XXVI SIMPÓSIO BRASILEIRO DE RECURSOS HIDRÍCOS

TESTES DE HIPÓTESES: LARGAMENTE EMPREGADOS, MAS MUITO MAL COMPREENDIDOS

Veber Afonso Figueiredo Costa [1] *& Dirceu Silveira Reis Jr.* [2]

Abstract: Hypothesis tests are commonly utilized in the practice of Statistical Hydrology. However, the interpretation of the tests' results is, more often than not, incompatible with the theoretical foundations underlaying distinct tests procedures, which leads to inappropriate conclusions and inconsistencies in water resources management-related decision making, under the false premise that such decisions are supported by solid statistical principles. In this paper, we present the theoretical principles, and the main uses of the Fisher significance tests and the Neyman-Pearson hypothesis tests, as well as the flawed logic of their hybridization in the null hypothesis significance tests (NHSTs), which aim at supporting substantive theories (e.g., stationarity, homogeneity and independence) solely on the basis of the numerical outcomes of the tests. Examples of the misinterpretation and erroneous utilization of NHSTs are discussed through trend tests and goodness-of-fit tests usually employed by researchers in water resources for, respectively, justifying nonstationary modeling of hydrological processes and model discrimination for frequency analysis of hydroclimatic variables. We intend to demonstrate, with the presented discussion, the needs for deeper understanding by engineers and hydrologists before utilizing statistical tools as a means to provide effective communication of risks to distinct stakeholders.

Resumo: Testes de hipóteses são correntemente utilizados na prática da Hidrologia Estatística. Não obstante, a interpretação dos resultados é, na maioria dos casos, incompatível com a fundamentação teórica subjacente aos distintos procedimentos de teste, o que leva a conclusões inadequadas e inconsistências na tomada de decisão associada à gestão de recursos hídricos, sob a falsa premissa de que tais decisões são amparadas por princípios estatísticos sólidos. Neste artigo, são apresentados os princípios teóricos e as principais utilizações dos testes de significância de Fisher e os testes de hipótese de Neyman-Pearson, bem como a lógica falha de sua hibridização nos testes de significância de hipótese nula (NHSTs), os quais visam sustentar teorias substantivas (e.g., estacionariedade, homogeneidade ou independência) unicamente a partir dos resultados numéricos do teste. Exemplos de interpretação equivocada e utilização errônea de NHSTs são discutidos a partir de testes de tendência e testes de aderência usualmente empregados por pesquisadores em recursos hídricos para, respectivamente, justificar a modelagem não estacionária de processos hidrológicos e discriminar modelos distributivos para análise de frequência de variáveis hidroclimáticas. Tenciona-se demonstrar, com a discussão apresentada, a necessidade de aprofundamento teórico de engenheiros e hidrólogos antes da utilização de ferramentas estatísticas, de maneira a prover a comunicação efetiva do risco aos diferentes *stakeholders*.

Palavras-Chave – Testes de significância, testes de hipótese, teste de significância de hipótese nula

1) Departamento de Engenharia Hidráulica e Recursos Hídricos (EHR). Universidade Federal de Minas Gerais, Belo Horizonte – MG. veber@ehr.ufmg.br.
2) Departamento de Engenharia Civil e Ambiental (ENC). Universidade Federal de Brasília, Brasília – DF. dirceureis@unb.br

INTRODUÇÃO

Teste de hipóteses constituem uma ferramenta comumente utilizada na prática da Hidrologia Estatística. Com efeito, procedimentos usuais de análise de frequência de variáveis hidroclimáticas são habitualmente precedidos por "verificações" de independência, homogeneidade e estacionariedade dos processos estocásticos subjacentes (Naghettini, 2017), as quais fundamentam a seleção de métodos para estimação paramétrica e quantificação de incertezas. Testes de hipótese são também empregados na avaliação de aderência de modelos distributivos a amostras de variáveis hidrológicas, uma vez que o raciocínio dedutivo raramente se aplica à prescrição de distribuições de probabilidades em tal área do conhecimento. É assim inegável que a correta utilização dessas ferramentas compreende um requisito para a adequada tomada de decisão no contexto de dimensionamento de estruturas hidráulicas e da gestão de recursos hídricos.

Em termos práticos, procedimentos de teste são empregados para se confrontar conjecturas acerca do comportamento populacional de uma variável e a evidência agregada por uma amostra observada – tal confronto fornece sustentação às premissas empregadas na análise de frequência e na estimativa de risco, em especial no contexto de extrapolação. Por outro lado, a interpretação dos resultados dos distintos procedimentos de testes por hidrólogos e engenheiros é quase sempre inconsistente com a teoria estatística subjacente a sua formulação. Essa constatação se manifesta de maneira óbvia na utilização indiscriminada, por pesquisadores em recursos hídricos, dos testes de significância de hipótese nula (NHSTs, do inglês "*null hypothesis significance tests*"), que combinam, sem amparo teórico, aspectos associados aos testes de significância propostos por Ronald Fisher e a abordagem probabilidade-consequência introduzida por Jerzy Neyman e Egon Pearson.

