Capítulo 13: **Nível de Significância**[1]

Hartmut Günther e Maria Fabiana Damásio[2]

O processo de pesquisa científica parte de uma pergunta qualitativa e requer, em última análise, uma resposta qualitativa. Porém, como chegar a uma resposta significativa e relevante? Este livro apresenta procedimentos *quantitativos* para chegar a respostas consistentes ao mostrar várias estratégias para operacionalizar perguntas e transformar categorias qualitativas em dados quantitativos. Desta maneira, podemos chegar a *estimativas* quanto à significância e relevância dos resultados das nossas pesquisas.

Os capítulos anteriores mostraram como adotar diferentes procedimentos estatísticos nas pesquisas em psicologia e verificar os níveis de significância. Neste capítulo, consideramos as suposições comuns a estes procedimentos. Implícito nestas suposições está o reconhecimento de que decisões baseadas nos dados obtidos no contexto de uma pesquisa são sujeitas a erros (vide 18 do livro em elaboração). Assim sendo, ao invés de chegar a resultados "definitivos", o pesquisador se defronta com a eterna tarefa de estimar e minimizar eventuais erros nas suas conclusões. Para tanto abordaremos (a) a pergunta básica e comum às pesquisas; (b) o Erro Tipo I; (c) o Erro Tipo II e poder do teste; (d) o tamanho de efeito; (e) a as implicações para o processo de pesquisa.

A Pergunta Básica de Pesquisa

Independentemente do método e da técnica de pesquisa utilizados, a opção para realizar uma pesquisa por meio de registros sistemáticos pressupõe que *há ordem* subjacente aos fenômenos sob estudo, e ainda, que se trata de uma ordem que é acessível, direta ou indiretamente, por meio de instrumentos apropriados de pesquisa. Assim, para fins deste capítulo, definimos o objetivo do processo de pesquisa como o e descobrir e elucidar a ordem dos fenômenos, dentro de um campo de interesse específico[3].

Tendo coletado dados de maneira sistemática (independente da técnica específica usada) e organizado os mesmos (seja qualitativa ou quantitativamente), o

---

[1] Trabalho preparado para o livro *Metodologias quantitativas de pesquisa científica*, organizado por C. Faiad, L. Pasquali & M. C. Ferreira, a ser publicado pela Editora Vozes.

[2] O primeiro autor é Professor Titular no Departamento de Psicologia Social e do Trabalho da Universidade de Brasília, onde leciona Psicologia Social, Psicologia Ambiental, Planejamento de Pesquisa coordenao Laboratório de Psicologia Ambiental e realiza pesquisa sobre qualidade de vida urbana; a segunda autora concluiu o seu doutorado na área de avaliação psicológica no programa de Psicologia Social, do Trabalho das Organizações da Universidade de Brasília.

[3] Mencionamos apenas *em passant* que existem alternativas quanto ao processo de pesquisa. Entretanto, não pretendemos entrar numa discussão epistemológica sobre perspectivas empíricas, construtivistas e outras, já que este livro pretende ter um caráter eminentemente prático.

pesquisador se defronta com as seguintes interrogações diante das suas, necessariamente limitadas, observações obtidas num determinado contexto de pesquisa empírica: Até que ponto é possível estender os achados para além daquela situação de pesquisa concreta? Será que é possível generalizar para além deste contexto e, em caso afirmativo, para qual contexto além daquele no qual foram coletados os dados? Será que os achados representam uma realidade subjacente mais ampla, ou será que a constelação dos dados encontrados é um simples acaso? Mais formalmente, a pergunta virá a ser: *Qual a probabilidade de poder obter este conjunto de dados e seus resultados por acaso?* Esta pergunta baseia-se no seguinte raciocínio: caso a probabilidade de obter os dados por acaso seja "suficientemente pequena", podemos supor que estamos diante de um fenômeno sistemático e não derivado do acaso. Caso contrário, devemos supor que o resultado constitui algo obtido por acaso, portanto, não representando um fenômeno sistemático. Embora esta questão seja mais facilmente respondida com base em dados quantitativos, por envolver estimativas probabilísticas, o raciocínio apresentado a seguir se aplica a qualquer tipo de pesquisa empírica.

Formulação de Hipóteses

Para responder à pergunta quanto a probabilidade de poder obter um determinado resultado por acaso, partimos da afirmação de que *não existe ordem* no fenômeno sob estudo. Consequentemente, o objetivo da coleta e da análise dos dados é demonstrar que a afirmação "*não existe ordem*" é *falsa*. Em outras palavras, ao invés de "provar" a existência de algum fenômeno, tentamos demonstrar a *inexistência do oposto* àquilo que nos interessa.

O contraponto entre *provar* e *demonstrar* nesta frase é fundamental para entender porque a palavra, o conceito "provar" deve ser banido do discurso científico. Por meio da pesquisa, coletamos e analisamos evidências que apoiam as afirmações acerca dos temas que nos interessam, porém, somente baseadas em probabilidades e não em constatações contundentes.

