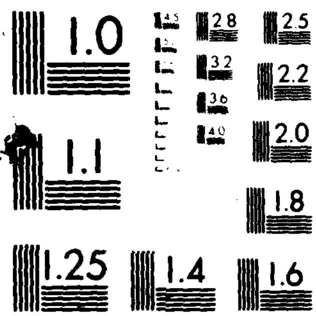
AD-A112 947 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 12/1
A BAYES PROCEDURE FOR SELECTING THE POPULATION WITH THE LARGEST--ETC(U)
JUL 81 K ALAM N00014-75-C-0451
UNCLASSIFIED N130 NL

END
DATE
FILMED
5 82
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A112947

(12)

A BAYES PROCEDURE FOR SELECTING THE POPULATION WITH THE LARGEST pth QUANTILE

Khursheed Alam*

Department of Mathematical Sciences
Clemson University

Technical Report #368

July, 1981

NR #130

*This work was supported by the U.S. Office of Naval Research under Contract No. NO. 00014-75-C-0451.

A BAYES PROCEDURE FOR SELECTING THE POPULATION WITH THE LARGEST pth QUANTILE

Khursheed Alam*
Clemson University

Abstract

The Bayesian approach has not been very fruitful in treating nonparametric statistical problems, due to the difficulty in finding mathematically tractable prior distributions on a set of probability measures. The theory of the Dirichlet process has been developed recently. The process generates randomly a family of probability distributions which can be taken as a family of prior distributions for the Bayesian analysis of some nonparametric statistical problems. This paper deals with the problem of selecting a distribution with the largest pth quantile value, from $k \geq 2$ given distributions. It is assumed à priori that the given distributions have been generated from a Dirichlet process.

---

* The work was supported by the U.S. Office of Naval Research under Contract No. 00014-75-C-0451.

1. Introduction. The Bayesian approach has been very useful in treating statistical problems of parametric nature. In treating nonparametric statistical problems, the scope of Bayes methods is limited due to the difficulty of finding mathematically tractable prior distributions on a set of probability measures. For practical application, the family of prior distributions should be such that (i) its support is large and that (ii) the posterior distribution, given a sample drawn from the true distribution, is analytically manageable. In a couple of papers, Ferguson (1973, 1974) has developed the theory of the Dirichlet process. The process generates randomly a family of probability distributions which can serve as a family of distributions, with the properties (i) and (ii). Therefore, the process can be used in the analysis of some statistical problems of nonparametric nature, by the Bayes method.

In this paper we consider a problem of ranking and selection. Given a sample of size n, drawn from each of k univariate distributions $Q_1,\ldots,Q_k$, we want to select the population associated with the largest pth quantile value, which we call the "best" population. A Bayes procedure for selecting the best population is given. It is assumed à priori that the given distributions are k independent realizations of a Dirichlet process. The given procedure may be considered as a Bayes solution of a nonparametric problem. In Section 2 we describe the Dirichlet process and its properties relating to the selection problem. In Section 3 we describe the selection rule and derive a formula for the probability of a correct selection.

The problem of selecting the population with the largest pth quantile value arises in various situations of practical interest. Gibbons, Olkin and Sobel (1977) have given a number of examples to illustrate the importance of the problem.

The problem of selecting the population with the largest pth quantile value, from k populations and the analogous problem of selecting a subset of the k populations which includes the best population, have been considered by Sobel (1967), Rizvi and Sobel (1967) and Desu and Sobel (1971). The thrust of these papers is to find a "least favorable" configuration of the populations for which the probability of a correct selection is minimized and to determine a minimum sample size n such that the probability of a correct slection is at least as large as a specified number less than 1. Rizvi and Saxena (1972) give a confidence interval for the largest pth quantile.

