UNIVERSIDADE DE BRASÍLIA

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO - FACE

DEPARTAMENTO DE ECONOMIA

INFERÊNCIA ESTATÍSTICA E A PRÁTICA ECONÔMICA NO BRASIL: OS (AB)USOS DOS TESTES DE SIGNIFICÂNCIA

CARLOS LEONARDO KULNIG CINELLI

BRASÍLIA

JUNHO DE 2012

CARLOS LEONARDO KULNIG CINELLI

INFERÊNCIA ESTATÍSTICA E A PRÁTICA ECONÔMICA NO BRASIL

OS (AB)USOS DOS TESTES DE SIGNIFICÂNCIA

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade de Brasília como requisito parcial à obtenção do grau de mestre em economia.

Orientador: Professor Dr. Bernardo P. M. Mueller

BRASÍLIA

JUNHO DE 2012

CARLOS LEONARDO KULNIG CINELLI

INFERÊNCIA ESTATÍSTICA E A PRÁTICA ECONÔMICA NO BRASIL: OS (AB)USOS DOS TESTES DE SIGNIFICÂNCIA

Brasília, _____ de _____ de _____.

BANCA EXAMINADORA

_____

Prof. Dr. Bernardo Pinheiro Machado Mueller – Orientador
Universidade de Brasília

_____

Prof. Dr. Donald Matthew Pianto
Universidade de Brasília

_____

Prof. Dr. Leonardo Monteiro Monastério
IPEA – Instituto de Pesquisa Econômica Aplicada

AGRADECIMENTOS

Agradeço ao Bernardo Mueller pela orientação acadêmica.

Agradeço a Stephen Ziliak, Deirdre McCloskey, Aris Spanos, Deborah Mayo e Walter Kramer pelas informações prestadas e dúvidas esclarecidas. Obviamente que qualquer interpretação errônea que tenha remanescido é de minha exclusiva responsabilidade.

Agradeço à minha família, à minha namorada, e aos meus amigos pelo apoio e paciência.

RESUMO

Esta dissertação trata da confusão entre significância estatística e significância econômica nos trabalhos econométricos aplicados. O capítulo teórico resgata alguns tópicos pertinentes ao entendimento da confusão entre significância estatística e significância econômica, expondo as principais diferenças entre os métodos de Fisher, Neyman-Pearson e Bayesianos para testes de hipótese. Além disso, discute-se a ideia do *p-valor* como medida de evidência e trabalham-se, por fim, as noções de erro real e erro amostral, bem como a distinção entre diferença estatística e diferença substantiva. O capítulo empírico resgata a literatura acerca do tema especificamente para a área da economia, com as evidências verificadas em outros países, como para os Estados Unidos – McCloskey e Ziliak (1996), Ziliak e McCloskey (2004a, 2008a) – ou a Alemanha – Kramer (2011): 70 a 79% dos artigos da *American Economic Review* nos anos 80 e 90, respectivamente, bem como entre 56 a 85% dos artigos da *German Economic Review* confundiram significância estatística com significância econômica. Em seguida, quantificamos o problema no Brasil, tomando como amostra todos os 94 artigos publicados na Revista Brasileira de Economia entre 2008 a 2011, dos quais 67 que utilizaram testes de significância foram detidamente analisados. Como principais resultados temos que: 64% dos artigos confundiram significância estatística com significância econômica; mais de 80% dos artigos ignoraram o poder dos testes utilizados; 97% dos artigos não discutiram o nível de significância adotado; 74% não demonstraram preocupação com a especificação ou adequação estatística do modelo; 40% não apresentaram estatísticas descritivas; mais da metade não discutiu o tamanho de seus coeficientes ou a conversa científica em torno da grandeza do parâmetro, entre outros números.

Palavras chave: teste de hipótese nula; significância estatística; significância prática; o culto da significância estatística; inferência estatística; revista brasileira de economia.

ABSTRACT

This dissertation deals with the confusion between statistical significance and economic significance in applied econometrics. The theoretical chapter brings some topics necessary to the understanding of the confusion between statistical and economic significance, outlining the main differences between Fisherian, Classical and Bayesian methods. In addition, we discuss the interpretation of the *p-value* as a measure of evidence and the notion of real error *versus* sampling error as well as the distinction between statistical and substantive difference. The empirical chapter discusses the literature about the subject specifically in economics. We show the evidence found in other countries like the United States - McCloskey and Ziliak (1996), Ziliak and McCloskey (2004a, 2008a) - and Germany - Kramer (2011): 70 and 79% of the papers published in the American Economic Review, in the 80's and the 90's, respectively, and between 56 to 85% of the papers published in the German economic Review conflate statistical and economic significance. We, then, quantify the problem in Brazil, taking a sample of all 94 papers published in Revista Brasileira de Economia, between 2008 and 2011, and carefully analyzing all 67 that used significance tests. Among other numbers, the main results are: 64% of them confused statistical significance with economic significance; more than 80% ignored the power of the tests; 97% did not discuss the significance level; 74% showed no concern about specification or statistical adequacy; 40% did not present descriptive statistics; more than half did not discuss the size of the coefficients; also more than half did not discuss the scientific conversation within which a coefficient would be judged large or small.

Key-words: null hypothesis significance testing; statistical significance; practical significance; the cult of statistical significance; statistical inference; revista brasileira de economia.

LISTA DE TABELAS

LISTA DE FIGURAS

Sumário

1) INTRODUÇÃO E JUSTIFICATIVA

Em março de 2011, a Suprema Corte dos Estados Unidos deliberou sobre assunto de interesse aos cientistas empíricos: o uso dos testes de significância estatística. O caso foi trazido por investidores da empresa *Matrixx Initiatives Inc*, fabricante do remédio para gripe *Zicam*. A acusação recaía sobre o fato de a empresa ter omitido que algumas pessoas, usuárias do remédio, sofreram de *anosmia*[1]. Quando a notícia veio a público, houve reação no mercado acionário, resultando em perda financeira por parte dos investidores. Entretanto, como o número de casos de *anosmia* nos indivíduos que tomaram o *Zicam* havia sido considerado estatisticamente insignificante, aos níveis "usuais" de significância estatística, a empresa alegou não existir obrigação de informar aos investidores o fato – pois este, estatisticamente, sequer existiria (SUPREME COURT OF THE UNITED STATES, 2011).

A linha de raciocínio da *Matrixx* havia sido acatada pela corte distrital, que opinou ***a favor*** do requerimento da significância estatística para discutir a materialidade da omissão da empresa. Tal opinião, provavelmente, não seria incomum na prática científica e inclusive poderia encontrar respaldo se remontássemos a um dos fundadores dos testes de significância, como Ronald Fisher. O estatístico afirmou ser

> [...] usual e conveniente para os pesquisadores tomarem 5% como um nível de significância padrão, no sentido de estarem preparados para ignorar todos os resultados que falhem em alcançar este padrão e, por este meio, eliminar de uma discussão mais aprofundada grande parte das flutuações que a sorte possa ter introduzido em seus resultados experimentais (FISHER, 1971, p.13)[2].

Contudo, a corte federal da região revisou a decisão da corte distrital, afirmando que a informação ligando o medicamento *Zicam* à *anosmia* seria relevante para os investidores, e que no presente caso a

> […] regra da *Matrixx* – de que os relatórios de eventos adversos em relação a produtos de uma empresa farmacêutica não apresentam materialidade na ausência de um número suficiente para estabelecer um risco estatisticamente significativo de que o produto está causando os eventos – estaria "artificialmente excluindo" informação que "de outra forma seria considerada significativa para a decisão de um investidor razoável" (SUPREME COURT OF THE UNITED STATES, 2011a, p. 2).

[1] Perda da capacidade olfativa.

[2] Todas as citações em língua estrangeira foram traduzidas pelo autor desta dissertação para o português.

Sobre esta posição que a Suprema Corte dos Estados Unidos tinha de ponderar – podendo afirmá-la ou negá-la. Para tanto, contou também com o depoimento *Amici Curiae*[3] de Deirdre McCloskey e Stephen Ziliak (2010), que corroborou o entendimento da corte federal. Os autores procuraram esclarecer à Suprema Corte que

> [...] ao se realizar um teste de significância estatística, o pesquisador deve cotejar os custos de aceitar hipóteses falsas com os custos de rejeitar hipóteses verdadeiras. Para reduzir as chances do último erro (erro tipo I), o pesquisador pode diminuir seu padrão de significância estatística, entretanto, isso resultaria em aumentar as chances do primeiro erro (erro tipo II). O balanço deve ser feito pelos pesquisadores em cada caso (MCCLOSKEY, ZILIAK, 2010, p. 04).

Ziliak e McCloskey ressaltaram que uma falha em se rejeitar uma hipótese nula previamente estabelecida não implica necessariamente em se aceitar que esta seja verdadeira. Além disso, lembraram o fato de que se o modelo estatístico está especificado corretamente, a melhor estimativa para o parâmetro, fornecida pelos dados, é aquela derivada do procedimento de estimação – e não a hipótese nula. Assim, assumir arbitrariamente que um efeito é zero ***apenas*** por não se rejeitar esta hipótese em particular – dentre várias outras que também não seriam rejeitadas caso testadas – careceria de fundamentos científicos. No caso específico de relatórios de eventos adversos, seria de interesse de uma das partes – a farmacêutica – ***não rejeitar*** a hipótese nula de que o medicamento causasse problemas. Dessa forma, a aparente objetividade deste teste de hipótese, com base em significância estatística aos níveis usuais (como 5% ou 1%), passaria a ser uma arbitrariedade nas mãos de quem tem interesse no resultado (MCCLOSKEY, ZILIAK, 2010).

O posicionamento da Suprema Corte foi coerente com a visão dos economistas e da corte federal. Qualquer método que adotasse um único fato ou ocorrência como critério para definir materialidade seria demasiadamente falho. Dados estatisticamente significantes não estariam sempre disponíveis e, desta forma, dever-se-iam considerar múltiplos fatores para se tomar uma decisão. Assim, conclui a Suprema Corte que

> [...] os consumidores provavelmente teriam visto o risco associado com o *Zicam* (possibilidade de perda de capacidade olfativa) como substancialmente maior do que o benefício de usar o produto (aliviar os sintomas da gripe), particularmente quando existem muitos medicamentos alternativos no mercado. Importante ressaltar que o remédio para gripe *Zicam* alegadamente respondia por 70% das vendas da Matrixx. Vendo os argumentos da acusação

[3] *Amicus Curiae* é um termo em latim que significa "amigo da corte". Refere-se à pessoa que não tem relação ou interesse com as partes envolvidas do caso, mas que tem interesse maior na questão jurídica sendo discutida e pode contribuir com esclarecimentos que auxiliem a corte no julgamento.

> como um todo, são fatos que sugerem risco significativo à viabilidade comercial do principal produto da Matrixx. É bastante provável que um investidor razoável veria esta informação "como algo que teria alterado significativamente o "mix total" de informação disponível" (SUPREME COURT OF THE UNITED STATES, 2010, p.18-19).

Interessante frisar que o raciocínio exposto é análogo à utilização de uma ***função perda*** para a ***decisão sobre a relevância*** do resultado encontrado, em oposição a uma regra binária a partir de um nível de significância estatística arbitrário. O julgamento anterior foi, portanto, afirmado[4].

Este caso é particularmente interessante, pois, a decisão da corte distrital, posteriormente corrigida nas instâncias superiores, ilustra o ***reflexo prático*** de algo bastante difundido nas ciências sociais – a confusão entre significância estatística e significância científica, política, econômica ou substantiva, bem como a falta de um exercício de inferência (ou decisão) mais cuidadoso. Mais especificamente no caso da ciência econômica, os mesmos autores do depoimento *Amici Curiae* constataram, repetidamente, que tal confusão ocorre em um dos melhores periódicos de economia dos Estados Unidos – o *American Economic Review (AER)*. McCloskey (1985) coletou uma amostra de 10 dos 50 artigos publicados no período de 1981 a 1983 que utilizaram análise de regressão e 70% da amostra "[...] deixou a significância estatística fazer o trabalho da significância substantiva" (MCCLOSKEY, 1985, p. 204).

Na década seguinte, McCloskey e Ziliak (1996) ampliaram a análise e examinaram todos os 182 artigos da *AER* que utilizaram análise de regressão publicados na década de 80. Os artigos tinham de responder a 19 questões sobre o uso da significância estatística e, em linha com o estudo anterior, cerca de 70% não fizeram distinção entre significância estatística e significância econômica, política ou científica – dentre outros erros. Posteriormente, a fim de verificar se esta constatação se mantinha (pois, segundo os autores, muitos alegavam que a prática havia melhorado ao longo da década de 90) Ziliak e McCloskey (2004a, 2008a) coletaram e analisaram 184 artigos publicados na *AER* de 1990 a 1999[5]. O resultado foi de encontro ao suposto avanço: desta vez, 79% das publicações confundiram significância estatística com significância econômica;

[4] Poder-se-ia, também, ter calculado a probabilidade de o remédio ter causado a ocorrência, como proposto emr Judea Pearl (2010).

[5] Na versão publicada em 2004, foram analisados 134 artigos; o livro (2008a) corrigiu a omissão de 50 artigos e analisou todos 184. Mais sobre isso será discutido no capítulo 3.

além disso, 78% consideraram que uma mera análise do sinal do coeficiente, sem se preocupar com o tamanho do efeito, era suficiente para a ciência[6].

Este fato não é monopólio da ciência econômica e parece ser bastante disseminado nas ciências sociais. Segundo Sterling (1959), aproximadamente 80% dos artigos dos principais periódicos de psicologia utilizavam apenas testes de significância estatística *fisherianos* para tomar alguma decisão. Já na área de recursos humanos, Callahan e Reio (2006) reportam que menos de 6% dos artigos interpretaram o tamanho do efeito de suas estimações; nos estudos de terapia lúdica, segundo Armstrong e Henson (2004), este número foi de apenas 5%. Nos estudos de gerenciamento, Seth et alii (2009) verificaram que 90% dos artigos do *Strategic Management Review* não fizeram distinção entre significância econômica e significância estatística.

Se, na citação anterior de Fisher, fizemos acreditar que estes resultados não seriam vistos com maus olhos, mais para o fim de sua vida o estatístico provavelmente retiraria suas palavras. Segundo Gigerenzer (2004, p.03), mesmo Fisher se oporia aos testes cegos de significância estatística[7], por pelo menos três motivos:

> [...] primeiro, porque a hipótese nula não se refere a uma diferença média nula ou a uma correlação zero, mas a uma hipótese a ser testada [...] segundo, porque [...] Fisher pensava que utilizar uma rotina de 5% de significância [em qualquer situação] indicava uma falta de sofisticação estatística [...] terceiro [porque], para Fisher, o teste da hipótese nula era o tipo de análise estatística mais primitivo e deveria ser utilizado somente para os problemas nos quais não temos qualquer ou muito pouco conhecimento.

A despeito disso, segundo os *surveys* realizados em diversas ciências, a prática continua (ZILIAK, MCCLOSKEY, 2008a). Uma das explicações para esta continuidade é um provável viés de publicação; pois, conforme Ziliak e McCloskey (2004a, p. 530), obter resultados estatisticamente significativos talvez seja, atualmente, a forma mais fácil de se publicar. Sterling (1959) encontrou que ***apenas 3%*** dos artigos de importantes periódicos da psicologia ***falhavam*** em rejeitar a hipótese nula. Conforme Ziliak e

[6] Cabe mencionar, entretanto, que houve melhoria em alguns quesitos, tais como: (i) na interpretação dos significados dos coeficientes estimados; (ii) na exibição das estatísticas descritivas das variáveis; (iii) na utilização do poder do teste, entre outros. Mas os autores fazem a ressalva de que, apesar de ter havido melhoria nestes pontos, os números ainda são preocupantes (ZILIAK, MCCLOSKEY, 2004a, p.533). Estes resultados serão apresentados no capítulo 3.

[7] Ziliak e McCloskey (2008a, p.59) reconhecem que Fisher "[...] próximo ao fim de sua vida [...] afastou-se de sua regra [da convenção de 5%], voltando-se ao procedimento defendido há muito tempo por Karl Pearson – os pesquisadores deveriam fixar seus próprios níveis de significância". Entretanto, esta flexibilização de Fisher é interpretada como uma estratégia frente à perda de espaço que o estatístico teve na comunidade acadêmica, mais ao fim da vida, após as contribuições de Neyman, Egon Pearson, Wald, Deming, Shewhart e Savage (ZILIAK, MCCLOSKEY, 2008a, p. 232-233).

McCloskey (2009, p. 2311-2312), um editor da área chegou a afirmar explicitamente que era improvável um artigo ser publicado a não ser que trouxesse resultados estatisticamente significantes: “significância ao nível de 5% era “mal, mas aceitável”; significância a 1% ou um nível “melhor” era considerado “altamente aceitável” e definitivamente digno de publicação”. Este viés gera um problema para a interpretação efetiva dos testes publicados, principalmente levando-se em consideração o poder dos testes em ciências sociais.

Em geral, grande parte dos estudos tem pouco poder estatístico. Mazen, Graf, Kellogg e Hemmasi (1987) sugerem que somente 6% a 9% dos estudos na área de administração tem poder suficiente para detectar efeitos menores; na área de negócios internacionais, Brock (2003) sugere que este valor é de cerca de 4% a 10%; em contabilidade ou na pesquisa em educação esse números também preocupam (LINDSAY, 1993; DALY E HEXAMER, 1983); o fato parece ser a regra nas ciências sociais em geral (ELLIS, 2010; ZILIAK, MCCLOSKEY, 2008a). Tendo isso em vista, conforme colocou Kruskal (1968), a tendência é serem observados muito mais estudos que apresentem resultados estatisticamente significativos do que esperado – informação esta que deveria modificar a interpretação dos testes de hipóteses efetivamente publicados, com maior probabilidade de *erro tipo I* do que a probabilidade nominal utilizada.

Até o presente momento, parece não haver estudo sobre o uso da inferência estatística – mais especificamente o uso do teste de significância estatística – e a confusão entre significância estatística e significância econômica na literatura brasileira. Dada a pertinência do tema e a lacuna identificada, justifica-se a realização deste trabalho, que buscará verificar em que grau os problemas apontados se apresentam nos artigos empíricos publicados no Brasil, conforme questionamentos propostos por McCloskey e Ziliak (1996). Para tanto, serão analisados os artigos acadêmicos de um dos principais periódicos de economia do país segundo classificação da QUALIS/CAPES e ranking REPEC – a ***Revista Brasileira de Economia (RBE)***.

A presente dissertação, portanto, está organizada da seguinte forma. Além desta introdução, o segundo capítulo, teórico, busca resgatar alguns tópicos pertinentes ao entendimento da confusão entre significância estatística e significância econômica. Primeiramente, são expostas as diferenças entre os métodos de Fisher, Neyman-Pearson e Bayesianos para testes de hipótese; discute-se, também, a ideia do *p-valor* como

medida de evidência. Trabalham-se, por fim, as noções de erro real e erro amostral, bem como a distinção entre diferença estatística e diferença substantiva.

O capítulo terceiro trata da parte empírica. Em um primeiro momento, resgata-se a literatura acerca do tema especificamente para a área da economia, com as evidências verificadas em outros países, como para os Estados Unidos – McCloskey e Ziliak (1996), Ziliak e McCloskey (2004a, 2008a) – ou a Alemanha – Kramer (2011). Em seguida, quantificamos e analisamos o problema no Brasil, tomando como amostra todos os 94 artigos publicados na RBE entre 2008 a 2011. Utiliza-se como base para o questionário os tópicos propostos por McCloskey e Ziliak (1996), levando em conta as críticas formuladas por Hoover e Siegler (2008a), bem como outros elementos apontados por Delong e Lang (1992), Gelman e Stern (2006), Keuzenkamp e Magnus (1995), McCloskey (1986), Wooldridge (2004), Zellner (1981), Ziliak e McCloskey (2004a) e Ziliak e McCloskey (2008a).

Ao final, são tecidas algumas considerações finais, indicando as limitações do trabalho e sugestões de pesquisas futuras.

2) O QUE (NÃO) É A SIGNIFICÂNCIA ESTATÍSTICA

Em seu livro *Statistics as Principled Argument* (1995, p. 54-55), Robert Abelson, professor de estatística por mais de 42 anos do departamento de psicologia da Universidade de Yale, traz a seguinte passagem pertinente ao nosso tema:

> [...] resultados negativos muitas vezes sequer são escritos. Estudantes chegam a abandonar dissertações porque a hipótese nula não pode ser rejeitada. Professores, editores de periódicos e outros leitores de manuscritos são menos receptíveis, em geral, a hipóteses nulas não-rejeitadas do que a rejeitadas [...] dadas as circunstâncias, é muito tentador para as pessoas tentar desesperadamente fazer com que seus resultados saiam estatisticamente significantes [...] isso é especialmente verdade para resultados quase significantes, digamos, $.05 < p < .10$. O jogo torna-se, então, em como empurrar os resultados para o nível convencional $p = .05$.

O leitor desta dissertação, que já tenha utilizado métodos da estatística clássica em trabalhos empíricos aplicados, provavelmente se identifica com a citação de Abelson. Pois, a importância (ou não importância) dada a resultados "(in)significantes", bem como a busca por tais resultados, também se verifica na economia[8]. Gigerenzer (1993, p.01-3) expõe esta prática sob a alcunha de *ritual nulo*, enfatizando que, nas ciências sociais, tem se visto o uso da estatística para mecanizar inferências, como uma solução algorítmica que funcionaria em qualquer situação ou contexto[9]. Este procedimento, entretanto, prejudica a intuição estatística do usuário, fazendo-o: (i) julgar um coeficiente como zero por ser estatisticamente insignificante[10]; (ii) julgar um coeficiente como importante por ser estatisticamente significante[11]; (iii) enfatizar a significância estatística "nominal" em contextos em que ela sequer faz sentido, como em modelos mal especificados ou quando outros erros não-amostrais são muito mais sérios e

[8] Wooldridge (2006), Goldberger (1989), Mayer (2006), Ziliak, McCloskey (2008a).

[9] Encontra-se em alguns livros de econometria de pós-graduação, quando tratam de testes de hipóteses, apenas algo como o seguinte algoritmo (HAYASHI, 2000, p. 89; GREENE, 2002, p. 51): (i) calcule a estatística de teste t, com relação à hipótese nula (H_0); (ii) se $t > t_\alpha$ onde t_α representa um valor crítico cuja probabilidade de obtê-lo (ou valores mais extremos), sob a hipótese nula, é igual a α, então rejeite H_0. Caso contrário, aceite-a. Após isso, costuma-se dar como exemplo o valor de α=5%, explicando a razão de este ter sido escolhido por ser "prática comum" (GREENE, 2002, p. 51; HANSEN, 2012, p. 159). Mais ainda, em relação à hipótese nula específica de que um coeficiente seja igual à zero, diz-se que se $t > t_\alpha$ então o coeficiente é "significante". Caso contrário, diz-se ser "insignificante".

[10] Ou julgar uma teoria adequada por não a rejeitar estatisticamente.

[11] Ou julgar uma teoria inadequada por rejeitá-la estatisticamente.

claramente presentes; (iv) ignorar a própria natureza econômica do problema, como as magnitudes dos coeficientes estimados; entre outros pontos[12].

Os livros-textos de estatística e econometria utilizados têm, de certo modo, contribuído para tanto, pois, apesar de o discurso padrão ser em torno do arcabouço de testes de hipótese de Neyman-Pearson, pouco ou nada se discute acerca de questões fundamentais, como: (i) formas de se calcular o poder dos testes (ou a severidade dos testes); (ii) como escolher o nível de significância tendo em vista crenças *a priori*, funções perda, testes múltiplos, buscas por especificação; (iii) ou até, algo mais básico, não se questiona a simples razão ou plausibilidade por trás de um teste de hipótese pontual ou exato[13]. Além disso, são raras as apresentações de métodos bayesianos, o que permitiria uma compreensão mais plural da estatística e evitaria a interpretação incorreta de alguns procedimentos. Deste modo, na prática, os usuários costumam acreditar que os testes de significância lhes dizem muito mais do que podem dizer – mais especificamente, costumam exagerar o que pode ser inferido de um *p-valor*. Por se tratar de hábito arraigado e difundido na literatura, os pesquisadores muitas vezes tomam estes hábitos como científicos e têm bastante resistência em questioná-los (ZILIAK, MCCLOSKEY, 2008a; GIGERENZER, 1993; FIDLER *et alii*, 2004).

A abordagem de muitos textos que discutem estes problemas costuma ser agressiva (HALLER, KRAUSS, 2002, p.14) e, talvez, isto tenha gerado o que em psicologia se denominou "efeito bumerangue"[14] – o que poderia criar ainda mais resistência a mudanças. Tentaremos evitar este tom neste trabalho, e buscar mostrar construtivamente sua pertinência. Daí a necessidade deste capítulo antes da exposição dos resultados

---

[12] Todos estes casos foram observados na amostra analisada nesta dissertação e serão apresentados no próximo capítulo.

[13] Lembre-se, por exemplo, que em uma distribuição contínua, a massa de probabilidade de um ponto, em geral, é igual a zero. Evidentemente que há livros que discutem algumas destas questões, como Kennedy (2008), com mais referências do que a discussão, Leamer (1978), ou Spanos (1993); contudo, tendo em vista os resultados empíricos encontrados, muito provavelmente não são estes que estão sendo utilizados.

[14] O "efeito bumerangue" foi discutido por Abelson e Miller (1967). Grosso modo, constatou-se que, em grupos em que o discurso era feito com insultos, o público tendia a ir ***contra*** a opinião de quem discursava, mesmo que concordasse com ela; quando ***o mesmo*** discurso era feito sem insultos, o público tendia a ir ***a favor*** da opinião de quem discursava. Em particular muitos autores [SPANOS (2008), ENGSTED (2009), KRAMER (2011), HOOVER E SIGLER (2008a), COBB (2009), ROBERT (2011)] acharam agressivo o "tom" do último livro de Ziliak e McCloskey (2008a). Robert (2011, grifo nosso) afirmou, por exemplo, que achou "[...] entristecedor um livro que trata de ***assunto tão importante*** deixar a agressividade, arrogância e a retórica ***arruinar o seu propósito***". Já Spanos (2008 p.155, p.163) achou que a retórica do livro acabou por utilizar a tática de "envergonhar acadêmicos notáveis" e denominou a atitude como "apontar o dedo".

empíricos encontrados. Pois, se a confusão entre significância estatística e significância econômica decorre do mau entendimento dos instrumentos de inferência utilizados, faz-se necessário explicitar – mesmo que sucintamente – o que estes instrumentos são e o que não são, o que podem e o que não podem responder. Do contrário, o leitor poderia tomar as críticas como gratuitas, infundadas, ou até pessoais, o que definitivamente não é o caso.

Neste capítulo, exporemos as principais diferenças entre três abordagens de testes de hipótese – as de Fisher, Neyman-Pearson e Bayes – e como a prática usual tem sido um híbrido das três. Em seguida, daremos ênfase à confusão com relação aos *p-valores* e à sua interpretação como medida absoluta de evidência. Discutiremos a diferença entre erro amostral e "erro real", o sentido de testes de hipóteses precisas e a decorrente confusão entre diferença estatística e diferença substantiva. Por fim, serão ilustrados meios de se amenizar alguns dos problemas tratados. Esta discussão serve de pano de fundo para entender a confusão entre significância estatística e significância econômica.

2.1. Fisher, Neyman-Pearson, Jeffreys e o Ritual Nulo

É comum verificar entre pesquisadores o desconhecimento sobre a natureza e a origem das diferentes teorias de inferência estatística. A compreensão dominante é de que haveria uma única e bem desenvolvida teoria clássica e isto estaria sendo propagado pelos próprios livros-textos utilizados nas universidades. Como apontam Hubbard e Bayarri (2003, p.01),

> [...] os livros modernos sobre a análise estatística nas ciências de negócios, sociais e biomédicas, seja em nível de graduação ou pós-graduação, geralmente apresentam o assunto como se fosse um evangelho: um único, unificado, e incontroverso método de inferência estatística.

Entretanto, as diferenças entre os métodos, em particular os testes de hipóteses, não são triviais[15]. A discordância entre três dos principais teóricos, Fisher, Jeffreys e Neyman, em uma situação tão simples quanto uma hipótese pontual sobre a média de uma distribuição, pode ser grande. Vejamos um exemplo baseado em Berger (2003, p.01).

Suponha uma amostra aleatória *i.i.d*, $\boldsymbol{X} = (X_1, \dots, X_n)$, proveniente de uma distribuição normal $N(\theta, \sigma^2)$, com a variância, σ^2, conhecida. Suponha que queiramos testar

[15] Há diversas tentativas de conciliação, como Berger (2003), que tenta reconciliar os três métodos, ou Mayo e Spanos (2011), que buscam conciliar os métodos de Fisher e Neyman-Pearson.

$H_0: \theta = 0$ contra $H_1: \theta \neq 0$. Dada uma realização específica dos dados, $x = (x_1, \dots, x_n)$, de tamanho $n = 10$, se tivermos que $z = \frac{\sqrt{n}\bar{x}}{\sigma} = 2.3$[16], então Fisher, Jeffreys e Neyman provavelmente divulgariam os seguintes resultados:

(i) Fisher reportaria o *p-valor*, isto é, $p = 0{,}021$;

(ii) Jeffreys reportaria a probabilidade *a posteriori* de H_0 , isto é, $\Pr(H_0|x_1, \dots, x_n) = 0{,}28$[17], considerando que ambas as hipóteses têm igual probabilidade *a priori* de 50% e utilizando uma distribuição *a priori* $Cauchy(0, \sigma)$ para distribuir a massa de probabilidade da hipótese alternativa;

(iii) Neyman teria escolhido, antes de realizar o teste, uma probabilidade de *erro tipo I*, representada por α, e uma probabilidade de *erro tipo II*[18], representada por β. Supondo que, por acaso, Neyman tivesse escolhido α = 0,05, ele reportaria simplesmente que rejeita H_0 com α = 0,05 (e com β em determinado valor para uma hipótese alternativa), independentemente de o *p-valor* obtido ter sido muito menor do que 5%.

Percebe-se a discrepância entre os números reportados– mas, mais divergentes são suas interpretações. Fisher ressaltaria um *p-valor* de 2,1% que usualmente seria considerado evidência bastante forte contra H_0, pois indica que, ou (i) H_0 é verdade e observamos um fato – a realização de $x = (x_1, \dots, x_n)$ – cujo valor, ou valor mais extremo, apenas ocorreria apenas 2,1% das vezes, ou (ii) H_0 é falsa. Já Jeffreys nos informaria uma probabilidade de 28% de a hipótese nula ser verdadeira, dado o resultado encontrado – note que esta probabilidade, dependendo da seriedade do caso, poderia ser considerada alta, e uma evidência não tão forte quando comparada com a argumentação de Fisher. Neyman, por sua vez, nos forneceria um nível de significância freqüentista de 5%, nos dizendo que se H_0 for verdadeira, apenas em 5% das vezes o nosso teste nos faria rejeitá-la como agora o fazemos, e que o poder do nosso teste, isto é, a probabilidade de

[16] $\bar{x}$ refere-se à média amostral e z trata-se da estatística *z*, que possui distribuição $Normal\ (0,1)$, que não depende do parâmetro θ.

[17] A probabilidade *a posteriori* é calculada utilizando o teorema de Bayes e as distribuições *a priori* mencionadas. Mais sobre o teorema de Bayes será dito a seguir.

[18] Em alguma hipótese alternativa de referência considerada cientificamente relevante.

rejeitar H_0 quando esta é falsa, para uma alternativa relevante de nosso interesse, é de $1 - \beta$.

A despeito das interpretações conflitantes, muitos livros textos não expõem devidamente estas distinções e o que elas significam – ou, caso fosse a intenção, em como conciliá-las de modo coerente. Ocorre que, atualmente, o método considerado clássico é o de testes de hipótese originalmente desenvolvido por Jerzy Neyman e Egon Pearson. Todavia, muitas vezes este é apresentado com elementos dos testes de significância popularizado por Fisher[19]. Este híbrido, com mais elementos de Fisher do que propriamente Neyman-Pearson – e, ainda, com uma interpretação bayesiana de alguns resultados – tem vigorado na prática e foi denominado por Gigerenzer (2004, p.02) de "ritual nulo", podendo ser resumido da seguinte forma:

(i) Estabelecer uma hipótese nula tais como "nenhuma diferença na média" ou "zero correlação", e nunca especificar as predições de sua hipótese de pesquisa ou tampouco qualquer outra hipótese alternativa relevante;

(ii) Utilizar o nível de significância convencional de 5% (ou 10%) para rejeitar a hipótese nula. Se for significante, aceitar a hipótese da pesquisa. Reportar os resultados como $p < 0{,}10$, $p < 0{,}05$, $p < 0{,}01$ ou $p < 0{,}001$ (ou o que vier mais próximo do *p-valor* obtido, como se fossem medidas α de *erro tipo I*. Este procedimento foi batizado, por Goodman (1993), como "***alfas itinerantes***");

[19] Conforme afirmam Keuzenkamp e Magnus (1995, p. 06), os cursos de econometria costumam apresentar os testes de hipóteses dentro do arcabouço formal de Neyman-Pearson. Já a prática costuma se assentar em Fisher, sem que os próprios usuários reconheçam o fato. Ou como coloca de modo mais direto Spielman (1974, p.211) "[...] muitos jovens investigadores em ciências sociais tendem a usar uma teoria híbrida dos testes, que são chamados 'testes de significância'. Este híbrido é essencialmente Fisheriano em sua lógica, mas diz da boca para fora que segue a teoria de testes de Neyman-Pearson (esta última é, atualmente, a teoria oficial dos testes, certificada pelos livros, na América do Norte). Alguns pesquisadores fazem uso da teoria de Neyman-Pearson em uma forma pura, mas eles constituem uma pequena minoria. Independentemente da sua fidelidade à terminologia verbal, a maioria dos pesquisadores [...] utilizam e / ou aceitam como válido um padrão de raciocínio indutivo que é característica de testes Fisherianos de significância. No entanto, as profundas lógica e estrutura desse padrão são quase que universalmente mal compreendidas". Sebastiani e Viali (2011, p.838), em estudo recente, no Brasil, afirmam que "[...] nenhuma ou quase nenhuma atenção é dedicada aos conceitos básicos dos testes, tais como erros do Tipo I (nível de significância), Tipo II (poder de um teste), região crítica e valor-p. Isto ainda é agravado por abordagens que misturam a de Neyman-Pearson com a de Fisher [...]".

(iii) Não reportar o poder do teste nem o que levou o pesquisador a escolher os níveis de significância utilizados, bem como não analisar o tamanho do efeito encontrado.

Trata-se de procedimento que acaba por resultar: (i) na confusão entre medidas de erro procedimentais (α) e uma variável aleatória, medida de "evidência" ou "erro hipotético", "contra factual" *(p-valor*); (ii) como métodos bayesianos são pouco trabalhados, na confusão entre as probabilidades freqüentistas utilizadas e probabilidades *a posteriori* de a hipótese nula ser verdadeira ou de se cometer um erro; e, principalmente, (iii) na confusão entre a significância estatística e a significância científica dos resultados da pesquisa (BERGER, 2003; GIGERENZER, 2004; GIGERENZER, GRAUSS E VITAO, 2004; HUBBARD, BAYARRI, 2003; ZILIAK, MCCLOSKEY, 2008a).

Nesta seção, portanto, apresentaremos de forma ***resumida e geral*** essas três abordagens de inferência[20]. A exposição a seguir é baseada em Casella e Berger (2002, p. 373-413), Hogg e Craig (1978, p.235-257), Lehman e Romano (2008, p.56-107) para o delineamento do método "clássico" moderno; em Cox (1958, 1977, 1982), Mayo e Cox (2006), Wagenmakers (2007), para os testes de significância e em Berger (1985), Berger (2003), Berger e Delampady (1987), Christensen (2005), Cohen (1994), Gigerenzer (1993), Gigerenzer (2004), Gigerenzer, Grauss e Vitao (2004), Hubbard e Armstrong (2006), Hubbard e Bayarri (2003), Hubbard, Bayarri, Berk e Carlton (2003), Kadane (2011), Keuzenkamp e Magnus (1995), Mayo e Spanos (2006; 2011), Spielman (1974) e Ziliak e McCloskey (2008a)[21] para o realce das diferenças. Dessa forma, para facilitar a leitura, as referências serão omitidas a não ser em casos específicos ou de citação direta.

[20] Dessa forma, tendo em vista o objetivo da seção, obviamente que diversos pontos relevantes dos métodos serão omitidos, principalmente de análise Bayesiana e de teoria da decisão.

[21] Também foram utilizadas as referências originais de Fisher (1971, 1973a, 1973b), Neyman (1950), Neyman e Pearson (1928a, 1928b, 1933), Pearson (1955, 1962) e Wald (1939, 1949). Todavia, a intenção desta parte do trabalho não é realizar um resgate histórico do desenvolvimento das teorias e, portanto, será exposto apenas o arcabouço mais geral de seus fundamentos sem adentrar em nuances e exceções apontadas pelos autores, que debateram vigorosamente entre si.

2.1.1. Testes de Significância de Fisher

Os testes de significância de Fisher procuram analisar a realização dos dados e verificar se esta é ***consistente com uma dada hipótese***, denominada hipótese nula (H_0). A hipótese nula, em geral, supõe que a amostra provém de uma ***população hipotética infinita*** com distribuição conhecida. O procedimento do teste de significância seria, assim, derivar a distribuição de uma estatística de teste, $T(X)$, caso H_0 fosse verdadeira, e definir a probabilidade de se observar um resultado tão extremo ***ou maior*** do que o observado sob estas circunstâncias. Tal probabilidade é uma estatística, usualmente chamada de *p-valor*, isto é:

$$p = Pr(|T(X)| \geq |t| \mid H_0) \qquad (01)$$

Quanto menor o *p-valor*, maior seria a "evidência" contra a hipótese testada[22]. Para Fisher, tal estatística forneceria ***evidência indutiva*** contra a hipótese nula, baseada no seguinte raciocínio de ***probabilidades disjuntas:*** ou a hipótese nula é verdadeira e nos defrontamos com um evento cujo valor, ou valor mais extremo, é raro (mas, frise-se, possível), ou a hipótese nula é falsa. Tendo em vista a perplexidade de se defrontar com um evento raro caso a hipótese nula seja verdadeira, haveria motivos "objetivos" para sua descrença. O teste de significância estatística, fundamentado no *p-valor*, seria, portanto, uma forma de ***inferência indutiva***.

Em seus trabalhos iniciais, Fisher propunha que, caso o *p-valor* fosse menor do que certo nível considerado bastante improvável para o pesquisador, chamado de ***nível de significância,*** a hipótese nula fosse considerada rejeitada. Como visto na introdução deste trabalho, o autor chegou a sugerir padrões para a comunidade cientifica, afirmando ser:

> [...] usual e conveniente para os pesquisadores tomarem 5% como um nível de significância padrão, no sentido de estarem preparados para ignorar todos os resultados que falhem em alcançar este padrão e, por este meio, eliminar de uma discussão mais aprofundada grande parte das flutuações que a sorte possa ter introduzido em seus resultados experimentais (FISHER, 1971, p.13).

Já em sua fase madura, Fisher afirmava que os pesquisadores não teriam de tomar uma decisão acerca da rejeição ou não da hipótese, apenas deveriam comunicar a informação

[22] Verificaremos na seção seguinte que este entendimento tem de ser contextualizado.

encontrada. Desta forma, dever-se-ia ***relatar o p-valor exato*** do teste, como uma ***propriedade dos dados*** frente a uma suposta hipótese.

É possível demostrar que, quando H_0 é verdade, em condições bem gerais, $p \sim uniforme\ (0,1)$. Isso significa que ***qualquer valor para p***, ou qualquer intervalo de mesmo tamanho ϵ, ***é equiprovável*** quando H_0 é verdadeira. Sob tais circunstâncias, obter um valor muito pequeno de p, como 2^{-1000} ***não é algo mais raro*** do que obter qualquer outro valor como, digamos, $p = 0{,}67$. O raciocínio para a estatística ser evidência contra a hipótese nula não é fundamentado na raridade do *p-valor*, que é uniformemente distribuído sob H_0, e sim na lógica de probabilidades disjuntas exposta anteriormente. Por exemplo, segundo Fisher, $p = 0{,}999$ também poderia ser considerado evidência contra H_0, pois:

$$Pr(|T(X)| \geq |t| \mid H_0) = 0{,}999 \rightarrow\ Pr(|T(X)| \leq |t| \mid H_0) = 0{,}001 \qquad (02)$$

o que significa que, ou a hipótese nula é verdadeira e nos deparamos com uma discrepância ***tão pequena ou menor*** do que seria observada uma em cada mil vezes, ou a hipótese nula é falsa[23].

Pelo exposto, percebe-se que os testes de significância não fazem referência a qualquer hipótese alternativa e tratam-se, deste modo, tão somente de testes acerca da consistência da hipótese nula – em um sentido estrito, probabilístico. Buscam apenas informar se os dados são coerentes com o modelo que se supôs gerá-los, sendo a medida de coerência sua improbabilidade. Resultados extremos seriam considerados muito discordantes com H_0, levando à sua desconfiança, realizando, desta forma, um ***simulacro*** da prova por contradição, ou ***modus tollens*** [24]. Quando os dados "contradizem" o modelo, no sentido probabilístico, haveria evidências de sua falsidade.

---

[23] Fisher (1973a, p. 81), com relação aos testes χ^2de grau de ajuste, afirma que "[...] valores acima de 0,999 tem algumas vezes sido reportados e, se a hipótese for verdadeira, ocorreriam apenas uma vez em mil testes [...] nesses casos, a hipótese é considerada definitivamente rejeitada como se p tivesse sido 0,001".

[24] Se $A \rightarrow B$ então, necessariamente, $não\ B \rightarrow não\ A$. Obviamente que no contexto probabilístico este raciocínio não é válido, daí a ideia de simulacro do ***modus tollens***. Se $A\ então\ provavelmente\ B$, isso não implica que $não\ B\ então\ provavelmente\ não\ A$. Um exemplo adaptado de Cohen (1994) ilustra facilmente a situação: se você é brasileiro, provavelmente não é membro do congresso; entretanto, se você é membro do congresso, isso não quer dizer que provavelmente você não seja brasileiro. Vide Spielman (1974) para crítica mais extensa da lógica subjacente aos testes de significância. Fisher justifica a aproximação da prova por contradição como um meio de ***inferência indutiva*** e de ***aprendizagem pelo erro***, vide Fisher (1971, 1973a, 1973b), Mayo (1985, 2005a), Mayo e Spanos (2006, 2011).

Todavia, não é claro qual o aspecto do modelo que está sendo rejeitado. Se $x_1, x_2, \dots, x_n$ são supostas independentes com distribuição $N(\mu, \sigma^2)$ e realizamos o teste de $H_0: \mu = 0$,

> [...] a rejeição pode significar que $\mu \neq 0$, ou pode significar que os dados não são independentes, ou pode significar que a distribuição não é normal, ou pode significar que as variâncias das observações não são iguais[25] [...] ele [o teste de significância] não especifica o que está errado[26] (CHRISTENSEN, 2005, p.124).

Se os dados não contradizem o modelo, apenas não se chega a nenhuma "prova" e o melhor que se pode dizer é que os dados são consistentes com a hipótese. Note que, neste caso, ***a decisão acerca da hipótese nula é assimétrica.*** O pesquisador estabelece qual o nível de significância que consideraria evidência suficiente para rejeitar H_0 e verifica se o resultado é estatisticamente significativo. Uma falha em rejeitar a hipótese nula não levaria a nenhuma conclusão, pois não haveria qualquer forma de confirmá-la. Não rejeitar não provaria que o modelo está correto, simplesmente porque podem existir diversas outras hipóteses consistentes com os mesmos dados – isto é, outras hipóteses que também não seriam rejeitadas. Daí a ideia de que nunca se deveria "aceitar" uma hipótese nula[27].

Vejamos alguns corolários do que foi discutido, que merecem destaque por aparecerem constantemente em exemplos práticos: (i) para o caso específico de $H_0: \theta = 0$, a ausência de significância ***não significa*** a ausência de efeito; (ii) se há dois estudos, um em que o resultado é "estatisticamente significante" e o outro em que o resultado é "estatisticamente *in*significante", isto ***não significa*** que os resultados são conflitantes – por exemplo, dois estudos com $H_0: \theta = 0$, $\widehat{\theta_{1,2}} \approx 2$, e $p_1 = 0{,}05$, $p_2 = 0{,}15$, apontam para a mesma direção, contra H_0; (iii) se vários estudos não encontraram significância estatística isto ***não significa*** que a evidência de que inexiste efeito foi corroborada –

---

[25] Ou, ainda, pode indicar inclusive que são ***os dados*** que estão mal mensurados, incompletos ou errados. Como bem lembra Gigerenzer (1993, p.326), "[...] durante o século XIX astrônomos utilizaram testes de significância para rejeitar *dados* (os chamados *outliers*), pressupondo, pelo menos provisoriamente, que suas hipóteses estavam corretas [...] a matemática não nos diz em quem os pesquisadores devem confiar ou quem eles devem refutar [se os dados ou as hipóteses]".

