

Introdução à análise de dados em Saúde



Módulo 3 Estatística descritiva – Medidas

Prof. Juliano Gaspar



Prof. D.r Juliano Gaspar

Email: julianogaspar@gmail.com

Orcid ID: <https://orcid.org/0000-0003-0670-9021>

Formação

- **Cientista da Computação** pela UNIVALI (SC)
- **Mestre em Informática Médica** pela UP (Portugal)
- **Doutor em Saúde Digital** pela UFMG
- **Pós-doutor em Tec. para Educação em Saúde** pela UFMG

Educação

Professor da Faculdade de Medicina da UFMG

- Introdução à Pesquisa Científica II
- Informação em Decisão em Saúde

Professor da Pós-Graduação da FM-UFMG

- Informática Médica

Professor da Especialização em Saúde Digital da UFG

- TCCs em Saúde Digital

Professor Grupo Ânima: Una e Unibh

- UDWMJ
 - Usabilidade, Desenvolvimento Web e Mobile
 - Vida & Carreira

Inovação, Pesquisa, Desenvolvimento e Extensão

- Coordenador do Núcleo de Pesquisa em Informática Aplicada à Saúde da UFMG
- **Membro do CINTESIS** (Centro de Investigação em Tecnologias e Serviços de Saúde da Faculdade de Medicina da Universidade do Porto FMUP), investigador na equipe HIS-EHR – Sistemas de Informação em Saúde e Registos de Saúde Eletrónicos.
- **Membro da SBIS** – Sociedade Brasileira de Informática em Saúde
- Membro do Comitê Científico Organizador do CBIS 2022
- Revisor de revistas científicas

Linhas de pesquisa e projetos

- PreemieTest – Detecção da prematuridade através da interação entre a luz e a pele neonatal
- SISMeter – Aplicativos e Sistemas de Informação para Atenção Materno Infantil
- Inteligência Artificial aplicada à Saúde.

Programas e projetos de extensão

- Informática e Saúde
- Prevenção da COVID-19 em APP
- Meu Pré-natal (APP)
- Projeto Educação Continuada em Informática



Introdução à análise de dados em Saúde

Conteúdo

Módulo 1 - Variáveis e Bancos de dados

- As variáveis Clínicas
- Banco de dados biomédicos
- Modelagem de dados
- Qualidade e consistência dos dados

Módulo 2 - Modelagem de dados

- Introdução ao SPSS
- Codificação de variáveis
- Recodificação de variáveis
- Criando variáveis

Módulo 3 - Estatística descritiva - Medidas

- Descrevendo as variáveis
- Medidas de tendência central e dispersão
- Distribuições de frequência
- Testes de normalidade (Kolmogorov-Smirnov)
- Select, split e sort cases

Módulo 4 - Inferências sobre variáveis categóricas

- Tabelas de contingência 2 x 2
- Risco Relativo e Razão de Chance
- Intervalos de confiança
- Testes de hipótese (independência)
- Testes de Qui-quadrado, Fisher e McNemar

Módulo 5 - Inferência sobre numéricas x categóricas

- Testes de médias para distribuição paramétrica
- Teste-t (student) Independente (2 categorias)
- Teste de Levene (teste de variâncias)
- Teste-t Pareado (2 categorias)
- ANOVA one way, Teste de Tukey (> 2 categorias)
- Testes pos-hoc

Módulo 6 - Inferência sobre numéricas x categóricas

- Testes de medianas para distr. não-paramétrica
- Teste U Mann-Witney (2 categorias)
- Teste Wilcoxon (2 categorias)
- Teste de Kruskal-Wallis (>2 categorias)
- Teste de Friedman

Módulo 7 - Inferência sobre variáveis numéricas

- Coeficiente de Correlação Intra-classe
- Correlação de Pearson
- Correlação de Spearman

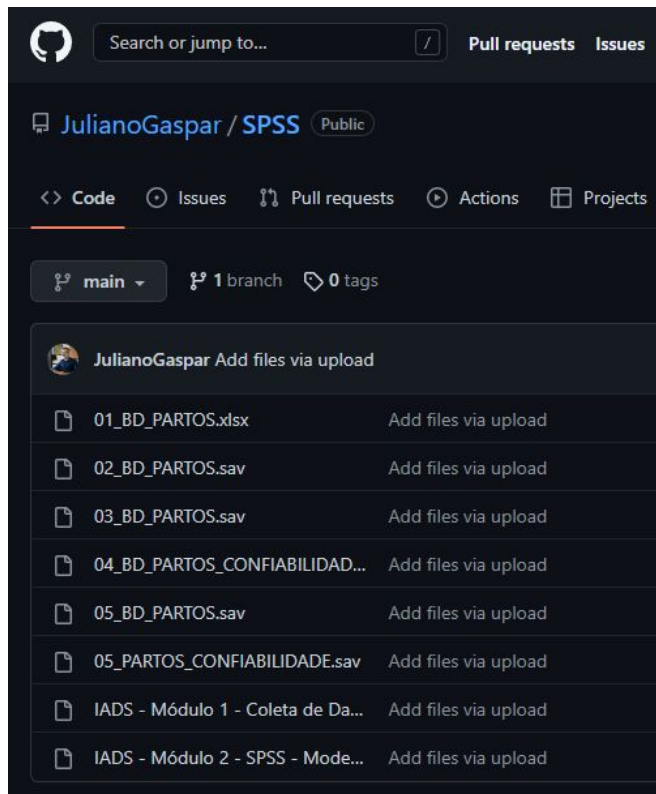
Módulo 8 - Estatística descritiva - Gráficos

- Gráficos para variáveis categóricas
- Gráficos para numéricas contínuas e discretas
- Gráficos de dispersão de variáveis



Introdução à análise de dados em Saúde

Baixar os arquivos do Github

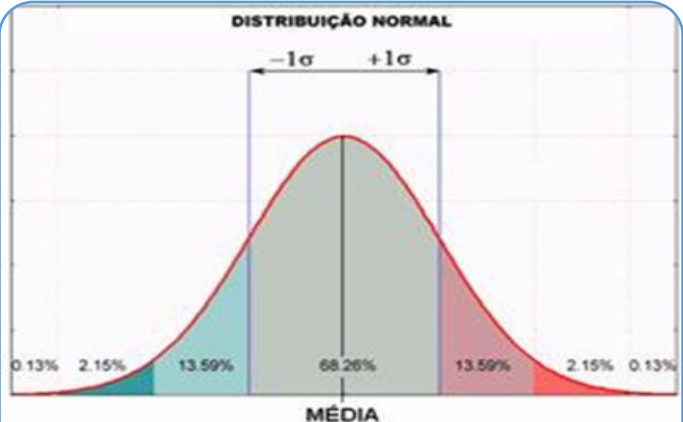


<https://github.com/JulianoGaspar/SPSS>



Introdução à análise de dados em Saúde

Medidas de Resumo por tipo de variável



Medidas-resumo
numéricas

TABELA 2.8

Frequências relativas e frequências relativas acumuladas de níveis séricos de colesterol para 2.294 homens dos Estados Unidos, 1976-1980.

