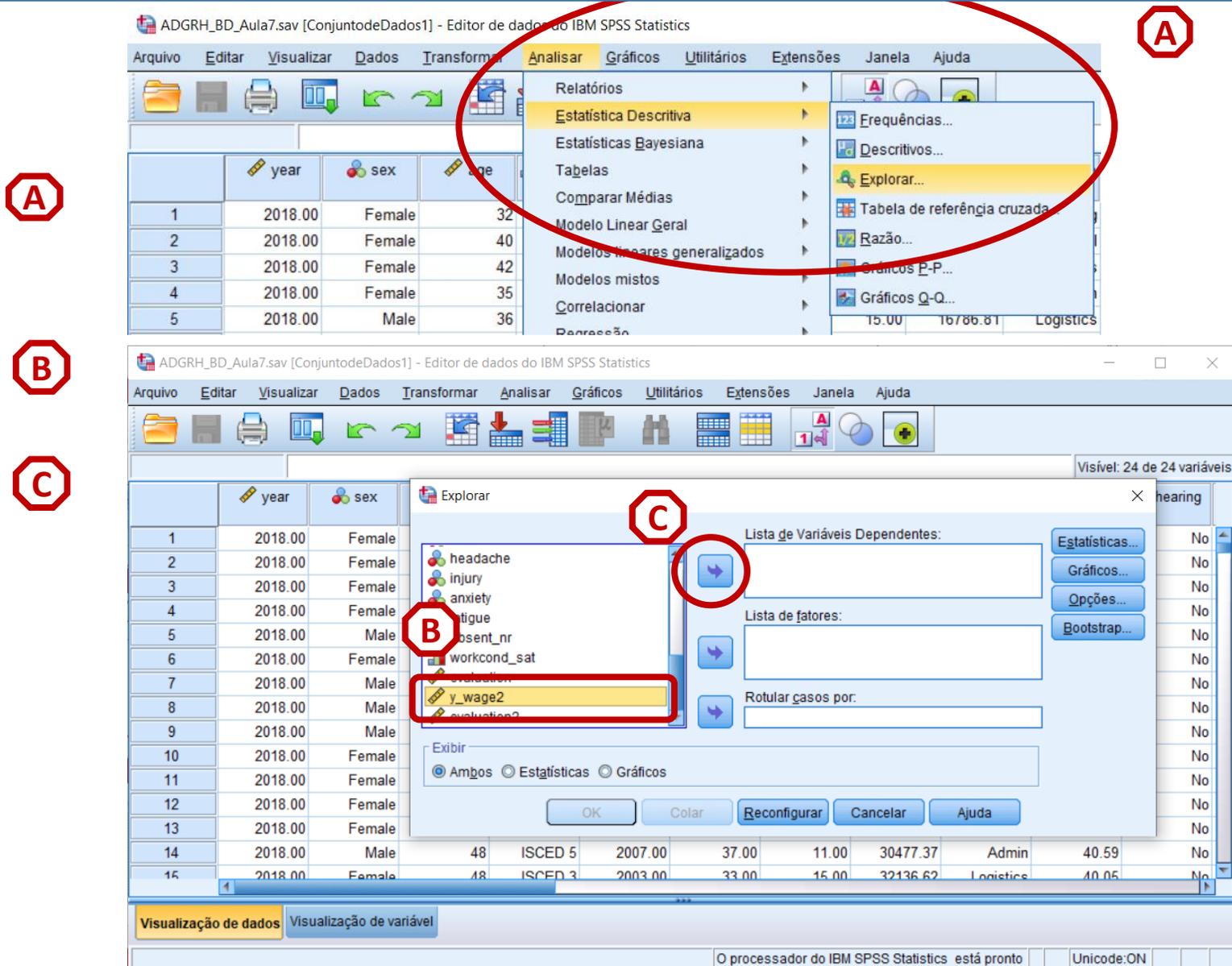
Intervalo de Confiança: Média

- Seleccionar ficheiro 'Database_Tutorial_7.sav' **A**
- Seleccionar 'Abrir' **B**



Intervalo de Confiança: Média

- Selecionar 'Analisar' / 'Estatísticas Descritivas' / 'Explorar'
- Selecionar a variável 'y_wage2'
- Colocar na caixa 'Lista de Variáveis Dependentes'



