WHAT’S ALL THE FUSS OVER P-VALUES?

Bruce Levin, Ph.D.

Biostatistics in Action: Tips for Clinical Researchers
Fall 2017 Lecture Series

October 9th, 2017

My Experience

• Lots of grey hair—I’ve been at Columbia since 1974 (and Columbia College ‘68 before that).
• Professor of Biostatistics, MSPH
• Chair of Biostatistics (2000-2011—“Levin’s eleven”)
• A consulting biostatistician for numerous projects including as an expert statistical witness in litigation
• Director of the Biostatistics, Epidemiology and Research Design (BERD) Resource at the Irving Institute for Clinical and Translational Research

Outline of today’s talk

• The p-value is a time-honored and useful statistical tool...as far as it goes.
• Many statisticians have argued for many years that it doesn’t go far enough.
• The p-value answers one, but only one, question about research hypotheses.
• If you ignore this last point you can fall into serious logical traps!
• In particular, the p-value provides a flawed measure of the weight of evidence about research hypotheses.

DISCLAIMER

• The ideas I am presenting today are not new. It’s important to repeat them often, though, because, for a variety of reasons, they are often ignored.
• 1947—first known use in print of the term “P-value” (Merriam-Webster “Time Traveler” tool)

The content of these slides are intended for educational and not-for-profit use only. You are free to share and adapt, but must give appropriate credit to the author, Bruce Levin, PhD, who is affiliated with the Irving Institute for Clinical and Translational Research and the Department of Biostatistics at the Mailman School of Public Health, and indicate whether any changes were made to the source material.
FURTHER READING


WHAT IS THE P-VALUE?

- “A p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.”
  —ASA Statement on P-Values
- Also known as the *attained significance level* in the sense that if a hypothesis were being tested at that level, it would “just” be rejected.

EXAMPLE

“The Lady Tasting Tea”


<table>
<thead>
<tr>
<th>Actual preparation</th>
<th>Lady says: “Milk before Tea” “Tea before Milk”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk before Tea</td>
<td><img src="chart" alt="Chart showing milk preparation" /></td>
</tr>
<tr>
<td>Tea before Milk</td>
<td><img src="chart" alt="Chart showing tea preparation" /></td>
</tr>
</tbody>
</table>
There are 70 different ways the Lady (Dr. Muriel Bristol) can assign her determinations but only one can be all correct.

There are 70 different ways the Lady (Dr. Muriel Bristol) can assign her determinations but only one can be all correct. Therefore the $p$-value = 1/70 and we’re entitled to claim “$p < 0.05$.”

**PRINCIPLES** from the ASA Statement on Statistical Significance and $P$-Values

1. **$P$-values can indicate how incompatible the data are with a specified statistical model.**
   - More precisely, the $p$-value tells you the probability that data will deviate from expectation under an assumed model in hypothetical replications of an experiment by as much or more so than they have been observed to deviate from expectation in the actual experiment.
   - $P$-values are good at this but note that they assume the truth of the model (typically called the “null hypothesis”).

**Actual preparation**

<table>
<thead>
<tr>
<th>Lady says:</th>
<th>“Milk before Tea”</th>
<th>“Tea before Milk”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk before Tea</td>
<td>4 cups</td>
<td>0 cups</td>
</tr>
<tr>
<td>Tea before Milk</td>
<td>0 cups</td>
<td>4 cups</td>
</tr>
</tbody>
</table>

4 cups 4 cups
PRINCIPLES
from the ASA Statement on Statistical Significance and P-Values

2. \textit{P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.}

- \textbf{LOGICAL FALLACY #1:} The probability of A given B is the same as the probability of B given A.

\textbf{Example 2:} In the card game of poker, if the deck is shuffled thoroughly and fairly, the probability of being dealt a royal flush is about one in 2.6 million. You sit down at a card table and behold! You are dealt a royal flush ($p<0.0001$)!

\textbf{Question:} What is the probability the cards were shuffled thoroughly and fairly? Is it one in 2.6 million?

\textbf{NO!!!} The answer to \textit{THAT} question depends on who was dealing.

Your mother? Almost certainly the cards were shuffled thoroughly and fairly.

An infamously card sharp? Almost certainly the cards were \textit{not} shuffled fairly!!!

So while we can say that it is very unlikely to get such a valuable hand as a royal flush under the null hypothesis of a fair deal ($P[A|B] < 0.0001$), it does \textbf{not} follow that the probability of a fair deal given the observed royal flush, $P[B|A]$, is either large or small. \textit{That} assessment depends on extrinsic factors.
The same logic underlies the familiar fact that a screening device or a diagnostic laboratory test which has excellent sensitivity and specificity may nevertheless have poor predictive value in a population where the attribute of interest has low prevalence.

- Sensitivity = $P[+ | D \text{ in actuality}]$
- Specificity = $P[− | \text{no } D \text{ in actuality}]$
- Positive Predictive Value = $P[D \text{ in actuality} | +]$
- Negative Predictive Value = $P[\text{no } D \text{ in actuality} | −]$

$PPV = \frac{P[+ | D]}{P[+ | \text{no } D]} \times \frac{P[D]}{P[\text{no } D]} = LR \times \text{Prevalence odds}$

**Example 3(a):**
- If a person is a Martian, then he is not a member of Congress.
- This person is a member of Congress.
- Therefore he is not a Martian.
  
