Nutri·Time

Revista Eletrônica

Vol. 18, Nº 01, jan/fev de 2021
ISSN: 1983-9006
www.nutritime.com.br

Valor P: o que fizemos dele na experimentação zootécnica?

Conclusões científicas, inferência, interpretação experimental.

Edenio Detmann

Zootecnista, D.Sc., Professor Titular, Departamento de Zootecnia, Universidade Federal de Viçosa. Pesquisador do CNPq e do INCT Ciência Animal. E-mail: detmann@ufv.br.

RESUMO

A experimentação zootécnica é um campo científico amplo e plural. Contudo, muitos aspectos adotados em seu escopo são extremamente conservadores. Nesse artigo, de forma particular, são discutidos aspectos da tomada de decisão em trabalhos de zootecnia com base no valor P obtidos em experimentos. São destacados, principalmente, os aspectos relativos ao que não se deve concluir a partir do valor P.

Palavras-chave: conclusões científicas, inferência, interpretação experimental.

P VALUE: WHAT DID WE DO WITH IT IN THE ANIMAL SCIENCE EXPERIMENTS?

ABSTRACT

Animal science experiments are a wide and plural scientific field. However, many aspects that has been adopted in their scope are extremely conservative. In this article, in particular, some aspects of decision making in animal science experiments are discussed based on the P value obtained in such experiments. Mainly, the aspects related to what should not be concluded from the P value are highlighted.

Keyword: experimental interpretation, inference, scientific conclusions.

A quem se predispôs a ler esse artigo, afirmo de antemão que se trata de um artigo de opinião e divulgação científica. Logo, alguns pontos devem estar bem claros. Primeiro, isso reflete minha opinião, com a qual você, caro leitor, pode ou não concordar. Segundo, não há aqui qualquer amarra tradicional com a redação científica formal e rígida. Na verdade, há aqui um chamamento à reflexão. Esse texto constitui uma tentativa de criar uma fagulha a qual, sei nesse momento, não se transformará em uma fogueira de forma instantânea. Talvez haja duas fogueiras aqui. A primeira, a qual eu gostaria de ver queimando para produzir luz e calor à ciência que praticamos. A segunda, já esperada, à qual me refiro metaforicamente aqui. Uma fogueira de desconfiança e, quem sabe, rejeição aos meus argumentos. Vocês entenderão essa segunda visão nos parágrafos que se seguem.

De certa forma, posso adjetivar a experimentação zootécnica como "conservadora" (no sentido de resistente a mudanças). Pagarei um preço por essa adjetivação, mas estou disposto a isso. Por que conservadora? Porque estamos anos luz atrás de outras áreas científicas. Nossa comunicação científica é antiga, pois estamos sempre nos baseando no conceito de "ensino". O que quero dizer com isso? Que escrevemos como se nosso leitor não soubesse de nada antes de nós. Usamos introduções prolixas, expondo coisas do tipo: "a pecuária nos trópicos é importante, pois...", "gramíneas são relevantes para a alimentação de bovinos, pois...", "a nutrição é importante para a produção animal, pois...", etc. Isso é conhecimento notório e, com probabilidade próxima ao absoluto, não acrescenta nada na descrição do problema estudado ou no entendimento de nossas hipóteses e objetivos. Isso não se restringe à seção introdução, infelizmente. Não nos comunicamos com objetividade e modernidade. Isso é um fato e faço aqui *mea culpa*, pois assim agi em demasia.

Sempre escrevemos trabalhos nos baseando no fato de sermos os primeiros a escrever sobre o assunto e, intrinsicamente, julgamos nosso leitor da mesma forma. Talvez haja algo mais agravante nesses fatos, pois, muitas vezes, os corretores (i.e., orientadores, examinadores de bancas, revisores) de nossos trabalhos nos cobram isso. Lamento, mas não está correto.

Não podemos mais proceder assim. A comunicação científica moderna é dinâmica, simples e direta. Escrevemos para leitores informados. Ao agirmos assim consumimos menos espaço, demandamos menos tempo para a leitura e nos expressamos de forma direta, reta e simples ao expormos nossa contribuição à comunidade científica (na verdade, a contribuição deveria ser dada à ciência). Se o leitor, supostamente um profissional da área de Zootecnia, não sabe o que é um pH ruminal, o problema não é nosso. Pode ser duro dizer, mas é isso mesmo. Ele que vá se (in)formar melhor. Imagine se a cada trabalho precisássemos explicar cada detalhe daquilo que medimos? Os trabalhos publicados ao longo dos anos seriam cada vez maiores e, possivelmente, desgostosos de serem devidamente apreciados. Sem apreciação pela comunidade científica, não haveria agregação ou aprimoramento de conhecimento na grande rede internacional chamada ciência.

Não pretendo aqui fazer apologia a nenhuma linha de pensamento ou, muito menos, realizar proselitismo em relação ao estudo da filosofia da ciência, epistemologia ou comunicação científica. Em primeiro lugar, e acima de tudo, não me considero detentor de conhecimento suficiente, muito menos necessário, para militar nessas áreas. Deixo isso para aqueles que possuem mais conhecimento do que eu. A questão é levantar alguns pontos que julgo relevantes no planejamento, análise e interpretação de experimentos (aqui incluo as conclusões, pois são essas que serão expostas à comunidade científica e que garantirão o correto "fazer ciência"). Será que estamos no caminho correto?

Nos últimos anos, uma certa revolução tem acontecido no campo da estatística experimental. Essa "revolução" não terá volta ou muito menos poderá ser ignorada. Muito dos pontos abordados (e questionados) estão intimamente associados ao que praticamos em nossos experimentos na área de Zootecnia. Em suma, como discutirei com maior profundidade posteriormente, o que se exigirá da construção de conclusões de trabalhos científicos é uma atitude mais "pensativa"[1] (WASSERSTEIN et al., 2019) ou, em melhor português, mais pensada.