Colocado de outra maneira, querendo verificar a existência de um efeito de pesquisa, que pode ser tanto uma diferença entre grupos quanto uma relação entre variáveis, tentamos demonstrar que o seu *oposto*, a ausência do efeito, não é verdadeiro. Querendo demonstrar o oposto nos leva à formulação de duas hipóteses formais:

H_0 – a *hipótese nula*, que afirma que não existe efeito

e

H_1 – a *hipótese alternativa*, que afirma que existe um efeito

Embora uma dada pesquisa parta do interesse em demonstrar a existência da hipótese alternativa, isto é, a existência de algum efeito, no processo de pesquisa parte-se da suposição de que não existe efeito, isto é, da hipótese nula, e tem a meta de rejeitá-la. Assim, é importante enfatizar que devemos falar somente em "conseguir" ou "não conseguir" rejeitar a hipótese nula; mas jamais em "provar" a existência da hipótese alternativa.

Efeito.

Ressaltamos que o termo *efeito* refere-se de maneira genérica a diferenças entre grupos e relações entre variáveis, uma vez que em qualquer delineamento de pesquisa tenta-se estabelecer uma relação, isto é, o efeito, de uma manipulação (por exemplo, em um experimento de laboratório) ou de algum acontecimento antecedente (por exemplo, em um experimento natural) ou de alguma variável concomitante (por exemplo, em uma pesquisa correlacional) com uma variável critério (ou, até, mais de uma variável). Para mais detalhamento de delineamentos experimentais e quase-experimentais, veja, por exemplo, Shadish, Cook & Campbell (2002). Em resumo, *efeito* refere-se, igualmente, à "diferença entre A e B" e à "relação entre X e Y".

Hipóteses *uni*direcional e *bi*direcional.

Ao comparar dois grupos, é necessário diferenciar entre hipóteses alternativas *uni*direcional e *bi*direcional. No caso da hipótese alternativa *bi*direcional – o exemplo deste capítulo apresentado logo a seguir – o contraste é entre a ausência de diferença (H_0) e qualquer diferença (H_1). Qualquer diferença quer dizer, neste contexto, tanto a média do grupo A é maior do que a média do grupo B (A > B), quanto a média do grupo B é maior do que a média do grupo A (B > A).

No caso de formular uma hipótese alternativa *uni*direcional, a mesma seria, média do grupo A é maior do que a média do grupo B (A > B), junto com uma hipótese nula de que a média do grupo A é menor ou igual a média do grupo B (A ≦ B).

O Exemplo usado neste Capítulo

Supondo que queremos verificar se a forma de pagar o ingresso no cinema influencia no atraso de chegar a sessão do filme. Definimos "forma de pagamento" em termos de dois grupos. O grupo A é constituído por pessoas que pagam em dinheiro, o grupo B por pessoas que pagam com cartão de crédito. "Chegar com atraso" será operacionalizado em termos do número de minutos após o começo do primeiro *trailer*. Neste exemplo, a nossa hipótese nula, H_0, é que não existe diferença na média de minutos, isto é, do tempo de atraso entre os dois grupos. A hipótese alternativa, H_1, afirma que existe uma diferença na média do tempo de atraso entre os dois grupos, sem, entretanto, especificar qual das duas médias é maior, isto é, qual dos dois grupos chega com maior atraso. Conforme explicitado acima, neste exemplo, a hipótese alternativa é de natureza *bi*direcional. Importante repetir e frisar que ao invés de "provar" uma diferença, visamos demostrar a *in*existência do *oposto*, isto é, a inexistência da ausência desta diferença.

Neste processo de pesquisa, contrastamos duas realidades: (1) como o fenômeno estudado de fato é na *ordem real* e (2) como ele aparece com base dos dados coletados, isto é em uma *ordem achada*. Partimos da hipótese nula – H_0 – de que não existe um efeito (no caso, uma diferença entre as médias de tempo dos grupos A e B). Este passo é possível uma vez que se trata de uma pergunta e de uma hipótese que pode ser *falsificada* uma vez que há a possibilidade de obter dados empíricos acerca da questão. Operacionalizamos os conceitos em termos de variáveis,

realizamos observações e coletamos dados apropriados para a questão. Por meio da análise estatística dos dados tentamos, então, *rejeitar* a hipótese nula. À medida que não é possível provar a inexistência de algo, há apenas o caminho da aproximação, isto é, verificar a possibilidade da sua inexistência. Daí a pergunta: *Qual a probabilidade de obter este conjunto de dados e seus resultados por acaso?* Caso a probabilidade de se obter estes dados por acaso seja "suficientemente pequena" rejeitamos a hipótese nula e supomos a existência do efeito de interesse; no caso, a diferença no tempo de atraso nos dois grupos de visitantes do cinema. Ao contrário, se a probabilidade de se obter os dados por acaso for "relativamente grande", afirmamos não ter sido possível rejeitar a hipótese nula e continuamos com a afirmação de que não existe o efeito.