A

B

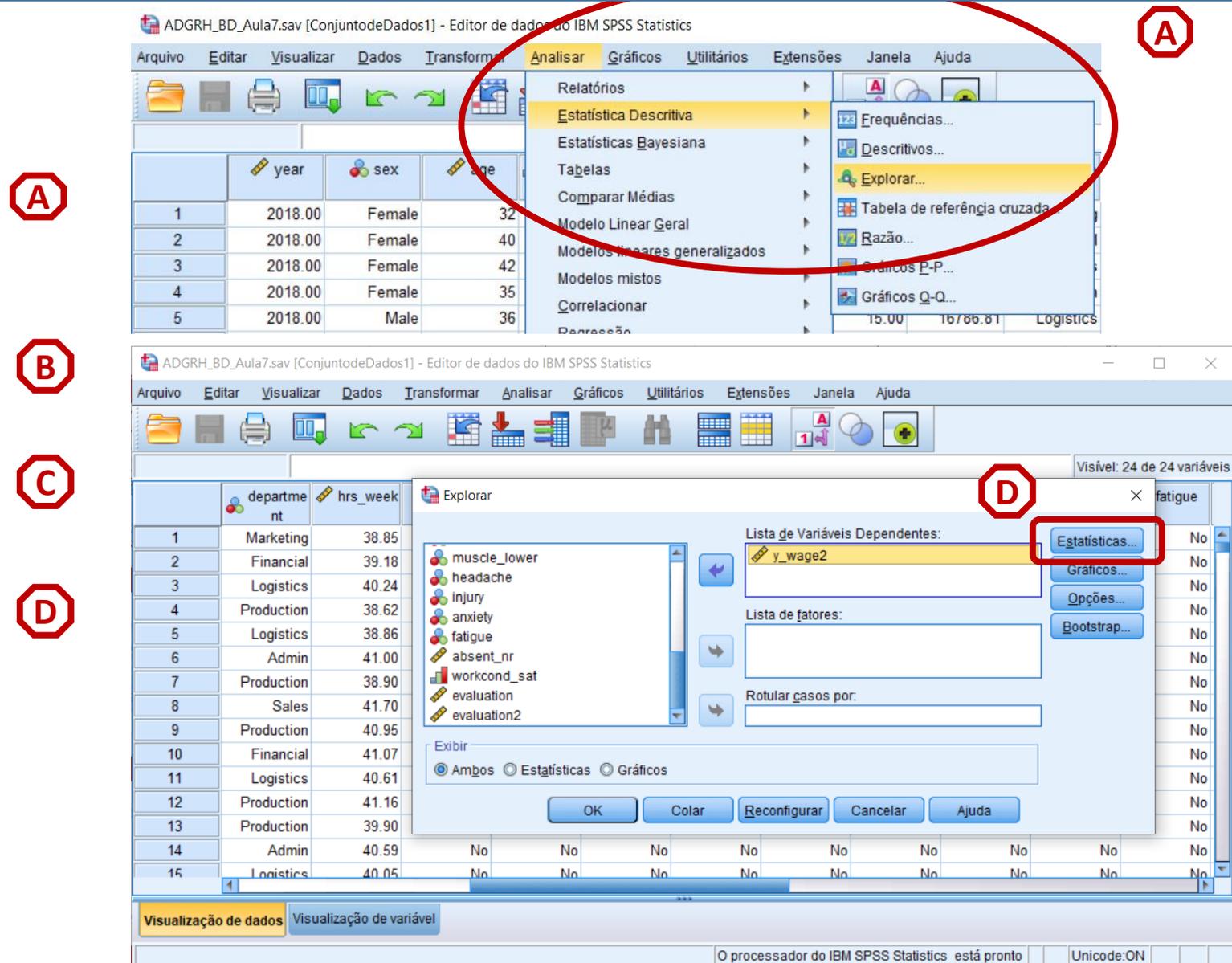
C

year	sex	age	
1	2018.00	Female	32
2	2018.00	Female	40
3	2018.00	Female	42
4	2018.00	Female	35
5	2018.00	Male	36

year	sex	headache	injury	anxiety	stigma	absent_nr	workcond_sat	evaluation	y_wage2	evaluation2
1	2018.00	Female								
2	2018.00	Female								
3	2018.00	Female								
4	2018.00	Female								
5	2018.00	Male								
6	2018.00	Female								
7	2018.00	Male								
8	2018.00	Male								
9	2018.00	Male								
10	2018.00	Female								
11	2018.00	Female								
12	2018.00	Female								
13	2018.00	Female								
14	2018.00	Male	48	ISCED 5	2007.00	37.00	11.00	30477.37	Admin	40.59
15	2018.00	Female	48	ISCED 3	2003.00	33.00	15.00	32136.62	Logistics	40.05

Intervalo de Confiança: Média

- Selecionar 'Analisar' / 'Estatísticas Descritivas' / 'Explorar'
- Selecionar a variável 'y_wage2'
- Colocar na caixa 'Lista de Variáveis Dependentes'
- Selecionar 'Estatísticas'



A

B

C

D

year	sex	age	
1	2018.00	Female	32
2	2018.00	Female	40
3	2018.00	Female	42
4	2018.00	Female	35
5	2018.00	Male	36

departme nt	hrs_week	
1	Marketing	38.85
2	Financial	39.18
3	Logistics	40.24
4	Production	38.62
5	Logistics	38.86
6	Admin	41.00
7	Production	38.90
8	Sales	41.70
9	Production	40.95
10	Financial	41.07
11	Logistics	40.61
12	Production	41.16
13	Production	39.90
14	Admin	40.59
15	Logistics	40.05