[1] Esse termo é uma tradução literal do inglês *thoughtful*.

Em palavras mais diretas, entende-se que raciocinar sobre os resultados será muito mais relevante para a construção de conclusões do que simplesmente aplicar-se uma regra binária de aceitar ou rejeitar uma hipótese estatística de nulidade (H_0). Isso pode ser um problema para áreas nas quais o "pensativo" não seja estimulado[2]. Acredito que esse seja o nosso caso, infelizmente.

Estamos (e digo "estamos" no plural, me incluindo no coletivo) tão viesados em regras "culturais" de nossa área que, muitas vezes, não nos damos conta de que o "cultural" pode não representar o mais correto. O "mundo" da avaliação estatística de resultados hoje tenta deixar bem claro que existem dois tipos de diferenças a serem apontadas em nossos resultados: as **diferenças estatísticas** e as **diferenças práticas**[3].

Diferenças estatísticas dizem respeito às diferenças que a estatística aponta como possivelmente existentes por intermédio de um teste de hipóteses. Contudo, nem sempre diferenças estatísticas querem dizer algo. Apontar diferenças depende de uma série de fatores. Por exemplo, o poder dos testes de hipóteses será maior na medida em que o *n* de nossos experimentos aumenta. Isso é um fato. Contudo, isso deveria ser bom, mas nem sempre pode ser salutar.

Imaginemos a seguinte situação. Você conduz um experimento com bovinos confinados e consegue um *n* de 1000 animais por tratamento[4]. Sua intenção é medir o incremento no desempenho produtivo proporcionado pela adição de um antibiótico à dieta. Digamos que hajam dois grupos experimentais: um com e outro sem o antibiótico. Você conduz o experimento e chega à seguinte conclusão (a partir dos resultados): o grupo controle (sem o antibiótico) mostrou GMD de 1,289 kg; e o grupo com o antibiótico exibiu GMD de 1,292 kg. A diferença entre os grupos foi de 0,003 kg/d. Devido ao elevado *n*, o teste aponta diferenças (talvez com P<0,01). Essa seria sua resposta estatística. Mas, como profissional da área, qual seria a sua conclusão?

Então, "estatisticamente" falando, de maneira informal, sua conclusão deveria ser: o antibiótico é bom, pois amplia o desempenho animal. Contudo, eu posso te perguntar: o que significam 3 g a mais no GMD de um bovino confinado? Possivelmente, sua resposta seria: nada. Talvez sua balança não tenha sensibilidade suficiente para medir essa diferença quando um animal é colocado sobre a mesma. Vamos avaliar a situação em termos práticos: o incremento no ganho médio diário significa um ganho monetário de menos de R$ 0,03/d com o uso de um antibiótico que nos custa R$ 0,40/d. O que se ganharia? A resposta seria óbvia: não se ganharia nada! Essa é a resposta prática. Em outros termos: devo relevar simultaneamente a resposta apontada pela estatística com sua relevância para a área de interesse (0,003 kg/d?), com aquilo que ganhamos em termos produtivos, monetários ou em termos de tempo. Antecipadamente, posso lhes direcionar um questionamento: qual seria a conclusão mais plausível? A estatística ou a prática? Devemos considerar, nesse exemplo em particular, que a estatística satisfaz o ego (afinal, apontaremos diferença!), mas a prática nos traz a uma dura realidade pautada pela produção de conhecimento científico válido e, quem sabe, aplicável. De antemão podemos entender que a coisa aqui é uma virtude aristotélica, ou seja, a boa conclusão deveria se centrar num meio termo entre o estatístico e o prático.

Talvez o ator principal nesse filme de mistério (ou, quem sabe, terror) seja o famigerado "valor P". Aquele que suporta a famosa regra decisória que nos alivia o fardo de dizermos "que sim ou que não" em nossos experimentos. O maior entrave para aceitação da teoria de testes de hipóteses é justamente sua principal estrutura: usamos o valor P, uma variável contínua, para construirmos uma conclusão que se baseia em uma variável discreta e binária (DETMANN, 2018). Há claramente uma certa

[2] Sugiro aqui a leitura do excelente artigo de BOSCH (2018), cuja ideia central é sobre formar pensadores e não somente especialistas em alguma área do conhecimento nos programas de pós-graduação.

[3] Entendam que o termo "prático" aqui constitui um termo técnico e não um termo pejorativo. Veja mais detalhes ao longo do texto que segue. Alguns autores também usam o termo "significância científica" quando se referem a diferenças práticas. Discussão adicional sobre esses termos pode ser encontrada em RYAN (2013).

[4] Vamos supor aqui que o animal seja a unidade experimental.

incompatibilidade na ideia, que é acentuada pela dogmatização de α = 0,05 como regra de ouro para tomadas de decisões. Isso tem levado muitos cientistas a recomendar a "aposentadoria" da chamada "significância estatística" (AMRHEIN et al., 2019).

