THE WEIGHTED MEDIAN TEST FOR FINITE POPULATION SAMPLES

Rick L. Williams and Rebecca L. Perritt, Research Triangle Institute

1. Introduction

Inference methods for medians estimated from a complex finite population sample (possibly stratified, clustered and unequally weighted) are not well developed. A method for constructing confidence intervals for medians from such samples was justified by Woodruff (1952) and is presented in Kish (1965, Section 12.9) and Hansen, Hurwitz and Madow (1953, Section 10.18 of Vol. I and Section 10.7 of Vol. II). This method is quite general and can be applied to most any survey where the variance is estimable. While the approach can be used to test the difference between any two subgroup medians (see Kish, 1953), it does not directly apply to the analysis of more than two subgroups. On the other hand, the well known Brown and Mood (1951) median test (see Lindgren, 1968 or Conover, 1980) provides a method for testing the equality of the medians from several different subgroups but is not directly applicable to complex finite population samples. Mood (1954) showed that for normal shift alternatives, the asymptotic relative efficiency of the median test compared to the t-test is $2/\pi$. This paper adapts the median test to complex finite population sample data.

2. The Median Test

The usual median test considers c populations from which random samples of size n(i) are obtained (i=1,...,c). The c populations are combined and the overall sample median, m, determined as the sample value exceeded by half the observations. To avoid unnecessary complications, assume that the median is unique and that n(1) + ... + n(c) = n = 2k where k is an integer. Next let a(1i) be the number of observations in the ith population which are less than or equal to the overall sample median and let a(2i) be the number of observations that exceed the overall sample median. These form a 2 x c contingency table with fixed marginal totals as follows:

	Population 1	 2	 ...	 c	 Total
≤ Median	a(11)	a(12)	...	a(1c)	k
> Median	a(21)	a(22)	...	a(2c)	k
Total	n(1)	n(2)	...	n(c)	n

Under the null hypothesis that all c populations have the same median, it is expected that a(1i) = a(2i) (i=1,...,c). This hypothesis can be tested using a large sample chi-square statistic with c-1 degrees of freedom.

To recast this problem for a finite population consider a population of N units consisting of c subgroups containing N(1), N(2), ..., N(c) units, respectively. Let M be the median of the overall population and M(i) be the median of subgroup-i. Like before, assume that N = 2K (K an integer) and that M is exceeded by the values of K units in the overall population. Next, define A(1i) to be the number of units in the ith subgroup with values less than or equal to the overall population median and A(2i) to be the number of units with values exceeding the overall population median. This yields the following 2 x c contingency table:

	Subgroup 1	 2	 ...	 c	 Total
≤ Median	A(11)	A(12)	...	A(1c)	K
> Median	A(21)	A(22)	...	A(2c)	K
Total	N(1)	N(2)	...	N(c)	N

Following the approach explained by Woodruff, define P(i) = A(1i)/N(i), which is the proportion of subgroup-i with values less than the overall median. The following two null hypotheses are equivalent:

$$H_o(1): \quad M(1) = M(2) = \ldots = M(c)$$

$$H_o(2): \quad P(1) = P(2) = \ldots = P(c) .$$

The second null hypothesis is amenable to analysis using the weighted least squares approach illustrated by Koch, Freeman and Freeman (1975) or Koch and Lemeshow (1972). This process proceeds by first estimating P(1), P(2),..., P(c) and their variance-covariance matrix. The null hypothesis $H_o(2)$ (and, hence $H_o(1)$) is then tested using a large sample chi-square test.

3. Implementation

Assume that a probability sample is drawn from a population and that the values of a variate z are observed. The only requirements placed on the design is that the estimates of subgroup means have asymptotically multivariate normal distributions with covariance matrices that can be consistently estimated from the sample. Further assume that the population and the sample can be divided into c subgroups and that we wish to test if the subgroup medians are all equal.

Given the distribution function F of z in the overall population, the overall population median is the value M such that $F(M-) < 1/2 \leq F(M)$. The overall median can be consistently estimated by considering the sample empirical distribution function

$$G(z) = \sum_j w(j)\, I_z[z(j)] \,/\, \sum_j w(j)$$

where z(j) and w(j) are the observed value and the sampling weight for sample unit-j, respectively, and I_z is the indicator function

$$I_z[z(i)] = \begin{cases} 1 \text{ if } z(j) \leq z \\ 0 \text{ if } z(j) > z. \end{cases}$$

A consistent estimate of M is the value m such that $G(m-) \leq 1/2 \leq G(m)$.

Once the overall sample median, m, is determined, $P(i)$ can be estimated by

$$p(i) = \sum_j w(j)\, x(j|i)\, I_m[z(j)] \;/\; \sum_j w(j)\, x(j|i)$$

where I_m is the indicator function defined above and $x(j|i)$ is the subgroup-i indicator for sample unit-j given by

$$x(j|i) = \begin{cases} 1 \text{ if unit-j is in subgroup-i} \\ 0 \text{ otherwise.} \end{cases}$$

