Journal of the KSQC
Vol. 18, No. 1, June 1990

1990년 6월 韓國品質管理學會誌 제18권 제1호

A Distribution-Free Confidence Interval for Difference between Treatment and Control

Sang-Gue Park*
Tai-Kyoo Kim*
Gyu-Jin Jeong*
Bae-Hyun Yoon*

ABSTRACT

The two-sample problem where an experimental treatment is compared with a control is considered. Without making any parametric model assumptions for the distributions(or survival distributions), a measure for summarizing the differences between the treatment and the control is introduced. A method for constructing a confidence interval for the proposed measure is given in cases of complete and right random censored data. This method is illustrated with two numerical examples.

I. Introduction

Researchers frequently compare the effects of an experimental treatment with a control treatment or a placebo. Part of the analysis of such data is oriented to the estimation of a measure of differences between the effects of the treatment and the control. Suppose that survival time is the response variable on which the data are available. Ameasure of interest is the probabilty of surviving beyond certain control percentile survival time under the experimental treatment. This paper deals with the estimaton of such a measure.

*Dept. of Applied Statistics Hannam University, Daijeon, Korea.

We assume that the responses(X_{01}, X_{02}, ···, X_0n_0) and (X_{11}, X_{12}, ···, X_1n_1) are independent random variables with continuous cumulative distributions F_0 and F_1 respectively. Let p be a positive fraction and ξ_0 be the $100p$-percentile of the distribution F_0. A measure of interest is

$$\theta(p)=F_1[\xi_0]. \tag{1}$$

We note that if this quantity is "sufficiently" smaller than p, researchers may tend to recommend to use of the experimental treatment or may decide to investigate further about the beneficial effects of the experimental treatment.

Gart(1963) pointed out that sometimes the parameter of interest is $F_1[\xi_0]$(i.e. the possible discrepancy between F_1 and F_0 is characterized in terms of $F_1[\xi_0]$. He considered the logit transformation of $F_1^*(X(m+1))$ and utilized the asymptotic distribution of the control median statistic U to find a confidence interval for $F_1[\xi_0]$ with approximate confidence interval $1-\alpha$, where m is the greatest integer not exceeding $[n_0p]$ and $F_i^*(x)$ is the empirical distribution of $F_i(x)$, for $i=0$, 1.

Chakraborti and Desu(1986) discussed a distribution-free confidence interval of difference between quartiles of two distributions with censored data. Further they also mentioned that a point estimator could be very biased estimator if censorings exist.

These motivate us to consider this confidence interval. We derive an alternative confidence interval of $\theta(p)$ for the complete and censored data case ; this can be view as a generalization of Gart's work. We see that such a confidence interval has not only the nice properties of Gart's interval but is much simpler to compute.

II. Confidence Interval for $\theta(p)$

We first find an exact distribution-free $100(1-\alpha)\%$ confidence interval for ξ_0 based on two order statistics of the control sample. From the work of Mac Kinnon(1964), we easily construct the confidence interval for ξ_0 ; choose an integer r, $0<r<[n_0p]+1$, such that

$$\sum_{i=r}^{n_0-r}\binom{n_0}{i}p_i(1-p)^{n_0-i}\geq 1-\alpha. \tag{2}$$

We assume that such an integer r exists. Then we can see that.

$$P(X_0(r)<\xi_0<X_0(n_0-r+1)]\geq 1-\alpha. \tag{3}$$

Hence $\{X_0(r),\ X_0(n_0-r+1)\}$ is an exact distribution free confidence interval for ξ_0 with confidence coefficient $1-\alpha$.

Theorem 1 : The interval

$$\{F_1^*(X_0(r)),\ F_1^*(X_0(n_0-r+1))\} \tag{4}$$

is a confidence interval for $F_1(\xi_0)$, with approximate confidence coefficient $1-\alpha$.

Proof :

Let $\delta = F_1(\xi_0)$, $P[F_1^*(X_0(r)) > F_1(\xi_0)] = P[U(r) > F_0 F_1^{*-1}(\delta)]$, where

$U(r)$ is the $r-$th order statistic in a random sample of size n_0 from the Uniform(0, 1) distribution. Further this probability is equal to

$$1-I(F_0F_1^{*-1}(\delta)\ ;\ r, n_0-r+1), \tag{5}$$

where $I(x\ ;\ a,\ b)$ is the incomplete beta function defined by

$$I(X\ ;\ a,\ b) = [B(a,\ b)]^{-1}\int^{x} t^{a-1}(1-t)^{b-1}\ dt,\ a>0,\ b>0\ ;\ 0<x<1.$$

Since $F_1^*(\delta)$ is the δ -th quantile of the X_1-sample, from the strong consistency of sample quantiles(Serfling(1980)), we have

$$F_1^{*-1}(\delta) \xrightarrow{a.s.} F_1^{-1}(\delta),\ \text{as}\ n_1\to\infty. \tag{6}$$