Nível de Colesterol (mg/100 ml)	Idades 25-34		Idades 55-64	
	Número de Homens	Frequência Relativa (%) Acumulada	Número de Homens	Frequência Relativa (%) Acumulada
80-119	1,2	1,2	0,4	0,4
120-159	14,1	15,3	3,9	4,3
160-199	41,4	56,7	21,6	25,9
200-239	28,0	84,7	37,3	63,2
240-279	10,8	95,5	22,9	86,1
280-319	3,2	98,7	10,4	96,5
320-359	0,8	99,5	2,9	99,4
360-399	0,5	100,0	0,6	100,0

Medidas resumo
não-numéricas



Introdução à análise de dados em Saúde

Medidas de Resumo

- o As Medidas de Resumo são ferramentas importantes para descrever a distribuição de uma variável quantitativa.
- o São valores que, de certa forma, e de maneira condensada, trazem consigo informações contidas nos dados estatísticos, sejam eles, populacionais ou amostrais.
- o As medidas de resumo nos informam o comportamento geral das observações estudadas.
- o Pode-se dizer que elas são como valores de referência, em torno dos quais, os outros se distribuem.



Introdução à análise de dados em Saúde

Medidas de Resumo

Medidas de
Tendência
Central

Medidas de
posição

Medidas de
Dispersão

Medidas de Resumo

Conjunto de estatísticas descritivas que permitem uma avaliação concisa de grandes quantidades de valores.

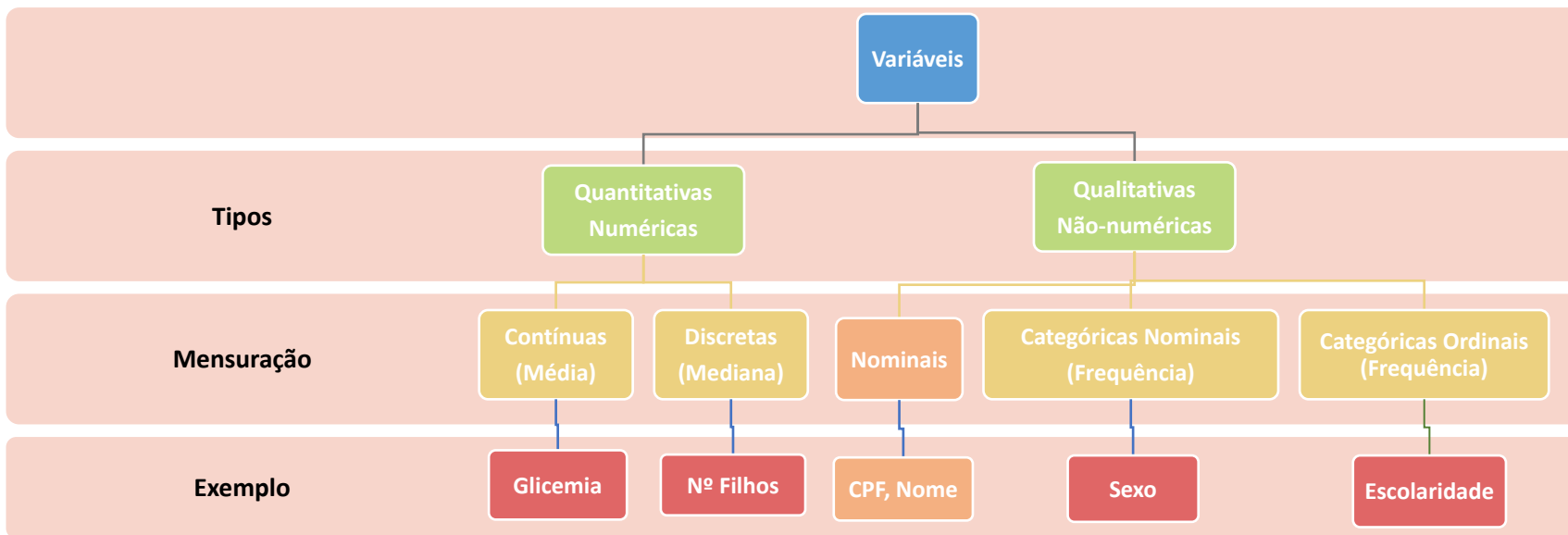




Introdução à análise de dados em Saúde

Tipos de dados e as medidas-resumo

Algumas medidas de resumo são mais indicadas para determinados tipos de Variáveis.





Média

- Soma de todos os valores absolutos / número de observações
- Acompanha-se geralmente de medida de dispersão (Desvio padrão)
- Vantagem: algebricamente definida
- Desvantagens: distorcida por valores extremos

$$\frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}$$

Mediana

- Refere-se ao valor do meio, a partir dos dados ordenados em ordem crescente (p50)
- Acompanha-se geralmente de medida de dispersão (amplitude)
- Se n=ímpar a mediana é o valor do meio. Se n=par, a mediana é a média dos valores centrais
- Vantagem: não é distorcida por valores extremos
- Desvantagens: leva em consideração a ordem e não os valores em si

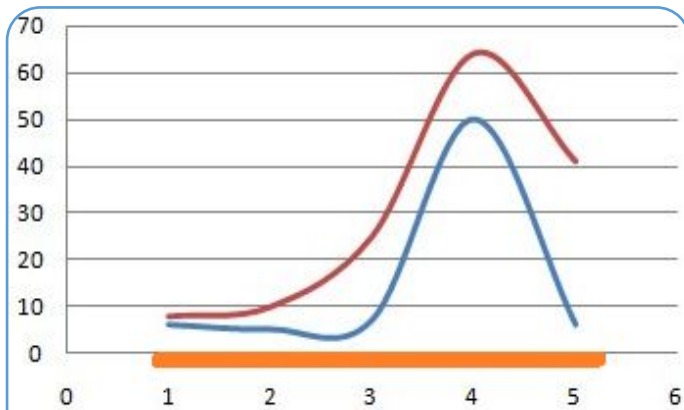
Moda

- É o valor mais frequente
- Pode ser usada para variáveis não-numéricas (categóricas ou nominais)
- O conjunto de dados pode ser amodal, unimodal, bimodal, multimodal