A consistent estimate of the covariance matrix of $\mathbf{p} = [p(1),\ldots,p(c)]'$, say $\mathbf{V}$, is also required. The estimation of $\mathbf{V}$ will vary from survey to survey, however, one method is to use a survey regression procedure like SURREGR (Holt, 1977, revised by Shah 1982) or SUPER CARP (Hidiroglou, Fuller and Hickman, 1979). These procedures will accommodate most stratified, multistage sample designs. To understand how this can be done, first define

$$y(i) = \begin{cases} 1 \text{ if } z(i) \leq m \\ 0 \text{ otherwise} \end{cases}$$

and $\mathbf{y} = [y(1),\ldots,y(n)]'$ where n is the total sample size. Also, define $\mathbf{x}_j = [x(1|j),\ldots,x(n|j)]'$ for $j=1,\ldots,c$, column vectors corresponding to each subgroup; $\mathbf{O}_n$ to be null column vector of length n; and $\mathbf{P} = [P(1),\ldots,P(c)]'$. Fitting the following model will yield the estimates $\mathbf{p}$ and $\mathbf{V}$:

$$E\begin{vmatrix} \mathbf{y} \\ \cdot \\ \mathbf{y} \end{vmatrix} = \begin{vmatrix} \mathbf{x}_1 & \cdots & \mathbf{O}_n \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \mathbf{O}_n & \cdots & \mathbf{x}_c \end{vmatrix} \quad \mathbf{p}\,.$$

Armed with these estimates, hypothesis $H_o(2)$ can be evaluated by first noting that $H_o(2)$ can be restated as

$$H_o(2):\ \mathbf{CP} = \mathbf{O}_{c-1} \quad \text{vs.} \quad H_a(2):\ \mathbf{CP} \neq \mathbf{O}_{c-1}$$

where $\mathbf{C} = [\mathbf{J}_{c-1} \mid -\mathbf{I}_{c-1}]$, $\mathbf{J}_{c-1}$ is a column vector of all ones and $\mathbf{I}_{c-1}$ is the (c-1) x (c-1) identity matrix. The contrast matrix $\mathbf{C}$ generates the c-1 linearly independent pairwise differences among the elements of $\mathbf{P}$. The test statistic

$$T = [\mathbf{Cp}]'[\mathbf{CVC}']^{-1}[\mathbf{Cp}]$$

is approximately distributed as a chi-square with c-1 degrees of freedom under H_o and the null hypothesis is rejected if T exceeds the appropriate critical value.

4. Example

These examples study the concentrations of benzene and carbon tetrachloride found in breath samples of 350 residents of Bayonne and Elizabeth, New Jersey. A stratified three stage sample survey design was used to collect the data. The first stage sampling units (FSUs) were blocks, block groups, enumeration districts, or combinations thereof as defined by the U.S. Bureau of the Census for the 1980 Decennial Census. A stratified sample of the FSUs was selected with probabilities proportional to the number of occupied housing units in the FSU based upon the 1980 Decennial Census. A compact cluster of housing units was selected at the second stage within each selected FSU. A short screening interview was conducted for each participating household. The purpose of the screening interview was to collect data from a knowledgeable adult household member concerning the age, sex, smoking status, and occupation of all household members. These data were used to stratify the third stage sample of individuals selected for personal exposure and body-burden monitoring. In addition to stratification by age, sex, and smoking status, the occupational data were used to stratify by suspected occupational exposure to the organic chemicals being monitored.

The selected persons were divided into current smokers and current non-smokers. For benzene, a common component in cigarette smoke, the estimated median concentration for current smokers was 20.0 ug/m^3, while for current non-smokers the median was 7.9 with an overall estimated median of 12.0. For carbon tetrachloride, which is not usually associated with cigarette smoke, the estimated medians were 0.63 and 0.70 for smokers and non-smokers, respectively, with an overall estimated median of 0.69. These results are summarized in Table 1.

For both chemicals, the percentage of persons with concentrations less than their respective overall estimated medians were estimated separately for smokers and non-smokers and are presented in Table 2. These estimates were obtained using the survey regression package SURREGR. We see that benzene exhibits a substantial difference in percentages less than the overall median between non-smokers and smokers (61% vs. 31%) with a highly significant chi-square statistic of 15.6 (p=.0015). On the other hand, carbon tetrachloride does not exhibit a significant difference with a chi-square statistic of .48 (p=.4977).

REFERENCES

Brown, G. W., A. M. Mood (1951). "On Median Tests for Linear Hypotheses," *Proceedings of the Second Berkeley Symposium on Mathematics, Statistics and Probability*, University of California.

Conover, W. J. (1980). Practical Nonparametric Statistics, Second Edition, John Wiley and Sons, New York.

Hansen, M. H., W. N. Hurwitz, W. G. Madow (1953). Sample Survey Methods and Theory, Vol. I and II, John Wiley and Sons, New York.

Hidiroglou, M. A., W. A. Fuller, R. D. Hichman (1979). SUPER CARP, Iowa State University, Ames, Iowa.

Holt, M. H. (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, Revised by B. V. Shah (1982), Research Triangle Institute, Research Triangle Park, NC.

Kish, L. (1965). Survey Sampling, John Wiley and Sons, New York.

Koch, G. G., D. H. Freeman, J. L. Freeman (1975). "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, Vol. 43, No. 1.

Koch, G. G., S. Lemeshow (1972). "An Application of Multivariate Analysis to Complex Sample Survey Data," Journal of the American Statistical Association, Vol. 67, No. 340.

Lindgren, B. W. (1968). Statistical Theory, MacMillan, New York.

Mood, A. M. (1954). "On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests," Annals of Mathematical Statistics, Vol. 25.

Woodruff, R. S. (1952). "Confidence Intervals for Medians and Other Position Measures," Journal of the American Statistical Association, Vol. 47.

Table 1. Weighted Median Concentrations (ug/m^3)

	Current Nonsmoker	Current Smoker	Overall
Benzene	7.90	20.00	12.00
Carbon Tetrachloride	0.70	0.63	0.69

Table 2.Summary of Tests for Equality of Medians

	Percent of Concentrations $\leq$ Overall Median			
	Current Nonsmoker	Current Smoker	Chi-square	P-value
Benzene	61	31	15.60	0.0015
Carbon Tetrachloride	49	56	0.48	0.4977