  (Correct logic and sensible conclusion.)

**Example 3(b):**
- If a person is an American, then he is not a member of Congress.
- This person is a member of Congress.
- Therefore he is not an American.
  
  The logic is still correct here, but the conclusion is nonsense because the initial premise is false. Can we modify the premise to make it reasonable?

**PRINCIPLES**

*from the ASA Statement on Statistical Significance and P-Values*

1. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

2. LOGICAL FALLACY #2: A correct logical syllogism stays approximately true if the word “probably” is inserted.

- If a person is an American, then he is probably not a member of Congress.
- This person is a member of Congress.
- Therefore he is probably not an American.

Here the initial premise becomes entirely reasonable by inserting “probably” but the logic is now wrong and the conclusion is nonsense.

But Example 3c is the hypothesis test fallacy:

- If the null hypothesis is true, then this result (statistical significance) would probably not occur.
- This result has occurred.
- Therefore the null hypothesis is probably not true.

PRINCIPLES from the ASA Statement on Statistical Significance and P-Values

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

- Doing so completely ignores a more nuanced assessment of whether a significant result is a type I error (error of commission) or whether an insignificant result is a type II error (error of omission). Does one lab result a diagnosis make?
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Therefore, p-values should ALWAYS be accompanied by point estimates indicating the size of the effect under study together with a confidence interval or some other assessment of the uncertainty or (un)reliability of the effect.

- Confidence intervals are based upon p-values...
- ...which shows that p-values are not inherently flawed or misleading as a tool.
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Example 4: Before you stand a large box containing balls numbered from $X=1$ to $X=N$. The null hypothesis is that $N=100$ and the alternative hypothesis is that $N=1,000$. You can’t see into the box; you are only allowed to pick one ball from the box at random without looking and observe $X$.

You reason that under the null hypothesis, the ball you pick will have a number $X$ averaging around 50, and that larger $X$ values would therefore provide evidence against the null hypothesis.

Furthermore, you are have a strong hunch that the box actually has 1000 balls, not 100; you would even give 4-to-1 odds that the box has 1000 balls.

You reach in and pull out the ball marked $X=100$. Under the null hypothesis, the probability of this extreme an outcome (above average) is 1 in 100. So you say, “The $p$-value is 0.01. This is strong evidence against the null hypothesis.”
EXAMPLE 4

• But wait a minute. That $p$-value didn’t pay any attention to the likelihood of drawing 100 under the alternative hypothesis. Are you really so sure the weight of evidence is against the null hypothesis?

EXAMPLE 4

• Under the alternative hypothesis, the probability of drawing any given $X$ is 0.001. Now if we had observed $X > 100$, the evidence would have been overwhelming in favor of $H_1$. As it stands, though, drawing $X=100$ is 10 times more likely to occur under $H_0$ than under $H_1$.

EXAMPLE 4

• This is actually moderate to strong evidence against the alternative hypothesis and in favor of the null !!!

• In fact, after observing $X=100$, fair betting odds would actually be 5-to-2 in favor of $N=100$ !!!

EXAMPLE 4

PRINCIPLES

from the ASA Statement on Statistical Significance and $P$-Values

6. By itself, a $p$-value does not provide a good measure of evidence regarding a model or hypothesis.

• Proper assessment of the weight of evidence in favor of a hypothesis must be considered relative to other hypotheses.
• The likelihood ratio (LR) is arguably the best measure of weight of evidence:
  \[ \text{Likelihood ratio} = \frac{\text{Probability of data under } H_1}{\text{Probability of data under } H_0} \]

• Betting odds:
  \[ \frac{P(H_1 \text{ after data})}{P(H_0 \text{ after data})} = LR \times \frac{P(H_1 \text{ before data})}{P(H_0 \text{ before data})} \]

• In Example 4, betting odds after seeing X=100 are \((1/10) \times 4 = 2/5\) in favor of \(H_1\) or \(5:2\) against.

---

**CONCLUDING REMARKS**

• A p-value should be the start of an investigation, not the end of it.

• Null hypothesis significance testing is sterile unless it is accompanied by estimates of effect sizes based on a scientific theory of what causes such effects together with an honest assessment of variability, uncertainty, error, bias, and reliability or the lack thereof.

---

**PRINCIPLES**

from the ASA Statement on Statistical Significance and P-Values

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

• In large samples a result with \(p=0.05\) actually represents at best only weak to moderate evidence against the null hypothesis (LR=6.8).

• When we assess the evidence as strong (LR>20), however, the probability we will be misled is usually very small (less than 1%).

---

**CONCLUDING REMARKS**

• \(P\)-values answer one question: how likely is it to observe such data assuming the null hypothesis is true? Other important questions include:

  • How large an effect do I have and how reliable is it? [Consistent estimates of effect, standard errors, sources of bias, other systematic errors.]

  • What is the strength of evidence in favor of or against the null hypothesis vis-à-vis a scientific alternative? [Use the LR, not \(P\)-value.]
CONCLUDING REMARKS

- What should I believe about the truth of my theory? [Betting odds, Bayesian ideas]
- What should I do next? [Decision theory—actions, consequences, losses, risks, policies]

THANK YOU VERY MUCH.