Statistics in Medicine: I Don't Need a p Value . . . What's a p Value Anyway?

By Kerry Barker, Ken Story, and J. Kyle Wathen

Would you bet $25 to win $100?

We would gladly take the bet if the chance of winning were 96 percent. We certainly would not change our minds if the odds dropped to 94 percent. However, many scientists make different decisions on the basis of the p values inherent in this example—whether p = 0.04 or p = 0.06.

Most scientists utilize Frequentists statistics to prevent abuse of data. Frequentists require that you specify your hypothesis, statistical test, and the criteria of success in advance. Often this criterion is a "significant" p value, especially in medical studies. While p ≤ 0.05 is often the criterion of choice, one may also choose 0.01/0.001 for "definitive studies" or 0.10/0.15 for "exploratory studies."

Much has been written about why this is a flawed strategy. Most statisticians admit the strategy is flawed in hallway conversations or even in articles such as this one. Clients of statisticians will state that they know this is only part of the conclusions from a data analysis. Yet, at the end of the day, p = 0.08 is deemed a failure and p = 0.04 is a success.

In this article, we will informally present the Frequentists dilemma and describe differences between the Frequentists and Bayesian schools of statistical thought.

The Frequentists dilemma

You are asked to analyze a study upon completion. Even though you get an interim data set, you wait until the study is done before analyzing because, a priori, you decided to wait until the study is done. You find significance, p = 0.05. However, in another analysis scheme, we a priori decided to perform an interim analysis. We tell you that significance was not obtained after adjusting for multiple looks. Who is right? Luckily, since both of our analytic strategies were stated "a priori," we are both right, even though our conclusions from the same data are completely different.

You and your friend develop a program for predicting winners in college basketball. It is only 50 percent accurate (a coin toss). Your friend looks further into the data and e-mails you stating that if you adjust for which team was at home and use only major conferences, the accuracy is now 99 percent. You smile at the foolishness of your friend, violating two important principles of statistics: adding a post-hoc analysis and analysis on a subgroup. Your friend decides to use the scheme to bet on games anyway. You meet one month later and remind your friend of his statistical foolishness. You also tell him that paying cash for a new Porsche with his winnings seems extravagant.

Have multiple looks at data been used inappropriately to declare significance? Yes. Have post-hoc analyses been used inappropriately? Yes. Have subgroup analyses been grossly misused? Yes, more times then we care to remember! Based on the cases above, the problems with Frequentists statistics seem obvious.

Bayesian methodology, disdained for years by Frequentists, uses prior information and subjective data. Worse, it is hard to do. However, in real life everyone uses prior information and subjectivity in making decisions. Buying a house? Hiring a person? Marrying? All use prior information and subjectivity. Given the housing bubble, human resource issues, and the divorce rate, many people are not good at this type of analysis. The solution is not to eliminate the use of this information, but to make better use of it.

Frequentists versus Bayesian analyses

Let **X** represent the observed data from a study comparing a placebo (P) and an experimental (E) treatment and Θ be an unknown parameter of interest. Assume that the larger the value of Θ, the bigger the improvement E provides over P and a value of zero when E and P are equivalent.

Both Frequentists and Bayesians have a common goal, making a statement about Θ.

For Frequentists, one calculates the probability of observing X or, more severely, assuming $\Theta = 0$, which can be represented as P(X or more extreme $\mid \Theta = 0$). If this probability is small (0.05), then the conclusion is that Θ is some value that is not zero. If the probability is not small, the conclusion is not that $\Theta = 0$; rather the conclusion is that there is not enough evidence to conclude $\Theta \neq 0$. Note neither statement is telling us much about Θ. A Bayesian analysis gives us the probability of Θ based on the observed data, which can be represented as P ($\Theta \mid$ X, for example, the probability that $\Theta > 0$ **given** X is 0.95, or in notation $Pr(\Theta > 0 \mid X) = 0.95$

While the goals of Frequentists and Bayesian analysis are the same, there is a big difference between

$$P(\text{ X or more extreme} \mid \Theta = 0) \text{ and } P(\Theta \mid X)$$

The Frequentists make a conclusion about Θ by assuming it is a particle value (here $\Theta = 0$) and calculate the probability of observing more extreme data under that assumption. This is a very roundabout way of analyzing Θ, and one must specify what is meant by "or more extreme." A Bayesian analysis is more straightforward, and the resulting probability directly references Θ, the parameter of interest. So why is not everyone a Bayesian? One reason is that $P(\Theta \mid X)$ is calculated using the following equation

$$P(\Theta \mid X) \approx P(\Theta \mid \Theta) * P(\Theta)$$

Where $P(\Theta \mid X)$ is called the posterior, $P(X \mid \Theta)$ is called the likelihood (sampling distribution), and $P(\Theta)$ is called the prior. The prior is what Frequentists have an issue with and what Bayesians devote much of their time to. Just as the name suggests, "the prior" is based on all knowledge prior to the data collection and may include historical data or "expert" opinion. There are many types of priors. Given that the choice of priors is based on judgment, a Frequentist would state that this introduces bias into the analysis.