Como será explicitado a seguir, em termos decisórios, podemos reformular a pergunta: *Qual a probabilidade de se obter este conjunto de dados e seus resultados por acaso?* A pergunta verá: *Qual o risco de se cometer um erro ao aceitar um resultado obtido por acaso?*

O que, entretanto, quer dizer probabilidade (ou risco) 'suficientemente pequeno' ou 'relativamente grande'? À medida que os dados e sua análise nos levam a concluir que existe uma certa probabilidade de se encontrar aqueles dados por acaso, admitimos o risco de chegarmos a uma conclusão errada, ou seja, que dados e análises futuros podem apontar resultados diferentes. Assim, a probabilidade "suficientemente pequena" versus "relativamente grande" é o tamanho da possibilidade de erro que admitimos cometer ao chegar à conclusão que algum fenômeno existe, quando, de fato, não existe.

As Quatro Relações entre a Ordem Real e a Ordem Achada

Na Tabela 13.1, resume-se as quatro possíveis situações, ao cruzar como o fenômeno estudado de fato é na *ordem real* e como ele aparece com base nos dados coletados, isto é na *ordem achada*.

(1) O fenômeno sob estudo de fato existe na ordem real.

Os dados obtidos em nossa pesquisa e os resultados deles derivados por meio de análise estatística apontam para a existência do efeito, uma vez que a probabilidade de obter estes resultados por acaso é "suficientemente pequena". Desta maneira, conseguimos rejeitar a hipótese nula e temos confiança que a hipótese alternativa seja correta, já que o risco de estarmos errado parece pequeno. Sendo que os dados correspondem a realidade, 'acertamos' a resposta correta. Uma outra maneira de falar desta situação é afirmar que o teste estatístico utilizado tem o poder para permitir que chegássemos ao resultado 'correto'. Este poder é definido como 1-β – vide item (4) logo a seguir, bem como a discussão na secção *Erro Tipo II e Poder de um Teste.*

(2) O fenômeno sob estudo de fato não existe na ordem real.

Os dados obtidos na pesquisa e os resultados deles derivados por meio da análise tampouco apontam para a existência do efeito, uma vez que a probabilidade de obter estes resultados por acaso é "relativamente grande". Por esta razão não conseguimos rejeitar a hipótese nula e preferimos não aceitar a hipótese alternativa, já

que o risco de estarmos errado parece grande. Uma vez que os dados correspondem à realidade 'acertamos' a resposta correta.

Tabela 13.1

		Na realidade, o fenômeno de interesse ...	
		… existe	… não existe
Os dados da pesquisa e os resultados daí derivados indicam que o fenômeno de interesse…	… existe	(1) Por meio da análise chegamos a rejeitar a H_0 e "acertamos" um resultado correto	(3) Por meio da análise chegamos a rejeitar a H_0 e cometemos o ***Erro Tipo I***, ao aceitar a existência de algo que não existe
	… não existe	(4) Por meio da análise não conseguimos rejeitar a H_0 e cometemos o ***Erro Tipo II*** ao aceitar a ausência de algo quando de fato existe	(2) Por meio da análise não conseguimos rejeitar a H_0 e "acertamos" um resultado correto
Observe: já que não temos acesso direto a ordem real, nossa conclusão, seja qual for, sempre será uma inferência acerca da ordem real.			

(3) O fenômeno sob estudo de fato não existe na ordem real.

Os dados obtidos em nossa pesquisa e os resultados deles derivados por meio da análise apontam para a existência do efeito, uma vez que a probabilidade de obter estes resultados por acaso é "suficientemente pequena". Assim, tendo a confiança em cometer um erro é pequeno, rejeitamos a hipótese nula. Porém, quando os dados obtidos e analisados ***não*** correspondem à realidade, cometemos o *Erro Tipo I* – rejeitamos a H_0 e confiamos na existência do efeito (H_1) quando de fato este não existe. Uma vez que os dados não correspondem a realidade, não 'acertamos' a resposta correta.

Erro Tipo I. Significa supor que existe um efeito onde de fato não existe. No nosso exemplo, o efeito refere-se a uma diferença entre as médias do tempo de atraso nos dois grupos de visitantes do cinema. Embora os dados da pesquisa apontem para a existência do efeito, o mesmo, de fato, não existe na realidade. A probabilidade do risco de cometer este Erro Tipo I é denominado α (alfa).

(4) O fenômeno sob estudo de fato existe na ordem real.

Os dados obtidos na pesquisa e os resultados deles derivados por meio da análise não apontam para a existência do efeito, uma vez que a probabilidade de obter estes resultados por acaso parece "relativamente grande". Assim, não tendo a confiança diante de um risco relativamente grande, não conseguimos rejeitar a hipótese nula. Sendo que os dados obtidos e analisados ***não*** correspondem à realidade, isto é, a existência de um efeito, cometemos o *Erro Tipo II* – não conseguimos rejeitar a H_0 e continuamos a não aceitar a existência de um efeito (H_1) quando este de fato existe. Por meio daquele conjunto de dados e sua análise não foi possível rejeitar a H_0 porque não espelhou de maneira adequada a realidade.

