ASYMPTOTICALLY DISTRIBUTION FREE STATISTICS SIMILAR TO STUDENT'S t

Z. W. BIRNBAUM
UNIVERSITY OF WASHINGTON

1. A class of statistics

1.1. Let X be a random variable with a continuous distribution function $F(x) = P\{X \leqq x\}$, $X_1, X_2, \cdots, X_n$ a random sample of X, and $X_{(1)} \leqq X_{(2)} \leqq \cdots X_{(n)}$ the corresponding ordered sample. For given $0 < \gamma < 1$, we consider the γ quantile of X

$$\mu_\gamma = F^{(-1)}(\gamma) \tag{1.1.1}$$

and the corresponding sample quantile

$$V_\gamma = X_{(k)}, \tag{1.1.2}$$

where

$$k = [\gamma n] + 1. \tag{1.1.3}$$

We now consider the statistic

$$S_\gamma = \frac{V_\gamma - \mu_\gamma}{X_{(k+r_2)} - X_{(k-r_1)}}, \tag{1.1.4}$$

where r_1, r_2 are integers such that $0 < r_1 < k$, $0 < r_2 \leqq n - k$.

1.2. The statistics of the form (1.1.4) have a structure somewhat similar to Student's t: the numerator is the difference between an estimate of a location parameter (sample quantile V_γ) and that location parameter (population quantile μ_γ), while the denominator is an estimate (sample interquantile range) of a scale parameter (population interquantile range). A more pertinent analogy with the t statistic is this: the statistic S_γ is invariant under linear transformations and hence, if the distribution function $F(\cdot)$ is given, S_γ has for fixed γ, n, r_1, r_2 a probability distribution independent of location and scale parameters; that is the same for all random variables with distribution functions $F((x - a)/b)$ with arbitrary real a and positive b. One could, therefore, choose, for example, $F(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2}\, dy$ and tabulate the probability distributions of S_γ for practically meaningful values of γ, n, r_1, r_2; these probability distributions

Research supported in part by the U.S. Office of Naval Research.

would be independent of the mean and the variance of X, and could be used in a manner analogous to that in which one uses the t statistic.

The S statistics defined by (1.1.4) can be computed and used when only the three order statistics $X_{(k)}$, $X_{(k+r_2)}$, $X_{(k-r_1)}$ are available, for example in situations when a number of more extreme order statistics have been "censored." This constitutes a possible practical advantage as compared with the t statistic, which can be computed only when the complete sample is available.

Another minor advantage of the S statistics is that, as soon as the order statistics $X_{(k)}$, $X_{(k+r_2)}$, $X_{(k-r_1)}$ are available, the calculation of S is quite simple.

1.3. Among the statistics (1.1.4), the one corresponding to $\gamma = \frac{1}{2}$ and hence $\mu_{1/2}$ = median of X is closest in structure and possible applications to the t statistic. Using the notations

$$\begin{aligned} n &= 2m + 1 \quad \text{(odd sample size)}, \\ \mu_{1/2} &= \mu = F^{(-1)}(1/2) = \text{median}, \\ V_{1/2} &= V = X_{(m+1)}, \end{aligned} \tag{1.3.1}$$

and, specializing even more, $r_1 = r_2 = r$,

we have the statistic

$$S = \frac{X_{(m+1)} - \mu}{X_{(m+1+r)} - X_{(m+1-r)}} \qquad 1 \leqq r \leqq m. \tag{1.3.2}$$

This statistic was considered in some detail in [1], while similar statistics were previously studied in [2]. From now on we shall limit our considerations to the statistic (1.3.2).

2. Exact distribution

2.1. We assume in the following the existence everywhere of the probability density $f(x) = F'(x)$. Using well-known expressions for the joint probability distribution of the order statistics $X_{(m+1-r)}$, $X_{(m+1)}$, $X_{(m+1+r)}$, one obtains

$$\begin{aligned} P\{S > \lambda\} = P(\lambda) = {} & \frac{(2m+1)!}{[(m-r)!\,(r-1)!]^2} \\ & \cdot \int_{v=0}^{+\infty} \int_{u=(\lambda-1)v/\lambda}^{v} \int_{w=v}^{u+v/\lambda} f(u)f(v)f(w) \\ & \cdot F^{m-r}(u)[F(v) - F(u)]^{r-1}[F(w) - F(v)]^{r-1}[1 - F(w)]^{m-r}\,dw\,du\,dv \end{aligned} \tag{2.1.1}$$

for $\lambda > 0$.

2.2. In [1], expression (2.1.1) was used to prove the following statement: if the probability density $f(x)$ is bell shaped about its median μ, that is, $f(\mu - x) = f(\mu + x)$ for $x \geqq 0$ and $f(\mu + x)$ is nonincreasing for $x \geqq 0$, then

$$P\{|S| > \lambda\} \leqq \binom{2m+1}{m-r}\binom{2r}{r}[\lambda(\lambda-1)]^{-r}\,2^{-(m+r)} \tag{2.2.1}$$

for $\lambda > 1$.