[26] Novamente, se A **e** B **e** $C \rightarrow D$ então a negação de D implica, necessariamente, e tão somente em $não\ A$ **ou** $não\ B$ **ou** $não\ C$. A negação de D não implica na negação de uma hipótese ou um conjunto de hipóteses em particular.

[27] Isto concerne à preocupação em não realizar a falácia de afirmação do consequente. Se $A \rightarrow B$ isso ***não*** quer dizer que $B \rightarrow A$. Poderíamos ter, por exemplo, que $C \rightarrow B$ e, dessa forma, observar B é consistente tanto com A quanto com C.

pode muito bem ser o contrário. Tome, por exemplo, quatro estudos independentes testando $H_0: \theta = 0$, com $p_1 = 0{,}27, p_2 = 0{,}26, p_3 = 0{,}23, p_4 = 0{,}31$. Neste caso, dentro do paradigma de testes de significância, o ***conjunto geral*** indica um resultado ***estatisticamente significativo***[28]; (iv) para se julgar com segurança a respeito da hipótese específica do parâmetro, $H_0: \theta = 0$, é preciso assegurar que as ***outras fontes de erro*** estejam controladas (isto é, que as outras hipóteses para se derivar a distribuição de $T(X)$ sejam razoavelmente válidas).

Este ponto pode ser ressaltado de uma perspectiva diferente, lembrando-se mais uma vez a definição de *p-valor:* trata-se da probabilidade de se observar um valor tão extremo ***ou mais extremo*** do que o observado sob H_0. Perceba, portanto, que a evidência que é gerada contra a hipótese nula não é baseada somente no que de fato foi observado, mas também no que ***poderia*** ter sido observado, pois utiliza a distribuição amostral[29]. Daí a seguinte observação de Jeffreys*:*

> [...] se o *p-valor* é pequeno, isso quer dizer que houve grandes desvios inesperados da predição. Mas por que isto deveria ser expresso em termos do *p-valor?* Este último nos dá a probabilidade de desvios, medidos em uma maneira particular, ***iguais ou maiores*** do que o conjunto observado, e a contribuição do valor atual obtido é quase sempre negligenciável. ***O que o uso do p-valor implica, portanto, é que uma hipótese que pode ser verdade pode ser rejeitada porque não previu resultados observáveis que não ocorreram*.** Isso parece um procedimento notável. (JEFFREYS, 1961, p. 385 grifo nosso)

Esta diferença, que pode não ser tão intuitiva a princípio, talvez seja mais facilmente ilustrada com um exemplo numérico. Suponha que você tenha várias urnas com a mesma proporção desconhecida θ de bolas brancas e a mesma proporção $(1-\theta)$de bolas vermelhas. Você, então, realiza dois experimentos: (i) no primeiro você escolhe doze urnas e retira de cada apenas uma bola; e (ii) no segundo você vai retirando uma bola de cada urna até que você tenha três bolas brancas. Agora suponha que o resultado do experimento (i) tenha sido três bolas brancas entre as doze retiradas e que o resultado

---

[28] Pode-se testar a significância conjunta de estudos independentes utilizando-se o resultado de que a soma do logaritmo de v.a.'s com distribuição uniforme tem distribuição χ^2.

[29] Este é um ponto bastante criticado por Bayesianos, pois fere o princípio da verossimilhança, ver Berger (1985), Edwards et alii. (1963), Kadane (2011). A explicação que Cox dá para se incluir informação não observada ao *p-valor* é nos seguintes termos: "suponha que nós tenhamos decidido que a observação é decisiva para rejeitar a hipótese nula. Então teríamos de aceitar dados mais extremos como evidência contra a hipótese" (COX, 1982, p.326). Outro argumento é que o princípio da verossimilhança seria adequado dentro do paradigma Bayesiano, mas não deveria ser um princípio frequentista, pois impediria o controle adequado das probabilidades de erro. Para este contraponto, ver, por exemplo, Mayo e Spanos (2011) ou Cox e Mayo (2010).

do experimento (ii) tenha sido que a terceira bola branca saiu na décima segunda urna. Em ambos os experimentos você tem 25% de bolas brancas, ou 3/12. Uma inferência estatística baseada nestes dados seria diferente para cada resultado?

Testemos $H_0: \theta = 50\%$. O *p-valor* para o primeiro experimento seria $p_1 \approx 7{,}3\%$, pois a ***distribuição amostral é binomial***; já no segundo experimento teríamos $p_2 \approx 3{,}3\%$, pois a ***distribuição amostral é binomial negativa***. Caso considerássemos o nível de significância de 5% como limiar, o segundo experimento nos levaria a rejeitar H_0 enquanto o primeiro experimento não. Ou, ainda, se considerássemos a ideia de Fisher em sua fase madura, e apenas reportássemos os *p-valores*, a evidência contra H_0, nestes termos, poderia ser considerada como "duas vezes maior" no segundo experimento (pois p_2 é menor do que a metade de p_1). Todavia, note que, em ambos os casos, temos três bolas brancas de doze bolas retiradas. A amostra obtida é a mesma. O que muda o resultado da inferência é algo que ***poderia ter sido observado,*** mas não foi, ou seja, ***a intenção*** de se ter pegado menos ou mais do que doze bolas no segundo experimento até se obter a terceira bola branca.

Agora, suponha que o pesquisador não saiba se o experimento realizado fora o da situação (i) ou o da situação (ii). Ao se deparar com a amostra de três bolas brancas e nove bolas vermelhas, o pesquisador teria de ***supor***, também, ***um modelo probabilístico*** que originou a amostra para poder realizar um teste de significância. Caso escolhesse o modelo (ii), defrontar-se-ia com o *p-valor* de 3,3%, o que aos níveis usuais de significância implicaria na rejeição de H_0, pois os dados indicam que o resultado encontrado é bastante improvável sob a hipótese nula. Como visto, contudo, improvável segundo qual aspecto de H_0? A contradição é de que $\theta = 50\%$ ou de que a distribuição amostral é binomial negativa? Como coloca Cox (1987, p.336),

> [...] ele [o teste de significância] serve como um aviso geral de que algo está errado (ou não), e não como um suporte explícito para uma explicação alternativa em particular. Portanto, tais testes têm um objetivo bastante limitado e em geral o pesquisador deveria fazer algo mais fortemente focado.

O pesquisador, portanto, tem de ter cautela para não julgar da inferência do teste de significância mais do que este é capaz de oferecer.

2.1.2. Testes de Hipóteses de Neyman-Pearson

Entremos agora no paradigma dos testes de hipóteses clássicos. Insatisfeitos com o método de Fisher[30], Neyman e Pearson buscaram aprimorar os testes de significância; porém, a contribuição dos autores acabou por diferir substancialmente da inferência indutiva anterior[31], fundamentando-se na busca por ***regras de decisão*** "ótimas" em contextos probabilísticos. Os testes de hipótese clássicos, neste sentido, são inerentemente ***dedutivos***. Na sua forma mais simples, a abordagem já se inicia com uma diferença: ao invés de somente uma, são estabelecidas *a priori* pelo menos duas hipóteses estatísticas concorrentes, a hipótese nula, H_0, e a hipótese alternativa, H_1. O teste consistiria em ***decidir pela aceitação de uma das duas***. Ou seja, neste contexto ***a decisão é simétrica, decide-se ou por H_0 ou por H_1.*** Para fazer a escolha, toma-se uma amostra aleatória $\boldsymbol{X} = (X_1, \dots, X_n)$ de $X \sim f(x;\theta)$, e define-se um subconjunto C do espaço amostral, denominado de ***região crítica,*** tal que se $\boldsymbol{X} \in C$ então se aceita H_1 e se $\boldsymbol{X} \notin C$ aceita-se H_0.

Uma vez definido C, tem-se, em especial, definidas também duas probabilidades condicionais ***a priori***, isto é, anteriores ao teste: a primeira, denotada por $\alpha = \Pr(\boldsymbol{X} \in C|H_0)$, é a probabilidade de a amostra aleatória pertencer à região crítica quando a hipótese nula é verdadeira – este valor também é usualmente chamado de ***tamanho do teste*** ou, infelizmente, para aumentar a confusão entre os métodos de inferência, ***nível de significância.*** Observe que, neste caso, o pesquisador cometeria um erro e rejeitaria H_0 quando H_0 é de fato verdadeira. Este erro é denominado de ***erro tipo I.*** Já a segunda probabilidade, denotada por $\beta = \Pr(\boldsymbol{X} \notin C|H_1)$, representaria as chances de a amostra aleatória não pertencer à região crítica quando a hipótese alternativa é verdadeira. Note que aqui também haveria um erro, pois o pesquisador aceitaria H_0 quando H_1 de fato é

---

[30] Como colocou Pearson (1962, p.395), "[...] o que Neyman e eu experimentamos [...] foi uma insatisfação com a base lógica – ou sua falta – que parecia amparar a escolha e construção de testes estatísticos [...] nós tentamos, portanto, desenvolver um conjunto de princípios com fundamentos matemáticos que, nos parecia, levariam a uma escolha racional de procedimentos estatísticos ao se defrontar com certos tipos de problemas de análise e interpretação de dados".

[31] Fisher (1955, p.69), por exemplo, afirmou que "a tentativa de reinterpretar os testes de significância comuns utilizados na pesquisa científica como se fossem um tipo de procedimento de aceitação que levassem a decisões no sentido de Wald se originou em várias incompreensões e, aparentemente, levou a muitas mais". As diferenças destacadas por Fisher dizem respeito à ideia de ***amostragem repetida de uma população definida***, à idéia de ***erro tipo II*** e à idéia de ***comportamento indutivo fundamentado em regras de decisão*** frente à noção de inferência indutiva e aprendizagem pelo erro. Deborah Mayo (1992) acredita que as diferenças entre os autores eram mais pessoais do que teóricas e alega que Pearson não concordou plenamente com a filosofia de "comportamento indutivo" proposta por Neyman.

verdadeira. Este erro é denominado de ***erro tipo II.*** Também surge daí a ideia de ***poder do teste,*** que é a probabilidade complementar do *erro tipo II*, $1 - \beta = \Pr(\boldsymbol{X} \in C|H_1)$ – a probabilidade de se rejeitar a hipótese nula quando a hipótese alternativa é verdadeira.

Contrariamente à ideia de uma população hipotética infinita de Fisher, o teste de hipótese de Neyman-Pearson presume o uso de ***amostras aleatórias repetidas*** de uma ***população definida***. É desta situação que decorrem as noções de α e β como limites freqüentistas de erro. A partir daí, uma vez montado o problema, caberia ao pesquisador encontrar ***uma regra a ser seguida*** que ***minimizasse os riscos dos erros tipo I e tipo II***. Contudo, uma vez fixado o tamanho amostral em n, a diminuição de α implica em um aumento de β, e não é possível minimizar ambos ao mesmo tempo. O que fazer então? A solução proposta pelos autores foi ***fixar $\boldsymbol{\alpha}$*** e encontrar a regra de decisão que ***minimizasse o erro tipo II***, ou, analogamente, que ***maximizasse o poder do teste, $\boldsymbol{1-\beta}$, dado $\boldsymbol{\alpha}$***. O lema fundamental de Neyman-Pearson consiste justamente na demonstração da existência e das condições necessárias e suficientes para o teste "mais poderoso" [32].

Neste contexto, cabe introduzir o conceito de ***função perda***[33]**.** Suponha que as hipóteses nula e alternativa digam respeito a um parâmetro θ da população. Uma vez definida uma ***função de decisão*** para cada realização amostral, $d = \delta(x)$, pode-se associar uma perda a cada decisão dado o verdadeiro valor de θ, isto é, $L(\theta, d)$. A partir daí é possível derivar uma ***função risco***, $R(\theta, \delta) = E_\theta\big[L\big(\theta, \delta(x)\big)\big]$, que represente o valor esperado da perda (no espaço amostral) quando θ é o verdadeiro parâmetro (por isso o subscrito θ na esperança). O teste de hipótese no contexto de Neyman-Pearson poderia ser visto com um ***problema de decisão*** com dois tipos de perda. Supondo que as hipóteses concorrentes sejam definidas por $H_0: \theta \in \Theta_0$ e $H_1: \theta \in \Theta_1$, e que d_0 e d_1 representem a decisão de aceitar e rejeitar H_0 respectivamente, poderíamos estabelecer

---

[32] A ideia de teste mais poderoso se aplica ao se confrontarem duas hipóteses simples. Ao se testar uma hipótese simples contra uma hipótese composta, tem-se a noção de teste "uniformemente mais poderoso". A partir da contribuição seminal de Neyman e Pearson, foram desenvolvidos outros conceitos de propriedades "desejáveis" para os testes, de modo a se buscar os "melhores" testes em condições mais gerais. Os testes t bilateral, e o teste F, por exemplo, comumente utilizados na econometria, são exemplos de testes, respectivamente, "uniformemente mais poderoso não-viesado" e "uniformemente mais poderoso invariante". Em Neyman-Pearson os testes são com tamanho de amostra fixo; já se avançou o estudo de testes sequenciais. O desenvolvimento dos conceitos de não-viés, consistência, invariância, quase-invariância, admissibilidade, testes sequenciais, testes múltiplos, entre outros, não é necessário para a exposição deste trabalho e podem ser facilmente encontrados nos textos elencados nas referências.

[33] Análise extensiva sobre funções perda foi primeiramente realizada por Wald (1939; 1949).

as seguinte perdas: caso a decisão tomada seja errada (um *erro tipo I* ou *erro tipo II*), perde-se 1; por outro lado, caso a decisão tomada seja correta, perde-se 0. Isto é:

$$L(\theta, d_i) = \begin{cases} 0, & se\ \theta \in \Theta_i \\ 1, & se\ \theta \in \Theta_j \end{cases} \quad (i \neq j) \tag{03}$$

Esta função perda é usualmente chamada de "0-1". Note que o risco quando H_0 é verdadeira é simplesmente a probabilidade de *erro tipo I:*

$$R(\theta_0, \delta) = E_{\theta_0}\big[L\big(\theta, \delta(x)\big)\big] = 1 * \Pr(\boldsymbol{X} \in C|H_0) + 0 * \Pr(\boldsymbol{X} \notin C|H_0) = \alpha \tag{04}$$

Já quando H_1 é verdadeira, o risco associado é simplesmente a probabilidade de *erro tipo II:*

$$R(\theta_1, \delta) = E_{\theta_1}\big[L\big(\theta, \delta(x)\big)\big] = 0 * \Pr(\boldsymbol{X} \in C|H_1) + 1 * \Pr(\boldsymbol{X} \notin C|H_1) = \beta \tag{05}$$

Neste caso, o teste de hipótese em Neyman-Pearson seria equivalente a um problema de decisão em que se busca encontrar uma regra, $\delta(x)$, função da distribuição amostral, que minimize o risco associado a quando a hipótese alternativa é verdadeira, impondo-se um limite máximo ao risco tolerado quando a hipótese nula é verdadeira, supondo-se, ainda, que ambos os erros *(tipo I e tipo II)* ou acertos tenham a mesma "perda" ou ganho caso cometidos (no caso 1 ou 0).

Portanto, fica claro em que sentido a abordagem de Neyman-Pearson é, na verdade, fundamentalmente ***dedutiva.*** Trata-se de um procedimento ***não evidencial*** que, com base no modelo utilizado, ***estabelece uma regra de decisão a priori*** controlando os riscos probabilísticos de tal empreitada. Dentro deste contexto, apenas um resultado em particular não precisa ter interpretação epistêmica como evidência contra ou a favor de uma teoria, pois, aceitar ou rejeitar uma hipótese não implica em acreditar que esta seja verdade, mas agir como se fosse verdade tendo em vista o critério de otimização adotado. O argumento parte do geral para o particular[34]. Neyman distinguiu este

[34] Cabe aqui esclarecer mais detidamente em que sentido se classifica o método de Fisher como indutivo e o de Neyman-Pearson como dedutivo. Evidentemente que, para se calcular o *p-valor,* devem-se estabelecer os pressupostos da distribuição populacional, definir-se a estatística de teste, e daí derivar-se, dedutivamente, a distribuição amostral da estatística. Entretanto, após isso, a interpretação epistêmica dada ao *p-valor,* em Fisher, é um passo indutivo; pois, caso se quisesse deduzir logicamente a improbabilidade da hipótese nula face à realização dos dados, seria necessário algo como a inversão de probabilidade pela regra de Bayes. O processo de Neyman-Pearson quando interpretado dentro do contexto de teoria da decisão, por outro lado, busca critérios "ótimos" de regras de comportamento para a

procedimento do raciocínio indutivo de Fisher, denominando-o de ***comportamento indutivo***. Em suas palavras,

> [...] o termo "raciocínio indutivo" permanece obscuro e é incerto se pode ser convenientemente usado para designar qualquer conceito claramente definido. Por outro lado [...] parece haver espaço para a expressão "comportamento indutivo". Esta pode ser utilizada para denotar o ajustamento do nosso comportamento a quantidades limitadas de informação. O ajuste é em parte consciente e em parte subconsciente. A parte consciente é baseada em certas regras (se eu vir isso acontecendo, então eu faço isso) que chamamos de regras de comportamento indutivo. Ao estabelecer essas regras, ambas a teoria da probabilidade e da estatística desempenham um papel importante, e há uma quantidade considerável de raciocínio envolvido. Como de costume, no entanto, o raciocínio é todo dedutivo (Neyman 1950, p 01).

Percebe-se, assim, que o teste de hipótese, no arcabouço de Neyman-Pearson, trata-se do estabelecimento de uma regra ótima no sentido estrito de minimizar β uma vez que α fora previamente definido. ***A teoria não fornece qual o balanço ótimo entre o nível de significância α e o poder do teste $1-\beta$***. Tal ponderação não se trata de um problema estatístico. Para tanto, o pesquisador deveria ter em mente quais são os custos associados a cada tipo de erro – algo específico ao seu problema – e estabelecer *a priori* os níveis de α *e* β com base em uma análise custo-benefício. Nas palavras de Pearson,

> [...] nós certamente estávamos cientes de que inferências devem usar informação *a priori* e que decisões devem levar em conta utilidades [...] nós deixamos uma lacuna em nosso modelo matemático para o exercício de um processo de julgamento pessoal mais intuitivo em tais questões [...] como a escolha da classe de hipóteses admissíveis, o nível de significância apropriado, a magnitude de efeitos relevantes e o balanço das utilidades (PEARSON, 1962, p.395-396).

Esta discussão, quando surge nos livros-textos de inferência, é muitas vezes em termos pouco formais. Em Lehmann e Romano (2008), por exemplo, os autores indicam que os padrões atualmente adotados para controle de *erro tipo I* surgiram como consequência

---

definição das áreas de aceitação e rejeição da hipótese nula sob incerteza. Poder-se-ia, frente aos riscos envolvidos, escolher-se, por exemplo, uma área de rejeição com $\alpha = 50\%$, e justificá-la por algum critério de "otimização" (como uma solução *mini-max*). Note que, neste caso, não haveria qualquer apelo a um simulacro do *modus tollens*, pois uma probabilidade de erro de 50% quando H_0 é verdadeira não teria força evidencial, indutiva, contra a hipótese nula. Entretanto, há quem recrimine este tipo de uso. Deborah Mayo (2004) acredita que a interpretação comportamental da teoria de Neyman-Pearson acaba por levar a decisões "rudes" de aceitação e rejeição. Deste modo, a autora busca reinterpretá-la dentro da filosofia indutiva de Fisher. Exporemos brevemente a abordagem ao final do capítulo, mas cabe mencionar, aqui, que alguns autores, como Casella (2004), ainda acham que esta reinterpretação é um pouco vaga; Ziliak e McCloskey (2008a), por sua vez, temem que a abordagem não leve em conta aspectos econômicos do problema, levando à mesma confusão entre significância estatística e econômica. É interessante ressaltar também que a análise de Neyman-Pearson fica, muitas vezes, circunscrita aos limites do modelo utilizado, enquanto os testes de significância podem permitir o escrutínio dos pressupostos do modelo – como um pressuposto de normalidade ou de linearidade nos Mínimos Quadrados Ordinários (GRAVES, 1978; KEUZENKAMP, MAGNUS, 1995; LOUCÃ, 2008; NEYMAN, PEARSON, 1933; SPANOS, MCGUIRK 2001).

das limitações computacionais da época em que os testes foram desenvolvidos e que, atualmente, é lamentável o fato de serem utilizados sem qualquer ponderação. Nas palavras dos autores,

> [...] a escolha do nível de significância α é de certo modo arbitrária, uma vez que na maioria das situações não há um limite preciso para a probabilidade de erro tipo I que pode ser tolerada. ***Valores padrões, como 0,01 ou 0,05, foram originalmente escolhidos para reduzir as tabelas necessárias para realizar vários testes.*** Pelo hábito e por conta da conveniência da padronização em prover uma referência comum, ***esses valores gradualmente ficaram entranhados como os níveis usuais*** a serem utilizados. ***Isto é lamentável, pois a escolha do nível de significância deveria levar em conta o poder que o teste irá alcançar contra as alternativas de interesse.*** Há pouco sentido em realizar um experimento em que se tem apenas uma pequena chance de se detectar o efeito procurado quando ele existe (Lehmann, Romano, 2008, p. 57, grifo nosso).

Em uma situação ideal, complementam Lehmann e Romano, para se aumentar o poder do teste ao nível desejado poder-se-ia aumentar o tamanho da amostra. Contudo, quando isso não é possível, é interessante refletir se um aumento na probabilidade de *erro tipo I* não compensaria a consequente redução na probabilidade de *erro tipo II*. Por outro lado, em situações de amostras grandes, na maior parte das vezes seria desejável diminuir ainda mais a probabilidade α, pois pouco se perderia em termos de poder. Uma questão mais subjetiva na determinação do tamanho do teste também é colocada pelos autores, quando afirmam que

> [...] outra consideração que pode entrar na especificação de um nível de significância é a ***atitude*** frente à hipótese antes de o experimento ser realizado. Se ***o pesquisador acredita firmemente*** que a hipótese é verdadeira, ***evidência extremamente convincente será requerida*** antes de se abandonar a crença e o nível de significância será fixado em nível ***bastante baixo*** (Lehmann, Romano, 2008, p. 58, grifo nosso).

Vejamos agora o que dizem Casella e Berger (2002), que apresentam perspectiva bastante diversa quando da estruturação do teste a ser aplicado. Caso o pesquisador acredite em dada hipótese, ao invés de estabelecê-la como H_0 e requerer α pequeno, os estatísticos recomendam que esta seja estabelecida como H_1, pois,

> [...] ao fixar o tamanho do teste, o pesquisador está apenas controlando as probabilidades de *erro tipo I*, e não as de *erro tipo II* [...] suponha que o pesquisador espere que um experimento dê suporte a uma hipótese em particular, mas não deseje afirmar isso a não ser que os dados deem suporte convincente. O teste pode ser montado então de forma que a hipótese alternativa seja aquela que se espera que o dado confirme [...] ao usar um teste de tamanho α, sendo α bastante pequeno, o pesquisador está se guardando contra dizer que os dados dão suporte à hipótese de pesquisa quando esta é falsa (CASELLA, BERGER, 2002, p. 386).

Mesmo com esta breve discussão, nota-se que há diversas formas de se montar o teste, tanto em relação a estabelecer qual será H_0 e qual será H_1 bem como quanto à ponderação entre as probabilidades de erro α e β. Estas são questões circunstanciais que ***fazem parte do problema*** e não poderiam ser simplesmente ignoradas ou omitidas. Ademais, além das discussões informais expostas acima, admitindo-se o caráter de teoria da decisão do procedimento de Neyman-Pearson, o pesquisador poderia formalizar a preocupação com os erros e buscar funções perda adequadas ao problema, bem como critérios de otimalidade para a escolha da região crítica, tais como minimizar o ***risco de bayes***[35] ou encontrar uma solução do tipo ***minimax***,[36] entre outras. Hoffmann (2001), por exemplo, discorre acerca da escolha de α nos moldes de Lehmann e Romano, entretanto com uma abordagem um pouco mais formal. E, ao final, conclui que

> [...] é fácil depreender que a escolha do nível de significância, em um dado problema, tem muito de arbitrário. A discussão apresentada tem por finalidade deixar clara a direção em que deve ser ajustado o nível de significância, conforme mudam a probabilidade *a priori* de H_0 ser verdadeira e a relação entre os custos de cometer *erro tipo I* e *erro tipo II* (HOFFMANN, 2001, p.175).

2.1.3. Contrastando α e *p-valor*

Uma vez que, ao se definir a região crítica em Neyman-Pearson, divide-se o espaço amostral em duas regiões exaustivas, uma de aceitação e outra de rejeição da hipótese nula, obviamente que é possível realizar o teste com qualquer função dos dados, isto é, com qualquer estatística, inclusive o *p-valor*. Desta forma, com base nas considerações acerca das circunstâncias em que o teste será aplicado – isto é, levando em conta o poder que o teste alcançará, os pesos de cada um dos erros envolvidos, entre outros fatores – ***uma vez definido α, a priori***, a regra de decisão para a rejeição de H_0 será, em geral, $p < \alpha$[37]. Note, contudo, que o ***nível de significância α*** é uma propriedade do

---

[35] Supondo uma distribuição π *a priori* para os parâmetros, o risco de bayes é o valor esperado da função risco, isto é, $E^{\pi}[R(\theta, \delta)]$. A solução de bayes é a regra de decisão $\delta(x)$ que minimiza o risco de bayes.

[36] Uma solução *minimax* é aquela regra de decisão $\delta(x)$ que minimiza o maior risco possível.

[37] Desta forma, não é incomum encontrar a definição de *p-valor, **no contexto de Neyman-Pearson***, como "o menor nível de significância em que a hipótese nula teria sido rejeitada" ou $p = \inf\{\alpha | \boldsymbol{X} \in C\}$. Entretanto, esta definição não tem uma interpretação de limite de erro de longo prazo, pois, como visto, o tamanho do teste é definido *a priori* e o *p-valor* é uma variável aleatória dependente da amostra. Fisher rejeitou este tipo de interpretação do *p-valor* (FISHER, 1971, p. 25; 1973b, p. 42-48;79-81; e FISHER 1955).

teste, ou seja, é fixo, e que o valor de p não importa a não ser pelo fato de pertencer à região crítica. Por conseguinte, o *p-valor* não é a probabilidade α de *erro tipo I* e, ao se observar um resultado como $p = 0{,}09$, não é válida a interpretação frequentista *a posteriori* de que H_0 é rejeitada com $\alpha < 10\%$[38].

Talvez a forma mais fácil de perceber esta diferença entre o *p-valor* e a probabilidade α de *erro tipo I* seja com ***testes randomizados.*** Para se alcançar um α arbitrário quando a distribuição de probabilidade não é contínua, é necessário o auxílio de fatores aleatórios que estão além do espaço amostral. Considere o exemplo baseado em Christensen (2005), representado na Tabela 1, abaixo:

Tabela 1 – Distribuição f(x|$\boldsymbol{\theta}$), $\boldsymbol{\theta = 0, 2}$

x	***1***	**2**	**3**	**4**
f(x|0)	0,980	0,005	0,005	0,010
f(x|2)	0,098	0,001	0,001	0,900

Para se conseguir um nível de significância arbitrário ao se testar $H_0: \theta = 0$ contra $H_1: \theta = 2$ é preciso realizar testes randomizados, que consistem em se construir áreas de rejeição C randômicas. Para $\alpha = 1{,}25\%$, por exemplo, três conjuntos possíveis que poderiam ser construídos seriam: (i) rejeitar quando $x = 4$, jogar uma moeda e, se der cara, rejeitar quando $x = 2$ (assim teríamos $\alpha = 1\% + \frac{0{,}5\%}{2} = 1{,}25\%$) ; (ii) rejeitar quando $x = 4$, jogar uma moeda e, se der cara, rejeitar quando $x = 3$ (assim teríamos $\alpha = 1\% + \frac{0{,}5\%}{2} = 1{,}25\%$); e, (iii) rejeitar quando $x = 2$ ou 3, jogar uma moeda duas vezes, e se der cara duas vezes, rejeitar quando $x = 4$ (assim teríamos $\alpha = 0{,}5\% + 0{,}5\% + \frac{1\%}{4} = 1{,}25\%$). Suponha que, por algum motivo, o pesquisador tenha escolhido a regra de rejeição em (iii). Observe que quando $x = 4$, a mesma "evidência", isto é, o mesmo *p-valor*, $p = 1\%$, uma em cada quatro vezes será considerado "significante" (probabilidade de se ter duas caras) e três em cada quatro vezes não será. Frise-se que neste caso p, inclusive, é menor do que a probabilidade de *erro tipo I* da regra de

[38] Entretanto, desde que corretamente calculado, o *p-valor* pode ter a intepretação contra-factual de severidade do teste (MAYO, SPANOS, 2006).

decisão (iii) que é $\alpha = 1{,}25\%$, independentemente de rejeitarmos ou não H_0 em um teste específico[39].

Assim, como dito, decerto é possível definir a mesma área de rejeição em termos de uma estatística $T(x)$ ou em termos da estatística específica do *p-valor*. Todavia, como ilustrado pelo exemplo, percebe-se que os testes de hipóteses não tem apelo ao simulacro da prova por contradição e, deste modo, o *p-valor* é apenas uma estatística que está ou não na área de rejeição definida *a priori*, C. O objetivo do teste é controlar um tipo de erro enquanto se minimiza o outro. Ocorre que muitas vezes, dentro do arcabouço de Neyman-Pearson, o nível de significância é tomado pelo pesquisador como o único fato de importância, sem qualquer reflexão quanto ao poder do teste. Isso é claramente incoerente e talvez o exemplo de Kadane (2011, p.439), apesar de irônico, ilustre de forma mais categórica o motivo:

> [...] considere, por exemplo, o seguinte teste trivial. Jogue uma moeda viciada que dê cara com probabilidade 0,95 e que dê coroa com probabilidade 0,05. Se a moeda der coroa, rejeite a hipótese nula. Uma vez que a probabilidade de rejeitar a hipótese nula quando ela é verdade é de 0,05, este é um teste com 5% de nível de significância válido. É também bastante robusto a erros nos dados; de fato, sequer depende dos dados.

Obviamente que este teste seria considerado absurdo. Mas tem um nível de significância de 5%; e, se a preocupação do pesquisador fosse de apenas controlar α, este é um teste que lhe dará absoluta certeza de que, caso H_0 seja verdadeira, ela apenas será rejeitada em 5% das vezes. Isto chama atenção à inferência realizada sem qualquer outra preocupação a não ser o nível de significância: ela é tão boa quanto permitir que uma moeda viciada realize automaticamente o julgamento científico no lugar do pesquisador. É neste sentido que Ziliak e McCloskey (2008a, p. 8-9) atentam ao fato de que ajustar os níveis de *erro tipo I* e *erro tipo II* é necessário para se lidar com diferentes atitudes frente ao risco e que, sem se atentar às perdas relativas e aos riscos envolvidos, um teste deste tipo não é "nem um pouco melhor do que uma tabela de números aleatórios" (ZILIAK, MCCLOSKEY, 2008a p. 8-9).

Em resumo, os testes de Neyman-Pearson e os testes de Fisher não são procedimentos, ***por si só***, comparáveis[40]. Enquanto o primeiro é projetado para otimizar a detecção de

---

[39] Para não entrar em contradição com a definição de *p-valor* como "menor nível de significância em que a hipótese nula teria sido rejeitada", é comum generalizá-la, no caso de testes randomizados, como "menor nível de significância em que a hipótese nula teria sido rejeitada com probabilidade 1".

alguma hipótese alternativa, o teste de significância não faz nenhuma referência a qualquer outra hipótese além da nula, inexistindo os conceitos de *erro tipo II* , aceitação da hipótese nula e poder do teste. Enquanto os testes de hipótese partem da premissa de amostras repetidas de uma população definida para a validade da interpretação de erros de longo prazo, definidos em α e β, o *p-valor* é uma variável aleatória, propriedade da amostra e da distribuição amostral, e refere-se à probabilidade de observar dados tão extremos, ou mais extremos, sob a hipótese nula, desempenhando um papel epistêmico como uma medida de "evidência indutiva" em experiências individuais. Essas questões, bem como o que fora exposto também anteriormente, são resumidas na Tabela 2.

Tabela 2 – Contraste entre *p-valor* e $\boldsymbol{\alpha}$

p-valor	$\boldsymbol{\alpha}$
Teste de significância	Teste de hipótese (regra de decisão "ótima")
Evidência contra $\mathbf{H_0}$	Limite de rejeição errônea de $\mathbf{H_0}$
Filosofia indutiva	Filosofia dedutiva
Inferência indutiva	Comportamento indutivo
Evidencial	Não evidencial
Variável aleatória	Pré-fixado antes do experimento
Aplicado a um estudo particular	Interpretação de longo-prazo em várias amostras
População hipotética infinita	Amostras aleatórias de uma população definida

Fonte: tabela baseada em Hubbard e Bayarri (2003).

2.1.4. Teorema de Bayes

Na breve incursão que fizemos anteriormente, foram apresentados os conceitos de *p-valor*, α e β, com base em propriedades da distribuição amostral. Viu-se que estes não se confundem entre si. Agora, cabe esclarecer como os três não são medidas de probabilidade *a posteriori*, isto é, tanto (i) o *p-valor* não é a probabilidade de a hipótese nula ser verdadeira, como (ii) α e β não são as probabilidades de se ter cometido um *erro tipo I* ou um *erro tipo II*, respectivamente. Para tais medidas, seria preciso fazer a inversão da ***probabilidade condicional da observação dada a hipótese*** para a ***probabilidade condicional da hipótese dada a observação***.

[40] Como fizemos referência anteriormente, Aris Spanos e Deborah Mayo buscam conciliar ambos os métodos, voltando-se com mais ênfase ao conceito de inferência indutiva de Fisher e introduzindo o conceito de severidade na análise posterior à coleta dos dados. Alegam os autores que "[...] embora a mistura de aspectos de testes de Neyman-Pearson e de Fisher seja frequentemente acusada de ser um híbrido incoerente [Gigerenzer, 1993], o guarda-chuva do ***erro estatístico***, ligado pela noção de severidade, permite uma mistura coerente de elementos de ambas as abordagens" (MAYO, SPANOS, 2011, p.164). Vide também Mayo e Spanos (2006).

Uma análise Bayesiana envolve o uso de informação *a priori* sobre os possíveis valores de θ e utiliza o teorema de Bayes para combiná-la com a informação fornecida pelos dados, encontrando, assim, a distribuição *a posteriori* dos parâmetros estudados. Suponha que queiramos testar $H_0: \theta \in \Theta_0$ contra $H_1: \theta \in \Theta_1$. Então, para uma inferência Bayesiana, bastaria calcular $\Pr(\Theta_0|x) = \alpha_0$ bem como $\Pr(\Theta_1|x) = \alpha_1$ e decidir acerca da validade de H_0 ou H_1 com base nas probabilidades *a posteriori* de cada hipótese.

Defina por $\pi(\theta)$ a distribuição *a priori*[41] do parâmetro de interesse. A distribuição conjunta de x e θ é, portanto, dada por:

$$h(x, \theta) = \pi(\theta) f(x|\theta) \tag{06}$$

Assim, a distribuição marginal *incondicional* de x pode ser escrita como:

$$m(x) = \int f(x|\theta)\pi(\theta)d\theta \tag{07}$$

Por conseguinte, a distribuição *condicional* de θ dado que se observou x, isto é, a distribuição *a posteriori* de θ é:

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} \tag{08}$$

E, consequentemente, a probabilidade *a posteriori* de qualquer conjunto Θ_i nada mais é do que:

$$\alpha_i = \Pr(\Theta_i|x) = \int_{\Theta_i} \pi(\theta|x)d\theta \tag{09}$$

Ou seja, uma vez obtida a observação x, é possível calcular a probabilidade *a posteriori* de uma hipótese em particular e, consequentemente, a probabilidade de se cometer um erro caso se aceite a hipótese como verdadeira. Não haveria necessidade de se realizar um teste propriamente dito, pois seria possível calcular, diretamente, qual a hipótese mais provável. Destarte, a validade ou não de uma hipótese sobre o parâmetro torna-se

[41] A validade de se supor distribuições *a priori* para os parâmetros não será discutida neste trabalho. A intenção aqui é contrastar a probabilidade frequentista com a probabilidade inversa para esclarecer erros frequentes de interpretação dos métodos clássicos, bem como a divergência que se pode ter entre os diferentes métodos de inferência. Para justificativas acerca da abordagem Bayesiana, ver Jeffreys (1961), Berger (1985), Kadane (2011), Gelman e Robert (2012).

claramente um problema de decisão no contexto Bayesiano: sabendo-se que H_0 tem a probabilidade α_0 de ser verdade e que H_1 tem a probabilidade α_1 de ser verdade, qual a melhor decisão a ser tomada?

Dentro do arcabouço do teste clássico de hipóteses, vimos que este também pode ser considerado como um problema de decisão utilizando uma função perda "0-1". Como ilustração, podemos traduzir este cenário em um contexto Bayesiano. A perda esperada quando se decide por H_0 é dada por:

$$E[L(\theta, d_0)] = \int L(\theta, d_0)\, \pi(\theta|x) d\theta = \dot{\int_{\Theta_1}} L(\theta, d_0)\pi(\theta|x) d\theta = Pr(\Theta_1|x) \tag{10}$$

Analogamente, a perda esperada quando se decide por H_1 seria igual a:

$$E[L(\theta, d_1)] = \int L(\theta, d_1)\, \pi(\theta|x) d\theta = \dot{\int_{\Theta_0}} L(\theta, d_1)\pi(\theta|x) d\theta = Pr(\Theta_0|x) \tag{11}$$

Em um contexto de perda "0-1", portanto, a decisão com menor perda esperada é simplesmente ***aceitar como verdade a hipótese mais provável a posteriori.*** Evidentemente, da mesma forma que no teste de hipótese clássico, considerar que ambos os erros tenham o mesmo peso nem sempre pode ser a escolha mais adequada. Para um cenário um pouco mais realista, por exemplo, em que se atribui uma perda k_i a cada tipo de erro, a hipótese nula seria ***rejeitada*** quando:

$$\frac{P(\Theta_0|x)}{P(\Theta_1|x)} < \frac{k_0}{k_1} \tag{12}$$

Diferentemente do caso clássico, aqui a disjunção entre crença e ação fica mais clara. O pesquisador pode chegar a um resultado *a posteriori* em que uma hipótese é mais provável do que outra e, mesmo assim, tendo em vista os riscos associados a cada tipo de decisão, escolher agir de modo diferente.

2.1.5. Voltando ao ritual nulo

Após a breve exposição sobre os métodos de inferência, percebe-se como o "ritual nulo" praticado é uma junção inconsistente de aspectos operacionais dos dois métodos clássicos e, muitas vezes, com interpretações bayesianas.

O primeiro passo é estabelecer apenas a hipótese nula, tal qual proposto por Fisher – exceto pelo fato de que na maioria das vezes a hipótese nula é, mecanicamente, zero. Note que este passo está em desacordo com Neyman-Pearson, em que, com base também em H_1, o pesquisador precisaria definir $\alpha, \beta \; e \; n$ anteriormente à realização do teste, levando em conta os riscos que está disposto a enfrentar (se n é dado então definir o balanço ideal de α e β). Já o segundo passo é um simulacro de Neynan-Pearson, tomando-se "decisões" de se aceitar ou rejeitar a hipótese nula, entretanto com base em um nível de significância arbitrário e muitas vezes com *alfas itinerantes*, considerando o *p-valor* como a medida α de *erro tipo I* (ou considerando-o uma medida *a posteriori* de erro). Note que, para Fisher, aceitar a hipótese nula no contexto dos testes de significância seria equivalente à falácia da afirmação do consequente. Ademais, em sua fase madura, Fisher sugeria apenas a comunicação do *p-valor* exato do teste, sem qualquer julgamento de aceitação acerca da hipótese (GIGERENZER, 1993, 2004).

Mais ainda, como expôs Cohen (1994, p.997), os testes de significância estatística, com seus *p-valores,* "[...] não nos dizem o que queremos saber, e nós queremos tanto saber o que queremos saber que, no desespero, nós ainda assim acreditamos que eles nos dizem [o que queremos saber]![42]"; isto é, a informação que o *p-valor* fornece ao cientista, como vimos, é a probabilidade de a estatística de teste ser tão grande ou maior do que a estatística efetivamente observada caso a hipótese nula fosse verdadeira. Não obstante, é a partir dela – e infelizmente, em muitos casos, somente a partir dela – que o pesquisador tira suas conclusões a respeito da veracidade ou não da hipótese nula, muitas vezes realizando uma inferência do tipo $Pr(\theta = \theta_0 | \, T(x) = t)$, que para ser obtida dever-se-ia recorrer a métodos bayesianos. Assim, atualmente, o *p-valor* é, ao mesmo tempo, uma das estatísticas mais difundidas nas ciências sociais e biomédicas e uma das menos compreendidas (GOODMAN, 2008; ZILIAK, MCCLOSKEY, 2008a)[43].

O *p-valor,* no seu uso diário, tem sido erroneamente interpretado como: (i) a probabilidade de a hipótese nula ser verdadeira; (ii) a probabilidade de o resultado ter

[42] Por outro lado, Mayo (2005a) diria que o que queremos realmente saber é se a hipótese passa em testes severos.

[43] Para um *survey* amplo sobre a literatura empírica acerca das incompreensões em testes de hipóteses, ver Castro Sotos *et alii* (2007). Mais referências serão dadas na próxima seção.

ocorrido caso a hipótese nula fosse verdadeira; (iii) o tamanho do teste α; (iv) a probabilidade de se cometer um *erro;* (v) a indicação da importância científica do resultado, caso $p < \alpha$; (vi) a confirmação da inexistência de efeito, caso $p > \alpha$; (vii) a probabilidade de replicação dos resultados - entre outras concepções enganosas (BERGER, DELAMPADY, 1987; GOODMAN, 2008). A confusão com relação ao uso dos testes de hipóteses é completa e tão disseminada que, conforme Hubbard e Bayarri (2003, p.17), sua aplicação, na maior parte dos casos, é um mero ritual automático.

Pelo exposto nas seções anteriores, percebe-se como cada uma dessas interpretações é incorreta. Porém, para o presente trabalho, vale discutir um pouco mais acerca do *p-valor* como evidência, o que veremos na próxima seção. Trataremos também das noções de erro amostral e erro real, além da distinção entre diferença estatística e diferença substantiva. Estes pontos têm bastante relação com a confusão entre significância estatística e significância científica, prática ou econômica dos resultados, que, conforme, Goodman (2008, p.138), é o equívoco que

> [...] engloba todos os outros. É equivalente a dizer que a magnitude do efeito não é relevante, que a única prova relevante para uma conclusão científica está na experiência em questão, e que tanto as crenças quanto as ações fluem diretamente a partir dos resultados estatísticos. A evidência de um dado estudo tem de ser combinada com aquelas de trabalhos anteriores para gerar uma conclusão. Em alguns casos, uma conclusão cientificamente defensável pode ser que a hipótese nula é ainda provavelmente verdade, mesmo depois de um resultado significativo, e em outros casos, um valor não significativo para *p* ainda poderia levar a uma conclusão de que um tratamento funciona [...] para justificar as ações, devemos incorporar a gravidade dos erros decorrentes delas juntamente com a chance de que as conclusões estejam erradas.

2.2. *P-valor* como evidência, erro real e diferença substantiva

A evidência que um *p-valor* gera com relação à hipótese nula, contra ou a favor, não é algo absoluto. Isso pôde ser visto dentro do contexto dos testes de hipóteses clássicos, em que, enquanto Casella e Berger afirmavam que a hipótese nula a ser testada deveria ser aquela em que o pesquisador pouco acredita, Lehmann e Romano sugeriam que, quando o pesquisador acreditasse firmemente na hipótese nula, fosse então requerida evidência muito mais exigente para abandoná-la. Dentro do próprio contexto do método de Neyman-Pearson, o pesquisador teria de levar em conta diversos aspectos (como a gravidade do *erro tipo I* ou do *erro tipo II*, o tamanho da amostra, o poder alcançado

pelo teste, as magnitudes dos coeficientes envolvidos e as crenças com relação à hipótese) para considerar um resultado como "significante" ou não.

A esse respeito, Savage (*apud* BERGER, 1985, p.02), consegue, com um exemplo bem simples, ilustrar como a mesma evidência "objetiva" empírica, fornecida por um *p-valor*, é capaz de ser interpretada de maneira bastante diferente dependendo do contexto em que é utilizada. Considere três experimentos estatísticos.