Erro Tipo II. Significa supor que não existe um efeito onde de fato existe. No nosso exemplo, o efeito refere-se a diferença nas médias do tempo de atraso nos dois grupos de visitantes do cinema. No caso, embora os dados da pesquisa não apontem para a existência do efeito, de fato existe na realidade. A probabilidade de cometer o Erro Tipo II é denominado β (beta).

Implicações

Salientamos que existe uma relação inversa entre o Erro Tipo I e o Erro Tipo II: Quanto mais exigente é o pesquisador na tentativa de minimizar o risco de cometer um Erro Tipo I, isto é, quanto menor estipula um valor α (alfa) que precisa ser alcançado para que se aceite um determinado resultado, maior a possibilidade de cometer o Erro Tipo II, isto é, de não conseguir detectar um determinado efeito quando, de fato, este efeito existe. Por outro lado, quanto mais cauteloso o pesquisador na tentativa de não cometer um Erro Tipo II, quanto menor o valor β (beta), maior a possibilidade de cometer um Erro Tipo I, isto é, supor um determinado efeito quando, de fato, não existe. Entretanto, esta relação não é linear, já que ambos os erros dependem de uma série de fatores, como explicitaremos mais adiante.

Como afirmamos anteriormente, até este ponto o raciocínio independe da técnica de pesquisa, seja ela qualitativa ou quantitativa, seja ela por meio de observação, de entrevista, de experimento, ou de análise de dados secundários. De qualquer maneira, resumimos os dados obtidos por meio de estatística *descritiva* e apresentamos os achados em termos de frequências; além de, caso seja apropriado, por meio de medidas centrais como médias, medianas e/ou modas, e de medidas de dispersão como percentis e/ou desvio padrão. Podemos organizar os dados em tabelas, gráficos, e/ou desenhos. Entretanto, o próximo passo, estimar até que ponto estamos lidando com algo sistemático, espelhando uma ordem real, ou algo que reflete uma constelação em base de dados obtidos por acaso, requer o uso de procedimentos da estatística *inferencial*. A seguir consideramos os passos necessários para estimar as probabilidades de cometer o Erro Tipo I e o Erro Tipo II.

Como verificado nos demais capítulos deste livro, o uso da estatística inferencial requer, inicialmente, a definição operacional de variáveis e o teste de hipóteses. Identificar as variáveis de uma pesquisa e indicar a maneira como as mesmas são analisadas constituem as bases para a escolha de modelos e

procedimentos estatísticos adequados. Desta maneira, podemos verificar até que ponto aquele conjunto de dados e as conclusões derivadas dele constituem um evento obtido por acaso ou representam uma relação sistemática.

O Erro Tipo I – Supor que Existe um Efeito onde de Fato não Existe tal Efeito

Retornamos à pergunta básica de pesquisa: *Qual a probabilidade de obter este conjunto de dados e seus resultados por mero acaso?* Caso esta probabilidade seja "suficientemente pequena", rejeitamos a H_0 e supomos que aqueles dados representam um fenômeno real. Caso contrário, se esta probabilidade for "relativamente grande", constatamos que não foi possível rejeitar a H_0, e supomos que aqueles dados não representam um fenômeno real, mas que foram obtidos por acaso.

Vale, entretanto, a pergunta: o que quer dizer "suficientemente pequeno" e "relativamente grande" quando falamos do risco de cometer o Erro Tipo I? Embora a literatura científica considere riscos menores de 5% ou de 1%, isto é $\alpha < 0,05$ ou $\alpha < 0,01$, em geral como "suficientemente pequenos", cabe ao pesquisador decidir se para sua pesquisa específica um determinado valor α (alfa) é, ou não é, pequeno ou grande, considerando, ainda, o Erro Tipo II e o poder do procedimento estatístico, e o tamanho de efeito. Para tanto, o pesquisador precisa levar em conta, antes de tudo, as implicações advindas da especificação de um determinado valor α (alfa).

Por Exemplo.

Erro Tipo I. Ao se testar a eficácia de um novo remédio, especifica-se um valor α muito exigente, para, com isto, evitar a possibilidade de se cometer um Erro Tipo I e declarar um novo remédio como eficiente, quando, no fundo, este não difere de um placebo. Por outro lado, corre-se o risco de não liberar um novo remédio eficaz, cometendo um Erro Tipo II, ao não detectar a diferença entre o novo remédio e o placebo.

Erro Tipo II. Após exame médico, o paciente pode ser declarado saudável (H_0) ou doente (H_1). Não conseguir rejeitar a hipótese nula, H_0 quando o paciente de fato está doente (Erro Tipo II) traz consequências mais sérias para o paciente, do que supor que ele está doente (e precisa de um remédio) quando de fato ele está saudável (Erro Tipo I) – supondo, obviamente, que o remédio indicado para tratar esta doença não adoece um paciente saudável.

Passos para o Teste de Hipótese

Diante desta situação, quais as opções para o pesquisador? Kvanli (1988) sugere os seguintes passos para a realização de um teste de hipótese:

(1) Formular as hipóteses nulas e alternativas. A hipótese nula (H_0) afirma que não existe efeito, a hipótese alternativa (H_1) afirma que existe efeito.