Com efeito, a incompatibilidade teórica e os diferentes objetivos das abordagens mencionadas são, há muito, reconhecidos por estatísticos, mas tais distinções conceituais são frequentemente ignoradas por pesquisadores de outras áreas, incluindo-se a Hidrologia. Os elementos dessa incompatibilidade são discutidos em Hubbart e Bayarri (2003). Por certo, ao se negligenciar aspectos teóricos, a interpretação dos resultados dos testes e a validade de suas conclusões se tornam questionáveis (Perezgonzalez, 2015). Não obstante, a imensa expansão da literatura associada à modelagem não estacionária de variáveis hidrológicas com base unicamente em testes de tendência nas últimas décadas, bem como a virtual ausência de validação desses modelos frente a novas evidências empíricas (e.g., mudanças nas direções das tendências com a incorporação de novas informações), constituem um exemplo claro da utilização inadequada de preceitos fundamentais da inferência estatística, sob a inverídica premissa de que as conclusões apresentadas são amparadas por tais preceitos – doutrina essa frequentemente denotada por "afirmar o consequente": se a consequência é verdadeira, então a causa deve necessariamente ser verdadeira (Serinaldi; Kilsby; Lombardo, 2018).

Neste artigo, um breve histórico e as principais distinções entre procedimentos de teste de hipóteses são apresentadas. Aspectos teóricos são discutidos de maneira a ilustrar vantagens e limitações das diferentes variantes. Por fim, uma síntese da utilização de NHSTs no contexto da Hidrologia Estatística, com ênfase em testes de tendência e de aderência, é apresentada com intuito de se prover perspectivas teoricamente consistentes dessas ferramentas para utilização tanto em contexto acadêmico quanto na prática da Engenharia Hidrológica.

VARIAÇÕES DE UM MESMO TEMA? COMO EQUÍVOCOS CONCEITUAIS SUBVERTEM A TOMADA DE DECISÃO SOB INCERTEZA

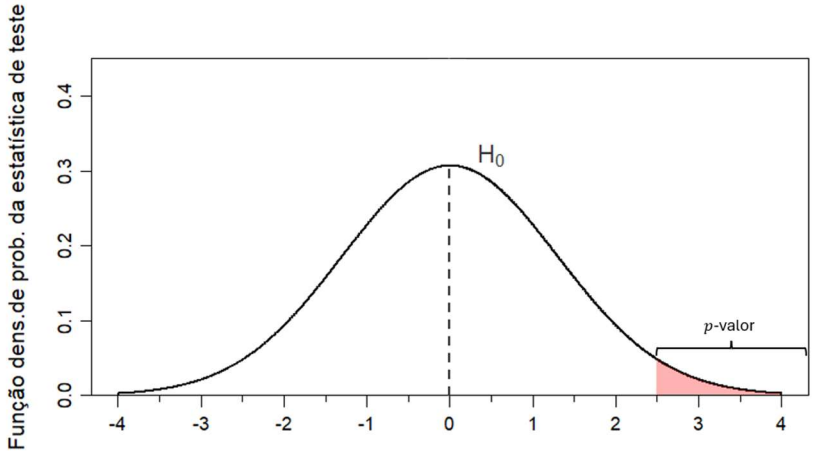
Procedimentos de teste, sob diferentes propósitos e fundamentações conceituais, têm sido empregados desde a década de 1920, quando Ronald Fisher, um dos pesquisadores mais influentes da Estatística Matemática, introduziu a sistemática dos testes de significância – uma abordagem geral de inferência (i.e., após a observação dos dados) destinada a "invalidar" a hipótese de que um dado resultado de pesquisa decorra meramente de flutuações amostrais (Fisher, 1925). Três anos depois, Neyman e Pearson (1928) apresentaram um procedimento alternativo de teste, voltado à tomada de decisão, no qual duas hipóteses explicitamente definidas são contrapostas e a rejeição de uma delas implica necessariamente a aceitação da outra. Tal rejeição é embasada pela quantificação prévia (i.e., antes da observação dos dados) dos possíveis erros decorrentes da natureza probabilística do teste e do tamanho das amostras disponíveis. Neyman e Pearson (1928) advogavam que sua teoria compreendia um avanço teórico com relação aos testes de significância, particularmente pela inclusão dos conceitos de poder de teste, que possibilita a escolha do teste mais efetivo para a situação em análise (Fisher fazia menção ao conceito de "sensibilidade", similar à noção de poder de teste, mas nunca o formalizara nos testes de significância (Perezgonzalez, 2015)), e "*effect size*", o qual permite estabelecer o tamanho amostral necessário à "percepção" de uma diferença de magnitude δ entre as duas hipóteses postuladas. Por outro lado, Fisher se opunha fortemente à concepção dos testes de hipótese, argumentando que a aceitação irrestrita de uma das conjecturas em análise é inconsistente com o raciocínio puramente indutivo sob o qual se fundamenta a inferência estatística frequentista.