1) Uma senhora, que adiciona leite ao seu chá, alega ser capaz de dizer se o chá ou o leite foram colocados na xícara primeiro. Em dez tentativas para testar sua alegação, ela acertou todas;

2) Um músico profissional alega ser capaz de distinguir uma página de uma peça de Haydn de uma página de uma peça de Mozart. Em dez tentativas para dizer de quem era a página, ele acertou todas;

3) Seu amigo bêbado diz que é capaz de prever o resultado do lançamento de uma moeda não viciada. Em dez tentativas que você lançou a moeda, ele acertou todas.

Nos três experimentos, defina θ como a probabilidade de cada um dos indivíduos acertar a predição que alega ser capaz de fazer. Perceba, neste caso, que a hipótese de nenhuma capacidade especial pode ser definida como $H_0: \theta = 50\%$, ou seja, a hipótese de que o sujeito seja tão bom em prever o evento quanto o acaso. Dessa forma, podemos testar $H_0: \theta = 50\%$ contra $H_1: \theta > 50\%$. Se aplicássemos um teste de significância estatística contra a hipótese nula, rejeitaríamos ao nível de significância unilateral de 2^{-10}. Nos três casos, haveria ***menos de 0,1% de chances*** de ocorrer uma série de 10 acertos. Todavia, é fácil perceber que este resultado ***não teria a mesma força como evidência*** para as três situações.

Com relação ao caso de número dois, é bastante provável que a evidência apresentada seja considerada prova quase irrefutável de sua alegação. Já com relação à situação de número três, a evidência, por mais impressionante que seja, provavelmente não seria interpretada como nada mais além de um grande lance de sorte. Nossa convicção com relação à hipótese nula de que $H_0: \theta_{amigo\_bêbado} = 50\%$ permaneceria inalterada. Por

fim, o primeiro caso provavelmente levaria a várias conclusões diferentes, a depender da plausibilidade e convicções prévias acerca da alegação da senhora[44].

Quando exposto desta forma, não é difícil perceber como o peso do *p-valor* é relativo; trabalhemos mais um pouco este tema contrastando-o com probabilidades *a posteriori* e discutindo sua *coerência.*

2.2.1. *P-valor* não é probabilidade *a posteriori*

Isto pode parecer um erro trivial, mas não é: a incorreta inversão da probabilidade do *p-valor* parece ser algo ***generalizado***[45]**.** Oakes (1986) encontrou, em pesquisa na Inglaterra, que 86% dos pesquisadores em psicologia confundiram o *p-valor* como a indicação da probabilidade de se cometer um erro ao se rejeitar a hipótese nula. Flak e Greenbaum (1995) repetiram o estudo em Israel e obtiveram resultado similar[46]. Haller e Krauss (2002), na Alemanha, chegaram à mesma conclusão. Outros estudos nos Estados Unidos (LINK, 2002; delMAS, 2007), na Argentina (RODRIGUEZ, 2006), na Espanha (CASTRO SOTOS *et alii*, 2009), e no Brasil (SEBASTIANI, VIALI, 2011) também verificaram a dificuldade de interpretação do *p-valor* em alunos universitários. Tendo em vista que a confusão entre $\Pr(H_0|\boldsymbol{X})$ e $\Pr(\boldsymbol{X}|H_0)$ parece ser bastante disseminada, o que atrapalharia o julgamento da evidência que a estatística *p-valor*

---

[44] Savage faz referência ao famoso experimento *"The lady tasting tea",* utilizado por Fisher em "*The Design of Experiments*" para ilustrar os diversos elementos por trás do desenho de um experimento e a realização de um teste de significância. O exemplo de Fisher se consistia, na verdade, em misturar 08 xícaras aleatoriamente, 04 com o chá colocado primeiro e 04 com o leite. Dessa forma, a chance de discernir as xícaras por sorte seria de 01 em 70 ou aproximadamente 1,4%. Segundo Salsburg (2001, p.08), o experimento de fato ocorreu e a senhora conseguiu discernir cada uma das xícaras. Neyman (1950) criticou a forma que Fisher escolheu para o experimento e modificou o problema no contexto de Neyman-Pearson, considerando que seriam apresentados à senhora n pares para experimentação, cada par com uma xícara de cada tipo. Neyman mostrou, então, que ao se considerar um teste $H_0: \theta = 50\%$ contra $H_1: \theta > 50\%$, poucas observações, como proposto por Fisher, resultariam em pouco poder para discrepâncias razoáveis como $\theta \in [60\%, 80\%]$. Para $n = 10$, por exemplo, e $\alpha = 0{,}0546$, ainda assim teríamos o poder para $\theta = 60\%$ igual a 0,1673, ***o que para Neyman pareceu poder bastante baixo para identificar a capacidade da senhora***. É interessante notar, também, que a mera rejeição da aleatoriedade, isto é $\theta = 50\%$, que é uma hipótese estatística, não implicaria na aceitação de uma hipótese alternativa do mundo real em particular (como a capacidade de a senhora discernir qual ingrediente foi colocado primeiro), pois, existem várias explicações consistentes com o resultado, tais como a trapaça (alguém haveria revelado quais eram as xícaras) ou a má elaboração do experimento (GIGERENZER 1993, NEYMAN, 1950).

[45] Segundo Kramer e Gigerenzer (2005, p.224), mesmo grandes matemáticos como d'Alembert e Leibniz já cometeram erros de interpretação de probabilidades devido às nuances de enumeração dos eventos favoráveis.

[46] Na verdade, os autores inclusive facilitaram o teste ao fazerem os alunos lerem o artigo de Bakan (1966), ***que avisa sobre estes problemas,*** antes de responderem o questionário.

fornece, é pertinente realizar breve discussão acerca do tema, ilustrando o quão discrepantes essas probabilidades podem ser.

Iniciemos com um exemplo de Cohen (1994, p 998-999), com relação à incidência de esquizofrenia. Suponha que queiramos testar a hipótese nula H_0: *o indivíduo é normal* contra a hipótese alternativa H_1: *o indivíduo tem esquizofrenia.* Ademais, suponha que exista um teste que tenha uma precisão para acusar um resultado positivo definida por $\Pr(positivo|H_1) = 0{,}95$. Isto é, dado que o indivíduo tenha esquizofrenia, o teste acusa positivo em 95% das vezes. Suponha, igualmente, que a precisão do teste para acusar um resultado negativo seja $\Pr(negativo|H_0) = 0{,}97$. Isto é, dado que o indivíduo seja normal, o resultado acusa negativo para esquizofrenia em 97% das vezes. Com essas informações, se tivermos um resultado positivo em mãos, o que inferir?

Como as chances de se ter um resultado positivo dado que o indivíduo seja normal é menor do que 3%, poderíamos, segundo os métodos usualmente estabelecidos, rejeitar a hipótese nula ao níveis de significância usuais. O raciocínio poderia ser feito tanto pela lógica disjunta de Fisher quanto pela lógica de tolerância de *erro tipo I* de Neyman-Pearson. Mas, isso quer dizer que a probabilidade de o indivíduo ter esquizofrenia seja igual, ou pelo menos similar, a 3%, como muitos usuários do *p-valor* interpretam? Ou que a probabilidade de cometermos um *erro tipo I*, quando o teste acusa positivo, seja de fato 3%? Não.

Acontece que a incidência de esquizofrenia na população de adultos é de cerca de apenas 2%. Ou seja, temos que, neste caso, as probabilidades *a priori* das hipóteses são $\Pr(H_0) = 0{,}98$ e $\Pr(\mathrm{H}_1) = 0{,}02$[47]. Dessa forma, dado que tivemos um resultado positivo, podemos calcular $\Pr(H_0|positivo)$ com uma simples inversão de probabilidade utilizando a regra de Bayes:

$$\Pr(H_0|positivo) = \frac{\Pr(H_0) * \Pr(positivo|H_0)}{\Pr(H_0) * \Pr(positivo|H_0) + \Pr(H_1) * \Pr(positivo|H_1)} \qquad (13)$$

[47] Mayo (2004, p.107; 2005b, p.812) argumentaria que este passo se trata de uma falácia, pois não se poderia dizer que, pelo simples fato de a distribuição de esquizofrênicos ser de 2% na população, esta seria a probabilidade *a priori* de se testar um esquizofrênico. Como dissemos anteriormente, tendo em vista o escopo da dissertação, não entraremos no debate acerca de como se definir adequadamente distribuições *a priori*. Para o argumento que segue, suponha-as válidas, a intenção aqui é contrastar as diferenças das medidas de evidência.

Assim, obtemos $\Pr(H_0|positivo) \approx 61\%$. Ou seja, a evidência do teste de hipótese que rejeitou H_0 a um nível de significância menor do que 5% – a princípio uma evidência bastante forte – torna-se uma evidência menos incisiva quando invertemos a probabilidade. Na verdade, de todos os indivíduos que testarem positivo, aproximadamente 61% ***não*** serão esquizofrênicos. Kramer e Gigerenzer (2005, p.224), relatam que, em uma pesquisa com médicos experientes da Alemanha, constatou-se a dificuldade dos profissionais em interpretar as probabilidades condicionais. Em um problema sobre um hipotético teste de câncer, em que $\Pr(\mathrm{H}_1) = 0{,}003$, $\Pr(positivo|H_1) = 0{,}50$, e $\Pr(negativo|H_0) = 0{,}97$, as respostas dos médicos para $\Pr(H_1|positivo)$ variaram entre 1% a 99%, sendo que metade das respostas ficaram em torno de 47% a 50%. Note que, neste caso, a resposta correta é em torno de 5% e, portanto, a resposta média foi em torno de dez vezes maior.

Voltando ao exemplo da esquizofrenia, o resultado da inversão de probabilidade significa que o teste realizado é inútil? Claro que não, pois sem o teste teríamos apenas 2% de chances de identificar um esquizofrênico e, após o teste acusar positivo, esta probabilidade passaria para algo próximo a 40%. O exemplo expõe, todavia, que ***a interpretação incorreta*** usualmente dada ao *p-valor* (isto é, considerá-lo como $\Pr(H_0|\boldsymbol{X})$) ***pode ser muito enganosa***, ainda mais quando se adotam níveis de significância arbitrários sem qualquer ponderação acerca de seu sentido no problema. Seguindo o exemplo exposto, seria interessante, portanto, fazer uma pergunta de modo mais geral: ao se observar um $p = 0{,}05$, que é um dos limiares mais adotados para se determinar a significância estatística de uma variável, qual seria, em condições gerais, a probabilidade de o pesquisador cometer um *erro tipo I na prática*[48]?

Vejamos, desta vez, com uma simulação. Para este exercício, defina uma variável aleatória proveniente de uma distribuição normal com média θ e desvio padrão 1 como $X_\theta \sim N(\theta, 1)$. Suponha que retiramos uma amostra aleatória de tamanho n de X_θ e que queiramos testar $H_0: \theta = 0$ contra $H_1: \theta \neq 0$. Suponha, também, que a proporção $\xi = 0{,}1, 0{,}2, (\ldots), 0{,}9$ diga respeito à quantidade de hipóteses nulas verdadeiras, isto é, a proporção de variáveis aleatórias $X_0 \sim N(0{,}1)$. Já o restante dos X_θ, provenientes de

[48] Note que agora nos referimos à ***probabilidade de uma hipótese rejeitada ser verdadeira*** (que é a interpretação errônea dada à probabilidade α ou ao *p-valor* que comumente também é confundido com α) e ***não*** à ***probabilidade de uma hipótese verdadeira ser rejeitada*** (que é a definição de α e que, como vimos, é propriedade de um teste e é diferente da variável aleatória *p-valor*, dependente dos dados).

distribuições em que $\theta \neq 0$, tem seus parâmetros distribuídos como $\theta \sim N(0,2)$. Para facilitar a interpretação, pode-se imaginar que estão sendo testados vários remédios diferentes e que ξ se refere à proporção dos que não funcionam. Para cada um dos medicamentos valeria a dicotomia freqüentista – ou a hipótese nula é verdadeira, ou é falsa. Mas, no conjunto, há aqueles que têm efeito e aqueles que não. Verificaremos se o grupo de remédios acusados como "significantes" a 5% são uma boa amostra daqueles que funcionam.

A simulação que iremos realizar, portanto, é a seguinte: dadas as condições acima, faremos *testes z* até que 2.000 destes atinjam o um *p-valor* entre 0,049 e 0,050[49]. Então veremos, dentre estes *p-valores*, quantos rejeitaram a hipótese nula quando ela era verdadeira. Isto nos daria certa "indicação frequentista" da probabilidade de, na prática, cometermos um *erro* quando nos deparamos com $p \approx 0{,}05$ e rejeitamos H_0, dadas as circunstâncias estabelecidas – seguindo a interpretação intuitiva, isto nas daria uma ideia de quantos ***remédios sem qualquer efeito*** foram ***acusados falsamente como eficientes***. Os resultados, para vários valores de ξ, considerando-se $n = 100$, são apresentados nas duas primeiras colunas da tabela 3.

Tabela 3 – Falsos positivos $\mathbf{p \approx 0{,}05}$

$n = 100$		$\xi = 0{,}1$	
Proporção de nulas (ξ)	**Falsos positivos (%)**	**Tamanho amostral (n)**	**Falsos positivos (%)**
10,0	25,0	50	19,2
20,0	39,8	100	25,0
30,0	57,4	150	26,9
40,0	65,8	200	31,7
50,0	75,7	250	34,8
60,0	80,6	300	35,0
70,0	87,2	1000	51,1

Fonte: *applet* disponível em http://www.stat.duke.edu/~berger/applet2/pvalue.html.

Façamos agora a simulação com a proporção de nulas fixa em $\xi = 0{,}1$ – ou seja, a hipótese nula é *a priori* muito improvável, o que reduziria bastante a probabilidade de cometer um *erro* na rejeição de H_0 – mas variando o tamanho amostral utilizado para o teste, iniciando com $n = 50$ até $n = 1000$. Os resultados são apresentados nas duas últimas colunas da Tabela 3.

[49] Simulações realizadas com *applet* de James Berger.

Disponível em http://www.stat.duke.edu/~berger/applet2/pvalue.html. Acesso em 17/02/2012.

Perceba que, em todas as simulações, o conjunto de testes com $p \approx 0{,}05$ contém maior proporção de hipóteses nulas verdadeiras do que a proporção ξ estabelecida para o universo. Isto dá um indício de quão temerário pode ser respaldar-se unicamente no *p-valor* para uma decisão acerca da veracidade ou não da hipótese nula: nas simulações em tela, o resultado foi pior do que aquele esperado naturalmente pelo acaso. Tomando a proporção $\xi = 0{,}5$ como exemplo, a simulação terminou com 75,7% falsos positivos, enquanto que o real valor de nulas na população era de 50%. Escolheu-se apresentar a simulação por ser mais elucidativa. Mas, é possível demonstrar analiticamente que este resultado é típico para vários valores de ξ ou p, em condições gerais (BERGER, 2003; BERGER, DELAMPADY, 1987; BERGER, SELKE, 1987; SELKE, BAYARRI, BERGER, 2001).

Supondo $\xi = 0{,}5$, por exemplo, poderíamos pensar em alterar as condições iniciais do experimento para tentar favorecer os resultados dos testes de significância; entretanto, como colocam Berger e Selke,

> [...] mesmo uma análise Bayesiana bastante enviesada para H_1 nos diz que a hipótese nula ***tem 22,7% de probabilidade de ser verdade***, evidência contra a nula que não seria considerada por tantas pessoas como sendo muito forte (BERGER, SELKE, 1987, p. 113, grifo nosso).

Ter-se-ia, aqui, uma proporção de nulas no conjunto em que $p \approx 0{,}05$ menor do que a proporção de nulas no universo – tal qual no exemplo da esquizofrenia – no entanto, ainda longe do que se poderia imaginar com uma interpretação equivocada do *p-valor*. As chances reais de um *erro* seriam cerca de quatro vezes e meia maiores do que os 5% nominais do nível de significância. Percebe-se, por conseguinte, (i) que a utilização de um mesmo *p-valor* de cerca de 5% como critério de rejeição de H_0 pode chegar a proporções de falsos positivos bastante diferentes – em nossas simulações variando desde 19,2% até 96,3% – dependendo do ***contexto a priori*** da probabilidade de H_0 e do ***tamanho da amostra*** utilizada; e que, portanto, (ii) a discrepância entre o *p-valor* e a probabilidade *a posteriori* de H_0 pode em algumas circunstâncias ser intolerável.

Para finalizar esta discussão, uma pergunta pertinente seria qual a probabilidade *a priori* que teria de ser dada à hipótese nula para que tenhamos $\Pr(H_0|\boldsymbol{X}) = 0{,}05$ quando observamos $t = 1{,}96$? Isto é, qual a probabilidade que o pesquisador teria de atribuir a H_0 para que, quando se realizasse a inversão, o *p-valor* fosse semelhante à probabilidade *a posteriori* da hipótese nula? Surpreendentemente, a resposta para tal

questionamento, em condições razoavelmente gerais, "[...] é que se deve dar a H_0 uma ***probabilidade inicial de 15%*** e então espalhar a massa de 0,85 (dada a H_1) de uma maneira simétrica que mais favoreça H_1" (BERGER, SELKE, 1987, p. 113, grifo nosso)[50]. Ou seja, dentro de um paradigma Bayesiano de inversão de probabilidade, a evidência que o *p-valor* fornece contra ou a favor de uma hipótese seria praticamente dominada pelas crenças estabelecidas anteriormente pelo pesquisador. A tal ponto que leva os autores a concluir que

> [...] este viés flagrante para H_1 seria dificilmente tolerado em uma análise Bayesiana; mas o pesquisador que quiser rejeitar não precisa ***parecer*** tão viesado – ele pode somente observar que $p = 0{,}05$e rejeitar pela "prática padrão" (BERGER, SELKE, 1987, p. 113).

Em outras palavras, a definição do *p-valor* como dada na equação (01) é objetiva, mas, no contexto real de sua aplicação, parece que isto se perde. Certamente não seria objetivo dizer ou pensar que a probabilidade de a hipótese nula ser verdade é de apenas 5% quando se observa $t = 1{,}96$, sem ao menos esclarecer o fato de que grande parte da evidência não provém da observação em si, mas sim de se ter dado uma probabilidade *a priori* ***bastante baixa*** para a hipótese nula.

Como visto em Casella e Berger (2002), os autores recomendavam que o pesquisador estabelecesse a hipótese de pesquisa, isto é, a hipótese em que o pesquisador acredita, como H_1 e, consequentemente, a hipótese em que o pesquisador não acredita, isto é, aquela "menos provável", como H_0. Isto fica ainda mais evidente na seguinte passagem dos autores, oriunda da discussão do texto de Berger e Delampady (1987) acerca do *p-valor* como evidência:

> [...] o propósito de um experimento é comumente contradizer H_0 e os pesquisadores não realizariam experimentos que acreditariam, a priori, falhar 50% das vezes. Nós ficaríamos surpresos se a maioria dos pesquisadores colocassem até mesmo 10% na probabilidade a priori de H_0 (CASELLA, BERGER, 1987b, p. 345).

Todavia, isto não é consenso, e, por exemplo, Lehmann e Romano (2008), em seu livro-texto clássico, inclusive recomendaram que a hipótese nula pudesse ser algo em que o pesquisador acreditasse firmemente e, portanto, exigisse evidência bastante crítica para abandoná-la. Deste modo, a abordagem de Casella e Berger parece não ser transparente quanto ao papel do *p-valor* como evidência, a não ser que, no contexto da aplicação de

[50] Para exemplos em que as probabilidades condicionais ***coincidem*** "naturalmente", ver DeGroot (1973) ou Casella e Berger (1987a).

testes de hipótese, o pesquisador explicitasse que grande parte da conclusão contra a hipótese nula não provém dos dados, mas sim de crenças *a priori* "menores do que 10%", o que quase nunca ocorre.

Em suma, ***a interpretação equivocada*** usualmente atribuída ao *p-valor* contém um viés grande e quase sempre não entendido ou não revelado. Note que esta aplicação está mais ligada à filosofia indutiva de Fisher do que ao contexto de teoria da decisão de Neyman-Pearson, uma vez que, neste último caso, os riscos da regra de decisão deveriam ser ponderados diante das circunstâncias da aplicação e, assim, o *p-valor* em si não passaria de uma estatística dentro ou fora da região crítica. Entretanto, quando se dá um papel epistêmico ***unicamente*** ao *p-valor*, é preciso ter cautela, pois este não tem o mesmo peso como evidência independentemente das circunstâncias e, definitivamente, não pode ser interpretado como a probabilidade *a posteriori* de H_0.

2.2.2. *P-valor* não é uma medida "coerente" de evidência

Schervish (1996), em breve artigo, apresentou como a interpretação do *p-valor* como medida de evidência pode levar a inconsistências lógicas. Suponha que uma hipótese H implique na hipótese H′, isto é, $H \rightarrow H'$. Uma medida de evidência coerente para H′ teria de ser ***tão grande ou maior*** do que uma medida de evidência para H, pois a rejeição de H′ implica logicamente na rejeição de H ($não\ H' \rightarrow não\ H$). O *p-valor* ***não satisfaz*** a este critério de coerência. Schervish traz um exemplo simples, um teste de média para a distribuição normal. Ao se observar $x = 2.18$, o *p-valor* para a hipótese $H_0: \mu \in [-0.5, 0.5]$ é igual a 0.0502 e para a hipótese $H_0': \mu \in [-0.82, 0.52]$ é igual a 0.0498. Note que, neste caso, rejeitar H_0' implica logicamente em rejeitar H_0. Pois se a média populacional não está no intervalo $[-0.82, 0.52]$, então ela também não poderia estar no intervalo $[-0.5, 0.5]$, já que trivialmente $[-0.5, 0.5] \subset [-0.82, 0.52]$. Não obstante, se o limiar de 5% fosse adotado, H_0' seria rejeitada enquanto H_0 não o seria, o que é uma ***contradição lógica.***

Patriota (2012, p.04-05) fornece outro exemplo interessante. Suponha uma amostra aleatória, $n = 100$, *i.i.d*, de uma distribuição normal bivariada, com médias μ_1 e μ_2, com uma matriz identidade de variância-covariância, cujas médias amostrais tenham resultado em $\overline{x_1} = 0{,}14$ e $\overline{x_2} = -0{,}16$. Ao se calcular o *p-valor* da estatística de *Wald* da hipótese nula $H_0': \mu_1 = \mu_2$, obtém-se $p = 0{,}03$ – valor usualmente considerado

evidência bastante forte. Já se o pesquisador resolvesse testar se $H_0: \mu_1 = \mu_2 = 0$, obteria $p = 0{,}10$. Note que $H_0 \rightarrow H_0'$; logo, não seria possível rejeitar H_0' sem que também se rejeitasse H_0. Deste modo, um nível arbitrário de significância para a decisão levaria a uma contradição lógica. Perceba que, teoricamente, não há contradição no cálculo do *p-valor*, pois, como bem ressalta Patriota (2012, p.05, grifo nosso) "[...] p-valores são apenas probabilidades de ***encontrar estatísticas não observadas tão grandes quanto as observadas***, a conclusão conflitante apresentada no exemplo acima não é uma contradição lógica do método frequentista". A contradição decorre da interpretação do *p-valor* como uma medida absoluta de evidência[51].

A conclusão de Schervish é a de que o único modo coerente de interpretar o *p-valor* como evidência seria em casos em que ***não há nenhuma comparação*** entre hipóteses e, mesmo assim, este valor ***deveria ser relativizado*** de acordo com a hipótese formulada. Vale aqui trazer outro exemplo simples dado pelo autor. Ainda no contexto da distribuição normal, suponha que se tenha observado $x = 2.7$. O *p-valor* para $H_0: \mu = 0.9$ é $p = 7{,}18\%$ enquanto que para $H_0': \mu \leq 1$ teríamos $p = 4{,}46\%$. Evidentemente que o dado observado dá mais suporte à hipótese H_0', muito embora seu *p-valor* seja menor.

***

Como vimos, há um desconhecimento a respeito dos métodos de inferência para testes de hipóteses e suas respectivas medidas de erro e de evidência. Particularmente, é comum se ignorar o peso que o *p-valor* ou a rejeição de uma hipótese nula têm diante de diferentes circunstâncias, o que leva à incompreensão acerca do real significado de um resultado "estatisticamente (*in*)significante". No fim, isso acaba por levar à ideia de que um resultado para ser importante teria de ser, primeiramente, estatisticamente significante (aos níveis usuais) ou, ainda, à ideia de que um resultado estatisticamente significante (aos níveis usuais) é, por conseguinte, importante. Entretanto, a significância estatística (aos níveis usuais) ***não é condição necessária, tampouco condição suficiente*** para a importância científica de um resultado.

[51] Patriota (2012) também busca criar uma medida freqüentista que respeita este critério de "coerência".

Como isso se manifesta nos trabalhos empíricos? Vejamos um exemplo retirado do *AER* por Ziliak e McCloskey (2008a, p. 34), que é bem característico[52]:

> [...] o coeficiente é significante ao nível de confiança de 99 por cento. Nem o choque monetário nem todos os 12 coeficientes como um grupo são estatisticamente diferentes de zero. O coeficiente de *c* é negativo e significante e a defasagem distribuída de *c* também é significante. Na coluna (2) nós mostramos uma regressão que omite as defasagens insignificantes do choque monetário. Agora, a defasagem distribuída de *c* é significante a 1 por cento [...] nós interpretamos isso como indicando que o fator primário determinando variações cíclicas na probabilidade de deixar o desemprego é provavelmente a heterogeneidade [...] entretanto, choques monetários não têm impacto significativo.

Note que as palavras significante e significativo são utilizadas com sentidos ambíguos. O que quer dizer "choques monetários não têm impacto significativo"? Que o efeito dos choques monetários inexiste? Que o efeito é negligenciável? ***Stricto sensu***, "estatisticamente significante" quer dizer que: dado que a ***única*** fonte de erro fosse a ***variação amostral***[53], então, ***caso o verdadeiro coeficiente fosse zero***, a probabilidade de se obter um coeficiente estimado ***tão grande ou maior*** do que o efetivamente estimado seria menor do que 5%. É evidente que isso ***não responde se o coeficiente tem impacto significativo ou não no sentido usual do termo***. Analogamente, "não ser estatisticamente diferente de zero" quer dizer que, dado que a ***única*** fonte de erro fosse a ***variação amostral***, então, ***caso o verdadeiro coeficiente fosse zero***, a probabilidade de se obter um coeficiente estimado ***tão grande ou maior*** do que o efetivamente estimado seria maior do que 5%. É ainda mais evidente que isso não nos diz se o impacto do choque monetário ***é relevante ou importante***. Perceba que não há qualquer menção ao tamanho do efeito bem como sua relação com os riscos da inferência que se realiza[54].

---

[52] Talvez a forma mais simples de descrever o fenômeno como aparece nos textos seja a seguinte: o pesquisador observa se o pacote econométrico coloca os asteriscos nos coeficientes da regressão. Aqueles que não têm asterisco são considerados zero. Aqueles que têm asterisco são importantes e o valor considerado para análise é apenas o próprio valor estimado, sem qualquer preocupação com o próprio erro amostral.

[53] Ou seja, o modelo está corretamente especificado, estatisticamente adequado, não há erros de mensuração, a amostra é aleatória etc.

[54] Interessante notar o que Tversky e Kahneman (1971, 1974) denotaram por "lei dos pequenos números". Os pesquisadores têm a expectativa de que "[...] uma hipótese válida sobre uma população será representada por um resultado estatisticamente significante na amostra – sem qualquer preocupação com seu tamanho. Como consequência, os pesquisadores depositam muita fé em resultados de amostra pequena e superestimam grosseiramente a replicabilidade de tais resultados" (TVERSKY, KAHNEMAN, 1974, p. 1126).

Nesta seção iremos discutir brevemente a confusão entre um resultado estatisticamente significante e um resultado importante (ou significante no sentido cotidiano do termo). Trabalharemos dois pontos: (i) as diversas outras fontes de erro que podem existir e que acabam sendo ignoradas pela busca por resultados significantes; e, (ii) o sentido de se testar hipóteses exatas quando sabemos que modelos não são cópias perfeitas da realidade, bem como a consequente confusão entre diferença estatística e diferença "substantiva" *stricto sensu*. Por fim, exporemos brevemente alguns métodos que possam indicar como amenizar os problemas de inferência tratados.

2.2.3. Erro amostral ou erro real

Como expuseram Ziliak e McCloskey (2008a, p. 07), "[...] a significância estatística não está preocupada com nenhuma de uma longa lista de fontes de erros não amostrais"; entretanto, parece haver uma constante busca por resultados "estatisticamente significativos" – não é incomum o pesquisador rodar várias regressões com várias especificações e covariadas diferentes até obter o "asterisco" na variável de interesse – como se isso fosse validar o resultado empírico encontrado. Ocorre que todas as demais fontes de erros que possam estar sendo ignoradas talvez sejam mais importantes do que os erros decorrentes da variação amostral.

Leamer (1983) trata de maneira intuitiva este ponto. Poderíamos decompor a variabilidade de um estimador $\hat{\beta}$ em dois componentes, um decorrente da variação amostral (S), e outro decorrente dos demais erros, como a má especificação do modelo (M). Isto é:

$$Var(\hat{\beta}) = S + M \tag{14}$$

Para n grande, quando a incerteza amostral (S) se torna pequena comparada com a incerteza da especificação (M), seria hora de o pesquisador buscar outras formas de evidência. O autor traz uma analogia simples e ilustrativa (1983, p.33 -34):

> [...] suponha que eu esteja interessado em medir a largura de uma moeda e eu entregue réguas para uma sala de voluntários. Após cada voluntário reportar sua medida, eu calculo a média e o desvio padrão, e concluo que a moeda tem largura de 1,325 milímetros com erro padrão de 0,013. Uma vez que esta quantidade de incerteza não me agrada, eu proponho encontrar três outras salas cheias de voluntários, multiplicando assim a amostra por quatro e dividindo o erro padrão pela metade. Isso é uma forma tola de conseguir uma medida mais precisa, porque já alcancei o ponto em que a incerteza amostral

> S é pequena comparada com a incerteza da má-especificação M. Se eu quero aumentar a verdadeira precisão da minha estimativa, é hora de considerar o uso de um micrômetro.

Tragamos alguns casos da teoria econômica. Leamer, ainda em seu texto de 1983, cita o exemplo do efeito de penas de morte sobre a taxa de homicídios. O autor listou 14 variáveis dependentes que poderiam ser utilizadas como controles, sendo incluídas ou não na regressão a depender das crenças prévias do econometrista. Combinações diferentes das covariadas poderiam resultar desde uma estimativa de que uma execução adicional ***deteria***, na média, quase 29 homicídios, até uma estimativa de que uma execução adicional ***aumentaria,*** na média, 12 homicídios. Diante disto, o autor concluiu que "[...] qualquer inferência, com estes dados, sobre o efeito dissuasivo da pena de morte é muito frágil para ser acreditada" (LEAMER, 1983, p. 42).

Mais recentemente, estudos continuaram divergindo com relação ao efeito dissuasivo da pena de morte. Alguns têm encontrado efeitos altos, enquanto outros, nenhum ou, ainda, efeitos ambíguos[55]. Trabalho recente de Durlauf, Fu e Navarro (2012) busca, deste modo, verificar como a ***incerteza quanto ao modelo*** conduz a estes resultados conflitantes. Reproduzimos aqui uma figura apresentada pelos autores, que exibe de maneira clara o quão discrepantes os resultados podem ser a depender do modelo utilizado.

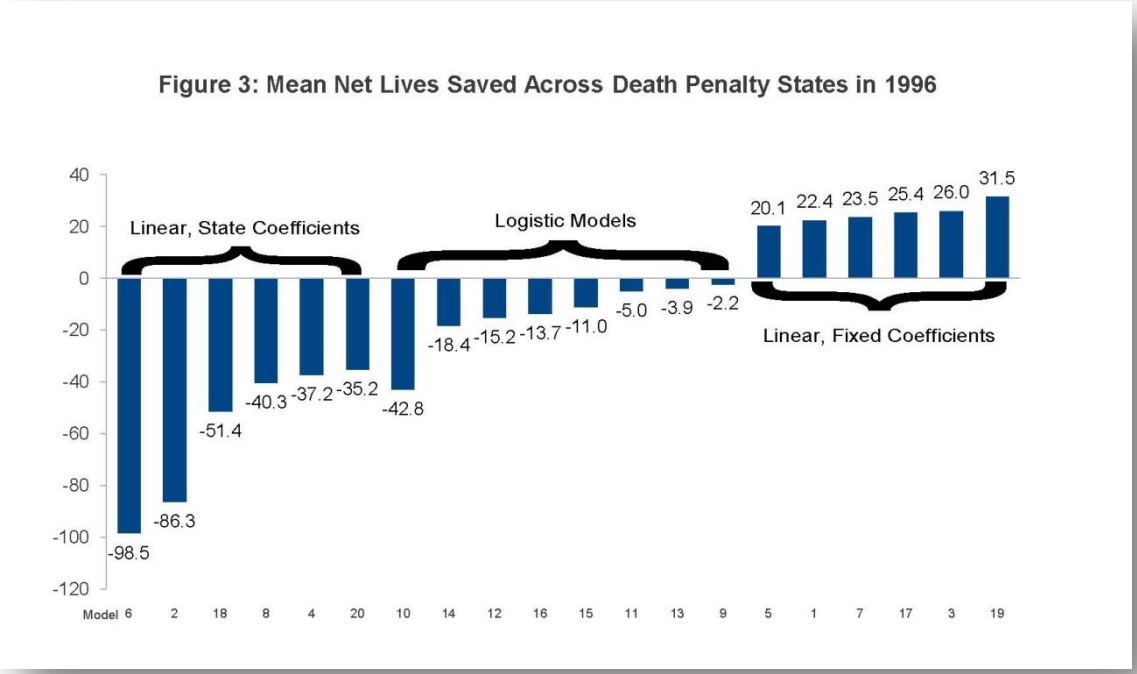
No caso apresentado, a escolha entre um modelo linear ou logístico, com coeficientes fixos ou individuais para cada estado dos Estados Unidos, faz com que as estimativas variem desde -98,5 (o que indicaria que penas capitais aumentam a criminalidade[56]) até 31,5 (o que indicaria que penas capitais diminuem a criminalidade), com o mesmo conjunto de dados.

---

[55] Dezhbakhsh, Rubin e Shepherd (2003), Zimmerman (2004), Donohue e Wolfers (2005), Durlauf, Navarro e Rivers (2010) e Shepherd (2005).

[56] Como os próprios Durlauf, Fu e Navarro (2012, p.21) explicam, este resultado pode ser abrangido pela teoria econômica de decisão racional. Um indivíduo que sabe que enfrentará a pena máxima por um assassinato não teria incentivos para reduzir a violência e poderia racionalmente escolher matar as testemunhas ou outras pessoas envolvidas no crime. Outra explicação para este resultado pode ser o efeito "brutalidade", em que a pena capital de certo modo legitimaria o assassinato, tornando-o moralmente menos custoso.

Figura 1 – Incerteza nos efeitos dissuasivos da pena de morte



Fonte: Durlauf, Fu e Navarro (2012, p. 28)

Percebe-se que a incerteza com relação ao modelo suplanta – e muito – uma possível incerteza com relação à variação amostral. A bem da verdade, para poder se avaliar com segurança a variabilidade amostral, seria preciso primeiramente se acordar em relação a um modelo razoavelmente satisfatório. Em uma situação como essa, procurar um modelo que resulte em estimativas "estatisticamente significativas" com "sinais corretos" seria, decerto, algo fora de propósito. É importante deixar claro que ***não*** se quer dizer aqui que o erro amostral deva ser ignorado; a questão é que, como expôs Leamer (2010, p. 37), "[...] uma cultura que insiste em estimativas estatisticamente significativas não é, naturalmente, receptiva a outra razão pela qual os nossos dados não são informativos".

Outro exemplo que podemos trazer é o debate em torno dos determinantes da diferença de renda entre países, cujo principal artigo talvez seja o de Acemoglu, Johnson e Robinson (2001). A discussão na literatura gira em torno da comparação de diferentes aspectos, como institucionais, culturais, geográficos, de política econômica, de capital humano – entre outros – para explicar a disparidade de renda *cross-section* observada

no mundo[57]. Em geral, a significância estatística da variável de interesse figura como um dos principais argumentos empíricos, chegando a afirmações ousadas como: "[...] nós apresentamos evidência de uma regressão que mostrou que, uma vez controlado o efeito das instituições econômicas sobre o PIB per capita, ***variáveis geográficas*** [...] ***não têm qualquer poder explicativo para a prosperidade atual***" (ACEMOGLU, JOHNSON, ROBINSON, 2004, p.28, grifo nosso).

A evidência a que os autores se referem trata-se de uma regressão linear por variáveis instrumentais em que o coeficiente de um índice que representaria as instituições se mostrou estatisticamente significante enquanto os coeficientes de variáveis geográficas, como a incidência de malária, não. Será que é realmente plausível que fatores geográficos tenham ***exatamente nenhum efeito direto*** sobre o desempenho econômico? E será que os dados fornecem respaldo a essa afirmativa? Discutiremos isto logo a seguir, na próxima seção. Antes, contudo, cabe ressaltar que há outras fontes de erro que são negligenciadas – e que talvez sejam mais sérias – como erros de especificação, a dependência de resultados assintóticos e, até mesmo, a própria definição das variáveis utilizadas. Quanto a este último ponto, Glaeser *et alii* (2004, p.13) ao analisarem as medidas que representariam "instituições" concluem que estas

> [...] não podem ser usadas para estabelecer causalidade. Essas medidas não são construídas para refletir restrições nos governos ou características permanentes do cenário político. Ao invés disto, elas são altamente voláteis e revertem à média. São pouco correlacionadas com as medidas objetivas disponíveis de restrições constitucionais aos governos. Ainda assim, são essas as variáveis utilizadas para mostrar que instituições causam crescimento.

Em um cenário como este, é difícil entender a ênfase na significância estatística como argumento empírico, a não ser se derivada de uma incompreensão sobre seu real significado. Estes exemplos ressaltam a pertinência do posicionamento de Deming,

> [...] na minha prática, eu prontamente me recuso a calcular ou discutir a interpretação do erro padrão quando grandes erros operacionais não amostrais estão obviamente presentes [...] é possível que um resultado seja útil e ainda possua um amplo erro padrão. Um resultado obtido por definições e técnicas que têm sido elaboradas com cuidado, e realizada por entrevista e supervisão excelentes, pode ter um amplo erro padrão porque a amostra era pequena; todavia, esse resultado pode ser bem preferível a outro obtido com uma maior amostra, com um menor erro padrão, mas cujas definições, técnicas e entrevistas estavam fora de linha com as melhores práticas e conhecimento do assunto (DEMING, 1961, p. 55-57).

[57] Por exemplo, Acemoglu, Johnson e Robinson (2001), Easterly e Levine (2003), Rodrik, Subramanian e Trebbi (2002), Sachs (2003), Gundlach e Carstensen (2006).

2.2.4. Diferença estatística ou diferença substantiva

Para iniciar a discussão acerca deste ponto, cabe colocar um paradoxo levantado por Berkson, ainda em 1938, quando os testes de significância estatística estavam sendo difundidos por Fisher. Berkson inicia sua exposição afirmando que qualquer estatístico experiente, que tenha aplicado muitos testes qui-quadrado de grau de ajuste, deverá ter percebido que, quanto maior o número de dados disponíveis, menores os *p-valores* tendem a sair. Tomando a curva normal como exemplo, afirma Berkson que, para qualquer que seja a variável utilizada, se o número de observações for extremamente grande – como, digamos, 200.000 – o *p-valor* de ajuste dos dados à curva normal, em geral, será menor do que qualquer limite usual de significância. Isto se daria, pois, conforme o autor,

> [...] ***podemos presumir que é praticamente certo que qualquer série de observações do mundo real não segue exatamente uma curva normal com absoluta exatidão em todos os aspectos***, e não importando o quão pequena seja a discrepância entre a curva normal e a curva de observações, o *p-valor* do qui-quadrado será pequeno se a amostra tiver um número suficientemente grande de observações. E eu suponho que seja consenso entre os estatísticos que uma amostra maior é sempre melhor do que uma amostra menor. Se, então, sabemos de antemão qual o *p-valor* que será resultado da aplicação de um teste qui-quadrado em uma amostra grande, não parece ter uso algum aplicá-lo em uma amostra pequena (BERKSON, 1938, p.526-527, grifo nosso).

Isto é, como já se sabe que, quando $n \to \infty$, o teste eventualmente rejeitará o ajuste dos dados à curva normal, então a aplicação a qualquer n não seria sequer um teste! Ao realizar um procedimento deste tipo, o pesquisador teria de ter ciência de que o seu modelo não passa de uma ***aproximação*** da realidade. A não ser que Deus ou qualquer outro ser onisciente tenha revelado quais as relações exatas prevalecentes na natureza, o modelo estabelecido para se realizar o teste ***é necessariamente aproximado,*** e, dessa forma, ***rejeitá-lo estatisticamente a um nível arbitrário de significância*** seria, na maior parte dos casos, apenas uma questão de tamanho da amostra[58].

Deste modo, o teste de uma hipótese precisa ou exata tem de ser feito com cautela, principalmente quanto à resposta que busca. Pois, muito provavelmente, a hipótese, por

[58] Fora da econometria, Kydland e Prescott (1982, p. 1360), por exemplo, deixam isso claro em seu modelo de *Real Business Cycle* ao afirmar que escolheram "[...] não testar o modelo contra o modelo menos restritivo de vetores auto-regressivos. O resultado mais provável é que o modelo seria rejeitado, tendo em vista os problemas de medida e a natureza abstrata do modelo. Nossa abordagem é focar em certas estatísticas em que os ruídos introduzidos por aproximações e erros de medida sejam provavelmente pequenos [...]".

natureza, ***é falsa;*** e, se a questão for constatar ou não sua veracidade – uma mera questão de existência – não rejeitá-la seria, ***necessariamente,*** cometer um *erro tipo II*. Como afirma Greene (2000, p. 227 grifo nosso), "[...] é um erro pensar os testes t ou F como testes sobre a verdade. Nós sempre sabemos a resposta correta à pergunta '*um coeficiente zero é uma hipótese razoável?* ' A resposta é não". Por exemplo, suponha uma amostra aleatória i.i.d, $\boldsymbol{X} = (X_1, \dots, X_n)$, proveniente de uma distribuição normal $N(\theta, 1)$. Ao se conduzir um teste de hipótese de tamanho $\alpha = 0{,}05$ de $H_0: \theta = 0$ contra $H_1: \theta \neq 0$, se obtermos $\sqrt{n}\bar{x} \geq 1{,}96$ rejeita-se H_0. Todavia, é bastante improvável que a hipótese nula seja exatamente verdade; na maioria dos casos, principalmente em ciências sociais, $\theta \neq 0{,}00000000000000000000$ (...) ***com certeza***[59]. Supondo-se que $\theta = 10^{-10}$ e que, ***para fins práticos, tal valor seja considerado irrelevante***, se tomarmos uma amostra muito grande, como $n = 10^{24}$, então a probabilidade de se rejeitar H_0 será bastante alta, ***muito embora a verdadeira média seja negligenciável***. E

> [...] este mesmo fenômeno existe não importando qual o tamanho de $\alpha > 0$ seja escolhido, e não importando quão pequena seja a diferença $\varepsilon > 0$ entre a média verdadeira e zero. Para uma amostra grande o suficiente, o teste clássico irá com virtual certeza rejeitar [a hipótese nula] (Berger, 1985, p. 20).

Isto é, se a intenção do teste é apenas constatar uma diferença, qualquer que seja, basta ter n grande. É possível verificar esta preocupação em vários autores, como Kadane (2011, p. 438):

> [...] quando o tamanho da amostra é pequeno, poucas hipóteses nulas são rejeitadas, enquanto que, quando o tamanho da amostra é grande, quase todas são rejeitadas [...], portanto, enquanto que o teste de significância pretende abordar (em certo sentido) se $\mu = 0$, de fato a aceitação ou rejeição da hipótese nula tem muito mais a ver com o tamanho da amostra do que com a extensão em que a hipótese nula é uma boa aproximação da verdade.

Leamer (1983, p.39),

> [...] testes de diagnósticos [...] sem hipóteses alternativas explícitas, são inúteis, uma vez que, se o tamanho da amostra é grande o suficiente, qualquer hipótese será rejeitada (por exemplo, nenhuma distribuição é exatamente normal). Tais testes, portanto, degeneram-se em rituais elaborados para medir o tamanho efetivo da amostra.

Ou até no livro de introdução à econometria de Maddala (2003, p.18)

---

[59] Evidentemente que, em algumas circunstâncias, podem existir situações em que a hipótese nula $H_0: \theta = 0$ seja desejável e plausível, como talvez, para os céticos, no caso de testes de percepção extra-sensorial. Todavia, em decorrências de perturbações, erros de medida, erros de especificação entre outros, mesmo estes casos dificilmente sobreviveriam a um teste de hipótese quando $n \to \infty$.

> [...] se o tamanho da amostra for grande o bastante, podemos rejeitar todas as hipóteses nulas. Isso é geralmente o que ocorre com os que usam grandes conjuntos de séries *cross-section* com milhares de observações. Quase todo coeficiente é significante ao nível de 5%.