Now as F_0 is continuous,

$$F_0(F_1^{*-1}(\delta)) \xrightarrow{a.s.} F_0(F_1^{-1}(\delta)) = p,\ \text{as}\ n_1\to\infty. \tag{7}$$

Therefore as $n_1\to\infty$,

$$P[F_1^*(X_0(r)) > F_1(\xi_0)] \to 1-I(p\ ;\ r,\ n_0-r+1). \tag{8}$$

Similarly as $n_1\to\infty$,

$$P[F_1^*(X_0(n_0-r+1)) < F_1(\xi_0)] \to 1-I(p\ ;\ n_0-r+1,\ r). \tag{9}$$

As $r < n_0-r+1$, we have

$$P[F_1^*(X_0(r)) < F_1(\xi_0) < F_1^*(X_0(n_0-r+1))]$$

$$= 1-p[F_1^*(X_0(r)) > F_1(\xi_0)] - P[F_1^*(X_0(n_0-r+1) < F_1(\xi_0)]. \tag{10}$$

In view of (8) and (9), the limit of this probabilty tends to

$$I(p\ ;\ r,\ n_0-r+1) - I(p\ ;\ n_0-r+1,\ r). \tag{11}$$

This quantity can be expressed as $\sum_{i=r}^{n_0-r}\binom{n_0}{i}p_i(1-p)^{n_0-i}$, which is at least $1-\alpha$ by(2). And the proof is complete.

Remark 1 : To determine r such that $(X_0(r),\ X_0(n_0-r+1))$ is a $100(1-\alpha)$ confidence interval for ξ_0, tables prepared by Mac Kinnon(1964) can be used. It is clear that the confidence interval $[F_1^*(X_0(r)),\ F_1^*(X_0(n_0-r+1))]$ is an asymptotically distribution free confidence interval. The interval has the desirable property that its left end point is ≥ 0 and its right end point is ≤ 1 for any set of data. Finally compared to Gart's interval this is much simpler to compute.

In case no tables of r are readily available, we may use the following approach of finding a confidence interval for $F_1(\xi_0)$. We start with a confidence interval for ξ_0 given by(*) in Serfling (1980, p104). Let $\{k(1,\ n_0)\}$ and $\{k(2,\ n_0)\}$ be sequences of integers satisfying

$$\begin{aligned}&1<k(1,\ n_0)<k(2,\ n_0)<n_0,\\&k(1,\ n_0)=n_0(p-pz_{\alpha/2}n_0^{-1/2}),\\&k(2,\ n_0)=n_0(p+pz_{\alpha/2}n_0^{-1/2}).\end{aligned}\tag{12}$$

Then as $n_0\to\infty$, the interval $[X_0(k(1,\ n_0)),\ X_0(k(2,\ n_0))]$ are distribution-free and has approximate confidence coefficient $1-\alpha$.

Theorem 2 : Let

i) ξ_0 be the unique x-solution of $F_0(x-)\leq p\leq F_0(x)$;

ii) F_0 be twice differentiable at ξ_0 with $f(\xi_0)>0$;

iii) $\{k(1,\ n_0)\}$ and $\{k(2,\ n_0)\}$ be two sequences of integers satisfying (12).

Then min $(n_0,\ n_1)\to\infty$, the interval $[F_1^*\{k(1,\ n_0)\},\ F_1^*\{k(2,\ n_0)\}]$ is an asymptotically distribution-free confidence interval for $F_1(\xi_0)$ with approximate confidence coefficient $1-\alpha$.

Proof :

Note that $P[F_1^*\{k(1,\ n_0)\}>F_1(\xi_0)]=P[F_1^*\{k(1,\ n_0)\}-F_1(\xi_0)>0]$.

We can rewrite $F_1^*\{k(1,\ n_0)\}-F_1(\xi_0)$ as, $A+B$, say, where

$$\begin{aligned}&A=[F_1^*\{k(1,\ n_0)\}-F_1^*(\xi_0)]\\&B=[F_1^*(\xi_0)-F_1^*(\xi_0)].\end{aligned}$$

First observe that as $n_1\to\infty$.

$$B\xrightarrow{a.s.}0.\tag{13}$$

Now A can be rewritten as $C+D$, say, where

$$C=[\{F_1^*\{k(1,\ n_0)\}-F_1^*(\xi_0)\}-\{F_1\{K(1,\ n_0)\}-F_1(\xi_0)\}]$$
$$D=[F_1\{k(1,\ n_0)\}-F_1(\xi_0)].$$

from theorem 2.5.1 of Serfling(1980),

$$X_0\{k(1,\ n_0)\}\xrightarrow{p}p,\ \text{as}\ n_0\to\infty.$$

From the definition of $k(i,\ n_0)/n_0$, $k(i,\ n_0)/n_0=p+0((\log\ n_0)^{1/2}n_0^{-1/2}$ as $n_0\to\infty$, $i=1,\ 2$. Hence by lemma C of Serfling(1980, p97), we get, WP 1,

$$|X_0(k(1,\ n_0))-\xi_0(p)|\leq a(n_0),$$

where $a(n_0)\sim E\ n_0^{-1/2}(\log\ n_0)^{1/2}$, $E>0$, as $n_0\to\infty$.