Intervalo de Confiança: Média

- Selecionar 'Analisar' / 'Estatísticas Descritivas' / 'Explorar'
- Selecionar a variável 'y_wage2'
- Colocar na caixa 'Lista de Variáveis Dependentes'
- Selecionar 'Estatísticas'
- Selecionar 'Descritivos'
- Definir um Grau de Confiança de '95%'
- Selecionar 'Continuar'/OK

A

B

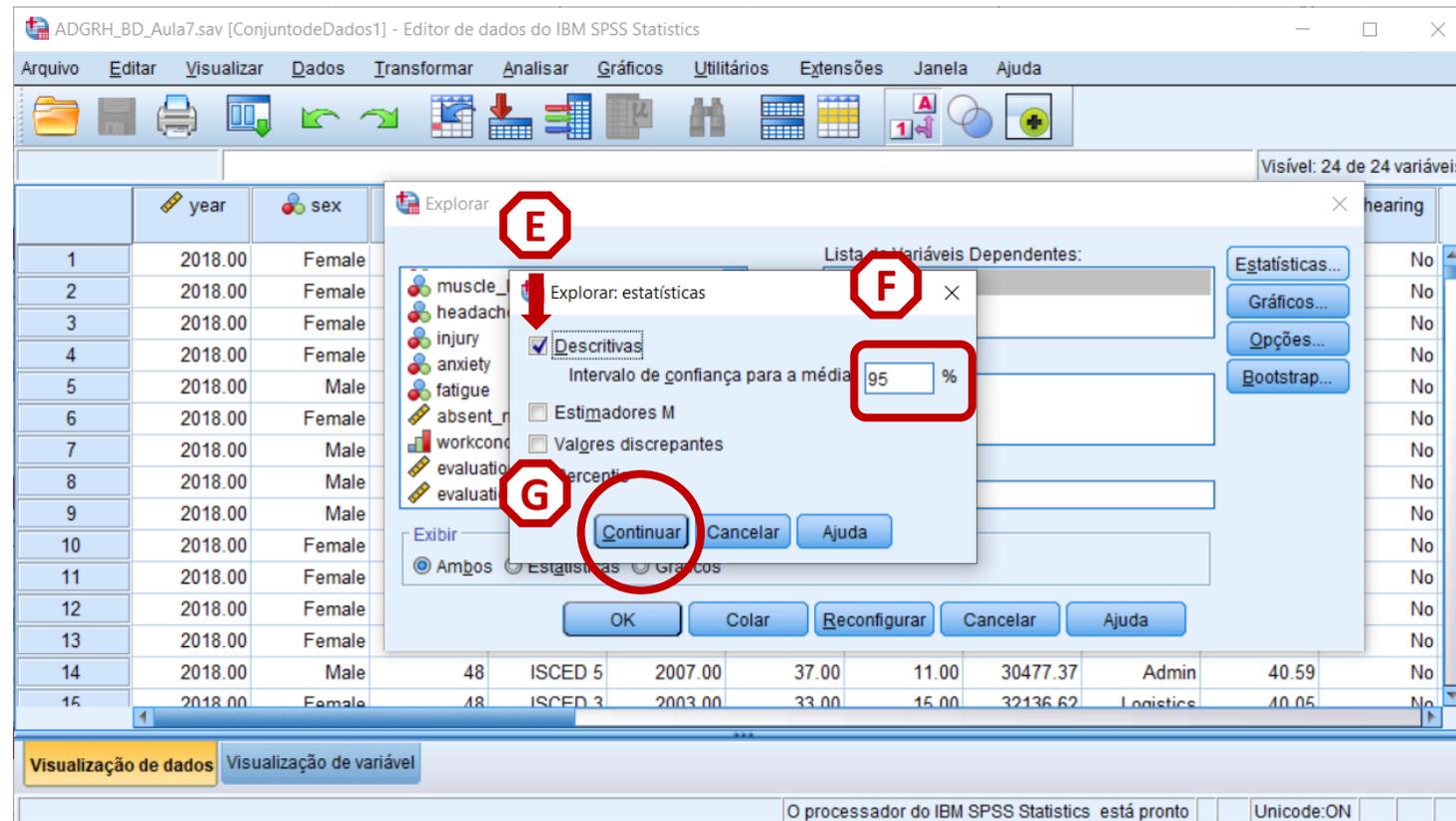
C

D

E

F

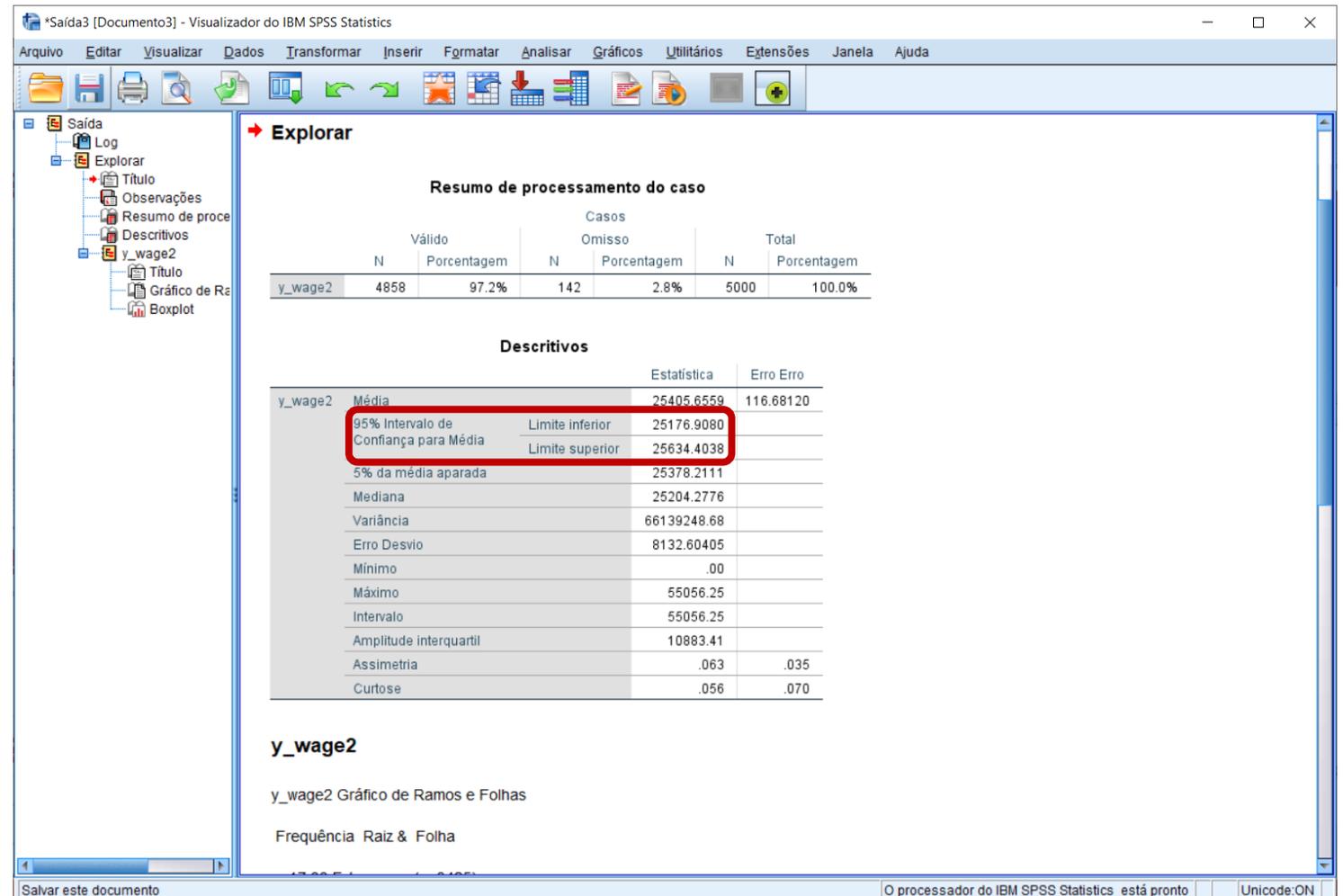
G



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.



*Saída3 [Documento3] - Visualizador do IBM SPSS Statistics

Arquivo Editar Visualizar Dados Transformar Inserir Formatar Analisar Gráficos Utilitários Extensões Janela Ajuda

Explorar

Resumo de processamento do caso

	Válido		Casos Omissos		Total	
	N	Porcentagem	N	Porcentagem	N	Porcentagem
y_wage2	4858	97.2%	142	2.8%	5000	100.0%

Descritivos

		Estatística	Erro
y_wage2	Média	25405.6559	116.68120
95% Intervalo de Confiança para Média	Limite inferior	25176.9080	
	Limite superior	25634.4038	
5% da média aparada		25378.2111	
Mediana		25204.2776	
Variância		66139248.68	
Erro Desvio		8132.60405	
Mínimo		.00	
Máximo		55056.25	
Intervalo		55056.25	
Amplitude interquartil		10883.41	
Assimetria		.063	.035
Curtose		.056	.070

y_wage2

y_wage2 Gráfico de Ramos e Folhas

Frequência Raiz & Folha

Salvar este documento

O processador do IBM SPSS Statistics está pronto

Unicode.ON



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 fazemos testes de significância estatística**
 - **Para variáveis binomiais (com valores 0 e 1)**
 - **Pressuposto: a média dessa variável representa o proporção de observações com valor 1**

