Capítulo 3
Organização e síntese de dados

3.1. Introdução

Neste capítulo, serão abordados alguns aspectos que podem ser utilizados para organizar, resumir e descrever um conjunto de dados. Os elementos básicos necessários para esta abordagem são: *tabelas de freqüência*, *gráficos* e *medidas descritivas*. Vale ressaltar que tais elementos devem considerar a natureza dos dados.

As técnicas estudadas neste capítulo permitem detectar anomalias e inconsistência nos dados, apresentá-los de forma que a tabela e a visualização proporcionem maior compreensão na interpretação e caracterizar o perfil dos pacientes.

3.2 Montagem do banco de dados e Classificação das Variáveis

A seguir será descrito, resumidamente, um exemplo utilizado no restante do capítulo para ilustrar os métodos estatísticos. Nota-se que este exemplo foi adaptado, ou seja, as informações contidas no banco de dados são hipotéticas, a fim de atingir os objetivos propostos para o capitulo.

Trata-se de um estudo retrospectivo, caso e controle, com informações dos prontuários de 39 pacientes com câncer de mama. Definiram-se os casos como aquelas pacientes associadas à gravidez ou lactação e os controles como aquelas pacientes que não estavam associadas à gravidez ou lactação. Todas estas pacientes foram acompanhadas no período compreendido entre janeiro de 1980 e dezembro de 2000. Integram o prontuário as seguintes *variáveis* listadas na tabela 3.1.

Após sua coleta nos prontuários, os dados devem ser inseridos em uma planilha eletrônica, em que cada linha indica um paciente e cada uma das colunas denomina uma *variável* que, como vimos anteriormente, é uma característica de interesse que é medida em cada paciente da amostra ou população. A tabela 3.2 representa a planilha das pacientes com câncer de mama contendo 39 linhas e 10 colunas. A última coluna, denominada sg1 expressa o intervalo de tempo desde a data do diagnóstico até a data da última consulta, em meses.

Tabela 3.1 - Variáveis medidas no estudo caso-controle: prognóstico do câncer de mama associado à gravidez ou lactação

NP	Número de prontuário
Idade da Paciente (IDE)	Medida em anos
Presença da gravidez, Caso e Controle	0 - Controle 1 - Caso
Data do diagnóstico (DDiag)	dd/mm/aa
Grau de Malignidade (GM)	0-G1 1-G2 2-G3 9-Ignorado
Tamanho do Tumor (T)	0-T0 1-T1 2-T2 3-T3 4-T4 5-TX 9-Ignorado
Número de Nódulos Linfáticos Axilares acometidos (N)	Medido em valor absoluto
Data da última consulta (DUCONS)	dd/mm/aa
Estado Atual (FUP)	0-Viva 1-Óbito

Fonte: dados hipotéticos.

De acordo com a tabela 3.2, por exemplo, a variável idade assume valores numéricos em anos. A presença de gravidez nas pacientes foi codificada como 1 se estiver associada à gravidez e 0 se não estiver associada. Isto não significa que a variável caso-controle apresente valores numéricos como da variável idade. Portanto essas duas variáveis têm naturezas distintas no que tange aos seus valores. Mediante este acontecimento, o primeiro passo para realizar as análises estatísticas será classificar a natureza das variáveis como quantitativa, qualitativa ou datas, como definidas no capítulo 2.

Podemos notar, no entanto, que a classificação da natureza das variáveis depende de certas particularidades. Exemplificando, a variável idade, medida em anos e meses, pode ser considerada como qualitativa ordinal, caso seja apurada no banco de dados em faixa etária (0 a 5 anos, 6 a 10 anos e acima de 10 anos). Por outro lado, a variável idade, medida em anos e meses, pode ser considerada como quantitativa discreta, caso seja apurada no banco de dados em anos completos.

As demais variáveis, da maneira que se encontram no banco de dados, podem ser classificadas como qualitativas (SCC, FUP, GM e T), datas (DDIAG, DUCONS) e quantitativa (N).

3.3 Tabelas de Freqüências e Gráficos

Recebe a denominação dados brutos, à reunião de toda a informação resultante da coleta de dados, e armazenada em uma planilha eletrônica. Evidentemente, extrair de imediato a informação a partir dos dados brutos seria uma tarefa árdua caso o número de linhas e de colunas da planilha fosse elevado.

Tabela 3.2 - Planilha do Banco de dados no estudo de Prognóstico do câncer de mama associado à gravidez ou lactação

NP	IDE	SCC	DDIAG	GM	T	N	DUCONS	FUP	sg1
Paciente 1	23,3	1	18-mar-1983	9	4	11	02-mar-1990	1	83,48
Paciente 2	34,2	1	15-dez-1985	9	1	5	14-abr-1987	1	15,93
Paciente 3	39,9	1	22-jul-1991	2	4	6	28-abr-1993	1	21,22
Paciente 4	40,6	1	20-mar-1994	9	2	0	04-out-2005	0	138,51
Paciente 5	41,1	1	06-set-1995	1	2	7	03-nov-2004	0	109,93
Paciente 6	27,4	1	13-ago-1980	9	4	11	02-set-1981	1	12,65
Paciente 7	35,8	1	18-fev-1981	9	4	10	22-mai-1982	1	15,05
Paciente 8	44,7	1	04-mai-1981	9	2	1	17-mar-2005	0	286,42
Paciente 9	37,6	1	16-abr-1984	9	4	9	27-abr-1985	1	12,35
Paciente 10	34,0	1	27-jun-1984	9	4	10	04-nov-1984	1	4,27
Paciente 11	31,0	1	12-ago-1985	9	4	10	09-abr-1987	1	19,88
Paciente 12	28,7	1	04-abr-1986	9	3	11	28-fev-1987	1	10,84
Paciente 13	36,1	1	29-dez-1988	9	3	0	15-jul-1990	1	18,50
Paciente 14	34,9	1	31-mai-1990	9	4	6	12-jul-1990	1	1,38
Paciente 15	39,0	1	07-mar-1991	9	3	4	11-mar-1994	1	36,14
Paciente 16	33,2	1	24-out-1991	9	2	2	14-jul-2004	1	152,67
Paciente 17	31,1	1	09-nov-1992	2	2	11	18-dez-1993	1	13,27
Paciente 18	37,5	1	15-jun-1994	2	3	5	26-fev-2004	0	116,40
Paciente 19	29,6	1	02-ago-1995	0	2	3	26-jan-1997	1	17,84
Paciente 20	35,0	1	18-abr-1996	0	1	1	23-mar-2005	0	107,14
Paciente 21	30,2	1	04-set-1987	9	4	7	02-out-1988	1	12,94
Paciente 22	40,8	1	04-nov-1999	1	4	11	21-set-2005	0	70,57
Paciente 23	37,2	0	15-jan-1981	9	5	8	28-jan-1984	1	36,40
Paciente 24	32,4	0	29-set-1984	1	1	0	02-mar-1998	0	161,05
Paciente 25	36,3	0	21-mar-1985	9	3	6	06-mai-1987	1	25,49
Paciente 26	35,0	0	01-out-1985	0	5	10	09-jan-1987	1	15,28
Paciente 27	35,0	0	05-nov-1986	9	9	4	29-nov-1998	1	144,79
Paciente 28	33,3	0	09-mar-1984	2	4	11	07-nov-1985	0	19,98
Paciente 29	32,6	0	23-fev-1987	2	3	2	03-jul-1990	1	40,28
Paciente 30	33,1	0	30-jun-1993	9	1	0	25-out-2005	0	147,84
Paciente 31	43,4	0	15-mai-1989	0	2	1	05-jan-1998	1	103,72
Paciente 32	43,2	0	25-ago-1995	1	1	6	06-dez-2004	0	111,41
Paciente 33	33,7	0	31-mai-1979	9	4	3	23-jan-1981	1	19,81
Paciente 34	43,1	0	10-set-1981	9	2	0	12-ago-1982	1	11,04
Paciente 35	36,5	0	03-abr-1982	9	3	0	12-jun-1982	1	2,30
Paciente 36	37,9	0	04-dez-1984	9	4	0	15-ago-1988	1	44,35
Paciente 37	38,7	0	13-fev-1985	2	2	0	03-mai-1985	0	2,60
Paciente 38	35,5	0	26-jul-1982	2	4	0	30-jun-1983	0	11,14
Paciente 39	35,2	0	15-jan-1982	2	9	9	26-mai-1983	1	16,30

Para melhor análise dos dados é necessário apresentá-los e descrevê-los de forma organizada e sucinta. As ferramentas utilizadas para esta tarefa são as tabelas, os gráficos e as medidas numéricas. Passaremos a estudá-los de acordo com a natureza dos dados.

3.3.1 Variáveis Qualitativas (Ordinais e Nominais)

Com base no banco de dados da tabela 3.2, a variável caso-controle, classificada como variável qualitativa nominal, será resumida por meio de uma tabela de freqüência. Denomina-se tabela de freqüência uma tabela que contém as categorias da variável representada em cada linha, Caso e Controle, neste exemplo. Para cada categoria da variável associamos na primeira coluna a contagem de ocorrências (freqüência absoluta) e para a segunda coluna, relacionamos em cada categoria os percentuais que essas contagens representam do total (freqüência relativa). Esse tipo de tratamento dos dados representa distribuição de freqüência das pacientes segundo a variável Caso-Controle, como descrito na tabela 3.3.

Tabela 3.3 - Distribuição da amostra segundo variável Caso-Controle

Status Caso-Controle	Freqüência Absoluta (n)	Freqüência Relativa (%)
Caso	22	56%
Controle	17	44%
Total	39	100.0%

Fonte: Dados da pesquisa

Compõe o banco de dados da tabela 3.2, uma amostra de 39 pacientes com câncer de mama composta por 22 mulheres grávidas (56%) e 17 mulheres sem a presença de gravidez (44%). A tabela 3.3 exibe essa distribuição.

Observe que, para variáveis cujas categorias apresentam ordenação (qualitativas ordinais), as linhas da tabela de freqüência devem ser dispostas na ordem existente das categorias. Nesse caso, faz sentido adicionar duas colunas contendo as *freqüências acumuladas* (*absoluta* e *relativa*). A freqüência acumulada até uma determinada categoria é calculada pela soma das freqüências de todas as categorias da variável, menores ou iguais à categoria considerada. Ilustrando, até um tamanho de tumor classificado por T4, foram encontrados 35 pacientes, o que corresponde 89,7% do total (tabela 3.4).

Tabela 3.4 - Distribuição da amostra segundo o tamanho do tumor

Tamanho do Tumor	Freqüência Absoluta (n)	Freqüência Relativa (%)	Freqüência Absoluta Acumulada(n)	Freqüência Relativa Acumulada (%)
T1	5	12,8%	5	12,8%
T2	9	23,1%	14	35,9%
T3	7	17,9%	21	53,8%
T4	14	35,9%	35	89,7%
Tx	2	5,1%	37	94,9%
Ignorado	2	5,1%	39	100,0%
Total	39	100,0%	----	----

Fonte: Dados da pesquisa

A utilização de recursos visuais na elaboração de gráficos para ilustrar as tabelas de freqüências pode ser mais facilmente compreendida, permitindo a interpretação rápida das suas principais características. Em função disto, abordaremos, neste momento, dois tipos de gráficos para variáveis qualitativas (*gráfico de setor* e *gráfico de colunas*).