Em meio ao intenso debate que envolveu as duas abordagens mencionadas – o qual perdura por décadas na comunidade estatística (Gigerenzer, 2004; Perezgonzalez, 2015) –, um sistema híbrido, que combina, sem fundamentação teórica, as abordagens de Fisher e Neyman-Pearson, foi apresentado na década de 1940, sob a denominação de NHSTs. NHSTs são formulados de maneira a contemplar uma hipótese alternativa, posta *a priori*, assim como na teoria de Neyman e Pearson. Tal hipótese, contudo, é postulada como sendo complementar e mutuamente excludente com relação à hipótese nula (i.e., as duas hipóteses são exaustivas), o que exclui da análise o "*effect size*" e o poder de teste. Diante disso, a significância estatística é avaliada *a posteriori*, a partir dos p-valores, e a aceitação (rejeição) da hipótese nula se dá, frequentemente de forma errônea, por meio do problema da probabilidade inversa: o p-valor é geralmente interpretado de forma equivocada como a probabilidade de que a hipótese nula seja verdadeira, dadas as observações $(P(H_0|D))$, e não como a probabilidade de observar aquela amostra em particular, dado que a hipótese nula é verdadeira $(P(D|H_0))$, como originalmente estabelecido por Fisher.

Apesar das inconsistências teóricas decorrentes da "hibridização" das abordagens de Fisher e Neyman-Pearson, os NHSTs se tornaram difundidos em distintas áreas do conhecimento, incluindo-se a Hidrologia Estatística. De fato, testes com intuito de aferir aleatoriedade, independência, homogeneidade, estacionariedade e aderência são correntemente utilizados na análise estatística de variáveis hidroclimáticas. Contudo, tais testes são usualmente inaptos a informar o que se postula, de maneira complementar e mutualmente exclusiva, nas hipóteses nula e alternativa – como essa última não é explicitamente definida, múltiplos processos físicos, por vezes de natureza incompatível (e.g., tendências monotônicas ou "saltos"), são agregados indistintamente em sua formulação (Serinaldi; Kilsby; Lombardo, 2018). Dessa maneira, ao se rejeitar a hipótese nula, não é possível identificar a população postulada na hipótese alternativa. Em outras palavras, a não rejeição apenas indica ausência de evidências para se excluir a hipótese nula da análise, mas a rejeição é virtualmente inconclusiva sob uma perspectiva física.

Os testes de significância de Fisher são essencialmente inferenciais e seu principal intuito é avaliar a probabilidade de se observar valores mais extremos que a estatística de teste D sob a hipótese nula H_0 (p-valor) – para propósitos práticos, H_0 é sempre assumida como verdadeira e a distribuição de amostragem resultante estabelece o padrão de variabilidade "esperado" para a estatística em análise; tal distribuição é deduzida de maneira teórica, mas alguns de seus parâmetros são inferidos da amostra observada. A significância estatística é então aferida, *a posteriori*, por meio dos p-valores – menores magnitudes dessa grandeza constituem evidências mais fortes contra a hipótese nula, i.e., quanto menores os p-valores, menores as probabilidades de que os resultados obtidos decorram meramente de flutuações aleatórias do processo postulado na hipótese de pesquisa (Figura 1).