Introdução à análise de dados em Saúde

Principais medidas de dispersão



**Intervalo e
Amplitude**

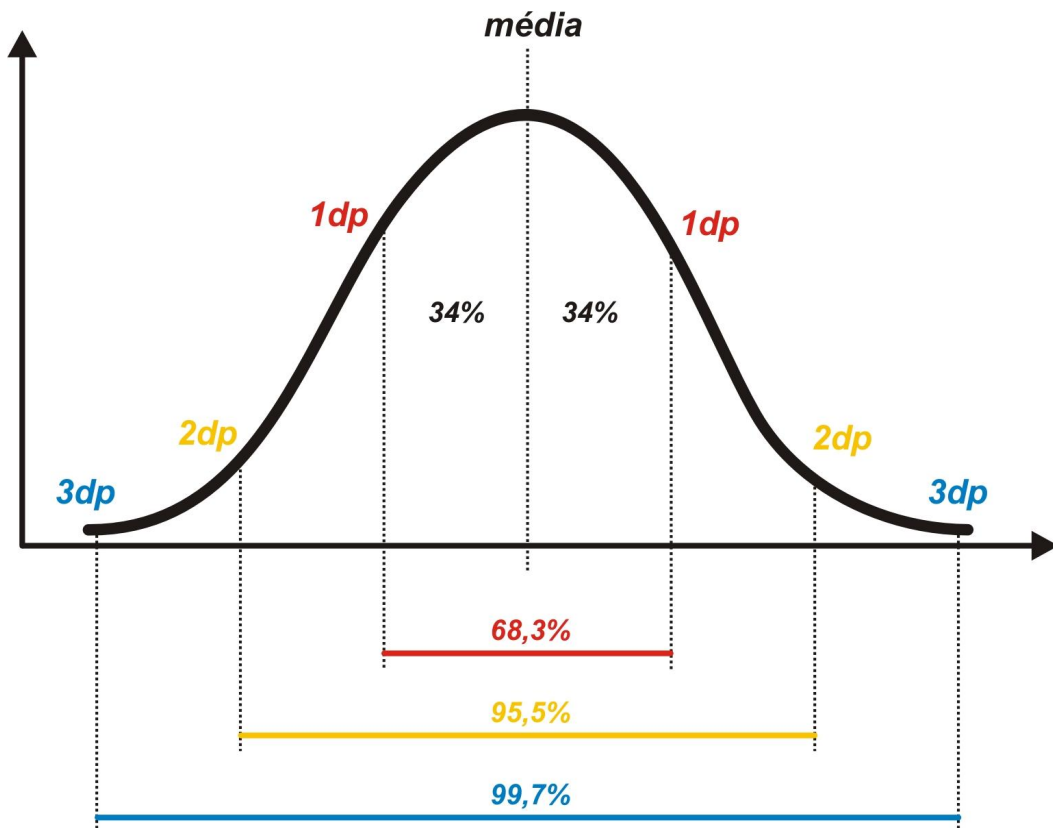


Desvio Padrão



Introdução à análise de dados em Saúde

Principais medidas de dispersão



Desvio padrão (σ)

68,3% ☐ 1

95,5% ☐ 2

99,7% ☐ 3

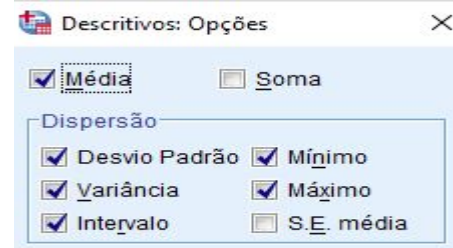


Medida	Vantagens	Desvantagens
Âmbito (amplitude) diferença entre o valor maior e menor	Fácil de calcular	Usa apenas dois valores Distorcido por valores extremos
Variância soma dos quadrados dos desvios à média dividido pelo N° casos menos um	Usa todos os dados Definida algebricamente	A unidade é o quadrado da unidade dos dados Sensível a valores extremos Não apropriada em distribuições enviesadas
Desvio padrão raiz quadrada da variância	Usa todos os dados Definida algebricamente Mesma unidade que os dados Fácil de interpretar	Sensível a valores extremos Não apropriada em distribuições enviesadas

Introdução à análise de dados em Saúde

Medidas de resumo no SPSS

Variáveis numéricas



Estatísticas descritivas

	N	Intervalo	Mínimo	Máximo	Média	Desvio Padrão	Variância
IG Obstetra	1622	42	0	42	38,01	3,193	10,194
N válido (de lista)	1622						



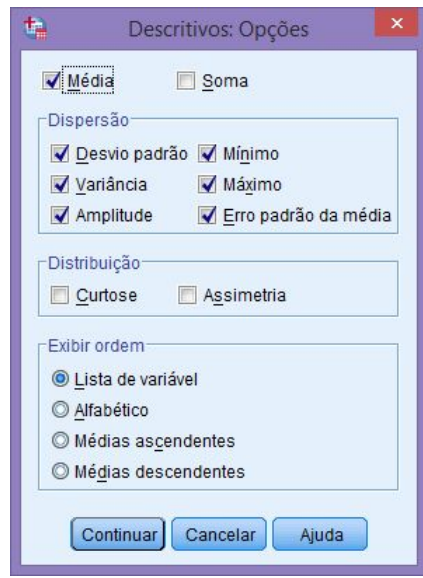
Comando SPSS: Analisar >> Estatística Descritiva >> Descritivos



Introdução à análise de dados em Saúde

Medidas de resumo no SPSS

Variáveis numéricas



Estatísticas descritivas

	N	Range	Mínimo	Máximo	Média		Desvio padrão	Variância
	Estatística	Estatística	Estatística	Estatística	Estatística	Modelo padrão	Estatística	Estatística
SP_PESO_NASCER	337	4215	520	4735	3010,01	32,753	601,262	361515,777
NU_GESTACOES	345	10	1	11	2,15	,079	1,464	2,144
N válido (de lista)	337							

Comando SPSS: Analisar >> Estatística Descritiva >> Descritivos



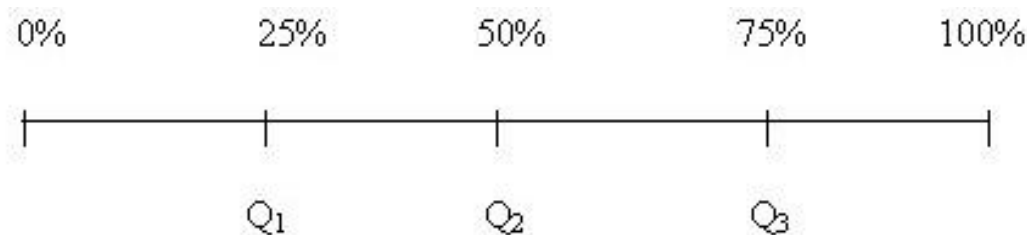
Introdução à análise de dados em Saúde

Medidas de Posição ou Separatrizes

É um tipo de separatriz que divide a série estatística em quatro partes iguais de 25% cada.

Possui três divisórias, que são Q1, Q2 e Q3, significando respectivamente:

- 1º quartil ou quartil inferior
- 2º quartil ou quartil médio
- 3º quartil ou quartil superior



As medidas separatrizes estão ligadas à **Mediana**.