O conceito de valor P foi moldado pelo professor Ronald Fisher nos anos 1920[5], mas não como uma ferramenta definitiva para obtenção de conclusões estatísticas. Sua ideia era de fornecer de uma forma fluida, e não rígida, um olhar informal para julgar as evidências empíricas e construir conclusões. Ou seja, olhar para os dados experimentais e avaliar se os resultados eram ou não coerentes com um padrão ocorrido ao acaso. Segundo seu método, o cientista deveria estabelecer uma hipótese de nulidade (e.g., não há diferença). Assim, assumindo-se que essa hipótese seria verdadeira, calcular-se-ia a chance de se obter um resultado no mínimo tão extremo como o que foi obtido no experimento. Quanto menor fosse esse valor numérico (i.e., o valor P), maior a chance da hipótese de nulidade ser falsa. Isso guiaria o cientista à construção de suas conclusões. Posteriormente, o matemático polonês Jerzy Neyman e o estatístico britânico Egon Pearson, rivais declarados de Fisher, desenvolveram seu trabalho alternativo de avaliação de hipóteses, definindo conceitos como poder, falso positivo (i.e., erro tipo I), falso negativo (i.e., erro tipo II), etc[6], mas não deram importância ao valor P definido por seu desafeto[7].

Contudo, enquanto os rivais trocavam farpas, outros pesquisadores perderam a paciência e começaram a escrever manuais de estatística para cientistas, o que acabou implicando na criação de um sistema híbrido que levou o valor P fluido de Fisher para dentro do sistema rigoroso de testes de hipóteses de Neyman e Pearson. Então surgiu o casamento entre valor P e α = 0,05 (NUZZO, 2014).

A partir desse breve histórico, para entendermos melhor a causa de "todos os males" e o porquê de tamanha revolução, devemos nos reportar com maiores detalhes às declarações realizadas pela *American Statistical Association* (ASA) sobre o valor P, as quais podem ser encontradas em WASSERSTEIN & LAZAR (2016) e WASSERSTEIN (2016), com interpretações mais aplicadas apresentadas por YADDANAPUDI (2016). Esses trabalhos servem de base para o texto adaptado que apresento a seguir.

A ideia em si se baseia em uma definição de valor P e nos princípios que suportam essa definição. Assim, definimos valor P como a probabilidade, sob um modelo matemático específico, que um sumário estatístico dos dados seria igual ou mais extremo que seu valor observado (WASSERSTEIN, 2016). Resumidamente, se sua hipótese de nulidade pressupõe que haja igualdade entre médias, um valor P próximo de 1 indicaria que os dados estão próximos dessa condição estabelecida. Ao contrário, na medida que o valor P se distancia de 1 (em direção a 0) os dados estariam mostrando divergência em relação ao estabelecido por H_0 (i.e., a igualdade não estaria retratando os fatos). Notem que a condição de "distanciamento" se associa a uma variável contínua e nada, absolutamente nada, na interpretação sugere algum "divisor de águas" na qual nossa opinião sobre a convergência ou não à condição estabelecida por H_0 deveria mudar. Essa é a grande questão.

Para entendermos a definição é necessário que abordemos os princípios que a norteiam. Abordarei a seguir os mesmos princípios conforme apresentados por WASSERSTEIN (2016).

Princípio 1: *O valor P indica o quão incompatíveis são os dados em relação a um modelo estatístico específico.*

Resumidamente, o valor P representa uma sumarização de quão os dados se distanciam da condição de igualdade estabelecida por H_0[8]. Assim, quanto menor o valor P, maior seria a incompatibilidade de nossas evidências empíricas com aquilo que é estabelecido por H_0. Contudo, dizer que há maior incompatibilidade não necessariamente

[5] Veja mais detalhes em FISHER (1958).
[6] Veja mais detalhes em NEYMAN & PEARSON (1933).
[7] Essa breve descrição histórica foi adaptada dos textos de NUZZO (2014) e LEHMAN (2011).

[8] Estou assumindo aqui que o "modelo" mais comum a ser avaliado em um processo estatístico de análise de dados seja a hipótese de nulidade.

quer dizer que há diferença. Lembrem-se da ideia de "divisor de águas" anteriormente ressaltada. Esse "divisor" pode ser arbitrário e transforma algo contínuo em algo discreto e binário.

Para entendermos melhor essa colocação, podemos analisar a simulação apresentada na Figura 1. Nessa, simulei uma situação específica experimental representada pela linha negra. Percebemos claramente que na medida que a diferença entre as médias de tratamentos se amplia, menor se torna o valor P. Considerando que a hipótese de nulidade estabelece a igualdade entre as médias, o aumento na diferença acentua a discrepância entre os dados e o modelo estabelecido por H_0. Contudo, a Figura 1 também ressalta que outras características do experimento afetam o valor P. A redução/ampliação da precisão experimental ou a variação no número de unidades experimentais afetará diretamente a dimensão do valor P. Logo, a mesma diferença entre médias pode ser julgada (considerando $\alpha = 0,05$) como "significativa" ou "não significativa" dependendo de como essas características são consideradas no experimento. Assim, diferentes experimentos podem apresentar as mesmas diferenças entre tratamentos (i.e., de mesmo valor), mas serem julgadas diferentes na construção da conclusão do trabalho. Isso será discutido novamente nesse artigo; contudo, há aqui um primeiro alerta (ou um novo alerta) de que o julgamento de um experimento somente com base no valor P constitui uma atitude arbitrária e, algumas vezes, perigosa. Como uma mesma diferença pode ser "existente" em um experimento e "inexistente" em outro?

Princípio 2: *O valor P não mede a probabilidade de que a hipótese estudada é verdadeira ou a probabilidade de que o comportamento dos dados tenha sido produzido meramente pelo acaso*.

Embora a intepretação frequentista do valor P seja comprovada em muitos casos, o conceito de aplicação não deve ir mais longe do que isso[9]. A ideia deveria ser simples. O valor P é uma expressão do comportamento dos dados em relação a uma possível explanação hipotética e não sobre uma explanação absoluta *per si*. Trocando em miúdos, o valor P não deveria ser usado para suportar nem a hipótese de nulidade, muito menos a hipótese alternativa (H_a). Esse deve ser visto apenas como uma medida de divergência frente aos dados obtidos e a hipótese de nulidade pré-estabelecida. Um elevado valor P não indica que a hipótese de nulidade seja verdadeira, assim com um baixo valor P jamais deveria ser visto como um atestado de veracidade para a hipótese alternativa.