2. Dirichlet Process. First we describe the Dirichlet distribution. Let $X_1,\ldots,X_k$ be k independent random variables, where X_i is distributed according to the gamma distribution with ν_i degrees of freedom and a common scale

parameter, $i = 1,\ldots k$. Let $Z_i = X_i / (\sum_{i=1}^{k} X_i)$. The Dirichlet distribution with parameter $(\nu_1,\ldots,\nu_k)$ is given as the joint distribution of $Z_1,\ldots,Z_k$. Marginally, Z_i is distributed according to the beta distribution with parameter $(\nu_i, \nu_1+\ldots+\nu_{i-1}+\nu_{i+1}+\ldots+\nu_k)$.

The Dirichlet process is defined on a general space, but for the purpose of this paper we consider only (R,B), where R denotes the real line and B denotes the σ-algebra of all Borel subsets of R. Let $\alpha(.)$ be a finite measure on (R,B), and let Q be a stochastic process indexed by the elements of B. We say that Q is a Dirichlet process with parameter α, and write $Q \in D(\alpha)$, if for every finite measurable partition $(B_1,\ldots,B_m)$ of R, the vector $(Q(B_1),\ldots,Q(B_m))$ is distributed according to the Dirichlet distribution with parameter $(\alpha(B_1),\ldots,\alpha(B_m))$. Thus Q is a random probability distribution on R, and Q(A) represents the probability measure of A under Q, for $A \in B$.

It is known that Q is discrete with probability 1, and that if $X_1,\ldots,X_n$ is a sample from Q then a posteriori,

$Q \in D(\alpha + \sum_{i=1}^{n} \delta_{X_i})$, where δ_x is a measure which assigns unit mass to the single point x. It is a drawback that Q is discrete with probability 1. We should have a prior that chooses a continuous distribution with probability 1. However, Ferguson (1973) points out that the discreteness of Q does not limit the use of the Dirichlet process as a family of prior distributions in certain problems, such as the estimation of the quantiles.

We give below, certain properties of the Dirichlet process which will be used in the sequel. The following result is due to Ferguson (1973). Let Q be a realization of the Dirichlet process with parameter α, and let $M = \alpha(R)$. For estimating Q let the loss function be given by

$$L(Q,\tilde{Q}) = \int_{-\infty}^{\infty}(Q(t) - \tilde{Q}(t))^2 \, dw(t)$$

where $\tilde{Q}$ denotes the estimate of Q, $Q(t) = Q((-\infty,t])$ and w is a given finite measure on (R,B). Then a Bayes estimator of Q is $EQ = Q^o$, say, where

$$(2.1) \qquad Q^o(t) = (\alpha(-\infty,t])/M.$$

The distribution Q^o is our prior guess of Q. If a sample of size n is drawn from Q, the Bayes estimator is given by

$$(2.2) \qquad \tilde{Q} = p_n Q^o + (1-p_n)F$$

where $p_n = M/(M+n)$ and F denotes the empirical distribution function of the sample.

Let ξ denote the pth quantile of Q, given by

$$Q((-\infty,\xi)) \le p \le Q((-\infty,\xi])$$

for $0 < p < 1$. It is known that ξ is uniquely determined with probability 1. For estimating ξ let the loss function be given by

$$L(\xi, \tilde{\xi}) = \begin{cases} q(\xi-\tilde{\xi}) & \text{for } \xi \ge \tilde{\xi} \\ (1-q)\ (\tilde{\xi}-\xi) & \text{for } \xi < \tilde{\xi} \end{cases}$$

where $\tilde{\xi}$ is an estimate of ξ and q is a given number, such that $0 < q < 1$. Any qth quantile of the distribution of ξ is a Bayes estimate of the realized value of ξ, under the given loss. Let $b(x;a,c)$ denote the beta density function and let $u = u(p,q,M)$ be a solution of the equation,

$$(2.3) \qquad I(p;uM,\ (1-u)M) = 1 - q$$

where

$$I(x;a,c) = \int_0^x b(y;a,c)dy$$

denotes the cdf of the beta distribution. Then the $u(p,q,M)$th quantile of Q^o is a Bayes estimate of ξ. Given a sample of size n drawn from Q, a Bayes estimate is the $u(p,q,M+n)$th quantile of $\tilde{Q}$, given by (2.2). Ferguson has tabulated the values of $u(p,q,M)$ for $q = .05(.05).95$, $p = .05(.05).95$ and $M = 1(1)10$.