Essas medidas, Quartis – Decis – Percentis, são juntamente com a Mediana, conhecidas pelo nome genérico de separatrizes.

Introdução à análise de dados em Saúde

Outras medidas de tendência central e dispersão

Geralmente usadas para dados numéricos contínuos, populacionais
Servem como padrão (dados de saudáveis tendem à distribuição normal)

Centro: mediana

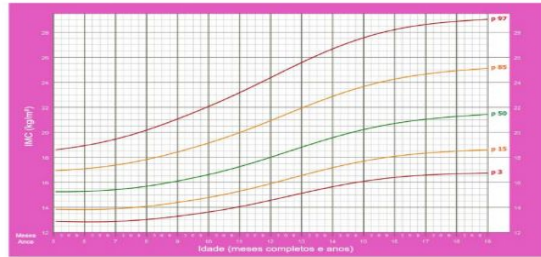
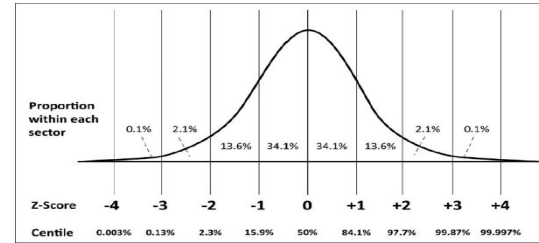


Figura 3. Gráfico de percentil do IMC para meninas entre 5 e 19 anos de idade.

Percentil: mais sensíveis

Centro: media

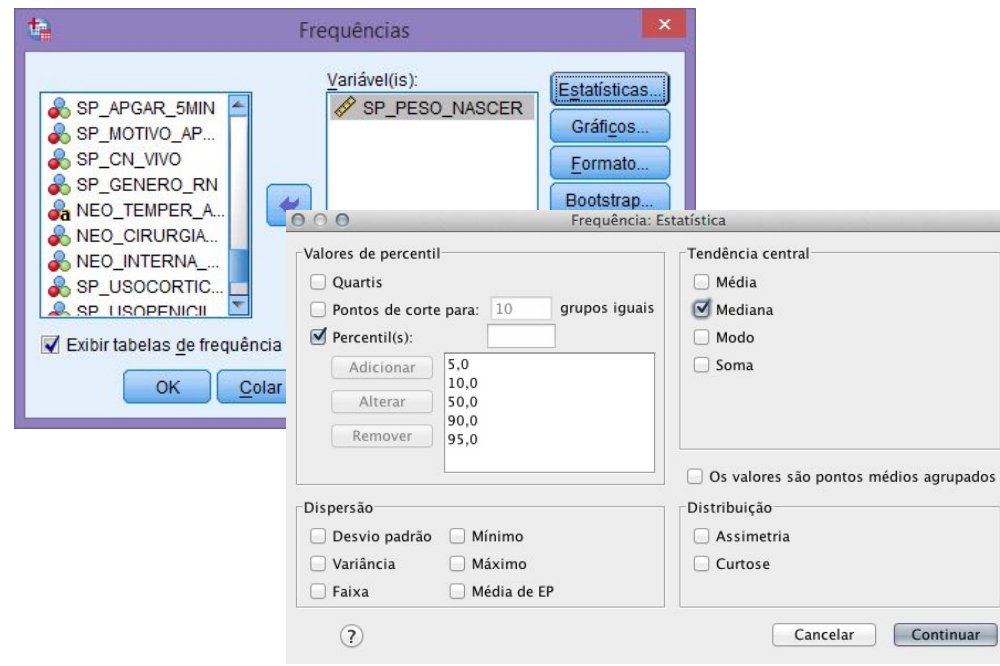


Percentil: mais específicas



Introdução à análise de dados em Saúde

Medidas de resumo no SPSS



Variáveis quantitativas

Statistics		
SP_PESO_NASCER		
N	Valid	337
	Missing	8
Median		3035,00
Percentiles	5	1916,00
	10	2270,00
	50	3035,00
	90	3718,00
	95	3892,00

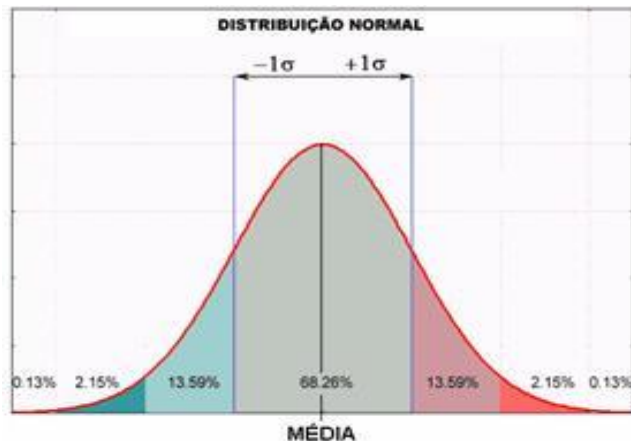
Comando SPSS: Analisar >> Estatística Descritiva >> Frequências



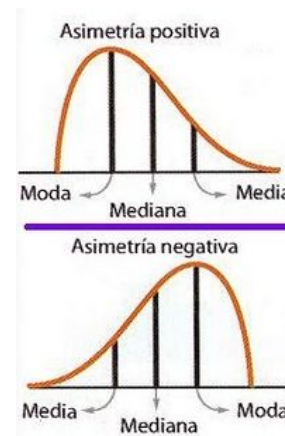
Introdução à análise de dados em Saúde

Gráfico de distribuição de frequência

Histograma de frequência é um gráfico de barras de mesma largura, adjacentes e em ordem crescente de valores. Na horizontal encontram-se as classes de valores e na vertical a sua frequência de ocorrência.



Distribuição simétrica (Gaussiana)

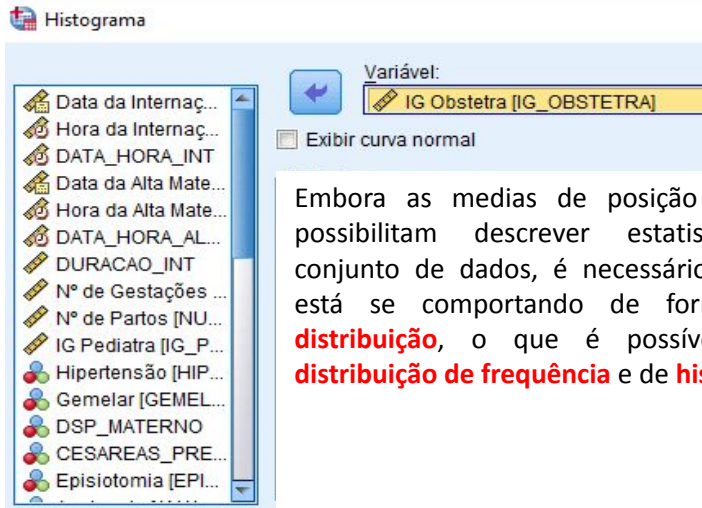


Distribuição assimétrica (não-Gaussiana)

