Capítulo 4
Testes de Hipóteses

Inferência estatística pode ser definida como um conjunto de procedimentos que nos permite tirar conclusões acerca de populações, a partir do estudo de amostras coletadas desta população.

No capítulo 3 aprendemos a quantificar características de uma população ou grupo de dados, fazer estimativas e, principalmente, saber a precisão destas estimativas.

Os testes de hipóteses, que fazem parte da inferência estatística, incluem um amplo conjunto de procedimentos, mas no que diz respeito à oncologia, o mais importante são os testes de significância estatística, por fornecer indicações sobre a extensão da diferença entre os valores médios de 2 ou mais agrupamentos de dados e qual a probabilidade desta diferença ser devida ao acaso.

Trata-se de ferramenta amplamente usada em várias áreas do conhecimento humano nas quais os dados envolvidos estão sujeitos à variabilidade.

No contexto deste Manual, estamos interessados na comparação entre dois, ou mais, grupos, como por exemplo, tratamentos, procedimentos diagnósticos, dietas, etc.

Este tema deverá ser subdividido nos seguintes tópicos:

4.1) Hipóteses a serem testadas
4.2) Critério de decisão
4.3) Erros tipos I e II, nível de significância e o poder do teste
4.4) Probabilidade de significância (valor p)
4.5) Hipóteses unilaterais X bilaterais (*one-sided* e *two-sided*)
4.6) Critérios de escolha e exemplos

4.1. Hipóteses a serem testadas

Uma *hipótese* é uma alegação a respeito de um determinado problema. Em termos estatísticos, uma hipótese é uma afirmação sobre um parâmetro de uma população, tais como média, proporção, desvio-padrão, coeficiente de correlação, etc. Uma vez definida a hipótese, esta deverá ser comprovada ou rejeitada. A ferramenta de comprovação é denominada *teste de hipótese.*

Na comparação entre dados numéricos, sujeitos a variabilidade, extraídos de amostras de uma mesma população, uma diferença real entre eles pode não ser evidente à primeira vista. A afirmativa inversa, ou seja, de que a ausência de diferença real pode não ser evidente à primeira vista, também é verdadeira.

Por convenção, podemos formular os problemas através de 2 hipóteses:

a.1) Hipótese nula

Quando temos um problema de comparação de dois tratamentos é usual fixar como hipótese de interesse a inexistência de diferença entre os dois tratamentos comparados. A hipótese a ser testada é chamada de hipótese nula (H_0).

a.2) Hipótese Alternativa

A hipótese nula deve ser comparada com uma hipótese alternativa, denominada H_1. Para

cada situação existem muitas hipóteses alternativas adequadas. Seguindo convenção, a hipótese alternativa é a inexistência de igualdade entre os tratamentos.

Exemplo:
$H_0 : pc = pt$ *versus* $H_1 : pc \neq pt$

Onde pc e pt são respectivamente as probabilidades de se observar a resposta de interesse entre os controles e entre os pacientes do grupo tratamento.

É importante notar que as hipóteses definidas se referem à comparação do parâmetro populacional dos pacientes controles com o parâmetro populacional do grupo tratamento. No entanto, para testar estas hipóteses são empregados resultados baseados em amostras. Logo, os resultados amostrais são generalizados, após o teste de hipótese, para todo o universo estudado.

Existem situações inerentes a alguns experimentos clínicos nos quais hipóteses diferentes podem ser utilizadas, como veremos no próximo capítulo.

4.2. Critério de decisão

Após decidirmos as hipóteses a serem testadas, teremos que construir um critério baseado no qual a hipótese H0 será julgada. O critério de decisão é baseado na estatística de teste. De uma forma bem genérica e intuitiva podemos dizer que a estatística do teste mede a discrepância entre o que foi observado na amostra e o que seria esperado se a hipótese nula fosse verdadeira. Rejeitaremos a hipótese nula se o valor da estatística de teste for "grande", o que traduziria uma discrepância entre os dados. Na prática para se decidir quão "grande" é o valor da estatística de teste é necessária a comparação entre o valor obtido e o valor estabelecido em uma distribuição hipotética de dados. Pequenas diferenças podem ser devido ao acaso em função da variabilidade dos dados; quanto maior a diferença menor é a probabilidade do acaso para sua explicação. Nesta circunstância uma relação de causa e efeito, ou de concomitância, pode ser inferida.

Para que cálculos estatísticos possam ser realizados, vários critérios necessitam ser definidos a priori, como veremos em seguida.

4.3 - Erros tipo I e II (nível de significância e poder do teste)

A decisão de rejeitar H0 quando de fato ela é verdadeira é chamada de erro tipo I. Para evitá-lo, escolhemos um critério de decisão que torna este erro pouco provável. Na literatura, a probabilidade de cometer esse erro recebe o nome de nível de significância do teste, sendo representado pela letra grega α(alfa).

Há um segundo tipo de erro, erro tipo II, que consiste em não rejeitar a hipótese nula sendo que ela é falsa. Isto implicaria na não liberação do novo tratamento, cujo efeito real não está sendo percebido. É representado por β(beta).

Convencionou-se que o erro mais sério seria o tipo I. O quadro 4.1 a seguir sintetiza os erros possíveis associados a cada decisão tomada em um teste de hipóteses.

QUADRO 4.1 - Erros possíveis associados a teste de hipóteses

Conclusão do teste	Situação H_0 verdadeira	Real H_0 falsa
Não rejeitar H0	decisão correta	erro tipo II
Rejeitar H0	erro tipo I	decisão correta

A capacidade de um teste identificar diferenças que realmente existem, ou seja, de rejeitar H_0 quando é realmente falsa, é denominada poder do teste e é definida como 1 - β.

4.4. Probabilidade de significância (valor p)

Existem duas opções para expressar a conclusão final de um teste de hipóteses. A primeira consiste em comparar o valor da estatística de teste com o ponto crítico a partir da distribuição teórica, específica para o teste, para um valor pré-fixado do nível de significância (por exemplo 5% ou 1%), conforme descrito na figura 4.1.

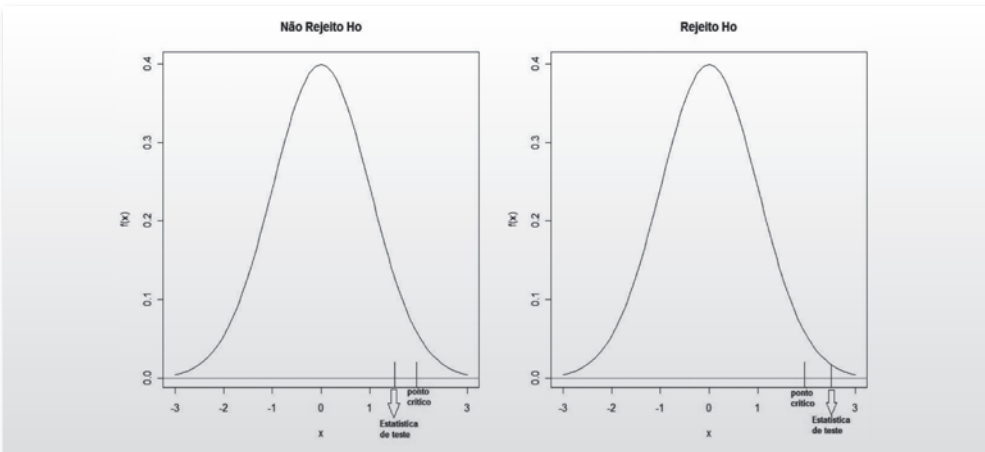

FIGURA 4.1 - Conclusão para um teste de hipótese. Na primeira curva, a estatística de teste se encontra fora da área de rejeição, logo não rejeito Ho. Para a segunda curva, a estatística de teste se encontra dentro da área de rejeição, logo rejeito H_0.