Ocorre que o resultado de um teste de significância de uma hipótese que se sabe ser falsa, a um nível arbitrário, não responde ao pesquisador se a hipótese é ***aproximadamente correta*** do ponto de vista científico**.** Não obstante, por algum motivo, os testes rotineiramente aplicados são sobre hipóteses do tipo $H_0: \beta = 0$ ou $H_0: X \sim Normal$.

Vejamos um exemplo da teoria econômica levantado por Ziliak e McCloskey (2008a, p. 94-97): a teoria paridade do poder de compra (PPP). Tomando os Estados Unidos como base, em geral o modelo utilizado para testar a PPP é da seguinte forma:

$$P_{eua} = \beta_0 + \beta_1 e/P^* + u \qquad (15)$$

Onde P_{eua} representa o índice de preços dos Estados Unidos, e representa a taxa de câmbio com um país exterior, P^* o índice de preços deste país e u o termo de erro. A PPP é derivada da lei do preço único que, no mundo real, é apenas aproximadamente válida, tendo em vista a existência de custos de transação, bens não transacionáveis, bens diferenciados, investimentos fixos entre outros fatores. Além disso, cabe enfatizar que a PPP é medida com o uso de índices de preços, que também não são calculados de forma exatamente igual para todos os países.

Em um mundo em que o modelo fosse representação exata de realidade, teríamos que $\beta_1 = 1$. No mundo real, não necessariamente β_1 será exatamente igual a 1. Mas não seria algo raro ver o pesquisador "testar" ***a validade do modelo*** com a hipótese $H_0: \beta_1 = 1$. Se não rejeitasse, diria que o modelo é boa aproximação da realidade; já se rejeitasse, diria que o modelo falha em explicar a realidade. Note, no entanto, que o teste é de algo ***que se sabe ser falso*** e apenas uma estatística t ou um *p-valor* não responderão se a PPP é adequada do ponto de vista econômico. Se, com uma amostra enorme, obtivermos o valor estimado $\hat{\beta}_1 = 0{,}9999$, com erro padrão muito baixo, o teste rejeitará H_0, com uma estatística t alta e $p < 0{,}000(\ldots)$. É, todavia, 0,9999 próximo o suficiente de 1? Estatisticamente, ao nível de significância de 5%, 1%, ou o valor mais próximo de p, não é, pois a estimativa é precisa e, por isso, a hipótese nula foi rejeitada. Já quanto à proximidade necessária para validar ou invalidar a teoria do ponto de vista econômico,

isto é uma questão que ***cabe ao pesquisador***, e não ao *software* econométrico, ***responder***. Se, neste caso, parece óbvio que talvez não devêssemos rejeitar que a PPP seja válida simplesmente porque β não é exatamente igual 1,000000(...), então o mesmo senso comum deveria ser aplicado – com maior dificuldade de ponderação, certamente – para casos com maior variabilidade amostral, em que $\hat{\beta}_1 = 0{,}75$ ou em que $\hat{\beta}_1 = 1{,}62$.

Por mais simples que pareça este conceito quando exposto desta forma, testes que parecem considerar hipóteses exatas, ***literalmente***, não são incomuns na literatura, como constatam Keuzenkamp e Magnus (1995, p.11):

> [...] mesmo os melhores periódicos, como o Journal of Econometrics, reportam testes da paridade do poder de compra ou de mercados perfeitamente eficientes, muito embora saibamos que estas teorias não são literalmente verdade. Não seria muito mais interessante, em tais casos, medir o quão próximo o mundo real está do mundo ideal destas teorias?

Ou ainda Delong e Lang (1992, p.1269, grifo nosso)

> [...] a maior parte dos economistas não age como se soubessem que suas hipóteses são falsas e estivessem apenas procurando estabelecer sua qualidade como aproximações. A prática da econometria sugere que os economistas tomam suas hipóteses a sério. Como um exemplo, lembramos que a literatura sobre "raiz unitária" tem visto uma grande quantidade de esforço dedicado para determinar a distribuição assintótica da estatística de teste sob a hipótese nula e testar a hipótese nula de que os coeficientes de um modelo univariado auto-regressivo do produto nacional bruto dos EUA somam ***exatamente um***. Tal enfoque sobre as implicações exatas do que é formulado ***como um subespaço de menor dimensão dos valores dos parâmetros possíveis para a estatística de teste*** é difícil de entender se a hipótese nula é vista como apenas uma aproximação.

Com efeito, um dos artigos analisados da *Revista Brasileira de Economia* estudou a PPP. O trabalho buscou verificar, dentre seis índices de preços diferentes, quais apresentavam maior evidência da teoria. Porém, o critério escolhido para tanto foi a rejeição da hipótese nula de raiz unitária em testes *Dickey-Fuller* aumentado (ADF) ou *Dickey-Fuller* aumentado em mínimos quadrados generalizados (DF–GLS) aos níveis de 1%, 5% ou 10% de significância. Pela discussão anterior, é certo que estes resultados não respondem qual índice de preços apresenta melhor evidência da PPP ***do ponto de vista da teoria econômica*** e que há, aqui, um claro equívoco acerca da função do teste de significância estatística.

Estes exemplos ilustram bem o caso relatado por Kadane (2011, p.438) que, ao testar estatisticamente uma teoria específica da psicologia, acabou por rejeitá-la ao nível de significância de 10^{-6}. Isto levou o autor a ponderar sobre o significado da rejeição:

> [...] eu tive de pensar se ficaria mais impressionado se fosse rejeitada, digamos, ao nível de 10^{-13}, e decidi que não. O problema é que tínhamos uma base de dados muito grande [...] um simples gráfico nos mostrou que, na verdade, a teoria era muito boa.

O mesmo raciocínio se aplica aos casos mais comuns de hipótese nula, acerca de um coeficiente de regressão, $H_0: \theta = 0$. Em geral, a pergunta que se quer respondida é se o coeficiente é negligenciável. Não obstante, se a intenção é demonstrar que o efeito é grande ou pequeno, como vimos nas seções anteriores, ***somente*** o *p-valor* com relação à H_0 não dará esta resposta e, portanto, valer-se ***somente*** de um limiar arbitrário como $p < 5\%$ para julgar a significância econômica do coeficiente é, justamente, incorrer na confusão da qual estamos tratando. Pois, dado um nível arbitrário de significância, poderíamos encontrar um resultado "estatisticamente significante", mas de magnitude negligenciável, ou poderíamos encontrar um resultado "estatisticamente *in*significante" sugerindo efeito substancial. Cox expõe o tema de maneira clara:

> [...] o ponto central é que a significância estatística é bastante diferente da importância científica e que, portanto, a estimativa, pelo menos aproximada, da magnitude dos efeitos é, em geral, essencial, independentemente do fato de uma discrepância estatisticamente significativa da hipótese nula ter sido alcançada (COX, 1977, p. 71).

Na mesma linha, diz Berger:

> [...] uma diferença 'estatisticamente significante' entre o parâmetro verdadeiro (ou o modelo) e a hipótese nula pode ser uma diferença insignificante na prática. Da mesma forma, uma diferença que não é estatisticamente significante pode, não obstante, ser bastante importante na prática (BERGER, 1985, p. 21).

Como exemplo, podemos voltar ao caso do impacto das instituições e da geografia sobre o crescimento econômico. Vimos uma passagem de Acemoglu, Johnson e Robinson que sugeria que fatores geográficos, como a ***incidência de malária***, não teriam ***qualquer efeito direto sobre a renda – isto é, efeito zero.*** Na realidade, no artigo original, os autores são menos contundentes, e alegam que as regressões ***não fornecem evidência de impacto importante*** (2001, p. 1392-1393, grifo nosso)

> [...] em contraste com os resultados de McArthur e Sachs, nós encontramos que ***apenas instituições são significantes***. A diferença se deve ao fato de que McArthur e Sachs incluem a França e Grã-Bretanha em sua amostra, que

> consiste apenas de ex-colônias (não há razão para a variação da taxa de mortalidade doméstica das tropas francesas e britânicas estarem relacionadas com seu desenvolvimento institucional). Verifica-se que, uma vez retiradas Grã-Gretanha e França da amostra, ***a especificação de McArthur e Sachs não gera qualquer evidência*** de que fatores de geografia/saúde tenham um ***impacto importante*** no desempenho econômico.

Supondo, por instante, que não existam quaisquer outras fontes de incerteza além da variação amostral, verifiquemos se a assertiva dos autores procede. Na tabela 7, coluna 7 (p. 1392), o coeficiente para a população sob risco de malária foi estimado em -0.62, com erro-padrão de 0,68. De fato, isto nos dá uma estatística t de 0,91 com *p-valor* bilateral de aproximadamente 36% (unilateral de 18%) – o que nos levaria a não rejeitar a hipótese nula de que o coeficiente seja zero.

No entanto, valores para o coeficiente do risco de malária de até -1,22 teriam menos de 50% de chances de serem detectados, caso fossem verdade[60]. Este efeito para o qual teríamos ***pouco poder*** estatístico de identificar poderia ser considerado aproximadamente zero, isto é, o efeito seria economicamente negligenciável? Utilizemos um exemplo dos próprios autores – a comparação entre um país como a Nigéria e um país como o Chile. A região de baixo poder do teste abarca efeitos tais que ***aumentariam a renda de um país como a Nigéria, caso erradicasse a malária (situação do Chile), em até 200%,*** em média. Do ponto de vista econômico, certamente isso não é negligenciável. Assim, a não ser que os autores sintam-se confortáveis em errar mais da metade das vezes quando exista impacto desta magnitude, não parece ser uma inferência segura declarar a ausência de efeito importante. Menos ainda de efeito exatamente igual a zero, algo bem implausível *a priori.*

Sobre este tema, é interessante citar uma passagem do próprio Neyman, referente ao caso do "*The lady tasting tea*", em que o matemático faz a ***distinção*** entre inexistência de efeito e efeito negligenciável, bem como considerações acerca do poder do teste para detectar discrepâncias ***substancialmente importantes***:

> [...] é óbvio que se p[61] tem um valor, digamos, $p = 0{,}50001$, então, estritamente falando, a senhora tem de fato a habilidade que alega ter, mas, operacionalmente, sua habilidade é nula. Portanto, ***o pesquisador [...] provavelmente não se arrependeria do fato de o teste rejeitar raramente*** $\boldsymbol{H_1}$ ***[a hipótese nula de que*** $\boldsymbol{p = 0{,}5}$***] se o real valor de*** $\boldsymbol{p}$ ***for*** $\boldsymbol{0{,}50001}$***.*** ***Neste***

[60] Para $\alpha = 5\%$, teste bilateral. Para $\alpha = 5\%$ unilateral, ou $\alpha = 10\%$ bilateral, a região de baixo poder abrangeria coeficientes até -1,12. Adotamos a estratégia de Andrews (1989) que será discutida a seguir.

[61] Aqui p refere-se à capacidade de predição da senhora, que representamos por θ na seção anterior.

> ***caso, a intensidade do fenômeno é muito fraca para se preocupar***. Por outro lado [...] o pesquisador pode estar interessado em "detectar" o fenômeno se sua intensidade for substancial [...] ***que valores de p excedem 0,5 consideravelmente e que valores excedem 0,5 apenas ligeiramente é uma questão subjetiva e depende das circunstâncias gerais do problema*** [...] se o pesquisador considera que $p \geq 0{,}6$ significa uma habilidade substancial de discriminação, então é natural que ele arranje o experimento de forma que, se o real valor de p for pelo menos $0{,}6$, as chances de detectar o fenômeno $p > 0{,}5$ sejam razoavelmente grandes (NEYMAN, 1950, p. 278).

Note que Neyman preocupa-se com a estruturação de um teste, com o balanço de α e β de modo que este não tenha tanto poder quando a diferença é "muito fraca para se preocupar", mas que tenha bastante poder quando a diferença indica "uma habilidade substancial de discriminação".

A este ponto do trabalho, deve ter ficado claro o sentido da citação de Goodman (2008, p.138) ao final da seção 2.1. O mau entendimento em relação aos testes de significância ou testes hipóteses leva o pesquisador a desconsiderar a magnitude dos efeitos e a observar apenas aspectos da incerteza referente à variação amostral – isto é, à precisão da estimativa. Como corretamente apontam Ziliak e McCloskey (2008a, p. 26), "[...] logicamente falando, uma medida de precisão amostral não pode ser o final do questionamento". Contudo, os pesquisadores "[...] falam como se estabelecer a significância estatística de um número no sentido *fisheriano* fosse a mesma coisa de estabelecer significância de um número no senso comum" (ZILIAK, MCCLOSKEY, 2008a, p.27). Mais ainda, mesmo com relação à variação amostral, muitos usuários esquecem-se dos conceitos probabilísticos e dos riscos envolvidos na inferência e utilizam um nível de significância arbitrário para declarar a importância ou não da variável, a existência ou não de um efeito, como se a única fonte de informação para tal julgamento fosse a amostra que tem em mãos, sem entender o peso da evidência calculada. Em alguns casos, ignora-se outras fontes de erro por se buscar a significância estatística como se esta fosse necessária para declarar a importância científica. E, por fim, há a mistura entre crença e ação, ignorando o fato de que, para este último caso, a ponderação da gravidade dos erros de julgamento bem como das chances de cada erro deveriam ser levadas em conta.

2.2.5. Há como minimizar o problema?

Tendo em vista o escopo desta dissertação, não serão discutidas aqui soluções que estejam muito fora da tradição da literatura. Nesta seção, mencionaremos ***brevemente***

algumas abordagens que possam elucidar maneiras de diminuir o risco destes problemas, de maneira simples e dentro do próprio paradigma frequentista[62].

Podemos resumir as falácias decorrentes da confusão entre diferença estatística e diferença material em duas: (i) a falácia da aceitação, em que uma falha em se rejeitar a hipótese nula leva o pesquisador a considerá-la como verdadeira, e (ii) a falácia da rejeição, em que um resultado estatisticamente significante é considerado como cientificamente importante. No caso de estimativas de coeficientes, talvez o modo mais direto e conhecido de se evitar um julgamento precipitado, que consubstancie significância estatística e econômica, seja a utilização de intervalos de confiança, ***não como meros substitutos dos testes de hipótese***, mas no sentido de se preocupar com as magnitudes e incertezas envolvidas, em seu sentido econômico[63].

Já mais especificamente para evitar falácias decorrentes da falha em se rejeitar uma hipótese nula, Andrews (1989) proveu um método de fácil aplicação, por meio de uma ***função poder inversa.*** Note que a não rejeição de H_0 nos daria indícios de que aquelas hipóteses alternativas com bastante poder, isto é, aquelas que, caso fossem verdadeiras, nos levariam a rejeitar a hipótese nula frequentemente, podem ser falsas. Por outro lado, a não rejeição de H_0 ***não*** nos passaria tanta segurança acerca da falsidade de discrepâncias cujo poder para detectá-las fosse demasiadamente baixo[64].

Por exemplo, para um teste $H_0: \theta = 0$ contra $H_1: \theta \neq 0$, poderíamos construir uma região em que o poder para detectar uma discrepância fosse maior ou igual a $1 - \alpha$.

---

[62] Não discutiremos em detalhes a validade dos métodos. A ideia é apenas ilustrar possíveis maneiras de se evitar falácias bastante comuns. É importante notar que a aplicação dos métodos pressupõe que as probabilidades de erro estejam corretas ou aproximadamente corretas, a critério do pesquisador; e, principalmente, que os métodos ***não são*** condição suficiente para que a confusão entre significância estatística e científica seja suprimida.

[63] Na economia, isto foi defendido por DeLong e Lang (1992, p.1269) "os economistas não deveriam reportar se rejeitam ou não a hipótese nula, mas se seu intervalo de confiança exclui ou não exclui (a) valores economicamente insignificantes ou (b) valores economicamente significantes". Cox também incentiva o uso: "[...] uma falha em se alcançar um nível interessante de significância estatística [...] não significa que diferenças de importância prática estejam ausentes. [...] é necessário calcular os chamados limites de confiança para a magnitude dos efeitos e não apenas os p-valores. Isto é de crucial importância. É uma prática muito ruim resumir uma investigação importante unicamente por um p-valor (COX, 1982, p. 327)". Mayo e Spanos (2011) julgam que intervalos de confiança acabam por ser um tanto rudes, por tratarem de maneira igual todos os valores dentro do nível de confiança escolhido.

[64] Evidentemente que o que é considerado alto ou baixo poder depende das circunstâncias de cada problema, bem como da avaliação do pesquisador. O autor sugeriu a área de alto poder como $\geq 1 - \alpha$ e a área de baixo poder como $\leq 50\%$; contudo, diante de toda a discussão, é fácil perceber que tais valores não devem ser interpretados como medidas absolutas.

Esta região seria, usualmente, da forma $S = \{\theta: |\theta| > c\}$, para algum $c > 0$. Note que uma falha em rejeitar H_0 seria equivalente a rejeitar com nível de significância α que $|\theta| > c$. Dessa forma, "[...] se c é próximo de zero em um sentido substantivo, então o teste provê evidência de que $|\theta|$ é zero ou aproximadamente zero, como desejado" (ANDREWS, 1989, p. 1060). Por outro lado, também poderíamos definir uma região de discrepâncias em relação à hipótese nula que tivessem baixo poder para serem detectadas, como, por exemplo, chances menores do que 50%. Em geral, esta região seria da forma $B = \{\theta: 0 < |\theta| \leq b\}$ para algum $b \in (0, c)$ e uma falha em se rejeitar H_0 trairia pouca evidência contra estes valores.

Para testes comumente utilizados[65] de tamanho α, $H_0: \theta = 0$ contra $H_1: \theta \neq 0$, as aproximações assintóticas seriam $b = \lambda_{1,\alpha}\left(\frac{1}{2}\right)\widehat{\sigma_\theta}$ e $c = \lambda_{1,\alpha}(1-\alpha)\widehat{\sigma_\theta}$, onde $\widehat{\sigma_\theta}$ é uma estimativa consistente do erro-padrão para $\hat{\theta}$, $\hat{\theta}$ é o estimador de θ, e $\lambda_{1,\alpha}(p)$ é a constante calculada para 1 restrição testada, teste de tamanho α e poder p[66]. Por exemplo, quando $\alpha = 0{,}05$ teríamos $\lambda_{1,0,05}\left(\frac{1}{2}\right) = 1{,}96$ e $\lambda_{1,0,05}(0{,}95) = 3{,}605$ e quando $\alpha = 0{,}01$ teríamos $\lambda_{1,0,01}\left(\frac{1}{2}\right) = 2{,}576$ e $\lambda_{1,0,01}(0{,}99) = 4{,}902$. Como afirma Andrews (1989, p.1060), "[...] com estas fórmulas, é trivial determinar as regiões de baixo e alto poder discutidas anteriormente".

Andrews (1989, p. 1061) cita o exemplo do artigo de Ashenfelter e Johnson (1972) sobre o efeito da sindicalização nos salários. O trabalho verificou que estudos anteriores estimaram efeitos salariais de 10% a 52% e questionou como, diante de efeito de tão alta magnitude, tantas classes de trabalhadores permanecem sem sindicatos. Diante disto, Ashenfelter e Johnson (1972) buscaram verificar se não haveria um viés de endogeneidade que havia sido ignorado. De fato, ao estimarem os efeitos por variáveis instrumentais, não se rejeitou que a *dummy* para sindicalização fosse diferente de zero ao nível de significância de 5%. Entretanto, os autores foram cautelosos e evitaram aceitar a hipótese nula de que os coeficientes fossem iguais a zero. Em suas palavras:

---

[65] Testes de Wald, Razão de Verossimilhança ou Multiplicador de Lagrange, por exemplo.

[66] As tabelas com os diversos valores para λ podem ser encontradas em Andrews (1989), páginas 1067-1071. Note que são valores assintóticos, dessa forma, a depender do tamanho da amostra em que se aplica, as regiões são aproximações "rudes" das verdadeiras regiões. Contudo, como justifica Andrews (1989, p.1072), as aproximações são de fácil aplicação e rapidamente disponíveis para o auxílio na inferência, um avanço em um contexto em que quase ninguém analisa a função poder.

> [...] em um nível empírico descobrimos que permitir a determinação simultânea de salários, sindicalismo, e qualidade do trabalho na estimação tende a produzir uma estimativa do diferencial de salário de equilíbrio entre sindicatos/não-sindicatos na gama de 0 a 20 por cento, mas esta nunca é significativamente diferente de zero. Dadas as limitações quantitativas e qualitativas dos dados, estamos preparados apenas a dizer que não temos certeza da magnitude do efeito dos sindicatos sobre as diferenças salariais intersetoriais (ASHENFELTER, JOHNSON, 1972, p.505).

Andrews afirma que a cautela foi com razão. A área de rejeição de alto poder englobaria diferenciais acima de 165%. Dessa forma, seria possível afirmar que o teste nos dá indícios de que o efeito da sindicalização seja menor do que este valor. Contudo, a área de baixo poder alcançaria diferenciais de até 65% e como este valor "[...] é um enorme diferencial salarial, é claro que o teste não é capaz de distinguir entre uma diferença de zero e aquelas diferenças que não são zero e que são interessantes sob uma perspectiva econômica".

Ilustremos abordagem semelhante que pode ser realizada: uma análise do poder obtido após a realização dos dados – ou de severidade (gravidade) dos testes a que as hipóteses são submetidas (MAYO, SPANOS, 2006; 2011) [67]. Deborah Mayo (2004, p.86) reconhece a pertinência da crítica do "ritual nulo" de Gigerenzer, e afirma que o autor

> [...] sem dúvida [...] está correto ao afirmar que os textos de estatística erroneamente omitem essas diferenças filosóficas e históricas entre os testes de NP, testes Fisherianos e métodos Bayesianos; e, sem dúvida, a estatística foi mal ensinada a muitos [...]

Todavia, a filósofa acredita que o problema não esteja nos métodos clássicos em si, mas no seu mau uso, e que seria necessária

> [...] uma interpretação de testes estatísticos que mostre como eles podem produzir um julgamento genuíno de evidência sem interpretar erroneamente as probabilidades de erro e sem serem utilizados como métodos mecânicos de um "livro de receitas" cuja saída sejam "atos" associados a "aceitar H" ou "rejeitar H".

Mayo afirma acreditar que a análise de severidade supra esta lacuna. A autora argumenta que o papel da probabilidade não é fornecer medidas de suporte a alguma hipótese em particular, mas medidas de erro com relação ao teste aplicado. A função da estatística não seria revelar quais hipóteses são mais prováveis, mas sim quais hipóteses foram submetidas a testes altamente probatórios. Segundo Mayo, pode-se afirmar que uma hipótese passa em um teste severo se, caso fosse falsa, o teste teria alta probabilidade de detectar sua falsidade.

---

[67] Os autores utilizam a palavra *"severity"* em inglês.

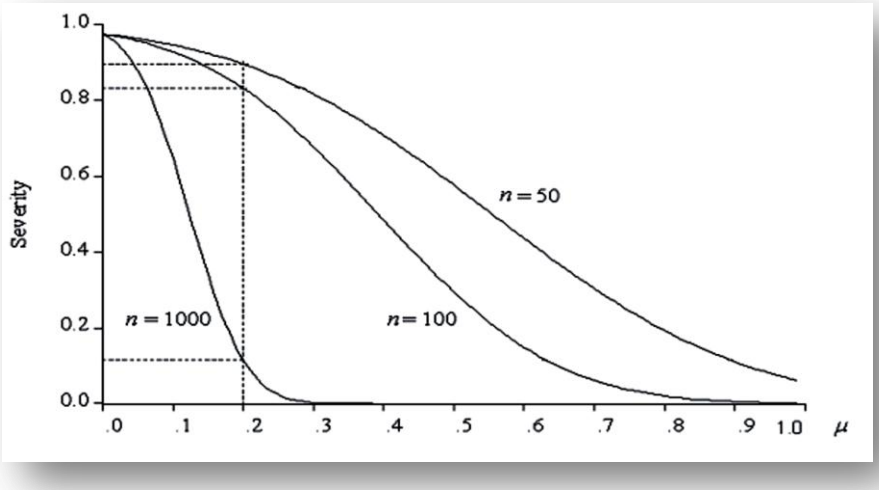
Antes da realização dos dados sabemos que: (i) a não rejeição de H_0 nos daria indícios de que podemos desconfiar da falsidade de hipóteses alternativas em que se teria alto poder caso fossem verdadeiras. Analogamente, portanto, uma rejeição de H_0 nos daria (ii) indícios de que há ***alguma*** discrepância (que pode ser trivial ou não trivial) em relação à hipótese nula. Ademais, após a realização dos dados, poderíamos "refinar" o processo de inferência e fazer os seguintes raciocínios contra factuais: (i) quando não rejeitamos H_0, que valores da hipótese alternativa H_1, caso fossem verdade, nos teriam dado, com alta probabilidade, uma estatística mais extrema do que a observada? Isto é, para algum $\theta' \in \Theta_1$, qual seria $\Pr(T(x) > t \mid \theta = \theta' \in \Theta_1)$? Quanto maior esta probabilidade, maior seria a evidência de que $\theta < \theta'$. E, também, no caso de (ii) rejeitarmos H_0, que valores da hipótese alternativa H_1, caso fossem verdade, nos teriam dado, com alta probabilidade, uma estatística menos extrema do que a observada? Isto é, para algum $\theta' \in \Theta_1$, qual seria $\Pr(T(x) < t \mid \theta = \theta' \in \Theta_1)$? Quanto maior esta probabilidade, maior seria a evidência de que $\theta > \theta'$. Vejamos alguns exemplos numéricos simples retirados de Mayo e Spanos (2006) e de Spanos (2008).

Considere uma amostra *i.i.d* $\boldsymbol{X} = (X_1, \dots, X_n)$ de tamanho n, em que cada X_i é normalmente distribuído com média μ e variância σ^2. Suponha que $\sigma = 2, n = 100$ e que $\alpha = 0{,}025$. Consideremos o teste $H_0: \mu \leq 12$ contra $H_1: \mu > 12$. Suponha que o resultado da média amostral tenha sido $\bar{x} = 12{,}3$. Note que a estatística de teste seria $T(x) = 1{,}5$ e ***não*** rejeitaríamos H_0. Concluir que $\mu = 12$ ou que, digamos, $\mu \leq 12{,}1$, seria uma inferência segura? Mayo e Spanos diriam que não, pois $\Pr(T(x) > t \mid \mu = 12.1) = 0{,}16$, o que significa que se a média populacional fosse 12,1, apenas em 16% das vezes a estatística seria maior do que a observada. Já a inferência de que $\mu \leq 12{,}3$ seria mais bem respaldada pelos dados? Sim, pois $\Pr(T(x) > t \mid \mu = 12.3) = 0{,}9997$, isto é, se a média populacional fosse de fato 12,3, seria quase certo termos observado valor maior para a estatística de teste – mas não observamos. Assim, fosse a discrepância de 0,1 magnitude relevante do ponto de vista econômico, então a insignificância estatística não se configuraria em insignificância econômica, pois a hipótese de que $\mu \leq 12{,}1$ não passa em um teste severo[68].

[68] Da mesma forma que na análise de Andrews, na análise de severidade ainda restará ao pesquisador definir o quão severo o teste tem de ser para considerar o resultado relevante ***do ponto de vista econômico.*** Expomos o método pois, ao discriminar diferentes tipos de inferência, isto talvez chame mais a atenção do usuário às magnitudes envolvidas. Ziliak e McCloskey (2008a, p.146-147), por um lado,

Considere novamente uma amostra i.i.d $\boldsymbol{X} = (X_1, \dots, X_n)$ de tamanho n, em que cada X_i é normalmente distribuído com média μ e variância σ^2. Realizemos um teste t para $H_0: \mu = 0$ contra $H_1: \mu > 0$ de tamanho $\alpha = 0{,}05$. Suponha que o resultado da média amostral tenha sido $\bar{x} = 0{,}02$, com, $n = 10.000$, e que o desvio-padrão amostral seja $s = 1.1$. O valor da estatística de teste seria $T(x) = 1{,}82$, nos levando a rejeitar H_0. Entretanto, veja que $\Pr(T(x) < t \mid \mu = 0{,}05) = 0{,}003$, isto é, se, por exemplo, o verdadeiro valor da média fosse 0,05, em apenas 0,3% das vezes o resultado da estatística de teste teria sido tão baixo quanto o observado. Deste modo, os dados ***não*** fornecem evidência forte de que $\mu > 0{,}05$, pois esta hipótese não passa em um teste probatório. Fosse este valor negligenciável do ponto de vista econômico, a significância estatística do resultado não se configuraria em significância econômica.

Figura 2 – Tamanho amostral e severidade para $\boldsymbol{p = 5\%}$



Fonte: Mayo e Spanos (2011, p.175)

Ressaltemos que esta análise corrobora o mesmo ponto discutido anteriormente acerca da relativização do *p-valor* como evidência, como nos resultados da tabela 3. A figura 2 ilustra diferentes julgamentos de severidade da rejeição da hipótese nula $H_0: \mu = 0$, com *p-valor* de 5%, em uma distribuição $N(\mu, 2)$, mas variando o tamanho amostral. Note que, quando $n = 50$, temos que $\Pr(T(x) < t \mid \mu = 0{,}2) = 0{,}895$, assegurando de certo modo a inferência de que $\mu > 0{,}2$. Já a mesma rejeição da hipótese nula com uma amostra de 1.000 observações nos daria $\Pr(T(x) < t \mid \mu = 0{,}2) = 0{,}115$, teste menos

elogiam a análise mais sistemática de erros proposta por Mayo; por outro lado, criticam o foco no uso de probabilidades contrafactuais para o julgamento de evidência, sem levar em consideração funções perda ou informação *a priori*, o que poderia acabar por resultar na má prática observada nos testes de significância puros.

severo para a hipótese $\mu > 0{,}2$. Em outras palavras, suponha que tomássemos $p = 5\%$ como evidência suficiente para inferir que $\mu > 0{,}2$. Note que, mesmo se a hipótese fosse falsa, isto é, mesmo quando $\mu = 0{,}2$, o teste somente nos forneceria resultado pior em 11,5% das vezes, possibilitando poucas chances detectar sua falsidade.

3) A SIGNIFICÂNCIA ESTATÍSTICA E A PRÁTICA ECONÔMICA

No capítulo anterior realizamos breve discussão teórica acerca da confusão entre significância estatística e significância econômica. Este capítulo trata da parte empírica – os pesquisadores têm, de fato, interpretado o *p-valor* como uma medida absoluta de evidência, dando o mesmo peso empírico a um resultado significante independentemente das circunstâncias, tais como o tamanho amostral? Uma vez que o arcabouço predominante é de testes de hipóteses de Neyman-Pearson, os pesquisadores têm se preocupado com o poder dos testes utilizados? Outras fontes de erro estão sendo ignoradas? Um resultado (in)significante tem sido tomado como (não) importante? O significado econômico da investigação – como as magnitudes estimadas – está sendo analisado?

Para responder essas perguntas, resgataremos, primeiramente, a literatura acerca do mau uso da significância estatística na economia, apontando alguns resultados encontrados em outros países, como a constatação do problema na *American Economic Review* e na *German Economic Review*. Além disso, até onde a revisão bibliográfica para este trabalho logrou alcançar, ***inexiste estudo publicado*** a respeito do tema para periódicos nacionais. Esta dissertação buscará suprir a lacuna. Assim, discutiremos o questionário aplicado por McCloskey e Ziliak (1996), fazendo algumas modificações em virtude dos resultados obtidos em uma amostra piloto, e apresentaremos os resultados de sua aplicação para os artigos da Revista Brasileira de Economia no período de 2008 a 2011.

3.1. Resgatando o debate na ciência econômica

Como pode ser visto em Ziliak e McCloskey (2008a), a literatura em periódicos internacionais acerca da confusão entre significância estatística e significância substantiva é extensa. O debate sobre o uso dos testes de significância estatística é antigo e está presente em diversas outras disciplinas, como psicologia, educação, negócios, direito, sociologia, arqueologia, biologia, epidemiologia, medicina e na própria estatística (ZILIAK, MCCLOSKEY, 2008a, p. 57-58). Não obstante, a discussão na economia se iniciou tardiamente, com poucos trabalhos que tratavam diretamente, ou mesmo tangencialmente, do tema. Tendo em vista que elementos

teóricos já foram expostos no capítulo anterior, nesta seção, buscaremos resgatar estes trabalhos de maneira cronológica, dando ênfase aos principais resultados.

3.1.1. A retórica da significância: α, β, hipóteses extravagantes, $\hat{\pi} = 0$?

Na economia, podemos remontar a discussão à Zellner (1981)[69]. O econometrista, em 1978, coletou uma pequena amostra de 22 artigos empíricos em periódicos importantes. O autor verificou o uso disseminado de níveis de significância de 1% ou 5% sem qualquer consideração com relação ao tamanho da amostra ou a outros fatores. Mais ainda, dos 22 artigos apenas 1 havia discutido o poder do teste utilizado. Zellner concluiu, modestamente, que haveria bastante espaço para a melhoria dos testes de hipótese e recomendou o uso de métodos Bayesianos.

Dois anos mais tarde, McCloskey (1983), em seu conhecido artigo *The Rhetoric of Economics*, foi quem primeiramente apontou de maneira explícita o uso da significância estatística como substituto do julgamento científico na ciência econômica. Segundo McCloskey (1983, p.497-498, grifo nosso),

> [...] a significância estatística parece prover um padrão para se julgar se uma hipótese é falsa ou verdadeira que é independente de qualquer consideração trabalhosa [...] o ponto não é que os níveis de significância são arbitrários. Claro que são. O ponto é que não se sabe se a amplitude abrangida pelo nível de significância ***afirma ou nega a hipótese***.

McCloskey afirma, em seguida, que se o estatístico poderia tentar alegar não conhecer o problema econômico a fundo, e assim se esquivar do julgamento econômico de quão boa ou ruim a teoria é como aproximação da realidade, o mesmo não pode ser dito, por exemplo, do Macroeconomista aplicado com especialização em comércio internacional que analisa a teoria da paridade do poder de compra. Neste caso, dever-se-ia ter "[...] padrões de argumentação que vão além da retórica inconclusiva provida pela cerimônia pseudocientífica hipótese-regressão-teste-publicação da maior parte da economia moderna" (MCCLOSKEY, 1983, p. 499).

---

[69] Arrow (1959) já trabalhava a diferença entre significância estatística e significância econômica. Neste texto o autor alertava que "[...] desde o trabalho clássico de Neyman e Pearson, tem sido evidente que, na escolha de um teste de hipótese, o poder do teste deveria ter papel coordenado com o nível de significância. Entretanto, até hoje, a escolha do nível crítico para uma estatística de teste é feita na prática por uma escolha arbitrária convencional de probabilidade de erro tipo I; dificilmente, quando nunca, se encontra uma justificativa explícita da escolha feita em comparação com a função poder" (ARROW, 1959, p.70). Todavia, como não houve levantamento empírico do problema, consideramos o início da discussão em Zellner, que foi seguido de outros textos em espaço mais curto de tempo.

Também foi neste mesmo ano que Leamer (1983) publicou outro artigo conhecido, *Let's Take the Con Out of Econometrics*, questionando seriamente as *hipóteses extravagantes* [70] feitas no uso da econometria. Leamer destacou a falta de reconhecimento da dependência dos resultados de tais pressupostos bem como a decorrente omissão de uma análise de sensibilidade ou de fragilidade das estimativas. O autor questionou a aparente "objetividade" das análises estatísticas na economia, alegando que

> [...] os economistas herdaram das ciências físicas o mito de que a inferência científica é objetiva e livre do julgamento pessoal. Isso é um total absurdo. Todo conhecimento é crença humana; mais precisamente, opinião humana (LEAMER, 1983, p.36).

E, adiante, lembra o leitor de que "[...] como tanto a distribuição amostral quanto a distribuição *a priori* são *opiniões* e não *fatos*, uma inferência estatística é e deve sempre permanecer uma *opinião*" (LEAMER, 1983, p.37).

Pouco depois, McCloskey (1985) examinou com mais cautela uma amostra de 10 dos 50 artigos que utilizaram análise de regressão na *AER* nos anos 1981, 1982 e 1983. A conclusão foi de que

> [...] aproximadamente três quartos dos autores do *American Economic Review* utilizam incorretamente o teste de significância. Eles o utilizam para se convencerem de que uma variável é importante. Mas o teste pode somente afirmar a probabilidade de ceticismo excessivo em face de erros resultantes de uma amostra muito pequena. ***O teste não diz ao economista se um coeficiente ajustado é grande ou pequeno em um sentido economicamente significante*** (MCCLOSKEY, 1985, p.201, grifo nosso).

A autora explicou como o processo de confusão usualmente ocorre:

> [...] o resultado que aparece na página 10 (estatisticamente) significante acaba por aparecer como (economicamente) significante na página 20. Nos piores casos, não há qualquer tentativa de mostrar quão grandes os efeitos são, ou se os testes estatísticos de sua grandeza são poderosos, ou qual padrão de grandeza alguém deveria usar [...] (MCCLOSKEY, 1985, p.204).

McCloskey (1986, p.06), em texto posterior, tenta colocar o ponto de uma maneira mais direta e óbvia:

> [...] suponha que você saiba o valor do coeficiente. Saiba com certeza. Deus te contou, sem qualquer disparate como um intervalo de confiança; o erro amostral é zero. A estatística t é infinita. Bem, então: a variável é importante? *Você ainda não sabe*. Para descobrir, você terá de perguntar e responder outras questões [...]

[70] Whimsical assumptions.

Em resposta aos artigos de Zellner e McCloskey, Andrews (1989) publica artigo com derivação de uma função poder inversa para aplicação no auxílio das inferências dos pesquisadores, método o qual mencionamos na seção 2.2.5. Conforme o autor

> [...] nós notamos que o cálculo de poder atualmente é pouco utilizado na pesquisa econométrica aplicada (por exemplo veja Zellner [...] e McCloskey [...]). Muitos praticantes não sabem como mobilizar informação para ajudar a analisar seus resultados ou pelo menos como fazer isso de uma maneira simples. (ANDREWS, 1989, p. 1061).

Na década de 90, DeLong e Lang (1992) publicam artigo curioso intitulado *Are all Eeconomic Hypothesis False?*. Como vimos na seção 2.1.1., a distribuição do *p-valor* sob a hipótese nula é uniforme (0,1). Dessa forma, caso H_0 seja verdadeira temos que

$$\Pr(P \geq p \mid H_0) = 1 - p \tag{16}$$

Isto, se a hipótese nula for verdadeira, em 10% das vezes nós veríamos valores maiores do que $p = 0{,}9$, e em 20% das vezes nós veríamos valores maiores do que $p = 0{,}8$ e assim por diante. Sob a hipótese alternativa, a distribuição do *p-valor* tem uma função de distribuição acumulada desconhecida $G(p)$, assim:

$$\Pr(P \geq p \mid H_1) = 1 - G(p) \tag{17}$$

DeLong e Lang presumem que a densidade $g(p)$ sob a hipótese alternativa é decrescente em p de tal forma que a razão $[1 - G(p)]/[1 - p]$ caia monotonicamente de 1 quando $p = 0$ até $g(1)$ quanto $p = 1$. Assim, considerando que π seja a proporção de hipóteses nulas verdadeiras, a probabilidade incondicional $\Pr(P \geq p)$ poderia ser escrita como:

$$\Pr(P \geq p\ ) = \pi(1 - p) + (1 - \pi)\big(1 - G(p)\big) \tag{18}$$

Como a distribuição acumulada $G(p) \leq 1\ \forall\ p \in [0{,}1]$, temos que:

$$\pi \leq \frac{\Pr(P \geq p\ )}{1 - p} \tag{19}$$

A equação (19) nos permitira estimar um ***limite superior*** para a proporção de hipóteses nulas que são verdadeiras. DeLong e Lang modificam um pouco o contexto da equação e estimam o limite superior para a proporção de nulas ***não rejeitadas*** que são verdadeiras. Para tanto, analisam 276 artigos de diversos periódicos da economia.

Destes, apenas 78 falharam em rejeitar a hipótese nula. Um dos resultados dos autores é que $\hat{\pi} = 0$, pois de todos os 78 *p-valores,* nenhum caiu no limiar entre 0,9 a 1. Ou seja, todas as hipóteses nulas não rejeitadas seriam falsas. A explicação mais plausível encontrada foi a da existência de um possível viés de publicação nos periódicos: haveria uma tendência de somente se publicarem resultados estatisticamente significantes ou aqueles resultados que falham em rejeitar uma hipótese nula que *a priori* era considerado de fato falsa (como não rejeitar que uma expansão monetária tenha impacto zero no produto de curto prazo, por exemplo). Assim, concluem os autores que,

> [...] em seu sentido mais simples, nossos resultados reforçam as solicitações anteriores para os economistas concentrarem-se nas magnitudes dos coeficientes e reportarem intervalos de confiança e não testes de significância. Se todas ou quase todas as hipóteses nulas são falsas, há pouco sentido em se concentrar se uma determinada estimativa é distinguível ou não de seu valor previsto sob a hipótese nula. Ao invés disto, nós deveríamos lançar luz em quais modelos são boas aproximações, o que requer que saibamos intervalos de valores para os parâmetros que são excluídos pelas estimativas empíricas [...] a pergunta central não deveria ser, posso rejeitar zero? Mas deveria ser, posso rejeitar todos os pequenos (ou todos os grandes) valores para este parâmetro? (DELONG, LANG, 1992, p.1271-72).

McCloskey (1992a, 1992b, 1993, 1995), por sua vez, continuou trazendo a discussão à tona no meio acadêmico, com tom cada vez mais provocante na tentativa de despertar a atenção dos pesquisadores:

> [...] o estatístico amador acenando o SPSS alega ter domínio da técnica. Procedimentos estatísticos, diferentemente dos números resultantes, são tomados pelos não estatísticos como técnicas para gerar verdades. Estudantes de doutorado em economia migram para o curso de econometria, porque eles acreditam que é o lugar para aprender a ciência econômica [...] seus professores têm vergonha de desiludi-los, pois eles não podem fornecer nenhuma outra fórmula para a ciência e os jovens procuram fórmulas. Os estudantes são atraídos pela ilusão de que técnicas de análise fatorial ou de variáveis instrumentais irão mecanizar a persuasão científica (MCCLOSKEY, 1993, p. 485).

Em meados da década, Keuzenkamp e Magnus (1995) explicaram, em periódico voltado para econometria, os diferentes tipos de hipóteses que podem ser testadas bem como as diferenças entre os métodos de Fisher e Neyman-Pearson. Os autores ainda pesquisaram 668 artigos do *Journal of Econometrics* observando com mais cautela 99 que utilizaram testes de significância. Nestes, verificou-se que, conforme havia constatado Zellner (1981), não há qualquer relação entre o tamanho da amostra e o nível de significância adotado: "[...] a escolha dos níveis de significância parece arbitrária e depende mais da convenção e, ocasionalmente, do desejo do investigador em rejeitar ou aceitar uma hipótese do que em uma avaliação bem-definida de perdas possíveis que

possam resultar de uma decisão errada" (KEUZENKSAMP, MAGNUS 1995, p.20). Assim, o debate em torno do uso da estatística, mais especificamente em torno do uso indiscriminado da significância estatística, que havia surgido na psicologia e em outras ciências sociais na década de 60 e 70, estava tomando forma na economia (MCCLOSKEY, 1993; ZILIAK, MCCLOSKEY, 2008).

3.1.2. O problema na *American Economic Review* nas décadas de 80 e 90

Em 1996, McCloskey e Ziliak realizaram o primeiro estudo sistemático e abrangente acerca da confusão entre significância estatística e significância econômica. À época, constataram que poucos livros-textos de econometria faziam a devida distinção e, muitas vezes, inclusive confundiam os alunos. Em um exemplo, um dos livros examinados testou se penas severas para a direção perigosa diminuiriam as mortes no trânsito. A conclusão foi de que o resultado era significante a 5% mas não era a 1%. McCloskey e Ziliak (1996, p.100) complementam,

> [...] mas as 100.000 vidas que seriam salvas pela redução medida não são reconhecidas como "significantes" [...] a que nível o nível de significância deveria ser estabelecido, considerando o custo humano de se ignorar o efeito de penas severas [...] não é trabalhado no livro.

McCloskey e Ziliak tomaram para análise todos os 182 artigos publicados nos anos 80 na *AER* que utilizaram análise de regressão e aplicaram, a cada, um questionário de 19 questões sobre a preocupação com a magnitude e a importância dos efeitos medidos bem como sobre o uso da significância estatística. Os resultados encontram-se na tabela de número 04, que se encontra ao final deste capítulo[71]. Os principais números foram: cerca de 70% dos artigos não diferenciaram significância estatística da significância econômica; 72% não discutiram o "quão grande é grande" ou o "quão próximo é próximo" para o problema que estudavam, isto é, não definiram nem conversaram com a literatura da área acerca do padrão adequado para julgar um modelo ou uma variável relevantes; 59% utilizaram a palavra significante de maneira ambígua; apenas 4% dos artigos consideraram o poder do teste e apenas 1% analisou a função poder; 69% dos artigos sequer reportaram suas estatísticas descritivas; 63% dos artigos praticaram a "econometria do sinal", observando apenas o sinal do coeficiente sem qualquer preocupação com sua magnitude.

