STATISTICAL SIGNIFICANCE AND PRACTICAL SIGNIFICANCE IN STATISTICS EDUCATION

KUMAR, Pranesh

Department of Mathematics and Statistics, University of Northern British Columbia, Prince George, BC, Canada

OBJECTIVE

- **Statistical Significance vs practical significance.**
- **Does the sample provide good evidence against a claim?**

BACKGROUND

Statistics null hypothesis testing (SNHT) indicates whether there is any evidence in favour of research hypothesis or not.

Statistical significance is measured p-value generated by conducting the statistical test of the null hypothesis.

Several interpretations of p-values are possible like

the probability that the results obtained were due to chance.

A small p- value would suggest that the observed mean difference was not due to chance and therefore, could be assumed significantly different.

p-value is affected by sample size and sometime can be made small by taking larger samples.

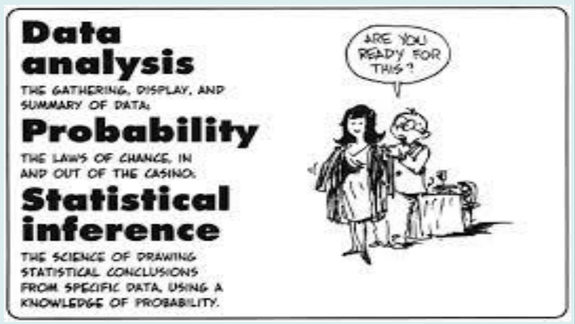
Practical significance is measured by effect size

Effect size is about the extent to which the research hypothesis is true or to the degree to which findings have practical significance in context of the study population.

Effect size quantifies the degree to which the study results should be considered negligible or important regardless of the size of the study sample.

Effect size has advantages over statistical significance testing because they are independent of the sample size and are scale-free.

Effect size measures can be uniquely interpreted in different studies regardless of the sample size and the original scales of the variables.



STATISTICAL SIGNIFICANCE

- In SNHT, the claim being tested is a null hypothesis H_0.
- A statistical test is conducted to assess the strength of the evidence against H_0.
- This test generates a P-value which is the probability, computed assuming H_0 true, that the sample outcome would be as extreme or more extreme than one actually observed.
- To make a decision, we decide how much evidence against H_0 we will accept. That means, how small a p-value we require. The decisive value of p is known as the significance level α. If the p-value is as small as or smaller than α, data are statistically significant at level α.
- It may be noted that significant in statistical sense does not imply important. It means: Not likely to happen just by chance.
- Several interpretations of p-values are possible.
- Sometimes p-value is viewed as the probability that the results were obtained due to chance.
- Another interpretation of the probability 1-p is that it signifies the reliability of the results and is the probability of getting the same result if the experiment was repeated.
- Significant differences are often termed as reliable under this interpretation.
- The p-value can also be interpreted as the probability that the null hypothesis is true.
- These interpretations are termed as fantasies about the statistical significance.
- None of them is actually true however they are treated as if they were true.
- Most H_0 under significance tests state that some parameter equals zero or some set of parameters are all equal.
- These hypotheses are almost invariably known to be "false" before any data are collected
- If such types of null hypotheses are not rejected, it is often because the sample size is too small
- What remains of interest is whether or not the sample size considered will be sufficient to detect the difference.
- If the null hypothesis truly is false, p-value can be made as small as one wish by selecting a large enough sample.

PRACTICAL SIGNIFICANCE: EFFECT SIZE

- Questions which interest practitioners:
  - What the magnitudes of sample effects are?
  - Whether these results will generalize?
- Statistical significance testing does not respond to such questions.
- Effect size quantifies the size of the difference between two groups.
- Effect size emphasizes the size of the difference rather than confounding this effect with sample size
- The statistical significance measured by p-value is the probability that a difference of at least the same size would have arisen by chance, even if there really were no difference between two populations.
- However statistical significance combines the effect size and sample size.
- The major concern in using statistical significance testing is that the P-value depends essentially on the effect size and the size of the sample.
- One may infer significant difference either if the actual effects were very large despite having only small samples, or if the samples were very large even if the actual effect sizes were small.
- We cannot ignore the statistical significance of a result since without it we may infer firm conclusions from studies where the samples are too small to justify such confidence.
- Effect size is defined as the standardized mean difference between two groups.
- Another feature of the effect size is that it can be directly converted into statements about the overlap between the two samples in terms of a comparison of percentiles.
- Another way to interpret effect size is to compare them to the effect sizes of differences that are familiar. For example, Cohen (1969) describes an effect size of 0.2 as small, an effect size of 0.5 is described as medium and an effect size of 0.8 as grossly perceptible and therefore, large.
- Margin of error in estimating effect sizes: Estimate using the confidence interval which provides the same information as is usually contained in a significance test. For example, a 95% confidence interval is equivalent to choosing a 5% significance level.

CONCLUDING REMARKS

- Use of statistical significance testing in scientific studies is debated.
- Statistical hypothesis testing tool is overused, misused and often inappropriate.
- Effect size can be considered as a metric of the extent to which the research hypothesis is true or to the degree to which the findings have practical significance in context of the study population.
- Effect size quantifies the degree to which the study results should be considered negligible or important regardless of the size of the study sample.
- Effect size measures can be uniquely interpreted in different studies regardless of the sample size and the original scales of the variables.

References

- Berger, J. 0. and Berry, D. A., Statistical analysis and illusion of objectivity, American Scientist, 76: 159-165, 1988.
- Berger, J. O. and Selke, T. , Testing a point null hypothesis: the irreconcilability of P values and Evidence, Journal of the American Statistical Association, 82:112-122, 1987.
- Carver, R.P., The case against statistical significance testing, Harvard Educational Review, 48: 378-399, 1978.
- Clark, C. A., Hypothesis testing in relation to statistical methodology, Review of Educational Research 33: 455-473,1963.
- Cohen, J., Statistical Power Analysis for the Behavioral Sciences, NY: Academic Press, 1969.
- Coe, R., It's the Effect Size, Stupid: What effect size is and why it is important, Annual conference of the British Educational Research Association, University of Exeter, England, 12-14, 2002.
- Johnson, D.H., The insignificance of statistical significance testing, Journal of Wildlife Management 63(3):763-772, 1999.
- Thompson, B., Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. Annual Meeting of the American Educational Research Association, Montreal, 1999.

2013 Joint IASE / IAOS Satellite Conference

Statistics Education for Progress, Macao, China, 22-24 August 2013