Na segunda opção, a mais usada, o interesse é quantificar a ocorrência do que foi observado ou de resultados mais extremos, sob a hipótese da igualdade dos grupos. Assim, essa opção, baseia-se na probabilidade de ocorrência de valores iguais ou superiores ao assumido pela estatística de teste, sob a hipótese de que H0 seja verdadeira, conforme mostrado na figura 4.2.

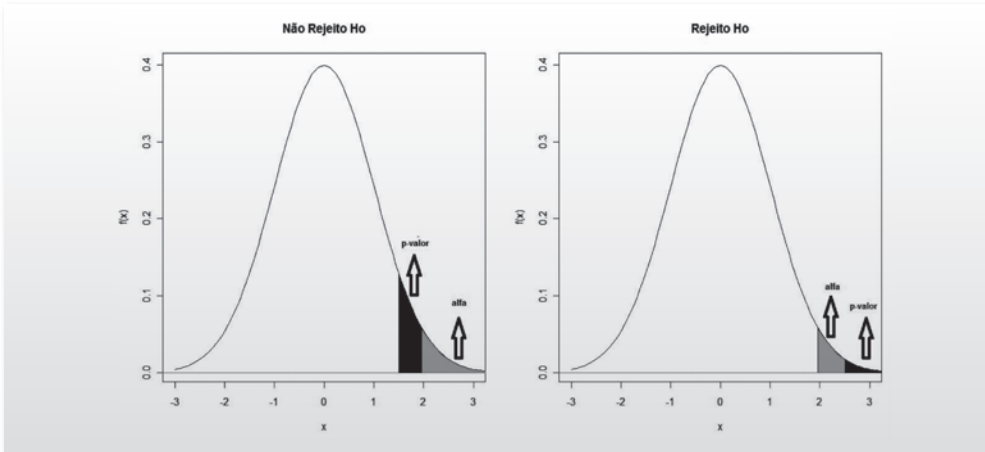

FIGURA 4.2 - CConclusão para um teste de hipótese. Na primeira curva, o valor p é maior do que o nível de significância (alfa), logo não rejeito H_0. Para a segunda curva, o valor p é menor do que o nível de significância(alfa), logo rejeito H_0.

Este número é chamado de probabilidade de significância ou valor p e freqüentemente é indicado apenas por p. Como o valor p é calculado supondo-se que H0 seja verdadeira, duas conjecturas podem ser feitas quando se obtém um valor muito pequeno. Um evento que é extremamente raro pode ter ocorrido ou a hipótese H_0 não deve ser verdadeira, isto é, a conjectura inicial e conservadora não parece plausível.

De um modo geral, na área médica, considera-se que valor p menor ou igual a 0,05 indica que há evidências para rejeitar H_0, ou seja, há diferença significativa entre os grupos.

Nas outras situações a diferença encontrada não é significante, do ponto de vista estatístico. Esses pontos de corte são arbitrários e não se deve dar uma importância exagerada a eles. É inaceitável que os resultados de dois estudos em que os valores p sejam 0,045 e 0,055 sejam interpretados de forma diferente para $\alpha = 0,05$. Esses valores devem levar a conclusões muito parecidas e não diametralmente opostas (significativo e não significativo).

4.5 Hipóteses bilaterais *versus* unilaterais

As hipóteses alternativas, respectivamente para o teste de comparação de proporções, de médias ou de medianas (no caso de testes não-paramétricos) são: $H_1: p_1 \neq p_2$ e $H_1: \mu_1 \neq \mu_2$

Mas podem ser desmembradas como: $H_1: p_1 > p_2$ ou $H_1: p_1 < p_2$ e $H_1: \mu_1 > \mu_2$ ou $H_1: \mu_1 < \mu_2$.

Estas hipóteses assumem, portanto, que qualquer um dos dois grupos pode ter uma proporção ou média maior do que o outro. Por isto este tipo de hipótese é denominada bilateral. O valor p bilateral é a probabilidade de se obter em qualquer direção uma diferença igual ou mais extrema do que a observada.

Existe também a possibilidade de se formular hipóteses alternativas unilaterais (H_1), como a seguir:

Situação	Proporções	Médias
(1)	$H_1: p_1 > p_2$	$H_1: \mu_1 > \mu_2$
(2)	$H_1: p_1 < p_2$	$H_1: \mu_1 < \mu_2$

Nestes casos, as comparações são estabelecidas em uma determinada direção. Assim, por exemplo, ao se comparar um procedimento novo com o padrão, estamos avaliando se a inovação deve ser recomendada. Portanto, a escolha de hipóteses unilaterais ou bilaterais influencia decisivamente a interpretação dos resultados da análise estatística.

As duas opções de testes (unilateral ou bilateral) estão disponíveis em programas de computador. Em geral, o valor p para teste bilateral é o dobro do valor p correspondente à hipótese unilateral (figura 4.3).

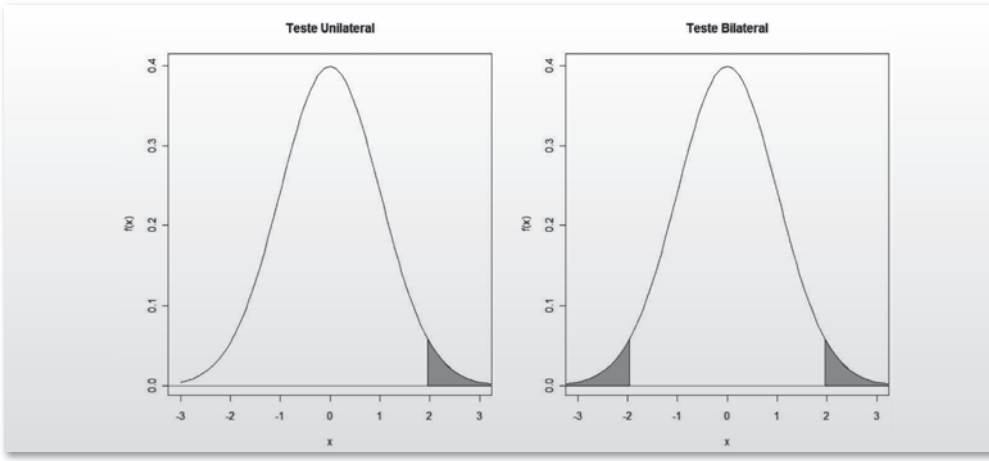

FIGURA 4.3 - Opções de testes de hipóteses (unilateral ou bilateral).