Princípio 3: *Conclusões científicas ou decisões comerciais ou políticas não deveriam se basear somente no fato de o valor P ultrapassar ou não um valor específico.*

Aqui critica-se diretamente a transformação de uma variável contínua (i.e., o valor P) por intermédio de uma regra binária (i.e., aceita-se ou rejeita-se). A prática corriqueira de reduzir todo o processo de análise de dados e, consequentemente, a inferência estatística a um "divisor de águas mecânico" (i.e., $P<0,05$) para suportar conclusões científicas pode, com grande probabilidade, levar a falsas crenças ou a uma desastrosa tomada de decisão (como veremos mais tarde neste artigo). Uma conclusão não se torna verdadeira ou falsa simplesmente pelo fato de o valor P se colocar sobre o que nós definimos como regiões de rejeição ou aceitação de H_0. Isso faz parte da realidade. Como discutido anteriormente, os cientistas devem trazer muitos outros fatos à tona, pois existe uma grande lacuna entre as diferenças estatísticas e práticas. Adicionalmente, devemos considerar que o valor P é apenas a "cereja do bolo". Para termos uma "cereja", devemos ter um bolo saboroso produzido segundo as boas práticas de confeitaria (DETMANN, 2018). Assim, elementos como o delineamento do estudo, a correta definição do número de unidade experimentais, a qualidade das mensurações realizadas, a definição correta do modelo matemático, etc., influenciarão o valor P. Tudo isso foi feito adequadamente? A interpretação correta do valor P poderá mudar se nossa resposta for sim ou não. Assim, como confiar na cereja se o bolo não está lá? Pragmaticamente, muitas de nossas considerações em trabalhos científicos demandam uma decisão binária (i.e., ou sim ou não). Contudo, essa necessidade não significa que um valor de P

[9] Uma demonstração prática e direta do conceito frequentista pode ser visualizada na Figura 1 (e sua adjacente discussão) do trabalho de REUTER & MOFFET (2016).

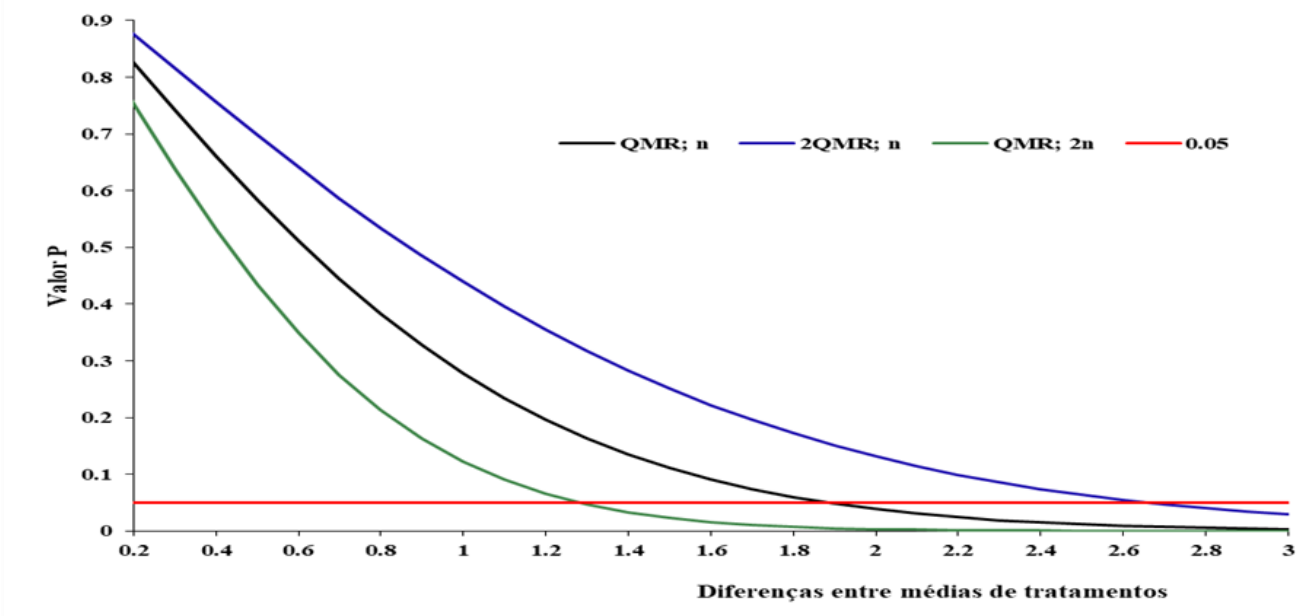

FIGURA 1. Simulação do comportamento do valor P em função de diferenças entre médias (experimento simulado em DIC com dois tratamentos; média 1 = 100; média 2 = 100 + X, sendo X a diferença entre tratamentos; QMR = 4; *n* = 10).

isoladamente assegure que nossa decisão estará ou não correta ou nos dará um salvo conduto para construção de conclusões adequadas.