3. Selection rule. Let ξ_i denote the pth quantile of Q_i, and let F_i denote the empirical distribution function of the sample drawn from Q_i. From (2.2) we have that

$$(3.1) \qquad \tilde{Q}_i = p_n Q^o + (1-p_n)F_i$$

is a Bayes estimator of Q_i and that $\tilde{\xi}_i$, equal to the $u(p,q,M+n)$th quantile of $\tilde{Q}_i$, is a Bayes estimator of ξ_i, where the function $u = u(p,q,M)$ is given by (2.3). Therefore, we select the population associated with the largest value of $\tilde{\xi}_i$ as the best population. If two or more of the $\tilde{\xi}_i$ are tied for the largest value we select one of them randomly. We shall ignore in the following discussion the occurrence of a tie.

The probability of a correct selection (PCS) *for* the given rule, is obtained as follows: Let $v = u(p,q,M+n)$ and

$$\gamma(x) = [n(v-p_n\, Q^o(x))/(1-p_n)]$$

where $[x]$ denotes the smallest integer greater than or equal to x.

Since the selection rule is symmetric, the PCS is given by

$$\text{(3.2)} \qquad PCS = KP\{\xi_i \le \xi_K,\ \tilde{\xi}_i \le \tilde{\xi}_K,\quad i = 1,\ldots,k-1\}$$
$$= K \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G^{K-1}(x,y)dG(x,y)$$

where

$$\text{(3.3)} \qquad G(x,y) = P\{\xi_i \le x,\ \tilde{\xi}_i \le y\}$$
$$= P\{Q_i(x) \ge p,\ \tilde{Q}_i(y) \ge v\}$$
$$= P\{Q_i(x) \ge p,\ F_i(y) \ge (v-p_nQ^o(y))/(1-p_n)\}.$$

We have

$$P\{F_i(y) \ge (v-p_nQ^o(y))/(1-p_n) | Q_i\}$$
$$= I(Q_i(y);\ \gamma(y),\ n+1-\gamma(y)).$$

Therefore

$$(3.4) \qquad G(x,y) = \int\int_{s \geq p} I(t;\gamma(y),\ n+1-\gamma(y))dH(s,t)$$

where $H(s,t)$ denotes the joint cdf of $s = Q_i(x)$ and $t = Q_i(y)$.

Now $(Q_i(x),\ Q_i(y)-Q_i(x),\ 1-Q_i(y))$ is distributed according to the Dirichlet distribution with parameter $(\alpha(x),\ \alpha(y)-\alpha(x),\ M-\alpha(y))$ for $x < y$, and $(Q_i(y),\ Q_i(x)-Q_i(y),\ 1-Q_i(x))$ is distributed according to the same distribution with parameter $(\alpha(y),\ \alpha(x)-\alpha(y),\ M-\alpha(x))$ for $x > y$. Given $Q_i(y)$, $Q_i(x)/Q_i(y)$ is conditionally distributed according to the beta distribution with parameter $(\alpha(x),\ \alpha(y)-\alpha(x))$ for $x < y$, and $(1-Q_i(x))/(1-Q_i(y))$ is conditionally distributed according to the beta distribution with parameter $(M-\alpha(x),\ \alpha(x)-\alpha(y))$ for $x > y$. Let $\phi(x,y) = 1(0)$ for $x <(>)\ 1$ and let $I(\omega;a,b) = 1$ for $\omega \geq 1$. From (3.4) we have

$$(3.5) \qquad G(x,y) = \int_0^1 I(\omega;\ \gamma(y),\ n+1-\gamma(y))\{(1-I(\tfrac{p}{\omega};\alpha(x),\alpha(y)-\alpha(x))) \phi(x,y) + (1-\phi(x,y))I(\tfrac{1-p}{1-\omega};M-\alpha(x),\alpha(x)-\alpha(y))\} b(\omega;\ \alpha(y),\ M-\alpha(y))d\omega$$

for $x \neq y$ and

$$(3.6) \qquad G(y,y) = \int_p^1 I(\omega;\gamma(y),\ n+1-\gamma(y))b(\omega;\ \alpha(y);\ M-\alpha(y))d\omega$$

The expression for the PCS given above, involves several parameters, namely, k, p, q and n, in addition to the measure α. Given k, p, q and α, a minimum value of n can be determined such that the value of the PCS is at least as large as a specified number between 0 and 1.