Há circunstâncias em que hipótese unilateral é a melhor forma de se descrever a questão de interesse. Em estudos comparando uma inovação de um procedimento padrão, a hipótese alternativa mais interessante é que a inovação é superior, um apelo à hipótese unilateral.

Um teste unilateral também pode ser justificado quando se pode afirmar que uma das direções contempladas pela hipótese bilateral é completamente inconcebível.

O argumento mais forte contra o uso de hipóteses unilaterais é que, por maior que seja a evidência de que um tratamento seja superior ao outro, nunca se tem certeza absoluta do que realmente pode acontecer. Embora existam razões para esperar que novas drogas, ou novos procedimentos, sejam melhores que os do grupo controle (caso contrário o estudo não estaria sendo realizado), ainda assim existe a possibilidade, mesmo que remota, de que seus resultados sejam piores. Na comparação de uma droga com o placebo, não se pode descartar a possibilidade de que a droga tenha um efeito deletério, e portanto, não deva ser recomendada.

Na escolha entre hipótese bilateral e unilateral os seguintes aspectos devem ser considerados:

1. O tipo de hipótese adotado deve preceder a análise dos dados, isto é, a escolha não deve ser influenciada pelo resultado da amostra.

2. O teste bilateral é mais conservador que o unilateral. Na maioria dos casos, testes unilaterais são vistos como uma maneira de se exagerar a força dos achados. Se houver qualquer dúvida, deve-se optar pelo teste bilateral.

3. Segundo alguns autores, a distinção entre testes unilaterais e bilaterais não é de fundamental importância na interpretação dos resultados, desde que fique claro qual foi usado.

4. Mesmo quando o teste unilateral pode ser justificado, pode-se encontrar resistência editorial para publicar tais achados.

5. Teste bilateral é a forma padrão, usada em periódicos médicos.

Alguns estatísticos e editores de jornais acreditam que o valor p unilateral nunca deva ser usado. O primeiro argumento é a uniformidade de apresentação dos resultados, tal que um determinado valor p tenha um mesmo significado em todos os artigos. Segundo, acreditam que situações que justificam o uso do testes unilaterais são extremamente raras. Terceiro, em estudos sobre importantes questões, como a regulamentação de uma droga, o valor p é apenas um fator usado na tomada de decisões. O critério de que o valor p seja menor que 0,05, em geral é insuficiente para estabelecer eficiência e pelo menos o teste bilateral é mais conservador.

Aceitando estes argumentos e considerando a padronização já existente na maioria dos periódicos médicos, recomendamos o uso rotineiro de hipóteses bilaterais.

4.6 Critérios de escolha e exemplos

4.6.1 Critérios de escolha

Dentre os inúmeros testes e técnicas estatísticos que se apresentam no contexto de um trabalho de pesquisa, é natural certo grau de desorientação inicial quanto à identificação daqueles que são ou não aplicáveis a cada situação. Para realizar as escolhas adequadas, é importante considerar alguns parâmetros básicos dos dados a serem analisados, tais como:

- *N° de Amostras*: O número de grupos distintos sendo analisados.
- *Relações Entre Amostras*: Refere-se a duas ou mais amostras consistirem ou não de múltiplas medidas das mesmas entidades ou de entidades relacionadas (serem ou não pareadas ou casadas).
- *Escala Numérica*: A forma que os dados foram registrados (escala qualitativa, quantitativa discreta e quantitativa contínua).
- *Distribuição*: A densidade de probabilidade (distribuição de probabilidade) dos dados (Normal ou Não-Normal).
- *Dependência Entre Variáveis*: O conhecimento de que uma variável pode contribuir ou não para o conhecimento de outras (respectivamente, serem associadas ou independentes entre si).

São estes os fatores que determinam quais os procedimentos gráficos e analíticos possíveis para cada combinação de número de amostras e tipos de dados.

A tabela 4.2 abaixo indica as técnicas estatísticas que podem ser aplicadas para a comparação entre os parâmetros de dois ou mais grupos de dados.

Tabela 4.2 - Testes estatísticos utilizados na comparação entre parâmetros de duas ou mais amostras

Comparações				
N° de Amostras	**Tipo de Relação**	**Distribuição**	**Escala Numérica**	**Análises Aplicáveis**
2	Pareadas	Normal	Quant. Contínua Quant. discreta	Teste t de Student Pareado
2	Pareadas	Não-Normal	Quant. discreta, Quant. contínua	Sign-Test, Wilcoxon *Matched-Pairs Test*
2	Pareadas	Não-Normal	Qualitativa Dicotômica*	Teste de McNemar
2	Não-Pareadas	Normal	Quant. Contínua Quant. discreta	Teste t de Student
2	Não-Pareadas	Não-Normal	Quant. discreta, Quant. contínua	Teste Mann-Whitney U
2	Não-Pareadas	Não-Normal	Qualitativa	Teste de Qui-Quadrado
≥3	Pareadas	Normal	Quant. Contínua Quant. discreta	ANOVA c/ Medidas Repetidas
≥3	Pareadas	Não-Normal	Quant. discreta, Quant. contínua	Teste de Friedman
≥3	Pareadas	Não-Normal	Qualitativa	Teste Q de Cochran
≥3	Não-Pareadas	Normal	Quant. Contínua Quant. discreta	ANOVA c/ Grupos Independentes
≥3	Não-Pareadas	Não-Normal	Quant. discreta, Quant. contínua	Teste de Kruskal-Wallis
≥3	Não-Pareadas	Não-Normal	Qualitativa	Teste de Qui-Quadrado

* Variável com apenas dois valores ou duas categorias (variável binária).

O quadro a seguir mostra as técnicas analíticas e procedimentos gráficos aplicáveis quando se quer verificar a existência e/ou caracterizar a relação entre duas ou mais variáveis.

Tabela 4.3 - Técnicas analíticas e procedimentos gráficos usados na determinação da relação entre duas ou mais variáveis

Relação / Associação				
N° de Variáveis	**Escala Numérica das Variáveis**	**Distribuição**	**Análises Aplicáveis**	**Gráficos Aplicáveis**
2	Quantitativa contínua	Normal	Correlação de Pearson, Regressão Linear Simples	Diagrama de Dispersão (X,Y)
2	Quant. discreta e/ou Quant. contínua	Não-Normal	Correlação de Spearman	Diagrama de Dispersão (X,Y)
2	Qualitativa	Não-Normal	Odds Ratio, Teste de Qui-Quadrado	---
≥3	Quantitativa contínua	Normal ou Não-Normal	Regressão Múltipla (Linear e Não-linear)	Diagrama Previsão vs. Observação
≥3	Qualitativa	Não-Normal	Análise Discriminante	---
≥3	Quantitativa contínua	Normal e/ou Não-Normal	Regressão Linear Múltipla, Regressão Não-Linear	---
≥3	Qualitativa dicotômica* (Variável-Resposta) e Qualitativa ou Quantitativa (variáveis explicativas)	Normal e/ou Não-Normal	Regressão Logística	---

* Variável com apenas dois valores ou duas categorias (variável binária).

