P value의 의미, 얼마나 절대적인가?

박 광 열
중앙의대

What is a *p* Value?

Kwang-Yeol Park, MD, PhD
Department of Neurology, Chung-Ang University, Seoul, Korea

What is a *p* value?

What is a *p* value?

For example, suppose that a drug study produced a *p* value of 0.001.

This *p* value indicates that

- If you reject the null hypothesis, there's a 0.1% chance that you're making a mistake.
- Assuming that the drug had no effect, you'd obtain the observed difference or more in 0.1% of studies due to random sampling error.
- Both explanations are true

2/26

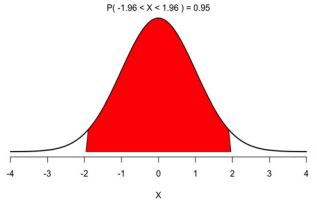
ASA Statement on Statistical Significance and P-values

1. P-values can indicate how incompatible the data are with a specified statistical model.

- A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data.
- **The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold.**
- This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

4/26

Null Hypothesis Significance Testing



5/26

2. 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.

- Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data.
- The p-value is neither.
- It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

6/26

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

- A conclusion does not immediately become "true" on one side of the divide and "false" on the other.
- Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.

7/26

4. Proper inference requires full reporting and transparency.

- P-values and related analyses should not be reported selectively.
- Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided.
- Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

8/26

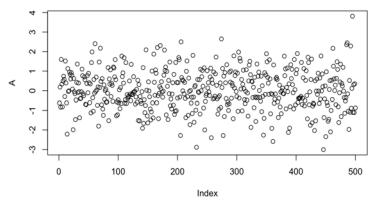
Multiple testing

```
set.seed(1)
A <- rnorm(500, 0, 1)
plot(A)

N <- 1000
pvals <- replicate(N, {
  control <- sample(A, 50, replace = T)
  treatment <- sample(A, 50, replace = T)
  t.test(control, treatment)$p.value
})
```

9/26

Multiple testing



10/26

Multiple testing

```
length(pvals[pvals < 0.05])

## [1] 54

length(pvals[pvals < 0.05/10])

## [1] 6

length(pvals[pvals < 0.05/100])

## [1] 0

length(pvals[pvals < 0.05/1000])

## [1] 0
```

11/26

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

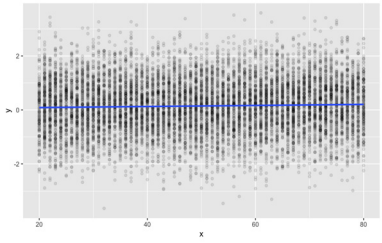
- Statistical significance is not equivalent to scientific, human, or economic significance.
- Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p- values do not imply a lack of importance or even lack of effect.

12/26

```
x <- sample(20:80, 10000, replace = T)
y <- 0.003*x + rnorm(10000, mean = 0, sd = 1)
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7478 -0.6714  0.0061  0.6748  3.4267
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.048686   0.029989   1.623 0.104524
## x           0.001953   0.000568   3.439 0.000586 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9941 on 9998 degrees of freedom
## Multiple R-squared:  0.001181,   Adjusted R-squared:  0.001082
## F-statistic: 11.83 on 1 and 9998 DF,  p-value: 0.0005865
```

13/26



14/26

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

- Researchers should recognize that a p-value without context or other evidence provides limited information.
- For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis.

15/26

Other approaches

- In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches.
- These include methods that emphasize estimation over testing, such as **confidence, credibility, or prediction intervals**; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and **false discovery rates**.

17/26

False discovery rate, or how not to make a fool of yourself with P values

http://www.dcscience.net/2014/03/24/on-the-hazards-of-significance-testing-part-2-the-false-discovery-rate-or-how-not-to-make-a-fool-of-yourself-with-p-values/

If you observe a P value close to 0.05, your false discovery rate will not be 5%. It will be at least 30% and it could easily be 80% for small studies.

18/26

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Review

Cite this article: Colquhoun D. 2014 An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. open sci.* **1**: 140216.
http://dx.doi.org/10.1098/rsos.140216

An investigation of the false discovery rate and the misinterpretation of *p*-values

David Colquhoun

Department of Neuroscience, Physiology and Pharmacology, University College London, Gower Street, London WC1 6BT, UK

19/26

What is the probability of diseases?

- Test
  - Sensitivity: 80%
  - Specificity: 95%
- If the test was positive, what is the probability of disease?

20/26

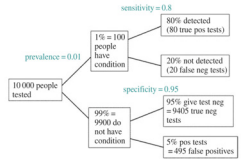


Figure 1. Tree diagram to illustrate the false discovery rate in screening tests. This example is for a prevalence of 1%, specificity 95% and sensitivity 80%. Out of 10 000 people screened, 495 + 80 = 575 give positive tests. Of these, 495 are false positives so the false discovery rate is 86%.

21/26

What is the probability of the true effect?

- If the significance level is 0.05 and power is 80%, what is probability of the true effect?

22/26

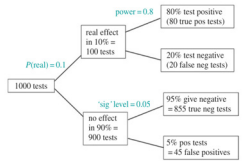
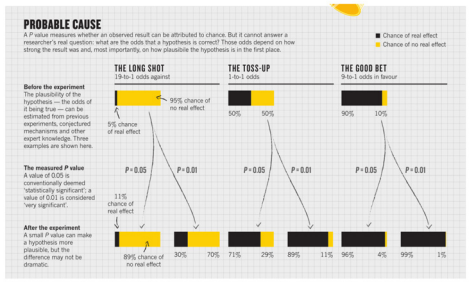


Figure 2. Tree diagram to illustrate the false discovery rate in significance tests. This example considers 1000 tests, in which the prevalence of real effects is 10%. The lower limb shows that with the conventional significance level, $p = 0.05$, there will be 45 false positives. The upper limb shows that there will be 80 true positive tests. The false discovery rate is therefore $45/(45 + 80) = 36\%$, far bigger than 5%.

23/26



24/26

Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.

No single index should substitute for scientific reasoning.

25/26

Take-Home Message

- If the test has a p-value below 0.05, we say the data supports the alternative hypothesis at the 0.05 significance level.
- If the p-value is above 0.05, it merely indicates that "there is not enough evidence to reject the null hypothesis." That lack of evidence does not prove that the null hypothesis is true.

26/26