Elementary Statistics: A History of Controversy

Joseph Manthey
Saint Joseph College
West Hartford, Connecticut
jmanthey@sjc.edu

I. Eugenics

"No degenerate and feeble stock will ever be converted into healthy and sound stock by the accumulated effects of education, good laws, and sanitary surroundings. Such means may render the individual members of the stock passable if not strong members of society, but the same process will have to be gone through again and again with their offspring, and this in ever-widening circles, if the stock, owing to the conditions in which society has placed it, is able to increase in numbers. The suspension of that process of natural selection which in an earlier struggle for existence crushed out feeble and degenerate stocks, may be a real danger to society, if society relies solely on changed environment for converting its inherited bad into an inheritable good."

Karl Pearson, The Grammar of Science (1892).

"From the moment that we grasp, firmly and completely, Darwin's theory of evolution, we begin to realize that we have obtained not merely a description of the past, or an explanation of the present, but a veritable key of the future … . The socially lower classes have a birth-rate, or, to speak more exactly, a survival rate, greatly in excess of those who are, on the whole, distinctly their eugenic superiors. It is to investigate the cause and cure of this phenomenon that the eugenic society should devote its best efforts."

Ronald Fisher, "Some hopes of a eugenist", Eugenics Review, v. 5, p. 309-315 (1914)

Recommendations

Read the primary sources. They are not as whitewashed as the current generation of textbooks and are by far a more interesting read.

Additional resources

1. Pearson, K. (1900). The Grammar of Science. London, Adam and Charles Black.
   An extremely influential and wide ranging book covering topics such as the philosophy of science, probability, space and time, motion, matter, life and evolution. Pearson's views on eugenics are confined to a couple of chapters.
2. Pearson, K. (1905). National Life from the Standpoint of Science London, Adam and Charles Black.
   A much more focused book on the policy implications of science, especially genetics. Pearson argues forcefully for policies which "check the fertility of inferior stock" and "encourage the fertility of the fitter stock".
3. Fisher, R. A. (1914). Some hopes of a eugenist", Eugenics Review, v. 5, p. 309-315.
   Fisher expresses concern that the lower classes have a higher birth rate than their "eugenic superiors" and that this is a trend that has led to the collapse of civilizations.
4. Fisher, R. A. (1924). The elimination of mental defect. Eugenics Review, 16, 114-116, 1924.
   The Hardy-Weinberg principle predicts that defective genes are very persistent in a population even if they are rare. This suggests that it would be difficult to remove defective genes from a population. Fisher argues that this is misleading and that sterilization is more effective than previously thought.

II. Two incompatible approaches to inference

Significance testing has been controversial since its inception and the source of much of this disagreement can be traced through time to the "fathers" of modern statistics including Ronald Fisher, Jerzy Neyman and Egon Pearson. Most modern elementary statistics blend elements of these incompatible approaches. Fisher

Inductive evidence (Fisher)

- **Purpose:** Measure the strength of the evidence against a single hypothesis.
- **Elements:** A single hypothesis H_0. A test statistic and its distribution under the assumption that the hypothesis is true.
- **Result:** An index called the *p*-value = P (D | H_0), interpreted as a measure of the evidence against the hypothesis. If the *p*-value is less than 0.05, the evidence is considered significant.
- **Scope:** Useful in the context of evaluating scientific hypotheses where it is not necessary or even desirable to make decisions on the basis of a single observation.

Decision making (Neyman-Pearson)

- **Purpose:** To select one of two hypotheses on the basis of an observation.
- **Elements:** Two hypotheses, the null H_0 and alternative Ha. The distribution of the test statistic under the assumption the null hypothesis is true. Type I error (false rejection) rate α and Type II error (false acceptance) rate β selected in advance of the test (usually $\alpha = 0.05$ and $\beta < 0.20$. Rejection region whose size depends on α. Statistical power $1 - \beta$.
- **Result:** A decision.
- **Scope:** Useful in quality control situations where an immediate decision must be made.