Os quadros acima apontam para as análises de dados possíveis nas diversas situações de pesquisa, porém, não indicam exatamente os procedimentos a serem adotados em cada situação. Isso ocorre devido ao fato de que a decisão final depende não apenas das restrições matemáticas, mas também dos objetivos do estudo e da própria natureza dos achados que vão sendo produzidos. É importante, contudo, ter em mente que as tabulações apresentadas constituem um mapa de referência que deixa claro espaço para ações, dentro do qual pode se manifestar a liberdade do pensador analítico.

4.6.2 Exemplos

I - Testes paramétricos

Ilustraremos alguns testes estatísticos da tabela 4.3 acima omitindo no entanto, o cálculo da estatística de teste, que é fornecido pelos programas estatísticos usuais.

Variável dicotômica: amostras independentes

Neste caso, a variável de interesse é a ocorrência de um determinado evento, como o desenvolvimento de uma doença, ou a presença de certo atributo, por exemplo, albinismo.

Usaremos exemplo citado por Siqueira e Teixeira (2002), a propósito do tratamento de pacientes aidéticos com AZT ou placebo e cujos resultados são descritos na tabela 4.4.

O problema da comparação das probabilidades de ocorrência do evento ou do atributo nos dois grupos é formulado através das hipóteses

H_0: p AZT vivo = p AZT morto = p Placebo vivo = p Placebo morto
H_1: pelo menos 1 grupo diferente

Tabela 4.4 - Número de sobreviventes tratados com AZT ou placebo

Grupo	Situação		
	Vivo	Morto	Total
AZT	144	1	145
Placebo	121	16	137
Total	265	17	282

Fonte: Siqueira e Teixeira (2002)

Calculado o valor da estatística do teste (teste do X^2), é preciso decidir se este é ou não um valor 'grande'. Assim, para se tomar uma decisão sobre a igualdade ou não das duas proporções, é preciso conhecer o comportamento, isto é, a distribuição estatística dos valores de X^2 quando as proporções são iguais. Esta distribuição foi obtida e recebeu o nome de qui-quadrado com 1 grau de liberdade, é indicada por X^2_1 e está sintetizada em tabelas de fácil utilização. A figura 4.4 ilustra a distribuição do X^2 com 1 grau de liberdade.

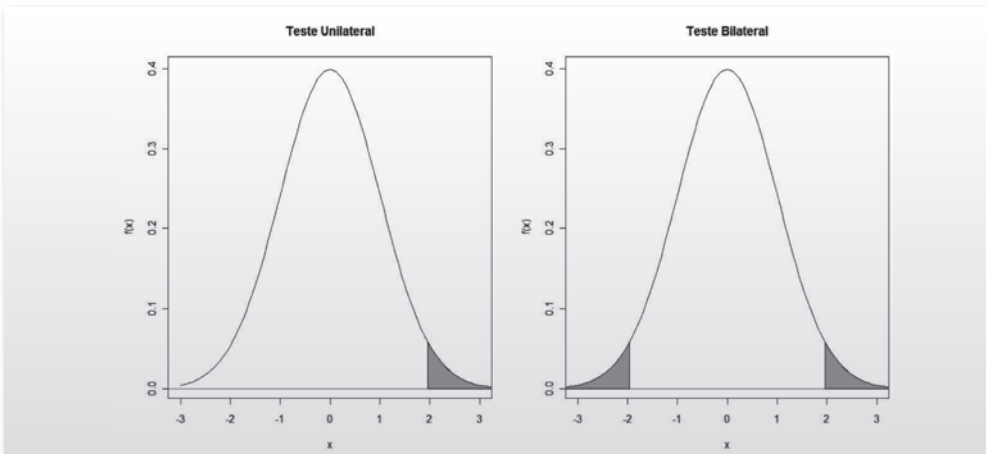

FIGURA 4.4 - Distribuição qui-quadrado e o valor da estatística de teste.

O valor da estatística de teste foi de 13,14. Como este valor é maior que 3,84, obtido da distribuição de X^2_1 para um nível de significância de 0,05, rejeita-se a hipótese de igualdade entre os grupos de tratamento e controle. Em outras palavras, decidimos com 95% de confiança que há evidência do efeito do AZT. Para obtermos a probabilidade de significância devemos calcular a probabilidade de encontrar valores maiores que 13,14, isto é, $P[X^2_1 \geq 13{,}14]$, sendo verdadeira a hipótese de igualdade das proporções. Da tabela da distribuição do X^2_1, vemos que este valor é aproximadamente 0,0001, ou seja, o valor p é 0,0001.

Baseado neste estudo, é possível afirmar, com 95% de certeza, que o AZT tem efeito de prolongar a vida de pacientes com AIDS, sendo esta a primeira evidência necessária para a liberação do medicamento.

Apresentaremos, agora, um teste para a comparação entre 2 proporções: o teste Z. Trata-se de um teste aproximado que requer grandes amostras para a sua aplicação. Um critério é exigir que n_1p_1 e n_2p_2 excedam o valor 5.

Queremos testar a hipótese de equivalência entre dois tratamentos:

$H_0: p_1 = p_2$ *versus* $H_1: p_1 \neq p_2$

Com o objetivo de comparar a eficácia de dois preventivos contra náusea, dividiu-se aleatoriamente uma amostra de 400 marinheiros em dois grupos de 200. Um grupo recebeu a pílula A e o outro a pílula B, sendo que no 1º grupo 152 não enjoaram durante uma tempestade e no outro grupo apenas 132. Há indicações de que a eficácia das pílulas A e B seja a mesma?

Sejam p_A e p_B as proporções de marinheiros que não enjoam, respectivamente com as pílulas A e B. Temos $n_A = 200$ e $n_B = 200$,

$$\hat{p}_A = \frac{152}{200} = 0{,}76 \quad \hat{p}_B = \frac{132}{200} = 0{,}66$$

O valor da estatística de teste é: 2,22

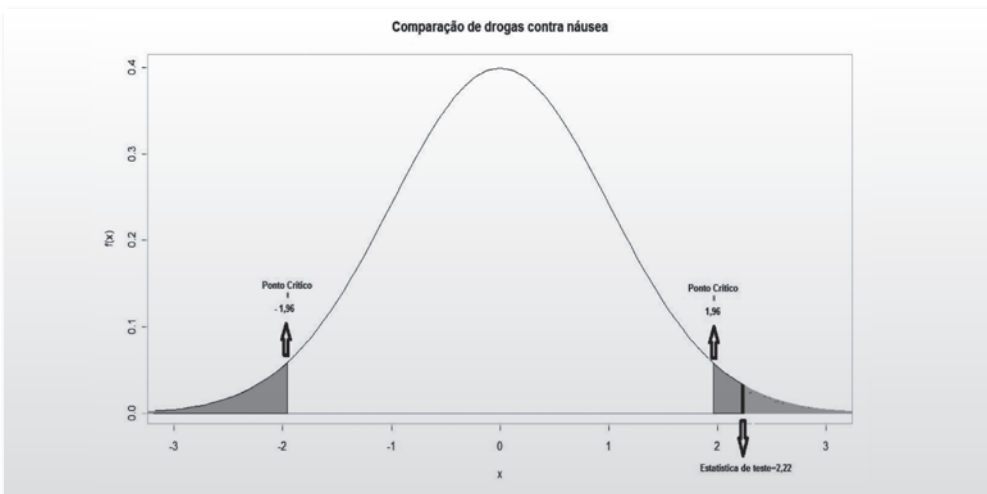

FIGURA 4.5 - Curva normal com o valor da estatística de teste e o ponto crítico.