Figura 1 – Ilustração de um teste de significância



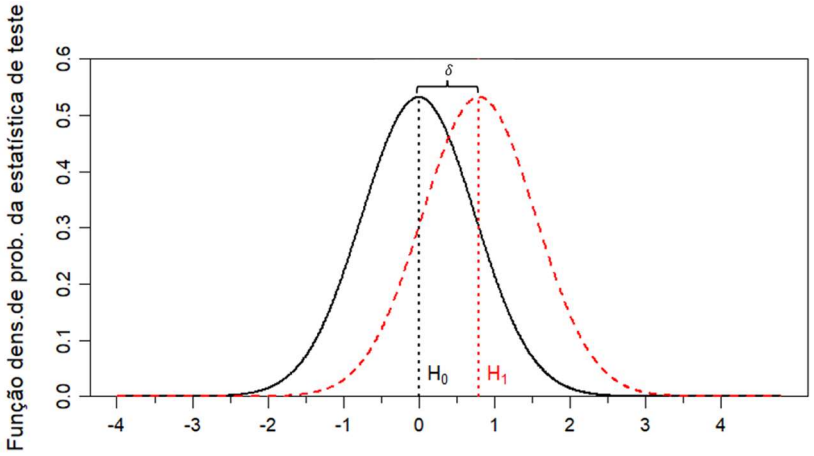
Cabe notar que, sob a abordagem de Fisher, não é necessário definir um nível de significância *a priori* e a decisão de quão reduzido deve ser o p-valor para que a H_0 seja rejeitada cabe estritamente ao pesquisador. Dessa maneira, é usual, quando do uso de testes de significância, que somente o p-valor seja reportado, o que confere flexibilidade (e também subjetividade) à tomada de decisão – p-valores iguais a 0,049 e 0,051 traduzem aproximadamente a mesma evidência contra H_0 (Perezgonzalez, 2015). Por outro lado, deve-se atentar que, caso o nível de significância seja especificado *a priori*, esse corresponde a um p-valor teórico sob a hipótese nula e não deve ser confundido, do ponto de vista conceitual, com o erro do tipo I discutido na abordagem de Neyman-Pearson – esse se refere à frequência de rejeições incorretas da hipótese nula no longo-termo (princípio da repetição do experimento).

O caráter puramente inferencial (*a posteriori*) dos testes de significância exclui as noções de hipótese alternativa, "*effect size*" ou poder de teste. Nesse caso, a rejeição de H_0 somente implica a revisão da hipótese de pesquisa. Por outro lado, tal fato permite que os testes de significância sejam realizados sob distintas hipóteses nulas para uma mesma amostra observada – essa é uma distinção fundamental com relação aos testes de hipótese de Neyman e Pearson, os quais devem ser realizados somente uma vez. A possibilidade de realização de múltiplos testes torna os testes de significância especialmente úteis a estudos exploratórios ou *ad hoc*, os quais tencionam estabelecer relações entre variáveis e, assim, prover possíveis descrições conceituais dos processos sob investigação (Serinaldi; Kilsby; Lombardo, 2018). Tal característica, por outro lado, faz com que os testes de significância sejam inadequados em estágios confirmatórios de pesquisa.

Neyman e Pearson, por sua vez, desenvolveram uma abordagem de teste com foco em tomada de decisão. Para tanto, assume-se, *a priori*, que a amostra a ser testada pode ter sido extraída de duas populações distintas: uma postulada na hipótese nula H_0 e que será, de fato, testada, e outra postulada

sob uma hipótese alternativa H_1, que materializa o "*effect size*" (Figura 2). Testes de hipótese são baseados na quantificação formal (*a priori*) das probabilidades de se tomar decisões incorretas diante do "*effect size*" estipulado: o erro do tipo I, usualmente denotado por α, que compreende a probabilidade de se rejeitar uma hipótese nula verdadeira e delimita a chama região crítica ou região de rejeição; e o erro do tipo II, denotado por β, que contempla a probabilidade de não se rejeitar a hipótese nula quando essa é falsa – o complemento de β com relação a 1 é denominado poder de teste e constitui um importante critério para a seleção do teste mais adequado ao problema em estudo (Naghettini, 2017). Definidos o "*effect size*", α e β, é possível determinar (também *a priori*) o tamanho da amostra necessário à realização do teste.

Figura 2 – Ilustração de um teste de hipótese



Em essência, a abordagem de Neyman-Pearson preconiza que a rejeição de H_0 acarreta a aceitação de H_1 e vice-versa; em outras palavras, uma das alternativas em análise é mais "compatível" com a evidência empírica. Ademais, como os possíveis erros e suas consequências são quantificados *a priori*, os testes de hipótese podem ser realizados uma única vez para uma amostra observada em particular. Tal fato os torna ferramentas adequadas para estágios confirmatórios de pesquisa, mas os exclui como alternativa em estudos exploratórios – as hipóteses de pesquisa são postas, necessariamente, antes da observação dos dados.

Além disso, diferentemente da abordagem de Fisher, é usual se reportar os valores calculado e crítico da estatística de teste (i.e., quantis da distribuição de amostragem sob H_0) quando testes de hipótese são empregados. Deve-se ressaltar, contudo, que a probabilidade de excedência da estatística de teste é numericamente equivalente ao p-valor calculado, para a mesma hipótese nula, em um teste de significância. Essa coincidência frequentemente ocasiona grande confusão na interpretação dos resultados dos diferentes tipos de teste: pesquisadores comumente reportam o nível de significância do teste em função do p-valor obtido (e.g., "o teste é significativo a um nível α arbitrário"). Porém, o nível de significância de um teste de hipótese é sempre posto *a priori* e, assim, os resultados possíveis somente informam se a estatística de teste é significativa ou não. Por fim, cabe notar que os erros do tipo I e II compreendem frequências de longo termo e somente possuem significado mediante a repetição do experimento; de fato, é impossível se estabelecer, para um teste em particular, qual dos erros foi cometido.