O *gráfico de setor*, popularmente conhecido como gráfico de pizza ou de torta, representado em um sistema de coordenadas polares, consiste na divisão de um disco em setores circulares

correspondentes às freqüências de cada categoria da variável analisada. Como exemplo, mostramos na figura 3.1 o gráfico de setor para a variável caso-controle, obtida a partir da tabela 3.3. Repare que as informações da figura 3.1 são as mesmas da tabela 3.3.

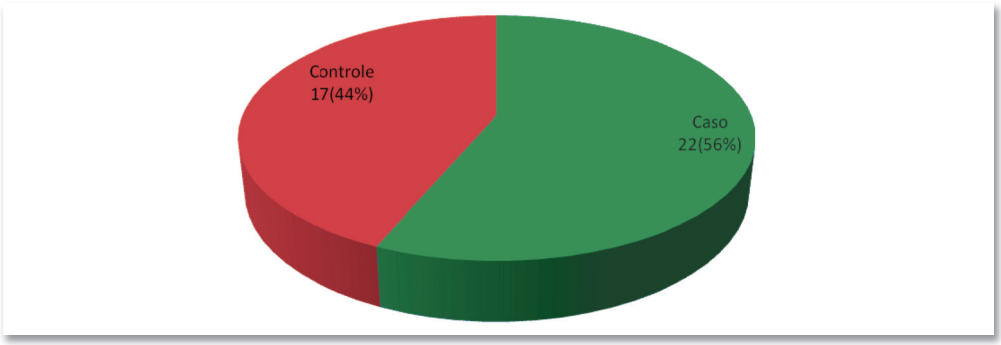

FIGURA 3.1 - Distribuição da amostra segundo a variável Caso-Controle

O gráfico de colunas é representado por um plano cartesiano onde no eixo das abscissas estão representadas as categorias da variável, enquanto no eixo das ordenadas estão representadas as freqüências (absoluta ou relativa). Neste gráfico, cada coluna representa uma categoria com altura associada a sua freqüência (absoluta ou relativa). A figura 3.2 apresenta o gráfico de colunas para a variável tamanho do tumor, obtida a partir da tabela 3.4. Note que as informações da figura 3.2 são as mesmas da tabela 3.4.

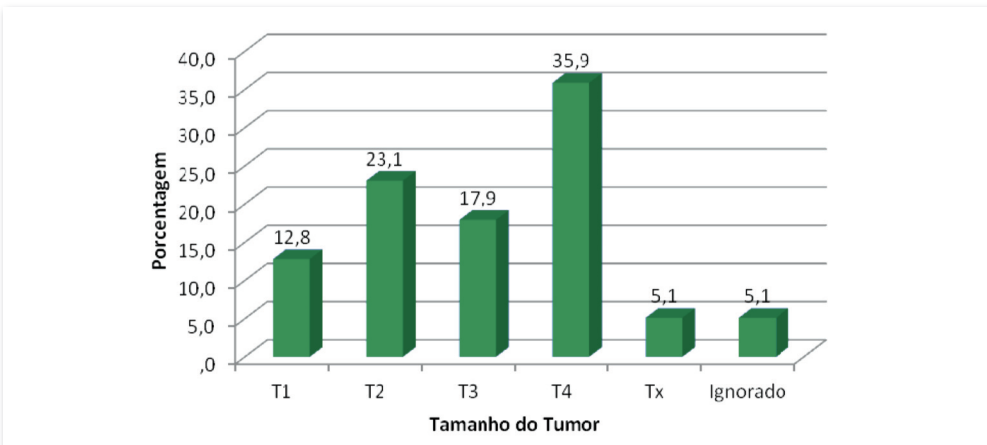

FIGURA 3.2 - Gráfico de Colunas segundo o tamanho do tumor

Um ponto importante a se dizer a respeito de ambos os gráficos é que as freqüências relativas das categorias devem somar 100%. Além disso, a construção do gráfico de setor se adapta melhor para variáveis qualitativas nominais, enquanto para variáveis qualitativas ordinais a sugestão seria o gráfico de colunas.

3.3.2 Variáveis Quantitativas (Discretas e Contínuas)

Particularmente, quando nos deparamos em situações em que a variável quantitativa discreta apresenta poucos valores, é comum adotarmos o mesmo procedimento realizado anteriormente, para as variáveis qualitativas ordinais, assumindo que cada valor é uma categoria e que exista uma

ordem natural entre as categorias. Exemplificando, a tabela 3.5 indica a distribuição do número de nodos linfáticos axilares acometidos, que assumiu onze valores distintos.

Tabela 3.5 - Número de nodos linfáticos axilares acometidos nas pacientes com câncer de mama

Número de Nodos Linfáticos Axilares Acometidos	Freqüência Absoluta (n)	Freqüência Relativa (%)	Freqüência Absoluta Acumulada(n)	Freqüência Relativa Acumulada (%)
0	9	23,1%	9	23,1%
1	3	7,7%	12	30,8%
2	2	5,1%	14	35,9%
3	2	5,1%	16	41,0%
4	2	5,1%	18	46,2%
5	2	5,1%	20	51,3%
6	4	10,3%	24	61,5%
7	2	5,1%	26	66,7%
8	1	2,6%	27	69,2%
9	2	5,1%	29	74,4%
10	4	10,3%	33	84,6%
11	6	15,4%	39	100,0%
Total	39	100,0%	----	----

Fonte: Dados da pesquisa

Analisando a tabela 3.5 e a figura 3.3, o maior percentual de nodos linfáticos axilares acometidos nas pacientes, é de 23,1% que corresponde a 0 nodos (nenhum nodo). Compõem o percentual restante, 15,4% de pacientes com 11 nodos, 10,3% de pacientes com 10 nodos, 10,3% de pacientes com 6 nodos e 7,7% de pacientes com 1 nodo, entre outros descritos naquela figura.

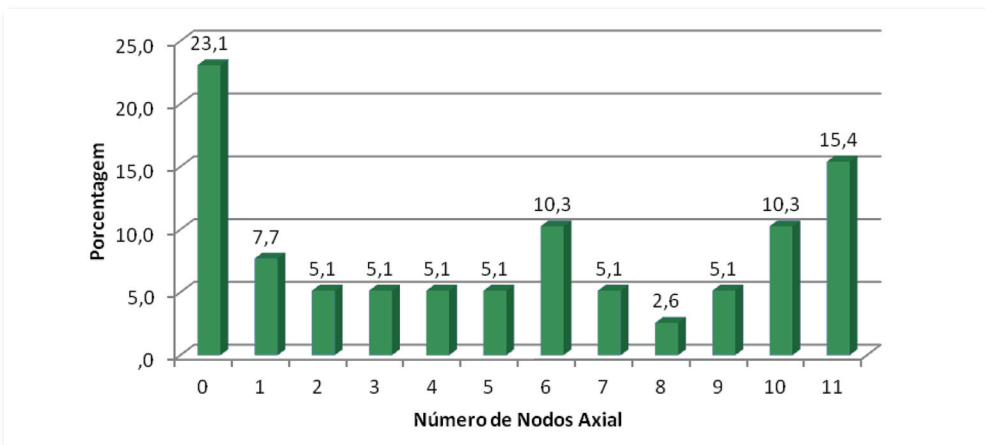

FIGURA 3.3 - Distribuição do número de nodos linfáticos axilares nas pacientes com câncer de mama

Por outro lado, se a variável é contínua ou, se é discreta, mas assume um grande número de valores distintos, considerar cada valor como uma categoria na tabela de freqüência e no gráfico de colunas ficaria inviável. Nestes casos, para se ter uma melhor visualização do seu comportamento de modo a facilitar sua compreensão, é conveniente agrupar os valores em classes ou intervalos. Normalmente, essas classes contém intervalos iguais.

Uma questão polêmica quanto à construção da tabela de freqüência para variáveis quantitativas seria a determinação do número de classes e a *amplitude da classe*. Repare que a distribuição de freqüência pode ser diferente quando mudamos o número e a *amplitude de classes* da tabela. Amplitudes muito grandes para as classes resumem demais a informação dos dados, pois poucas classes são construídas. Entretanto, amplitudes muito pequenas gerariam muitas classes, dificultando a interpretação dos dados. Uma sugestão para estabelecer o número de classes, adequadamente, é utilizar a fórmula desenvolvida pelo matemático Sturges; muitos programas estatísticos adotam este critério. Portanto toma-se como número de classes o inteiro mais próximo encontrado pela seguinte fórmula:

Fórmula de Sturges: $i = 1 + 3{,}3 \log n$

Onde i = número de classes
n = número total de dados
log= *logaritmo na base 10*

Esta fórmula é utilizada como referencial, mas ajustes no número das classes são permitidos para tornar a tabela mais clara.

A tabela 3.6 ilustra a representação da variável quantitativa idade da Tabela 3.2 em uma variável qualitativa faixa etária.

Tabela 3.6 - Freqüência para Idade

Faixa Etária	Freqüência Absoluta (n)	Freqüência Relativa (%)	Freqüência Absoluta Acumulada(n)	Freqüência Relativa Acumulada (%)
22⊢ 26	1	2,56 %	1	2,56 %
26⊢ 30	3	7,69 %	4	10,26 %
30⊢ 34	9	23,08 %	13	33,33 %
34⊢ 38	16	41,03 %	29	74,36 %
38⊢ 42	6	15,40 %	35	89,76 %
42⊢ 46	4	10,24 %	39	100,00 %
Total	39	100,00 %	----	----

Fonte: Dados da pesquisa

Em relação aos elementos da tabela de freqüência da Tabela 3.6, podemos enumerar as classes, que são os agrupamentos de valores num intervalo de abrangência. Para o exemplo da Tabela 3.6 encontramos seis classes. Cada classe é constituída de um *limite inferior* e um *limite superior*. O símbolo "⊢" estabelece inclusão do valor do limite inferior e exclusão do valor do limite superior num intervalo de classe. A *amplitude* de um intervalo de classe é a diferença entre o limite superior e inferior de uma classe, que, nesse exemplo, é 4. A *freqüência absoluta* é a quantidade de observações de uma classe. Finalizando, a *freqüência relativa* é obtida em termos percentuais da *freqüência absoluta.*

A representação visual da distribuição de freqüência de uma variável quantitativa é realizada por meio de um gráfico denominado *histograma*, mostrado na Figura 3.4. *Histograma* é um conjunto de retângulos justapostos com as bases sobre um eixo dividido em classes do mesmo tamanho e altura igual à freqüência absoluta ou relativa da classe correspondente.