(2) Definir a estatística apropriada para os dados e a questão a ser respondida, conforme apresentado nos demais capítulos deste livro.

(3) Definir o valor crítico (V_{crit}) da estatística, que define se o valor calculado (V_{cal}) pelo procedimento estatístico indica que a hipótese nula pode ser rejeitada, ou não. Este V_{crit} depende do número de participantes da pesquisa e do nível α (alfa) escolhido pelo pesquisador.

(4) Calcular o V_{cal} pelo procedimento previsto na estatística escolhida.

(5) Comparar os valores V_{crit} e V_{cal}.

(6) Formular a conclusão em termos de
rejeitar a H_0, quando o V_{cal} é maior do que o V_{crit}
ou
não rejeitar a H_0 quando o V_{cal} é igual ou menor do que o V_{crit}.

(7) Formular a conclusão em termos do problema original, sem recorrer ao jargão estatístico, sumarizando apenas o resultado da análise.

Duas Estratégias para o Teste de Hipótese

Existem duas estratégias no processo decisório para a rejeição, ou não, da hipótese nula (H_0). O procedimento elencado acima supõe que se determina o nível α (alfa) e o V_{crit} ao mesmo associado *antes* de começar a pesquisa, a coleta de dados e a escolha do procedimento estatístico. Estabelece-se, de antemão, qual será a probabilidade "suficientemente pequena" do risco de cometer um Erro Tipo I. Pré-estabelecendo o nível α (alfa) e o V_{crit} – baseado no tamanho da amostra e na probabilidade que consideramos "suficientemente pequena" para conseguir rejeitar a H_0 – significa que não podemos, nem devemos, mudar de opinião, a depender dos resultados. Em outras palavras, não é aceitável mudar o nível de α e o V_{crit} associado a ele, caso o V_{cal} não permita a conclusão "desejada" de conseguir rejeitar a hipótese nula.

Uma segunda estratégia segue o procedimento acima nos passos 1, 2, 4 e 7 e difere, especialmente no passo 6. À medida que não se define um V_{crit} de antemão (passo 3), não haverá o passo 5. A decisão mencionada no passo 6 baseia-se não mais numa aplicação quase "mecânica" do procedimento chamado NHST (*Null hypothesis significance testing* ou teste de significância por meio de hipótese nula) [4] com um nível α predeterminado, mas parte de uma reflexão, quase que qualitativa, sobre as implicações de um determinado resultado. Ao final da pesquisa, se chega a um V_{cal} e uma probabilidade associada a este. Somente neste momento, será determinado se o V_{cal} e a probabilidade associada a ele podem ser considerados "suficientemente pequenos", isto é, se o consideramos "aceitável" para rejeitar a H_0 neste nível de significância α (alfa).

Seja qual for a estratégia do processo decisório tomada, o cálculo do V_{cal} da estatística relativa à pergunta da pesquisa em curso permite determinar a

[4] Para a discussão deste procedimento e suas implicações, vide, por exemplo, Kline (2004), Nickerson (2000), ou Rozeboom (1960[!]).

probabilidade associada a este valor e, desta maneira, responder à pergunta básica de pesquisa: *Qual a probabilidade de obter este conjunto de dados e seus resultados por acaso?*

Por exemplo.

No exemplo deste capítulo queremos saber se existe uma diferença entre as médias de tempo de atraso de entrar no cinema entre um grupo de pessoas que paga em dinheiro (grupo A) e as de um outro grupo que paga com cartão de crédito (grupo B). Por se tratar de uma simples comparação de médias de tempo de dois grupos, podemos usar um teste *t* (vide Capítulo 3 deste livro). Usando dados fictícios de 40 pessoas, sendo 20 em cada uma dos grupos, chegamos ao resultado de $V_{cal} = t_{df=39} = 2,62$. O que pode ser inferido deste resultado?

Nível de significância preestabelecida

Caso pré-estabeleçamos um nível de significância $\alpha = 0,01$, isto é, uma probabilidade "suficientemente pequena" de 1% ou, colocado de outra maneira, aceitarmos como maior risco 1% para cometer um Erro Tipo I, devemos comparar o V_{cal} obtido com o V_{crit} que pode ser encontrada numa tabela de significância dos valores de *t*. Constatamos que o V_{crit} da estatística *t*, para 40 sujeitos e um $\alpha = 0,01$ é 2,704. Seguindo os passos (5) e (6) do processo decisório de Kvanli elencado acima, há de se concluir que não conseguimos rejeitar a hipótese nula H_0, uma vez que o V_{cal} é menor do que o V_{crit}.

Observamos acima que um valor α (alfa) e seu V_{crit} pré-estabelecidos não devem ser modificados *ex post fato*, isto é, dependendo do resultado. Desta maneira, uma vez chegado a este final da pesquisa, não há mais o que fazer: não foi possível rejeitar a H_0, portanto não há razão para se supor que existe uma diferença entre os dois grupos.

Reflexões adicionais ao preestabelecer o nível de significância.