Princípio 4: *Uma inferência adequada exige exposição completa dos dados e transparência.*

Esse princípio toca em uma ferida complicada e, muitas vezes, estimulada por especialistas da área de comunicação científica. Não é difícil de se encontrar sugestões do tipo “mostre somente os dados que suportem sua conclusão”. Isso pode ser um problema. Entendo perfeitamente que não estamos mais lidando com ciência do século XVII. Dados per si não suportam conclusões na direção de leis universais. É preciso interpretar e, para que haja interpretação, faz-se necessária a atuação do ator principal no processo: o cientista. Talvez, essa seja a solução e, ao mesmo tempo, o problema. Os dados (i.e., as evidências empíricas) suportam as conclusões e conclusões são o que importam, pois essas agregam verdadeiro valor epistemológico (ou, algumas vezes, metodológico) ao trabalho científico. Contudo, a ligação entre dados e conclusões pode ser feita de duas maneiras. Podemos concluir de forma geral, usando todo o escopo de informações que levantamos, ou podemos concluir de forma seletiva, usando aquilo que medimos e que suporte nossos desejos ou anseios. Isso pode ser um problema. Por exemplo, podemos medir a atividade de 15 enzimas, mas somente duas ou três suportam nossa hipótese de pesquisa. A partir daí, temos duas decisões possíveis: ou expomos apenas a atividade das duas ou três enzimas ou expomos a atividade das 15 enzimas. No primeiro caso, filtramos nossos interesses. No segundo caso, diluímos nossos interesses. Interesses são pessoais, mas o conhecimento gerado deveria independer de interesses pessoais[10].

[10] Não estou afirmando aqui que a ciência independa de pessoas. Não é isso. Para que a ciência seja feita, há necessidade do cientista. O que aqui argumento é que as conclusões (i.e., conhecimento científico gerado) são redigidas pelo cientista, mas baseando-se nas evidências empíricas. O que não pode ocorrer é construção de conclusões baseadas exclusivamente no que o cientista acha ao invés do que ele interpreta com lógica a partir dos dados e do suporte bibliográfico disponível. Se somente a opinião bastasse, não haveria necessidade de experimentos ou de um processo argumentativo (i.e., discurso lógico-científico sobre o comportamento dos resultados).

A questão é: qual das vias é mais correta? Não importa o que os despreparados revisores digam, a segunda será sempre a mais transparente[11]. Nossa interpretação deve guiar a redação, mas nunca deveríamos restringir a interpretação do leitor. "Filtrar" dados para suportar nossas conclusões pode significar andar no limiar entre a fraude e o real. A necessidade de apontar resultados "significativos" pode estar causando um viés extraordinário no que é publicado. Se temos diferenças, publica-se; caso contrário, engaveta-se[12]. Seria esse o caminho? Tenho plena consciência que não. Haverá sempre um componente aleatório em cada experimento que não pode ser contemplado pela avaliação de um experimento isoladamente. Essa é a ideia de visão meta-analítica dos dados. Somente por intermédio de uma meta-análise poderemos formular uma ideia mais clara se a coisa funciona ou não. Entretanto, um banco de dados viesado somente pode resultar em conclusões integradas viesadas. Ocultar resultados que não nos agradam ou que, momentaneamente, não suportam nossas conclusões, pode acarretar em uma visão global viesada sobre o assunto e a aplicação dessas ideias pode ser desastrosa no futuro. Uma intepretação criteriosa dos resultados por parte do leitor deve se calcar ao menos na informação de quantas hipóteses foram avaliadas (i.e., quantas variáveis-resposta). Mesmo que você não exponha nominalmente todas as variáveis, a informação de que as mesmas foram avaliadas mas não expostas e o como e o porquê essa seleção foi realizada devem constar no trabalho.

Princípio 5: *O valor P (ou a "significância estatística") per si não mede o tamanho de um efeito ou a importância do resultado.*

Na verdade, esse princípio resume um pouco o que discutimos no começo desse artigo. Existe uma clara e necessária distinção entre a diferença observada ser estatisticamente significante e ser relevante. Um baixo valor P não quer dizer que ser resultado é "forte" no campo específico que você trabalha. Seguir essa "linha de pensamento" (i.e., valores P seriam inversamente relacionados à dimensão do efeito) é limitar-se a não raciocinar. Nunca devemos esquecer que a estatística em si, na nossa área de atuação, é uma ferramenta de **orientação** na tomada de informação e na avaliação dos dados e um potente **auxílio** para a construção de conclusões. Em outras palavras, a estatística não produz resultados nem constrói conclusões por si mesma. É você, enquanto cientista, que faz isso. A intepretação de um efeito, em termos de tamanho ou importância é realizada pelo cientista, com base em todo o seu preparo técnico-intelectual, considerando ainda o estado da arte (i.e., todo o conhecimento prévio produzido por outros cientistas).

Talvez a grande questão aqui é que muitos querem se eximir de responsabilidades ou de se preparar cientificamente. Afinal, é muito mais cômodo delegar a conclusão ao teste de hipóteses do que a si mesmo. Caso algo dê errado, seria cômodo ter em quem colocar a culpa: na estatística. Isso está errado! O ator principal é o homem (i.e., o cientista). Nunca devemos esquecer ou omitir esse fato. A estatística e suas ferramentas **orientam** e **auxiliam**, mas cabe ao cientista, e somente a ele, o encargo e o dever de tomar decisões para interpretação dos dados e construção de conclusões.

[11] Os periódicos, atualmente, têm propiciado uma boa alternativa para esses casos. Caso você opte por não discutir todas as variáveis, você pode expô-las na forma de material *on line* suplementar. Assim, você consegue poupar espaço no artigo, mas evita de não expor publicamente aquilo que foi medido na totalidade. Atualmente, há também periódicos que se propõem a publicar apenas dados, os quais podem ser usados por outros autores, principalmente em abordagens meta-analíticas. Isso parece ser uma tendência, pois, teoricamente, não há limite para o armazenamento digital de dados; além de atribuir mais transparência aos trabalhos de pesquisa.