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

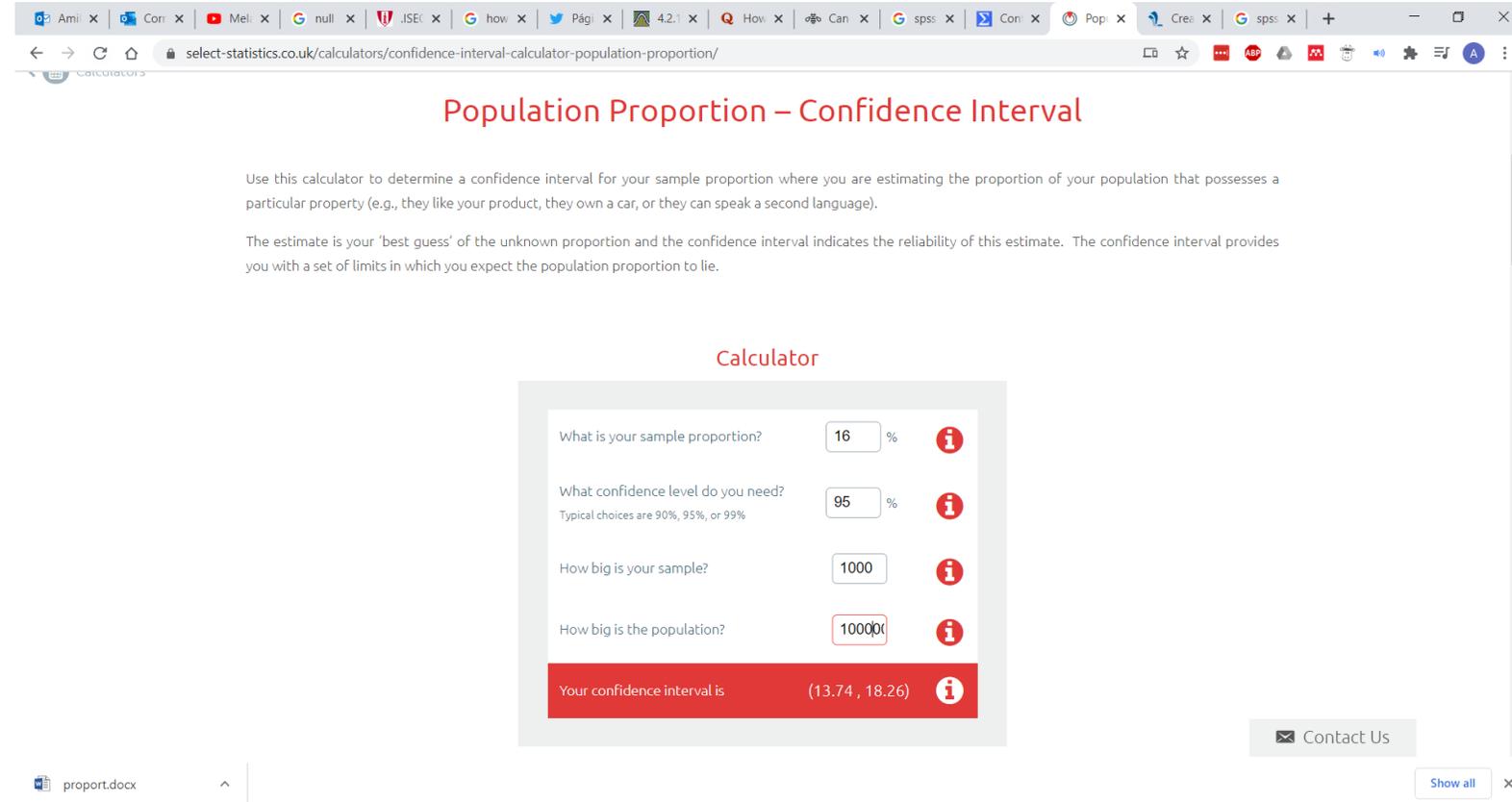
Em alternativa, o Intervalo de Confiança pode ser calculado à mão...

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

- 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%..



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

Population Proportion – Confidence Interval

Use this calculator to determine a confidence interval for your sample proportion where you are estimating the proportion of your population that possesses a particular property (e.g., they like your product, they own a car, or they can speak a second language).

The estimate is your 'best guess' of the unknown proportion and the confidence interval indicates the reliability of this estimate. The confidence interval provides you with a set of limits in which you expect the population proportion to lie.