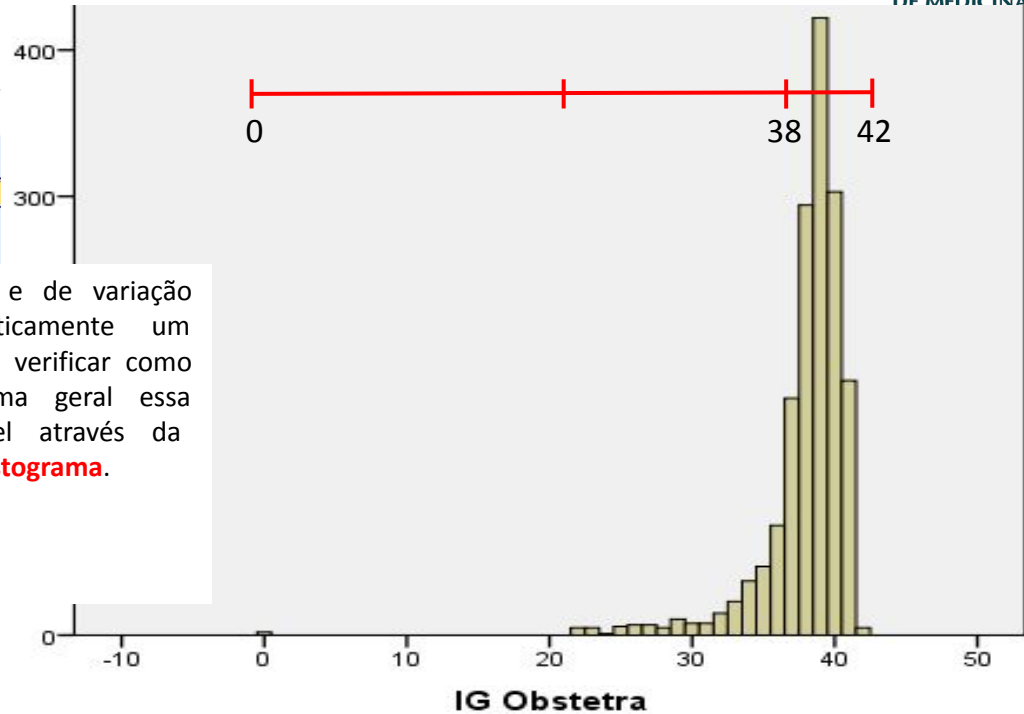
Introdução à análise de dados em Saúde

Medidas de resumo no SPSS

Variáveis numéricas



Embora as medias de posição e de variação possibilitam descrever estatisticamente um conjunto de dados, é necessário verificar como está se comportando de forma geral essa **distribuição**, o que é possível através da **distribuição de frequência** e de **histograma**.

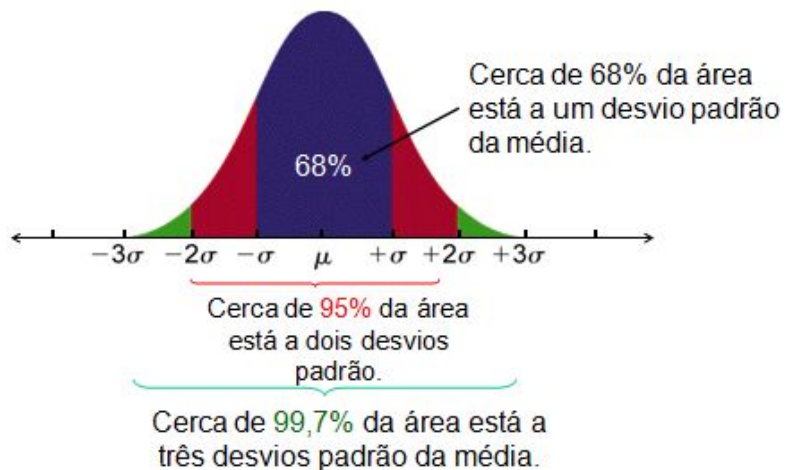


Descritivos

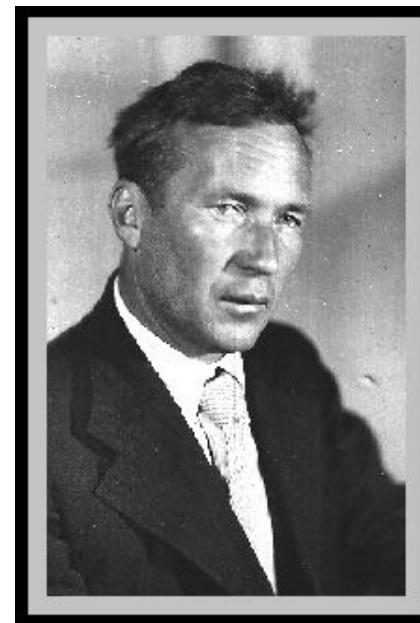
Estatísticas descritivas

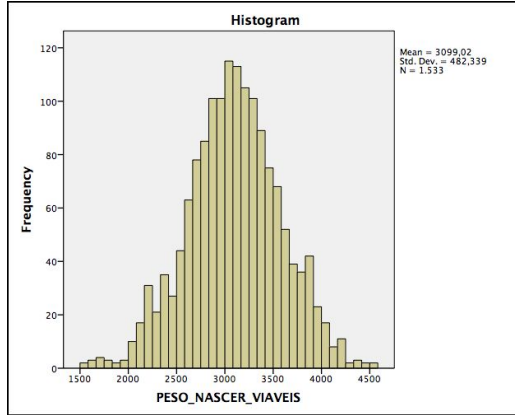
	N	Intervalo	Mínimo	Máximo	Média	Desvio Padrão	Variância
IG Obstetra	1622	42	0	42	38,01	3,193	10,194
N válido (de lista)	1622						

Comando SPSS: Analisar >> Estatística Descritiva >> Descritivos

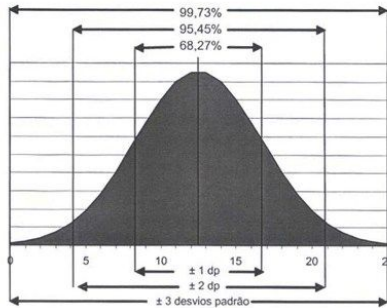


Ho: a distribuição é normal





Distribuição Normal



Características da distribuição normal:

- A curva é simétrica em torno da média
- A média, mediana e a moda coincidem
- As extremidades se estendem infinitamente
- Coeficiente de assimetria e curtose, padronizados pelo seu erro padrão estão entre -1,96 e + 1,96
- Testes de normalidade

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
PESO_NASCER_VIAVEIS	,019	1533	,200 [*]	,998	1533	,089
*. This is a lower bound of the true significance.						
a. Lilliefors Significance Correction						

Ho: a distribuição é normal

- Para amostra com **mais de 30 casos**, recomenda-se o **Kolmogorov-Smirnov** com a correção de Lilliefors.
- Para **amostras pequenas** (< 30 casos), recomenda-se o teste de **Shapiro-Wilk**.