[12] Interessante, que a preocupação do impacto desse comportamento não é recente. Por exemplo, isso foi muito bem discutido por T.D. Sterling em 1959! Recentemente, críticas severas têm sido apontadas por diversos autores, entre os quais sugiro a leitura de IOANNIDIS (2005) e HEAD et al. (2015). Esses últimos autores destacam de forma preocupante trabalhos onde busca-se de todas as "maneiras possíveis" encontrar resultados que apontem diferenças e seus consequentes impactos no conhecimento científico. De fato, a proporção de trabalhos publicados que apontam diferenças tem aumentado progressivamente, incluindo aqui as ciências agrárias (vide FANELLI, 2012). Será esse o caminho correto? Estamos expondo ciência sem o devido viés ou estamos expondo o que é causado pela desenfreada busca por $P<0,05$?

Como exposto na Figura 1, a detecção de diferenças pode ser afetada pelo tamanho do experimento e por sua precisão. Caso haja alteração nessas características, um efeito de grande dimensão pode se associar a valores P muito altos; ao passo que efeitos muitas vezes diminutos podem produzir valores P muito próximos de zero. Valor P não é adjetivo para importância de efeito detectado ou não. É preciso mais. É preciso pensar.

Essa questão é exemplificada na Figura 2. Podemos perceber que os experimentos A e B apontaram a mesma diferença entre as médias de tratamentos, o que indica a mesma diferença prática. Contudo, devido à diferença na precisão de ambos, o pesquisador responsável pelo experimento A concluirá pela ausência de diferença (o intervalo de confiança para a diferença entre médias inclui o valor paramétrico 0), ao passo que o pesquisador responsável pelo experimento B concluirá pela presença de diferença entre tratamentos (o intervalo de confiança para a diferença entre médias não inclui o valor paramétrico 0). A questão que levanto é: duas diferenças numericamente idênticas em duas situações experimentais distintas serão julgadas distintamente? Um é e outra não é? Raciocine comigo: faz sentido? A resposta é de certa forma complexa, mas o raciocínio base parte do princípio que esse aparente julgamento foi realizado apenas com base no valor P. Entendeu o perigo?

A Figura 2 pode ainda gerar outros questionamentos. Contrastemos os experimentos A e C. O experimento A apontou uma grande diferença entre tratamentos, mas com alto valor P (i.e., não significativo). De outra forma, o experimento C apontou uma diferença bem inferior entre os tratamentos, porém com baixo valor P (ou seja, significativo). Chegamos com isso a duas constatações. Primeiro, que o valor P não é diretamente proporcional ao tamanho da diferença. Isso se dá pelo fato de outras características experimentais (e.g., tamanho da amostra, precisão das mensurações) afetarem o comportamento dos dados. Segundo, poderíamos concluir que o menor é melhor que o maior! Imagine que a diferença represente diferenças no GMD dos animais. Entendeu, pela segunda vez, o perigo?

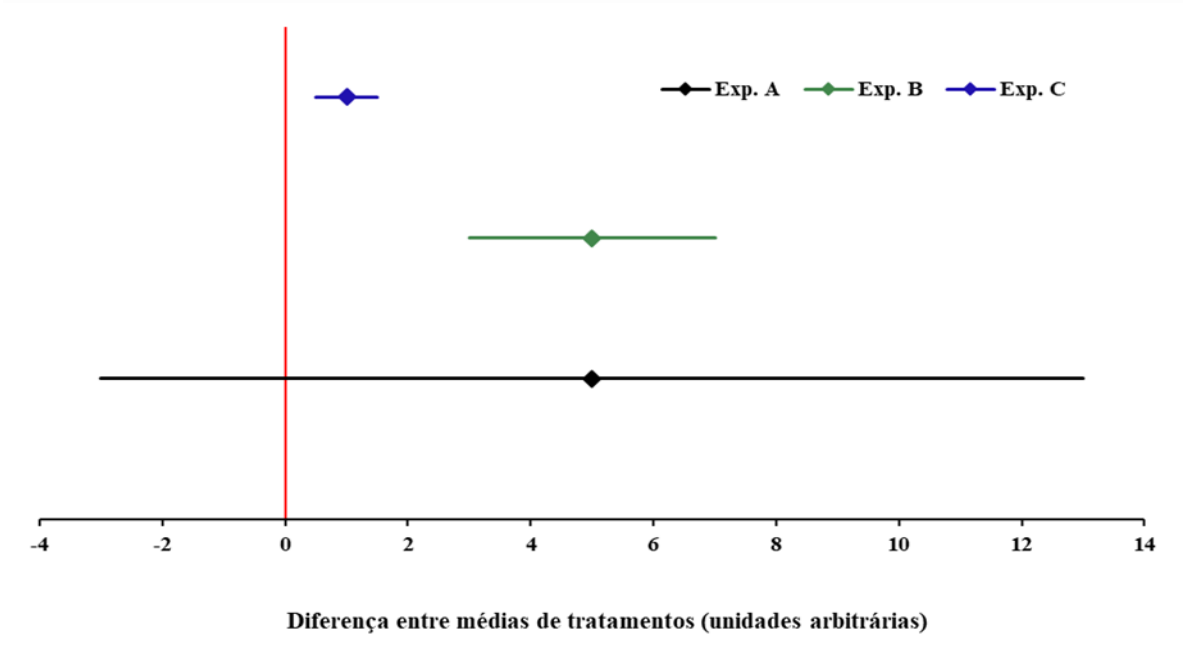

FIGURA 2. Diferença entre as médias de dois tratamentos, com o seus respectivos intervalos de confiança, para três diferentes experimentos. Hipoteticamente, em todos os experimentos foram avaliados os mesmos tratamentos e a mesma variável resposta foi mensurada[13]. Interpretação simplificada: Exp. A - diferença não significativa, valor P elevado; Exp. B e C - diferenças significativas, baixo valor P. A significância estatística é apontada quando o intervalo de confiança para a diferença entre médias não inclui o valor 0.