É interessante notar que, mesmo diante das diferenças conceituais e de objetivos dos testes de significância e de hipóteses, a impossibilidade de se determinar a probabilidade de H_0 era reconhecida tanto por Fisher – que admitia que o conhecimento advindo de raciocínio indutivo é provisório, uma vez que as conclusões podem ser invalidadas à medida que novas observações da variável em análise são obtidas –, quanto por Neyman e Pearson – que somente preconizavam a escolha da hipótese mais

plausível frente à informação disponível, sem a pretensão de se determinar se alguma delas é, de fato, verdadeira. Em outras palavras, tanto Fisher quanto Neyman e Pearson reconheciam que, em função da natureza aleatória dos fenômenos modelados, procedimentos de teste não podem ser conclusivos. Não obstante, o advento dos NHSTs, em grande medida, subverteu tal constatação: ao se postular duas hipóteses exaustivas e mutuamente excludentes, e necessariamente se aceitar uma delas, o resultado do teste forçosamente indicaria a veracidade de uma das alternativas em análise – um exemplo clássico de se "afirmar o consequente". Tal condição é especialmente crítica em casos de rejeição, tendo em vista o estado de conhecimento incompleto acerca dos processos físicos acomodados na formulação de H_1.

A lógica inerente à construção dos NHSTs é essencialmente falha: ao se postular duas hipóteses concorrentes, tal como nos testes de Neyman-Pearson, a rejeição de H_0 deveria ser baseada em um nível de significância α definido *a priori* (em função do teste (Serinaldi; Kilsby; Lombardo, 2018)). Contudo, uma vez que somente H_0 é explicitamente definida na formulação do teste, o nível de significância é avaliado *a posteriori*, por meio dos p-valores (em função dos dados), tal como testes de Fisher. Diante disso, um p-valor baixo é erroneamente entendido como uma probabilidade reduzida de que H_0 seja verdadeira (o problema da probabilidade inversa) e, como nenhuma outra hipótese além de H_1 é contemplada, tal resultado indicaria que a hipótese alternativa é provavelmente verdadeira. Obviamente, p-valores, por definição, são inaptos a fornecer evidências acerca da veracidade tanto de H_0 quanto de H_1. Com efeito, a probabilidade condicional de H_0, que compreende o real interesse de pesquisa, e o p-valor são usualmente diferentes e podem ser relacionados pelo teorema de Bayes. Formalmente:

$$P(H_0|\mathcal{D}) = \frac{P(\mathcal{D}|H_0)P(H_0)}{P(\mathcal{D}|H_0)P(H_0)+P(\mathcal{D}|H_1)P(H_1)} \quad (1)$$

A Equação 1 evidencia a distinção conceitual entre $P(H_0|\mathcal{D})$ e $P(\mathcal{D}|H_0)$: um p-valor reduzido não implica, necessariamente, probabilidade condicional reduzida. Dessa maneira, o p-valor, sozinho, não constitui elemento para embasar a exclusão de H_0 em favor de H_1.

Outro aspecto crítico associado aos NHSTs é que esses, em sua maioria, não se propõem a testar hipóteses estatísticas (e.g., $H_0: \mu = 0 \times H_1: \mu = 1$), que investigam diferenças entre parâmetros populacionais, mas teorias substantivas relacionadas a mecanismos causais ou relações entre variáveis (e.g., H_0: não há tendência no processo estocástico $\times$ H_1: há tendência no processo estocástico; (Koutsoyiannis, 2023; Serinaldi; Kilsby; Lombardo, 2018). Nesse contexto, a hipótese de pesquisa é colocada *a posteriori*, como uma consequência do resultado do teste. Em outras palavras, o NHST é, inapropriadamente, utilizado em estágio exploratório (e.g., o p-valor sugere que H_1 é mais "compatível" com a evidência observada) e como instrumento confirmatório (e.g., o modelo desenvolvido sob H_1, para a mesma amostra observada, é mais adequado que aquele obtido sob H_0). Tal lógica, contudo, mostra novamente a incoerência de se "afirmar o consequente", uma vez que rejeições incorretas podem se originar da incompatibilidade entre o processo postulado na hipótese nula e a realidade física do fenômeno em análise, e não da conformidade dessa realidade física com a hipótese alternativa. Por exemplo, a rejeição em um teste de tendências pode ser devida à consideração de independência na formulação de H_0 e não à existência de não estacionariedade no processo estocástico subjacente (Serinaldi; Kilsby; Lombardo, 2018). Ademais, como indicado por Serinaldi, Kilsby e Lombardo (2018), significância estatística pode ser detectada em um dado teste somente em decorrência do tamanho elevado da amostra utilizada na inferência – ainda que o "*effect size*" e o poder de teste sejam excluídos da análise, a variância da distribuição de amostragem da estatística de teste sob a hipótese nula decresce com a maior disponibilidade de pontos amostrais. A significância física, por outro lado, não pode ser estabelecida por meio dos resultados de NHSTs: a

natureza aleatória dos processos em análise impede a prescrição de relações determinísticas de causa e efeito, as quais requerem raciocínio dedutivo.

