Asian Journal of Psychiatry 2 (2009) 74–79

Contents lists available at ScienceDirect

Asian Journal of Psychiatry

journal homepage: www.elsevier.com/locate/ajp

Discussion

# Statistical versus clinical significance in psychiatric research—An overview for beginners

Jagadisha Thirthalli *, Ravi Philip Rajkumar

*Department of Psychiatry, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore 560029, India*

ARTICLE INFO

*Article history:*
Received 28 February 2009
Received in revised form 19 March 2009
Accepted 9 April 2009

*Keywords:*
Statistical significance
Clinical significance
Odds ratio
Effect size
Number needed to treat
Confidence intervals

## 1. Introduction

This article is intended to be a practical guide for psychiatric trainees and clinicians, who are often unfamiliar with the statistics reported in research papers. We will discuss two apparently opposing concepts, statistical significance and clinical significance. We will first introduce the philosophy of inferential statistics in biomedical research. The concept of *statistical significance* is central to this philosophy. As with any other concept, the notion of statistical significance has its own limitations. Methods to overcome these problems have assumed prominence in the last two decades. These deal with the concept of *clinical significance*. We will discuss them in the next sections. When teaching medical students in general, and psychiatry students in particular, a common complaint is that statistics are hard to understand. This is partly because statistical concepts are often taught using examples that may not have much relevance to their clinical practice. In this article, illustrations from published studies in psychiatry have been used whenever possible, as such examples are more relevant to students and improve their ease of understanding statistical concepts.

## 2. Statistical significance

Tests of statistical significance measure the probability that a given finding is due to chance. For example, a research paper may report that drug A is "significantly more effective" than drug B, and reports a probability value, or *p*-value, of 0.01. How should a clinician reading the paper interpret this value? Research in the biomedical field involves examining data from a sample and making inferences about the population, which the sample is expected to represent. This is known as *inferential statistics*. The value of a variable observed in an experiment is thus considered the best *estimate* of the true value of that variable in the population. Suppose, in a study, the mean (SD) weight gains in subjects who received antipsychotics A ($n = 41$) and B ($n = 41$) were 1.9 (2.9) kg and 0.5 (2.4) kg, respectively. There is a greater weight gain (of 1.4 kg) in subjects treated with antipsychotic A. This figure is actually an *estimate* of the actual difference in the population, if the entire population were treated with these two antipsychotics. However, the actual difference in the population might be different from this estimated value. When samples are examined, the values obtained are likely to be different from the population value. These differences are due to sampling variation. It is possible that the actual difference in the population is 0 (i.e., there is no difference in weight gain among individuals using antipsychotic A or B) and a value of 1.4 kg is obtained because the researcher has examined a sample and not the population. Tests used in inferential statistics would help us in obtaining a probability that a value as large as

* Corresponding author. Tel.: +91 80 26995350; fax: +91 80 26564830/2121.
*E-mail addresses:* jagatth@yahoo.com (J. Thirthalli), ravi.psych@gmail.com (R.P. Rajkumar).

1876-2018/$ – see front matter 
doi:10.1016/j.ajp.2009.04.004

1.4 kg or more is obtainable because of sampling error even when the population difference is actually 0.

Inferential statistics involves testing hypotheses. In this case, the hypothesis might be "antipsychotic A causes more weight gain than B". In other words, the hypothesis is that the difference in mean weight gain in the population is more than 0. If there is a high probability that a difference as large as 1.4 kg or more could be obtained because of sampling error, when the actual difference in the population is 0, then it is reasonable to infer that there is no difference in weight gain between drugs A and B. Conversely, if this probability is very low, then it is reasonable to believe that the population value for the difference in mean weight gain is greater than 0. What probability value should be considered as a cut-off to be reasonably sure that the population value is greater than 0? Conventionally, a probability of 1 in 20 (0.05) is taken as an arbitrary cut-off. If the probability value given by the statistical test is greater than 0.05, then one cannot believe *beyond reasonable doubt* that the population value is greater than 0. In such a case, the difference of 1.4 kg obtained in the sample could be considered statistically not significant. Conversely, if the probability is less than 0.05, then it could be considered statistically significant.