[13] Essa Figura foi baseada no exemplo apresentado por AMRHEIN et al. (2019; p.306)

Por fim, ao contrastarmos os experimentos B e C chegamos à conclusão que ambos são estatisticamente similares, pois ambos apontarão diferenças com um baixo valor P. A minha questão é: a interpretação prática entre as duas diferenças apontadas será a mesma? Uma é "grande" e a outra é "pequena". Julgaremos as mesmas zootecnicamente similares? Entendeu, pela terceira vez, o perigo?

Não atribua a sua decisão a um objeto inanimado (i.e., valor P). Pense! Afinal, espera-se que você seja um cientista preparado em sua área de atuação. A vida é maior e mais ampla que o valor P.

Princípio 6: *Isoladamente, um valor P não fornece uma boa medida ou evidência a respeito de um modelo ou hipótese.*

Esse princípio, na verdade, resume os cinco anteriores. O que posso afirmar aqui é a citação completa da sentença de WASSERSTEIN (2016): *os pesquisadores devem reconhecer que um valor P fora de contexto ou sem outras evidências provê informação limitada.* Precisamos pensar sobre os dados. Raciocinar, considerando que o conhecimento de nossa área de atuação é muito mais importante que um simples diagnóstico estatístico. Não estou afirmando que a estatística seja dispensável ou um grande mal. Não é isso. Mas também devemos entender que a estatística não constitui uma panaceia universal. O que quero atentar é que a **ferramenta de auxílio** jamais pode se transformar na decisão em si. Você é maior que a estatística, embora nunca maior que a natureza. Isso se aplica especialmente àqueles responsáveis por formar pessoas ou por julgar o trabalho de outrem (especialmente os famigerados revisores *ad hoc*). Há muito mais na ciência do que valores P. O raciocínio lógico indutivo ou, algumas vezes, dedutivo, rege a construção do conhecimento. Um valor P não significa, nem poderia significar, uma conclusão de um trabalho científico. Não atrapalhem as mentes que querem exercer sua função vital e natural: pensar.

Podemos sumariar a aplicação desses princípios a partir do excelente trabalho de MATTHEWS (2000). Nesse artigo, o autor analisa a associação entre a população de cegonhas e o número de nascimentos de bebês em 17 países da Europa. De forma proposital, o autor se baseia na estória infantil de que cegonhas trazem os bebês. Assim, estatisticamente, a hipótese de nulidade se baseia na ausência de associação entre essas variáveis. Você pode argumentar a *priori* que seria óbvia a ausência de associação. Eu concordo, pois a biologia por detrás do nascimento de bebês é bem clara e nada tem a ver com o uso de cegonhas (bom, podem existir fetiches bastante incomuns, não tenho preconceito. Mesmo assim, os fetiches não alteram a biologia da fusão dos gametas). Contudo, supreendentemente, os dados evidenciaram associação[14] entre a população de cegonhas e o nascimento de bebês com valor P de 0,008!

Vamos às duas faces da coisa. Primeiro, vamos interpretar esse resultado como se fôssemos escravos do valor P. Nesse sentido, poderíamos dizer que teríamos 0,8% de chance de a hipótese de nulidade estar certa, o que nos levaria a concluir, pelo dogmático α de 0,05 que H_0 não se sustenta. Por outro lado, alguns poderiam indicar que há 99,2% de H_a representar o comportamento da população. Assim, sendo escravos do valor P isoladamente, teríamos que concluir que cegonhas realmente entregam os bebês aos pais e todos os livros já produzidos sobre reprodução animal (lembrem-se que o *Homo sapiens* é um animal) seriam apócrifos. Entendeu a cilada de se avaliar um valor P isolado e fazer com que ele represente a conclusão que nós, enquanto cientistas, teríamos que construir?

Obviamente, tudo que especulei no parágrafo anterior está incorreto[15]. Por favor, não cite o parágrafo anterior em seu seminário ou artigo científico como se fosse verdade (e, por favor, não me cite!). Com isso, vamos abordar a segunda face. Primeiro, o valor P nos indica que existe 0,8% de chance de encontrarmos uma amostra mais divergente em relação às condições estabelecidas por

[14] A associação, nesse caso, foi avaliada pela significância do coeficiente de correlação de Pearson, o qual assumiu valor estimado de 0,62. Nesses casos, expressamos as hipóteses: H_0: $\rho = 0$; H_a: $\rho \neq 0$. Somente lembro, o que é evidenciado pela discussão no texto, que associação não representa relação de causa e efeito

[15] Maiores detalhes sobre como estas intepretações estão equivocadas podem ser obtidos em GREENLAND et al. (2016).

H_0 do que a que encontramos (ou 1 em 125 amostras; 1/125 = 0,008). Essa é a interpretação inicial e correta. Em primeiro plano, isso deve nos chamar a atenção, pois o comportamento observado em nossa amostra sugere ser incomum, pois aproxima-se do difícil ou do raro. Contudo, isso não significa uma conclusão, apenas uma sugestão de direcionamento para discussão. A indicação é de que a divergência em relação à hipótese de nulidade foi intensa. A partir desse ponto, devemos fazer a pergunta essencial a todos os trabalhos de investigação científica: por quê?

