July 2003: Very non-significant *P*-values are very significant (New Rule, 1.17).

[*At times I encounter information that suggests a useful new rule—evidence that not all the rules have been covered in the book. I will number such new rules according to the chapter in which the rule fits best. So far I have not found rules for which I would create a new chapter, but that possibility is not excluded either, of course*.]

Introduction

It is customary to draw conclusions, take actions, or make decisions when *P*-values are very small, that is "significant." There is another use of such values: when they are close to one.

Rule of Thumb

When a *P*-value is greater than 0.95, or so, examine the data and the model on which the *P*-value is based.

Illustration

Fisher analysis of Mendel's data. Fisher (1936) reanalyzed data by Mendel published in 1866. Fisher applied chi-square tests to separate, independent sets of Mendel's data and then summed the chi-squares. The observed value of $X^2 = 41.6056$ with 84 degrees of freedom has a *P*-value of 0.99993 (Table V in Fisher, 1936). Fisher considered the fit too good, and trying to explain it, writes, "Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected." In other words, the unusually good fit of the data to the theory was inconsistent with a randomly generated set of data. In this case Fisher did not question the model but the data.

Basis of the rule

Every statistical analysis is based on a model of the data in terms of structure and variability. The validity of tests of significance (and hence *P*-values) depends on the validity of data and the model used to summarize the data. Violations of the assumptions may invalidate one or more aspects of the analysis.

Discussion and Extensions

The issue can be illustrated by coin tossing. The probability of getting exactly *n* heads in 2*n* tosses of a coin becomes smaller and smaller as *n* becomes larger. For example, the probability of exactly half heads in two tosses is 1/2, in four tosses is 3/8, and in ten tosses is 36/1024. Hence if someone claimed to have tossed a coin 500 times and gotten exactly 250 heads (resulting in a X^2 of zero and a *P*-value of 1!) you would question the claim that this was a randomly generated sequence of outcomes. In fact, outcomes in some small interval around 250 heads would lead to suspicion.

Fisher's analysis continues to generate discussion. One ironic twist is that the model on which he based his analysis has been questioned. For example Seidenfeld (1998) writes, "Mendel's experiments include some important sequential design features that Fisher (and others to my knowledge) ignore." Additional reasons posited by Seidenfeld include that the form of Fisher's analysis is an example of meta-analysis and subject to current concerns and criticisms, and an attempt to modify the Mendelian model. Pilgrim (1984, 1986) asserted that Fisher was wrong and drew some pointed criticism from A.W. F. Edwards (1986). Snedecor and Cochran (1989) write, "Thus, the agreement of the results with Mendel's law looks too good to be true." All this indicates that the issue has not been resolved satisfactorily. To find the current status of the discussion type in "R.A. Fisher and Mendel's data" into the Google search engine and many references will pop up.

The same point about "too good" P-values is made in Christensen (2003). In fact, his article prompted this month's rule. He discusses under what circumstances, in the general linear model, a very small *F* statistic will be observed—leading to a *P*-value close to one. He demonstrates at least three scenarios where this may occur:

1. Carrying out an incorrect analysis of variance. For example, analyzing a randomized block design as if it were completely randomized. The block variability goes into the error term, and if there is no treatment effect, the treatment mean square estimates the pure error term. This will lead to a small *F* statistic.
2. The data are negatively correlated.
3. There is heterogeneity of variance. For example, in the one-way analysis of variance, the variances within treatments are not homogeneous and more observations are taken within the treatments with the larger variances.

In all these instances, and the one cited in the illustration, there are violations of the assumptions on which the analysis is based. Hence the rule of thumb that the observations of *P*-values close to one should spur you to examine the data and the model used for analyzing the data.

References

Christensen, R. (2003). Significantly insignificant *F* tests. *The American Statistician*, **57:** 27-32.

Pilgrim, I. (1984). A solution to the too-good-to-be-true paradox and Gregor Mendel. *The Journal of Heredity*, **75:** 501-502. Comments by A.W.F. Edwards, **77:** 138; response by Pilgrim, **77:** 138. On the web at http://www.mcn.org/c/irapilgrim/. Accessed July 23, 2003.

Seidenfeld, T. (1998). P’s in a pod: some recipes for cooking Mendel’s data. http://philsci-archive.pitt.edu/archive/0000156/ Accessed July 23, 2003.

Snedecor, G.W. and Cochran, W.G. (1989). *Statistical Methods*, eight edition. Iowa State University Press, Ames, IA. Page 210.