Note que é opcional a determinação da freqüência absoluta ou freqüência relativa na construção do histograma, pois a escolha não muda a forma da distribuição. É preferível o uso da freqüência relativa no histograma, pois ela facilita a comparação com outros histogramas, ainda que apresentem tamanhos de amostras distintos. Outra vantagem do uso da freqüência relativa é estabelecer uma relação entre o histograma e a *função de distribuição Normal*.

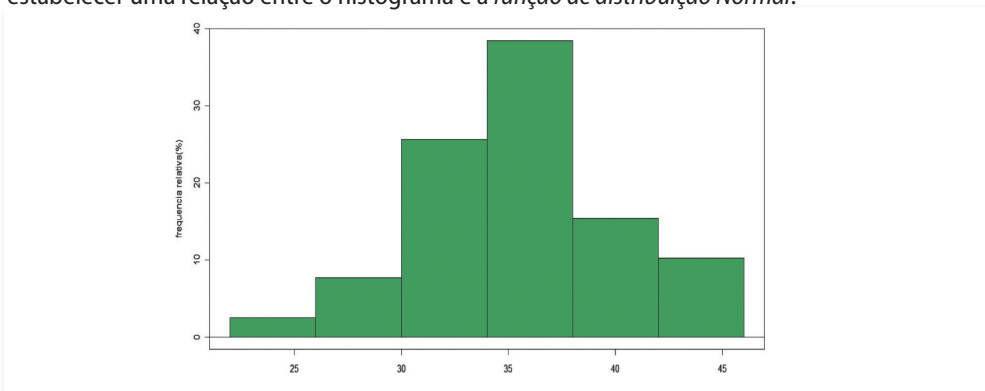

FIGURA 3.4 - Histograma da idade

Os resultados apontaram, conforme mostram a tabela 3.6 e a figura 3.4, que 64,11% das pacientes com câncer de mama, nesta amostragem, possuem idade entre 30 a 38 anos, sendo que deste percentual, 41,03% apresentam idade entre 34 a 38 anos.

Ao se construir o histograma da idade na figura 3.4, obtém-se uma poligonal, aproximadamente, simétrica. Em situações deste tipo, é comum adotarmos a *função de distribuição Normal* (ou *gaussiana*) para descrever o fenômeno estudado. O objetivo de se aproximar uma função de densidade aos dados (neste exemplo utilizou-se a função normal) é devido à facilidade do cálculo de área e esta área corresponde à probabilidade de interesse. A figura 3.5 ilustra dados hipotéticos de idade de pacientes sendo ajustados pela curva da *distribuição normal*; nela está assinalado que a probabilidade de pacientes com idade igual ou maior do que 35 anos é dada pela área sombreada.

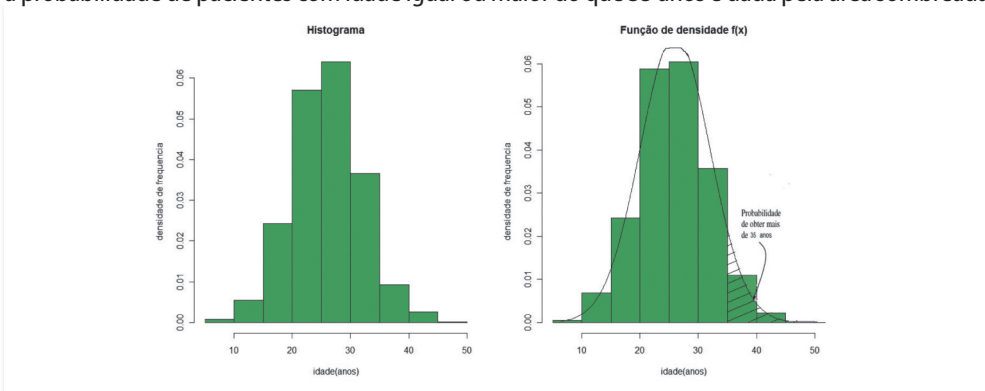

FIGURA 3.5 - Histograma de dados hipotéticos da idade de pacientes sendo ajustado pela curva de distribuição normal

A *distribuição de probabilidade normal* desempenha papel preponderante em inferência estatística. Nesta área da estatística, a média amostral é a variável de maior interesse e conhecer a sua distribuição de probabilidade é de grande relevância. Supondo uma coleta de amostra superior a 30 pacientes, podemos usar a *distribuição normal* como modelo adequado para descrever os resultados da média amostral, mesmo se a população de onde a amostra foi retirada não seguir a distribuição normal. Esse é o resultado do *Teorema Central do Limite* (principal teorema na Estatística) e que mostra a grande importância da distribuição normal.

Em se tratando da curva de distribuição normal (figura 3.6), entende-se que dois parâmetros devem ser pré-especificados para que possa calcular as probabilidades de interesse. O primeiro parâmetro é a *média* (μ), que determina o valor do centro da curva, enquanto que o *desvio-padrão* (σ) é o segundo e este determina a largura da curva normal. Assim, quanto menor o valor do *desvio-padrão,* menor variabilidade dos dados e, portanto, menor a largura da curva.

Com relação às características da distribuição normal, pode-se dizer que:

- A *média* (μ) da distribuição corresponde ao valor da mediana e moda;
- A curva normal é assintótica ao eixo x em ambas as direções, ou seja, suas extremidades prolongam para o infinito;
- A curva normal, além de ter uma área total igual a 1, é simétrica em torno da média.

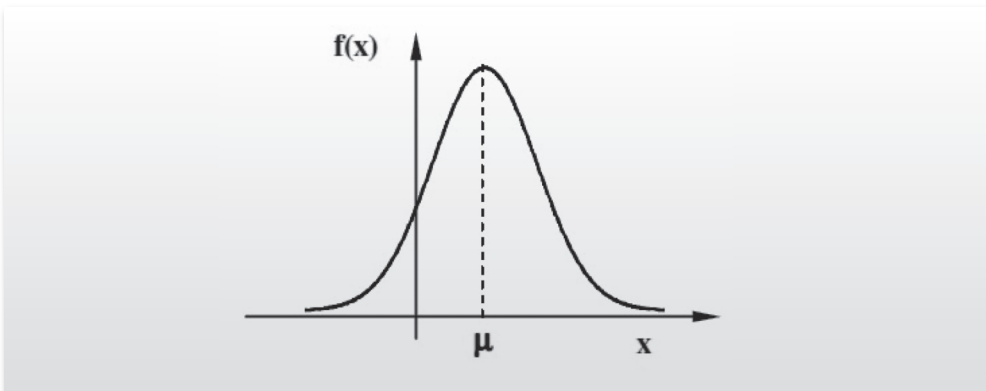

FIGURA 3.6 - Curva de distribuição normal

Muitos métodos estatísticos baseiam-se na suposição de normalidade dos dados, tais como *teste t*, ANOVA (*análise de variância*), *coeficiente de correlação de Pearson*, *análise de regressão*, etc. Caso a suposição de normalidade da variável de estudo seja violada, classificamos a variável como assimétrica, ou seja, a variável não apresenta distribuição normal, e, sendo assim, devemos escolher testes *não-paramétricos* para a análise estatística, quando não for possível corrigir esta violação ou quando não for possível propor outra *distribuição de probabilidade*. Os testes estatísticos *não-paramétricos* exigem menos pré-requisitos, mas produzem testes de significância com menos poder de detecção, quando comparados com os testes *paramétricos*.

A suposição de normalidade dos dados é avaliada por meio de testes específicos disponíveis em programas estatísticos. Os dois mais comuns são o teste Shapiro-Wilks e o teste de Kolmogorov-Smirnov. Cada um calcula o nível de significância para as diferenças em relação a uma distribuição normal (HAIR *et al*., 2009). Se este nível de significância, calculado pelo programa estatístico, apresentar valor p>0,05, por exemplo, podem ser empregados testes paramétricos na análise dos dados.

3.4 Medidas Descritivas

A descrição dos dados coletados em uma amostragem ou obtidas de toda a população-alvo, permite uma idéia da sua distribuição, mas não fornece valores numéricos necessários aos cálculos estatísticos. Isto é feito pelas medidas descritivas.

3.4.1 Medidas de Tendência Central

Uma maneira de descrever os dados de uma forma mais condensada do que usando as tabelas de freqüência para variáveis quantitativas é representar por um valor único. Este valor único é um número que seja o mais semelhante possível aos demais números do conjunto. Assim, define-se este número como uma medida central ou que tende ao centro.

Existem três medidas de tendência central para representar as variáveis quantitativas do banco de dados: a *média*, a *mediana* e a *moda*. Definiremos cada uma dessas medidas de forma sucinta e abordaremos seus pontos positivos e negativos.

I . Média

A medida de tendência central mais usual é a média aritmética, calculada pela soma de todas as observações de um conjunto de dados dividida pelo tamanho do mesmo.

II . Mediana

A mediana é definida como sendo o valor, em um conjunto de dados ordenados de maneira crescente, que os separa em dois subgrupos de mesmo tamanho. Entende-se que é um valor tal que a metade dos valores do banco de dados são maiores ou iguais à mediana, enquanto a outra metade é menor ou igual a ela.

III . Moda

O valor mais freqüente de um conjunto de dados é denominado *Moda*. Quando dois valores aparecem com a mesma freqüência máxima, cada um deles é uma moda, e o conjunto se diz *bimodal*. Se mais de dois valores ocorrem com a mesma freqüência máxima, cada um deles é uma moda, e o conjunto é *multimodal*. Quando não existe um valor mais freqüente que os demais, o conjunto não tem moda (*amodal*).

Nos recursos visuais, no caso específico, o histograma, a moda ocorre representada por um pico de freqüência. Em algumas situações, observam-se histogramas com dois picos, sendo classificada como distribuição bimodal. Neste caso, há indícios de que a população estudada é, de fato, um cruzamento de duas populações estatísticas. Exemplificando, suponha que a variável altura dos pacientes de uma clínica seja coletada, considerando conjuntamente os homens e mulheres, e, em seguida, representada visualmente por um histograma. Pela figura 3.7, o histograma apresentou dois picos de freqüência nas classes, demonstrando a existência de duas populações, uma vez que, em âmbito geral, os homens são mais altos do que as mulheres.

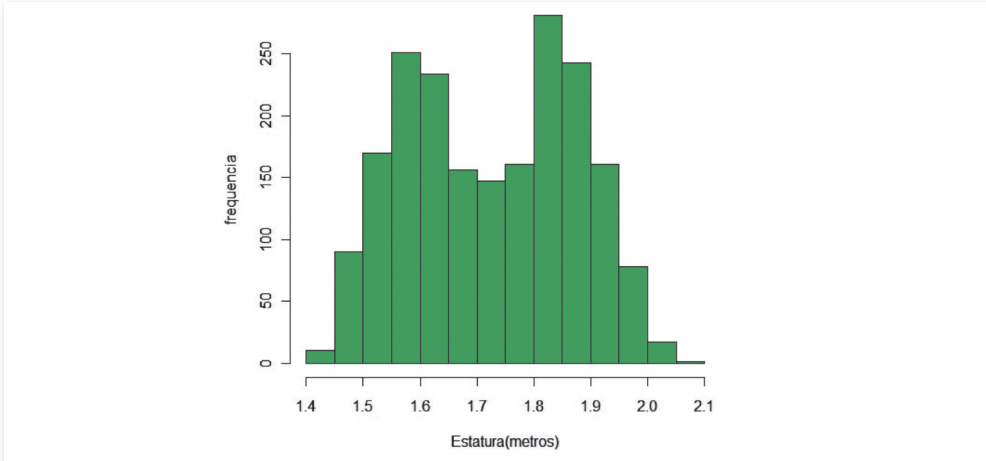

FIGURA 3.7 - Histograma das estaturas (metros)

IV . Exemplo de medidas de tendência central

É imprescindível apresentar os valores de todas as medidas de tendência central, simultaneamente, em uma tabela. A título de ilustração apresentamos, na tabela 3.7, as medidas de resumo para a variável idade das pacientes com câncer de mama do banco de dados da tabela 3.2.