[71] A discussão detalhada sobre cada questão será feita na seção seguinte, em que serão trazidos exemplos da amostra analisada.

Entretanto, o impacto do levantamento não foi aquele esperado pelos autores. Pois as reações foram, em geral, em dois sentidos: (i) aqueles economistas mais sofisticados afirmavam que reconheciam a existência do problema, mas diziam que *eles* não cometiam tais erros e que isso era coisa de amadores; e, (ii) os demais ignoravam a existência do problema, afirmando ser absurdo conceber que algo que todo mundo fazia estivesse tão errado. Não entendiam o ponto da crítica alegando ser o ataque contra a matemática e a estatística em si ou ser o ataque algo gratuito e pessoal contra aqueles economistas dos artigos analisados. Deste modo, pouco foi feito para se mudar a situação, levando McCloskey (1997a, 1997b, 1998, 1999, 2002) a afiar mais sua crítica:

> [...] eu quero que você tente esquecer as características do mensageiro que possam te distrair. Tente esquecer que sou uma mera historiadora econômica, que moro no centro-oeste, que não estou em Princeton, que sou uma libertária feminista da escola de Chicago, que sou transexual, que eu gaguejo [...] pense que seja possível que a significância estatística e teoremas de existência têm sido pior do que inúteis nestes últimos cinquenta anos [...] ***não é o uso*** da matemática ou da estatística que está em jogo. ***É o dramático mau uso*** delas em testes de significância e provas, que a despeito da retórica de números são desconectadas da ciência real. Nenhuma questão de Quão Grande foi alguma vez respondida consultando um teste de significância ou teorema de existência (MCCLOSKEY, 1998, p. 04-05, grifo nosso).

Mais adiante, o discurso predominante foi de que os autores estavam certos em terem trazido o problema à tona, mas que aquilo era coisa do passado, da década de 80, e que a ciência econômica havia avançado (ZILIAK, MCCLOSKEY, 2008a, p. 79).

De encontro a esta afirmativa, Mayer (2001) verificou que artigos da *American Economic Review* e da *Review of Economic and Statistics*, nos anos 1999 e 2000, confundiram uma falha em se rejeitar a hipótese nula como uma confirmação desta hipótese. Diante disto, uma recomendação imediata foi a de que os pesquisadores poderiam

> [...] relatar seus *p-valores* ou intervalos de confiança, de modo que eles – e seus leitores – possam decidir a partir desta informação, em combinação com a informação *a priori*, o quão crível a hipótese é. Apesar da subjetividade, isto é preferível a afirmar erroneamente que a falha de um teste de significância em refutar uma hipótese ao nível de 5 por cento implica que esta hipótese foi confirmada. E também é melhor do que o pesquisador decidir nos bastidores se o *p-valor* assegura a defesa da hipótese (MAYER, 2001, p.09).

Ziliak e McCloskey (2004a), por outro lado, replicaram seu estudo a todos artigos da *AER* na década de 90[72]. Os resultados encontram-se na tabela de número 04 e não foram animadores. Desta vez, 79% dos artigos confundiram significância estatística com significância econômica. Pioraram igualmente, dentre outras, a prática da "econometria do asterisco" – classificar a importância de um coeficiente pelo tamanho de sua estatística de teste – bem como a prática da "econometria do sinal", com 69% e 78% respectivamente. Dentre as práticas em que se viram melhorias podem ser citadas a exposição das estatísticas descritivas, o uso de simulação para verificar a razoabilidade dos coeficientes, e a interpretação cuidadosa dos coeficientes, com 66%, 32% e 81% respectivamente.

Neste trabalho, Ziliak e McCloskey (2008a, p.91-92) classificaram os ***autores dos artigos*** segundo suas pontuações no questionário. Entre os que tiraram boas notas encontram-se Joshua Angrist, que obteve três artigos com alta pontuação, ou Gary Solon e David Zimmerman. Mas, com artigos de baixíssima pontuação, também se encontram autores renomados como Gary Becker, Ben Bernanke e Alan Blinder. Foi uma medida de certa forma controversa que, se por um lado tinha o intuito de mostrar que a questão não era problema de amadores ou maus economistas e sim que se tratava de prática generalizada, por outro pode ter sido capaz de gerar mais resistência por parte de alguns economistas, como visto na introdução do capítulo 2. Não obstante, a partir daí a questão passou a ter maior repercussão, saindo matéria inclusive na revista The Economist (2004).

Zellner (2004) ao comentar o artigo de Ziliak e McCloskey, mostrou-se surpreso por não ter havido melhoria no que diz respeito aos usos dos testes de hipótese entre os anos 80 e anos 90. O autor se pergunta por que pesquisadores inteligentes e altamente treinados continuariam indo tão mal em testes de hipóteses, sua resposta foi

> [...] que eles estão muito confusos com relação as metodologias de testes. A maioria deles não sabe qual o conceito de probabilidade que eles estão usando, têm dificuldade em interpretar os *p-valores*, não sabem o que são funções poder e não sabem como usá-las, especialmente porque eles não têm ideia de que valores de parâmetros usar, e não sabem como escolher um nível de significância conforme o tamanho da amostra se altera (ZELLNER, 2004, p. 583)

[72] Ziliak e McCloskey (2004a) analisaram 134 artigos, omitindo, sem perceberem, cerca de 50 artigos, como foi apontado por Hoover e Siegler (2008). Posteriormente, Ziliak e McCloskey (2008a) consolidaram os resultados com todos os 184 artigos publicados na década de 90 não obtendo diferença significativa nos percentuais. Apresentaremos estes resultados.

Wooldrige (2004), por sua vez, não se disse impressionado com os resultados. Como disse o econometrista, "[...] eu participo de muitos *workshops* empíricos em que os tamanhos dos coeficientes não são discutidos. O estado das coisas é ainda pior para modelos não lineares" (WOOLDRIDGE, 2004, p. 577). Wooldridge, entretanto, faz a ressalva para não interpretar o chamado de Ziliak e McCloskey como prestar atenção apenas ao tamanho da estimativa do coeficiente, esquecendo-se da incerteza amostral, pois "[...] muito foco na significância econômica pode ser tão perigoso quanto muito foco na significância estatística" (WOOLDRIDGE, 2004, p. 579). Já Thorbecke (2004, p.571), ao comentar o artigo, afirma que a mensagem

> [...] é clara e convincente. Existe confusão entre os pesquisadores econômicos entre ajuste estatístico e a importância do efeito (por exemplo, a magnitude dos coeficientes correspondentes), fazendo falsas hipóteses serem aceitas e hipóteses verdadeiras serem rejeitadas. Muitas vezes os economistas não fazem distinção entre significância econômica e estatística.

Elliot e Granger (2004) e Horowitz (2004) também concordam com o ponto principal de Ziliak e McCloskey, de que significância estatística não é necessária nem suficiente para significância econômica e que o mau uso tem sido generalizado. Mas, ambos chamam a atenção para não interpretar este mau uso dos testes como a inutilidade de se testar em qualquer circunstância. Os autores frisam que não se pode ignorar a variação amostral como fonte de erro e Horowitz (2004) traz à tona o uso de testes de hipótese para erros de especificação.

O comentário de Leamer (2004), tal qual seu texto de 1983, é bastante crítico a toda a cultura atual vigente nos trabalhos aplicados. O econometrista afirma que "[...] modelos não são nem verdadeiros nem falsos. Eles são algumas vezes úteis e algumas vezes enganosos. O objetivo de um economista empírico não deveria ser determinar a veracidade de um modelo, mas o domínio de sua utilidade" (LEAMER, 2004, p. 556). Entretanto, os alunos em geral não conseguem entender este ponto:

> [...] é difícil treinar um computador para entender uma metáfora, e é igualmente difícil treinar nossos alunos a entenderem as metáforas da economia, os nossos modelos. Nossos alunos fazem o que qualquer um não familiarizado com um idioma faz: tomam os modelos literalmente. O objeto da econometria é extremamente prejudicial a esse respeito, uma vez que se baseia fundamentalmente na ideia de que nossos modelos são descrições literais da realidade (LEAMER, 2004, p. 556).

Dessa forma, na visão de Leamer, o que seria preciso seriam medidas de utilidade e não medidas de veracidade dos modelos e, sem mudar o paradigma vigente, pouco adiantaria chamar atenção às magnitudes.

Hoover e Siegler (2008a) foram os únicos a levantarem uma crítica sistemática aos trabalhos de McCloskey e Ziliak e McCloskey. Não obstante, iniciam o artigo afirmando que ***concordam incondicionalmente*** com o ponto principal levantado:

> [...] para evitar qualquer mal-entendido, vamos declarar desde o início que nós aceitamos o ponto principal, sem qualificação: um parâmetro ou outra quantidade estimada pode ser estatisticamente significativa e, ainda, economicamente sem importância ou pode ser economicamente importante e estatisticamente insignificante (HOOVER, SIEGLER, 2008a, p. 02).

O que ambos alegaram é que este ponto é desinteressante e incontroverso e que, assim, os economistas não cometeriam tais erros generalizadamente: a evidência apresentada por Ziliak e McCloskey seria fraca. O problema é que a questão aqui seria empírica. Constatar se os pesquisadores cometem ou não os erros não é uma questão de percepção com base em seu círculo profissional, como às vezes colocam os autores: "[...] isso não corresponde ***à nossa*** própria prática ou à de praticamente todos ***os economistas aplicados que conhecemos***, que estão geralmente preocupados com a escolha de modelos econômicos e estatísticos adequados" (HOOVER, SIEGLER, 2008b, p.463, grifo nosso). Para tanto, seria necessário investigar o que de fato ocorre nos trabalhos aplicados, e Hoover e Siegler não tomaram nenhuma amostra representativa para alegar o contrário, mesmo que fosse com metodologia diferente.

Aparentemente o discurso dos autores dá a entender que a intenção dos levantamentos é constatar se os economistas sabem ou não, ***subjetivamente***, a diferença entre significância econômica e estatística. Contudo, obviamente que não é este o ponto, pois, com exemplos simples, quase qualquer economista consegue entender a explicação e perceber a diferença. Mas, ainda assim, isso não necessariamente irá se refletir na prática. Por exemplo, como colocou Mayer, "[...] embora possa ser bem conhecida a proposição de que a incapacidade de rejeitar ao nível de 5 por cento não implica na confirmação ao nível de 5 por cento, em princípio, ***a prática é outra coisa***" (MAYER, 2001, p. 06, grifo nosso).

Além disso, como, em geral, os exemplos utilizados para a explicação são óbvios, pode parecer que a diferença entre significância estatística e significância substantiva seja

sempre algo trivial – mas não é. Vide, por exemplo, o problema dos testes de raiz unitária levantado por DeLong e Lang: o quão diferente um coeficiente tem de estar de uma raiz unitária para fins relevantes na análise de séries temporais? Note que a mera constatação de rejeição ou não rejeição da hipótese nula a um nível arbitrário de significância não fornece essa informação. Ou em um teste paramétrico de normalidade, como o teste Jarque-Bera: o quão distante tem de estar a distribuição da distribuição normal (em termos de curtose e assimetria) para que haja consequências relevantes nos propósitos da análise? Este caso é mais fácil de ser determinado, mas ainda assim é uma pergunta que poucos economistas saberiam responder prontamente.

A despeito de não trazerem levantamento empírico, Hoover e Siegler apontaram algumas questões metodológicas importantes. A principal delas foi uma falha grosseira de Ziliak e McCloskey, que não teriam, na década de 90, coletado todos os artigos aplicáveis. Todavia, apesar do erro, como visto posteriormente em Ziliak e McCloskey (2008a, p.79-88), a incorporação destes artigos omitidos não modificou os maus resultados encontrados. Outra crítica pertinente foi o caráter binário e impreciso das questões (que são respostas de *sim* ou *não*) o que exacerbaria a subjetividade[73] na codificação. Isso talvez tenha decorrido dos poucos exemplos utilizados por Ziliak e McCloskey para explicar o sentido de cada pergunta, levando Hoover e Siegler a se indagarem, por exemplo: quando os coeficientes seriam ditos como "interpretados cuidadosamente"? Ou, como se classificaria quem menciona o poder do teste, o autor teria de mencionar diretamente a palavra poder? Para suprir esta lacuna, nós traremos aqui vários exemplos de nossa amostra. Espera-se que, durante a leitura, perceba-se que estas questões, apesar de ***trabalhosas*** para a avaliação, não são tão complicadas.

Além disso, os autores questionam a aparente redundância de alguns pontos – como várias perguntas sobre magnitude, e várias sobre a confusão entre significância estatística e significância econômica. Isto se mostraria mais problemático quando da classificação dos autores em um *ranking*, pois poderia haver múltiplas contagens de um mesmo erro. De fato, a classificação de autores feita por Ziliak e McCloskey foi

[73] Com relação à subjetividade, Ziliak e McCloskey afirmam que convidaram Hoover e Siegler para discutir as questões e esclarecer possíveis dúvidas, mas os autores declinaram. Ao invés disso, segundo Ziliak e McCloskey, Hoover e Siegler exigiram que os autores escrevessem cada classificação com as respectivas citações que fundamentavam as notas. Em virtude do custo de oportunidade de tal empreitada, Ziliak e McCloskey negaram a tarefa, mas convidaram os autores a examinarem as fotocópias dos artigos originais com as anotações realizadas. A proposta foi negada (ZILIAK, MCCLOSKEY, 2008a, 2008c).

bastante imponderada e não faremos isso neste trabalho. Apenas iremos exibir a distribuição de frequência do percentual de "sim" obtidos pelos artigos com o simples intuito de sintetizar resultados. Em suma, o questionário proposto por Ziliak e McCloskey realmente apresenta bastante espaço para melhorias; para amenizar os possíveis problemas apontados por Hoover e Siegler – e outros – o questionário foi primeiramente aplicado em uma amostra piloto, e a partir daí foram feitos os ajustes que se mostraram necessários (por exemplo, a questão da redundância não se mostrou tão importante, como será visto a seguir), mas buscando não descaracterizá-lo. Pois, apesar dos defeitos, ainda assim trata-se do questionário mais abrangente, bem como o pioneiro da ciência econômica, com já duas amostras para comparação, sendo, portanto, a escolha mais natural para este trabalho.

Por fim, cabe mencionar aqui que Hoover e Siegler buscaram justificar o uso de testes de significância, sendo talvez a parte mais problemática de sua resposta. Pois, na discussão, os autores sequer mencionaram como se determinar o nível de significância adequado frente às diversas circunstâncias em que um teste pode ser aplicado, ou qual o verdadeiro sentido do *p-valor* como medida de evidência. Neste sentido, não há como diferenciar a aplicação sugerida do mau uso difundido e constatado por Arrow, Zellner, DeLong e Lang, Ziliak e McCloskey, Keuzenkamp e Magnus ou Mayer. Os autores afirmam, por exemplo, que "[...] a função do teste de significância é a de dar a qualidade da mensuração, de nos passar uma ideia da força do sinal [medido]. O princípio envolvido quando $n = 1\ ou\ 5$ não é diferente de quando $n = 10.000$" (HOOVER E SIEGLER, 2008a, p. 16). Salvo melhor juízo, isto dá a entender que a informação que um *p-valor* ou uma rejeição de uma hipótese nula fornece é a mesma independentemente do tamanho amostral ou da magnitude estimada, o que claramente não é verdade em virtude de todo o exposto nas seções 2.1.1, 2.1.2, 2.2.1, 2.2.2, e 2.2.5. Isto lança sérias dúvidas acerca do entendimento dos autores sobre a confusão entre significância econômica e estatística. Como colocaram Ziliak e McCloskey (2008b, p.49) "[...] Hoover e Siegler declaram que, com pequenas amostras, "o ruído ultrapassa o sinal." Mas não existe um padrão absoluto de 'ultrapassar'. Isto depende. Também não existe um padrão absoluto de 'pequenez' de amostras".

3.1.3. O livro de Ziliak e McCloskey e o "culto" na *German Economic Review*

O último trabalho de Ziliak e McCloskey (2008a) compila os dois estudos anteriores realizados na *AER* bem como os argumentos utilizados em outros artigos. O livro traz, ainda, exemplos e referências de como aparece o problema da significância estatística em outras áreas das ciências sociais e biomédicas, e também busca identificar as origens históricas do ritual estatístico atualmente adotado, encontrando suas origens em Fisher, que se sobrepôs aos métodos sugeridos por Gosset, Neyman, Pearson e Wald. Tendo em vista tudo que já foi exposto no capítulo anterior, não é necessário adentrar em detalhes da obra. Cabe, aqui, apenas mencionar os comentários de Aris Spanos (2008) e Tom Engsted (2009) ao livro, bem como reportar os resultados do artigo de Walter Kramer (2011), que fez análise recente do problema na *German Economic Review (GER).*

Engsted (2009 p.395) concorda com o ponto principal de Ziliak e McCloskey e no início do texto o autor faz questão de frisar o ponto:

> [...] eu gostaria de afirmar, desde o início, que eu concordo plenamente com o ponto de McCloskey e Ziliak de que (in)significância estatística não implica necessariamente em (in)significância econômica, e que uma boa pesquisa empírica em economia deve discutir o significado econômico de uma maneira ou de outra [...] não há método objetivo ou padrão (como o nível de significância de 5%) que por si mesmo pode decidir por nós.

Na verdade, o principal objetivo do texto de Engsted é chamar a atenção para áreas da ciência econômica em que os pesquisadores, reconhecendo as limitações e o caráter aproximado e inerentemente mal especificado de seus modelos, não se respaldam em testes de significância para avaliá-los. Tais áreas seriam a de modelos de equilíbrio geral dinâmico estocástico (DSGE), ciclos reais de negócios (RBC) e modelos de expectativa racional linear (LRE). Nestes campos, é explicitamente reconhecido que os modelos não pretendem ser replicações da realidade e que testes de significância seriam pouco ou quase nada informativos. Engsted defende, diga-se, com certa propriedade, essas linhas de estudo, afirmando que fazem exatamente o que recomendam Ziliak e McCloskey. O argumento do autor é, portanto, que não seria verdade que "quase todos" os economistas confundem significância econômica com significância estatística. Apesar da validade de chamar a atenção para estes campos – que abandonaram os testes de significância, pois, como já afirmou Sargent, a significância estatística estaria rejeitando muitos bons modelos – a discussão de Engsted tem pouco sentido empírico para esta dissertação, uma vez que desconsidera a existência do problema onde ela foi

apontada (nos estudos econométricos aplicados, que representam cerca de 70% de todos os trabalhos publicados na RBE, por exemplo) e não realizou qualquer levantamento para quantificar o problema.

Já Aris Spanos (2008) reconhece que a questão dos testes de significância foi levantada na economia tardiamente e dá a Ziliak e McCloskey crédito por fazerem disto um problema na área. Spanos, todavia, preocupa-se com o tom e o modo como a discussão foi trazida, que, em suas palavras, "[...] ofusca as questões envolvidas" (SPANOS, 2008, p. 156). O autor afirma que, de fato, os métodos de Fisher e de Neyman-Pearson são muito suscetíveis às falácias de aceitação e de rejeição, mas não acredita que os métodos vez ou outra pincelados por Ziliak e McCloskey sejam adequados para resolver o problema[74]. Spanos sugere, por conseguinte, a análise da severidade dos testes a que as hipóteses são submetidas (tratada sucintamente na seção 2.2.5.). Outro ponto levantado é com relação ao uso dos testes de significância para abordar problemas de especificação, alegando, inclusive, serem modelos mal especificados problema pior do que as falácias de aceitação e de rejeição[75]. Em suas palavras,

> [...] o problema de má especificação estatística não é apenas mais fundamental, mas os pesquisadores sabem, há algum tempo, como lidar com ele usando os testes de má especificação e reespecificação. Além disso, testes de significância de Fisher desempenham um papel crucial na validação do modelo [...] na verdade, me pergunto quantos artigos aplicados publicados na *American Economic Review*, nos últimos 30 anos, são suscetíveis de passar nos testes de adequação estatística; eu arriscaria um palpite de menos de 1% [...] (SPANOS, 2008, p.163).

Não obstante, vale lembrar que mesmo testes de especificação não têm como fugir das falácias de aceitação e de rejeição sem uma métrica de quão grande é um desvio em relação à hipótese nula. Ademais, se os autores dos textos analisados por Ziliak e McCloskey fizeram inferência estatística em seus modelos, supõe-se que fizeram porque achavam que poderiam fazer, isto é, porque achavam que seus modelos eram estatisticamente adequados. Deste modo, um erro não justificaria o outro, e se os pesquisadores além de confundirem significância estatística com significância

[74] É interessante notar que o livro de Ziliak e McCloskey não é um livro técnico, e tem um caráter informal de prosa. Dessa forma, realmente não há no livro qualquer desenvolvimento de método para solução dos problemas apontados, apenas referências sobre onde buscá-las, inclusive referências ao próprio Spanos.

[75] Na verdade este é também um dos pontos levantados por Ziliak e McClokey (2008a e 2008c, p. 166).

econômica, ***o fizeram em um contexto em que qualquer teste seria inerentemente falho, a situação seria, na verdade, muito pior***.

Por fim, cabe discutir o trabalho mais recente encontrado na área, de Walter Kramer (2011). Kramer analisou todos os artigos publicados na *GER* desde seu lançamento em 2000, totalizando 258 trabalhos. Destes, 110 utilizaram testes de significância, e foram objeto de análise mais detalhada. Kramer (2011, p. 462) confirma os resultados encontrados por Ziliak e McCloskey na *AER* – 56,4%, isto é, mais da metade dos artigos cometeram a falácia da rejeição, bem como 28,2% cometeram a falácia da aceitação. Além disso, 20,4% adicionaram ou excluíram variáveis do modelo com base unicamente na significância estatística. Kramer enfatiza também a questão da falta de discussão acerca da adequação estatística do modelo utilizado. Mais de 70% dos artigos não discutiram o fato de que a "significância" dos resultados depende da especificação correta dos modelos. Além disso, 57,1% sequer deram uma justificativa, seja teórica ou por meio de testes de diagnóstico, para a especificação utilizada. Assim, Kramer conclui "[...] que as intermináveis tabelas de valores *t* que adornam a maioria dos trabalhos empíricos de hoje são de fato o que Ziliak e McCloskey as denominam - um desperdício desnecessário de tempo e espaço (KRAMER, 2011, p. 466)". Kramer finaliza com a ressalva de que não se deveria abandonar por completo o uso dos testes de significância, e sim utilizá-los principalmente para verificar se os modelos estão corretamente especificados, sem, contudo, perder de vista que, como não existe um modelo exatamente correto, também seria necessária "[...] uma distinção entre "incorreção" no sentido estatístico e no sentido econômico" (KRAMER, 2011, p. 469).

Passemos agora à análise dos artigos publicados na Revista Brasileira de Economia.

Tabela 04 – O culto da significância estatística na *American Economic Review*
Décadas de 1990 e 1980 – Percentual de "sim"

O artigo…	***AER (90's)***	***AER (80's)***
Q8 - Menciona o poder do teste?	8,0	4,4
Q6 - Evita reportar todos os testes quando irrelevantes?	9,6	8,3
Q16 - Considera mais do que a significância estatística para um argumento decisivo do ponto de vista empírico?	20,9	29,7
Q11 - Evita a "econometria do sinal"?	21,9	46,7
Q14 - Evita escolher variáveis para o modelo unicamente por meio de significância estatística?	27,3	68,1
Q15 - Após o ponto principal, evita usar a significância estatística como o critério de importância científica?	27,8	40,7
Q10 - Evita a "econometria do asterisco"?	31,0	74,7
Q17 - Utiliza "simulação" para verificar se os coeficientes são razoáveis?	32,6	13,2
Q19 - Evita utilizar a palavra significante com sentidos ambíguos?	37,4	41,2
Q7 - Quando no primeiro uso, considera a significância estatística como apenas um entre outros critérios de importância?	39,6	47,3
Q9 - Caso mencione o poder do teste, faz algo em relação a isso?	44,0	16,7
Q13 - Discute a "conversa científica" na qual um parâmetro seria considerado grande ou pequeno?	53,5	28,0
Q18 - Nas conclusões ou considerações finais, separa significância estatística de significância econômica, política ou científica?	56,7	30,1
Q2 - Apresenta estatísticas descritivas?	66,3	32,4
Q1 - Utiliza amostra pequena?	71,1	85,7
Q12 - Discute o tamanho dos coeficientes?	78,1	80,2
Q5 - Interpreta cuidadosamente os coeficientes?	81,0	44,5
Q4 - Fez a hipótese nula adequada?	83,9	97,3
Q3 - Apresenta coeficientes em formas economicamente interpretáveis?	86,9	66,5

***Fonte**: todos os artigos completos publicados na American Economic Review nas décadas de 1980 e 1990, conforme análise de Ziliak e McCloskey (2008a).*

3.2. Significância estatística nos artigos empíricos: RBE 2008 - 2011

A amostra analisada constituiu-se dos artigos que utilizaram inferência estatística publicados na Revista Brasileira de Economia, nos anos de 2008 a 2011. No total, foram publicados 94 artigos, sendo que destes 84 eram trabalhos empíricos e 10 teóricos. Dos trabalhos empíricos, 67 utilizaram testes de inferência estatística, em especial, testes de significância, para fundamentar suas conclusões – estes últimos foram os artigos analisados. Os dados citados encontram-se resumidos na tabela de número 05.

Tabela 05 – Artigos empíricos x teóricos na RBE 2008-2011

Classificação	***Empíricos com testes***	***Empíricos sem testes***	***Teóricos***	***Total***
Artigos	67	17	10	94
Percentual	71%	18%	11%	100%

Fonte: *todos os artigos da Revista Brasileira de Economia, 2008-2011.*

Os artigos empíricos sem testes consistiram-se, em sua maioria, de trabalhos com exercícios de análise descritiva de dados, análises históricas, bem como simulação com modelos econômicos, tais como modelos de equilíbrio geral dinâmico estocástico, modelos de equilíbrio geral computável entre outros.

Note que, mesmo considerando todas as publicações no período, ***a amostra analisada representa 71% dos trabalhos publicados***. Ademais, dentro do ***universo de trabalhos empíricos***, os artigos submetidos à análise representam ***cerca de 80% do total***. Por conseguinte, percebe-se que, ainda que se excluam todos aqueles artigos que se enquadram no que foi discutido por Engsted (2009), ***o universo de análise é***, de longe, ***a maior parte dos trabalhos aplicados*** – o que evidencia a relevância do presente estudo. Com relação ao tipo de análise de dados utilizada nos artigos da amostra, verificou-se que a maioria realizou análise de séries temporais (42%), seguida de análise de dados em painel (34%) e, por fim, de corte transversal (24%). O resumo dos dados pode ser visto na tabela de número 06.

Tabela 06 – Tipo de análise dos artigos publicados na RBE 2008-2011

Classificação	***Séries Temporais***	***Painel***	***Corte Transversal***	***Total***
Artigos	28	23	16	67
Percentual	42%	34%	24%	100%

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia 2008-2011. Considerou-se o tipo de análise de dados predominante para a classificação.*

3.2.1. Os ajustes no questionário

Os artigos objetos desta dissertação foram analisados segundo uma bateria de perguntas baseadas principalmente em McCloskey e Ziliak (1996), Ziliak e McCloskey (2004a) e Ziliak e McCloskey (2008a). Escolheu-se trabalhar com este modelo por ser o questionário pioneiro utilizado na ciência econômica, bem como por este já ter sido aplicado a duas amostras, o que permitiria uma comparação, mesmo que imperfeita, dos resultados obtidos. Ademais, por ser o trabalho de referência sobre o assunto, é possível que *surveys* futuros também o tomem como base, o que facilitaria comparações posteriores. Como exposto na seção anterior, o questionário original dos autores era composto das dezenove perguntas elencadas na tabela de número 04. Não obstante, antes de realizarmos a codificação de todos os artigos, tendo em vista possíveis redundâncias e dificuldades de classificação – como as apontadas por Wooldridge (2004), Hoover e Siegler (2008a), Spanos (2008) e Cobb (2009) – foi realizada análise prévia com a aplicação do questionário integral em uma amostra piloto, com o intuito de realizar ajustes, caso necessário.

A avaliação dos artigos se mostrou deveras trabalhosa. Todavia, com relação à redundância de perguntas aparentemente similares, verificou-se que tal crítica não se aplica; antes disso, elas permitem maior flexibilidade de classificação e ajudam a ***reduzir*** a subjetividade de codificação. Por exemplo, as questões de número 02, 03, 05, 10, 11, 12 e 13 decerto tratam de um mesmo tema: a preocupação com magnitudes. Entretanto, um artigo que discute o tamanho dos efeitos estimados em seu trabalho (questão 12) não necessariamente trará a discussão da literatura da área sobre o tema (questão 13) e pode também esquecer-se de apresentar as estatísticas descritivas (questão 02). Por outro lado, um artigo que apresente as estatísticas descritivas (questão 02), e interprete cuidadosamente os coeficientes estimados (questão 05), pode não discutir a magnitude dos efeitos encontrados em seu trabalho (questão 12). E assim por diante. Tivéssemos apenas uma questão genérica sobre magnitude para classificar os artigos, a codificação estaria mais sujeita a variações por conta de pequenas nuances. E, diferentemente do que afirmam Hoover e Siegler (2008a), durante a aplicação do questionário ficou evidente que não é possível considerar questões sobre a utilização de formas economicamente interpretáveis e sobre a interpretação cuidadosa dos

coeficientes como meras "preferências pessoais" – elas se mostraram fundamentais, como será visto nos exemplos que serão dados mais adiante.

Da mesma maneira, as questões referentes ao uso da significância estatística, de números 07, 15, 16 e 19, por exemplo, permitem maior flexibilidade na classificação dos artigos. Isto ocorre porque, de fato, há trabalhos em que a significância estatística aparece como o ponto principal quando primeiramente utilizada (questão 07) e, além disso, a palavra "significante" é utilizada no decorrer do texto de maneira ambígua (questão 19); entretanto, o autor preocupa-se em trabalhar outros argumentos além da significância estatística, recebendo "sim" para a questão 15. Ademais – e o que é mais importante – independentemente das respostas anteriores, a significância estatística pode ter sido, ou não, o elemento decisivo do ponto de vista empírico do artigo. Note, assim, que as questões não perguntam se os autores "[...] confundem significância estatística e econômica [...] na p.1, p.2, p.3 ... p. n [...] transformado um erro em vários", conforme colocaram Hoover e Siegler (2008a, p.05). Se todas as perguntas fossem juntadas em apenas uma, casos como o citado acima seriam, na verdade, mais difíceis de serem codificados. Isso ficará mais claro com os exemplos trazidos na discussão das questões.

Já com relação às dificuldades de classificação – ou ganho marginal pequeno com sua codificação – algumas questões foram excluídas (ou substituídas). Antes de discutirmos estas questões, é muito importante, desde já, deixar claro que a intenção aqui ***não é,*** de forma alguma, criticar pessoalmente os autores dos trabalhos analisados. A intenção é expor ***práticas muito comuns e difundidas na profissão***, já feitas inclusive pelo próprio autor desta dissertação (CINELLI, 2010; 2011). Por conseguinte, a identidade dos autores será preservada tendo em vista que, na maior parte dos casos, a citação ou o exemplo escolhido é apenas um entre vários semelhantes que também poderiam ter sido elencados[76]. Igualmente, vale lembrar que ***não*** é apenas por conta de um erro ou um acerto citado que um artigo deve ser considerado bom ou ruim.

No que segue, os artigos analisados serão citados por numeração criada para referência desta dissertação, que não necessariamente guarda relação com a ordem ou ano de publicação. Vejamos as questões ***excluídas*** do questionário final.

[76] A fonte de alguma citação em particular ou a classificação do artigo serão fornecidas mediante solicitação ao autor.

Q1) O artigo utiliza um número pequeno de observações, de tal forma que diferenças estatisticamente significantes não foram encontradas meramente por a amostra ser grande?

A justificativa para este questionamento é a de que, com amostras muito grandes, praticamente qualquer coeficiente será estatisticamente diferente de qualquer hipótese pontual, justamente por se ter um menor erro padrão amostral e poder se estimar com maior precisão. Neste caso, portanto, usar a significância estatística da variável como ***fator decisivo*** de uma análise é ainda mais irrelevante. Sabe-se, de antemão, que virtualmente todo coeficiente será estatisticamente diferente de zero – ou de praticamente qualquer outro valor exato – à medida que $n \rightarrow \infty$.

Ziliak e McCloskey (1996, p.101-102) trazem o exemplo de um artigo publicado na *American Economic Review* cuja amostra era de 34.414 casas e 46.004 pessoas; entretanto, a despeito do grande número de observações, o artigo se ancorou nos padrões usuais de significância estatística. Conforme lembram Ziliak e McCloskey (2008a, p.67), "[...] com amostras deste tamanho, uma variável que é economicamente irrelevante aparecerá como estatisticamente significante pela simples força de um grande N".

A relevância desta pergunta – que a princípio pode soar estranha, como manifestaram Hoover e Siegler (2008) – reside no fato de que se basear apenas na significância estatística aos níveis usuais em uma situação de amostras enormes pode ser forte indício de confusão entre esta e a significância econômica. Durante a aplicação na amostra piloto, apenas um artigo apresentou amostra grande o suficiente para chamar a atenção. Trata-se do artigo de número 10, que buscou verificar o efeito da liberalização comercial no crescimento da produtividade brasileira. Para tanto, os autores utilizaram dados em painel da Pesquisa Industrial Anual do IBGE, com 2.072 firmas, de 1988 a 1998, totalizando 17.736 observações. Porém, mesmo neste contexto de amostra grande, o teste de significância estatística, aos níveis usuais, foi determinante na análise dos autores. Já na amostra completa, verificaram-se números de observações ainda maiores, como o caso do artigo de número 19, com 502.627. Apesar disso, a significância estatística aos níveis usuais ***foi determinante*** nas análises.

Nota-se, assim, que a essência da pergunta é pertinente. Todavia, apesar de em muitos casos ser fácil determinar quando o tamanho da amostra é "grande", sem um completo

entendimento do tema e das circunstâncias de cada trabalho analisado, o "quão grande" uma amostra tem de ser para ser considerada "grande" pode ser difícil de julgar – para alguns temas, uma amostra de 1.000 ou mais observações é mais do que suficiente para reduzir a preocupação com o erro amostral; para outros casos, 1.000 pode ser número irrisório. Deste modo, ***substituiu-se*** esta questão por uma mais abrangente: "***o artigo discute o nível de significância adotado tendo em vista as circunstâncias em que os testes serão aplicados?".*** Esta redação incorpora tanto a preocupação inicial de Ziliak e McCloskey, quanto todos os demais fatores que levariam os autores dos artigos a pensar acerca do nível de significância adotado.

Q4) A hipótese nula adequada foi feita?

Esta questão é um pouco mais delicada e pode ter uma interpretação mais subjetiva, sujeita à discordância. Há, decerto, casos claros de erro na formulação da hipótese nula. Um exemplo é o ponto já alertado por Gelman e Stern (2006), de alegar que dois coeficientes são diferentes porque um é estatisticamente significante e o outro não – neste caso, a hipótese nula adequada seria formular que os dois são iguais, e não que cada um é diferente de zero e depois compará-los. Isto pôde ser observado no artigo de número 02 (p. 31-32), que buscou verificar a sensibilidade dos investimentos das empresas com ações negociadas na Bovespa ao seu fluxo de caixa.

Já um ponto mais complexo é com relação a qual deveria ser a hipótese nula e qual deveria ser a hipótese alternativa a serem estabelecidas no teste. Ziliak e McCloskey (1996, p.102) afirmam que é comum colocar como nula aquela hipótese em que o pesquisador acredita. Este posicionamento, como vimos, foi sugerido por Lehmann e Romano (2008); todavia, como também vimos, alternativa contrária foi proposta por Casella e Berger (2002). Nota-se, portanto, que não há consenso com relação à forma de estruturação da hipótese a ser testada. Logo, optou-se por ***excluir*** esta questão. Perceba que ***sua exclusão não prejudica o tema principal*** do trabalho, pois o abuso dos testes de significância estatística tem ocorrido ***independentemente*** de como a hipótese nula é formulada.

Q6) O artigo evita reportar todos os testes t's e F's ou erros-padrão quando tal informação é irrelevante?

A intenção de Ziliak e McCloskey (1996, 2004a, 2008a) com esta questão era buscar indícios de que o *software* econométrico esteja substituindo o julgamento do pesquisador. Com o avanço computacional, os *softwares* atuais calculam rotineiramente todos os testes *t's* contra a hipótese nula de que os coeficientes sejam zero, mas nem por isso estas informações são relevantes para o problema. Mais recentemente, por exemplo, Kramer (2011, p. 461-462) contabilizou a aplicação de testes nos artigos da *German Economic Review*, totalizando ***10.575*** testes de significância, cerca de ***1.000*** testes por volume. Kramer (2011, p. 461), diante do resultado, endossou a constatação de Ziliak e McCloskey (2008a, p.112) de que "[...] testes *t* baratos, tornando-se cada vez mais baratos com a redução dos custos computacionais, têm, em equilíbrio, um produto marginal científico igual ao seu custo".

Não obstante a intenção dos autores, em nossa amostra piloto tal prática mostrou-se ***tão disseminada*** que acabou tornando-se difícil discernir um artigo que se preocupou, em algum grau, em não reportar tudo, de um artigo que não se preocupou com a questão e reportou todos os testes automaticamente. Dessa forma, ***excluiu-se*** esta questão do questionário final a ser aplicado para se obter ***maior parcimônia*** na codificação, sem perdas relevantes na informação obtida.

Q14) O artigo evita escolher variáveis para o modelo unicamente por meio de significância estatística?

A exclusão de uma variável do modelo apenas por não obter significância estatística aos níveis usualmente estabelecidos, sem levar em conta seus possíveis efeitos, pode deixá-lo mal especificado e, consequentemente, as estimativas serão enviesadas. Por exemplo, o artigo de número 01 (p.13) o fez, ao desconsiderar a análise de outro modelo de concorrência por votos devido à insignificância estatística das variáveis:

> [...] também foi considerada a inclusão da diferença entre as taxas de crescimento do produto estadual e nacional, com o intuito de testar se os eleitores avaliam os governantes com base na performance econômica relativa, em consonância com os modelos de "yardstick competition". Entretanto, nenhuma dessas variáveis apresentou significância estatística e, por este motivo, foram excluídas dos modelos.

Também admitiram terem feito o procedimento os artigos de número, 03, 30, 34, 39, 49, 59 e 60. É importante mencionar que alguns destes trabalhos utilizaram explicitamente uma abordagem *general-to-specific*, ou similar, para determinar a quantidade de defasagens das variáveis em sua regressão. Evidentemente que esta se trata de abordagem qualitativa e filosoficamente diferente de uma de exclusão de variável *ad-hoc* (desde que devidamente ajustados os níveis de significância) e, assim, poderíamos não penalizar o procedimento.

O problema principal nesta questão foi, na verdade, a confiança na sua apuração: como visto, apenas podemos saber se o autor realizou tal procedimento ***se este admitir explicitamente*** em seu texto. Sabe-se que a busca por variáveis "significantes", ou especificações que as tornem "significantes", é bastante comum[77]; mas, nem todos que realizam esta "busca" esclarecem o que foi feito no artigo – acabaríamos, desta maneira, contabilizando apenas quem foi honesto e não a extensão da prática. Destarte, o indicador que teríamos ***seria bastante inexato***, e optou-se por ***excluir*** esta questão.

Q17) O artigo utiliza simulação para verificar se os coeficientes são razoáveis?

Houve bastante dificuldade de verificar este ponto. Conforme Ziliak e McCloskey (2008a), a intenção não era ser rigoroso, mas apenas verificar se os autores fizeram exercícios de simulação, quaisquer que fossem, para averiguar a razoabilidade das estimativas. Todavia, como pertinentemente colocou Wooldridge (2004, p.577-578), esta pergunta não parece ser tão justa com muitos artigos, tendo em vista que pode não ser trivial pensar em como realizar uma simulação. Dessa forma, optou-se por ***excluí-la*** do questionário final.

Q18) Nas conclusões ou considerações finais, a significância estatística é separada da significância econômica, política ou científica?

As questões de número 07, 15, 16 e 19 se mostraram suficientes para classificar os artigos quanto ao foco na significância estatística. Na maior parte das vezes, as conclusões ou considerações finais são um pequeno resumo do artigo e esta questão

[77] Como afirmou Wooldridge (2006, p.612), "[...] virtualmente todos os pesquisadores examinam vários modelos antes de encontrar o "melhor" deles". Ou, ainda, Goldberger (1989, p.153) "[...] você lê um artigo e você sabe que houve uma muita "pesca" e você é cético quanto aos resultados reportados". Vide também Abelson (1995), Leamer (1983) ou Kramer (2011).

agregaria pouco aos principais resultados. Assim, optou-se por ***excluí-la*** para a ***parcimônia*** da codificação.

***

Como os próprios autores enfatizam (2008a, p. 73), o objetivo do questionário não é ser uma descrição completa da melhor prática em trabalhos empíricos; antes de tudo, busca-se focar nas perguntas que tratam da preocupação com a ***relevância econômica*** das estimações, isto é, da preocupação com as ***magnitudes*** e adequação econômica dos modelos estimados; e, principalmente, identificar se a análise de inferência estatística e o julgamento científico do pesquisador estão sendo substituídos por ***um ritual***.

Além da exclusão das questões mencionadas anteriormente, foram incluídas mais duas que os próprios Ziliak e McCloskey (2008a, p.66) se arrependeram de não terem feito em seu questionário original – uma referente à utilização de ***intervalos de confiança*** e outra referente à preocupação com ***erros de especificação (ou adequação estatística)***. Ficamos, assim, com dezesseis questões, as quais, buscando tornar a exposição mais clara, foram separadas em ***dois grandes grupos,*** com oito questões em cada. O primeiro grupo refere-se às perguntas mais intimamente ligadas à preocupação com ***magnitudes***. São elas (o número em parênteses se refere à numeração original de Ziliak e McCloskey):

GRUPO A – Questões de "Magnitude"

> ***A1 –*** *(Q2) As unidades e estatísticas descritivas das variáveis são devidamente apresentadas?*
>
> ***A2 –*** *(Q3) Os coeficientes são apresentados em elasticidades ou outra forma relevante para o problema, consistente com a teoria econômica, de forma que os leitores possam discernir o impacto econômico das variáveis?*
>
> ***A3 –*** *(Q5) Os coeficientes são cuidadosamente interpretados?*
>
> ***A4 –*** *(Q12) O artigo discute o tamanho dos coeficientes em algum momento?*
>
> ***A5 –*** *(Q13) O artigo discute a "conversa científica" na qual um parâmetro seria considerado grande ou pequeno?*
>
> ***A6 –*** *(Q10) O artigo evita a "econometria do asterisco"?*
>
> ***A7 –*** *(Q11) O artigo evita a "econometria do sinal"?*
>
> ***A8 –*** *(incluída) O artigo constrói os intervalos de confiança, utilizando-os para interpretar a significância econômica e não meramente como substituto da significância estatística?*

Já o segundo grupo aborda as perguntas que correspondem mais diretamente ao exercício automático de um ***ritual de inferência estatística***, sem ponderação acerca das circunstâncias da aplicação e sem se dar conta de sua extensão como argumento empírico. São elas:

GRUPO B – "Culto" da significância estatística

> ***B1*** *– (substitui Q1) O artigo discute o nível de significância adotado tendo em vista as circunstâncias em que os testes serão aplicados?*
>
> ***B2*** *– (Q8) O artigo menciona o poder do teste?*
>
> ***B2.1*** *– (Q9) Caso o artigo mencione o poder do teste, faz algo com relação a isso?*
>
> ***B3*** *– (Q7) A significância estatística, quando primeiramente utilizada, é apenas um entre outros critérios de importância?*
>
> ***B4*** *– (Q15) Depois do ponto principal do artigo, este evita usar a significância estatística como o critério de importância científica?*
>
> ***B5*** *– (Q16) O artigo considera mais do que a significância estatística para um argumento decisivo do ponto de vista empírico?*
>
> ***B6*** *– (Q19) O artigo evita usar a palavra "significante" em sentidos ambíguos, como significância estatística e influência econômica?*
>
> ***B7*** *– (incluída) O artigo demonstra preocupação com a especificação ou com a adequação estatística do modelo?*

Cabe esclarecer que não analisaremos ***se*** os modelos estavam estatisticamente adequados para a realização dos testes de significância, conforme sugerido por Spanos (2008) e Kramer (2011). Tal análise demandaria a base de dados dos autores e, muito provavelmente, conforme já constataram Kramer *et alii* (1985), ***revelaria situação ainda pior.*** Partiremos do pressuposto de que se o autor do artigo analisado faz inferência estatística é porque acredita que isto seja adequado na situação da análise. Entretanto, como exposto, a questão B7 refere-se à ***manifestação*** do autor com relação à adequação estatística do modelo utilizado.