Through a minor modification of the selection rule given above, we can obtain a procedure for selecting a random subset of the given populations which includes the best population. The above result can be used to obtain the probability of a correct selection for that procedure.

For an illustration of the given result, it would be convenient to consider the special case in which α represents the uniform distribution on (0,1). For this case we have

$$\begin{aligned} G(x,y) = \int_o^1 & I(u;[(n+1)v-y],\ n+1-[(n+1)v-y]) \\ & \{(1-I(\tfrac{p}{u};x,y-x))\phi(x,y)+(1-\phi(x,y)) \\ & I(\tfrac{1-p}{1-u};1-x,x-y)\}\ b(u,y,1-y)du, \qquad x \neq y \end{aligned}$$

$$\begin{aligned} G(y,y) = \int_p^1 & I(u;[(n+1)v-y],\ n+1-[(n+1)v-y]) \\ & b(u;y,1-y)du. \end{aligned}$$

4. Conclusion. In this paper we have developed the theory of the Dirichlet process for application to a non-parametric problem of ranking and selection. The given results are mainly of theoretical nature. For practical application it would be necessary to justify or test the assumption that the underlying distributions have been generated from a Dirichlet process. There are also questions relating to the prior specification of the parameter α and to the robustness of the statistical procedures resulting from the given theory, as compared to certain parametric statistical procedures. We shall discuss the application side of the selection problem in another paper.

References

1. Desu, M.M. and Sobel, M. (1971). Nonparametric procedure for selecting fixed-size subsets. Statistical Decision Theory and Related Topics (Ed. Gupta, S.S. and Yackel, J.), Academic Press, New York.

2. Ferguson, T.S. (1974). A Bayesian analysis of some nonparametric problems. Ann. Statist. (1) 209-230.

3. Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. Ann. Statist. (2) 615-629.

4. Gibbons, Olkin and Sobel (1977). Selecting and Ordering Populations - A New Statistical Methodology. John Wiley, New York.

5. Rizvi, M.H. and Saxena, K.M.L. (1972). Distribution-free interval estimation of the largest α-quantile. Jour. Amer. Statist. Assn. (67) 196-198.

6. Rizvi, M.H. and Sobel, M. (1967). Nonparametric procedures for selecting a subset containing the population with the largest α-quantile. Ann. Math. Statist. (38) 1788-1803.

7. Sobel, M (1967). Nonparametric procedures for selecting the t populations with the largest α-quantiles. Ann. Math. Statist. (38) 1804-1816.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NR 130	2. GOVT ACCESSION NO. AD-A112 947	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Bayes Procedure for Selecting the Population with the Largest pth Quantile		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER Technical Report #368
7. AUTHOR(s) Khursheed Alam		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 365-049
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code 434 Arlington, Va. 22217		12. REPORT DATE July, 1981 13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Ranking and selection, Bayes procedure, Dirichlet process, Population quantile.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Bayesian approach has not been very fruitful in treating nonparametric statistical problems, due to the difficulty in finding mathematically tractable prior distributions on a set of probability measures. The theory of the Dirichlet process has been developed recently. The process generates randomly a family of probability distributions which can be taken as a family of prior distributions for the Bayesian analysis of some nonparametric statistical problems. ~~This paper deals with the problem of~~

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) selection a distribution with the largest pth quantile value, from $k \geq 2$ given distributions. It is assumed à priori that the given distributions have been generated from a Dirichlet process.

DA

FILM