Tabela 3.7 - Medidas de tendência central para a variável idade das pacientes com câncer de mama

Variável	n	Média	Mediana	Moda
Idade (anos)	39	35,58	35,2	35

Fonte: Dados da pesquisa

Para representar a idade das pacientes com câncer de mama do banco de dados, usando a média, pode-se dizer que a idade média das pacientes é de 35,58 anos. Quanto à mediana, interpreta-se que a metade das pacientes tem idade menor ou igual a 35,2 anos e a outra metade tem idade maior ou igual a 35,2 anos. No conjunto de dados existe uma moda, apenas um valor que se repete com maior freqüência, a idade de 35 anos. Assim, conforme ficou evidente a partir dos resultados da tabela 3.7, as três medidas de tendência central apresentam valores semelhantes entre si. Mas isso só acontece quando a variável segue uma distribuição de freqüências específica (distribuição gaussiana, também denominada de Normal).

V . Vantagens e Desvantagens de medidas de tendência central

A média é uma das medidas mais utilizadas no quesito resumo de medidas, pois apresenta propriedades estatísticas mais interessantes, no que diz respeito ao assunto métodos de estimação. O cálculo da média leva em consideração todos os valores do banco de dados. Por este motivo a média é sensível a valores extremos (muito grande ou muito pequeno), ou seja, o valor calculado desloca a representação do centro. Em situações desse tipo é aconselhável utilizar-se da mediana, pois não é afetada pelos extremos do conjunto.

Apesar da moda não ser uma medida de tendência central muito conhecida, ela apresenta pontos positivos em relação às demais. Especificamente, em situações onde a variável de interesse possui distribuição de freqüências bimodais ou multimodais.

Observe que as medidas de tendência central podem ser usadas como uma medida-resumo, tanto para as medidas discretas como para as contínuas.

3.4.2 Medidas de Dispersão ou de Variabilidade

Nem sempre uma única medida é capaz de resumir, satisfatoriamente, um conjunto de dados. Suponha uma situação em que dois grupos de pacientes, caso e controle, estão sendo avaliados em relação à sua idade. É natural utilizarmos como medida de resumo o cálculo da média para representar cada grupo. Entretanto, percebe-se que ambos os grupos apresentaram a mesma idade média. Neste caso, torna-se necessário construir uma medida que permita efetuar uma análise do grau de dispersão dos dados.

Nesta seção, abordaremos três medidas de dispersão ou de variabilidade (amplitude total, desvio-padrão e coeficiente de variação), apresentando seus pontos positivos e negativos.

I . Amplitude Total

Amplitude total é a diferença entre o maior e o menor valor observado no conjunto numérico.

Apesar de ser uma medida fácil de calcular, a amplitude total possui limitações, pois considera apenas os extremos do conjunto de dados (máximo e mínimo), desprezando todos os outros valores.

II . Variância e Desvio-Padrão

Se por um lado há limites para o uso da amplitude total para a obtenção do grau de dispersão é, então, razoável propor uma medida que leve em consideração todas as diferenças do conjunto de dados.

Por convenção, adota-se a média como valor referencial para calcular as diferenças dos valores do conjunto em relação a ela. Note que teremos um desvio (diferença) para cada elemento do banco de dados. Se, por ventura, arriscássemos calcular o desvio médio, o resultado daria sempre zero. A explicação a este fato é que a soma de desvios negativos com positivos se anulam. Por este motivo, se fez necessário, como sugestão, elevar ao quadrado cada desvio.

Para sintetizar, a *Variância* é definida como a média aritmética de todos os desvios ao quadrado.

A *Variância* representa uma medida de variabilidade, porém esta medida é expressa em unidade diferente da unidade dos dados originais. Por esta razão utilizaremos o Desvio-Padrão (D.P) que soluciona tal problema.

O *Desvio-Padrão* (D.P) exige o calculo prévio da Variância para que seja extraída desta a raiz quadrada. Um ponto importante a se dizer sobre o *Desvio-Padrão* é que o valor calculado é sempre positivo.

Pode-se dizer que a interpretação do desvio-padrão representa a distância típica (padrão) dos dados em relação à média. Isto significa que quanto maior o desvio-padrão, maior heterogeneidade existe entre os dados.

III . Coeficiente de Variação

Ao realizar o cálculo do desvio-padrão, ocasionalmente, nos deparamos com a dificuldade de classificá-lo como uma medida de baixa variação ou de alta variação. Por exemplo, um desvio-padrão de 10 unidades pode ser classificado como baixa variação se a média é de 1000 unidades; entretanto, se a média é igual 100 unidades, um desvio-padrão de 10 unidades significa uma alta variação.

Uma medida de variabilidade que condensa as duas informações (média e desvio padrão) é o coeficiente de variação, que consiste na divisão entre o desvio-padrão (D.P) e a média aritmética ($\bar{\chi}$) multiplicado por 100.

Assim, entende-se que quanto menor o valor do coeficiente de variação, menor é a sua dispersão, ou seja, os dados são mais homogêneos.

Como o *Coeficiente de Variação* não possui unidade de medida, ou seja, é adimensional, permite a comparação das variabilidades de diferentes conjuntos de dados.

IV . Intervalo de Confiança de 95%

Além dessas medidas de dispersão, em estatística, existe outra medida muito usada em oncologia que é o *Intervalo de Confiança* de 95%. O fato das estimativas pontuais serem pouco confiáveis impõe ao pesquisador o uso de estimativas intervalares. Restringir-nos-emos em definir, apenas, seu conceito, uma vez que em cada tipo de situação existe uma fórmula específica para o cálculo do *Intervalo de Confiança* de 95%. Denomina-se *Intervalo de Confiança* de 95% ao intervalo de valores entre um parâmetro amostral (tipos de parâmetros amostrais existentes: média, mediana proporção, desvio-padrão, coeficiente de correlação, risco relativo, *odds ratio*, *hazard ratio*, etc) nos quais, com uma probabilidade (ou nível de confiança) de 95%, se situará o parâmetro populacional. Para compreender melhor como é realizado o cálculo, é necessário que o leitor examine os conceitos de *distribuição normal, erro-padrão do parâmetro, nível de confiança, valor crítico e nível de significância* (α) em livros estatísticos.

V . Exemplo de medidas de variabilidade

Vamos supor que estejamos interessados em saber qual grupo, entre casos ou controles, é mais semelhante entre si com relação à idade das pacientes. Essa informação é obtida por meio de medidas de dispersão ou variabilidade. O grupo controle é, em média, 2 anos mais velho do que o grupo dos casos. Ao avaliarmos a medida de variabilidade dos dois grupos utilizando o desvio-padrão, arriscaríamos a dizer que o grupo de casos é menos homogêneo quanto à idade do que o grupo controle. Ao realizarmos essa suposição, estamos esquecendo que, mesmo que comparando unidades iguais, as medidas de idade dos dois grupos variam em escalas distintas. Para suprir esta questão, utilizaríamos a medida de coeficiente de variação. Nesta, percebe-se que o grupo dos casos é um pouco mais heterogêneo (disperso) quanto à idade do que o grupo controle (tabela 3.8).

Em âmbito geral, podemos considerar como um parâmetro de homogeneidade dos dados um coeficiente de variação menor do que 25%. Em casos onde se espera uma dispersão maior entre os pacientes, essa faixa de homogeneidade dos dados deve ser redefinida.

Tabela 3.8 - Estatística Descritiva para idade por grupo de caso-controle

Grupo Caso-Controle	Casos	Média	Variância	D.P	Coef. Variação	I.C 95% Média
Caso	22	34,80	27,28	5,22	15%	[32,62- 36,98]
Controle	17	36,60	13,25	3,64	9,95%	[34,87- 38,33]

Fonte: Dados da pesquisa

No grupo caso a idade está situada, em 95% das pacientes entre 32,6 e 37,0 anos e no grupo controle entre 34,8 e 38,3 anos. Como as médias estão contidas em ambos os intervalos de confiança, há grande probabilidade (95%) de que não exista diferença significativa entre os grupos, no que diz respeito à idade.

3.4.3 Medidas de Posição

Verificamos que a mediana separa o conjunto de dados em duas partes de mesmo tamanho, em que cada parte contém o mesmo número de elementos. Contudo, um mesmo conjunto de

dados pode ser dividido em mais partes que contenham a mesma quantidade de elementos. Exemplos de medidas de posição:

- mediana: divide o conjunto de dados em duas partes iguais (M_d).
- quartis: divide o conjunto de dados em quatro partes iguais ($Q\downarrow 1$, $Q\downarrow 2$, $Q\downarrow 3$) .
- decis: divide o conjunto de dados em dez partes iguais($D\downarrow 1$, $D\downarrow 2$, $D\downarrow 3$, $D\downarrow 4$, $D\downarrow 5$, $D\downarrow 6$, $D\downarrow 7$, $D\downarrow 8$, $D\downarrow 9$).
- percentis: divide o conjunto de dados em 100 partes iguais ($P\downarrow 1$, $P\downarrow 2$, $P\downarrow 3$, $P\downarrow 4$, $P\downarrow 5$, $P\downarrow 6$, $P\downarrow 7$, $P\downarrow 8$... $P\downarrow 99$).

Entende-se que os percentis estabelecem uma relação de equivalência com os decis e quartis, veja na tabela 3.9.

Tabela 3.9 - Relação de equivalência entre percentis, decis e quartis

Quartis	Decis
$Q_1 = P_{25}$	$D_1 = P_{10}$
$Q_2 = P_{50}$	$D_2 = P_{20}$
$Q_3 = P_{75}$	$D_3 = P_{30}$
	$D_4 = P_{40}$
	$D_5 = P_{50}$
	$D_6 = P_{60}$
	$D_7 = P_{70}$
	$D_8 = P_{80}$
	$D_9 = P_{90}$

A utilidade principal das *medidas de posição* é ajudar a estabelecer pontos de corte com uma determinada freqüência nos valores da variável. Vejamos, na tabela 3.10, as interpretações do primeiro quartil ($Q\downarrow 1$) e do percentil noventa e cinco (P_{95}) quanto à variável idade das pacientes de câncer de mama do banco de dados. Observa-se que 25% das pacientes apresentam idades menores ou iguais a 33,1 anos, enquanto que 75% das pacientes apresentam idades maiores ou iguais a 33,1 anos, no que se refere ao primeiro quartil ($Q\downarrow 1$). Já para o percentil noventa e cinco (P_{95}), 95% das pacientes apresentam idades menores ou iguais a 43,4 anos, enquanto que 5% das pacientes apresentam idades maiores ou iguais a 43,4 anos.