Calculator

What is your sample proportion?	<input type="text" value="16"/> %	
What confidence level do you need? <small>Typical choices are 90%, 95%, or 99%</small>	<input type="text" value="95"/> %	
How big is your sample?	<input type="text" value="1000"/>	
How big is the population?	<input type="text" value="10000"/>	
Your confidence interval is	(13.74 , 18.26)	

proport.docx

Contact Us

Show all

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

A outra alternativa é representar os intervalos de confiança visualmente

Gráfico de Barras com Erros

- Selecionar 'Gráficos' / 'Construtor de Gráfico'



- Selecionar 'Não mostrar este diálogo novamente'



- Selecionar 'OK'

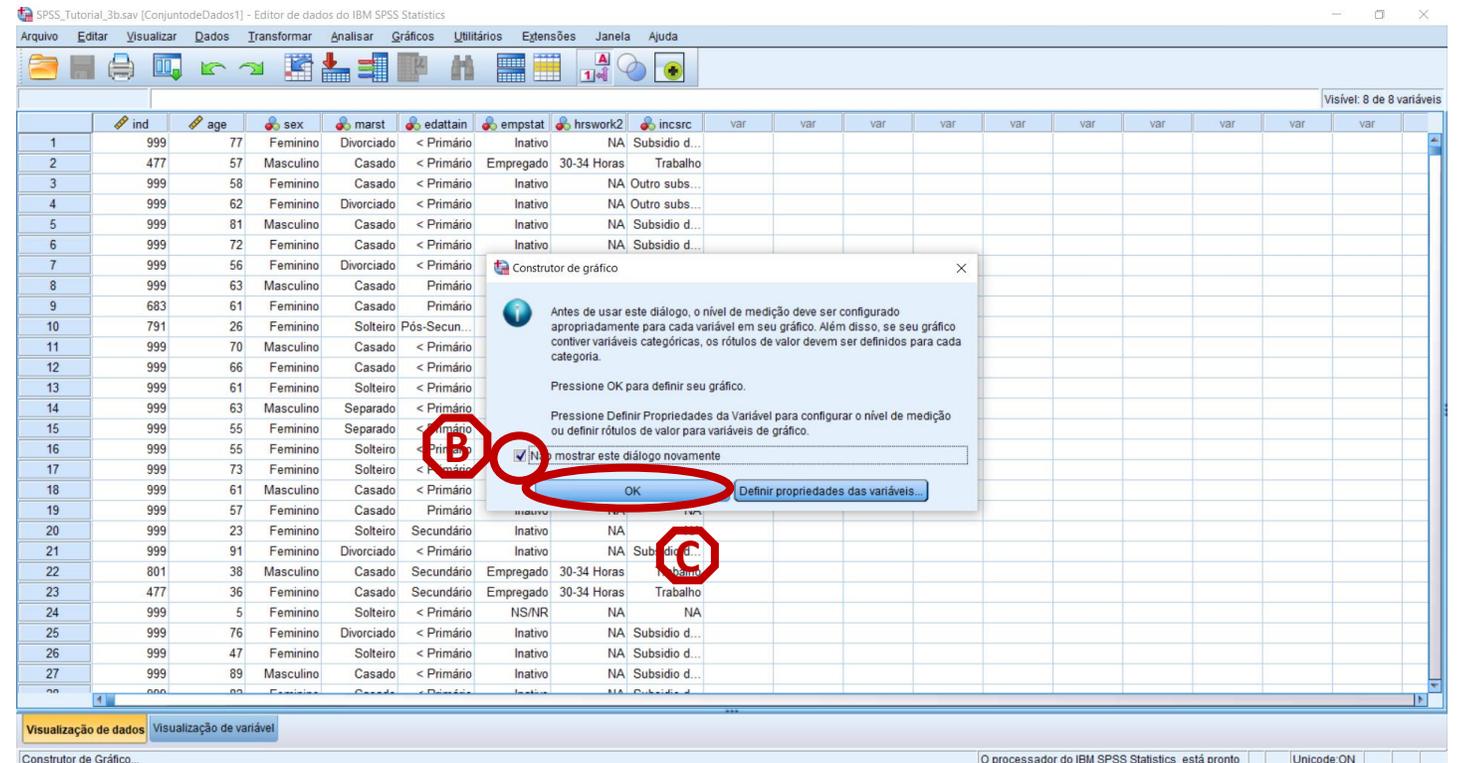
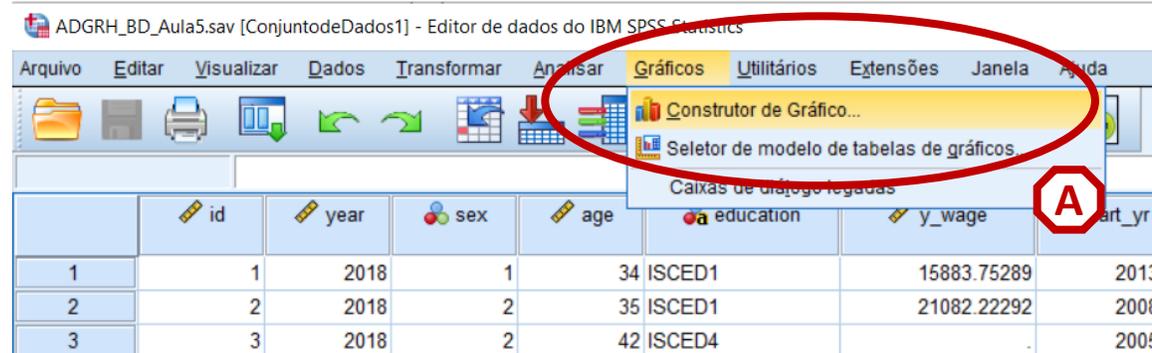