Suppose that the weight gains associated with antipsychotics A and B are normally distributed. If two random samples of any size ($n$) chosen from populations receiving antipsychotics A and B are examined, the mean weight gains in each sample and the difference between these means can be calculated. As discussed earlier, this difference is the *sample* difference of means. Hypothetically, if samples of the same size are chosen repeatedly and the sample differences of means are plotted, then it can be shown that they would form a normal distribution. The mean of this distribution will be equal to the population mean. The properties of a normal distribution will provide the probability of obtaining a difference equal to or greater than 1.4 kg, even when the actual difference in the population is 0. Typically we use Student's $t$-test to decide this. The $t$-test gives the value of $t$ according to the formula shown in Box 1. If $t$ is large, it means that it is highly improbable (i.e., probability is less than 1 in 20 or 0.05) that one would obtain a difference equal to or greater than 1.4 kg even when the population difference is 0. This allows us to decide with reasonable confidence that the population difference is "greater than 0". In other words, there is a statistically significant difference in mean weight gain between antipsychotics A and B.

### 2.1. The meaning of statistical significance

A few points are worth noting here: (1) "statistically significant difference" simply means that the difference of means is likely to be greater than 0; it does not indicate the magnitude of the difference. (2) As can be seen from the formula, the value of $t$ depends on three factors: the magnitude of the difference of means, the standard deviations of the samples and the number of subjects studied. We will discuss the differences of means and standard deviations at greater length later. Considering the number of subjects, it is evident from the formula that for any given difference of means and standard deviation, the value of $t$ increases as the sample size increases. In other words, for the same magnitude of difference, the value of $t$ would be greater if a greater number of subjects were included. Thus, two studies done on the same population may show identical differences in means, but the results may be statistically significant or insignificant depending on the number of subjects studied.

Consider the example of a hypothetical intervention that aims to improve children's IQ. Suppose a population of children has a mean IQ of 100 with a standard deviation of 17.5. An intervention is introduced to improve their IQ. It would be prescribed to all if an experiment shows that the claim is true, i.e., if the IQ of the children who undergo the intervention is *statistically significantly* better than those who do not undergo it. Note from the formula in Box 1 that a large value of '$t$' (indicating statistically significant superiority of the intervention) can be obtained if the intervention produces a large increase in mean IQ. Alternatively, even a small increase can result in a large '$t$' value if the study includes a large number of subjects. Suppose 6 students undergo the intervention and 6 do not. Then, it can be calculated that the intervention will be considered statistically significant if the intervention produces at least a 22.51-point increase in the mean IQ (assuming a constant SD of 17.5). Similarly, if 15 children are studied in each group, the intervention should produce a 13.09-point increase in the mean IQ. Table 1 shows the different cut-offs for the sample difference of means for the intervention to be considered statistically significant with different sample sizes. Thus, if 1500 children are studied in each group, the intervention would be considered to significantly increase the IQ even if it produces a change as small as 1.25 IQ point.

This example illustrates the limitation of relying only on statistical significance in making clinical decisions. Statistical tests in inferential statistics are, in general, designed to answer the question "how likely is the difference found in a sample due to chance (when actually no such difference exists in the population, the null-hypothesis)?". This is the only purpose they serve—the calculation of a probability value.

### 2.2. Why 0.05?

If a result is statistically significant, then it is highly unlikely to be due to chance, and is therefore likely to be replicable. Statistical tests serve a very important purpose in biomedical research, where any hypothesis is considered to be false unless proved otherwise beyond reasonable doubt. By convention, a probability value of less than 0.05 is considered to prove a point beyond reasonable doubt. However, as the above example illustrates, this may not be terribly useful for clinical purposes. Further, it should be noted that Ronald Fisher, who was the first to introduce the cut-off of 0.05, chose its

**Box 1.** Formula for calculating the $t$ value*.

$$t = \frac{\overline{m_1} - \overline{m_2}}{\sqrt{(s_1^2 + s_2^2)/N}}$$

$\overline{m_1}$ = mean weight gain in the first group.
$\overline{m_2}$ = mean weight gain in the second group.
$s_1$ = standard deviation of weight gain in the first group.
$s_2$ = standard deviation of weight gain in the second group.
$N$ = number of subjects in either group.
*This formula is applicable only when the number of subjects is the same for both groups. Different formulas are used if the numbers in the two groups are different, but the principle of calculating the value of $t$ remains the same.

**Table 1**
Sample sizes and the sample mean difference in IQ points for the intervention to be considered statistically significant at $p < 0.05$. IQ of the comparison sample = 100; SD = 17.5.

| Sample size in each group | Difference in mean IQ |
|---|---|
| 6 | 22.51 |
| 15 | 13.09 |
| 25 | 9.70 |
| 100 | 4.85 |
| 200 | 3.43 |
| 400 | 2.43 |
| 900 | 1.62 |
| 1500 | 1.25 |