Tabela 3.10 - Medidas de posição dos percentis, decis e quartis quanto à idade das pacientes com câncer de mama

Variável	P5	D1	Q1	D3	Mediana	Q3	D9	P95
Idade	27,4	29,6	33,1	33,3	35,2	38,7	43,1	43,4

Fonte: Dados da pesquisa

3.4.4 Medidas de Risco

Entendemos como risco, a relação proporcional entre as grandezas que correspondem à medida de ocorrência de um evento em relação a outro.

Trata-se de medidas que permitem a comparação entre diferentes populações e, eventualmente, a combinação de resultados de diferentes estudos.

Apresentaremos nessa seção as duas principais medidas de risco (*risco relativo* e *razão das chances*) para análise de *Tabelas de Contingência* do tipo 2x2.

Tabelas de Contingência do tipo 2x2 são tabelas em que as contagens correspondem a duas variáveis qualitativas, e cada uma delas possui duas categorias. As categorias de uma variável estão presentes nas linhas da tabela enquanto as categorias da outra estão presentes nas colunas, como pode ser visto na tabela 3.11.

Tabela 3.11 - Contingência 2x2 Genérica

Grupo	Presença da doença		Total
	Sim	Não	
Caso	a	b	a+b
Controle	c	d	c+d
Total	a+c	b+d	n

I . Risco Relativo

Imaginem que os pacientes de uma determinada população sejam classificados segundo o *Grupo*, Casos e Controle, e a presença ou ausência de uma determinada doença, denotados por Sim e Não, respectivamente, conforme a tabela 3.11.

Logo, para se obter o Risco Relativo, devemos calcular primeiramente:

Estimativa do risco da Presença da doença no grupo Caso: $\frac{a}{a+b}$

Estimativa do risco da Presença da doença no grupo Controle : $\frac{c}{c+d}$

A divisão entre o risco da presença da doença no grupo Caso e o risco da presença da doença no grupo Controle é denominada *Risco Relativo de doença (RR)*, matematicamente definido por:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Note que a estimativa do *Risco Relativo* só pode ser feita para estudos prospectivos, estudos de coorte e experimentos clínicos aleatorizados, pois os grupos formados são previamente definidos pelo pesquisador.

Tomemos como exemplo um estudo coorte que examina os fatores de risco para o câncer de mama entre as mulheres que participaram do 1º Levantamento Nacional de Exame de Nutrição e de Saúde. Nesse estudo há dois grupos: mulheres que deram à luz pela primeira vez com 25 anos ou mais e mulheres que deram à luz pela primeira vez com menos de 25 anos. Em uma amostra de 4.540 mulheres que deram à luz seus primeiros filhos antes de 25 anos, 65 desenvolveram o câncer de mama. Das 1.628 mulheres que deram à luz seus primeiros filhos com 25 anos ou mais, 31 desenvolveram o câncer de mama, tais informações estão sintetizadas na tabela 3.12.

Tabela 3.12 - Exemplo de Tabela de Contingência 2x2

Faixa Etária para primeira gestação a termo	Diagnóstico de câncer de Mama		Total
	Sim	Não	
Menos de 25 anos	65	4475	4.540
25 ou mais anos	31	1597	1628
Total	96	6072	6168

Fonte: Pagano e Gauvreau, 2004

Empregando a notação sugerida, o risco do grupo de mulheres com mais de 25 anos apresentar câncer de mama é de 1.90%, enquanto o risco de câncer de mama no grupo de mulheres com idade menor que 25 anos resulta 1.43%. Portanto, o *risco relativo* é de 1,33. Este valor indica que as mulheres que deram à luz pela primeira vez com 25 anos ou mais têm uma probabilidade de desenvolver câncer de mama 33% maior do que aquelas que deram à luz com menos de 25 anos.

Vale ressaltar que, normalmente, a medida de *risco relativo* é maior que 1,0, pois, hipoteticamente, a exposição ao *fator de risco* deve aumentar a prevalência da condição. No entanto, quando o risco relativo é inferior a 1,0, o fator passa a ser denominado *fator de prevenção*. Esse mesmo argumento é válido para a medida razão das chances, que será definida na próxima seção.

Finalizando, se o *risco relativo* (assim como a *razão das chances*) é próximo de 1,0, a pesquisa apresentará indícios que o fator não se relaciona com a condição estudada.

II . Razão das Chances (*odds ratio*)

Em estudos retrospectivos, do qual faz parte o estudo de caso e controle, o tamanho dos grupos não é conseqüência de sua incidência real na população, mas uma decisão do pesquisador baseado na questão científica proposta. Sendo assim, não se aplica o cálculo do *risco relativo* e, por isso, utilizaremos a medida *razão das chances*.

Chance pode ser definida como o número de vezes que um evento ocorreu dividido pelo número de vezes em que ele não ocorreu. Na tabela 3.11 a chance de doença no grupo caso é dada por a/b e no grupo controle por c/d.

Razão das chances expressa a relação de ocorrência da doença nos grupos caso e controle e é dada por a/b÷c/d, ou de forma simplificada:

$$RC = \frac{axd}{bxc}$$

Vejamos um exemplo de aplicação da *razão das chances* para o banco de dados de mulheres grávidas com câncer de mama. Nesse estudo, as pacientes apresentavam ausência e presença de gravidez, tinha como finalidade observar o estado atual (vivo ou óbito) nestes dois grupos. As informações desse estudo estão resumidas a seguir:

Tabela 3.13 - Exemplo de Tabela de Contingência 2x2 para pacientes com câncer de mama

Grupo	Estado Atual		Total
	Óbito	Vivo	
Caso	16	6	22
Controle	11	6	17
Total	27	12	39

Fonte: dados da pesquisa

Empregando a notação sugerida de RC, pode-se dizer que a *razão das chances* do estado atual da tabela 3.13 é de 1.45. Este valor indica que a chance de ocorrência de óbito no grupo de mulheres grávidas (caso) é 1.45 vezes a chance no grupo de mulheres não grávidas (controle). Vale ressaltar que este valor bruto, sem nenhuma avaliação da sua variabilidade (como, por exemplo, seu intervalo de confiança de 95%), não nos permite tirar conclusões.

3.4.5 Medidas de Sobrevida

Para apurar a medida de sobrevida em um banco de dados, são necessários dois componentes: o *tempo até a ocorrência de um evento determinado* e o *tipo de evento final*.

Em relação ao *tempo até o evento*, os três elementos básicos para o seu cálculo são o tempo inicial, a escala da medida e o tempo em que o evento final ocorreu. Para o primeiro elemento, tempo inicial, é comum utilizarmos a data do início do tratamento de doenças ou do diagnóstico. Quanto ao segundo elemento, normalmente, é utilizado o mês como escala de medida. Contudo, em algumas situações clínicas, é usual utilizarmos a escala dias ou anos. Por último, o tempo em que o evento final ocorreu pode ser a data do óbito (curva de sobrevida global), a data de recidiva de uma neoplasia (curva de sobrevida livre da doença ou de recidiva) ou a data em que a progressão de uma doença foi documentada (sobrevida livre de progressão). Assim, a partir da diferença entre as datas do terceiro e primeiro componente, com base na medida de escala definida pelo pesquisador, obtém-se a variável *tempo até evento*.

Em relação ao evento final, pode tratar-se do óbito do paciente, da recidiva ou progressão de uma doença ou do que é denominado de *censura*.

É comum que os resultados dos estudos clínicos sejam relatados antes que todos os pacientes incluídos apresentem qualquer tipo de evento considerado falha. Isto pode ocorrer por perda de acompanhamento do paciente no decorrer do estudo ou por ausência de falha até o término da pesquisa. Estes pacientes são chamados censurados, porque entende-se que o tempo de falha desses pacientes é superior ao tempo registrado até o último acompanhamento. Note que, mesmo que alguns pacientes sejam censurados, todas as informações provenientes de um estudo de sobrevida devem ser apuradas na análise estatística. Portanto, para se obter a variável *evento final*, de natureza dicotômica, cada paciente do banco de dados deverá ser classificado pela presença da censura, codificada por 0, ou ocorrência de falha, codificada por 1.

Desta forma, a variável de interesse em análise de sobrevivência é representada por duas colunas (tempo até evento e tipo de evento final) na planilha eletrônica que constitui o banco de dados.

I . Função Sobrevivência

A importância de métodos de análise de sobrevida está em saber a chance de sofrer o desfecho em cada ponto no tempo, já que o prognóstico expresso por uma taxa sumária, como por exemplo, sobrevida em 5 anos, não contém essa informação.

Um grande problema quando se usa variável função de sobrevivência é que os pacientes entram em momentos diferentes no estudo, frequentemente ao longo de anos. Mas os resultados são analisados em um só tempo, e neste momento, os pacientes têm diferentes períodos de seguimento.

O que se deseja é achar uma forma do paciente contribuir para a curva de sobrevida por todo o tempo em que estiver sendo seguido.

O modelo mais utilizado, em oncologia, é o Estimador de Kaplan-Meier para a *função de sobrevivência*. Entende-se que a *função de sobrevivência* é a probabilidade de um paciente sobreviver a um tempo especificado. Em oncologia, a função de sobrevivência pode ser denominada de *sobrevida global, sobrevida livre de recidiva, sobrevida livre de progressão, etc*.

A título de ilustração, a tabela 3.14, exibe, desde o primeiro até o vigésimo oitavo intervalo de tempo de falha, os cálculos da estimativa de Kaplan-Meier para a sobrevida global das mulheres com câncer de mama. Repare que a última coluna dessa tabela apresenta a sobrevida global das pacientes para variados intervalos.

Todas as pacientes estavam vivas no período inicial (t = 0) e se mantêm até a primeira morte que ocorre em 1,38 meses. Logo, a estimativa da sobrevida global é 1,00 no intervalo entre 0 a 1,38

meses exclusive. No segundo intervalo, (1,38 - 2,3), existem 39 pacientes que estavam vivas (sob risco) antes de 1,38 meses e 1 paciente morreu. Dessa forma, a probabilidade de uma paciente sobreviver no segundo intervalo é de 97,4%. Assim, analogamente, para qualquer intervalo especificado, a sobrevida global foi calculada em termos de probabilidade.

Observe que a *sobrevida global* tanto no 26º mês quanto no 36º mês são iguais (0,486), pois a *sobrevida global* é uma função escada com saltos somente nos tempos de falha.