By lemma E *of Serfling*(1980, *p*97), *we get*, WP 1,

$$|C|=0(n_1^{-3/4}(\log\ n_1)^{3/4}),\ \text{as}\ \min(n_0,\ n_1)\to\infty.$$

Since $n_1^{-3/4}(\log\ n_1)^{3/4}\to 0$, *as* $n_1\to\infty$, we have

$$|C|\xrightarrow{a.s.}0,\ \text{as}\ \min(n_0,\ n_1)\to\infty. \tag{14}$$

Therefore by theorem (*ix*) of Rao(1973, p101) we have that the limiting distribution of $F_1^*(X_0(k(1,\ n_0)))-F_1(\xi_0(P))$ is the same as of D as $\min(n_0,\ n_1)\to\infty$. Thus as $\min(n_0,\ n_1)\to\infty$,

$$P[F_1^*\{k(1,\ n_0)\}>F_1(\xi_0)]\to P[F_1^*\{k(1,\ n_0)\}-F_1(\xi_0)>0]$$
$$=P[X(k(1,\ n_0))>\xi_0]\to\alpha/2$$

Serfling(1980, p104.). Similarly we can show that $\min(n_0,\ n_1)\to\infty$,

$$P[F_1^*\{k(2,\ n_0)\}<F_1(\xi_0)]\to\alpha/2.$$

Hence, it follows that, as $\min(n_0,\ n_1)\to\infty$

$$P[F_1^*\{k(1,\ n_0)\}<F_1(\xi_0)<P[F_1^*\{k(2,\ n_0)\}]\to 1-\alpha,$$

and the proof is complete.

Remark 2 : When there are k treatments with a control and a simultaneous confidence interval for $\{F_1(\xi_0),\ F_2(\xi_0),\ \cdots,\ F_k(\xi_0)\}$ might be of interest, the above theorems can be easily applied to obtain it by using Bonferroni inequality.

III. Confidence Interval for $\theta(p)$ with right random censoring

Suppose that X_{i1}, X_{i2}, ···, $X_i n_i$ are te true survival times of n_i individuals in a random sample. For $i=0$, 1, the variables X_{ij}, $j=1$, 2, ···, n_i, are independent and identically distributed with life-time distribution function F_i and survival function $S_i=1-F_i$. Due to the presence of right censorship, the period of follow-up for the j-th individual in the i-th sample is restricted to T_{ij} The observed survival time for the j-th individual in the i-th sample is then $Z_{ij}=\min(X_{ij}, T_{ij})$. One would also observe δ_{ij} which indicates if X_{ij} is censored or not. If $\delta_{ij}=0$, the observation is said to be censored ; otherwise, the observation is said to be uncensored. Let $Z_i(1)<Z_i(2)<\cdots<Z_i(n_i)$ to be ordered Z_{ij}'s and let $\delta_i(j)$ be the δ-value associated with $Z_i(j)$, $j=1$, 2, $\cdots n_i$, $i=0$, 1.

The $100p$-th percentile of F_0 is $\xi_0=F_0^{-1}(p)=\inf\{t : F_0(t)\geq p\}$. Let $\hat{F}_i$ be the Kaplan-Meier estimator of F_i and let $\hat{\xi_0}=\hat{F}_0^{-1}(p)=\inf\{t : \hat{F}_0(t)\geq p\}$ be an estimator of ξ_0. Let V_0 be the asymptotic variance of $\sqrt{n_0}[\hat{F}_0(\xi_0)-p]$. From theorem 5 of Breslow and Crowley(1974) it follows that

$$V_0=[1-p]^2\int_0^{\xi_0}[1-F_0^*]^2\,d\bar{F}_0, \tag{15}$$

where F_0^* is the distribution function of observed lifetimes in the contol sample, i. e. $F_0^*(t)=P(X_{0j}\leq t)$ and $\bar{F}_0$ is the distribution function of the true observed lifetimes, i. e. $\bar{F}_i=P(X_{0j}\leq t, \delta_{0j}=1)$. A consistent estimator of V_0 Greenwood, 1926) is given by

$$\hat{V}=(1-p)^2\sum\frac{n_0 d_0(j)\delta_0(j)}{R_0(j)\{R_0(j)-d_0(j)}, \tag{16}$$

where $d_0(j)$ is the observed number of deaths at $X_0(j)$ and $R_0(j)$ is the number of patients at risk at $X_0(j)$. The summation is over all values of $j(1, 2, \cdots, n_0)$ such that each distinct $X_0(j)$ is less than or equal to $\hat{\xi}_0$.

