A NEW PARADIGM FOR HYPOTHESIS TESTING IN MEDICINE, WITH EXAMINATION OF THE NEYMAN PEARSON CONDITION

ABSTRACT. In the past, hypothesis testing in medicine has employed the paradigm of the repeatable experiment. In statistical hypothesis testing, an unbiased sample is drawn from a larger source population, and a calculated statistic is compared to a preassigned critical region, on the assumption that the comparison could be repeated an indefinite number of times. However, repeated experiments often cannot be performed on human beings, due to ethical or economic constraints. We describe a new paradigm for hypothesis testing which uses only rearrangements of data present within the observed data set. The token swap test, based on this new paradigm, is applied to three data sets from cardiovascular pathology, and computational experiments suggest that the token swap test satisfies the Neyman Pearson condition.

Key Words: Computer simulation, Hypothesis test, Neyman Pearson lemma, Token swap test.

INTRODUCTION

Traditionally, statistical hypothesis testing in medicine has employed the paradigm of the repeatable experiment. While this paradigm is appropriate for some scientific studies, indefinitely repeated experiments cannot be performed on humans; and animal rights activists are questioning the excessive use of experimental animals. There is a fundamental paradox. If one learns how to improve patient care while performing a clinical experiment, then one is obliged to discontinue any patient selection constraints within the experiment and offer the improved care to all patients (Taylor et al., 1984, p. 1363; Ellenberg, 1984, p. 1404; Angell, 1984, p. 1385). In the theory of statistics, there is a conservative hypothesis, known as the 'null hypothesis,' and any number of 'alternative hypotheses.' An 'unbiased' sample is drawn from a source population, often infinite in size with a known mathematical form (Gaussian, binomial, etc.); in principle, additional samples can be redrawn indefinitely many times. A prescribed formula is used to calculate a 'statistic' (numerical value) from information present in each sample. If the value of the statistic falls within a preassigned 'critical region', then the null hypothesis is rejected. The 'size' of a critical region is a value between 0 and 1, typically 0.05, 0.01, 0.001, etc. A critical region has the property that, if the null hypothesis is true, then repeated samples drawn from the source population will (erroneously)
reject the null hypothesis at a frequency of less than 5% (or 1%, 0.1%, etc.). Since many critical regions may be specified for a given critical region size, a second constraint is often placed on the critical region: namely, the probability of accepting the null hypothesis, when in fact the null hypothesis is false, should be minimized. The 'Neyman Pearson Lemma' asserts that if a distribution satisfies a certain numerical inequality, then the probability of erroneously accepting the null hypothesis is minimized. This numerical inequality, known as the Neyman Pearson condition, is satisfied by all the commonly used statistical hypothesis tests (Student t, analysis of variance, chisquare, Fisher exact test, etc.) (Hoel, 1971, pp. 97, 190).

In this report, we describe a new paradigm of hypothesis testing which does not involve a concept of repeated or unbiased sampling (Moore, et al., 1986, p. 182). Rather, the test uses a paradigm of misclassifications (fortuitous rearrangements) of data present within the observed data set, and does not appeal to a larger source population. We describe the application of a 'token swap' test, based on this paradigm, to three data sets relevant to cardiovascular pathology, and we present computational experiments in which the test satisfies an analogue of the Neyman Pearson condition.

CONTINGENCY TABLE ANALYSIS

In a statistical hypothesis test, the observed data are used to compute a statistic, which is then compared to a standard distribution representing the behavior of the presumed source population. The comparison of the computed 'statistic' with a critical region of the standard distribution yields a $p$-value, or probability that the observed data are drawn from a source population satisfying the null hypothesis. This probability is a long term average of proportions, assuming that the observed data are drawn from an indefinitely repeatable set of sample drawings from the source population. The $p$-value also represents the probability of committing a Type I error, i.e., the probability of asserting that the null hypothesis is false when in fact the observed data are actually drawn from a population satisfying the null hypothesis. If $p$ is small enough (i.e., the probability of committing a Type I error is sufficiently remote), then we 'reject' the null hypothesis. That is, the statistic is said to be 'significant' if its value is more extreme than all but 5% (or 1%, etc.) of the values in the chosen distribution.

The two by two contingency table is a widely used format for presenting the effects of a given treatment or the associations of two disease