Fixando-se o nível de significância em 5%, rejeita-se H_0. O valor p encontrado para Z é 0,026. Portanto pode-se concluir, com confiança de 95%, que as duas pílulas não são igualmente efetivas. Há indicações de que a pílula A oferece maior proteção contra náuseas comparada à pílula B.

Variável dicotômica: amostras pareadas

Foram avaliados 100 doentes com cefaleias frequentes. Os mesmos 100 doentes tomaram durante um mês um determinado medicamento A e no mês seguinte o medicamento B. Pediu-se aos pacientes que registrassem se durante cada mês tiveram ou não dores de cabeça.

Tabela 4.5 - Pacientes com cefaléias frequentes usando dois tipos de medicamentos

Medicamento B	Medicamento A		Total
	Sem cefaléia	Com cefaléia	
Sem cefaléia	45	4	49
Com cefaléia	17	34	51
Total	62	38	100

Fonte: http://medicina.med.up.pt/im/im2004/teoricas/categoricas.ppt

O teste apropriado para esta situação é o teste de McNemar.

As hipóteses são:

H_0: A percentagem de doentes com cefaléias usando o medicamento A é igual a percentagem de doentes com cefaleias usando o medicamento B.

H_1: A percentagem de doentes com cefaléias usando o medicamento A é diferente da percentagem de doentes com cefaleias usando o medicamento B.

O valor da estatística do teste de McNemar é: 6,86, conforme ilustrado na figura 4.6.

$$X^2_{McN} = \frac{((4-17)-1)^2}{4+17} = 6{,}86$$

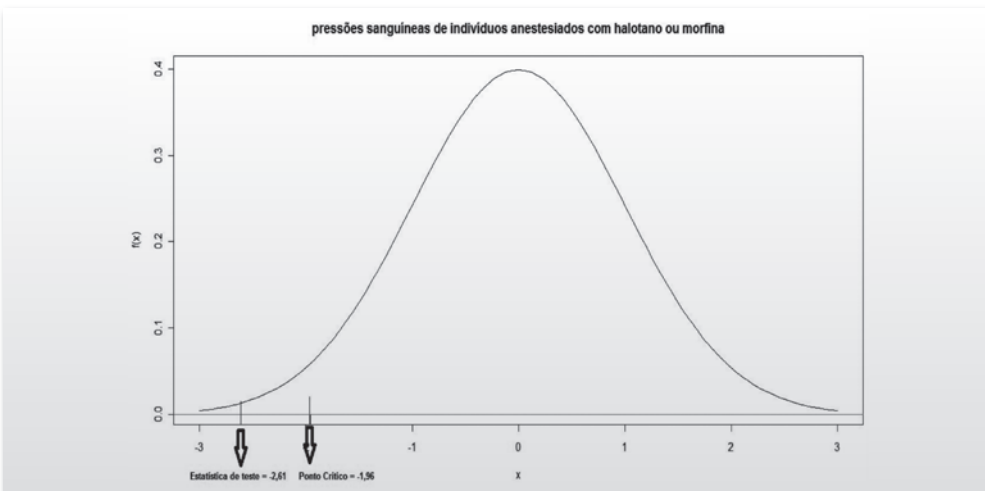

FIGURA 4.6 - Estatística de teste do teste de McNemar.

Este valor deve ser comparado com 3,84 para um nível de significância de 5%. Ou seja, com uma confiança de 95% a percentagem de doentes com cefaléias usando o medicamento A é diferente da percentagem de doentes com cefaleias usando o medicamento B.

Resposta Contínua: amostras independentes

Agora apresentaremos a metodologia para comparar dois grupos de pacientes (por exemplo, doentes vs não doentes) em relação a uma resposta contínua, por exemplo pressão sistólica. Testa-se então, nesse caso, a igualdade das médias das respostas de dois tratamentos.

Sejam μ_1 e μ_2 as médias da variável estudada para os dois grupos, respectivamente. As hipóteses a serem testadas são:

$H_0: \mu_1 = \mu_2$ *versus* $H_1: \mu_1 \neq \mu_2$

Vamos apresentar agora, o teste mais conhecido (o teste t para duas amostras) e adequado para situações em que as respostas aos dois tratamentos são variáveis quantitativas com distribuição normal (gaussiana) para os dois grupos. Assim, as suposições para se usar este teste são que as variáveis estudadas têm distribuições gaussianas com o mesmo desvio padrão. Para isso, deve-se realizar o teste de normalidade (Kolmogorov-Smirnov) e o teste de duas variâncias (teste de Fisher).

Um estudo relata os resultados de um ensaio clínico aleatorizado, duplo-cego, realizado com o objetivo de comparar a tianeptina com o placebo. Participaram desse estudo pacientes de Belo Horizonte, Campinas e Rio de Janeiro.

Sucintamente, o ensaio consistiu em administrar a droga a dois grupos de pacientes, compostos de forma aleatória, e quantificar a depressão através da escala de MADRS, em que os valores maiores indicam maior gravidade da doença. O escore foi obtido para cada paciente 7, 14, 21, 28 e 42 dias após o início do ensaio.

Pelo planejamento adotado, os dois grupos não diferiam em termos de depressão no início do ensaio. Assim, uma evidência sobre o efeito da tianeptina é obtida comparando-se os dois grupos ao fim de 42 dias.

A Tabela 4.6 apresenta os escores finais dos pacientes dos dois grupos admitidos em Belo Horizonte.

Tabela 4.6 - Escore final na escala MADRS de pacientes dos dois grupos admitidos em Belo Horizonte

Grupo	Escores
Placebo	6 33 21 26 10 29 33 29 37 15 2 21 7 26 13
Tianeptina	10 8 17 4 17 14 9 4 21 3 7 10 29 13 14 2

Fonte: Siqueira e Teixeira (2002)

Para se efetuar o teste t é preciso usar as seguintes informações:
$n_1 = 15 \quad \bar{x}_1 = 20{,}53 \quad s_1 = 11{,}09$
$n_2 = 16 \quad \bar{x}_2 = 11{,}37 \quad s_2 = 7{,}26$

A estatística de teste encontrada foi igual a 2,74, que comparando com o valor de $t_{29;0,975} = 2{,}045$ leva à rejeição da igualdade entre os dois grupos no nível de 5% (figura 4.7). O valor p encontrado foi 0,0104.