Os critérios escolhidos para as questões, conforme apontam Ziliak e McCloskey (2008a, p.66), não são, em princípio, controversos. São padrões que seriam considerados minimamente razoáveis por inclusive aqueles que não concordam com a virtual inutilidade dos testes de significância estatística ***na forma como estão sendo aplicados atualmente.*** Todavia, estas são codificações que envolvem uma escolha dicotômica entre "sim" e "não" e, evidentemente, sempre haverá alguma subjetividade. Como colocaram Ziliak e McCloskey (2008a, p.66) "[...] outro economista pode codificar este

ou aquele artigo de uma maneira um pouco diferente. Nós não achamos que isso irá afetar de maneira relevante nossas descobertas [...]". Após a conclusão da análise dos artigos da RBE, este trabalho conclui de maneira semelhante– a prática é muito disseminada para que os principais resultados sejam afetados por algumas (ou até muitas) discordâncias de codificação.

Ainda assim, para evitar críticas como as formuladas por Hoover e Siegler (2008) sobre uma possível subjetividade exagerada na análise das questões, cada uma delas será discutida, buscando-se trazer exemplos ilustrativos da amostra analisada. Desta forma, discordâncias quanto às classificações podem ser esclarecidas e debatidas com maior transparência. Iniciemos pelo ***Grupo A***.

3.2.2. GRUPO A – Questões de magnitude

A1 – (Q2) As unidades e estatísticas descritivas das variáveis são devidamente apresentadas?

Se o leitor do artigo não sabe quais são as unidades de medida das variáveis, é simplesmente impossível interpretar o significado dos coeficientes estimados. Não obstante, a omissão – ou confusão – sobre as unidades de medida costuma aparecer em artigos publicados em revistas renomadas, como a já citada *American Economic Review*. Por outro lado, a apresentação de algumas estatísticas descritivas também é importante para se julgar a relevância dos coeficientes; afinal, o efeito estimado é muito ou pouco considerando a amostra analisada? O coeficiente é plausível? Essas perguntas são mais fáceis de serem respondidas tendo alguns parâmetros para comparação, como medidas de centralidade, dispersão ou assimetria.

Desse modo, ignorar tanto a unidade de medida quanto as estatísticas descritivas pode ser indício de um descuido com a análise ***quantitativa e econômica*** do problema – em outras palavras, se o autor sequer apresenta algumas estatísticas descritivas de seus dados, é provável que não esteja dando tanta atenção à análise da magnitude dos efeitos estimados. A análise não foi rigorosa neste quesito, bastava apresentar algumas médias ou outras medidas que o autor julgasse relevantes para receber "sim". Todavia, em nossa amostra, cerca de 40% dos artigos não as apresentaram.

Como um exemplo de trabalho que ***apresentou*** suas estatísticas descritivas de maneira interessante, temos o artigo de número 05. Os autores buscavam evidências acerca da convergência de renda entre os estados brasileiros, tomando como referência o estado de São Paulo. As principais características de seus dados foram resumidas com *Box Plot* – aliás, este foi o único artigo que utilizou a técnica, que permite em uma única imagem passar uma ideia da dispersão, assimetria, *outliers* entre outras características dos dados.

Alguns artigos trouxeram seção específica para a descrição dos dados e apresentação de algumas estatísticas descritivas, o que se mostrou prática interessante para facilitar a leitura e o entendimento das variáveis[78]. Entre eles podemos citar, por exemplo, o artigo de número 51, que buscou mensurar o impacto das alíquotas do imposto de importação sobre a evasão fiscal nas importações oriundas dos Estados Unidos, cuja seção 3 dedicou-se à apresentação dos dados e de suas estatísticas descritivas, resumidas em uma tabela. O artigo 58, que versou sobre os determinantes do milagre econômico brasileiro, também dedicou curta seção, antes dos resultados econométricos, à apresentação das estatísticas descritivas. Citemos ainda o artigo de número 64, que testou o modelo de Mincer para o Brasil. O trabalho dedicou seção para a explicação do desenho amostral e exposição de algumas estatísticas descritivas, além de elencar várias tabelas mais completas no apêndice.

Outro exemplo mais simples, mas não menos eficiente, é o artigo de número 09, que buscou testar três teorias diferentes sobre o comportamento do consumo (teoria do comportamento otimizador, teoria da restrição de crédito ou "miopia" dos agentes) para a realidade brasileira. O autor apresentou gráficos dos logaritmos das séries temporais utilizadas e, em uma seção em que buscava compreender os resultados econométricos, apresentou um gráfico de dispersão entre consumo e renda.

> ***A2 – (Q3) Os coeficientes são apresentados em elasticidades ou outra forma relevante para o problema, consistente com a teoria econômica, de forma que os leitores possam discernir o impacto econômico das variáveis?***

Muitas vezes, para se ter ideia da relevância do impacto das variáveis sobre o problema de interesse, é preciso apresentá-las em maneiras diferentes de sua unidade de medida

[78] Pode parecer algo trivial, mas em alguns artigos ***sequer foi possível definir a quantidade de observações*** (e consequentemente os graus de liberdade) utilizada nas regressões. Isto também foi constatado no levantamento de Keuzenkamp e Magnus (1995) no *Journal of Econometrics*. Também houve casos de não se conseguir distinguir ***a origem*** de alguns dados.

natural. Uma dessas formas, por exemplo, é a ***elasticidade***, bastante utilizada pelos economistas. Outra saída, quando as dimensões das variáveis são de difícil interpretação, pode ser a utilização dos ***coeficientes beta*** ou ***padronizados***. Ou, ainda, podem-se dar ***exemplos ilustrativos*** quando, mesmo apresentados em elasticidades ou outras formas, os efeitos não sejam diretamente óbvios.

Neste quesito, a maior parte dos trabalhos recebeu "sim". Um trabalho que ***não*** apresentou os coeficientes de forma fácil para a interpretação foi o artigo de número 01, que, buscando verificar os determinantes dos votos nas eleições estaduais brasileiras, apresentou, dentre outras, uma regressão *logit* (tabela 4), mas não calculou efeitos marginais ou outras formas de fácil interpretação – na realidade, este foi um problema para quase todos os artigos que trabalharam com modelos *probit* ou *logit*. Outro trabalho que pode ilustrar esta questão é artigo de número 04, que amplia o modelo de crescimento de Solow introduzindo a qualidade institucional como variável explicativa. Os autores criam um índice de desempenho institucional que é uma média de dois outros índices, um de "políticas sólidas" e outro de "abertura econômica". Assim, o significado do coeficiente da variável não é auto evidente e exercícios que facilitassem o discernimento do impacto econômico da variável para o leitor seriam bem-vindos.

A3 – (Q5) Os coeficientes são cuidadosamente interpretados?

Diferentemente da questão três, que apenas requeria a apresentação do coeficiente em formas em que o discernimento de seu impacto econômico fosse mais fácil – como em elasticidades – esta questão exige que o autor interprete em seu texto os coeficientes obtidos frente à amostra que obtém. Eles fazem sentido? Como se dá seu impacto? Deixar de se atentar ao significado dos coeficientes é, de fato, indício de que não há muita preocupação com a magnitude dos efeitos estudados, mas principalmente com encontrar ou não a significância estatística aos níveis usuais.

Iniciemos com uma ***exceção***, que foi o artigo de número 54. O trabalho buscou verificar alguns fatos estilizados sobre choques da política monetária no Brasil. Em certo momento, o artigo ***justificou*** porque não iria interpretar alguns dos coeficientes estimados:

> [...] nós não vamos apresentar e discutir as estimativas dos parâmetros do modelo devido às dificuldades associadas à sua interpretação, especialmente as estimativas da função de reação do banco central. Ver Christiano et alii.

> (1999) para uma discussão sobre as armadilhas na interpretação de regras de política monetária estimadas (p.137).

Evidentemente que os autores demonstraram preocupação com o tema, mesmo que esta preocupação os tenha levado a não interpretar o coeficiente. Deste modo, o artigo ***recebeu "sim".***

Vejamos alguns trabalhos que interpretaram seus coeficientes. O artigo de número 03 buscou verificar se o efeito *pass-through* do câmbio para a taxa de inflação no Brasil é afetado pelo ambiente de alta ou baixa inflação. Ao estimar as equações, o artigo (p.52) frisa o significado dos coeficientes apresentados, como, por exemplo: "[...] podemos interpretar os coeficientes como multiplicadores: uma depreciação, digamos, de 10% resultaria em 0,5% de inflação no próximo trimestre, e em inflação de 2,1% no longo prazo". Já o artigo de número 35 buscou verificar o impacto de choques de demanda e de oferta sobre a produção e preços agrícolas. Ao apresentar seus resultados, os autores são claros (p.49):

> [...] medimos que um aumento de 10% no rendimento iria - no longo prazo - elevar a produção em 4,8% e reduzir os preços em 1,6%, de modo que o rendimento agrícola acabaria aumentando em torno de 3,1%. Isto, obviamente, estimula o uso contínuo de novas tecnologias de aperfeiçoamento da produção [...] nós medimos que uma desvalorização de 10% na taxa de câmbio iria - no longo prazo - aumentar a produção agrícola em 3,7%.

O artigo de número 04, por sua vez, ***não*** interpreta o significado do principal coeficiente de sua estimação, referente ao impacto da qualidade institucional no crescimento econômico. Em uma passagem, dizem os autores "[...] todos os sinais estão iguais aos esperados pela teoria, agora com a influência expressiva das instituições (1,8923)" (p. 62). Perceba que o número citado, 1,8923, é o coeficiente estimado para a variável de interesse. Mas o que significa 1,8923? Os autores não interpretam o número. Mais a frente, ao comparar a estimativa do coeficiente de variáveis instrumentais (2,68524) com a estimativa de mínimos quadrados ordinários (1,89236), os autores simplesmente afirmam "[...] o impacto das instituições é maior levando em consideração outros fatores como históricos e geográficos" (p.64).

Também é interessante citar aqui o artigo de número 06 (p.170-173), pois demonstra como a simples menção do efeito não é, em muitos casos, uma interpretação cuidadosa do coeficiente estimado. Os autores buscam medir o impacto que o Sistema de

Seguridade Social Rural tem sobre a pobreza rural e estimam a primeira diferença da seguinte equação:

$$\ln[P_{k,it}] = \beta_0 + \beta_1 \ln[P_{k,it-1}] + \beta_2 \ln[PIB_{it}] + \beta_3 \ln[Apos_{it}] + \beta_4 \ln[Gini_{it}] + \beta_5 \ln[Ame_{it}] + \beta_6 \ln[Pesdes_{it}] + v_t + u_{it} \quad (20)$$

Onde,

> [...] $P_{k,it}$ representa o k-ésimo índice de pobreza, $k = 0,1,2$, PIB_{it} é o produto interno bruto agropecuário per capita, $Gini_{it}$ é o índice de Gini, $Apos_{it}$ são os valores *per capita* dos benefícios recebidos pelos aposentados, Ame_{it} são os anos médios de estudo dos indivíduos e $Pesdes_{it}$ é o número de pessoas desocupadas com mais de 10 anos respectivamente, v_t os efeitos fixos não observáveis dos indivíduos e u_{it} o erro idiossincrático. Os subscritos i e t se referem, respectivamente, a região rural do i-ésimo estado no ano t.

Como os próprios autores estabeleceram um modelo dinâmico, uma alteração permanente em alguma das variáveis explicativas, digamos, um aumento permanente de 10% sobre os anos médios de estudo, tem dois impactos: o impacto imediato em P_0, medido pelo coeficiente de elasticidade β_i, e um impacto de longo prazo tendo em vista o efeito da mudança de P_0 em P_1, deste último em P_2 e assim sucessivamente. Dessa forma, uma interpretação mais cuidadosa dos efeitos estimados teria de informar o leitor sobre esta diferença e, caso fosse irrelevante para o problema, apontar o porquê. Entretanto, os autores se atentam somente aos valores estimados de β. Além disso, os coeficientes representam efeitos *ceteris paribus*. Isto é, β_5, por exemplo, mede o efeito da educação sobre a pobreza mantendo-se tanto o PIB, quanto o desemprego, quanto a desigualdade de renda *constantes*. Ora, caberia aqui perguntar como um aumento da educação média traria uma redução na pobreza sem: (i) aumentar a produção; (ii) reduzir a desigualdade de renda; ou (iii) aumentar o emprego? Por que canais se daria este efeito[79]? Assim, por estes motivos, o trabalho ***recebeu "não"*** nesta questão.

[79] Um exemplo simples para ilustrar esta questão é o seguinte: se rodarmos um modelo de preço hedônico dos preços das casas contra o número de cômodos, é provável que o coeficiente seja ***positivo***: quanto mais cômodos, maiores os preços. Todavia, se rodarmos a regressão contra (i) o número de cômodos ***e*** (ii) o tamanho do imóvel, é provável que o coeficiente seja ***negativo***: dado o tamanho fixo do imóvel, quanto maior o número de divisões, menores os tamanhos dos quartos, e isso pode prejudicar o valor do imóvel. Note que uma interpretação cuidadosa do coeficiente da regressão teria de explicar porque o coeficiente é negativo neste caso. Analogamente, o valor medido pelos autores é o efeito da educação considerando PIB, desemprego e desigualdade constantes. O valor foi positivo, o que é o "senso comum" e talvez por isso não tenha sido interpretado mais a fundo. Note, entretanto, que o valor do número de cômodos com o tamanho do imóvel fixo também poderia sair positivo e daí caberia a pergunta: como mais cômodos,

A4 – (Q12) O artigo discute o tamanho dos coeficientes em algum momento?

Além de reportar o coeficiente de forma economicamente mais palatável e interpretar o impacto da variável cuidadosamente, o artigo discute o tamanho do coeficiente e sua significância econômica? O autor diz ao leitor do artigo por que aquele coeficiente estimado é importante? O artigo explica por que algumas variáveis são economicamente relevantes enquanto outras não são tão importantes assim, fundamentando-se no tamanho dos efeitos encontrados? São essas as perguntas que deveriam ser respondidas nesta questão.

Por exemplo, o artigo pode estimar uma função *log-log*, e apresentar o coeficiente 0,001, dizendo que se trata da elasticidade-preço do produto, ressaltando que uma variação de 1% na variável explicativa causa uma variação de 0,001% na variável dependente. Desse modo, receberia "sim" para a questão A2 e "sim" para a questão A3. Contudo, se o artigo não explicar se o tamanho deste efeito é ***grande ou pequeno***, ***importante ou negligenciável***, receberá "não" na presente questão, a despeito de ter imprimido o valor do coeficiente no papel – o importante aqui não é apresentar o valor, e sim ***discutir*** esse valor. Importante ressaltar que, caso a magnitude do coeficiente fosse irrelevante para o estudo (o que dificilmente é o caso), e isso ***tivesse sido explicitado no texto com a devida justificativa,*** o trabalho também receberia "sim".

Citemos primeiramente alguns trabalhos que ***fizeram*** a devida discussão. O artigo de número 03, como vimos, apresenta as magnitudes dos coeficientes de maneira clara e ressalta o impacto econômico do câmbio sobre a inflação, que depende do contexto inflacionário. Mais ainda, o autor contextualiza os resultados afirmando que

> [...] estão de acordo com a literatura, sugerindo algum grau de influência do ambiente inflacionário sobre o *pass-through* da taxa de câmbio no Brasil, que estimamos variar no longo prazo ***em cerca de 8% a 40%,*** dependendo do nível de inflação. Mesmo que este resultado não exclua outras possíveis explicações para o declínio do *pass-through* da taxa de câmbio observado no Brasil na década de 1990, ele ***sugere que o ambiente de menor inflação do período teve um papel importante em tal diminuição, e, portanto, foi erradamente negligenciada*** (p. 50).

Isto é, o tamanho do efeito estimado é discutido e o autor apresenta por que acredita ser importante, tendo por base a literatura da área.

---

mantendo o tamanho fixo, valorizariam o imóvel? Da mesma forma, cabe perguntar: como a educação por si só reduziria a pobreza mantendo os demais controles fixos?

Outro texto que faz esta discussão é o artigo de número 07, que também versa sobre o *pass-through* da taxa de câmbio na inflação brasileira. Os autores apresentam em que sentido suas estimativas – e seus efeitos – importam para a prática econômica, em termos de magnitude do repasse inflacionário, como na seguinte passagem (p. 239):

> [...] no segundo trimestre de 1999 (imediatamente após a flutuação cambial), por exemplo, apesar da depreciação de 39% da taxa de câmbio no trimestre anterior, a inflação livre foi de apenas 0,49% e a inflação situou-se em 1,05%, ambas abaixo dos valores dos trimestres anteriores***. Nesse período, o hiato foi abaixo do limite estimado (desaceleração econômica), o que implica, de acordo com o modelo, um baixo repasse para a inflação***. As depreciações no terceiro trimestre de 2000 e durante 2001, por sua vez, foram acompanhadas por ***aumentos maiores na taxa de inflação. Naquele período, o hiato foi mais elevado do que o limiar estimado***.

Uma discussão mais simples e direta foi do artigo de número 42, que buscou calcular a alíquota de contribuição de equilíbrio atuarial do para a previdência brasileira. Ao realizar suas estimativas, os autores concluem que (p.174),

> [...] a alíquota de 31%, ao contrário do que acontecia até a aprovação da legislação do fator previdenciário – que diminuiu o valor da aposentadoria em relação às regras anteriormente vigentes – ***parece ser relativamente elevada.***

Outro exemplo interessante é o artigo de número 64 que, já no resumo do trabalho, apresenta a discussão da magnitude dos resultados encontrados (p.407, grifo nosso):

> [...] ao relaxar tais hipóteses, estimamos as taxas internas de retorno de Becker (1975) e obtemos ***vieses que chegaram a 14 pontos percentuais*** em relação ao coeficiente "minceriano". Assim, ***a magnitude destes retornos é bem menor*** do que os estudos baseados no modelo de Mincer.

E este foco permanece no decorrer do texto: o autor mostra como os modelos anteriores, baseados em Mincer, superestimavam em grande montante os retornos educacionais.

Podemos citar ainda o artigo de número 56. Os autores estimaram que os países que adotaram o regime de metas de inflação tiveram menores custos de desinflação, em termos de pontos percentuais do PIB, e julgaram o resultado encontrado como economicamente relevante (p. 185, grifo nosso):

> [...] neste artigo assumimos a tarefa de avaliar ***se metas de inflação são importantes*** para reduzir os custos da desinflação. Nossos resultados sugerem que ***sim***. A julgar pelo coeficiente de nossa especificação mais preferida, os países que adotaram metas de inflação ***pouparam cerca de 4 pontos percentuais no PIB (nível) em relação à sua tendência para cada ponto de queda da inflação***. ***Este é um efeito economicamente importante***. Além disso, a adição de uma dummy de mercado emergente interativa para o nosso modelo revela que os benefícios das metas de inflação parecem ser maiores para as economias desenvolvidas.

Citemos agora alguns trabalhos que ***não fizeram*** a devida discussão. O artigo de número 08 (p. 254) exemplifica a mera reprodução dos valores encontrados na estimação, sem qualquer discussão acerca da relevância da magnitude dos efeitos. Sobre os possíveis determinantes de as mulheres investirem em ações, os autores apenas transcrevem os resultados de sua tabela da estimação *Probit*:

> [...] o fator educação (EDUC) também é estatisticamente significante. O sinal é positivo e quanto maior o nível de escolaridade, maior a probabilidade de investir em ações. Há 64,53% de probabilidade de mulheres com pós-graduação investirem em ações. A variável RELIG é estatisticamente significante e com sinal positivo. Mulheres que interagem num grupo formado pela mesma religião apresentam 37.64% maior probabilidade de investir em ações. A interação com vizinhos, grupos e associações ou clubes e academias não são estatisticamente significantes, conforme resultados encontrados por Hong et alii (2004). A variável DEPRE é estatisticamente significante e com sinal negativo. Mulheres que responderam sentir algum sintoma de depressão diminuem a probabilidade de investir em ações em 11,13%,

Cabe notar também o descuido na interpretação dos efeitos, que são apresentados como probabilidades incondicionais, o que ensejou o "não" na questão A3, apesar de o artigo ter obtido o "sim" na questão A2.

Exemplo semelhante pôde ser visto no artigo de número 16. O trabalho buscou medir a eficácia das intervenções do Banco Central sobre a volatilidade cambial. Em uma passagem, os autores afirmam que (p.84-85):

> [...] o único instrumento de intervenção que afetou a volatilidade da taxa de câmbio nominal foi a intervenção via títulos cambiais, cujo valor foi significativo ao nível de 10%. Seu coeficiente é negativo (igual a –0.000165), indicando uma redução da variância condicional da taxa de câmbio nominal.

Note que o valor estimado (–0.000165) não é discutido. Temos também o artigo de número 21 (p. 188) que buscava verificar a relação entre a distribuição dos rendimentos do trabalho e a escolaridade dos pais dos trabalhadores. Os autores apresentam, em vários momentos, os coeficientes estimados, mas não os discute. Por exemplo:

> [...] diferença observada entre o logaritmo da média de rendimentos para indivíduos com mães mais escolarizadas e trabalhadores cujas mães não alcançaram 4 anos de estudo é igual a 1,19. De acordo com o resultado da decomposição, diferenças nas características entre indivíduos nos dois grupos são responsáveis pela variação em 0,80, enquanto 0,39 do hiato de rendimentos pode ser atribuído a diferenças nos coeficientes estimados. A parcela correspondente a diferença entre fatores não-observados é próxima de zero. Já a diferença observada entre o logaritmo dos rendimentos dos trabalhadores com mães mais escolarizadas e aqueles cujas mães completaram entre 4 e 7 anos de estudo é igual a 0,55. Nesse caso, a diferença de 0,43 é atribuída às características dos indivíduos nos dois grupos, enquanto as diferenças nos coeficientes são responsáveis por 0,12. Mais uma vez, a diferença nos fatores não-observados se mostra irrelevante.

A5 – (Q13) O artigo discute a "conversa científica" na qual um parâmetro seria considerado grande ou pequeno?

Nesta questão, o artigo deveria apresentar a literatura pertinente sobre o assunto e a discussão científica sobre os efeitos esperados; elencar o debate prévio sobre a relevância ou irrelevância das relações que busca encontrar, frisando magnitudes. Todavia, caso o artigo mencionasse que não há estudo quantitativo prévio acerca do tema, receberia "sim" nesta questão por justamente abrir precedente na literatura de sua área. De fato, este foi o caso do artigo de número 03 (p.50, grifo nosso) quando afirmou: "[...] ***com que extensão*** um ambiente de menor inflação contribuiu para este declínio [do *pass-through* cambial] é uma questão ***que ainda precisa ser investigada***". Perceba que o autor frisa uma questão de magnitude, isto é, o que interessa não é saber apenas ***se*** a inflação influencia o *pass-through*, mas principalmente ***quanto***. Temos ainda o artigo de número 50, que ao avaliar o grau de desigualdade educacional no Brasil, constatou que

> [...] surpreendentemente, até onde vai nosso conhecimento, ***nenhum estudo deste tipo parece ter sido feito para o Brasil***. Nosso objetivo neste trabalho é contribuir para ***preencher esta lacuna:*** nós fornecemos uma avaliação da desigualdade de escolaridade no Brasil, utilizando diferentes indicadores, tais como o coeficiente de Gini da educação, o desvio padrão da educação e do número médio de anos de escolaridade (p.32, grifo nosso).

Podemos citar também o artigo de número 51, que constatou não existir trabalho que estimasse a magnitude da evasão fiscal no imposto de importação:

> [...] ***embora não existam, a nosso conhecimento, trabalhos que estimam a magnitude da evasão do imposto de importação***, é razoável supor que o sistema tarifário, inserido no caótico sistema tributário nacional, presta-se bem à sonegação, como sugerem notícias freqüentes na imprensa, envolvendo grandes importadores, que, supostamente, teriam sido flagrados praticando evasão [...] o objetivo desse trabalho é, pois, investigar o impacto das tarifas sobre a evasão fiscal nas importações brasileiras oriundas dos Estados Unidos da América (p.79).

O artigo de número 07, por sua vez, além de ter discutido a magnitude das estimativas e sua importância, também trouxe exemplos da literatura que ajudariam o leitor a colocar as estimativas em perspectiva. Ao encontrarem uma diferença de grande magnitude entre o *pass-through* do câmbio para a inflação em períodos de baixa volatilidade e alta volatilidade cambial, de 80% para 7% respectivamente, os autores trazem dois trabalhos que também encontram mudanças drásticas. Conforme se lê na página 239,

> [...] Muinhos e Alves (2003), por exemplo, encontraram uma redução de coeficiente de 51% para 6% após a mudança no regime cambial, e

> Albuquerque e Portugal (2005), utilizando um modelo de filtro de Kalman, têm valores estimados de cerca de 42 % e 4%, respectivamente.

O artigo de número 11 (p. 301) igualmente compara os valores estimados em sua pesquisa com a literatura internacional da área:

> [...] os coeficientes estimados para A1 e A2 sugerem que o fator de desconto é de 0,69, menor do que o de 0,92 encontrado por Blanchard e Gali (2007) para a economia dos EUA. O valor mais baixo explica-se pelo fato de que a economia brasileira tem tido uma das maiores taxas de juros real no mundo. A economia dos EUA, por outro lado, está entre os países com menor taxa de juros real.

Ou ainda, após encontrar a rigidez salarial de 92% para o Brasil, valor que afirmam ser próximo ao calibrado para a economia americana no trabalho de Blanchard e Gali – que foi utilizado como referência – os autores discutem o resultado frente a outros trabalhos da área, afirmando que o dado

> [...] está de acordo com Camargo e Reis (2007), que argumentam que a recente estabilização da inflação aumentou a rigidez do salário real, porque a inflação permite uma maior flexibilidade para o salário real em situações onde o salário nominal não está legalmente autorizado a ser reduzido. Resultados por Arbache e De Negri (2004) e Orellano et alii. (2009), utilizando micro-dados, também sugerem que a estrutura salarial brasileira é rígida e insensível ao ciclo econômico.

Um bom exemplo é a revisão encontrada no artigo de número 44, que tratou sobre urbanização e diferença de rendimentos, utilizando casos brasileiros. Os autores trazem as estimativas encontradas na evidência internacional:

> [...] Acemoglu et alii. (2002) mostram a relação entre urbanização e renda por meio de regressões do logaritmo natural da renda per capita em função da taxa de urbanização. Os resultados encontrados apontam que, em 1995, um país com uma taxa de urbanização 10 pontos percentuais superior tinha, em média, uma renda per capita 43% mais elevada. Os autores fazem, ainda, um exercício empírico, a partir de dados históricos, considerando um painel para diferentes países no período de 1750 a 1913, incluindo dummies de países e período. Nesse caso, um país com taxa de urbanização 10 pontos percentuais mais elevada, tinha, em média, uma renda per capita 30% maior. [...] Glaeser and Mare (2001) encontram um prêmio salarial médio de 24,9%, controlando para características individuais, para aqueles que residem em áreas densamente povoadas. Quando incluem efeitos fixos para cada indivíduo [...] as estimativas de prêmio salarial nas grandes cidades ficam em torno de 10,9 e 4,5%, dependendo da base de dados utilizada.

Todavia, vale mencionar que, infelizmente, este mesmo artigo pouco discutiu os próprios valores encontrados, recebendo "não" na questão anterior, e priorizou a significância estatística e a direção dos efeitos como argumento empírico, recebendo "não" na questão B5.

Outro bom exemplo é retirado do artigo de número 20 (p.150, grifo nosso), que busca identificar qual o peso que os consumidores que utilizem uma regra de comportamento "rule of thumb" têm na economia. Revisando a literatura da área, o autor menciona que

> [...] o debate sobre ***a importância quantitativa*** do comportamento "rule of thumb" está longe de ser resolvido. Alguns estudos sugerem que consumidores "rule of thumb" respondem por uma parcela grande da renda disponível. Campbell e Mankiw (1989, 1990) mostram que aproximadamente 50% da renda disponível é de consumidores "rule of thumb".

O artigo também cita outros estudos que encontram uma importância quantitativa para o "rule of thumb", com valores em torno de 30% a 60%; por outro lado, haveria também aqueles trabalhos que não encontram participação tão relevante, com estimativas de 15% a 23% (p.157). Assim, no decorrer do texto, o artigo de número 20 compara suas estimativas com a literatura elencada (p.157, grifo nosso):

> [...] tanto as estimativas por OLS quanto por IV sugerem que algo em torno de 70% a 80% da renda disponível é de domicílios "rule of thumb" [...] ***esses valores são consideravelmente mais elevados do que o que foi encontrado anteriormente na literatura.***

Citemos ainda o artigo de número 64, que em momentos do texto compara os resultados de suas estimativas da taxa de retorno da educação no Brasil com os obtidos na literatura, como na seguinte passagem (p.422):.

> [...] destacamos ainda que as TIRs da abordagem dos anos de estudo são próximas das obtidas por Barbosa Filho e Pessoa (2008), com exceção do ginasial e do superior que diferiram em maior magnitude. Além disso, apresentam alguma semelhança com os estudos baseados no modelo de Mincer, como Blom e Verner (2001), pois estes autores obtiveram retornos do primário e ginásio bem menores em relação ao secundário e superior.

Tratemos agora de trabalhos que trouxeram uma discussão da literatura de seu tema, ***mas sem qualquer menção à magnitude***, recebendo, portanto, ***"não".*** Lembre-se que, como vimos na seção 2.1.1., não basta elencar se um artigo anterior constatou ou não a significância estatística de uma variável, ou apenas elencar a direção do efeito estimado, pois isto pode ser bastante enganoso. A revisão feita pelo artigo de número 06 (p.162, grifo nosso) preocupou-se apenas com o aspecto qualitativo, sem fornecer bases para uma comparação quantitativa dos resultados que foram obtidos, como na seguinte passagem:

> [...] Hoffmann (2006), por exemplo, verificou que as aposentadorias e pensões pagas pelo governo federal no Brasil ***constituem um "freio***" para a redução da pobreza, pois no período 1998-2005 os efeitos das aposentadorias contribuíram para aumentar a desigualdade no período. Por outro lado, Schwarzer (2000) discorda dessa visão ao analisar os impactos

> socioeconômicos do sistema de aposentadoria rural [...] o referido autor afirma que o programa de aposentaria rural brasileiro ***é eficiente*** no combate à pobreza, como talvez poucos outros consigam ser no mundo.

Note que ***não se obtém informação quantitativa para comparação*** do efeito estimado das aposentadorias rurais sobre a pobreza com as pesquisas já realizadas na literatura. Sabe-se apenas que um estudo conclui que as aposentadorias não são eficientes e que o outro conclui que são.

O artigo de número 32, que estuda as respostas assimétricas dos estados brasileiros a choques de política monetária, procede de maneira similar. Ao trazer a revisão de literatura, os autores trazem os resultados qualitativos, como nesta passagem (p.415, grifo nosso):

> [...] Araújo Jr. (2004), com o objetivo de comparar as respostas dos produtos da Região Sul com a Região Nordeste a choques monetários, estima um modelo VAR, e com base na função de impulso-resposta obtém indicações de que ***a Região Sul é mais fortemente influenciada*** pela política monetária. Bertanha e Haddad (2008) estimam um modelo VAR, controlando pela dependência espacial entre os estados, para analisar a repercussão da política monetária sobre diferentes unidades geográficas do país. Eles utilizam a variável "emprego" como proxy do nível da atividade produtiva. Os resultados dos autores indicam que o nível de emprego dos estados localizados nas Regiões Norte e Nordeste ***é mais vulnerável*** a choques na política monetária, e esse ***impacto é mais significativo*** quando não é considerada a inter-relação de dependência espacial entre as unidades da federação.

A6 – (Q10) O artigo evita a "econometria do asterisco"?

Segundo Ziliak e McCloskey (1996, 2004a, 2008a), a "econometria do asterisco" ocorreria quanto o artigo classifica os coeficientes estimados segundo o ***tamanho de sua estatística*** de teste, ao invés de se atentar ao tamanho e à importância de seu ***efeito***[80].

O artigo de número 01 (p. 13), por exemplo, classifica os coeficientes como "significantes", "marginalmente significantes" e "insignificantes", como no trecho a seguir:

> [...] observe que a variável oportunismo mostrou-se significante em todos os casos, evidenciando que os eleitores respondem positivamente aos ciclos oportunistas nas variáveis fiscais [...] a parcela de competência dos ciclos, por outro lado, mostrou-se marginalmente significante apenas nos casos das

[80] É importante ressaltar que a econometria do asterisco ***não é*** o ato de colocar asteriscos nos coeficientes e não haveria problema em fazer isso. Trata-se de se classificar a importância econômica do coeficiente pelo tamanho de sua estatística.

> receitas total e corrente, enquanto que a parcela prevista das variações fiscais mostrou-se insignificante em todas as regressões.

De maneira similar, o artigo de número 56 (p.180) qualificou alguns coeficientes como "highly significant" ou "borderline significant".

Já o artigo de número 04 (p. 62), considera um dos modelos como bastante significativo, por apresentar todos os parâmetros "[...] estatisticamente diferentes de zero ao nível de 5%". Também os autores observam uma "queda na significância estatística da restrição do parâmetro" (p. 64) como evidência a favor da suposição de que os instrumentos sejam não correlacionados com o termo de erro. Ademais, o artigo classifica os modelos segundo sua estatística F ou grau de ajuste, ressaltando que no modelo (2) de seu trabalho, "[...] a estatística F (191,290) aumentou, comparando com o modelo (1) (139,560), assim como também a qualidade do ajustamento do modelo (0,8535) melhorou em comparação ao outro (0,7290)" (p. 62).

O artigo de número 60 (p.292) ao reestimar sua equação principal com novos instrumentos, enfatiza que na nova equação "[...] a significância estatística é consideravelmente mais alta para [o coeficiente de] construções". Como último exemplo, podemos citar a seguinte passagem do artigo de número 64 (p.417, grifo nosso):

> [...] sob todas as especificações estimadas, rejeita-se a hipótese nula de que os coeficientes nos termos não lineares sejam nulos. Além disso, para todas as especificações nota-se que ***o valor da estatística apresenta uma tendência de crescimento, o que nos leva a concluir que a hipótese de linearidade do modelo de Mincer tem se tornado cada vez mais inadequada.***

Como vimos no capítulo anterior, a estatística de teste por si só não é suficiente para indicar a inadequação do modelo. Por exemplo, a distribuição normal pode ser menos adequada a uma amostra pequena que rejeite a normalidade com *p-valor* de 0,09 do que a uma amostra enorme que rejeite a normalidade com *p-valor* de 0,01.

A7 – (Q11) O artigo evita a "econometria do sinal"?

A "econometria do sinal" ocorreria quando os autores preocupam-se apenas com a direção do efeito estimado. Ocorre que a direção, na maior parte das vezes, apenas importa se o efeito também for economicamente relevante. Uma elasticidade preço estimada em -0, 000001%*** é estatisticamente significante, com o sinal esperado pela teoria, mas, para a maior parte dos casos, poderia ser considerada igual a zero. Não

obstante, é comum encontrar nos estudos o descaso com a magnitude e a ênfase apenas na direção do coeficiente.

No artigo de número 13 (p.453), por exemplo, após a estimação de um modelo VAR com as variáveis de interesse, os resultados são apresentados da seguinte forma:

> [...] assim, ao analisar o modelo que relaciona o superávit primário às suas expectativas, nota-se que o efeito dos choques das variáveis sobre si próprias é positivo e temporário. A resposta do superávit primário às suas expectativas é desprezível. A resposta da expectativa do superávit primário a impulso no superávit primário é significativa e positiva entre o terceiro e o sétimo mês.

Isto é, em nenhum momento o artigo se preocupa com a magnitude, mas tão somente com a direção da resposta ao choque.

Outro exemplo pode ser retirado do artigo de número 10 (p. 283), sobre os efeitos da liberalização comercial na produtividade. Ao apresentarem os principais resultados, não é dada a devida atenção aos valores estimados dos coeficientes, e o foco principal é na significância estatística e na direção do efeito, como na seguinte passagem:

> [...] o coeficiente de participação de mercado é muito forte e significativo [...] o coeficiente da variável crescimento de capital foi estimado negativo [...] mais importante, o coeficiente de tarifas de importação de insumos foi preciso e negativo [...].

O artigo de número 17, que analisa o impacto da redução dos encargos trabalhistas sobre a formalização das empregadas domésticas, (p.106), ao expor os resultados de uma estimação *Probit*, também procede de maneira semelhante:

> [...] observa-se na Tabela 8 que os resultados são parecidos com a estimação utilizando dados da PME. As diferenças consistem em que nesse caso, ser chefe tem impacto positivo sobre a probabilidade de ser mensalista e ser cônjuge e ter mais idade impactam negativamente.

Os exemplos são muitos, como pode ser visto no artigo de número 21 (p.185):

> [...] a dummy para mães com educação entre 4 e 7 anos é positiva e significativa, mas a interação dessa variável com a escolaridade do indivíduo não se mostra significativamente diferente de zero. A dummy para pais com pelo menos 8 de estudo é negativa, como na coluna (4), enquanto a dummy para mães com 8 anos de estudo ou mais não é significativamente diferente de zero. No entanto, as interações entre essas dummies e a escolaridade do indivíduo são ambas positivas e significativas.

Ou no artigo de número 28 (p.345) "[...] a variável "crise" mostra-se significante e com sinal positivo, indicando que crises econômicas e/ou políticas elevam o risco de colapso do regime democrático – o que é bastante intuitivo". Também no artigo de número 48 (p. 415):

> [...] os resultados encontrados para uma regra de Taylor padrão encontram-se na primeira coluna da Tabela 4. Os termos de suavização de taxa de juros de primeira e de segunda ordem são significantes a 1%. O coeficiente do hiato do produto possui sinal correto, mas não é estatisticamente significante. O coeficiente da variável de desvio da expectativa da meta possui o sinal correto e é estatisticamente significante a 10% de significância.

Ou, ainda, no artigo de número 61 (p.320)

> [...] pode-se notar que os efeitos da aposentadoria domiciliar per capita sobre a taxa de participação são negativos e significativos para os não-qualificados e os semi-qualificados, ou seja, aumentos na aposentadoria domiciliar per capita reduzem a taxa de participação. Para os qualificados os coeficientes não são significativamente diferentes de zero. Os resultados também mostram que para os não-qualificados maiores taxas de inflação levam a aumentos na participação, mas as variações no PIB não são significativas. Já para os semi-qualificados e os qualificados os coeficientes da taxa de inflação são não-significativos, enquanto as variações no PIB passam a apresentar efeitos positivos sobre a taxa de participação.

Esta forma de exposição, com pequenas variações, repetiu-se em diversos artigos de nossa amostra.

A8 – (incluída) O artigo constrói os intervalos de confiança, utilizando-os para interpretar a significância econômica e não meramente como substituto da significância estatística?

O uso de intervalos de confiança chama a atenção para a magnitude dos coeficientes; contudo, sua mera construção com base em um nível de significância arbitrário, para verificar se a hipótese nula pertence ao intervalo, não é considerada nesta questão.

Surpreendentemente, ***dos 67 artigos analisados, apenas 1*** preocupou-se em construir intervalos de confiança para a análise das magnitudes. Trata-se do artigo de número 33 (p. 18), que buscou estimar a disposição a pagar para reduzir o risco de morte associado à poluição do ar em São Paulo. Outros artigos que construíram intervalos de confiança utilizaram-nos apenas como substitutos dos testes de significância. Por exemplo, o artigo de número 13 não obteve nenhuma informação do intervalo a não ser o fato de este abranger ou não o valor zero. O artigo de número 05, da mesma forma, teve como objetivo somente verificar se o intervalo continha o valor unitário para o parâmetro de interesse. Já o artigo de número 14 chegou a mencionar que o parâmetro estimado estava aproximadamente entre 0 e 2%, mas, por fim, a única função do intervalo de confiança foi verificar se este excluía valores negativos.

Em geral, a "análise de robustez" de todas as estimativas, quando feitas, era assentada na ***permanência da significância estatística*** da variável, e não na variabilidade da magnitude estimada. Em nenhum momento, em nossa amostra, o pesquisador parou para se perguntar sobre a amplitude das estimativas, como, por exemplo, questionar se as magnitudes de um mínimo ou de um máximo do intervalo de confiança levariam a conclusões diferentes a respeito da ***significância econômica*** da variável.

Por fim, cabe citar uma ***exceção,*** o artigo de número 15. O trabalho buscou verificar como choques nas taxas de juros e na taxa de câmbio impactam na distribuição dos preços relativos do IPCA, desagregando o impacto em 512 subitens. O texto ***não*** construiu os intervalos de confiança, mas ***justificou*** por que não o fez. Nas palavras dos autores (p.56, grifo nosso),

> [...] toda a análise que se segue está baseada na estimação pontual dos coeficientes, ***sem considerar o intervalo de confiança*** sobre o qual se fazem as previsões. Optamos por proceder desta forma, em primeiro lugar, pelo fato de que, por estarmos tratando de 512 índices de preços, uniformizamos as regressões, ***sem nos preocupar com a significância estatística*** de cada um dos coeficientes. Além disso, dado o período amostral reduzido, ***os intervalos de confiança tendem a ser grandes, o que impossibilitaria a análise subsequente.***

Pela passagem citada acima, é possível perceber que a análise subsequente do artigo foi com relação aos efeitos econômicos estimados e que este se preocupou com as magnitudes, apenas não construindo o intervalo de confiança por justamente impossibilitar a análise proposta. Deste modo, o artigo recebeu "sim". Vale adiantar que o artigo também recebeu "sim" nas questões B3, B4 e B5.

Resultados – Questões de Magnitude

Os resultados das questões do "***Grupo A"*** encontram-se elencados na tabela de número 07. Apenas 61,2% dos artigos apresentaram suas estatísticas descritivas, número bastante similar ao encontrado para a *American Economic Review* nos anos 90. Já a não apresentação de coeficientes em formas economicamente interpretáveis mostrou-se ligeiramente menos grave (77,6%), com resultados também em ordem de grandeza similar aos obtidos na *AER* dos anos 80 e 90. Durante a leitura dos textos, verificou-se que, mais do que uma questão de estilo, a apresentação criteriosa de algumas estatísticas descritivas, bem como dos coeficientes em formas economicamente interpretáveis ***minimizam o impacto de outros problemas***, ao facilitar o julgamento dos resultados pelo leitor. Nota-se que, na maioria das vezes, ***estes são pontos de fácil solução*** –

acredita-se que pequenas medidas, como uma maior atenção dos revisores e editores, possam facilmente elevar ambos os critérios a um percentual de "sim" maior do que 90%. Deste modo, apesar de os percentuais terem sido altos em relação às demais questões, é difícil considerar o resultado encontrado como um bom sinal.

Tabela 07 – Questões de magnitude na RBE 2008-2011, AER 90's e 80's

O artigo…	***Percentual "sim"****	***AER (90's)***	***AER (80's)***
A1 – (Q2) Apresenta estatísticas descritivas?	61,2	66,3	32,4
A2 – (Q3) Apresenta coeficientes em formas economicamente interpretáveis?	77,6	86,9	66,5
A3 – (Q5) Interpreta cuidadosamente os coeficientes?	58,2	81,0	44,5
A4 – (Q12) Discute o tamanho dos coeficientes?	41,8	78,1	80,2
A5 – (Q13) Discute a "conversa científica" na qual um parâmetro seria considerado grande ou pequeno?	43,3	53,5	28,0
A6 – (Q10) Evita a "econometria do asterisco"?	73,1	31,0	74,7
A7 – (Q11) Evita a "econometria do sinal"?	47,8	21,9	46,7
A8 – (incluída) Constrói intervalos de confiança para interpretar a significância econômica?	3,0	n.a.	n.a.

Fonte*: todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011. Todos os artigos completos publicados na American Economic Review nas décadas de 1980 e 1990, conforme análise de Ziliak e McCloskey (2008a). *percentual de artigos que receberam "sim" dentre os artigos em que a questão se aplica.*

As questões seguintes foram bem mais problemáticas. Mais da metade dos artigos não discutiu o tamanho dos coeficientes, não apresentou a conversa científica em torno da qual se comparar as magnitudes estimadas e a importância econômica do modelo utilizado, ou, ainda, não evitou a "econometria do sinal". Ademais, pouco menos da metade dos artigos não interpretou com cuidado seus coeficientes. Estes resultados foram similares aos obtidos na *AER*, com exceção da questão A4, que foi substancialmente menor tanto para os anos 80 quanto para os anos 90. Isto nos permite concluir que, de fato, grande parte dos trabalhos aplicados tem ignorado as magnitudes estimadas e sua relevância econômica para o problema que se propõe discutir[81]. Tendo em vista os resultados que serão apresentados na seção seguinte, uma possível causa para tanto é a crença de que a constatação ou não da "significância estatística" da variável bastaria para a investigação científica.