Two Frequentists would obtain the same results analyzing a data set (if they used the same method), while two Bayesians could get different results using different priors.

Bayesian methodology is not as well known, even to statisticians, and seems more complex. However, new software tools are making common Bayesian analysis more accessible.

Points to remember

- Using Bayesian methods should be methodological, not ad hoc.
- Statistical methodology is not a substitute for planning. Careful consideration of practical issues that may influence the design should be discussed with the statistician at the early phase of product development. Under the Bayesian paradigm, these practical issues can easily be included and their impact studied via simulation.
- In the Bayesian paradigm, you are not penalized for looking at the data and can easily include external data that arise during a study. This makes Bayesian particularly useful for performing adaptive designs.
- Stopping rules have no consequences for a Bayesian, but make a world of difference for Frequentists.
- There is no silver bullet—increasing sample size is still the best way to improve a study.

Operating characteristics of Bayesian analysis: are you a closet Bayesian?

In the Frequentists paradigm, one often obtains the operating characteristics (OCs) of a design that includes simple summaries such as the average number of patients enrolled, false positives, and power without considering potential departures from the assumptions made when designing the trial.

Bayesian statisticians routinely perform evaluations under varied and realistic scenarios taking into account a multitude of practical aspects involved before a study begins. In general, Bayesians use simulation to better understand the decision-making process and to study the impact that deviations from assumptions will have on product development. Due to technical difficulties, the impact of deviations from model assumptions are rarely investigated under Frequentists methods even though departures can seriously degrade the properties of the design.

These are critical components contributing to the superior performance of Bayesian methods. While the gain in terms of OCs may not always be substantial, one gains a much clearer understanding of how the design will perform in practice. In addition, the resulting analysis and conclusions from a Bayesian analysis are much simpler to interpret and are often how the non-statistician interprets a Frequentist's results.

Subjectivity is pervasive in everyday life. And although scientific objectivity is crucial in separating science from intuition, that does not mean subjective opinion does not also arise in the scientific literature. Even if only the "planned" Frequentists analysis is presented, many other analyses have probably been done. It is important to note that there are real costs to waiting for more data—sometimes in lives lost or lives spared. We must also consider the value of rejecting the null hypothesis unless it is in favor of some alternative. Finally, full disclosure of what was done before, during, and after is important regardless of method.

> *"The human understanding, once it has adopted an opinion, collects any instances that confirm it, and though the contrary instances may be more numerous and more weighty, it either does not notice them or else rejects them, in order that this opinion will remain unshaken" – Francis Bacon, 1620.*

Kerry Barker and Ken Story are with the department of biostatistics at Baxter Healthcare in Round Lake, Ill., and J. Kyle Wathen is with the department of biostatistics at the University of Texas M.D. Anderson Cancer Center in Houston.

Suggested Reading

1. Chatfield C. Confessions of a pragmatic statistician. *Statistician* 2002; 51:1–20.
2. Demining WE. *Out of the Crisis*. Cambridge (UK): Cambridge University Press; 1994.
3. Efron B: Why isn't everyone a Bayesian? *Am Statistician* 1986; 40:1–7.
4. Nelder JA. Statistics, science and technology. *J R Statist Soc A.* 1986; 149:109–212.
5. Pocock S. Discussion on "Bayesian approaches to randomized trials." *J R Statist Soc A*. 1994; 157:338–390.
6. Savage SL. *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. Hoboken (NJ): Wiley; 2009.
7. Leach P. *Why Can't You Just Give Me the Number? An Executive's Guide to Using Probabilistic Thinking to Manage Risk and to Make Better Decisions*. Sugar Land (TX): Probabilistic Publishing; 2006.
8. Mlodinow L. *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Pantheon Books; 2008.
9. Radding A. Give me a number—introducing the DIST. Big Fat Finance Blog. http://bigfatfinance-blog.com/2009/08/04/give-me-a-number-introducing-the-dist. Posted August 4, 2009. Accessed January 25, 2010.
10. Thall PF, Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *Eur J Cancer* 2007; 43:859–866.