Para aplicarmos o teste t é necessário que os dois grupos comparados tenham a mesma variabilidade, o que nem sempre ocorre na prática. No caso de amostras grandes (n_1 e $n_2 \geq 30$) dispomos de um teste em que não é necessária qualquer suposição adicional sobre σ^2_1 e σ^2_2, ou seja, as variâncias podem ser iguais ou diferentes: o teste Z para comparação de médias.

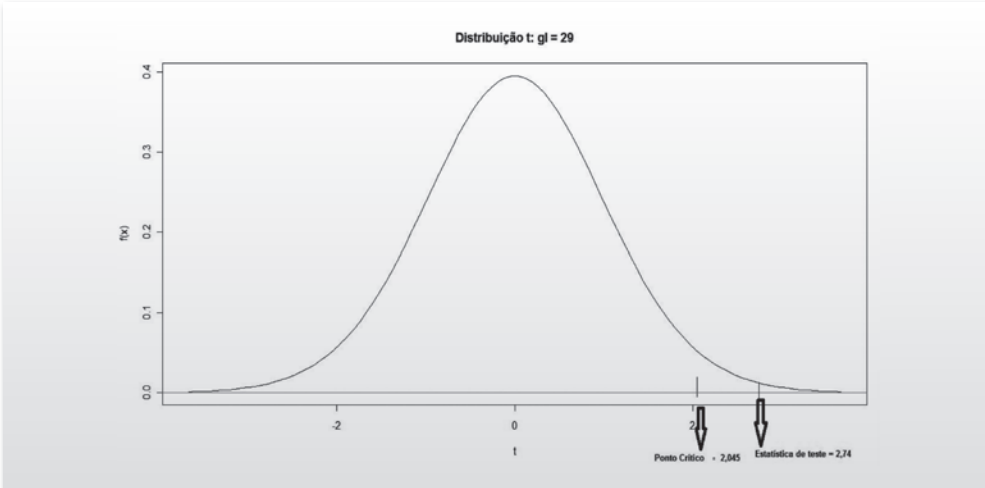

FIGURA 4.7 - Distribuição de t com grau de liberdade de 29 para nível de significância de 5,0%.

Um estudo foi feito em um grande número de pacientes para comparar dois agentes anestésicos, o halotano, de efeito poderoso, mas que pode causar complicações em pacientes com problemas cardíacos e a morfina, que tem pequeno efeito na atividade cardíaca. Todos pacientes foram submetidos a uma cirurgia de rotina para reparo ou substituição da válvula cardíaca. Para obter duas amostras comparáveis, eles foram alocados aleatoriamente a cada tipo de anestesia.

A fim de estudar o efeito desses dois tipos de anestesia, foram registradas variáveis hemodinâmicas, como pressão sanguínea antes da indução da anestesia, após a anestesia, mas antes da incisão, e em outros períodos importantes durante a operação. A questão que surge é se o efeito do halotano e da morfina na pressão sanguínea é o mesmo. A média e o desvio-padrão dos dois grupos encontram-se a seguir:

Tabela 4.7 - Média e desvio-padrão da pressão sanguínea (mmHg) segundo o tipo de anestesia

Informações sobre a amostra	Anestesia	
	Halotano	Morfina
Média	66,9	73,2
Desvio-padrão	12,2	14,4
n	61	61

Fonte: Siqueira e Teixeira (2002)

Nas condições do problema, as hipóteses são:

H_0: $\mu_1 = \mu_2$ e H_1 :$\mu_1 \neq \mu_2$

Isto é, devemos testar a diferença entre as pressões sanguíneas médias de indivíduos anestesiados com halotano ou morfina.

Como as amostras são grandes, podemos usar o teste Z, cujo valor da estatística do teste é de -2,61

Adotando um nível de significância de 5%, o resultado é estatisticamente significativo, já que $|-2,61| > 1,96$ (figura 4.8). Além disso, o valor p = 0,009, que é menor que o valor de α estipulado, indicando que os dois anestésicos não são equivalentes.

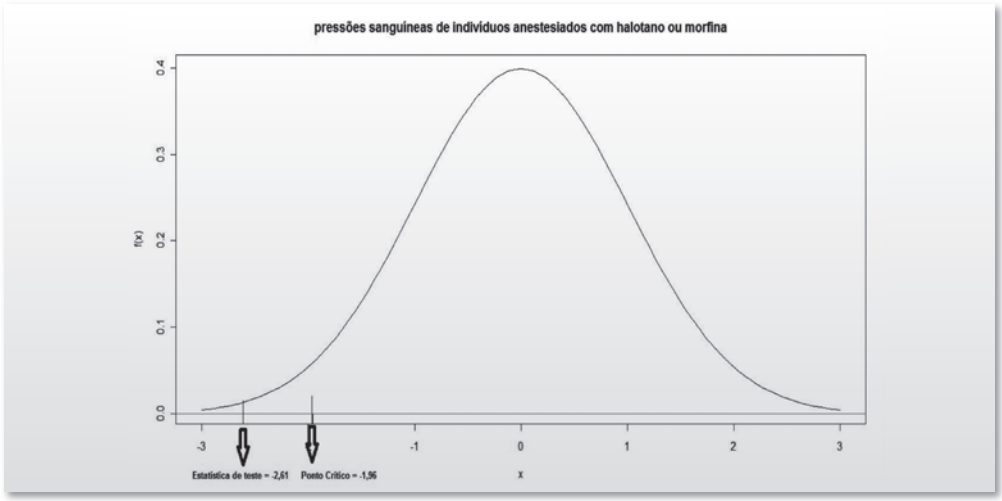

FIGURA 4.8 - Teste da diferença entre as pressões sanguíneas médias de indivíduos anestesiados com halotano ou morfina.

Variável contínua ou discreta: amostras pareadas

Consideremos a seguinte pergunta: será que o armazenamento da amostra do sangue influencia o valor da medida do colesterol e do triglicérides?

Neste caso, o problema de interesse é uma comparação entre dois grupos de medidas: de triglicérides, por exemplo. É razoável supor, e existem evidências empíricas neste sentido, que a distribuição estatística do nível de triglicérides é normal (gaussiana). No entanto, é aconselhável usar o teste de normalidade para o nível de triglicérides. Se o possível efeito do armazenamento se dá apenas no aumento ou decréscimo na média da distribuição, não na sua variabilidade, então as hipóteses a serem testadas são:

$H_0: \mu_1 = \mu_2$ e $H_1: \mu_1 \neq \mu_2$

Onde μ_1 e $_2$ são as médias antes e depois do armazenamento. A escolha de H_0 implica que, na ausência de outras evidências, consideremos que o armazenamento não tem efeito. Intuitivamente, o critério de decisão, a ser utilizado para testar H0, deve ser baseado nas diferenças entre os níveis encontrados de triglicérides nas duas ocasiões das medidas. Se houver influência do armazenamento, então essas diferenças devem ser diferentes de zero.

