THE TWO SAMPLE PROBLEM WITH CENSORED DATA

BRADLEY EFRON
STANFORD UNIVERSITY

1. Introduction

A medical investigator attempting to compare two different treatments for, say, prolongation of life among disease victims, often finds himself in the following situation: at time T, when it is necessary to end the experiment, or at least evaluate the results up to that time, a certain number of the patients in each treatment group will still be alive. His data will then be represented by two sets of numbers which might look like $x_1, x_2, x_3+, x_4, x_5+, x_6, \cdots, x_m$ and $y_1, y_2+, y_3+, y_4, \cdots, y_n$. Here x_1 and x_2 would represent actual lifetimes, while x_3+, a "censored" observation, represents a lifetime known only to exceed x_3. If all the patients in both treatment groups were treated at time 0, then every $+$ value would be equal to T, a situation that has been investigated by Halperin [1]. Frequently, however, patients enter the investigation at different times after it has begun, and the $x+$ and $y+$ values may range from 0 to T. Such a situation, of course, complicates the comparison of the two treatments, particularly if the mechanism censoring the x values is different from that censoring the y values. This may happen, for instance, if the x sequence was run some time ago, so that nearly all the patients have been observed to their death times, while the y sequence is begun later, and contains many censored observations.

Gehan [2] and Gilbert [3] have independently proposed the same extension of the Wilcoxon statistic as a solution to the two sample problem with censored data. In this paper the problem is discussed further, and a different test statistic is proposed, which is shown to be, in some ways, superior to the Gehan-Gilbert statistic.

2. A statement of the problem and some notation

Suppose $x_1^0, x_2^0, \cdots, x_m^0$ are independent, identically distributed random variables, having $F^0(s) = P\{x_i^0 \geqq s\}$ as their common right sided cumulative distribution function (c. d. f.). (Because of the censorship from the right, this is a more convenient function to deal with than the usual left c. d. f. Note that $F^0(s)$ is a left continuous, nonincreasing function of s, and that $F^0(-\infty) = 1$, $F(\infty) = 0$.) Likewise, let $y_1^0, y_2^0, \cdots, y_n^0$ be independent, identically distributed