Differences between *p*-values and α-levels

Here is a table which summarizes some of the differences between Fisher's p-value and the Neyman/Pearson Type I error rate α.

<u>*P*-values</u>

- Evidence against H0
- Inductive inference
- Data based random variable
- Short term measurement applies to any experiment

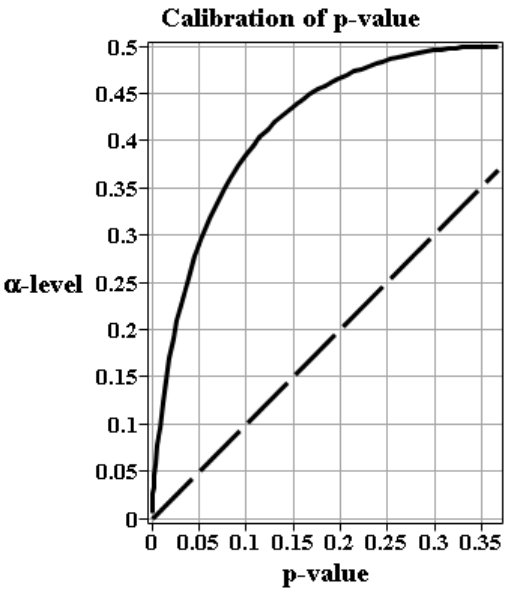
<u>α-levels</u>

- Type I error rate
- Inductive behavior
- Fixed in advance
- Long term error rate that applies to many repetitions of an experiment

In addition to the philosophical differences, there are numerical differences between p and α. To see this use the Java Applet (http://www.stat.duke.edu/~berger/p-values.html) and consider the lower bound on the Type I error (α) derived by Berger (2001)

$$\alpha(p) = \left(1 + \left[-e\, p \log(p)\right]^{-1}\right)^{-1}, \; p < 1/e$$

Here is a plot of the α-level as a function of p. The α-level is the solid curve and the p-value is dashed and shown only for comparison purposes.



Recommendations

1. Read the primary literature and some of the more recent literature on the distinction between the Fisher and Neyman-Pearson approaches to statistical inference.
2. Become familiar with the fundamentals of the major schools of statistical inference (frequentist and Bayesian).
3. The current recommendations of important professional organizations are nudging us in the direction of Fisher. Here is a quotation from the Statistical Methods in Psychology Journals: Guidelines and Explanations.

   "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p-value or, better still, a confidence interval.

Additional resources

1. Neyman, J., Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society of London, v. 231, pp. 289-337.
   Neyman and Pearson argue that one should select the test which minimizes the Type II error subject to a bound on the Type I error.
2. Lehmann, E.L. (1993). "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?", Journal of the American Statistical Association, Vol. 88, No. 424, pp. 1242-1249.
   An attempt to reconcile the Fisher and Neyman Pearson approaches to inference.
3. Goodman S (1999). "Toward evidence-based medical statistics. 1: The *P* value fallacy.". Ann Intern Med 130 (12): 995–1004.
   An explanation of the differences between the Fisher and Neyman/Pearson approaches to statistical inference which addresses the implications for medical research.
4. Goodman S (1999). "Toward evidence-based medical statistics. 2: The Bayes factor.". Ann Intern Med 130 (12): 1005–13.
   A follow on to the previous paper, this paper present the Bayes Factor as an alternative to p-values. Bayesian methods have been making their inroads into medical research and in the future, it is likely that elementary statistics textbooks will reflect this shift.
5. Sellke, T., Bayarri,M . J., and Berger,J . 0. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," The American Statistician, 55, 62-71.

Explains the differences between *p*-values and α-levels and shows how to estimate the α-levels from the *p*-values using simulations and via $\alpha(p) = \left(1 + \left[-e\, p \log(p)\right]^{-1}\right)^{-1}$, $p < 1/e$.

6. Hubbard R. & Bayarri M.J., (2003). "Confusion Over Measures of Evidence (ps) Versus Errors (alphas) in Classical Statistical Testing," The American Statistician, American Statistical Association, vol. 57, pages 171-178.
   This paper explains the difference between Fisher's evidential *p*-value and the Neyman-Pearson Type I error rate α.