Entretanto, para fins deste capítulo, este resultado proporciona algumas reflexões, diante das demais informações na mesma tabela de significância.

Importância do número de participantes na pesquisa.

Se tivéssemos feito a pesquisa com 120 ao invés de 40 pessoas, o V_{crit} para $\alpha = 0,01$ é 2,17. Se esta pesquisa tivesse tido o mesmo resultado V_{cal} de $t = 2,62$, a comparação entre o V_{cal} e o V_{crit} nos levaria a concluir que existe uma diferença estatisticamente significativa nesta amostra maior de participantes. Na secção sobre o tamanho de efeito e poder decisório voltaremos a mencionar as implicações do tamanho de uma determinada amostra.

Importância do nível de significância estabelecido.

Se tivéssemos pré-estabelecido um valor de $\alpha = 0,05$ ao invés de $\alpha = 0,01$, o V_{crit} da estatística *t* para uma pesquisa com 40 sujeitos seria 2,021. Desta maneira, a

comparação entre o V_{cal} = 2,62 e o V_{crit} = 2,021 nos levaria a conseguir rejeitar a H_0 e a supor que existe uma diferença estatisticamente significativa entre os dois grupos. Isto leva à indagação se estamos sendo rigorosos demais ao não conseguir rejeitar a H_0 em nível de α= 0,01. Ao tentar evitar o Erro Tipo I, estaríamos cometendo o Erro Tipo II? Vide a seguir, na secção sobre o Erro Tipo II.

Importância do tipo de hipótese alternativa.

Há de se comentar, ainda, sobre a diferença entre hipóteses alternativas *uni*- e *bi*direcional. Neste exemplo, a questão foi simplesmente se "há uma diferença" entre os dois grupos, sem especificar, se o grupo A (pagar com dinheiro) esteja mais atrasado do que o grupo B (pagar com cartão de crédito), ou se o grupo B está mais atrasado do que o grupo A. Em outras palavras, a hipótese alternativa, neste caso, foi *bi*direcional.

Vamos supor que a partir de pesquisas anteriores temos *informação adicional* sobre o comportamento de pessoas que pagam com dinheiro ou com cartão de crédito. Estes estudos sugerem que pessoas do grupo B (pagar com cartão) são menos atrasados do que do grupo A, razão pela qual formulamos uma hipótese alternativa *uni*direcional: a média do tempo de atraso do grupo B é menor do que a média de atraso do grupo A. Neste caso, a hipótese nula é que não há diferença entre os grupos, ou, ainda, que a média do grupo A é menor do que a média do grupo B.

Portanto, o V_{crit} associado a um nível de α = 0,01 *uni*direcional com 40 participantes, é 2,42. Sendo o $V_{cal} = t_{df=39}$ = 2,62 maior do que o V_{crit}, a decisão a ser tomada é que existe uma diferença estatisticamente significativa entre os dois grupos no sentido de que, em média, pessoas que pagam com cartão se atrasam menos do que pessoas que pagam com dinheiro.

Nível de significância não estabelecido previamente

Caso não tivéssemos pré-estabelecido um nível de significância, o programa de computação utilizado teria nos fornecido a probabilidade exata de p = 4,2 % ou α = 0,042 associada a um V_{crit} de t = 2,62 baseada em uma amostra de 40 sujeitos. Qual a implicação deste resultado? Obviamente saberíamos que este resultado seria estatisticamente significativo no nível tradicional de α = 0,05 e não seria significativo no nível tradicional de α = 0,01. Entretanto, sem os tradicionais níveis de significância, cabe ao pesquisador apresentar os parâmetros, no caso, o número de sujeitos N, a estatística t, o nível de significância p e, ainda, o tamanho de efeito. Desta maneira pode se chegar a um julgamento se a probabilidade associada ao V_{cal}, é "suficiente pequena" ou "relativamente grande" para tomar uma determinada decisão a respeito do resultado encontrado. Vide a última secção sobre implicações.

O Erro Tipo II e o Poder de um Teste

O Erro Tipo II ocorre quando o pesquisador, a partir das evidências empíricas da sua pesquisa – seus dados e suas análises – não consegue rejeitar a hipótese nula, quando, na verdade, deveria rejeitá-la porque, de fato, existe um efeito. Como

indicado acima, o Erro Tipo II é denominado β (beta). O seu complemento – (1 – β) – é o *poder* do teste, isto é, o poder de não cometer o Erro Tipo II e ser capaz de rejeitar a hipótese nula, quando ela é falsa.

Qual então a relação entre o Erro Tipo I – aceitar a existência de um efeito quando não existe, e o poder de um testes (1 – β) – conseguir detectar um efeito quando este existe? Observamos acima que existe uma relação inversa entre o Erro Tipo I e o Erro Tipo II – quanto maior um, menor tende a ser o outro.

Quanto ao poder de um procedimento estatístico (1 – β), este depende de uma série de fatores:

(1) tamanho da amostra: quanto maior o número de participantes em uma determinada pesquisa, maior o poder de rejeição da hipótese nula, quando ela é falsa.