Tabela 3.14 - Sobrevida global das pacientes com câncer de mama

Intervalo (meses)	Número de pacientes sob risco	Número de Falhas	Número de Censuras	Sobrevida Global
[0 - 1,38)	39	0	0	1,00
[1,38 - 2,3)	39	1	0	0,974
[2,3 - 4,27)	38	1	1	0,949
[4,27 - 10,84)	36	1	0	0,922
[10,84 - 11,04)	35	1	0	0,896
[11,04 - 12,35)	34	1	1	0,870
[12,35 - 12,65)	32	1	0	0,842
[12,65 - 12,94)	31	1	0	0,815
[12,94 - 13,27)	30	1	0	0,788
[13,27 - 15,05)	29	1	0	0,761
[15,05 - 15,28)	28	1	0	0,734
[15,28 - 15,93)	27	1	0	0,707
[15,93 - 16,3)	26	1	0	0,679
[16,3 - 17,84)	25	1	0	0,652
[17,84 - 18,5)	24	1	0	0,625
[18,5 - 19,81)	23	1	0	0,598
[19,81 - 19,88)	22	1	0	0,571
[19,88 - 21,22)	21	1	1	0,544
[21,22 - 25,49)	19	1	0	0,515
[25,49 - 36,14)	18	1	0	0,486
[36,14 - 36,4)	17	1	0	0,458
[36,4 - 40,28)	16	1	0	0,429
[40,28 - 44,35)	15	1	0	0,401
[44,35 - 83,48)	14	1	1	0,372
[83,48 - 103,7)	12	1	0	0,341
[103,7 - 144,7)	11	1	5	0,310
[144,8 - 152,7)	5	1	1	0,248
[152,7 - 300)	3	1	2	0,165

Fonte: Dados da pesquisa

Conforme a tabela 3.14, a probabilidade de uma paciente jovem com diagnóstico de câncer de mama estar viva aos 20 meses é de 0,544 (ou seja, 54,4%).

Diante dos dados obtidos na tabela 3.14, a construção de um gráfico pode ser mais facilmente compreendida. Este gráfico é elaborado mantendo o valor da sobrevida constante entre os intervalos. A figura 3.8A apresenta o gráfico da sobrevida global das pacientes com câncer de mama. Note que a sobrevida global não atinge o valor zero; isto ocorre em situações nas quais o maior tempo observado na amostra for uma censura. As censuras são representadas, na figura 3.8A, por

pequenos segmentos verticais ao longo do período analisado, [0-300]. Por exemplo, entre o período [150-300], encontramos dois pequenos segmentos verticais, ou seja, existem 2 censuras.

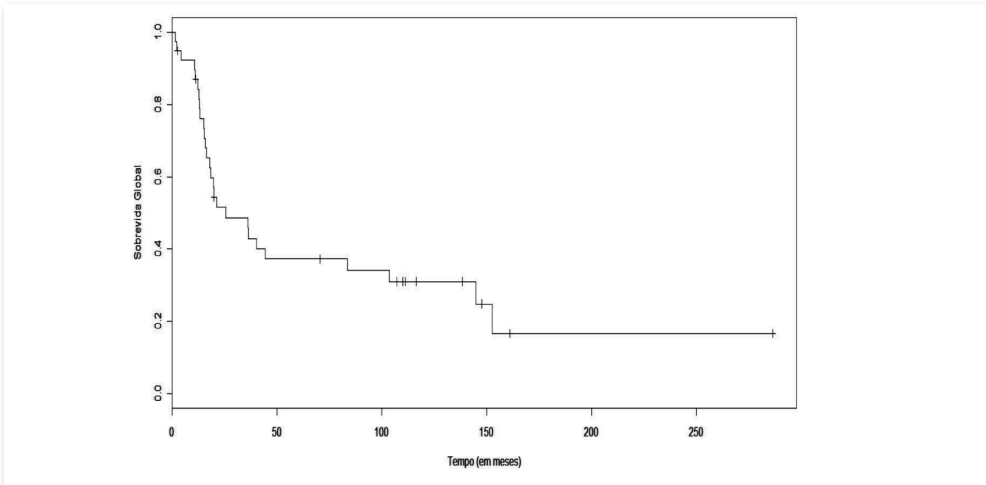

FIGURA 3.8 A - Sobrevida global das pacientes com câncer de mama (Gráfico de Kaplan-Meier).

A partir dos resultados obtidos pelo método de Kaplan-Meier é interessante obter estimativas dos percentis. Um exemplo de percentil é o *tempo mediano de vida* que é bastante usado na prática. O cálculo da mediana é realizado por meio de uma *interpolação linear*. INTERPOLAÇÃO LINEAR é uma técnica de cálculo que permite apurar, por aproximação, um valor desconhecido que se encontra entre dois valores fornecidos. Freqüentemente, as tabelas de sobrevivência não fornecem o valor exato necessário para efetuar os cálculos solicitados pelo pesquisador – daí a importância do método de interpolação linear: através deste, contornamos essa dificuldade, obtendo, mediante uma proporção simples, o valor desconhecido por meio de outros valores próximos, presentes na tabela.

Fórmula da interpolação linear:

$$\frac{b - a}{S(b) - S(a)} = \frac{x - a}{S(x) - S(a)}$$

Onde: **a** e **b** são pontos conhecidos da tabela, menor valor e maior valor, respectivamente.
S(a) e **S(b)** são as curvas de sobrevivências nos pontos **a** e **b**, respectivamente.
x é o ponto desconhecido entre **a** e **b** e **S(x)** é a curva de sobrevivência no ponto **x**.

Vejamos como se calcula o tempo mediano de vida para a Tabela 3.14. Entende-se que o *tempo mediano de vida* (**x**, ponto desconhecido) representa o tempo em que 50% dos pacientes sobrevivem, logo **S(x)** = 0,50. Os valores de sobrevida, da tabela 3.14, próximos de 0,50 são: 0,486 e 0,515 que correspondem **S(b)** e **S(a)**, respectivamente. Os pontos **a** e **b** associados as suas respectivas sobrevidas são: 21,22 meses e 36,14 meses. Assim, uma vez definido todos os parâmetros, substituímo-nos na fórmula da interpolação linear:

$$\frac{36{,}14\ meses - 21{,}22\ meses}{0{,}486 - 0{,}515} = \frac{X\ meses - 21{,}22\ meses}{0{,}50 - 0{,}515} = 28{,}94\ meses$$

Portanto, 28,94 meses é uma estimativa do tempo em que 50% das pacientes sobrevivem. Esta abordagem de estimar o tempo mediano é semelhante a conectar por retas as estimativas de Kaplan-Meier, em vez de se utilizar a sobrevida na forma de escada. Esta abordagem, geralmente, produz uma melhor representação da distribuição contínua dos tempos até ocorrência de um evento, razão pela qual deve ser preferida (COLOSIMO *et al*., 2002). Note que os programas estatísticos não baseiam o cálculo do tempo mediano ou outro tempo neste critério descrito.

Repare que a fórmula da interpolação aplicada para o *tempo mediano de vida* também pode ser apurada para outros percentis. Exemplificando, suponha que desejamos encontrar o tempo de vida que 25% dos pacientes permanecem vivos. Assim, substituímos a probabilidade de 50% para 25% na fórmula de interpolação linear, temos:

$$\frac{152{,}67\ meses - 103{,}72\ meses}{0{,}248\ -\ 0{,}310} = \frac{X\ meses - 103{,}72\ meses}{0{,}250\ -\ 0{,}310} = 151{,}1\ meses$$

Portanto, 151,1 meses é uma estimativa do tempo em que 25% dos pacientes sobrevivem.

II . Função taxa de falha

Além da função de sobrevivência, existe a função taxa de falha, também denominada de função de risco, e utilizada, geralmente, como uma medida de síntese para a sobrevida.

Podemos definir como taxa da ocorrência de falha em um determinado intervalo de tempo $[t]_1, t_2)$ à probabilidade de que a falha ocorra no intervalo especificado, considerando que esta ainda não ocorreu antes do tempo $[t]_1, t_2)$. Logo, a taxa de falha no intervalo t_1 é calculada em termos da função de sobrevivência e expressa por:

$$\lambda[t]_1, t_2) = \frac{[S(t]_1) - S(t_2)}{[(t]_2 - t_1)S(t_1)}$$

Onde: t_1 e t_2 são tempos especificados, menor valor e maior valor, respectivamente.

S(t_1) e S(t_2) são as curvas de sobrevivências nos tempos t_1 e t_2, respectivamente.

$\lambda[t]_1, t_2)$ é a taxa de falha no intervalo $[t]_1, t_2)$.

Note que se considerarmos um intervalo de tempo muito pequeno para $[t]_1, t_2)$, a taxa passa a ser denominada *taxa de falha instantânea* no tempo *t* condicional à sobrevivência até o tempo t. A função *taxa de falha instantânea* é muito utilizada na prática para descrever o comportamento do tempo de vida dos pacientes. A figura 3.8B mostra a comparação entre curvas de função de risco de dois grupos de pacientes (mulheres grávidas e não grávidas) com câncer de mama. O comportamento crescente das curvas indica que a taxa de falha dos dois grupos de pacientes aumenta com o decorrer do tempo.

A partir da razão da função de risco entre dois grupos, mulheres grávidas e não grávidas (Figura 3.8B), calcula-se a razão de risco instantânea no tempo *t* (*hazard ratio*). Ela equivale ao risco relativo aplicado à variável data e é muito útil em cálculos estatísticos.

Para o exemplo do banco de dados de pacientes com câncer de mama, o valor encontrado da medida de riscos proporcional (*hazard ratio*) foi de 1,22. Este valor indica que o risco de uma mulher grávida com câncer de mama falecer é 1,22 vezes maior, em comparação com uma mulher não grávida com câncer de mama. Nesse caso, como o risco proporcional é próximo de 1, há indícios que o fator gravidez não se relaciona com risco de óbito, mas para se fazer uma afirmação com grau conhecido de certeza, seria necessário o cálculo do intervalo de confiança de 95% e o valor p.

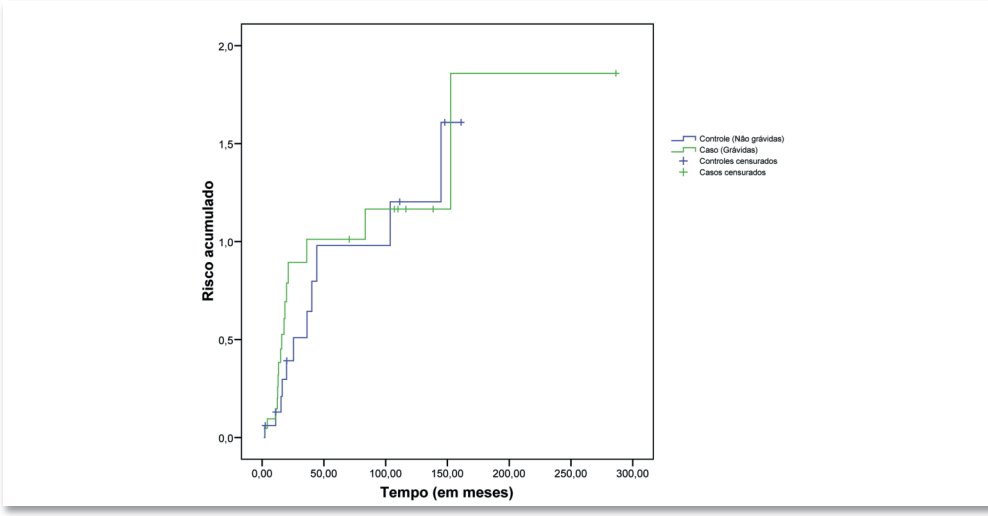

FIGURA 3.8 B - Comparação da função taxa de falha das pacientes com câncer de mama em dois grupos (grávidas e não grávidas).