O resultado que chamou mais a atenção foi o de que apenas um artigo construiu intervalo de confiança para a discussão das magnitudes estimadas e sua correspondente

[81] É importante lembrar aqui que, como explicado anteriormente, caso estivesse claro no artigo que a magnitude do efeito era irrelevante para a questão, este também teria recebido "sim".

incerteza amostral, conforme as suspeitas de DeLong e Lang (1992) e Ziliak e McCloskey (2008a). Caso os testes de significância estivessem sendo realmente utilizados para se precaver da incerteza em relação à variação amostral, seria coerente preocupar-se com esta incerteza mesmo quando o coeficiente é "estatisticamente significante a 5%". Entretanto, isto não ocorre. Uma vez que o zero é excluído do intervalo de confiança (*ad-hoc* de 99%, 95% ou 90%), então a estimativa passa a ser pontual, sem se preocupar com as diferentes magnitudes sugeridas pelo intervalo. Além da ênfase na "significância estatística" como critério de "importância científica", outra possível justificativa para este resultado é que, em geral, os intervalos de confiança são grandes, o que poderia revelar a fragilidade de muitas conclusões com relação à significância econômica das variáveis.

Tabela 08 – Resumo dos resultados da avaliação: questões de magnitude (RBE 2008-2011)

Percentual de "sim"	***Média***	***Mediana***	***DP***	***Mínimo***	***Máximo***
Percentual	50,7	50,0	27,9	0,0	100

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011.*

A tabela de número 08 apresenta as principais estatísticas descritivas dos percentuais de "sim" obtidos nas questões de magnitude. Percebe-se que os percentuais médio e mediano estão bastante próximos, em 50%, e que o desvio padrão é relativamente alto, indicando grande dispersão nas classificações – houve artigos com percentual zero bem como artigos com percentual máximo nestes quesitos.

Tabela 09 – Classificação dos artigos: Questões de Magnitude (RBE 2008-2011)

Percentual de "sim"	***(0 – 20)***	***(20 – 40)***	***(40 – 60)***	***(60 – 80)***	***(80 – 100)***
Artigos	11	14	12	21	9
Percentual	16,4	20,9	17,9	31,4	13,4
Acumulado	16,4	37,3	55,2	86,6	100

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011.*

Deste modo, observando-se a tabela de número 09, apesar de mais da metade dos artigos encontrar-se com percentual de "sim" abaixo de 60%, nota-se que a classificação com maior percentual de artigos foi a de valores entre 60% e 80% (com 31,4% da amostra). Tendo em vista que, tirando as duas principais questões (A4 e A5), todas as

demais não se mostraram difíceis de ser solucionadas, acredita-se que seja possível tornar esta distribuição mais concentrada nas duas últimas classes.

Vejamos agora as questões do ***Grupo B.***

3.2.3. GRUPO B – O culto da significância estatística

B1 – (substitui Q1) O artigo discute o nível de significância adotado tendo em vista as circunstâncias em que os testes serão aplicados?

O nível de significância a ser estabelecido para a inferência estatística deve ser ponderado pelo pesquisador. Isso faz parte do problema a ser discutido, ou seja, é uma das variáveis a serem pensadas diante das circunstâncias, e não um padrão exógeno a ser arbitrariamente alcançado. Em uma amostra enorme, a variação amostral pode ser irrisória e, deste modo, a significância estatística a 5% seria um exercício trivial. Já em uma amostra bastante pequena, a variação amostral pode ser relevante a ponto de um nível de 5% acabar por se tornar demasiadamente pequeno e arbitrário para o julgamento científico do valor encontrado. Em muitos casos, como vimos, a hipótese nula de efeito zero pode sequer ser plausível, e uma analise mais produtiva seria estabelecer quais discrepâncias podem ser inferidas ou não dos dados. Ou ainda, como expõem Kramer (2011, p.459) e Cox e Mayo (2010, p.267), quando vários testes são realizados e a hipótese nula escolhida é aquela que apresenta uma estatística de teste grande, se o pesquisador deseja manter a probabilidade geral de um *erro tipo I* em um nível predeterminado α, tem então de levar em conta esta multiplicidade[82].

Virtualmente ninguém questionou o tamanho do teste. Dos 67 artigos analisados, apenas 02 "discutiram" o nível de significância adotado. Um deles foi o artigo de número 09, que antes de apresentar os resultados econométricos, fez a seguinte ressalva (p. 267): "[...] como a amostra não é muito grande, especialmente quando se considera a taxa de juros, nas análises subsequentes é utilizado o nível de significância de 10% em todos os testes de hipótese". Porém, note que a discussão, como esperado, não levou em conta qualquer elemento de teoria da decisão, como uma função perda ou outro tipo de

[82] Por exemplo, ao se testarem 20 hipóteses independentes, a probabilidade de se achar pelo menos uma significante ($p < 5\%$), mesmo quando todas são falsas, é de aproximadamente 64% (COX, MAYO, 2010, p.269).

ponderação quantitativa entre os erros de *tipo I e tipo II* – independentemente, disto o artigo recebeu "sim".

Discussão um pouco mais elaborada surgiu no artigo de número 54 (p.132), que explicou a razão de ter utilizado um nível de significância de 20%, tomando como referência outro trabalho. Além disso, os autores experimentaram valores ao redor do nível escolhido para verificar se os resultados eram alterados:

> [...] com base em testes de simulação com DAGs aleatórios, SGS sugere definir o nível de significância em 20% para tamanho de amostra menor do que 100; em 10% para tamanho de amostra entre 100 e 300, e em 0,5% (ou menor) para amostras de maiores dimensões. Seguimos sua sugestão e definimos o nível de significância em 20%. Testamos diferentes níveis de significância na vizinhança do nível escolhido (20%) e notamos que as relações de causalidade contemporâneas atribuídas pelo TETRAD não mudaram.

Em suma, ***somente 3%*** dos artigos discutiram o nível de significância. É interessante ressaltar que a quantidade de observações de cada trabalho era ***bastante heterogênea***, variando de 27 até 502.627 – entretanto, todos os artigos aplicaram invariavelmente o mesmo limiar de 1%, 5% ou 10% ao nível de significância estatística[83]. Como artigos com temáticas tão díspares e amostras tão diferentes poderiam todos assentar-se em um mesmo critério automático de "validade" empírica?

B2 – (Q8) O artigo menciona o poder do teste?

Uma vez que a maior parte dos testes de hipótese é realizada sob o paradigma clássico, entender o poder do teste frente a hipóteses alternativas relevantes do ponto de vista econômico seria importante para se ponderar os riscos da inferência que está sendo realizada. Não obstante, como visto, a tradição nas análises econômicas – bem como em outras ciências – é ignorar por completo a função poder. Desta forma, esta questão não foi rigorosa, e requereu apenas que o autor demonstrasse algum conhecimento ou preocupação com o poder dos métodos que está utilizando.

Em nossa amostra piloto, apenas 11 artigos mencionaram o poder do teste em algum momento. É importante ressaltar que, como previsto, os artigos ***não*** calcularam o poder do teste e, dessa forma, não utilizaram esta informação para realizar a inferência estatística em um paradigma de Neyman-Pearson, por exemplo. Os autores apenas

[83] E a escolha entre 1%, 5% ou 10% dava-se, na maior parte das vezes, de maneira *ad-hoc*, segundo a conveniência de se rejeitar ou não a hipótese.

demonstraram preocupação com o fato de os testes utilizados apresentarem "baixo poder" ou "alta sensibilidade". Podemos citar o artigo de número 09 (p.262), que demonstra preocupação ao afirmar que "[...] a limitada disponibilidade de dados pode ter impactado o poder dos testes utilizados".

Os artigos de número 05, 11, 20, 26, 35, 37 e 52 preocuparam-se com o poder de alguns testes de raiz unitária, como o teste ADF. Já o artigo de número 40 apontou para o fato de o método *momentum threshold autoregressive cointegration* (M-TAR) ser, em geral, mais poderoso do que o modelo TAR, para se detectar bolhas racionais. Algumas discussões são bem marginais, como a do artigo de número 53 (p.109) que, no mesmo contexto de modelos TAR e M-TAR do artigo de número 40, apenas menciona a possibilidade de uma perda de poder devido à adição de alguns coeficientes. Ou o artigo de número 46, cuja menção foi em nota de rodapé.

B2.1 – (Q9) Caso o artigo mencione o poder do teste, faz algo com relação a isso?

Esta questão apenas se aplicou aos artigos que obtiveram "sim" na questão anterior. E, da mesma forma, a análise não foi rigorosa. Bastava o autor demonstrar alguma atitude corretiva frente à constatação de baixo poder, como reduzir explicitamente os níveis de significância exigidos, ou utilizar testes que, na média, seriam "mais ou menos poderosos", para receber o "sim". A maioria dos artigos buscou alguma "solução" nestes moldes. Os artigos de número 05, 11, 26, 35 e 37 buscaram testes de raiz unitária alternativos aos rotineiramente aplicados. O artigo de número 09, por sua vez, utilizou nível de significância de 10%.

B3 – (Q7) A significância estatística, quando primeiramente utilizada, é apenas um entre outros critérios de importância?

Em geral, segundo Ziliak e McCloskey, (1996, 2004a, 2008a) o primeiro uso da significância estatística ocorre no ponto principal do artigo, em que o autor acredita fazer seu argumento empiricamente mais importante. Todavia, algo estaria errado com a argumentação caso a significância estatística fosse o critério mais importante avaliado, pois, como visto, ela não é nem necessária tampouco suficiente para a significância substantiva.

O artigo de número 02 traz uma passagem interessante. Como trabalham com dados em painel, os autores estimam o modelo por efeitos fixos (EF) e por efeitos aleatórios (EA). Ao comparar as estimativas dos dois modelos, afirmam os autores que,

> [...] no caso de se concluir que tais estimativas não são estatisticamente iguais, deve-se empregar o método do EF, por este ser consistente independentemente da hipótese adicional do método do EA. Se forem iguais, o método do EA é mais eficiente. Recorre-se, assim, ao teste de Hausman somente se os resultados de EF e EA ***divergem em relação à relevância estatística*** das variáveis de maior interesse (p.30).

Isto é, a investigação sobre qual o modelo mais apropriado apenas se daria caso a variável revelasse significância estatística em um, mas não no outro. Por exemplo, o coeficiente que mede o efeito do fluxo de caixa na demanda por investimento de empresas médias é estimado em 0,093 em efeitos fixos e em 0,142 em efeitos aleatórios – esta diferença de cerca de 50% na magnitude é relevante? Como a significância estatística foi o único critério de importância utilizado, não se sabe.

O artigo de número 14 também classificou como seu principal resultado a "significância estatística" do coeficiente estimado. Os autores verificaram a possibilidade de a paridade descoberta da taxa de juros, no Brasil, estar de "cabeça para baixo". Isto seria evidenciado pela estimativa de um coeficiente positivo. Assim, como explicitado pelos próprios autores: "[...] nosso principal resultado é [que] [...] nosso α estimado é cerca de 1 e podemos rejeitar ao nível de 5% confiança [sic] que α seja negativo". Perceba que ao nível de significância de 1%, por exemplo, não se rejeitaria que α seja negativo. Por que o nível de 5%? O valor estimado de 1 é relevante economicamente para se considerar que a paridade descoberta esteja de fato de "cabeça para baixo"? Como a significância estatística foi o único critério empírico escolhido, também não se sabe.

Outro exemplo é o artigo de número 13, que buscou verificar se o mercado brasileiro é "míope" com relação à política fiscal. Os autores, por meio de um modelo VAR, analisam se o mercado, ao formar suas expectativas sobre a dívida líquida do setor público ou sobre o superávit primário, levaria em conta outros indicadores que dariam uma medida mais acurada da situação fiscal do governo. Antes de iniciar a análise, contudo, os autores expõem qual o critério utilizado para identificar a importância do choque:

> [...] na análise das funções de resposta a impulso, ***deve-se observar que somente são consideradas as respostas a impulsos significativas***, ou seja, aquelas que se encontram dentro do intervalo de confiança de dois desvios

> padrão, completamente acima ou completamente abaixo do número zero no eixo das abcissas. Em outras palavras, nos apêndices que mostram as funções de resposta a impulso, quando as linhas pontilhadas que denotam o intervalo de confiança abrange o eixo das abcissas, diz-se que não existe resposta ao respectivo impulso, ou seja, o efeito do choque da variável sobre a outra é desprezível (nulo) (p. 453, grifo nosso).

Ou seja, fica evidente na passagem citada a confusão entre (in)significância estatística e (in)significância econômica. Se o autor encontrasse um efeito desprezível do ponto de vista econômico, mas com erro padrão amostral pequeno o suficiente para que o intervalo de confiança não ultrapassasse o eixo das abcissas, diria que o resultado é "significante". Já se o autor encontrasse um efeito relevante do ponto de vista econômico, mas cujo intervalo de confiança fosse um pouco maior e abrangesse o zero, diria simplesmente que o efeito inexiste. Ademais, por fundamentar-se apenas na significância estatística durante todo o texto, podemos adiantar que o trabalho também recebeu "não" nas questões B4 e B5.

O artigo de número 16 (p.85) buscou medir a eficácia dos instrumentos de intervenção do Banco Central do Brasil sobre a volatilidade condicional da taxa de câmbio nominal. Da mesma maneira que no trabalho anterior, o único critério utilizado para verificar a importância das variáveis era sua significância estatística, a 1%, 5%, ou 10% conforme pode ser visto a seguinte passagem:

> [...] no Período B (julho/1999 a maio/2002), o único instrumento de intervenção não significativo ao nível de 5% foi a intervenção via Swaps cambiais. A taxa de juros Selic, as intervenções via títulos cambiais e intervenções à vista foram significativas ao nível de 5%. Todos os instrumentos de intervenção adicionaram volatilidade à taxa de câmbio nominal visto que seus coeficientes são positivos e significativos. No Período D (março/2003 a dezembro/2003), nenhum instrumento de intervenção afetou de forma significativa a volatilidade cambial. Rejeitamos a hipótese de assimetria dos choques para este período, mas não rejeitamos a existência de resistência auto-regressiva condicional e a persistência dos choques, ambos são significativos ao nível de 1% e 10% respectivamente.

Vejamos agora exemplos de trabalhos que ilustram como a significância estatística pode ser apenas um entre outros fatores de importância. O artigo de número 03 (p.52-53) estima um modelo linear e um modelo não linear para o *pass-through* da taxa de câmbio e os compara utilizando o grau de ajuste e testes de diagnóstico tais como teste de especificação RESET. Como o modelo não linear apresenta maior grau de ajuste (0,71 contra 0,57) e passa em todos os testes, este é considerado mais adequado – ou seja,

aqui de fato entrou a significância estatística como um dos critérios de avaliação[84]. Todavia, o autor não interrompe sua análise neste ponto. O artigo ressalta que a soma dos coeficientes de não linearidade é alta sugerindo uma forte influência da inflação no *pass-through* cambial. Assim, a diferença nos modelos também é ***economicamente importante***, e resume o autor:

> [...] voltemos nossa atenção para o *threshold*. Ele mostra que o *pass-through* da taxa de câmbio entrará no regime de inflação alta quando a taxa de inflação trimestral é de cerca de 2,56%, ou cerca de 10,6% a uma taxa anualizada. Quando a inflação está acima do limiar e a função de transição (G) é igual a 1 [...] nessas circunstâncias, uma depreciação de, digamos, 10% levaria a 2,7% de inflação no próximo trimestre, e quase 4% no longo prazo. Por outro lado, quando a inflação é bem abaixo do limiar, e a função de transição é igual a 0, a longo prazo o *pass-through* da taxa de câmbio será tão baixo quanto 0,08 e, portanto, um depreciação de 10% levaria a apenas 0,8% de inflação no longo prazo (p. 53, grifo nosso).

Podemos citar também o artigo de número 07 (p. 237, grifo nosso), que não presume que a ausência de significância estatística é igual a efeito zero. Os autores afirmam que os resultados encontrados

> [...] indicam que o efeito de curto prazo das variações cambiais sobre a inflação é assimétrico. No caso de grandes depreciações cambiais, o pass-through estimado para o trimestre seguinte é de cerca de 11%, enquanto que apreciações ou depreciações pequenas não têm um efeito estatisticamente significativo O teste de Wald rejeita a hipótese nula de que ambos os coeficientes são iguais (ao nível de significância de 10%). Portanto, a passagem é maior quando depreciações trimestrais são iguais ou maiores do que 2,1%. Embora os resultados sobre o efeito de uma apreciação em relação ao trimestre anterior sobre a inflação corrente não tenham sido estatisticamente significativos***, não devemos inferir que apreciações não são transmitidas para os preços***. Esta transmissão pode ter lugar com mais desfasamentos do que no caso de depreciação

Além disso, em outro momento, os autores encontram que os efeitos do câmbio sobre a inflação de fato variam segundo a volatilidade do período e não ignoram o resultado encontrado a despeito de não ter se obtido a significância estatística. Eles apenas sugerem maior cautela ao leitor. Nas palavras dos autores (p. 239, grifo nosso),

> [...] ***em termos de magnitude***, as estimativas pontuais ***indicam um maior pass-through*** em períodos de baixa volatilidade do que em momentos de volatilidade elevada (80% e 7%, respectivamente). No entanto, a estimativa do pass-through não é estatisticamente significativa no regime de baixa volatilidade, embora seja significativa no outro regime e ***os valores dos parâmetros são semelhantes aos relatados na literatura para os períodos de taxas de câmbio administrada e flutuante***. A divisão da amostra resultante atribuiu a maior parte das observações do câmbio administrado para o regime de baixa volatilidade. As observações correspondentes a valores abaixo do

[84] Muito embora tenha sido de um modo considerado positivo por muitos autores, como Spanos (2008) e Kramer (2011), pois um modelo estatisticamente inadequado não permitiria inferências seguras.

> limiar compreendem o período 1995:4-1998:2 No entanto, de acordo com o teste de Wald, não podemos rejeitar a hipótese nula de que ambos os coeficientes sejam iguais, e os resultados desta especificação da curva de Phillips são menos robustos do que os dos dois modelos anteriores. Portanto, esses resultados devem ser considerados com mais cuidado.

Estes exemplos ilustram de maneira simples a diferença entre fundamentar a análise apenas na significância estatística, como realizado nos quatro primeiros textos, ou considerá-la apenas um aspecto dentre outros para se julgar a relevância econômica dos resultados, como feito nos dois últimos. Veremos mais alguns casos nas duas próximas questões.

B4 – (Q15) Depois do ponto principal do artigo, este evita usar significância estatística como o critério de importância científica?

Muitas vezes o artigo apresenta a significância estatística como ponto central por exigência da tradição da literatura, da pressão dos editores ou a pedido dos revisores. Assim, poderia ser que o autor desse destaque à significância estatística no ponto principal do texto e, mais a frente, argumentasse com outros critérios sobre a importância científica de seu trabalho. Iremos abordar dois exemplos ilustrativos que diferenciam a presente questão das questões B3 e B5. Abordaremos um artigo que recebeu "não" nas questões B3, B4 e B5; e outro que recebeu "sim" na questão B3, "sim" na questão B4, mas "não" na questão B5.

O primeiro exemplo trata-se do artigo de número 04 (p. 64). O trabalho teve como ápice do estudo a significância estatística da variável de qualidade institucional, recebendo "não" na questão anterior. Ainda, no decorrer do texto, nenhum outro critério foi explorado, recebendo, portanto, "não" nesta questão. Por fim, chegou-se apenas à conclusão meramente qualitativa de que a qualidade das instituições tem impacto positivo no PIB. Nas palavras dos autores:

> [...] o estudo mostrou que a inserção da qualidade institucional no modelo de Solow contribui para aumentar o poder de explicação do modelo. A qualidade das instituições de um país apresenta um impacto positivo no PIB "per capita" e pode ser considerada de grande importância para a acumulação da riqueza e de um crescimento sustentável.

O ideal nesta questão seria exigir critérios mais substantivos complementares à significância, principalmente aspectos quantitativos ou outras abordagens estatísticas. Além disso, seria fundamental que o autor evitasse remeter à significância como critério de importância. Entretanto, sendo muito rigoroso na avaliação, poucos artigos

receberiam "sim" em nossa amostra. Dessa forma, análises que apresentaram algum outro elemento como importante, dando o devido destaque, foram consideradas.

Vejamos o caso do artigo de número 01. Quando do primeiro uso dos testes de significância estatística (p. 13), o trabalho, em sua argumentação, considera algo a mais do que o teste aplicado. Um exemplo seria o destaque dado ao coeficiente de emenda da reeleição:

> [...] os resultados relativos à Emenda da Reeleição indicam que a nova lei introduziu um estímulo adicional para as manipulações eleitoreiras. No lado das receitas, as estimativas apontam que ela implicou em um aumento na amplitude dos ciclos na receita total (27%), na receita corrente (14%) e nas transferências correntes (22%). No lado das despesas, os resultados indicam que governantes candidatos à reeleição elevam a despesa total (29%), a despesa corrente (21%) e as despesas de custeio (47%) acima dos demais governantes (não-candidatos à reeleição). Estes resultados parecem fazer bastante sentido, principalmente, quando se consideram as características das instituições políticas e eleitorais brasileiras. A falta de um maior grau de transparência política e a inexperiência do eleitorado, entre outros aspectos, criam elevados incentivos para que os governantes se distanciem das políticas socialmente ótimas.

Assim, o artigo recebeu "sim" na questão B3. No decorrer do artigo os autores dão primazia à significância estatística e aos sinais das variáveis, realizando uma comparação de magnitudes apenas de modo qualitativo. Isto é, as magnitudes em si não importaram na análise, apenas se eram maiores ou menores, e somente quando estatisticamente significantes. Por exemplo:

> [...] a Tabela 6 apresenta os resultados para as variáveis da despesa orçamentária. De forma semelhante ao obtido anteriormente, a variável oportunismo mostrou-se significante na maioria das regressões, com o seu coeficiente apresentando uma nítida tendência de queda ao longo do tempo [...] a variável competência, por outro lado, não se mostrou significante em nenhuma das regressões (p.16).

Entretanto, os autores consideraram o fato de os coeficientes reduzirem como algo importante ***e argumentaram em cima disto***, concluindo que "[...] (1) a falta de um controle político adequado tem induzido a produção de ciclos nas despesas orçamentárias; e (2) o nível de controle político tem se elevado ao longo do tempo." (p.16). Assim, a despeito da ausência de uma análise quantitativa, considerou-se ***"sim"*** na presente questão. Perceba, todavia, como ***a significância estatística foi o critério decisivo do ponto de vista empírico do artigo*** – sua ausência impediu qualquer análise posterior. Além disso, as magnitudes, na segunda parte do artigo, forem solenemente ignoradas, bastando a significância estatística para a decisão acerca da importância dos

resultados. Sem diferenciar a significância econômica da significância estatística, recebeu ***"não"***, portanto, na questão seguinte.

B5 – (Q16) O artigo considera mais do que a significância estatística para um argumento decisivo do ponto de vista empírico?

Esta questão diz respeito a qual o argumento decisivo utilizado pelo autor para a retórica de suas descobertas. Por exemplo, alegar que o coeficiente é zero por não ser estatisticamente significante é um destes casos; dessa forma, quando o autor descarta por completo o resultado de sua investigação empírica apenas por não se obter significância estatística – e geralmente a níveis usuais sem qualquer preocupação com o limiar adotado – recebe-se "não". O mesmo ocorre quando o resultado encontrado é considerado importante simplesmente por ser estatisticamente significante. O erro deste tipo de inferência já deve ter ficado claro neste ponto do trabalho. Por todo o exposto nas questões anteriores, é possível constatar diversos casos em que este problema surge. Tentaremos trazer aqui apenas mais alguns exemplos, separados por categorias, de modo a ilustrar as diferentes formas de ocorrência.

Insignificante, portanto negligenciável.

Alguns artigos consideraram a ausência de significância estatística como prova da ausência de efeito, sem levar em consideração o poder (ou a severidade) do teste analisado. O artigo de número 06 estimou o coeficiente para aposentadorias rurais em de -0,0578, com *p-valor* de 0,34. O fato de o coeficiente não ter sido "estatisticamente significante" aos níveis usuais fizeram os autores considerar que "[...] os resultados do modelo permitem ***concluir*** que os benefícios da aposentadoria per capita ***não impactaram*** a pobreza rural no Brasil" (p.161, grifo nosso), afirmando este ter sido o principal resultado do artigo. Note, entretanto, que um intervalo de confiança no coeficiente abarcaria valores tão grandes quanto o da variável PIB, que foi considerada importante pelos autores.

Artigo de número 17 procedeu de maneira similar. Os autores buscaram medir o impacto da redução dos encargos trabalhistas sobre a formalização das empregadas domésticas. Quando da análise do coeficiente de interesse, o trabalho considerou que a ausência de significância correspondia a efeito nulo, como na seguinte passagem (p.104):

> [...] os resultados, que podem ser observados na Tabela 5, evidenciam que a variável de interesse (lei*mensalista) é positiva quanto a formalização, porém, não significativa, o que indicaria que ***a mudança na legislação não surtiu efeito*** sobre a taxa de formalização das trabalhadoras domésticas no período analisado.

E, por fim, o artigo concluiu que

> [...] levando em consideração as evidências aqui mostradas, não se pode classificar o esforço do governo em reduzir os encargos trabalhistas incidentes sobre esta ocupação como instrumento eficaz de aumento da formalização devido aos resultados inconclusivos encontrados; em alguns casos houve aumento na formalização e em outros, não.

Note que ambos tratam de políticas públicas e que, se suas conclusões fossem tomadas ao pé da letra, poderiam levar a decisões de consequências relevantes, como o encerramento do programa de aposentadorias rurais bem como o aumento dos encargos trabalhistas das empregadas domésticas.

Significante, portanto importante.

Similarmente aos casos anteriores, mas na direção contrária, muitos trabalhos consideravam a significância estatística como fato suficiente para a importância científica dos resultados. O artigo de número 19 (p.142, grifo nosso), ao medir as relações entre ciclo econômico e oferta de trabalho no Brasil, encontrou um "efeito desalento" estatisticamente significante e concluiu que:

> [...] o efeito desalento também ***se mostrou relevante***, influenciando as decisões de participação das esposas cujos maridos permaneceram empregados. Para esse grupo de mulheres, uma redução de 10% na média dos rendimentos em relação ao valor de referência está associada a uma probabilidade de participação menor em 0,39 ponto percentual.

Com uma base de dados de 502.627 observações, um efeito desta magnitude é relevante apenas por ser estatisticamente significante aos níveis usuais?

O artigo de número 59 estudou os determinantes do protecionismo brasileiro tendo por base uma teoria de proteção endógena. Uma das variáveis de interesse era a quantidade de trabalhadores do setor com curso superior completo, denominada de INST, e a suposição era a de que setores mais intensivos em mão-de-obra qualificada demandariam maior proteção. Como esta variável foi a que apareceu como significante na maioria das especificações, o autor a considerou a mais importante (p.271, grifo nosso):

> [...] como se nota, a variável INST é a única que mantém o mesmo sinal em todas as regressões. Também é a variável que com maior freqüência aparece como significativa: em seis dos oito modelos. Essa consistência de sinais e significância sugere que, do conjunto de regressores utilizados neste trabalho, ***INST é a variável mais importante*** e deveria ser incluída em trabalhos futuros que pretendam estudar o protecionismo brasileiro a partir de uma abordagem econométrica.

O artigo de número 66 buscou identificar se há um conflito entre quantidade e qualidade nas publicações dos pesquisadores de economia, bolsistas do CNPq, no Brasil. Na maior parte das vezes, o autor apenas comenta se as diferenças entre médias são significantes ou não[85]. Como, por exemplo, na seguinte passagem (p. 475-477):

> [...] as duas últimas linhas do painel A mostram que a publicação média das 10 principais escolas ortodoxas nos periódicos internacionais de maior impacto é estatisticamente maior do que a das demais escolas (p-valor de 0,000) [...] o painel B da Tabelã [sic] 3 mostra que a publicação média dos heterodoxos nos periódicos 101 a 144 é estatisticamente maior do que a dos ortodoxos (p-valor de 0,000), enquanto que a publicação nos periódicos de maior impacto é estatisticamente menor (p-valor de 0,000).

Mais ainda, o argumento principal do artigo é que, ***como a diferença*** entre as médias de publicações entre os pesquisadores do CNPq e os pesquisadores americanos ***é estatisticamente significativa, então existe o viés*** de quantidade em detrimento da qualidade. Vide, por exemplo, o próprio resumo e a introdução (p.467-469, grifo nosso):

> [...] o número médio total das publicações dos pesquisadores no Brasil ***é estatisticamente maior***, sugerindo um sacrifício de qualidade para aumentar o número de publicações. [...] na ortodoxia, os pesquisadores do CNPq publicaram entre 1999 e 2004, em média, 5,2 artigos, enquanto os pesquisadores ortodoxos nos EUA publicaram 4,3 artigos; ***uma diferença estatisticamente significativa com um p-valor de 0,083***. Na heterodoxia, o viés de quantidade parece ser ainda mais severo. Enquanto os pesquisadores heterodoxos no Brasil publicaram, em média, 5,1 artigos, os heterodoxos nos EUA publicaram apenas 1,8 artigo; ***uma diferença estatisticamente significativa a 1%. Os resultados constituem evidência forte de que tanto os ortodoxos como os heterodoxos no Brasil sacrificam qualidade (prestígio do periódico de publicação) em nome de quantidade.***

Ora, tomando o caso dos economistas ortodoxos como exemplo, a diferença de 0,15 artigo por ano, ***na média,*** dos pesquisadores do CNPq (cuja amostra, frise-se, era de 57, contra 1.123 pesquisadores americanos), que foi "estatisticamente significante" ao nível de 10% (mas não ao nível de 5%) realmente ***é uma diferença importante***?

[85] O artigo não menciona qual teste de médias foi utilizado.

Uma problemática significante.

Muitos artigos, já de início, definiam o próprio problema do trabalho como encontrar ou não a significância estatística de uma variável. O artigo de número 16, que já havíamos citado na questão B3, utilizou a significância estatística a um nível arbitrário como único critério de relevância científica e isso pôde ser visto na própria definição do problema pelos autores (p. 84, grifo nosso):

> [...] iremos testar duas hipóteses: a primeira hipótese é se as intervenções impactam a volatilidade cambial. Caso positivo, os coeficientes [...] ***precisam*** ser significantemente diferente de zero. A segunda hipótese a ser testada é se as intervenções conseguem diminuir a volatilidade da taxa de câmbio nominal, com intuito de transformar a relação real-dólar mais estável, especialmente em momentos de crise. Para isso os coeficientes [...] ***precisam*** ser significativos e ter sinais negativos.

O artigo de número 30 definiu de maneira similar o seu problema. Para estudar a convergência de renda entre os municípios do estado do Mato Grosso, os autores definiram seu problema da seguinte forma (p.361, grifo nosso): "[...] portanto, para verificar a ***existência*** de convergência, analisa-se o parâmetro β_1^* , para haver convergência de renda per capita, tem-se um $\beta_1^* > 0$, estatisticamente significativo a 5% de significância". Se o autor obtivesse ***p-valor de 6%*** a convergência ***não existiria***?

Significância, necessidade e prova.

O artigo de número 12 (p. 434) trouxe uma situação interessante; o trabalho, em alguns momentos, sequer realizou o exercício de comparar os coeficientes estimados, pois acreditou que ***a falta de significância estatística ao nível usual comprometeria qualquer julgamento***. Na palavra dos autores, "[...] no caso da equação 31.a (modelo 2.a), os valores não são comparáveis, pois não são estatisticamente significativos para a esfera federal." Ora, neste caso fica evidente a primazia da significância estatística como argumento decisivo do ponto de vista empírico, pois sua ausência inviabilizaria até a comparação de estimativas.

A (in)significância estatística chegou a ser considerada ***prova irrefutável*** em alguns momentos, como no caso do artigo de número 04 (p. 63, grifo nosso). Os autores afirmam que "a validade dos instrumentos é ***comprovada*** pela estatística de Sargan (OVERID test) 3,5550", complementando com a afirmação de que "[...] de fato, os instrumentos utilizados não estão correlacionados com a perturbação". Percebe-se que a

argumentação teórica sobre os instrumentos serem ou não exógenos foi sumariamente substituída por um teste estatístico de sobreidentificação.

Houve quem diferenciou.

Finalizemos com dois artigos que ***fizeram*** a devida separação entre a significância estatística e a significância econômica dos resultados. O artigo de número 64, que citamos em outros momentos, buscou testar as hipóteses do modelo de Mincer para a estimativa da taxa de retorno da educação ao Brasil, rejeitando-as. Todavia, especificamente com relação à função *spline*, ***apesar de o autor ter rejeitado*** o pressuposto de paralelismo, considerou-se que esta ***fornece boa aproximação*** para a estimativa da taxa de retorno. Conforme o autor (p.422-423, grifo nosso):

> [...] vale notar que as TIRs não linear (terceira linha) e não paramétrica (última linha) ***diferem pouco***. Por exemplo, quando se incorpora o desenho amostral, este viés chega a no máximo 2.08 p.p. na comparação S4-S0 (9.18% - 7.11%) de 2001, e a 1.08 p.p. (7.07% - 5.98%) quando se compara EF8- EF4 de 2003. Em relação ao Censo, chega a quase 2.4 p.p. de viés comparando os dois maiores níveis escolares (S17+-S15) em 2000 e, para as séries, com exceção de 1970, o viés não chega a 1.5 p.p., em termos absolutos. ***Isso nos leva a crer que, apesar de rejeitarmos o paralelismo, a função spline é uma boa aproximação ao se estimar as TIRs***.

Isto é, o modelo foi rejeitado estatisticamente, mas considerado economicamente razoável. Podemos elencar ainda uma citação do artigo de número 22, que se trata de um ***exemplo claro*** em que houve separação entre a significância estatística e econômica da variável. O artigo buscou verificar se as medidas de núcleo da inflação brasileira são enviesadas, e em certa passagem afirma que:

> [...] não apenas os vieses são estatisticamente significativos, mas seus tamanhos ***são economicamente relevantes***. Enquanto o núcleo de exclusão tem um viés de 1,2 p.p. durante o período de metas de inflação, o viés da média aparada atinge 1,6 p.p. Além disso, ***embora não estatisticamente significativa,*** a média de núcleo suavizada e aparada ***fornece um viés de meio ponto percentual*** durante o período de metas de inflação, ***uma magnitude que é economicamente relevante (ou seja, é suficiente para interferir tanto no planejamento dos agentes quanto na política monetária do banco central).***

B6 – Q(19) O artigo evita usar a palavra "significante" em sentidos ambíguos, como significância estatística e influência econômica?

Este caso não trata da confusão em si, pois não é apenas por utilizar inadvertidamente a palavra significante com sentido ambíguo que o autor irá receber "não" em algumas das três questões discutidas anteriormente. Contudo, a constância com que a ambiguidade

ocorre surpreende bastante; além disso, em muitos casos, o uso ambíguo é, de fato, indício da confusão entre a significância econômica e estatística. Tendo em vista a similaridade das situações em que a ambiguidade ocorre, citemos apenas um trabalho. O artigo de número 09 (p. 261) logo em seu resumo utiliza a palavra em sentido ambíguo "[...] os resultados econométricos sugerem que os juros não são significativos". Em outras passagens a ambiguidade se repete: "Hansen e Singleton (1983) não obtêm estimativas significativas [...] Reis *et alii* (1998) também concluem que a resposta da taxa de crescimento do consumo à taxa de juros é não significativa" (p.263). Nestes casos, há a ambiguidade entre o sentido de "não significativo" como "não importante" ou como "estatisticamente insignificante".

B7 – (incluída) O artigo demonstra preocupação com a especificação ou adequação estatística do modelo?

Como visto, dentro de um paradigma de Neyman-Pearson, a adoção de um nível de significância de 5% para um teste estatístico estabelece que, se a hipótese nula for verdadeira, ainda assim esta seria rejeitada em 5% das vezes. Contudo, este número apenas é (aproximadamente) válido se os pressupostos estatísticos utilizados no teste também forem (aproximadamente) válidos. Caso contrário, o nível nominal de 5% pode ser bastante diferente do nível real (menor ou maior), e o teste de hipótese baseando-se neste nível nominal tornar-se-ia um exercício cego e incoerente. Esta seria uma forma de "erro real", apontada por Ziliak e McCloskey (2008a), cuja preocupação também foi levantada por Kramer (2011), com maior ênfase na especificação dos primeiros momentos, Spanos e McGuirk (2001), Spanos (1993, 2008, 2010), com ênfase em todos os pressupostos adotados[86]. Este último autor e Leamer (2010) enfatizam ainda que a confiança em estimadores de variância assintóticos "robustos" à heterocedasticidade e à autocorrelação não é justificável em grande parte dos casos.

Esta questão, portanto, buscou verificar se os autores ***demonstravam alguma preocupação*** com a especificação ou adequação estatística de seu modelo, de modo a assegurar a confiabilidade das estimativas e dos testes utilizados. É importante deixar claro que, infelizmente, não foi possível verificar a adequação estatística de fato e, deste

[86] Para o modelo clássico de regressão linear a redução probabilística dos pressupostos poderiam ser resumidas à (i) normalidade, (ii) linearidade, (iii) homocedasticidade, (iv) invariância temporal e (v) independência. Spanos critica a correção *ad-hoc* da violação de algum desses pressupostos proposta nos livros-textos. O autor afirma que, em muitos casos, "a correção" torna a inferência ainda menos confiável.

modo, um "sim" para esta questão não significa que as inferências realizadas pelo autor eram seguras, mas tão somente que este se preocupou com o problema. Em muitos casos, é possível, inclusive, que as "correções" adotadas não tenham sido adequadas, mas não entraremos neste mérito devido ao escopo deste trabalho. Ademais, para identificar problemas de especificação com testes estatísticos, o julgamento de "quão grande é grande" um desvio em relação ao pressuposto deveria ser igualmente ponderado, algo que também foi relaxado nesta questão, pois, do contrário, quase ninguém receberia "sim".

Para ilustrar como esta preocupação poderia se manifestar, citemos alguns casos[87]. O artigo de número 30 (p.365), por exemplo, preocupou-se com a heterocedasticidade e com os efeitos espaciais dos erros, buscando correções para os problemas e verificando se estas eram satisfatórias:

> [...] analisando os resultados da estimação e dos testes de especificação do modelo, constataram-se problemas de heterocedasticidade, multicolinearidade e ainda efeitos espaciais nos erros. O próximo passo foi corrigir o modelo estimando-o considerando-se os efeitos espaciais, com o intuito de investigar se os problemas estavam sendo causados por omissão desses efeitos. Porém, os resultados permaneciam com problemas de heterocedasticidade e multicolinearidade. Investigou-se qual variável estaria causando a heterocedasticidade no modelo. Após identificar algumas variáveis que poderiam estar acarretando esse problema, buscou-se corrigir o modelo conforme o método de correção de White. Porém, este método também não permitiu um bom resultado.

Analogamente, o artigo de número 32 (p.419) buscou verificar o quão satisfatórios eram os pressupostos de autocorrelação, heterocedasticidade e normalidade de seus modelos, tentando corrigir quando necessário:

> [...] uma vez estimados os modelos para cada um dos estados, os resíduos de cada modelo foram testados para a presença de autocorrelação e para a presença de heterocedasticidade condicional. Quando se detectou a presença de resíduos autocorrelacionados, foi feita uma tentativa de se obter resíduos melhores através da introdução de uma ou duas defasagens. Adicionalmente, nós testamos para a normalidade dos resíduos. As Tabelas A-2, A-3 e A-4 no Apêndice A apresentam os resultados desta análise. De um modo geral, os resultados são pobres em termos de normalidade, mas são relativamente

[87] Modelos que foram utilizados apenas para ***previsão*** não foram considerados, recebendo a classificação "não se aplica" e, portanto, não entrando no cálculo do percentual. Considerou-se "não" a simples realização de alguns testes já clássicos sem outras preocupações com o comportamento do erro, tais como: ***testes de Hausman*** na análise de modelos de ***efeitos fixos vs efeitos aleatórios***, ***testes de raiz unitária*** em análises de ***séries temporais***, ou ***testes de sobreidentificação*** na análise de regressão por ***variáveis instrumentais***. Entretanto, nestes casos, se o artigo complementasse a análise com algum outro teste, como o teste RESET, de independência, de normalidade, de quebra estrutural, de heterocedasticidade – entre outros – receberia "sim", mesmo se não julgasse a relevância econômica do desvio.

> satisfatórios em termos de autocorrelação e de heterocedasticidade condicional.

Uma preocupação bastante evidente com o impacto de erros de especificação nos exercícios de inferência foi encontrada no artigo de número 22 (p. 213, grifo nosso). Os autores questionaram um modelo – considerado "inocente" – utilizado na literatura para verificar o poder preditivo do núcleo da inflação, que provavelmente estaria mal especificado. Em suas palavras:

> [...] um problema importante aqui é que [a equação] (10) muito provavelmente está ***mal especificada***, ***impedindo qualquer inferência confiável***. Por exemplo, ela não inclui nem mesmo desfasamentos da inflação ou do núcleo da inflação. Por isso, não é de se estranhar que para a maioria dos casos e países Catte e Slok (2005) encontraram um coeficiente insignificante. ***A falta de significância não diz muito***, já que as mudanças do núcleo da inflação poderiam realmente ser úteis na previsão da inflação uma vez que outras variáveis relevantes fossem adicionadas ao modelo. Com efeito, nos poucos casos onde o regressor foi significativo, o seu sinal era teoricamente errado, um sintoma típico do problema de variável omitida.

Preocupação semelhante – e mais extrema – com a confiabilidade das inferências em um modelo mal especificado foi encontrada no artigo de número 39. Em certo momento do trabalho, ao realizaram testes de autocorrelação de Breusch-Godfrey, e testes de heterocedasticidade de Bartlett, Levene e Brown-Forsythe, os autores concluem que (p. 39):

> [...] em virtude do problema de erro de especificação [...], que causa heterocedasticidade e autocorrelação, a variância dos parâmetros capital e trabalho não é mínima, não se podendo fazer nenhuma inferência sobre eles.

Resultados – Culto da Significância Estatística

Os percentuais de "sim" para as questões do "***Grupo B***" encontram-se elencados na tabela de número 10.

Em geral, os números não são animadores. Iniciando com a questão "principal" (B5), cerca de 64% dos artigos na Revista Brasileira de Economia, entre os anos de 2008 a 2011, confundiram significância estatística com significância econômica. Este valor é bastante similar aos 70 e 79% encontrados para o *American Economic Review* nos anos 80 e 90, respectivamente, bem como ao intervalo de 56-85% verificado no *German Economic Review.* Como os Estados Unidos, atualmente, detêm grande parte das publicações e periódicos de maior impacto internacional[88], era de se esperar que os resultados dos demais países fossem de certo modo semelhantes. Observando-se as

[88] Conforme medido pelos índices REPEC, por exemplo.

duas questões auxiliares, percebe-se que, ao se considerar o primeiro uso dos testes (B3), a ênfase na significância estatística torna-se ainda maior, representado 74% dos artigos; e, mesmo levando-se em conta outros argumentos apresentados no decorrer do texto (B4), este número ainda representa mais da metade da amostra analisada, com 52%.

Tabela 10 – Significância estatística na RBE 2008-2011, AER 90's e 80's e GER

O artigo…	***Percentual "sim" RBE****	***AER (90's)***	***AER (80's)***	***GER*** **
B1 – (incluída) Discute o nível de significância utilizado?	3,1	n.a.	n.a.	n.a
B2 – (Q8) Menciona o poder do teste?	16,9	8,0	4,4	n.a.
B2.1 – (Q9) Caso mencione o poder do teste, faz algo em relação a isso?	81,8	44,0	16,7	n.a
B3 – (Q7) Quando no primeiro uso, considera a significância estatística como apenas um entre outros critérios de importância?	23,9	39,6	47,3	n.a.
B4 – (Q15) Após o ponto principal, evita usar a significância estatística como o critério de importância científica?	47,8	27,8	40,7	n.a.
B5 – (Q16) Considera mais do que a significância estatística para um argumento decisivo do ponto de vista empírico?	35,8	20,9	29,7	15,4 – 43,6
B6 – (Q19) Evita utilizar a palavra significante com sentidos ambíguos?	20,9	37,4	41,2	n.a.
B7 – (incluída) O artigo demonstra preocupação com a especificação ou adequação estatística do modelo?	26,2	n.a.	n.a.	23,6

***Fonte**: todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011. Todos os artigos completos publicados na American Economic Review nas décadas de 1980 e 1990, conforme análise de Ziliak e McCloskey (2008a). Todos os artigos publicados na German Economic Review, conforme análise de Kramer (2011). *percentual de artigos que receberam "sim" dentre os artigos em que a questão se aplica.** Kramer classificou dois erros, a falácia da aceitação e a falácia da rejeição. Conforme correspondência com o autor (KRAMER, 2012, comunicação pessoal), entretanto, não é possível somar os erros, pois possivelmente há artigos que cometeram ambos e, à época que realizou a classificação, Kramer não os separou. Deste modo, o limite mais otimista é que apenas 43,6% dos artigos do GER não confundiram significância estatística com econômica e este valor poderia chegar a expressivos 15,4%.*

Além disso, os demais indicadores corroboram o quadro apresentado. Com relação à questão B1, apenas 02 artigos "discutiram" o nível de significância adotado, em concordância com resultados observados por Zellner (1981) ou Keuzenkamp e Magnus (1995). A despeito de ter se observado amostras de tamanhos diferentes, métodos de estimação diferentes (com propriedades assintóticas diferentes), multiplicidade de testes em alguns casos e em outros não, virtualmente todos os artigos utilizaram o mesmo nível de significância[89]. Já com relação à questão B7, observou-se que apenas 26,2% dos artigos demonstraram preocupação com a correta especificação ou com a adequação

[89] Além disso, conforme também verificaram Keuzenkamp e Magnus (1995), houve artigos que confundiram "nível de significância" com "nível de confiança". Também foram vistas tabelas com asteriscos nos coeficientes sem legenda adequada.

estatística do modelo, algo bastante similar ao encontrado por Kramer (2011) na *GER*. Em conjunto estes resultados são bastante preocupantes e reforçam a ideia de que os testes de significância tem servido mais como um ritual do que como genuína preocupação com o erro amostral na inferência estatística.