O problema de escolha de um critério de decisão reduz-se a escolher uma forma de verificar se as diferenças são provenientes de uma distribuição com média zero.

Um exemplo semelhante pode ser ilustrado pelo estudo cujo objetivo era avaliar a efetividade de uma dieta combinada com um programa de exercícios físicos na redução do nível sérico de colesterol.

A tabela abaixo mostra os níveis de colesterol de 12 participantes no início e no final do programa.

Tabela 4.8 - Níveis de colesterol no início e no final do programa

Programa		Diferença	Desvio	Desvio ao quadrado
Início (x_1)	Final (x_2)	d = x1-x2	d- $\bar{d}$	$(d-\bar{d})^2$
201	200	1	-19,16	367,36
231	236	-5	-25,16	633,36
221	216	5	-15,16	230,03
260	233	27	6,83	46,69
228	224	4	-16,16	261,36
237	216	21	0,83	0,69
326	296	30	9,83	96,69
235	195	40	19,83	393,36
240	207	33	12,83	164,69
267	247	20	-0,16	0,03
284	210	74	53,83	2898,03
201	209	-8	-28,16	793,36

Fonte: Arango (2005).

Quanto maior o valor $\bar{d}$ (que representa a média das diferenças x_1-x_2) maior a evidência de que o programa reduz o nível de colesterol; quanto menor a variabilidade das diferenças individuais, maior a chance de se detectar um efeito médio significativo, isto é, uma redução significativa do colesterol devido à ação do programa e não ao acaso. Estes aspectos podem ser avaliados através do teste t.

Sejam μ_a e μ_d respectivamente as médias dos níveis de colesterol antes e depois do programa. Para testar a hipótese de que o programa altera o nível de colesterol (H_0: $\mu_a = \mu_d$ x H_1: $\mu_a \neq \mu_d$) será aplicado o teste t (11 graus de liberdade).

Apenas dois participantes tiveram o nível de colesterol aumentado após o programa, mas por pequenas quantidades (5 e 8 mg/dl). As médias antes e depois do programa são respectivamente 244,25 e 224,08, correspondente a uma redução média de 20,12 mg/dl ($\bar{d}$ = 20,17). A estatística de teste foi de 3,02 (p = 0,012), isto é, há evidência de que, em média, o programa altera o nível de colesterol (figura 4.9).

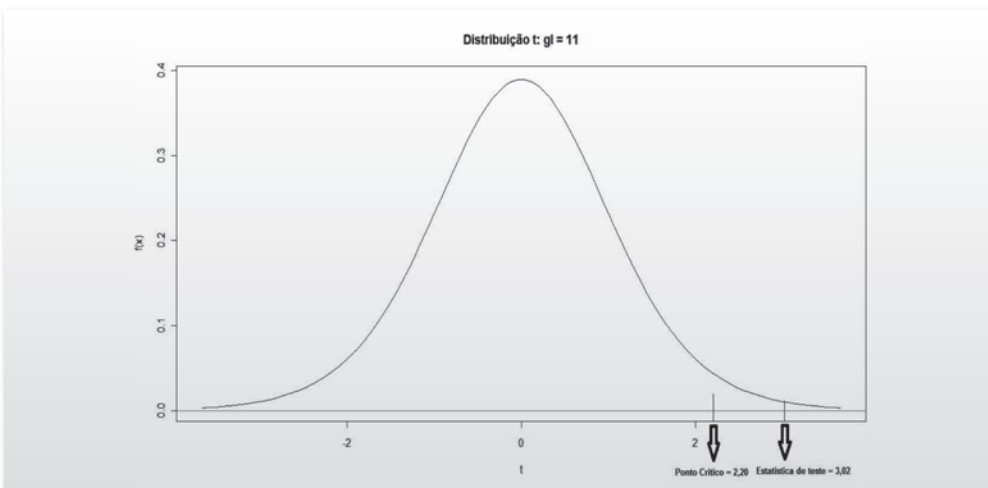

FIGURA 4.9 - Teste t para as médias antes e depois do programa

II - Testes Não Paramétricos

Os testes estudados até agora envolviam problemas nos quais a distribuição da população em estudo era conhecida, ou pelo menos nunca colocada em causa, e as hipóteses testadas apenas envolviam parâmetros populacionais.

No entanto, outros tipos de problemas podem ser colocados: se a distribuição de uma população é desconhecida e se pretender testar a hipótese de uma distribuição particular para aquela população, que fazer?

Os testes denominados *testes não paramétricos* ou t*estes de distribuição livre* constituem uma alternativa para este e outros tipos de problemas.

O termo "distribuição livre" é popularmente usado para indicar que os métodos são aplicáveis independentemente da forma da distribuição.

Estes métodos são, em geral, fáceis de aplicar, pois podem ser usados quando as hipóteses exigidas por outras técnicas não são satisfeitas.

Relembramos aqui que os testes paramétricos estudados até agora comportam uma diversidade de suposições fortes a que o seu emprego deve subordinar-se: as observações devem ser extraídas de populações com distribuição normal, as variáveis em estudo devem ser medidas em escala intervalar ou de razão, de modo a que seja possível utilizar operações aritméticas sobre os valores obtidos das amostras (adição, multiplicação, obtenção de médias, etc.).

Apesar de haver certas suposições básicas associadas à maioria das provas não paramétricas, essas suposições são em menor número e mais fracas do que as associadas às provas paramétricas.

Servem para pequenas amostras e, além disso, a maior parte das provas não paramétricas aplica-se a dados medidos em escala ordinal e, alguns, a dados em escala nominal.

Dentre uma vasta gama de testes não paramétricos disponíveis, foram selecionados, para análise neste capítulo, apenas alguns testes de utilização freqüente ou que complementam, de alguma forma, os testes paramétricos discutidos anteriormente.

Em resumo, nos exemplos mencionados anteriormente, os testes aplicados foram baseados em distribuições de probabilidade, denominado testes paramétricos; contudo, abordaremos nos próximos exemplos testes não paramétricos, ou seja, testes que não possuem distribuição de probabilidade.

Resposta Contínua ou Discreta: duas amostras independentes

O teste de Mann-Whitney é a versão não paramétrica do teste t. Sendo assim, o interesse é testar se as medianas são iguais ou diferentes entre si.

A tabela ao lado exibe a taxa de uréia de pacientes renais e sua condição quanto à presença de insuficiência renal aguda (IRA).

Neste tipo de situação, cruzamento de uma variável quantitativa (uréia) com uma variável dicotômica (IRA), é viável empregar o teste de Mann-Whitney.

Usando um programa de estatística adequado, temos que a estatística de teste foi de -2,76. Este valor corresponde a um valor p = 0,00289. Como o valor-p é menor que o nível de significância de 5%, logo pode-se dizer que existe uma diferença, significativa, entre a taxa de uréia de portadores ou não portadores de IRA.