Por fim, a seleção *a posteriori* da significância estatística também introduz inconsistências teóricas aos resultados dos NHSTs. De fato, é usual que a significância seja comunicada com base em níveis diversos e arbitrários, como, por exemplo, 0,10, 0,05 ou 0,01. Tal "flexibilidade" compreende uma manifestação inequívoca da confusão entre o erro do tipo I, da abordagem de Neyman-Pearson, e o p-valor, obtido nos testes de Fisher – a quantificação prévia daquele tem por intuito explicitar as consequências de uma escolha incorreta antes que a decisão seja tomada. Assim, o resultado do teste somente pode indicar se a estatística de teste é significativa ou não em relação àquele nível especificado *a priori*. Caso tal lógica seja invertida, a região de rejeição pode ser "ajustada", após a observação dos dados (i.e., em função do p-valor obtido), para se embasar ou refutar uma determinada hipótese de pesquisa, o que, por certo, invalida os propósitos de um teste.

Em visto do exposto, apesar da natureza rejeição-aceitação dos NHSTs, os resultados desses testes são, como aqueles dos testes de Fisher e de Neyman-Pearson, largamente inconclusivos: p-valores de maior magnitude somente indicam que não há elementos para se excluir a hipótese nula enquanto a rejeição dessa hipótese, com base no p-valor, não implica a veracidade de H_1. Nesse contexto, NHSTs são insuficientes para sustentar teorias substantivas (e.g., modelagem não estacionária), somente refletem diferenças com relação a processos específicos postulados na hipótese nula (e.g., variáveis independentes e igualmente distribuídas (IID)) e são inaptos a acomodar a ideia de poder de teste, o que exclui a busca por um teste mais efetivo. Assim, como recomendado por Busuioc e von Storch (1996), NHSTs não devem ser tratados como testes de fato, mas como meras ferramentas para investigação preliminar – de fato, a construção de teorias substantivas deve necessariamente envolver raciocínio dedutivo. A seguir, as limitações inerentes aos NHSTs são exploradas no contexto da Hidrologia Estatística, a partir de exemplos associados a testes de tendência e de aderência.

NHSTS E A HIDROLOGIA ESTATÍSTICA

A modelagem probabilística de processos hidroclimáticos é, via de regra, a etapa inicial para estudos hidrológicos e projetos de estruturas hidráulicas. Nesse contexto, duas vertentes de testes de hipótese são usualmente empregadas – aqueles que objetivam descrever os processos de referência específicos que nortearão a inferência estatística (e.g., variáveis IID) e aqueles que visam avaliar a adequabilidade de modelos distributivos para a estimativa de quantis ou probabilidades de interesse.

Como exemplo da primeira vertente, Naghettini (2017) sugere que a análise de frequência de variáveis hidroclimáticas seja precedida de testes de aleatoriedade, independência, homogeneidade e estacionariedade – todos envolvem teorias substantivas a respeito dos processos (não das séries temporais) em análise. Em particular, a estacionariedade tem sido objeto de inúmeras investigações em virtude de potenciais efeitos de mudanças climáticas antropogênicas no ciclo hidrológico.

O conceito de estacionariedade, por outro lado, é frequentemente incompreendido por engenheiros e hidrólogos, que, em geral, o associam a "tendências", sejam elas monotônicas ou abruptas, em séries temporais. Não obstante, tal conceito envolve a distribuição conjunta das variáveis aleatórias em análise, a qual, em um contexto estacionário, é mesma para qualquer tamanho de amostra e qualquer *lag* considerados (note-se que a amostra aqui referida não é a amostra observada e sim um conjunto de variáveis aleatórias descritas por uma distribuição marginal e uma estrutura de dependência temporal). Dessa maneira, a não estacionariedade é necessariamente relacionada a alterações na distribuição conjunta das variáveis (ou em alguma de suas medidas descritivas populacionais) ao longo do tempo – a noção de estacionariedade não se aplica a séries temporais, que

constituem realizações do processo estocástico subjacente (i.e., são variáveis comuns) e são afetadas por erros de amostragem.