Os resultados da questão B2, no geral, também foram baixos. Todavia, surpreendentemente, a "preocupação" com o poder do teste foi de 16,9%, bastante acima dos 4,5% encontrados por Zellner (1981), dos 4,4% na *AER* nos anos 80, e também dos 8,0% para o mesmo periódico nos anos 90. O mesmo padrão pode ser visto com relação à questão B2.1, que apresentou 16,7% na *AER* nos anos 80, 44,0% nos anos 90 e, para os anos de 2008-2011, na RBE, chegou a 81,8%. Tal resultado, mais do que uma diferença entre os periódicos, pode ser indício de uma evolução, ***ao longo do tempo***, da conscientização da importância de se observar o poder dos testes utilizados. Mais especificamente, o número geral (16,9%) esconde um ***resultado que pode ser relativamente animador***. Nota-se que 91%[90] dos artigos que se preocuparam com o poder dos testes utilizaram análise de ***séries temporais***, e o resultado da questão B2 para este subgrupo seria de 32,1%, melhora ainda mais expressiva. Aliás, 64% das preocupações concerniram a testes de raiz unitária, com os artigos, em geral, referindo-se à "ampla literatura" acerca do poder dos testes utilizados (como *Augmented Dickey-Fuller* ou *Phillips-Perron*). Interpreta-se isso como um indício de que ***existe demanda para a análise de poder*** e que, muito provavelmente, a área de séries temporais, em especial de testes de raiz unitária, tem tido maior ***avanço e divulgação*** de métodos para supri-la. Se realmente for este o caso, frentes de pesquisa que desenvolvam métodos de fácil aplicação para análise de poder (ou severidade), ou até simples *surveys* que recuperem trabalhos já feitos na área e os tornem mais acessíveis aos pesquisadores podem ter impacto significativo na melhoria deste indicador.

A tabela de número 11 apresenta as principais estatísticas descritivas dos percentuais de "sim" obtidos nas questões do culto da significância estatística. Percebe-se forte assimetria à direita, com tanto a média e a mediana bastante baixas, bem como desvio padrão relativamente mais alto do que nas questões de magnitude, com coeficiente de variação próximo a 100% – da mesma maneira que nas questões de magnitude, houve artigos com percentual zero bem como artigos com percentual máximo obtido.

[90] Na verdade um desses artigos teve como análise predominante dados em painel, mas a preocupação com o poder ocorreu no contexto da análise de série temporal.

Tabela 11 – Resumo dos resultados da avaliação: culto da significância
(RBE 2008-2011)

Percentual de "sim"	*Média*	*Mediana*	*DP*	*Mínimo*	*Máximo*
Percentual	26,7	14,3	24,7	0,0	100

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011.*

Em conformidade com a tabela anterior, verifica-se na tabela de número 12 que mais da metade dos artigos encontra-se na primeira classificação, com percentual de "sim" abaixo de 20%, chegando, por fim, a apenas uma observação na classificação de 80 a 100%.

Tabela 12 – Classificação dos artigos: Culto da Significância
(RBE 2008-2011)

Percentual de "sim"	*(0 – 20)*	*(20 – 40)*	*(40 – 60)*	*(60 – 80)*	*(80 – 100)*
Artigos	35	11	13	7	1
Percentual	52,2	16,4	19,4	10,5	1,5
Acumulado	52,2	68,6	88,0	98,5	100

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011.*

3.2.4. Viés de publicação e o poder dos testes

Os artigos também foram classificados conforme tenham rejeitado ou aceitado a hipótese nula do trabalho. Como a maior parte das pesquisas utilizou regressões múltiplas, para classificá-las adotou-se o seguinte método: sempre que houvesse uma hipótese dominante claramente especificada, esta foi considerada como a hipótese nula. Caso esta não estivesse claramente especificada, mas fosse possível identificá-la pelo contexto, também se considerou esta hipótese como H_0, seguindo DeLong e Lang (1992). Quando a arbitrariedade de escolher a hipótese principal fosse demasiadamente grande, o estudo foi classificado como rejeição da hipótese nula caso metade ou mais da metade dos resultados tenham sido estatisticamente significantes.

Tendo em vista o baixo número de artigos classificados como H_0, não se mostrou adequado estimar a proporção de hipóteses nulas não rejeitadas que são verdadeiras, como feito por DeLong e Lang (1992). Ao invés disso, tomou-se o conjunto de hipóteses nulas não rejeitadas e verificou-se a existência de hipóteses economicamente

relevantes na região aproximada de baixo poder, isto é $(1-\beta) < 50\%$, seguindo o método proposto por Andrews (1989). Vale frisar que não se quer dizer que realizar testes com baixo poder seja algo ruim *per se* – afinal, nem sempre amostras grandes ou com pouco ruído estão disponíveis. A ideia aqui seria apenas ter alguma medida, mesmo que rude, do poder dos testes que falharam em rejeitar a hipótese nula nos artigos empíricos aplicados[91].

Conforme pode ser visto na tabela de número 13, apenas 15,6% artigos da amostra não rejeitaram H_0. Este número é substancialmente menor do que os 28% encontrado, para a economia, por DeLong e Lang (1992). Mas ainda bastante superior aos verificados em outras ciências como 3% na psicologia (STERLING, 1959) ou 7,8% em Marketing (HUBBARD, ARMSTRONG, 1992). Nota-se, ainda, que 80% das hipóteses nulas não rejeitadas tinham poder menor do que 50% para alternativas econômicas relevantes. E todas elas tratavam de problemas em que a crença *a priori* de algum efeito era bastante alta. Isto, de certo modo, corrobora a conclusão de DeLong e Lang (1992, p.1270):

> [...] os autores, portanto, enfrentam um dilema: artigos que não conseguem rejeitar a hipótese nula central serão publicados apenas quando os editores acharem que eles são especialmente interessantes, mas os editores acharão que eles são especialmente interessantes apenas quando a hipótese nula que eles testam realmente é falsa.

Tabela 13 – "Viés" de publicação e poder dos testes

Artigos H_0	***Artigos H_1***	***Artigos H_0 com "baixo" poder****
15,6%	84,4%	80,0%

Fonte: *todos os artigos que utilizaram inferência estatística na Revista Brasileira de Economia, 2008-2011. A classificação sobre a rejeição ou não de H_0 foi feita conforme método utilizado por DeLong e Lang (1992), Hubbard (1992) e por Sterling (1959). Artigos com "baixo poder" referem-se a artigos classificados como H_0, em que havia uma hipótese alternativa relevante em região em que a função poder seria menor ou igual a 50%, segundo método proposto por Andrews (1989). Caso o artigo explicitasse uma hipótese alternativa pontual, esta seria considerada. *As regiões somente foram calculadas para aqueles artigos que forneceram erro-padrão para tanto, o que reduziu a amostra, infelizmente, para apenas 05 artigos.*

Evidentemente que apenas o fato de poucos resultados "nulos" terem sido publicados não é evidência conclusiva sobre o viés de publicação – pode ser simplesmente que, de fato, todas (ou a maior parte) das hipóteses nulas em economia sejam falsas. Entretanto, tendo em vista a cultura em torno da busca por resultados significantes, e a confusão entre significância econômica e significância estatística, o viés de publicação parece ser

[91] Ademais, da mesma forma que α, para $(1-\beta)$ ser (aproximadamente) válido, também os pressupostos dos modelos tem de ser (aproximadamente) válidos, o que nem sempre é o caso.

natural. Somem-se a isso trabalhos como os de Kerr, Tolliver, e Petree (1977), Rowney e Zenisek (1980) ou Atkinson, Furlong, e Wampold (1982) que confirmaram diretamente esta hipótese para as áreas de gerenciamento, sociologia e psicologia, o "viés" se torna ainda mais plausível[92].

Julga-se que estes resultados, juntamente com a discussão anterior de cada uma das questões, com seus respectivos exemplos, expõem com clareza a situação do uso dos testes de significância estatística na literatura econômica e a necessidade de se atuar para reverter este quadro.

[92] Os dois primeiros trabalhos realizaram *surveys* com editores e revisores. Já o último solicitou a editores que analisassem artigos que diferiram apenas com relação à significância estatística reportada.

4) CONSIDERAÇÕES FINAIS

Esta dissertação buscou discutir o uso da significância estatística nos trabalhos econométricos aplicados – mais especificamente, a confusão entre significância estatística e significância econômica. Para tanto, na introdução, trouxemos um exemplo de como esta confusão pode ter consequências práticas, ilustrando a importância real de se fazer a distinção – o erro poderia ter levado uma corte a requerer a significância estatística para discutir a materialidade jurídica de um caso. Viu-se também que o fato se encontra disseminado em várias ciências sociais, inclusive na economia, mas que ainda não havia trabalho específico para a literatura brasileira – o que justificaria a realização deste estudo.

No segundo capítulo, abordamos algumas noções teóricas necessárias para a discussão do tema. Diferenciamos as abordagens de testes de hipóteses de Fisher, Neyman-Pearson e Bayes, e como o "ritual" de inferência atualmente praticado trata-se de um híbrido que acaba por levar a falácias de aceitação, falácias de rejeição e à interpretação errônea do *p-valor*, com a consequente confusão entre significância estatística e significância científica. Detivemo-nos, em seguida, à discussão do *p-valor* como evidência, à distinção entre erro amostral e erro real bem como entre diferença estatística e diferença substantiva. Ao final, buscou-se ilustrar alguns métodos que pudessem amenizar os problemas de inferência mencionados.

O terceiro capítulo tratou da parte empírica. Resgatamos os principais levantamentos tangencialmente ou diretamente relacionados à confusão entre significância estatística e econômica feitos na literatura internacional, como os de Zellner (1981), McCloskey (1985), DeLong e Lang (1992), Keuzenkamp e Magnus (1995), McCloskey e Ziliak (1996), Mayer (2001), Ziliak e McCloskey (2004a, 2008a) e Kramer (2011). Verificou-se o fato, de maneira disseminada, tanto na *American Economic Review* quanto na *German Economic Review*. Desta forma, adaptamos o questionário de McCloskey e Ziliak (1996) e analisamos os artigos publicados na Revista Brasileira de Economia, no período de 2008 a 2011.

Em virtude dos resultados encontrados, difícil não se lembrar de alguns dos "dez mandamentos" da econometria aplicada elencados por Peter Kennedy (2002), tais como[93]:

- ***Tu inspecionarás seus dados***;
- ***Tu estarás ciente dos custos de data-mining;***
    - ***Corolário***: tu não adorarás o R^2;
    - ***Corolário***: tu não caçarás a significância estatística com uma espingarda;
    - ***Corolário***: tu não adorarás o nível de significância de 5%;
- ***Tu não confundirás significância estatística com substantiva***;
    - ***Corolário:*** tu não ignorarás poder;
    - ***Corolário:*** tu não testarás hipóteses precisas;
    - ***Corolário:*** tu procurarás evidência adicional;
- ***Tu confessarás na presença da sensitividade.***

Dentro da analogia de Kennedy, estamos "pecando" bastante: entre outros números, cerca de 40% dos artigos analisados não apresentaram suas estatísticas descritivas, pouco mais da metade interpretou cuidadosamente seus coeficientes e menos da metade discutiu a magnitude de suas estimativas. Apenas 3% dos artigos discutiram o nível de significância adotado e o mesmo número foi observado na construção de intervalos de confiança para se discutir magnitudes; mais de 80% ignoraram o poder dos testes, aproximadamente 64% dos trabalhos tomaram a significância estatística como argumento decisivo do ponto de vista empírico e apenas 26% demonstraram preocupação com a especificação ou adequação estatística do modelo.

Como afirma Abelson (1995, p. xii), estudantes não parecem ter tanta dificuldade com o aspecto matemático e operacional da estatística, mas principalmente em entender o que estão fazendo – em geral, estes se tornam presos a regras e passam a encarar a estatística como um "rito médico ou religioso". Assim, apesar de esta dissertação não ter o intuito de prescrever soluções, parece ser necessário modificar o ensino da econometria nos cursos de pós-graduação, enfatizando-se mais a análise descritiva e exploratória de dados, a meta-análise, a análise de sensibilidade, a conexão entre hipóteses estatísticas e substantivas, bem como os fundamentos metodológicos e filosóficos da estatística; também parece ser produtiva a contraposição de métodos clássicos e bayesianos e suas

[93] Para comentários acerca do texto de Peter Kennedy, ver Magnus (2002) e Hendry (2002).

interfaces com a epistemologia científica e a teoria da decisão. Isto deve ser sobreposto ao ensino de um único e simples algoritmo de inferência, como o "ritual nulo", permitindo ao futuro pesquisador a ponderação dos pontos fortes e fracos de cada abordagem e munindo-o da capacidade de escolher o melhor método segundo o problema, ou até mesmo de definir adequadamente o seu problema. Além disso, também parece ser necessário alterar os critérios de publicação, eliminando os incentivos à busca por resultados "significantes". Se os resultados encontrados nesta dissertação, por um lado, são preocupantes, por outro, em alguns momentos dão indícios de que há demanda para este tipo de abordagem, e que um esforço neste sentido pode ter resultados muito positivos.

Talvez a principal fragilidade deste estudo seja a possibilidade de erros de codificação, tanto por conta da subjetividade envolvida, quanto pela exaustiva tarefa de apenas uma pessoa ler e codificar todos os artigos. Após a finalização deste trabalho, ficou claro o sentido da seguinte passagem de Ellis:

> [...] codificação é difícil, um trabalho de entorpecimento mental. Ela começa divertida, mas muitas vezes termina com o revisor abandonando o projeto por frustração ou fadiga. Muitos daqueles que conseguem terminar o processo de codificação não desejam repetir a experiência (ELLIS, 2010, capítulo 5).

Realmente, o autor desta dissertação não pretende realizar outro levantamento deste tipo tão cedo. Frisa-se, entretanto, que se buscou ao máximo identificar inconsistências e corrigi-las. Além disso, a discussão extensiva dos exemplos teve o intuito de minimizar ambiguidades e de trazer casos concretos que deixem claro como o problema se manifesta. Ademais, ainda que codificações ambíguas ou erradas tenham persistido, basta o leitor fazer um breve exercício contra factual para perceber que seria necessária alteração bastante drástica em quase todas as questões, simultaneamente, para mudar o quadro apresentado. Argumenta-se aqui que a discussão e os exemplos elencados na seção 3.2., em conjunto com os resultados obtidos na literatura internacional apresentados na seção 3.1., e a discussão teórica do capítulo 2, são suficientes para o convencimento de que isto é implausível.

Como sugestão de pesquisa futura, volta-se para o fato de a educação ser uma das prováveis causas das questões levantadas. Ziliak e McCloskey (2009, p. 2308), sobre os Estados Unidos, afirmam que "nos departamentos de economia quase todos os professores de probabilidade, estatística e econometria alegam que significância estatística é a mesma coisa de significância científica". Como vimos, Oakes (1986) no

Reino Unido, Flak e Greenbaum (1995) em Israel e Haller e Krauss (2002) na Alemanha verificaram que o desconhecimento sobre o significado do *p-valor* é generalizado, inclusive nos professores que ensinam estatística.

Desta forma, há espaço para pesquisas que examinem qual o método de ensino e livros-textos adotados nos cursos de graduação e pós-graduação em economia, e que averiguem diretamente qual o grau de compreensão dos mestrandos, doutorandos e pesquisadores sobre o tema, principalmente quanto à interpretação dos testes de significância estatística baseado em *p-valores*, tomando por base os pontos levantados em Oakes (1986), Flak e Greenbaum (1995), Haller e Krauss (2002), Gigerenzer (2004), Gigerenzer, Krauss e Vitouch (2004), Goodman (2008), Gelman e Stern (2006), Mayo e Spanos (2011) e Ziliak e McCloskey (2008a). Uma análise sobre este tema talvez seja capaz de prover um diagnóstico mais preciso de onde e como atuar na produção ou sugestão de novos livros-textos, no desenvolvimento de novas rotinas que venham como *default* nos *softwares* econométricos, na sugestão de padrões para o auxílio da revisão de artigos aplicados entre outras atividades.

5) REFERÊNCIAS BIBLIOGRÁFICAS

ABELSON, R. **Statistics as Principled Argument.** New York: Psychology Press, 1995.

ABELSON, R; MILLER, J. Negative Persuasion via Personal Insult. **Journal of Experimental Social Psychology, 3,** p. 321-333. 1967.

ACEMOGLU, D.; JOHNSON, S.; ROBINSON, J.A. The Colonial Origins of Comparative Development: an empirical investigation. **American Economic Review,** v 91(5), dezembro, 2001.

ACEMOGLU, D.; JOHNSON, S.; ROBINSON, J.A. **Institutions as the Fundamental Cause of Long-Run Growth.** Prepared for the *Handbook of Economic Growth.* 2004.

ANDREWS, D. W. K. Power in Econometric Applications. **Econometrica,** v. 57(5), setembro, p. 1059-1090. 1989.

ARMSTRONG, S.A.; HENSON, R.K. Statistical and practical significance in the IJPTP: a research review from 1993-2003. **International Journal of Play Therapy,** 13(2), p. 9-30. 2004.

ARROW, K. J. **Decision Theory and the Choice of a Level of Significance for the *t*-test.** In: Olkin et alii., Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford: Stanford University Press. p. 70-78. 1960.

ATKINSON, D. R.; FURLONG, M. J; WAMPOLD, B. E. "Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is There a (Statistically) Significant Relationship?" **Journal of Counseling Psychology**, 29, p. 189-194. 1982.

BAKAN, D. The Test of Significance in Psychological Research. **Psychological Bulletin,** 66(6), p. 423-437, 1966.

BERGER, J. O. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? **Statistical Science,** v. 18(1), p. 1-32, 2003.

BERGER, J. O. **Statistical Decision Theory and Bayesian Analysis.** New York: Springer-Verlag, 1985.

BERGER, J. O.; DELAMPADY, M. Testing Precise Hypotheses. **Statistical Science**, v.2(3), p. 317-352, 1987.

BERGER, J. O.; SELLKE, T. Testing a point null hypothesis: The irreconcilability of P values and evidence. **Journal of the American Statistical Association**, v.82(397), p. 112-122, 1987.

BROCK, J. The Power of international business research. **Journal of International Business Studies.** 34(1), p. 90-99. 2003.

CALLAHAN, J.L.; REIO, T.G. Making Subjective judgments in quantitative studies: The importance of using effect sizes and confidence intervals. **Human Resource Development Quarterly,** 17(2), p. 159-173. 2006.

CASELLA, G.; BERGER, R. L. Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. **Journal of the American Statistical Association**, 82, p. 106-111, 1987a.

CASELLA, G.; BERGER, R. L. Testing Precise Hypotheses: Comment. **Statistical Science**, v.2(3), p. 344-347, 1987b.

CASELLA, G.; BERGER, R. L. **Statistical Inference.** EUA: Thomson Learning, 2002.

CASTRO, SOTOS *et alii.* Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. **Educational Research Review**, 2, p. 98–113. 2007.

CASTRO, SOTOS *et alii.* How confident are students in their misconceptions about hypothesis tests? **Journal of Statistics Education**, v. 17, n.2, 2009.

CHRISTENSEN, R. Testing Fisher, Neyman, Pearson, and Bayes. **The American Statistician**, Vol. 59, No. 2, p. 121-126. 2005.

CINELLI, C. L. K**. Eficiência fraca no Brasil durante crises**: evidências de um teste de razão de variâncias. Trabalho apresentado para obtenção de grau de especialista em Controladoria e Finanças. FUCAPE, Vitória, 2010.

CINELLI, C. L. K. Transferências voluntárias e corrupção municipal no Brasil: evidências preliminares do cadastro de contas irregulares do TCU. **Revista Economia e Tecnologia,** Ano 07, V. 27, p. 89-98, 2011.

COBB, G. Book Review. **Journal of the American Statistical Association,** Volume 104, Issue 488, p. 1716-1720, 2009.

COHEN, J. The Earth is Round ($p < 0{,}05$). **American Psychologist,** 49, p. 997-1003. 1994.

COX, D. R. Some problems connected with statistical inference. **Ann. Math. Statist**., 29, p. 357-372. 1958.

COX, D. R. Statistical Significance Tests. **British Journal of Clinical Pharmacology**, 14, 325-331, 1982.

COX, D. R .Testing Precise Hypotheses: Comment. **Statistical Science**, v.2(3), p. 335-336, 1987.

COX, D. R. The role of significant tests (with discussion). **Scandinavian Journal of Statistics, 4**, p. 49–70. 1977.

COX, D. R.; MAYO, D. A Statistical Scientist Meets a Philosopher of Science: A Conversation between Sir David Cox and Deborah Mayo (as recorded, June, 2011). **RMM**, Vol. 2, 103–114, 2011.

COX, D. R; MAYO, D. **Frequentist Statistics as a Theory of Inductive Inference**. In: MAYO, D.; SPANOS, A. (eds) Error and Inference. New York: Cambridge University Press, 2010.

DELMAS, R. C. *et alii.* Assessing Students' conceptual understanding after a first course in statistics. **Statistics Education Research Journal,** 6(2), p. 28-58. 2007.

DELONG, B. J.; LANG, K. Are all Economic Hypotheses False? **Journal of Political Economy**, Vol. 100, No. 6, Centennial Issue, p. 1257-1272, 1992.

DEMING, W. E. **Sample Design in Business Research.** New York: Wiley, 1961.

DEGROOT, M. H. Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio. **Journal of the American Statistical Association**, 68, p. 966-969, 1973.

DEZHBAKHSH, H.; RUBIN, p; SHEPHERD, J. "Does Capital Punishment Have a Deterrent Effect? New Evidence from Postmoratorium Panel Data." **American Law and Economics Review 5**: p. 344-376. 2003.

DONOHUE, J.; WOLFERS, J. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." **Stanford Law Review 58**: p. 791-846. 2005.

DURLAUF, S.; FU, C. NAVARRO, S. Capital Punishment and Deterrence: Understanding Disparate Results. **Human Capital and Economic Opportunity: A Global Working Group Working Paper Series, WP 2012-005**. 2012.

DURLAUF, S.; NAVARRO, S; RIVERS, D. "Understanding Aggregate Crime Regressions." **Journal of Econometrics** 158: 306-317. 2010.

EASTERLY, W.; LEVINE, R. Tropics, germs, and crops: how endowments influence economic development. **Journal of Monetary Economics,** v. 50, p. 3-39, 2003.

ECONOMIST, THE. **Signifying Nothing?**, The Economist, 370(8360) January 31, 2004, 71, 2004.

EDWARDS, W.; LINDMAN, W.; SAVAGE, L. Bayesian Statistical Inference for Psychological Research. **Psychological Review,** 70, p. 193-242, 1963

ELLIS, P. D. **The essential guide to Effect Sizes, Statistical Power, Meta-Analysis, and the interpretation of Research Results.** New York: Cambridge University Press, 2010.

ELIOT, G.; GRANGER, C. W. J. Evaluating Significance: Comments on "Size Matters". **Journal of Socio-Economics,** 33(5): p. 547-550. 2004.

ENGSTED, T. Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak. **Journal of Economic Methodology**, 16:4, p. 393-408, 2009.

FALK, R.; GREENBAUM, C. W. Significance tests die hard. **Theory and Psychology**, v. 5(1), p. 75–98. 1995.

FIDLER, F; CUMMING, G; BURGMAN, M; THOMASON, N. Statistical Reform in Medicine, Psychology and Ecology. **Journal of Socio-Economics 33(5),** p. 615-630. 2004.

FISHER, R. A. **Statistical Methods for Research Workers**. New York: Hafner Publishing Company, 14 ed. 1973a. In: BENNET, J. H. (ed.). Statistical Methods, Experimental Design and Scientific Inference. Oxford: Oxford University Press, 1993.

FISHER, R. A. Statistical Methods and Scientific Induction. **Journal of the Royal Statistical Society,** Series B (Methodological), Vol. 17, nº1, p. 69-78, 1955.

FISHER, R. A. **Statistical Methods and Scientific Inference**. New York: Hafner Publishing Company, 3 ed. 1973b. In: BENNET, J. H. (ed.). Statistical Methods, Experimental Design and Scientific Inference. Oxford: Oxford University Press, 1993.

FISHER, R. A. **The Design of Experiments**. New York: Hafner Publishing Company, 8 ed. 1971. In: BENNET, J. H. (ed.). Statistical Methods, Experimental Design and Scientific Inference. Oxford: Oxford University Press, 1993.

GIGERENZER, G.; KRAUSS, S.; VITOUCH, O. **The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask**. In: Kaplan, D. (Ed.), Handbook on quantitative methods in the social sciences. Thousand Oaks, CA: Sage, p. 389-406. 2004.

GIGERENZER, G. **The Superego, the Ego, and the Id in Statistical Reasoning," in A Handbook for Data Analysis in the Behavioral Sciences**. In: KEREN, G. LEWIS, C.A. (eds), Methodological Issues. Hillsdale, NJ: Erlbaum, 1993.

GIGERENZER, G. Mindless Statistics. **Journal of Socio-Economics,** 33(5): p. 587-606, 2004.

GELMAN, A. ROBERT, C. **"Not only defended but also applied"**: The perceived absurdity of Bayesian inference. 2012. Disponível em: http://arxiv.org/abs/1006.5366v4. Acesso em: 25/05/2012.

GELMAN, A.; STERN, H. The difference between 'significant' and 'not significant' is not itself statistically significant. **The American Statistician,** 60(November): p.328-331. 2006.

GLEASER, E. *et alii.* Do institutions Cause Growth***. NBER Working Paper No. 10568***, 2004.

GOLDBERGER, A. S. The ET Interview: Arthur S. Goldberger. **Econometric Theory, 5,** p. 133-160. 1989.

GOODMAN, S. P-Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. **American Journal of Epidemiology,** 137, 485-496. 1993.

GOODMAN, S. A dirty dozen: twelve p-value misconceptions. **Seminars in Hematology.** V.45. p. 135-140. 2008.

GRAVES, S. On the Neyman-Pearson Theory of Testing. **British Journal for the Philosophy of Science** 29: 1-23. 1978.

GREENE, C. I Am Not, Nor Have I Ever Been a Member of a Data-Mining Discipline. **Journal of Economic Methodology** 7:2, p. 217-230. 2000.

GREENE, W. H. **Econometric Analysis.** New Jersey: Prentice Hall, 2002.

GUNDLACH, E.; CARSTENSEN, K. The Primacy of Institutions Reconsidered: Direct Income Effects of Malaria Prevalence. **The world bank economic review,** vol. 20, no. 3, p. 309–339. 2006.

HALLER, H.; KRAUSS, S. Misinterpretations of significance: A problem students share with their teachers? **Methods of Psychological Research Online.** v.7(1), p. 1–20. 2002.

HANSEN, B. **Econometrics.** 2012. Disponível em: http://www.ssc.wisc.edu/~bhansen/. Acesso em 02/02/2012.

HAYASHI, F. **Econometrics.** New Jersey: Princeton University Press, 2000.

HENDRY, D. Applied Econometrics Without Sinning. **Journal of Economic Surveys, 16,** p. 591-604. 2002.

HODGES, J. L.; LEHMANN, E. L. Testing the Approximate Validity of Statistical Hypotheses. **Journal of the Royal Statistical Society**. Series B (Methodological), Vol. 16, No. 2, p. 261-268, 1954.

HOFFMAN, R. **Estatística para economistas**. São Paulo: Pioneira Thompson Learning, 2001.

HOGG, R. V.; CRAIG, A. T. **Introduction to Mathematical Statistics.** New York: Macmillan Publishing Co, 1978.

HOOVER, K.; SIEGLER, M. V. Sound and Fury: McCloskey and significance testing in economics. **Journal of Economic Methodology,** v. 15(1), março, p.1-37. 2008a.

HOOVER, K.; SIEGLER, M. V. The rhetoric of 'Signifying nothing': a rejoinder to Ziliak and McCloskey. **Journal of Economic Methodology,** v. 15(1), março, p.57-68. 2008b.

HOROWITZ, J. L. Comments on "Size Matters". **Journal of Socio-Economics,** 33(5): p. 551-554. 2004.

HUBBARD, R.; ARMSTRONG, J.S. Why We Don't Really Know What "Statistical Significance" Means: A Major Educational Failure. **Journal of Marketing Education**, Volume 28, Issue 2, p. 114-120. 2006.

HUBBARD, R.; ARMSTRONG, J.S. Are Null Results Becoming an Endangered Species in Marketing?. **Marketing Letters,** 3 127-136, 1992.

HUBBARD, R.; BAYARRI, M. J. **P-values are not error probabilities**. Working Paper Universitat de Valencia. 2003.

HUBBARD, R.; BAYARRI, M. J.; BERK, K.; CARLTON, M. A. Confusion over Measures of Evidence (p's) versus Errors (α's) in Classical Statistical Testing. **The American Statistician**, Vol. 57, No. 3, pp. 171-182. 2003.

JEFFREYS, H. **Theory of probability.** London: Oxford University Press, 1961.

KADANE, J. B. Testing Precise Hypotheses: Comment. **Statistical Science**, v.2(3), p. 347-348, 1987.

KADANE, J. B. **Principles of Uncertainty**. EUA: CRC Press, 2011.

KENNEDY, P. **A Guide to Econometrics.** 6ed. Malden: Blackwell Publishing, 2008.

KENNEDY, P. Sinning in the Basement: What are the rules? The ten commandments of Econometrics. **Journal of Economic Surveys, 16,** p. 569-589. 2002.

KERR, S; TOLLIVER , J; PETREE, D. "Manuscript Characteristics Which Influence Acceptance for Management and Social Science Journals," **Academy of Management Journal,** 20, 132-141, 1977.

KEUZENKAMP, H. A.; MAGNUS, J. R. On tests and significance in econometrics. **Journal of Econometrics** 67, p. 5-24, 1995.

KRAMER, W. The Cult of Statistical Significance: What economists should and should not do to make their data talk. **Schmollers Jahrbuch** 131, p. 455 – 468, 2011.

KRAMER, W. GIGERENZER, G. How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. **Statistical Science**, Vol. 20, No. 3, p. 223–230, 2005.

KRAMER, W; SONNBERGER, H; MAURER, J; HAVLIK, P. Diagnostic checking in practice, **Review of Economics and Statistics** 68, p. 118–123. 1985.

KRUSKAL, W. S. **Tests of Statistical Significance**. In: SILLS, David (ed). International Encyclopedia of the Social Sciences. V.15. MacMillan: New York, 1968.

LEAMER, .E.E. Are the roads red? Comments on "Size Matters". **Journal of Socio-Economics,** 33(5): p. 355-358. 2004.

LEAMER, .E.E. Let's take the con out of econometrics. **The American Economic Review,** v. 73, nº 01, p. 31-43, 1983.

LEAMER, E.E. **Specification Searches:** Ad Hoc Inference with Nonexperimental Data. New York: John Wiley, 1978.

LEAMER, .E.E. Tantalus on the Road to Asymptopia. **Journal of Economic Perspectives,** v.24, nº 02, p. 31-46, 2010.

LEHMANN, E. L.; ROMANO, J. P. **Testing Statistical Hypothesis.** New York: Springer, 2008.

LEVINE, M.; SCHERVISH, M. J. Bayes Factors: What They Are and What They Are Not. **The American Statistician**, Vol. 53, No. 2 p. 119-122, 1999.

LINDSAY, R. M. Incorporating statistical power into the test of significance procedure: a methodological and empirical inquiry. **Behavioral Research in Accounting.** V5. P. 211-236. 1993.

LINK, C. W. An Examination of Student Mistakes in Setting Up Hypothesis Testing Problems. **Proceedings of the Louisiana-Mississippi Section of the Mathematical Association of America**. Louisiana, Spring, 2002.

LOUÇÃ, F. The Widest Cleft in Statistics - How and Why Fisher opposed Neyman and Pearson. **School of Economics and Management, Technical University of Lisbon, WP 02/2008/DE/UECE,** 2008.

MADDALA, G.S. **Introdução à Econometria.** 3ed. Rio de Janeiro: LTC, 2003.

MAGNUS, J. The Missing Tablet: Comment On Peter Kennedy's Ten Commandments **Journal of Economic Surveys, 16,** p. 605-609. 2002.

MAYER, T. A Frequent Misuse of Significance Tests. **CESifo Working Paper No. 549**, 2001.

MAYER, T. The Empirical Significance of Econometric Models. **UCDAVIS Department of Economics Working paper Series, Paper 06-20**, 2006.

MAYO, D. G. **An Error-Statistical Philosophy of Evidence**. In: M. Taper and S. Lele (eds.), The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Consideration, Chicago, IL: University of Chicago Press, p. 79–97. 2004.

MAYO, D. Behavioristic, Evidentialist, And Learning Models Of Statistical Testing. **Philosophy of Science**, 52, p. 493-516. 1985.

MAYO, D. Did Pearson Reject the Neyman–Pearson Philosophy of Statistics? **Synthese,** 90, p. 233–62. 1992.

MAYO, D. **Evidence as Passing Severe Tests**: Highly Probed vs. Highly Proved. In Scientific Evidence, P. Achinstein (ed.), Johns Hopkins University Press, 2005a.

MAYO, D. **Philosophy of Statistics**. In: S. Sarkar and J. Pfeifer (eds.) Philosophy of Science: An Encyclopedia, London: Routledge, p. 802–15. 2005b.

MAYO, D.; SPANOS, A. **Error Statistics**. In: BANDYOPADHYAY, P.S.; FORSTER, M. R. Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics. EUA: Elsevier, 2011.

MAYO, D.; SPANOS, A. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. **British Journal for the Philosophy of Science**, 57 (2): 323-357. 2006.

MAZEN, A.M.; GRAF, L.A., KELLOG, C.E.; HEMMASI, M. Statistical power in contemporary management research. **Academy of Management Journal,** 30(2), p. 369-380. 1987.

MCCLOSKEY, D. N. Other things equal: The bankruptcy of statistical significance. **Eastern Economic Journal,** 18, 3, 1992a.

MCCLOSKEY, D. N. The Art of Forecasting: From Ancient to Modern Times. **Cato Journal,** 12(1), 1992b.

MCCLOSKEY, D.N. Other Things Equal: Cassandra's Open Letter to Her Economist Colleagues. **Eastern Economic Journal**, 25(3), p. 357–363. 1999.

MCCLOSKEY, D.N. **The Secret Sins of Economics**. Chicago: Prickly Paradigm Press, 2002.

MCCLOSKEY, D.N. The Insignificance of Statistical Significance. **Scientific American**, 272(4), p. 32–33, 1995.

MCCLOSKEY, D.N. **Rhetoric within the citadel: statistics.** In WENZEL et alii (eds) Argument and Critical Practice: Proceedings of the Fifth SCA/AFA Conference on Argumentation reprinted in C. A. Willard and G. T. Goodnight, eds., Public Argument and Scientific Understanding p. 485-490, 1993.

MCCLOSKEY, D.N. Other Things Equal: Aunt Deirdre's Letter to a Graduate Student. **Eastern Economic Journal**, 23(2), p. 241–244, 1997a.

MCCLOSKEY, D.N. **The Vices of Economists; The Virtues of The Bourgeoisie**. Amsterdam: University of Amsterdam Press, 1997b.

MCCLOSKEY, D.N. **Two Vices:** proof and significance. Speech for the AEA session at Chicago, 1998.

MCCLOSKEY, D. N. The loss function has been mislaid: the rhetoric of significance tests. **American Economic Review,** 25(2): p. 201-205, 1985.

MCCLOSKEY, D. N. The rhetoric of economics. **Journal of Economic Literature,** 21, p. 481-517, 1983.

MCCLOSKEY, D. N. Why Economic Historians Should Stop Relying on Statistical Tests of Significance and Lead Economists and Historians into the Promised Land. **Newsletter of Cliometrics Society,** v. 2, nº 02, 1986.

MCCLOSKEY, D. N.; ZILIAK, S. T. **Brief of amici curiae statistics experts professors Deirdre N. McCloskey and Stephen T. Ziliak in support of respondents:** Matrixx Initiatives, Inc, et al. v. James Siracusano and NECA-IBEW Pension Fund. The Supreme Court of the United States, 2010.

MCCLOSKEY, D. N.; ZILIAK, S. T . The Standard Error of Regressions. **Journal of Economic Literature**, 34, p. 97-114, 1996.

NEYMAN, J. **First Course in Probability and Statistics**, New York: Holt. 1950.

NEYMAN, J. Note on an Article by Sir Ronald Fisher. **Journal of the Royal Statistical Society.** Series B (Methodological), v. 18, nº 02, p. 288-294, 1956.

NEYMAN, J.; PEARSON, E. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. **Biometrika**, Vol. 20A, No. 1/2, p. 175-240. 1928a.

NEYMAN, J.; PEARSON, E. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. **Biometrika,** Vol. 20A, No. 3/4, p. 263-294. 1928b

NEYMAN, J.; PEARSON, E. On the Problem of the Most Efficient Tests of Statistical Hypotheses. **Philosophical Transactions of the Royal Society of London**. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 231, p. 289-337. 1933.

OAKES, M. **Statistical inference**: A commentary for the social and behavioral sciences. New York: Wiley. 1986.

PATRIOTA, A. **A classical measure of evidence for general null hypotheses**. 2012. Disponível em: http://arxiv.org/abs/1201.0400v1. Acesso em: 05/05/2012.

PEARL, JUDEA. **Causality:** models, reasoning and inference. New York: Cambridge University Press, 2010.

PEARSON, E. Statistical Concepts in the Relation to Reality. **Journal of the Royal Statistical Society.** Series B (Methodological), v. 17, nº 02, p. 204-207, 1955.

PEARSON, E. Some Thoughts on Statistical Inference. **The Annals of Mathematical Statistics**, Vol. 33, No. 2, p. 394-403. 1962

ROBERT, C. **The Cult of Significance.** 2011. Disponível em: http://xianblog.wordpress.com/2011/10/18/the-cult-of-significance/. Acesso em: 04/03/2012.

RODRIK, D.; SUBRAMANIAN, A.; TREBBI, F. Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development. **NBER Working Paper 9305.** 2002.

RODRÍGUEZ, M. I. Estudio Teórico y Experimental sobre Dificuldades en la Comprensión del Contraste de Hipótesis en Estudantes Universitarios. **Acta Latinoamericana de Matemática Educativa**, México, v. 19, s/n, p. 162-168, 2006.

ROWNEY, J. A; ZENISEK, T. J. "Manuscript Characteristics Influencing Reviewers' Decisions," **Canadian Psychology**, 21, p. 17-21. 1980.

SACHS, J. 2003. Institutions Don't Rule: Direct Effects of Geography on Per Capita Income. **NBER Working Paper 9490**. 2003.

SALSBURG, D. **The lady tasting tea: how statistics revolutionized science in the twentieth century**. New York: Freeman and Company, 2001.

SCHERVISH, M. J. P values: What They Are and What They Are Not. **The American Statistician**, Vol. 50, No. 3, p. 203-206, 1996.

SEBASTIANI, R. G.; VIALI, L. Teste de Hipóteses: uma análise dos erros cometidos por alunos de engenharia. **Bolema**, Rio Claro (SP), v. 24, n. 40, p. 835-854. 2011

SELKE, T. BAYARRI, M. J. BERGER, J. Calibration of p Values for Testing Precise Null Hypotheses. **The American Statistician**, Vol. 55, No. 1, p. 62-71, 2001.

SETH, A.; CARLSON, K.D.; HATFIELD,D.E.; LAN, H.W. **So what? Beyond statistical significance to substantive significance in strategy research.** In: BERGH, D.D.; KETCHEN, D.J. Research in Methodology in Strategy and Management, V.5. Emerald, p. 3-27. 2009

SHEPHERD, J. "Deterrence Versus Brutalization: Capital Punishment's Differing Impacts Across States**." Michigan Law Review** 104: 203-255. 2005.

SPANOS, A. Review of S. T. Ziliak and D. N. McCloskey's The Cult of Statistical Significance. **Erasmus Journal for Philosophy and Economics**, 1 (1), p. 154-164, 2008.

SPANOS, A. **Statistical Foundations of Econometric Modeling.** New York: Cambridge University Press, 1993.

SPANOS, A. Statistical adequacy and the trustworthiness of empirical evidence: Statistical vs. substantive information. **Economic Modelling 27**, p. 1436–1452. 2010.

SPANOS, A.; MCGUIRK, A. The Model Specification Problem from a Probabilistic Reduction Perspective. **Journal of the American Agricultural Association** 83, p. 1168– 1176. 2001.

SPIELMAN, S. The Logic of Tests of Significance. **Philosophy of Science**, Vol. 41, No. 3, p. 211-226. 1974.

STERLING, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. **Journal of the American Statistical Association**, 54, 30-34. 1959.

SUPREME COURT OF THE UNITED STATES. **Matrixx initiatives, Inc., et alii. V. Siracusano et alii.** Certiorari to the United States Court of Appeals for the Ninth Circuit, 2011.

TVERSKY, A. KAHNEMAN, D. Belief in the law of small numbers. **Psychological Bulletin**, v. 76, p. 105-110. 1971.

TVERSKY, A. KAHNEMAN, D. Judgment under Uncertainty: Heuristics and Biases. **Science,** New Series, Vol. 185, No. 4157, p. 1124-1131, 1974.

WALD, A. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. **The Annals of Mathematical Statistics**, Vol. 10, No. 4, p. 299-326. 1939.

WALD, A. Statistical Decisions Functions. **The Annals of Mathematical Statistics**, Vol. 20, No. 2, p. 165-205. 1949.

WAGENMAKERS, EJ. A practical solution to the pervasive problems of p values. **Psychonomic Bulletin & Review**, 14 (5), p. 779-804. 2007.

WOOLDRIDGE, J. M. **Introdução à Econometria:** uma abordagem moderna. São Paulo: Pioneira Thomson Learning, 2006.

WOOLDRIDGE, J. M. Statistical Significance is Okay Too: Comments on "Size Matters". **Journal of Socio-Economics,** 33(5): p. 577-580. 2004.

ZELLNER, A. Posterior odds ratios for regression hypotheses: General considerations and some specific results. **Journal of Econometrics,** v 16, nº 01, p. 151-152, 1981.

ZELLNER, A. To Test or not to Test and if So, How? Comments on "Size Matters". **Journal of Socio-Economics,** 33(5): p. 581-586.2004.

ZILIAK, S. T.; MCCLOSKEY, D. N. Size Matters: The Standard Error of Regressions in the American Economic Review. **Journal of Socio-Economics,** 33(5): p. 527-46, 2004a.

ZILIAK, S. T.; MCCLOSKEY, D. N. Significance Redux. Replies to comments by Elliot, Granger, Horowitz, Leamer, O'Brien, Thorbecke, and Zellner. **Journal of Socio-Economics,** 33(5): p. 665-75, 2004b.

ZILIAK, S. T.; MCCLOSKEY, D. N. **The Cult of Statistical Significance:** How the Standard Error Costs Us Jobs, Justice and Lives. Ann Arbor: The University of Michigan Press, 2008a.

ZILIAK, S. T.; MCCLOSKEY, D. N. Science is judgment, not only calculation: a reply to Aris Spanos's review of The cult of statistical significance. **Erasmus Journal for Philosophy and Economics,** Volume 1, Issue 1, p. 165-170, 2008b.

ZILIAK, S. T.; MCCLOSKEY, D. N. Signifying nothing: reply to Hoover and Siegler. **Journal of Economic Methodology,** v. 15(1) , março, p.39-55. 2008c.

ZILIAK, S. T. ; MCCLOSKEY, D. N. The Cult of Statistical Significance. **JSM, Section on Statistical Education**: 2302-19. 2009.

ZIMMERMAN, P. "State Executions, Deterrence, and the Incidence of Murder." **Journal of Applied Economics** 7: 163-193. 2004.