Tabela 4.9 - Pacientes segundo a taxa de uréia (mg/100ml) e a presença ou não de IRA

Paciente	Uréia	IRA
01	92	Sim
02	120	Sim
03	68	Sim
04	70	Sim
05	77	Sim
06	63	Sim
07	26	Não
08	33	Sim
09	38	Não
10	25	Não
11	21	Não
12	15	Não

Fonte: Arango (2005).

Resposta Contínua ou Discreta: três ou mais amostras independentes

O teste de Kruskall Wallis é utilizado quando não é possível aplicar a Anova, pois os dados não seguem distribuição normal. Sendo assim, as hipóteses são definidas pela mediana e não pela média.

A tabela abaixo mostra o Índice de Massa Corporal (IMC) e o grau de estadiamento do câncer colorretal em 18 pacientes submetidos a cirurgia. O objetivo é verificar se o grau de estadiamento desta doença se relaciona o IMC.

Tabela 4.10 - IMC de três grupos de pacientes

Estadiamento I	Estadiamento II	Estadiamento III
22.41	22.26	20.83
27.99	28.24	22.31
19.57	18.37	18.22
19.56	22.10	20.88
19.15	7.33	18.73
22.21	22.21	21.27

Fonte: Dados Fictícios

Usando um *software* adequado, temos que a estatística de teste foi de 0,758. Este valor corresponde a um valor-p = 0,685. Como o valor-p é maior que o nível de significância de 5%, pode-se dizer que o estadiamento do câncer colorretal não se correlacionou com o Índice de Massa Corporal.

Resposta Contínua ou Discreta: duas amostras pareadas

O teste de Wilcoxon é utilizado quando não é possível aplicar o teste t pareado, pois os dados não seguem distribuição normal, ou seja, é a versão não paramétrica do teste t pareado. Sendo assim, o interesse é testar se as medianas são iguais ou diferentes entre si.

A tabela abaixo mostra o nível máximo de concentração (NMC) de 12 pacientes selecionados aleatoriamente, antes e depois da ingestão de determinada droga. O objetivo deste estudo era testar a eficácia desta droga em relação à capacidade de aprendizado.

Tabela 4.11 - Níveis máximos de atenção/concentração, em segundos, em uma amostra de 12 indivíduos, antes e depois da ingestão da droga de teste

Paciente	NMC Antes	NMC Depois
01	9	14
02	16	22
03	12	18
04	28	23
05	5	11
06	33	40
07	17	15
08	13	18
09	18	22
10	12	31
11	26	19
12	14	8

Fonte: Arango (2005).

Usando um programa de estatística adequado, temos que a estatística de teste foi de 1,44. Este valor corresponde a um valor-p = 0,074 para o teste unilateral. Como o valor-p é maior que o nível de significância de 5%, pode-se dizer que o uso da nova droga não aumenta a capacidade de aprendizado. Recomenda-se fazer o teste com um número maior de pacientes para se ter uma melhor conclusão sobre o efeito real da droga.

Resposta Contínua ou Discreta: três ou mais amostras pareadas

O teste de Friedman é uma generalização do teste de Wilcoxon para situações de mais de duas opções na comparação de dados. Este teste é utilizado quando não é possível aplicar o teste Anova com medidas repetidas, pois os dados não seguem distribuição normal. Sendo assim, as hipóteses são definidas pela mediana e não pela média.

A tabela 4.12 mostra dados fictícios sobre 25 pacientes, com diagnóstico de metástase em coluna vertebral. Para cada paciente, foi aplicado o questionário VAS (Escala Analógica Visual), em uma escala de 0 a 10, para avaliação de dor no período pré-operatório, pós-operatório e um ano após a cirurgia. O objetivo é avaliar a evolução da dor em pacientes com lesão metastática em coluna vertebral operados por abordagem posterior.

Usando um *software* adequado, temos que a estatística de teste foi de 7,96. Este valor corresponde a um valor-p = 0,000 para o teste bilateral. Como o valor-p é menor que o nível de significância de 5%, pode-se dizer que os valores de VAS diferem entre si quando comparados estatisticamente os períodos pré-operatório, pós-operatório e um ano após a cirurgia.

Observe que o teste de Friedman apenas conclui que pelo menos uma situação difere das demais. Neste caso, é necessário realizar comparação de dois a dois grupos para identificar a hierarquia desta diferença.

Tabela 4.12 - Escala Analógica Visual (VAS) para avaliação da dor no pré-operatório, pós-operatório e um ano após cirurgia, em uma amostra de 25 indivíduos

Paciente	VAS pré-operatório	VAS pós-operatório	VAS um ano após cirurgia
01	5	1	0
02	6	0	0
03	9	0	0
04	7	5	0
05	8	8	8
06	7	2	1
07	9	9	9
08	6	6	6
09	3	3	0
10	7	7	5
11	8	8	0
12	8	8	8
13	9	0	0
14	10	6	0
15	8	0	0
16	4	2	2
17	10	10	10
18	10	0	5
19	10	8	8
20	8	8	5
21	8	7	8
22	7	7	6
23	8	6	0
24	10	4	8
25	7	7	7

Fonte: Arango (2005).

A Tabela 4.13 mostra os resultados da comparação de dois a dois entre Escala Analógica Visual (VAS) para avaliação da dor no pré-operatório, pós-operatório e um ano após cirurgia.

Tabela 4.13 - Teste de comparação múltipla entre a Escala Analógica Visual (VAS) para avaliação da dor no pré-operatório, pós-operatório e um ano após cirurgia

Comparação Múltipla de VAS	Resultados		
	Mediana	Valor p	Conclusão
VAS pré-operatório (1º)	8.00	**0.000****	1º > 2º
VAS pós-operatório (2º)	6.00		
VAS pré-operatório (1º)	8.00	**0.000****	1º > 3º
VAS um ano após cirurgia (3º)	5.00		
VAS pós-operatório (2º)	6.00	0.074	2º = 3º
VAS um ano após cirurgia (3º)	5.00		

Nota: – As probabilidades de significância (valor p) referem-se ao teste de *Wilcoxon*
– Valor p em negrito indica diferença significativa.
– Os resultados significativos foram identificados com asteriscos, de acordo com o nível de significância das comparações múltiplas, a saber: valor p < 0.0167** .

Em análise comparativa da Escala Analógica Visual (VAS), verificou-se que a Escala Analógica Visual no pré-operatório é maior do que os demais momentos, pois o valor $p < 0,05$. Enquanto que a Escala visual no pós-operatório e no ano após a cirurgia são iguais, valor $p > 0,05$ (Tabela 4.13).

Para as comparações múltiplas, o nível de significância fica dividido por três ($\alpha/3$), por se tratar de comparações entre 3 grupos, ou seja, será considerado significativa aquela comparação cujo valor p for inferior a 0,0167.

Para casos em que existam *n* comparações o nível de significância fica dividido por *n* (α/n).

Referências

1. Siqueira AL. Teixeira FJS. Introdução a Estatística Médica. 2ed. Belo Horizonte: COOPMED, 2002.

2. Triola M.Introdução à Estatística. 10ed. Rio de Janeiro: LTC, 2008, p.722-801.

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