Vale ressaltar que a função taxa de falha é mais informativa do que a função sobrevivência. Suponhamos que determinado paciente com câncer de pulmão tenha sobrevivido por 2 anos após o diagnóstico. Qual o prognóstico deste paciente? A curva de sobrevida não nos dará esta resposta à primeira vista, mas ela é facilmente visualizada em uma curva de função de risco. Por outro lado, a diferença entre curvas de sobrevida agrega informação de grande importância clínica, que é a magnitude da diferença.

Matematicamente, a função de risco é a negativa da inclinação da curva de sobrevida quando esta é construída em escala logarítmica, e fornece a variação do risco ao longo do tempo.

O uso da função de risco é fundamental para o modelo de riscos proporcionais de Cox (modelo de Cox), como veremos em capítulo posterior.

Alguns exemplos da função da taxa de falha são descritos na figura 3.9, onde a curva A representa risco constante ao longo do tempo; na curva B o risco é crescente e na C é decrescente. A curva D representa o risco da população geral. [32]

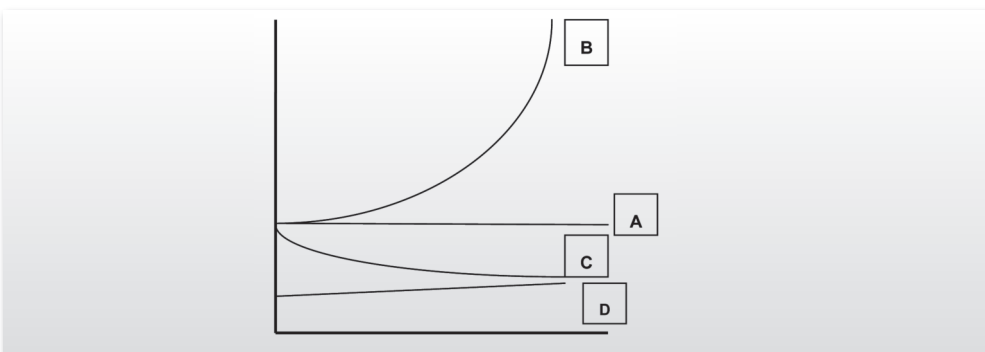

FIGURA 3.9 - Curvas da função da taxa de falha

3.5 Gráficos para o Cruzamento de Variáveis

Não raro em uma pesquisa clínica desejamos estabelecer relações ou associações entre duas ou mais variáveis. Para compreender melhor o tipo de relação entre tais variáveis, mencionaremos, nesta seção, ferramentas gráficas apropriadas em cada situação, que serão descritas a seguir.

3.5.1 Gráfico de dispersão (2 variáveis quantitativas)

O *gráfico de dispersão* é um gráfico em que são representados, em um plano cartesiano, os diversos pares de valores observados em duas variáveis quantitativas. Este gráfico permite uma avaliação, por meio das nuvens de pontos, de uma provável relação (do tipo: linear, quadrática, polinomial, exponencial, etc) entre as variáveis ou uma adequação de uma expressão matemática. Além disso, é útil para comparar o efeito de dois tratamentos no mesmo paciente, desde que as duas variáveis estudadas sejam quantitativas.

Vejamos um exemplo da utilização do *gráfico de dispersão* baseado no banco de dados das pacientes grávidas. Tendo em vista que este banco apresenta somente 2 variáveis quantitativas contínuas, sobrevida global e idade, portanto o eixo horizontal do gráfico representa a variável idade e o eixo vertical representa a variável sobrevida global. Na figura 3.10 mostramos a relação entre sobrevida e idade, de acordo com todas as pacientes grávidas.

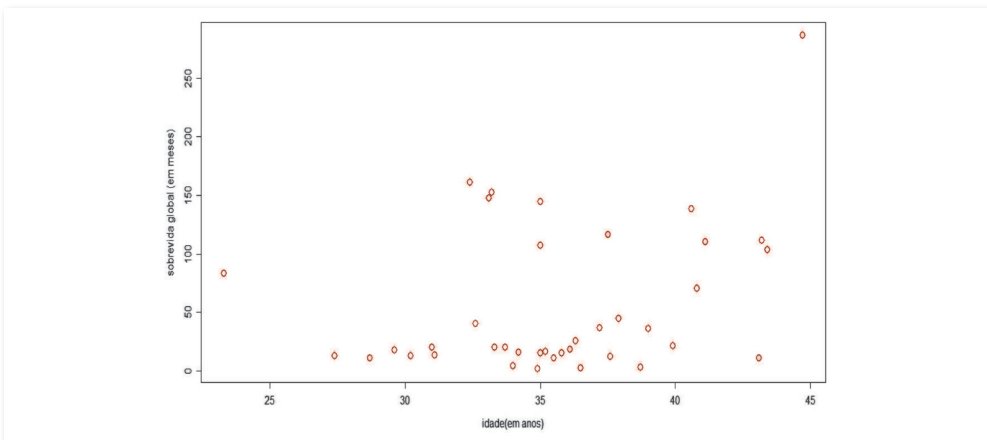

FIGURA 3.10 - Diagrama de dispersão entre idade e sobrevida

Avaliando o gráfico de dispersão, entendemos que não existe nenhum tipo de relação entre idade e sobrevida das pacientes, logo, seria inviável propor algum tipo de expressão matemática neste caso. A razão dessa conclusão é devido ao fato de que os pontos do gráfico não exibem nenhum padrão de valores crescentes, ou decrescentes, de idade que correspondem a valores crescentes da sobrevida, ou seja, o gráfico não apresenta qualquer padrão definido. Contudo, as conclusões embasadas nesse tipo de gráfico tendem a ser subjetivas, necessitando, portanto, de técnicas estatísticas (Correlação e Análise de Regressão).

Vejamos um exemplo de comparação entre dois tratamentos. Para tal, foram examinados 15 pacientes, tendo sido medidos os volumes de refluxos na veia poplítea, através de ultrassonografia, nas posições de pé e deitado (tabela 3.15). Deseja-se verificar se a posição (em pé ou deitado) influi na medição do volume de refluxo.

As informações de cada posição (em pé ou deitado) são classificadas como *dados emparelhados* (ou *pareados*), pois os mesmos pacientes foram utilizados na mesma amostra. Logo, por meio do gráfico de dispersão (figura 3.11), podemos verificar a diferença entre as duas posições.

Tabela 3.15 - Volumes de refluxos (litros por minuto medida em 15 pacientes em pé e deitado, avaliados pela ultrassonografia.

Pacientes	Volume do Refluxo em pé (litr/min)	Volume do Refluxo deitado (litr/min)
01	0,703	0,42
02	0,376	0,15
03	0,281	0,24
04	0,435	0,12
05	0,225	0,11
06	0,229	0,16
07	0,091	0,07
08	0,413	0,17
09	0,122	0,07
10	0,277	0,25
11	0,182	0,08
12	0,541	0,33
13	0,623	0,4
14	0,385	0,15
15	0,285	0,18

Fonte: Dados hipotéticos.

Conforme descrito na figura 3.11, a reta traçada no diagrama de dispersão corresponde à situação em que o volume de refluxo do paciente é o mesmo nas duas posições Como os pontos estão abaixo dessa reta, significa que, em todos os indivíduos, o volume de refluxo na posição em pé é maior do que na posição deitado.

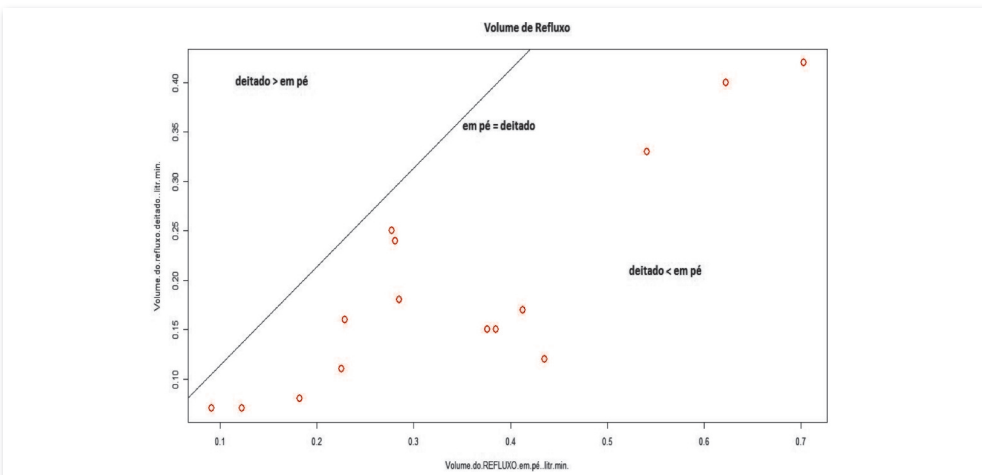

FIGURA 3.11 - Gráfico de dispersão dos volumes de refluxos (litro por minuto) medida em 15 pacientes em pé e deitado avaliado pela ultra-sonografia.

3.5.2 Box-plot (1 variável quantitativa e 1 variável qualitativa)

O gráfico de *Box plot* é um gráfico simbolizado por uma ou mais caixas. O nível superior da caixa é representado pelo *terceiro quartil* (3Q) enquanto para o nível inferior é representado pelo *primeiro quartil* (1Q). Já o traço no interior da caixa é definido pela mediana (2Q). Além disso, consta como informação o máximo e o mínimo representados por segmentos de reta. Este gráfico nos dá entendimento a respeito das medidas de tendência central, medidas de variabilidade e detecta diferenças entre os grupos do banco de dados analisado. Exemplificando, o cruzamento da variável Idade com a variável Caso-Controle é apropriado para construir tal gráfico. O resultado é apresentado na figura 3.12, onde se percebe que as mulheres do grupo controle apresentam idade mediana superior ao das mulheres do grupo caso; no entanto, as mulheres grávidas (caso) apresentam maior variabilidade de idade, pois o comprimento de sua caixa é maior.