O peso ao nascer tem distribuição de probabilidade normal?

PESO_NASCER



Hipotese nula: a distribuição é normal

Estatística do teste

Valor critico (p)

$p\text{-valor} < 0,05$ (rejeito H_0)

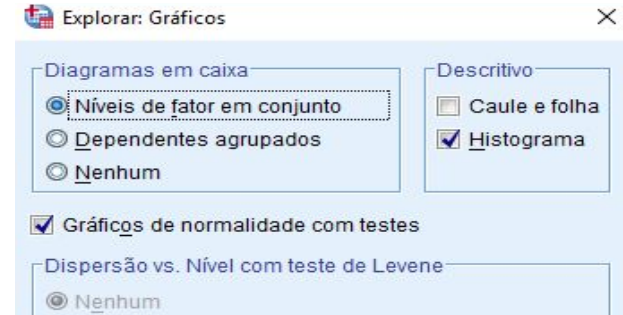
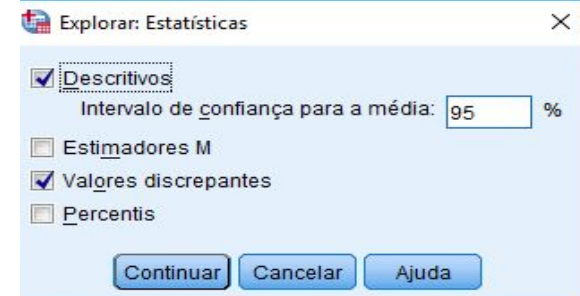
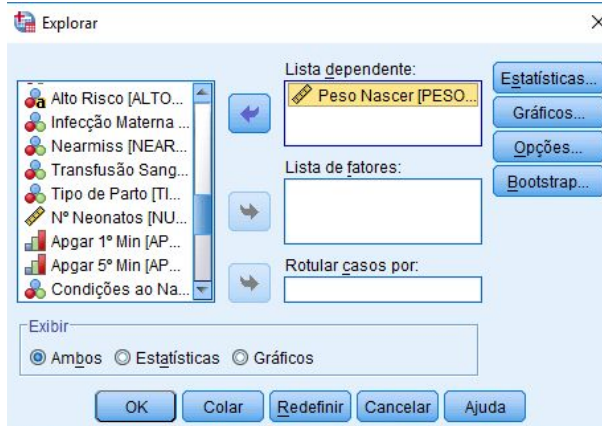
$p\text{-valor} \geq 0,05$ (aceito H_0)

Introdução à análise de dados em Saúde

Medidas de resumo no SPSS – Gráficos Histogramas

Crie um histograma de frequência para a variável **PESO_NASCER**

Variáveis quantitativas



Comando SPSS: Analisar >> Explorar>> Teste de normalidade / histograma (Testes de normalidade)



Introdução à análise de dados em Saúde

Medidas de resumo no SPSS

variável **PESO_NASCER**

Descritivos

			Estatística	Erro Padrão
Peso Nascer	Média		2977,88	16,853
	95% Intervalo de Confiança para Média	Limite inferior	2944,82	
		Limite superior	3010,93	
	5% da média aparada		3023,66	
	Mediana		3065,00	
	Variância		477463,927	
	Desvio Padrão		690,988	
	Mínimo		270	
	Máximo		5625	
	Intervalo		5355	
	Intervalo interquartil		680	
	Assimetria		-1,123	,060
	Curtose		2,538	,119

Valores Extremos

			Número do caso	Valor
Peso Nascer	Mais alto	1	18	5625
		2	219	4920
		3	1522	4820
		4	297	4735
		5	388	4635
	Mais baixo	1	1703	270
		2	1674	275
		3	1708	300
		4	1692	300
		5	13	350

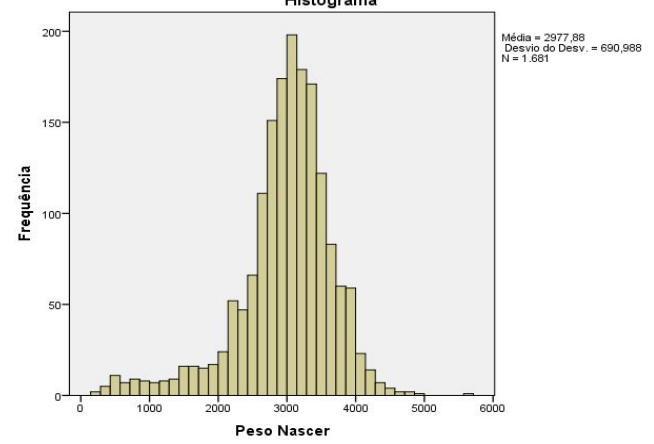
Testes de Normalidade

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estatística	df	Sig.	Estatística	df	Sig.
Peso Nascer	,101	1681	,000	,927	1681	,000

a. Correlação de Significância de Lilliefors

H0: a distribuição é normal
Neste caso, rejeita H0, a distribuição não é normal.

Histograma



Comando SPSS: Analisar >> Explorar>> Teste de normalidade / histograma (Testes de normalidade)



Introdução à análise de dados em Saúde

Medidas de resumo no SPSS – Exercício

Explore as variáveis:

IG_PEDIATRA

DURACAO_INT

NU_PARTOS

Medidas de tendência central (média e mediana)

Medidas de dispersão (desvio padrão e intervalo)

Medidas de posição (quartis, escala de percentil 5, 10, 50, 90 e 95)

Calcular o teste de normalidade

Qual é a melhor maneira de apresentar cada uma destas variáveis e porque?