No cultivo da escravidão ao valor P, onde o cérebro funciona com economia de bateria, concluiríamos que cegonhas realmente entregam os bebês. Contudo, quando ativamos o cérebro, já de início, evidenciamos que isso é um completo absurdo. Devemos pensar, avaliar o contexto. Nesse caso específico, há uma variável não considerada no enunciado do problema: o tamanho do país. Quanto maior a área do país, maior o número de nascimentos e maior a área para servir de hábitat para as cegonhas. Logo, países maiores possuem mais cegonhas e mais bebês nascendo. Isso é o que significa ser "pensativo". Esse é a forma de se caminhar para construção de uma discussão; ou seja, começamos a ir além do valor P, usando a lógica/raciocínio. Assim, apesar do diagnóstico dado pelo valor P, a nossa conclusão seria óbvia e lógica: não há nada de mágico nas cegonhas. São animais belos, mas em nada tem a ver com a intensidade da concepção em seres humanos.

Mas, o que será do futuro? Essa é uma pergunta interessante e necessária. Nós trabalhamos em uma área bastante conservadora e que tem uma resistência acima do normal em relação a mudanças. Mas o mundo está mudando e, cedo ou tarde, a Zootecnia terá que mudar.

Como já citado anteriormente, a preocupação a respeito do que o valor P se transformou[16] gerou medidas sobre a ASA. Esse movimento, que se tornou mais proeminente a partir de 2014, não cessou. As coisas estão evoluindo e não há retorno. Eu poderia aqui citar o título em português do clássico filme de Jean-Claude Van Damme de 1986: "Retroceder nunca, render-se jamais". Em 2019, a ASA publicou um número especial do periódico *The American Statistician*, trabalho resultante de um simpósio específico de 2017, com mais de 40 artigos versando sobre alternativas ao uso corriqueiro do valor P. Os artigos, em si, são plurais, versando desde a substituição até o uso com cautela em conjunto com outras características experimentais. Abordar todos aqui seria imprudente e eu jamais faria isso. De qualquer forma, independentemente da postura, todos os artigos apontam em uma direção comum: chega de sermos escravos do valor P e, principalmente, do dogmático α = 0,05. O mundo é plural e amplo. Há mais em sua avaliação do que um nível de significância. Isso parece estar sendo perdido pelos cientistas ao longo do tempo. Fomos escravizados pela significância estatística e esquecemos de olhar para aquilo que realmente buscamos entender: a natureza.

[16] Melhor seria dizer: o que nós fizemos dele.

AGRADECIMENTOS

Ao CNPq e ao INCT Ciência Animal pelo suporte. Aos Drs. Tadeu Eder da Silva (Cargill/Nutron), João Paulo P. Rodrigues (UNIFESSPA), Janaína Martuscello (UFSJ) e Kelly S. C. Detmann pela avaliação crítica do manuscrito.

REFERÊNCIAS

AMRHEIN, V.; GREENLAND, S.; McSHANE, B. Retire statistical significance. **Nature**, v.567, p.305-307, 2019.

BOSCH, G. Train PhD students to be thinkers not just specialists. **Nature**, v.554, p.277, 2018.

DETMANN, E. **Não seja como as vaquinhas!** Uma abordagem informal sobre as formalidades dos experimentos com animais de produção. 2 ed. Viçosa: Edenio Detmann, 2018. 373p.

FANELLI, D. Negative results are disappearing from most disciplines and countries. **Scientometrics**, v.90, p.891-904, 2012.

FISHER, R.A. **Statistical methods for research workers**. 13 ed. London: Oliver & Boid, 1958. 351p.

GREENLAND, S.; SENN, S.J.; ROTHMAN, K.J.; CARLIN, J.B.; POOLE, C.; GOODMAN,S.N.; ALTMAN, D.G. Statistical tests, P-values, confidence intervals, and power: a guide to

misinterpretations. **The European Journal of Epidemiology,** v.31, p.337-350, 2016.

HEAD, M.L.; HOLMAN, L.; LANFEAR, R.; KAHN, A.T.; JENNIONS, M.D. The extent and consequences of p-hacking in science. **Plos Biology**, v.13, e1002106, 2015.

IOANNIDIS, J.P.A. Why most published research findings are false? **Plos Medicine**, v.2, p.e124, 2005.

LEHMAN, E.L. **Fisher, Neyman, and the creation of classical statistics**. New York: Springer, 2011. 115p.

MATTHEWS, R. Storks deliver babies (P = 0.008). **Teaching Statistics**, v.22, p.36-38, 2000.

NEYMAN, J.; PEARSON, E.S. On the problem of the most efficient tests of statistical hypotheses. **Philosophical Transactions of the Royal Society A**, v.231, p.289-337, 1933.

NUZZO, R. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. **Nature**, v.506, p.150-152, 2014.

REUTER, R.R.; MOFFET, C.A. Designing a grazing experiment that can reliably detect meaningful differences. **The Professional Animal Scientist**, v.32, p.19-30, 2016.

RYAN, T.P. **Sample size determination and power**. Hoboken: John Wiley & Sons, 2013. 374p.

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**, v.54, p.30-34, 1959.

WASSERSTEIN, R.L. ASA statement on statistical significance and P-values. **The American Statistician**, v.70, p.131-133, 2016.

WASSERSTEIN, R.L.; LAZAR, N.A. The ASA's statement on p-values: context, process, and purpose. **The American Statistician**, v.70, p.129-131, 2016.

WASSERSTEIN, R.L.; SCHIRM, A.L.; LAZAR, N.A. Moving to a world beyond "p<0.05". **The American Statistician**, v.73, p.1-19, 2019 (Special Supplement).

YADDANAPUDI, L.N. The American Statistical Association statement on P-values explained. **Journal of Anesthesiology and Clinical Pharmacology**, v.32, p.421-423, 2016.