Gráfico de Barras com Erros

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

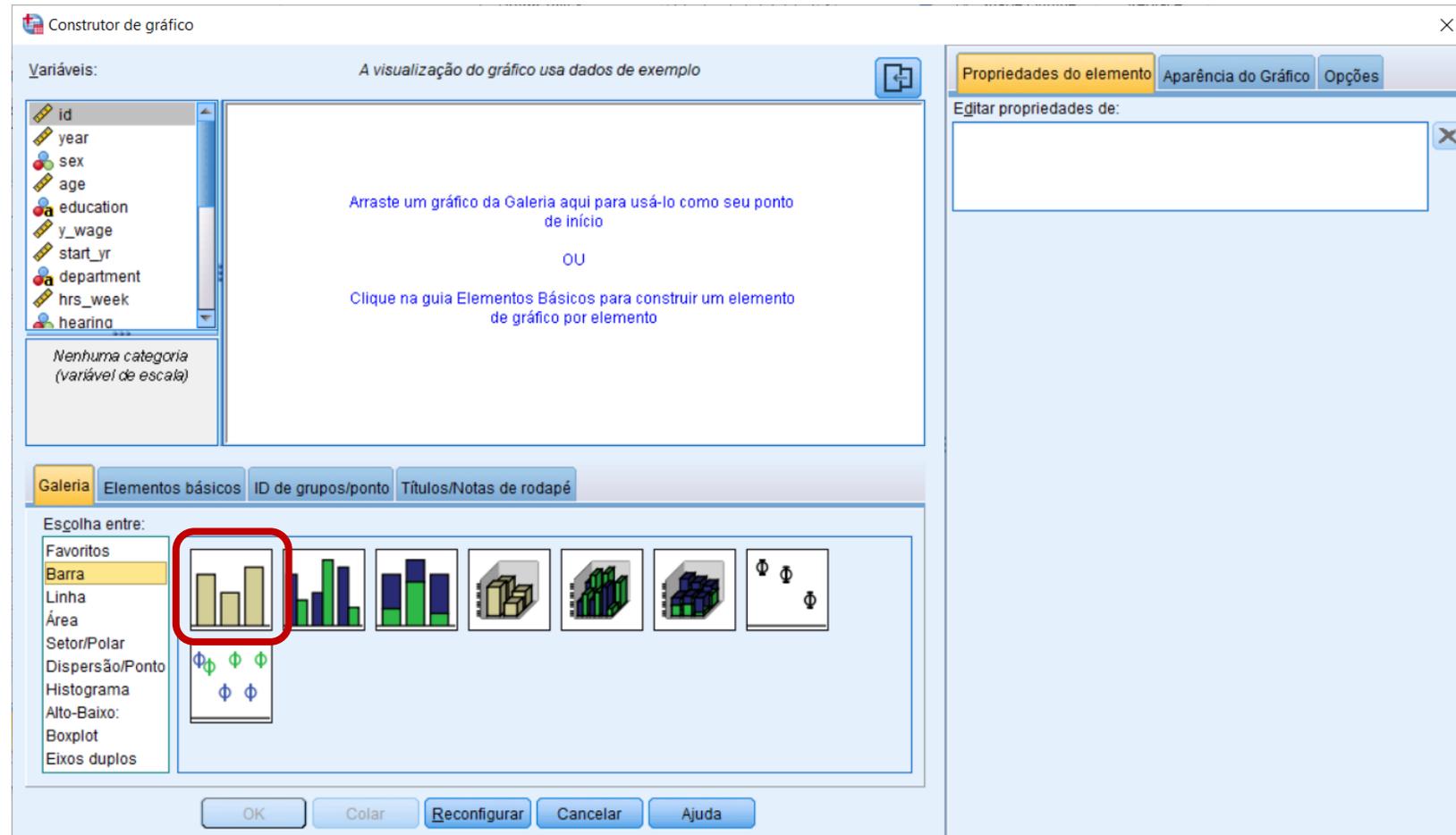


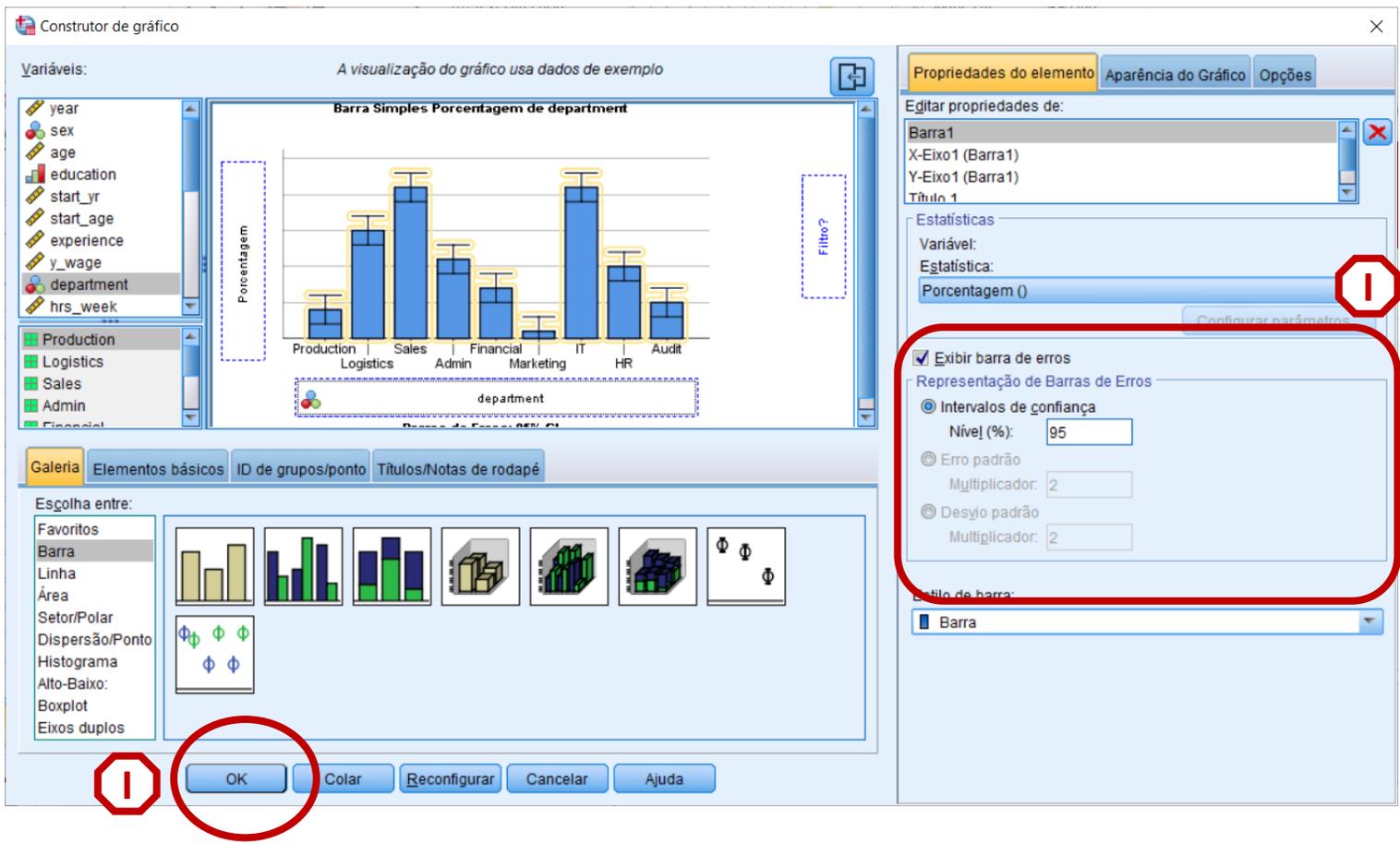
Gráfico de Barras com Erros

- Selecione 'Barras'
- Selecionar (com duplo-clique) o Gráfico de Barras (simples)
- Selecionar a variável 'department'
- Colocar a variável 'department' no 'Eixo X'



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'

Construtor de gráfico

A visualização do gráfico usa dados de exemplo

Variáveis:

- year
- sex
- age
- education
- start_yr
- start_age
- experience
- y_wage
- department
- hrs_week

Production
Logistics
Sales
Admin
Financial
Marketing
IT
HR
Audit

Barra Simples Porcentagem de department

Porcentagem

Filtro?

department

Galeria | Elementos básicos | ID de grupos/ponto | Títulos/Notas de rodapé

Escolha entre:

- Favoritos
- Barra
- Linha
- Área
- Setor/Polar
- Dispersão/Ponto
- Histograma
- Alto-Baixo:
- Boxplot
- Eixos duplos

Propriedades do elemento | Aparência do Gráfico | Opções

Editar propriedades de: Barra1

X-Eixo1 (Barra1)
Y-Eixo1 (Barra1)
Título 1

Estatísticas

Variável:
Estatística:
Porcentagem ()

Exibir barra de erros

Representação de Barras de Erros

- Intervalos de confiança
 - Nível (%): 95
- Erro padrão
 - Multiplicador: 2
- Desvio padrão
 - Multiplicador: 2

Estilo de barra:

Barra

OK | Colar | Reconfigurar | Cancelar | Ajuda

Gráfico de Barras com Erros

- O gráfico é publicado no 'Visualizador de Resultados'