Comando SPSS: Analisar >> Estatística Descritiva >> Explorar

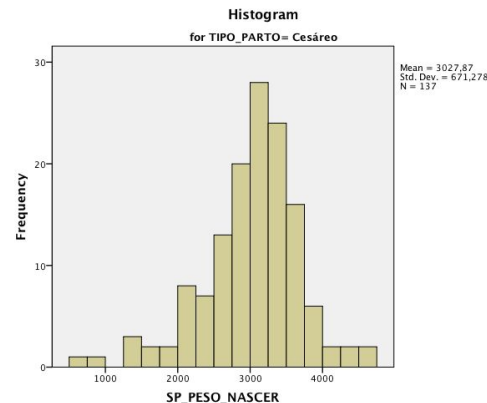
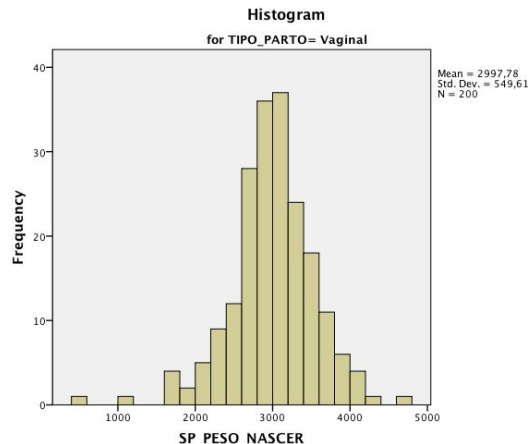
Comando SPSS: Analisar >> Estatística Descritiva >> Frequências

Introdução à análise de dados em Saúde

Peso ao nascer X Tipo de parto

Descriptives

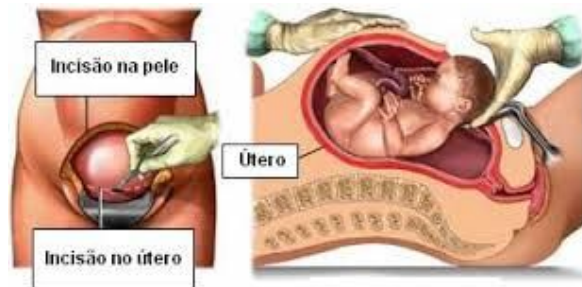
TIPO PARTO				Statistic	Std. Error
SP_PESO_NASCE	Vaginal	Mean		2997,79	38,863
		95% Confidence Interval for Mean	Lower Bound	2921,15	
			Upper Bound	3074,42	
		5% Trimmed Mean		3012,92	
		Median		3012,50	
		Variance		302071,044	
		Std. Deviation		549,610	
		Minimum		525	
		Maximum		4635	
		Range		4110	
		Interquartile Range		601	
		Skewness		-,577	,172
		Kurtosis		2,326	,342
		Mean		3027,87	57,351
	Cesáreo	95% Confidence Interval for Mean	Lower Bound	2914,45	
			Upper Bound	3141,28	
		5% Trimmed Mean		3054,89	
		Median		3120,00	
		Variance		450614,556	
		Std. Deviation		671,278	
		Minimum		520	
		Maximum		4735	
		Range		4215	
		Interquartile Range		708	
		Skewness		-,803	,207
		Kurtosis		1,843	,411





Crie tabela de frequência para tipo de parto e condições de alta da mulher

DESFECHO_MATERN0					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	15	,9	,9	,9
	1	1660	98,8	98,8	99,7
	2	2	,1	,1	99,8
	3	2	,1	,1	99,9
	4	1	,1	,1	100,0
	Total	1680	100,0	100,0	



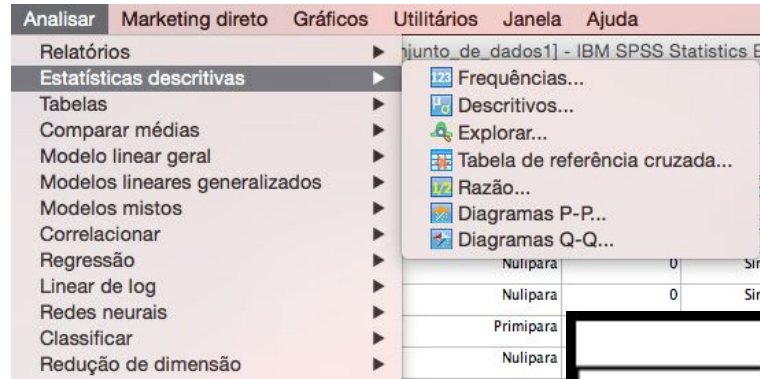
TIPO_PARTO					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Parto cesariana	624	37,1	37,1	37,1
	Parto normal	1056	62,9	62,9	100,0
	Total	1680	100,0	100,0	



Introdução à análise de dados em Saúde

Tabelas de frequência

- Criar tabela de frequência simples para as variáveis:
- Número de cesarianas prévias (numérica discreta)



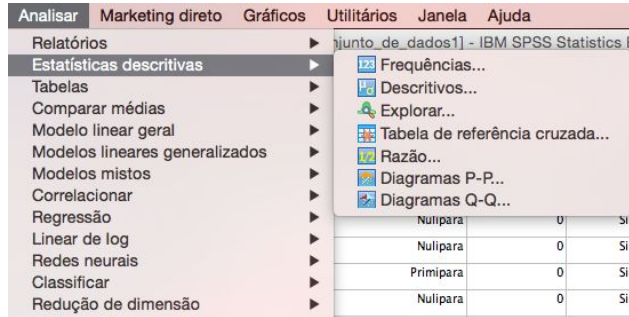
NU_CESAREAS_PREVIAS					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1350	80,4	80,4	80,4
	1	253	15,1	15,1	95,4
	2	62	3,7	3,7	99,1
	3	8	,5	,5	99,6
	4	6	,4	,4	99,9
	5	1	,1	,1	100,0
Total		1680	100,0	100,0	

Introdução à análise de dados em Saúde

Tabelas de frequência

Criar tabela de frequência simples para as variáveis:

- Idade gestacional, agrupada em:
 - Termo > 38 semanas
 - Termo precoce 37 a 38 semanas
 - Prematuro (Pre-termo) < 37 semanas



TERMO_TERMOPRECOCE_PREMATURO					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Termo	914	54,4	55,1	55,1
	Termo precoce	463	27,6	27,9	83,1
	Prematuro	281	16,7	16,9	100,0
	Total	1658	98,7	100,0	
Missing	System	22	1,3		
Total		1680	100,0		

Comando SPSS: Analisar >> Estatística Descritiva >> Frequências



Em alguns casos é necessário fazer análises de casos selecionados:

- Obter informações apenas dos doentes / ou apenas dos saudáveis

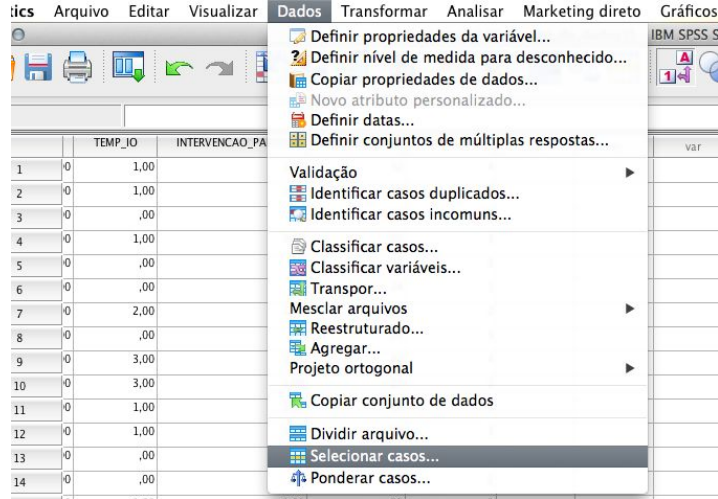
Sorteio aleatório :