Não obstante as inconsistências conceituais, NHSTs têm sido correntemente empregados para se detectar "não estacionariedades em séries temporais de variáveis hidrológicas" e, com isso, legitimar a utilização de modelos não estacionários para estimação de riscos futuros. Há aqui vários equívocos. Em primeiro lugar, a não estacionariedade é uma característica do processo estocástico e, como tal, deve ser estabelecida por meio de dedução, antes da inferência estatística. Com efeito, assumir a não estacionariedade com base unicamente em um teste de tendência é tão somente "afirmar o consequente"; um novo ponto amostral pode até mesmo inverter o sentido da tendência inicialmente observada, indicando que tais tendências compreendem somente flutuações aleatórias. Dessa maneira, para que a definição de não estacionariedade se aplique, é necessário que as mudanças nas medidas descritivas do processo estocástico sejam inteiramente previsíveis (Koutsoyiannis, 2023). Em outras palavras, o processo somente será não estacionário se suas medidas descritivas forem expressas como funções determinísticas do tempo.

Cabe ressaltar ainda que, comumente, distintos testes, como Pettitt, Mann-Kendall (Kendall, 1970; Mann, 1945) e Spearman, são empregados para que, em conjunto, corroborem a existência de não estacionariedade no processo estocástico em estudo. Por outro lado, os processos postulados nas hipóteses nulas de cada um dos testes são essencialmente diferentes e, dessa maneira, distintas características populacionais são testadas para se sustentar uma única teórica substantiva – essa é uma óbvia decorrência da indefinição do poder de testes dos NHSTs. Assim, a fundamentação inerente a cada teste faz com os resultados não sejam comparáveis, sob uma perspectiva teórica, e a utilização de vários testes somente enviesa o processo de inferência – o modelador pode recorrer unicamente aos testes cujos resultados amparam sua hipótese de pesquisa. O viés pode ser amplificado nesses casos pela seleção *a posteriori* do nível de significância do teste. Com efeito, é relativamente comum que pesquisadores recorram a erros do tipo I com maior magnitude (novamente, em uma confusão com o p-valor) para que a hipótese alternativa, que contempla tendências em algumas das medidas descritivas populacionais da variável em análise, seja aceita em um ou mais testes.

Por fim, a rejeição da hipótese nula em testes de tendências tem sido empregada como justificativa para a modelagem não estacionária, por meio de abordagens como os modelos aditivos generalizados (GAM, do inglês "*generalized additive models*") ou os modelos aditivos generalizados para posição, escala e forma (GAMLSS, do inglês "*generalized additive models for location, scale and shape*") empregando o tempo *per se* como covariável. Tal expediente, contudo, é afetado pela mesma lógica falha inerente aos testes de tendência: uma vez que os modelos de regressão capturam a "variabilidade temporal" da série observada *in lieu* da possível não estacionariedade no processo estocástico subjacente (o que envolve dedução), a eficiência desses modelos é restrita ao período utilizado na estimação paramétrica. Ao se empregar modelos *data-driven* não estacionários sob a perspectiva de validação, as predições, em geral, se mostram incompatíveis com a evolução temporal de fato observada para a variável em análise. Novamente, a doutrina de se "afirmar o consequente" pode ter severas implicações para a gestão de recursos hídricos, especialmente no que concerne ao *trade-off* segurança-custo: medidas de adaptação baseadas somente em resultados de NHSTs podem ser bastante onerosas, ao passo que a quantificação apropriada do risco, por meio de modelos estacionários apropriadamente construídos para acomodar as características de processos hidrológicos mais usuais, como a dependência de longo termo (Koutsoyiannis, 2023), pode constituir uma abordagem mais interessante para se garantir níveis de segurança adequados e custos de adaptação coerentes com as receitas dos entes responsáveis.

A segunda vertente mencionada compreende os testes de aderência, os quais têm por objetivo avaliar a "adequação" de um modelo distributivo postulado na hipótese nula e uma amostra observada. Frequentemente, testes de aderência avaliam a distância máxima entre a probabilidade de não excedência empírica e aquela derivada do modelo teórico – expedientes para penalização de maiores distâncias nas regiões das caudas da distribuição de probabilidades são comuns para "melhorar o poder do teste" em relação aos quantis de interesse prático. Por outro lado, a distribuição da estatística de teste sob H_0 depende tanto do modelo teórico quanto do vetor de parâmetros correspondente, o qual é necessariamente estimado a partir da amostra observada. Em outras palavras, enquanto em grande parte dos NHSTs a distribuição da estatística de teste independe tanto do modelo distributivo quanto de seus parâmetros, os testes de aderência requerem sua incorporação explícita para estimação da distribuição de amostragem sob H_0. Ademais, testes de aderência comumente não consideram o número de parâmetros dos modelos distributivos em sua formulação. Não obstante, a incerteza epistêmica decorrente da inferência estatística é fortemente afetada pela complexidade dos modelos, em função das covariâncias entre os parâmetros (Naghettini, 2017). Depreende-se do exposto que os testes de aderência não constituem ferramentas para discriminação de modelos; eles se prestam somente para indicar a exclusão ou não da distribuição postulada na hipótese nula, tendo-se em conta as limitações inerentes aos NHSTs discutidas anteriormente.