(2) nível de significância: quanto mais leniente o nível de significância α (alfa), menor o β (beta), portanto mais poderoso o teste;

(3) tipo de hipótese alternativa: um teste *uni*direcional é mais poderoso do que um teste *bi*direcional;

(4) tipo de teste utilizado: testes paramétricos são mais poderosos do que testes não-paramétricos, inclusive por exigirem, em geral ao mínimo, uma amostra maior de participantes de pesquisa. Além do mais, entre testes não-paramétricos existem diferenças de poder (vide Siegel, 1957/1977);

(5) qualidade dos dados: qualquer dado registrado em uma pesquisa é composto de dois elementos: (1) um componente é o 'verdadeiro' valor, que espelha de maneira correta a realidade que está sendo mensurada e (2) um erro de mensuração. Erros de mensuração podem ter várias origens, dependendo do tipo de pesquisa realizada. De maneira geral, estes podem ser resumidos como sendo erros de fidedignidade e/ou de validade no delineamento da pesquisa (Shadish, Cook & Campbell, 2002). Quanto menor a parte de erro em uma mensuração, maior a componente 'verdadeiro' da medição e, assim, melhor os dados refletem a realidade que está sendo estudada. Consequentemente, mais poderoso será o teste e seus resultados;

(6) *tamanho de efeito*: um elemento fora de controle direto do pesquisador, vide secção a seguir.

No exemplo apresentado acima, na secção referente ao Erro Tipo I, demostramos alguns efeitos destes antecedentes do poder de um teste. Um mesmo V_{cal} baseado em uma amostra maior permite a decisão de rejeitar a hipótese nula [item (1) acima]. Um mesmo V_{cal} pode ser estatisticamente significativo em nível de $\alpha = 0,05$, isto é, um nível mais leniente, mas não em nível de $\alpha = 0,01$, isto é, um nível mais exigente [item (2) acima]. Um mesmo V_{cal} pode ser significativo quando usado para testar uma hipótese *uni*direicional, mas não quando usado para testar uma hipótese *bi*direcional [item (3) acima].

Quanto ao item (4), observamos anteriormente que, para realizar um teste *uni*direcional, o pesquisador se baseia em informação adicional e que, desta maneira, contribui para obter um poder maior deste tipo de teste. De maneira correspondente,

há mais informação embutidas na mensuração requerida para o uso de testes paramétricos, por se tratar de dados em nível intervalar ou de razão. Já testes não-paramétricos exigem menos informação, por utilizar dados em nível nominal ou ordinal (vide, por exemplo, Siegel, 1957/1977).

Igualmente, quanto ao item (5), a questão da informação embutida na mensuração constitui a base do maior poder, isto é, quanto maior a qualidade de dados, maior o poder do teste que se baseia nestes dados. Conforme colocado anteriormente, qualidade de dados significa, justamente, pouco erro de mensuração.

Cohen (1988) sugere que as consequências de cometer um Erro Tipo I são quatro vezes maiores do que as de cometer um Erro Tipo II, o que implica em uma razão de $\alpha = 0,05 / \beta = 0,20$. Colocado de outra maneira, deve-se assegurar um poder de 80% em um procedimento estatístico (vide, também, Lipsey & Hurley, 2008).

Embora programas de estatística como SPSS incluam a informação sobre o poder de um determinado procedimento, vale apontar que um programa a parte G*Power está disponível, gratuitamente, pelo site

http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/ de

Descrições e detalhamentos sobre o seu uso podem ser encontrados em Faul, Erdfelder, Buchner e Lang (2007, 2009).

O Tamanho de Efeito e o Poder Decisório

Como primeiro ponto da discussão anterior sobre o poder do teste, afirmamos que, quanto maior a amostra, maior o poder do teste. Porém, há um teto para esta relação quando o tamanho da amostra é igual ao número total da população. Isto provoca a seguinte questão: Existe uma amostra de tamanho "suficientemente grande" para assegurar que se evite o Erro Tipo I dentro de determinadas margens de erro? Em caso afirmativo, é possível assegurar ao mesmo tempo um poder do procedimento estatístico, expressado por $1 - \beta$ (Erro Tipo II)? (cf. Serlin & Lapsley, 1985). Inversamente, podemos falar num tamanho de efeito de um dado resultado com uma dada amostra? (cf. Kline, 2004). Isto nos remete ao início deste capítulo, onde observamos

> O processo de pesquisa científica parte de uma pergunta qualitativa e requer, em última análise, uma resposta qualitativa. Porém, como chegar a uma resposta significativa e relevante?