- Amostragem aleatória em grandes bases de dados

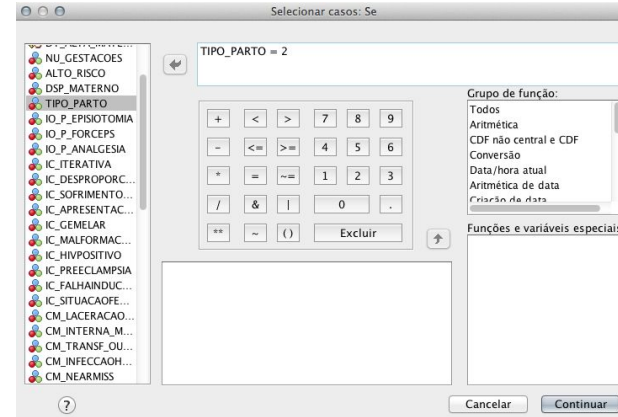
Ordenar a base de dados segundo uma das variáveis:

- Consistência de dados / Avaliar outliers

Introdução à análise de dados em Saúde



Quando é necessário selecionar apenas alguns casos:



Exemplos:

1. Selecionar apenas os partos cesáreos

- Explorar **IG_OBSTETRA**
- Voltar para a opção TODOS OS CASOS e usar SPLIT
- Refazer o comando explorar **IG_OBSTETRA** para partos normais e para partos cesáreos separadamente

Comandos SPSS: Dados >> Selecionar Casos



Introdução à análise de dados em Saúde

Explorar IG_Obstetra

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
TEMPO_INT	139	99,3%	1	0,7%	140	100,0%

Descriptives

			Statistic	Std. Error
TEMPO_INT	Mean		5,01	,981
	95% Confidence Interval for Mean	Lower Bound	3,07	
		Upper Bound	6,95	
	5% Trimmed Mean		3,22	
	Median		3,00	
	Variance		133,746	
	Std. Deviation		11,565	
	Minimum		2	
	Maximum		118	
	Range		116	
	Interquartile Range		1	
	Skewness		7,850	,206
	Kurtosis		70,026	,408

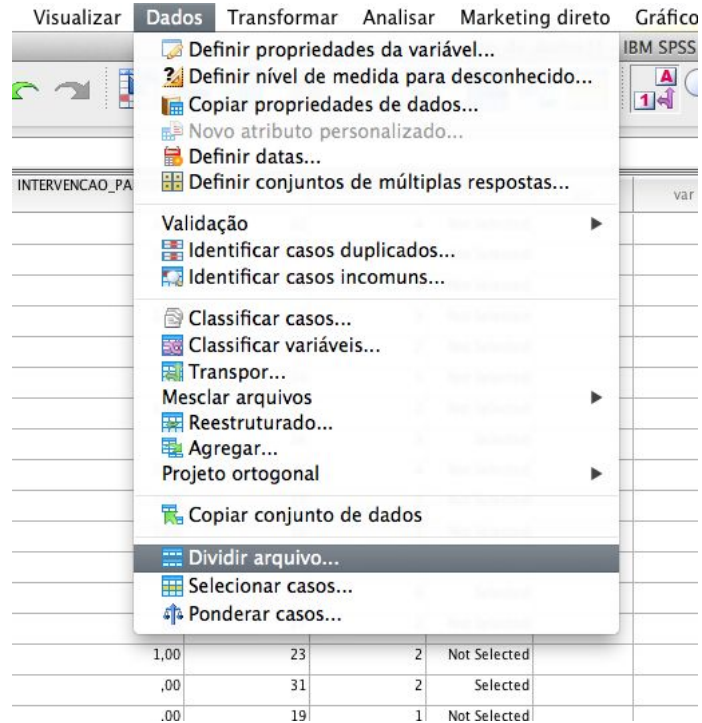
Neste caso, apenas os partos
cesariana são contabilizados



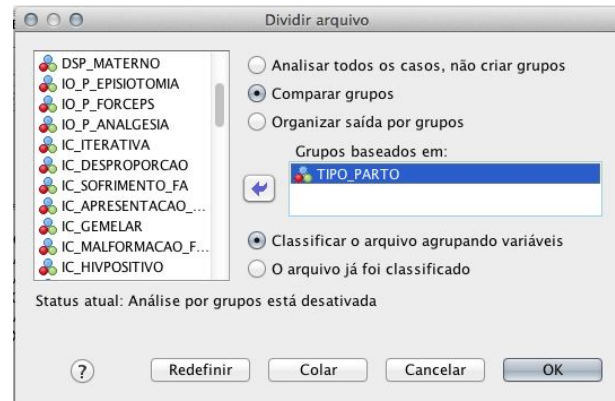
Introdução à análise de dados em Saúde

Comando: Split (Dividir planilha)

Universidade Federal de Minas Gerais



Obter informações comparativas por grupo de interesse



Comandos SPSS: Dados >> Dividir Arquivo



Introdução à análise de dados em Saúde

Explorar IG_INTERNACAO

Descriptives

TIPO PARTO			Statistic	Std. Error
1	TEMPO_INT	Mean	21,70	19,696
		95% Confidence Interval for Mean	Lower Bound -17,13 Upper Bound 60,54	
		5% Trimmed Mean	1,86	
		Median	2,00	
		Variance	79140,782	
		Std. Deviation	281,320	
		Minimum	-1	
		Maximum	4020	
		Range	4021	
		Interquartile Range	1	
		Skewness	14,282	,170
		Kurtosis	203,989	,339
2	TEMPO_INT	Mean	5,01	,981
		95% Confidence Interval for Mean	Lower Bound 3,07 Upper Bound 6,95	
		5% Trimmed Mean	3,22	
		Median	3,00	
		Variance	133,746	
		Std. Deviation	11,565	
		Minimum	2	
		Maximum	118	
		Range	116	
		Interquartile Range	1	
		Skewness	7,850	,206
		Kurtosis	70,026	,408

Case Processing Summary

TIPO PARTO		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
1	TEMPO_INT	204	99,5%	1	0,5%	205	100,0%
2	TEMPO_INT	139	99,3%	1	0,7%	140	100,0%

Agora, o tempo de internação será descrito em cada grupo separadamente:

- parto normal (1)
- partos cesariana (2)

Comando SPSS: Analisar >> Estatística descritiva >> Explorar



**"We are here to learn,
to make a difference and
to have fun."**

**William Edwards
DEMING**





Obrigado!



Prof. D.r Juliano Gaspar

julianogaspar@gmail.com

<http://lattes.cnpq.br/3926707936198077>