III. The problem of size

P-values do not reveal the size of an effect and are not useful for determining the practical significance of a finding.

Example 1

Suppose we administer an IQ test to 1,000 randomly selected males and 1,000 randomly selected females. The males are found to have an average IQ of 98 with a standard deviation of 15. The females are found to have an average IQ of 100 with a standard deviation of 15. Conduct an independent samples *t*-test.

Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-2	0.6708204	1998	-2.9814239	0.0029

As seen in the computer output, the low *p*-value of 0.0029 indicates a statistically significant difference. However, the difference between an IQ of 98 and 100 would not have any practical significance.

Example 2

Suppose we administer an IQ test to 5 randomly selected males and 5 randomly selected females. The males are found to have an average IQ of 95 with a standard deviation of 15. The females are found to have an average IQ of 110 with a standard deviation of 15. Conduct an independent samples *t*-test.

Difference	Sample Mean	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-15	9.486833	8	-1.5811388	0.1525

As seen in the computer output, the relatively high *p*-value of 0.1525 indicates that there is not a statistically significant difference. However, the difference between an IQ of 95 and 110 would be very significant from a practical perspective.

Effect sizes

The examples illustrate that statistical significance and practical significance are not the same. It is possible to have statistical significance without practical significance. It is also possible to have practical significance without statistical significance. The fundamental problem can be seen by studying the form of the test statistic. For example, consider the test statistic for an independent sample *t*-test.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the sample sizes are large, the denominator of the test statistic becomes small, leading to a large magnitude for the test statistic and small *p*-value. This can lead to statistical significance, even if there is no practical difference between the two sample means. Jacob Cohen introduced an index which does not suffer from this problem, now called Cohen's *d*.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

where σ was understood to be the standard deviation of either population, since they are assumed to be equal. Other authors have made the calculation of the standard deviation more explicit by providing the following formula for the spooled standard deviation

$$\sigma = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2}}$$

The denominator in this case does not become small when the sample size increases. Cohen describes an effect size of 0.2 as small, an effect size of 0.5 as medium and an effect size of 0.8 as large. Of course these terms are subjective and one needs to pay attention to the context when interpreting effect sizes.

Example 3

The effect sizes for Examples 1 and 2 are -0.133 and -1.12. Clearly, the effect size is an effective tool for establishing practical significance.

Recommendations

Here is a quotation from the Statistical Methods in Psychology Journals: Guidelines and Explanations.

> Always provide some effect size estimate when reporting a *p*-value. Cohen (1994) has written on this subject in this journal. All psychologists would benefit from reading his insightful article. Always present effect sizes for primary outcomes. It helps to add brief comments that place these effect sizes in a practical and theoretical context.

I recommend making a strong distinction between statistical significance and practical significance. Make a habit of asking students to address both the statistical and practical significance after every significance test. Students are open to the concept that size matters. For example, consider a weight loss medication study that results in a low p-value but an average weight loss of 3 pounds. This medication is not the answer to the nation's obesity epidemic.

Additional resources

1. Carver, R. P. (1978) "The Case Against Statistical Significance Testing." *Harvard Educational Review*, 48(3): 378-398.
   This is one of many papers discussing the limitations of significance testing.

2. Wilkinson L. and the Task Force on Statistical Inference (APA Board of Scientific Affairs). (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations, American Psychologist, Vol. 54, No. 8, 594-604.
   Explores some of the controversies surrounding the application of statistical methods including significance testing, provides recommendations and considers some alternatives made possible by the increases in computing power. These recommendations are influential since many disciples adhere to APA guidelines.
3. Kain, Z.N. (2005), "The legend of the P value", Anesthesia and Analgesia, Nov; 101(5): 1454-6.
   A paper emphasizing the importance of explaining the clinical significance of results published in medical journals.
4. Ziliak, S.T. and Deirdre N. McCloskey, D.N. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives, University of Michigan Press (2008).
   A book filled with examples illustrating the difference between statistical and practical significance.