P-values: Significance vs Hypotheses Testing
History and Present

Victor Kipnis

Biometry
National Cancer Institute, USA

Introduction

- "A growing chorus of concerns, from scientists and lay people, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring" – Collins & Tabak, Nature (2014)
- Could statistical testing of biomedical effects contribute to problems with reproducibility?

Significance testing - definition

- Although statistics similar to p-values and probabilistic reasoning akin to significance tests existed well before , Fisher's "Statistical Methods for Research Workers" formalized the concept in 1925 and expanded its reach to experimenters
- Formally, let empirical data x be observed value of a random $X \sim F(x)$, H_0 be a null hypothesis specifying the model $F(x) = F_0(x)$, and test statistic $T(x)$ be summary of the data such that the larger $T(x)$ the more inconsistent is x with H_0
- Then p-value is $p = Pr\{T(X) \geq T(x)|H_0\}$, i.e., the probability under the null that a test statistic would be *equal to or more extreme* than its observed value and regarded as measure of concordance with H_0

Significance testing - interpretation

- p-value $p(\boldsymbol{X}) = 1 - F_0(\boldsymbol{X})$ is *data dependent*, thus is a *random variable* which has a uniform distribution $p \sim U[0, 1]$
- p-value is just a convenient transformation of test statistic $T(\boldsymbol{X})$ to a probability scale and therefore carries the *same information* as $T(\boldsymbol{X})$
- Significance test is based on *inductive logic*: either H_0 is true and a rare event has occurred or H_0 is false
- Real issue: whether the *actual magnitude* of p-value could be given formal quantitative interpretation in terms of evidence against the null

Significance testing - weaknesses

- "Rare event" under H_0 is NOT the event of observing *actual data* but includes "more extreme" data that *have not been observed*
- "Rare event" under H_0 maybe even more rare under *alternative model* $F_A(x)$ which is not formally specified in significance testing
- While $p \sim U[0,1]$ for ANY test statistic $T(\boldsymbol{X})$, ANY sample size n, and ANY alternative model $F_A(x)$, its distribution under possible $F_A(x)$ depends on $T(\boldsymbol{X})$ and sample size n
- Since the choice of $T(\boldsymbol{X})$, alternative $F_A(x)$, and sample size n are not specified, p-values are difficult to formally interpret

Hypothesis testing

- Neyman-Pearson theory of *hypothesis testing* dismissed the notion of *inductive logic*, considering statistical tests as ways of *making decisions*
- N-P introduced two hypotheses, the null H_0 and the alternative H_A
- Testing procedure involves *deciding* between two courses of action: to proceed as if H_0 is true or as if H_A is true
- The decision is based on critical region C : if $T(\boldsymbol{X}) \in C$ then reject H_0 and accept H_A, otherwise accept H_0

Hypothesis testing

- Decision involves two types of error: Type I error $\alpha = Pr\{T(\boldsymbol{X}) \in C|H_0\}$ and Type II error $\beta = Pr\{T(\boldsymbol{X}) \notin C|H_A\}$
- Decision rule: choose C so that for a prespecified *test size* α, the *power,* i.e., the probability $1-\beta$ to detect the alternative is maximized
- Ideally, $T(\boldsymbol{X})$ is chosen to produce the most powerful test
- Given $T(\boldsymbol{X})$, C depends on n and effect size, so N-P framework can be used to design studies with adequate sample size and power
- Unlike *data dependent* p-values which have no role in N-P theory, α and $1-\beta$ are NOT random variables and their values should be prespecified *before data are observed*

Current testing framework

- Nowadays, statistical testing is a hybrid of Fisher's *significance testing* and Neyman-Pearson *hypothesis testing*
- N-P procedure is used in designing a study with adequate Type I error and power, but p-values are also calculated and their actual magnitude is often interpreted as quantitative evidence against H_0 in favor of H_A (significant effect, highly significant, border line significant, etc)

Current testing & its reproducibility

- Suppose an intervention is evaluated by measuring the effect of applying it to one of two groups and comparing the result with the other group
- Assume that the measured difference $X \sim N(\mu, \sigma^2),\ H_0 : \mu = 0,$ $H_A : \mu \neq 0$, and the test statistic is $T(X) = X/\sigma$
- The sample size n is specified for the test to have size $\alpha = 0.05$ (two-sided) and power of 90% to detect a difference δ of interest

Current testing & its reproducibility

- Assume that this experiment produced a statistically significant result, i.e., $p < \alpha$
- The probability of repeating this result requires knowing true effect μ, so assume that it is equal to the observed difference $X = x$
- Repeating this experiment under identical conditions, what is the probability of observing another statistically significant result in the same direction as the first?

Current testing & its reproducibility

Table: Replication probabilities of statistical significance ($p \leq \alpha = 0.05$) as a function of the p-value of the initial experiment

Probability of $p \leq 0.05$ in repeat experiment	
p-value of initial experiment	$\mu = x$ in initial experiment
0.05	0.50
0.03	0.58
0.01	0.73
0.005	0.80
0.001	0.91

Discussion

- The replication probabilities are rather low and are *not in accord* with the informal credibility of the null hypothesis based on initial p-values
- In N-P test, the precise location of the observed $T(x)$ within the critical region (as indicated by the corresponding p-value) is irrelevant
- Thus, knowing only *whether* (not *where*) $T(x)$ fell in a critical region, the replication probability under H_A is *always* greater or equal to the predefined power of the test (90% in this case), therefore eliminating dissonance between intuition and the actual probabilities
- There is nothing wrong in calculating p-values, but there proper role in hypothesis testing is to indicate whether or not $p \leq \alpha$