The proposed confidence interval for $\theta(p)$ is constructed in two steps. The first step is to find a confidence interval for ξ_0. The second step is to utilized this confidence interval to obtain a confidence interval for $\theta(p)$. Hence one first chooses two positive fractions $\hat{p}_-(n_0)$ and $\hat{p}_-(n_0)$ given by

$$\hat{p}_-(n_0)=p-z_{\alpha/2}(\hat{V}/n_0)^{1/2} \tag{17}$$

$$\hat{p}_+(n_0)=p+z_{\alpha/2}(\hat{V}/n_0)^{1/2} \tag{18}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-th quantile of the standard normal distribution. Now a confidence interval for ξ_0 is

$$[\hat{\xi}_-,\ \hat{\xi}_+]=[\hat{F}_0^{-1}(\hat{p}-(n_0)),\ \hat{F}_0^{-1}(\hat{p}_+(n_0))]\ ; \tag{19}$$

the confidence coefficient of this interval is approximately $100(1-\alpha)\%$.

From(19), a confidence interval for $\theta(p)$ is

$$[\hat{\theta}_-,\ \hat{\theta}_+]=[\hat{F}_1(\hat{\xi}_-),\ \hat{F}_1(\hat{\xi}_+)]. \tag{20}$$

IV. Numerical Examples

1. The following data appears in Lehmann(1975, problem 1.25).

Suppose that a new postsurgical treatment is being compared with a standard treatment by observing the recovery times of 9 treatment Subjects and of 9 controls.

Control : 20, 21, 24, 30, 32, 36, 40, 48, 54

Treatment : 19, 22, 25, 26, 28, 29, 34, 37, 38

Lep $p=0.5$ and $\alpha=0.05$. For this data, one finds $\hat{\xi_0}=32$, $\hat{p}_-(9)=0.173$ and $\hat{p}_+(9)=0.826$. Thus, an approximately 95% confidence interval for ξ_0 is [21, 48]. Hence an approximately 95% confidence interval for $\theta(1/2)$ is

$$[\hat{F}_1(21),\ \hat{F}_1(48)]=[0.18,\ 1].$$

2. The following data reported in Miller(1981, p.49), gives the length of remission(in weeks) for two groups of patients. The first group received maintenance chemotherapy ; the second or control group did not. The first group objective of the experiment was to see if the maintenance chemotherapy prolonged the length of remission.

Maintained group : 9, 13, 13_+, 18, 23, 28_+, 31, 34, 45_+, 48, 161_+

Nonmaintained group : 5, 5, 8, 8, 12, 16_+, 23, 27, 30, 33, 43, 45.

Here "+" indicates a censored observation.

Suppose that $p=0.5$ and $\alpha=0.5$. For this data $\hat{\xi_0}=23$ and $\hat{V}=0.2786$. Therfore $\hat{p}_-(11)=0.2013$ and $\hat{p}_+(12)=0.7987$. Thus, an approximately 95% confidence interval for ξ_0 is [8, 33]. Hence an approximately 95% confidence interval for $\theta(1/2)$ is $[[\hat{F}_1(8),\ \hat{F}_1(33)]=[0.0,\ 0.51]$.

REFERNECES

1. Breslow, N. and Crowley, J. (1974), "*A large sample study of the life table and product limit estimates under random censorship*", Ann. Statist., 2, 437-453.
2. Chakraborti, S. and Desu, M.M. (1986), "*A distribution-free confidence interval for the difference between quantiles with censored data*", Statistica Neerlandica, 40, 93-98.
3. Gart, J.J. (1963), "*A median test with sequential applications*", Biometrika, 50, 55-62.
4. Kaplan, E.L. and Meier, P. (1958), "*Nonparametric estimation from incomplete observations*", J. Amer. Statist. Assoc., 53, 457-481.
5. Lehmann, E.L. (1975), "*Nonparametrics*", San Francisco, Holden-Day.
6. Mackinnon, W.J. (1964), "*Tables for both the sign test and distribution-free confidence intervals of the median for sample sizes to* 1000", J. Amer. Statist. Assoc., 59, 935-956.
7. Miller, R.G. (1981), "*Survival Analysis*", New York, John Wiley.
8. Rao, C.R. (1973), "*Linear Statistical Inference and its Applications*", New York, John Wiley.
9. Serfling, R.J. (1980), "*Approximate theorems Mathematical Statistics*", New York, John Wiley.