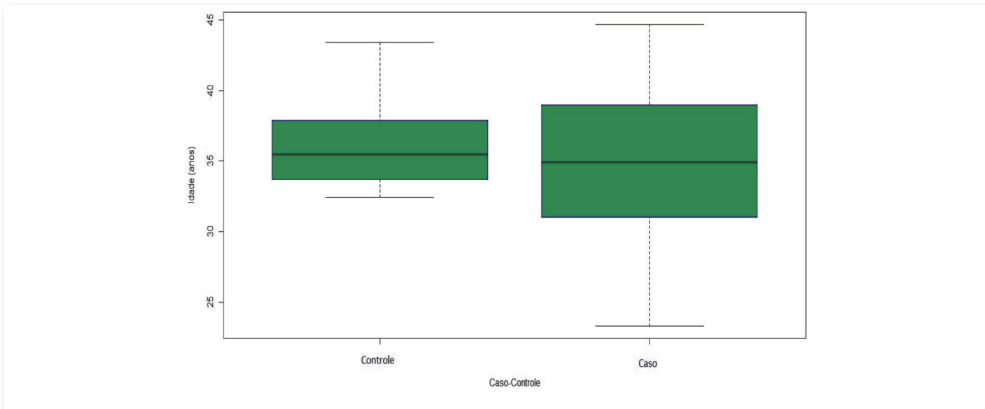

FIGURA 3.12 - *Box–plot* do cruzamento entre idade e caso-controle das pacientes com câncer de mama.

3.5.3 Gráfico de Colunas múltiplas (2 variáveis qualitativas)

É a representação simultânea de dois fenômenos de natureza qualitativa num mesmo gráfico. Essa simultaneidade tem como finalidade permitir a comparação entre os fenômenos estudados. Vejamos a construção do gráfico de colunas. A figura 3.13 descreve a situação do evento final (óbito ou vivo) nos casos e controles. Entende-se que o grupo de mulheres grávidas (caso) apresenta maior freqüência de óbitos do que o das mulheres não grávidas (grupo controle).

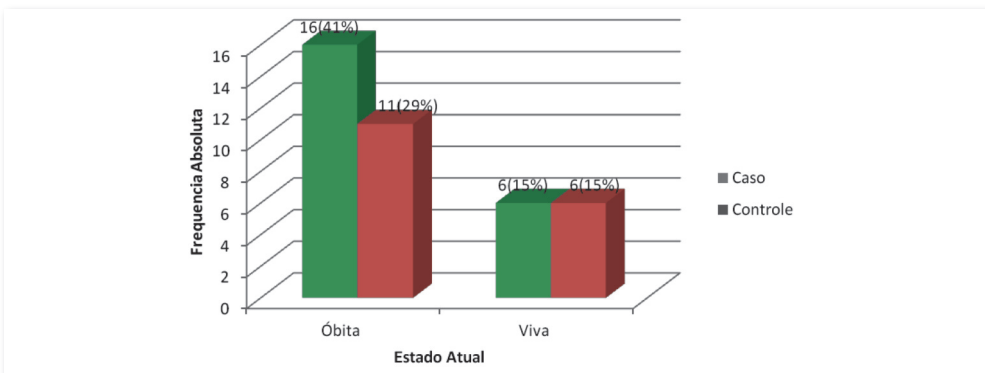

FIGURA 3.13 - *Box–plot* do cruzamento entre idade e caso-controle das pacientes com câncer de mama.

3.5.4 Gráfico de Linhas (1 variável quantitativa e 1 variável data)

O *gráfico de linha* descreve o comportamento de um conjunto de valores de uma mesma variável quantitativa (discreta ou contínua) no decorrer do tempo. O indicador de tempo é representado no eixo horizontal do gráfico de linha, enquanto a variável quantitativa é definida no eixo vertical. Este gráfico é de grande utilidade quando se deseja analisar a evolução temporal (aumento, estabilidade e declínio dos valores) da variável estudada, pois permite visualizar diferenças entre um período e os outros períodos subseqüentes. Na figura 3.14, nota-se que o número de mulheres grávidas que realizaram a última consulta ao longo dos meses da pesquisa é maior nos meses de janeiro e março.

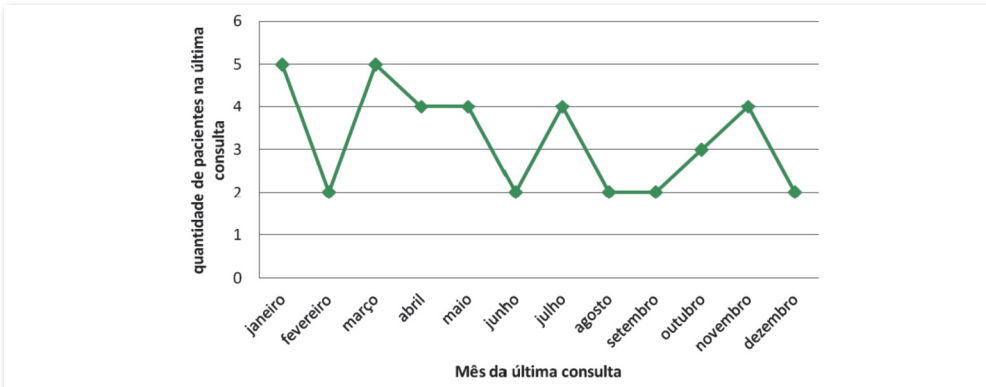

FIGURA 3.14 - Gráfico de linha entre a variável mês da última consulta e quantidade de pacientes na última consulta.

Um aspecto importante a ser ressaltado na construção deste gráfico é a definição da escala de valores do eixo vertical. Se alterarmos o final da escala de valores do eixo vertical, tanto para pequenos valores quanto para grandes valores, encontraremos comportamentos distintos na linha. Exemplificando, se definimos o eixo vertical finalizado no ponto quarenta (figura 3.15), a variação da linha ao longo do tempo poderá ser menos abrupta do que a variação da linha ao longo do tempo considerando um eixo vertical finalizado com um valor de seis (figura 3.14).

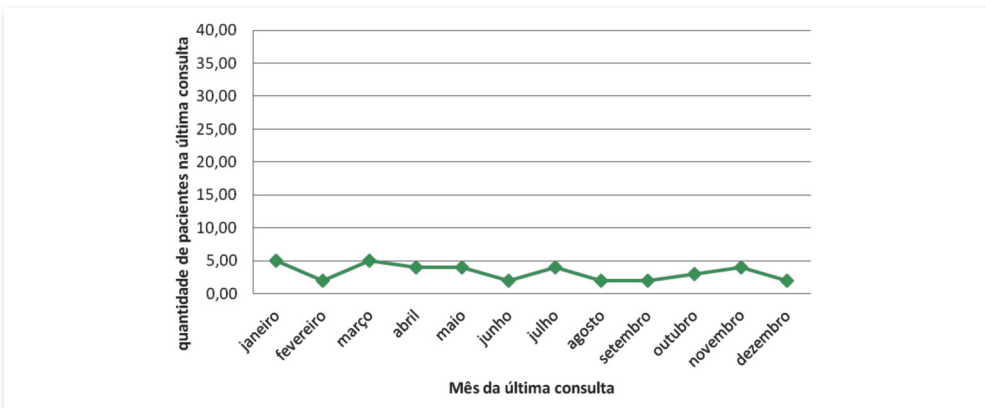

FIGURA 3.15 - Gráfico de linha entre a variável mês da ultima consulta e quantidade de pacientes na última consulta.

3. 6 Resumo

3.6.1 Classificação das Variáveis

Para cada tipo de variável existem técnicas mais apropriadas para resumir as informações, daí a importância de classificar corretamente cada variável. Uma classificação muito utilizada é:

3.6.2 Síntese dos dados

Alguns procedimentos adequados a cada tipo de variável:

Para as variáveis **qualitativas nominais**: Tabelas (distribuição de freqüência absoluta e relativa, tabela de dupla entrada), Gráficos (setores e colunas simples ou múltiplas) e Medidas (moda, risco relativo e razão das chances).

Para as variáveis **qualitativas ordinais**: Tabelas (distribuição de freqüência absoluta e relativa, freqüência absoluta acumulada, freqüência relativa acumulada, tabela de dupla entrada), Gráficos (setores e colunas simples ou múltiplas) e Medidas (mediana, moda, risco relativo e *odds ratio*).

Para as variáveis **quantitativas**: Tabelas (distribuição de freqüência absoluta e relativa, freqüência absoluta acumulada, freqüência relativa acumulada, tabela de dupla entrada), Gráficos (histograma, gráfico de dispersão, box-plot e gráfico de linhas) e Medidas (média aritmética, mediana, primeiro e terceiro quartil, percentil, variância, desvio-padrão, coeficiente de variação).

Para as **variáveis que medem o tempo até a ocorrência de um evento**: tabela (tabela de sobrevida), gráfico (gráfico de Kaplan-Meier) e medida (mediana).

Referências

1. Arango HG. Bioestatística: teórica e computacional. 2 ed. Rio de Janeiro: Guanabara Koogan, 2005.

2. Colosimo, E. Análise de Sobrevivência Aplicada. São Paulo: Blucher,2001.

3. Colosimo, EA, Ferreira, FF, Oliveira, MD, Souza, CB. Empirical Comparisons between Kaplan-Meier and Nelson-Aalen Survival Functions Estimators. J. Statist. Comput. Simul., 2002; 72(4): 299-308.

4. Crespo AA. Estatística Fácil.São Paulo: Saraiva, 2000.

5. Freund JE, Simon GA. Estatística Aplicada. 9ed. Porto Alegrel:Bookman, 2000.

6. Hair JR JF, Anderson RE, Tatham RL, Black WC. Análise Multivariada de dados. 6ed. Porto Alegre: Bookman, 2009.

7. Huff D. How To Lie With Statistics. New York:W.W. Norton & Company, 142 p.1982.

8. Lopes PA. Probabilidades e Estatística. Rio de Janeiro :Reichmann e Affonso Editores, 174 p.1999.

9. Magalhães MN, Lima ACP. Noções de Probabilidade e Estatística. 7ed. São Paulo: USP, 2010.

10. Reis EA, Reis IA . Análise Descritiva de Dados: Síntese Numérica. 2002. Relatório Técnico, Departamento de Estatística-UFMG. Disponível em:http://lattes.cnpq.br/3773191587995244.

11. Reis IA, Reis E A. Associação entre Variáveis Qualitativas: Teste Qui-quadrado, Risco Relativo e Razão de Chances. 2001. Relatório Técnico, Departamento de Estatística-UFMG. Dis´ponível em:http://lattes.cnpq.br/3773191587995244.

12. Reis EA, Reis IA. Análise Descritiva de Dados- Tabelas e Gráficos. 2001. Relatório Técnico, Departamento de Estatística-UFMG. Disponível em: http://lattes.cnpq.br/3773191587995244.

13. Simes RJ, Zelen M.Exploratory Data Analysis and the Use of Hazard Function for Interpreting Survival Data: An Investigator's Primer. J Clin Oncol, 1985; 3:1418-31.

14. Soares JF, Comini C. Introdução à Estatística. 2ed. Rio de Janeiro: LTC, 2002, 340 p.

15. Soares JF, Siqueira AL. Introdução à Estatística Médica. 2ed. Belo Horizonte: COOPMED, 2002.

16. Triola MF. Introdução à Estatísitica. 7 ed. Rio de Janeiro: LTC,2005.

17. Vieira S. Introdução à bioestatística. 3ed. rev. Ampl. Rio de Janeiro: Elsevier, 1980.