É usual, porém, que pesquisadores tentem justificar a escolha de um determinado modelo probabilístico em função dos p-valores estimados em testes de aderência, argumentando que maiores valores dessa grandeza são relacionados a menores "desvios" com relação ao comportamento esperado para uma dada distribuição postulada H_0. Não obstante, como a distribuição da estatística de teste em qualquer teste de aderência não é independente do modelo distributivo prescrito na hipótese nula, os p-valores correspondem a diferentes quantis de diferentes distribuições e, dessa maneira, os testes não são comparáveis entre si. Também nesse contexto, os NHSTs têm sido empregados erroneamente, o que, em grande medida, oculta as enormes dificuldades para a discriminação de modelos para a análise de frequência. Idealmente, essa tarefa deve ser alicerçada na junção de argumentos dedutivos, como, por exemplo, o princípio de máxima entropia para se postular, *a priori*, o tipo de decaimento esperado para as caudas da distribuição, e dos princípios de raciocínio indutivo, como a utilização de estimadores pouco ou não enviesados e mais eficientes. Tal abordagem provê mecanismos mais efetivos para a seleção de um modelo teórico e para a estimação de parâmetros, o que, por certo, permite a extrapolação com maior confiabilidade.

CONSIDERAÇÕES FINAIS

Testes de hipótese são empregados com frequência na prática da Hidrologia Estatística, mas a incompreensão da fundamentação teórica das diferentes vertentes de teste, bem como a disseminação dos NSHTs para "verificações" de teorias substantivas, ainda que esses não sejam aptos a fazê-lo, tem originado severas inconsistências na interpretação dos resultados dos testes usualmente utilizados na análise de frequência de variáveis hidrológicas e como subsídio à gestão de recursos hídricos. Identificar tais inconsistências a partir dos preceitos teóricos dos diferentes procedimentos para teste constitui o principal objetivo deste artigo.

A partir de exemplos relacionados a testes de tendência, cujo emprego aumentou consideravelmente nas últimas décadas em decorrência de possíveis efeitos de mudanças climáticas antropogênicas, e testes de aderência, comumente (e erroneamente) empregados como ferramentas para discriminação de modelos, buscou-se ilustrar a enorme incompatibilidade entre as teorias substantivas sob análise e o que é efetivamente comunicado a partir dos resultados dos testes de hipótese. Tal incompatibilidade revela a necessidade de aprofundamento teórico de engenheiros e pesquisadores em recursos hídricos antes da utilização de ferramentas estatísticas. Somente tal

aprofundamento permitirá a estimação consistente e a comunicação eficiente de riscos para a tomada de decisão sob incerteza.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelas bolsas concedidas durante a realização deste estudo.

REFERÊNCIAS

BUSUIOC, Aristita; STORCH, Hans Von. Changes in the winter precipitation in Romania and its relation to the large-scale circulation. **Tellus A**, v. 48, n. 4, p. 538–552, ago. 1996.

FISHER, Ronald Aylmer. **Statistical methods for research workers**. Edinburgh: Oliver and Boyd, 1925.

GIGERENZER, Gerd. Mindless statistics. **The Journal of Socio-Economics**, v. 33, n. 5, p. 587–606, nov. 2004.

HUBBARD, Raymond; AND BAYARRI, M. J. Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing. **The American Statistician**, v. 57, n. 3, p. 171–178, 1 ago. 2003.

KENDALL, Maurice G. **Rank correlation methods**. 4th ed ed. London: Griffin, 1970.

KOUTSOYIANNIS, Demetris. **Stochastics of Hydroclimatic Extremes - A Cool Look at Risk**. 3. ed. Athens: Kallipos, Open Academic Editions, 2023.

MANN, Henry B. Nonparametric Tests Against Trend. **Econometrica**, v. 13, n. 3, p. 245, jul. 1945.

NAGHETTINI, Mauro (ORG.). **Fundamentals of Statistical Hydrology**. Cham, Switzerland: Springer International Publishing, 2017.

NEYMAN, J.; PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. **Biometrika**, v. 20A, n. 3–4, p. 263–294, 1928.

PEREZGONZALEZ, Jose D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. **Frontiers in Psychology**, v. 6, 3 mar. 2015.

SERINALDI, Francesco; KILSBY, Chris G.; LOMBARDO, Federico. Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. **Advances in Water Resources**, v. 111, p. 132–155, jan. 2018.