Ao apresentar o conceito do tamanho de efeito, acrescentamos a esta observação inicial a pergunta, "o que é, no frigir dos ovos, uma diferença ou relação significativa e relevante?" No caso do nosso exemplo sobre a diferença na média do tempo de atraso em função de pagar com dinheiro ou cartão, antes de tudo, a questão não seria apenas se uma diferença encontrada é estatisticamente significativa, mas se é suficientemente grande para merecer alguma mudança por parte do cinema quanto à política do recebimento do pagamento. Um tamanho de efeito de cinco minutos, mesmo se for estatisticamente significativo, faria uma diferença para a qualidade da experiência no cinema? Porém, se o nosso exemplo tivesse sido 'média de tempo para

atender a uma chamada de emergência', sendo o Grupo A composto por ambulâncias sempre com o motor ligado *versus* um Grupo B, composto por ambulâncias que somente liguem seus motores quando há uma chamada, uma diferença de cinco minutos de atraso pode fazer uma diferença 'significativa e relevante' no atendimento dos que necessitam de assistência emergencial.

O que, então, é um efeito substancial? Kline (2004) lista cinco determinantes para a estimativa do tamanho de efeito: (1) a variável critério ser arbitrária versus significativa, quer dizer, o quanto a variável critério é válida; (2) a consistência da medida da variável critério oriunda de diferentes pesquisas, problema este solucionável por meio de transformação das medidas em escore *z*; (3) determinação, *a priori*, do tamanho da amostra e, portanto, do poder da análise estatística; (4) informação advinda de estudos anteriores por meio de meta-análise; (5) uso isolado do teste de estatística sem considerar a análise de poder do mesmo.

Kline observa (p. 134) que a determinante (3) é a mais complicada, uma vez que depende da especificidade teórica, prática ou clínica da pesquisa, em outras palavras, do conhecimento específico da área, por parte do pesquisador. Apesar desta especificidade, existem sugestões na literatura científica sobre o que pode ser considerado um efeito pequeno, médio ou grande e como calcular tais efeitos dependendo do tipo de estatística usado e do tamanho da amostra (cf. Cohen, 1988; Field, 2009; Kline, 2004). Quanto à relação entre o poder de um teste estatístico e o tamanho de efeito, deve ficar claro que quanto maior for estipulado um tamanho de efeito, maior o poder do teste (isto é, conseguir demonstrar a diferença, quando a mesma existe), mantendo os demais parâmetros iguais. Aplicado ao exemplo deste capítulo, supondo uma diferença nas médias dos dois grupos de 20 minutos ou invés de cinco minutos, maior o poder do teste que demonstra a diferença.

Embora programas de estatística como SPSS incluam a informação sobre o tamanho de efeito de um determinado procedimento, vale apontar que existem referências a programas e procedimentos sobre como calcular o tamanho de efeito de dados alheios, como, por exemplo, o *Practical Meta-analysis Effect Size Calculator* de David Wilson, disponível na internet no site

http://www.campbellcollaboration.org/resources/effect_size_input.phpum

Um documento didático explicando como realizar cálculos simples do tamanho de efeito é o de Thalheimer & Cook (2002), igualmente disponível na internet.

Implicações para o Processo de Pesquisa

Podemos oferecer o seguinte resumo quanto ao teste de significância de um procedimento estatístico:

(1) Existe uma relação recíproca entre o Erro Tipo I (supor efeito estatístico, quando este não existe) e o Erro Tipo II (não conseguir detectar um efeito, quando este, de fato, existe).

(2) Tanto o nível de significância alfa (α), isto é, o risco de cometer um Erro Tipo I, quanto o poder de um teste 1-beta (1-β), isto é, não correr o risco de não conseguir detectar um efeito quando ele existe, são estabelecidos por convenção.

(3) Como estes valores representam convenções, há de se observar que o procedimento tradicional, especialmente o passo 3 do processo de teste de hipótese relatado à página 8 (deste manuscrito), seguindo a proposta de Kvanli (1988), constitui um guia inicial importante, mesmo considerando a discussão acerca de NHST aludido às páginas 13/14 (deste manuscrito).

(4) Porém, cabe ao pesquisador justificar tanto a escolha do V_{crit} quanto da aceitação de um dados resultado V_{cal}, levando em conta o poder do procedimento estatístico utilizado bem como do tamanho de efeito – e sua relevância – alcançado em termos das consequência desta escolha, conforme alertado às páginas 9/10 e 21 (deste manuscrito).

Referências

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Erlbaum.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Field, A. (2009). Descobrindo a estatística usando o SPSS. Porto Alegre, RS: ArtMed.

Kline, R. B. (2004). *Beyond signifiance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychologial Association.

Kvanli, A. H. (1988). *Statistics: A computer integrated approach.* St. Paul, MN: West Publishing Comp.

Lipsey, M. W., & Hurley, S. M. (2008). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog (Eds.), *The Sage handbook of applied social research methods* (pp. 44-76). Thousand Oaks, CA: Sage.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*, 73-85.

Shadish, W. R., Cook, Th. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Belmont, CA: Wadsworth.

Siegel, S. (1977). *Estatística não-paramétrica para as ciências do comportamento.* São Paulo: McGraw-Hill. (originalmente publicado em 1957).

Thalheimer, W., & Cook, S. (2002, August). How to calculate effect sizes from published research articles: A simplified methodology. Retirado em 9 de setembro de 2012 de http://education.gsu.edu/coshima/EPRS8530/Effect_Sizes_pdf4.pdf