2.3. For every family of probability distributions with a location and scale parameter determined, say, by the probability density $f(x)$ for the standardized random variable X, $E(X) = 0$, $\mathrm{Var}\,(X) = 1$, expression (2.1.1) and a similar expression for $\lambda < 0$ could be used to compute numerically all probabilities needed for practical use. The evaluation of the triple integrals in (2.1.1), however, appears to be quite time consuming. Fortunately, the following simple intuitive argument makes it clear that for m large and r/m small the probability distribution of S will be practically independent of $f(x)$; for such values of m and r, all three order statistics $X_{(m+1-r)}$, $X_{(m+1)}$, $X_{(m+1+r)}$ fall with probability close to one very close to μ, where $F(x)$ is approximately equal to $1/2 + f(0)(x - \mu)$. Therefore S is approximately distributed as if the sample was obtained from a random variable with uniform distribution on $[\mu - 1/2f(0),\ \mu + 1/2f(0)]$, and since a change of location and of scale does not affect the probability distribution of S, it is approximately distributed as if the sample came from a random variable distributed uniformly on $[-1/2,\ 1/2]$.

3. Limiting distributions

3.1. THEOREM 3.1. *If $f(\mu) > 0$, and $f'(x)$ exists and is continuous in an interval $(\mu - \delta,\ \mu + \delta)$ for some $\delta > 0$ then, for r fixed, the probability distribution of the statistic* (1.3.2) *has the limit*

$$\lim_{m\to\infty} P\left\{\sqrt{\frac{2}{m}}\,S \leqq s\right\} = \frac{1}{(2r-1)!}\int_0^\infty \Phi(zs)z^{2r-1}e^{-z}\,dz \tag{3.1.1}$$

where $\Phi(\cdot)$ is the standardized normal distribution function.

An elementary proof of (3.1.1) appears in [3]. In view of the intuitive argument in Section 2.3, there is reason to believe that the right side of (3.1.1) is a good approximation, already for moderate values of m, when $f(x)$ is reasonably close to a linear function in a neighborhood of μ; by contrast, a good approximation need not be expected, for example, for $f(x) = \frac{1}{2}e^{-|x|}$.

3.2. When not only m, but also r is large, the following argument yields an even simpler result.

The right side of (3.1.1) is

$$R(s) = \int_0^\infty \Phi(zs)\,\gamma_{2r}(z)\,dz, \tag{3.2.1}$$

where $\gamma_{2r}(z)$ is the probability density of a gamma distribution with parameter $2r$, hence with mode $2r - 1$ and variance $2r$. The change of variable

$$y = \frac{z}{2r-1}. \tag{3.2.2}$$

yields

$$R(s) = \int_0^\infty \Phi[(2r-1)sy]\, \gamma^*_{2r}(y)\, dy, \tag{3.2.3}$$

where $\gamma^*_{2r}(y) = \gamma_{2r}[(2r-1)y]\,(2r-1)$ is a probability density with mode at $y = 1$ and variance $2r/(2r-1)^2$, hence with its mass concentrated about $y = 1$. One concludes from this by a routine argument that $R(s)$ is approximated by $\Phi[2r-1)s]$, hence

$$P\left\{\sqrt{\frac{2}{m}}\, S \leqq s\right\} \sim \Phi[(2r-1)s] \tag{3.2.4}$$

for r large and r/m small.

4. Some numerical computations and comparisons

4.1. One sided critical values for the limiting distribution (3.1.1) were calculated with the aid of a high speed computer by Richard Erickson [4] for the significance levels $\alpha = 0.10, 0.05, 0.025, 0.01, 0.005$, and $r = 1\ (1)10$, and for $\alpha = 0.001, 0.0005$ and selected values of r. It would be desirable to compute the corresponding critical values for a specific family of probability distributions, preferably the normal family, from the exact expression (2.1.1), and to determine how large m must be to make the difference between these exact critical values and those obtained from (3.1.1) negligible. As mentioned before, these calculations are quite time consuming, and have not yet been carried out.

4.2. Monte Carlo estimates of the exact probabilities (2.1.1) were obtained by Professor Jean Tague at Memorial University, St. John's, Newfoundland, for X normally distributed and the following values of the arguments: $m = 1(1)10$, r $= 1(1)m$, $\lambda = 0.0(0.1)5.0$ and some selected large values of λ. These Monte Carlo values [5] are the best information available at the present on the exact values of (2.1.1).

4.3. Let $\lambda_{\alpha,m,r}$ be the exact "critical value" determined by (2.1.1), under the assumption that X has normal distribution, so that for sample size $2m + 1$ one has $P\{S > \lambda_{\alpha,m,r}\} = \alpha$. Three different approximations to these exact values are available:

$\lambda^*_{\alpha,m,r}$ = the Monte Carlo estimates obtained by interpolation from [5],

$\lambda^{**}_{\alpha,m,r}$ = the approximations obtained in [4] from (3.1.1)

$\lambda^{***}_{\alpha,m,r}$ = the approximations obtained from (3.2.3).

From the manner in which these values have been obtained, one would expect that $\lambda^*_{\alpha,m,r}$ is generally a good approximation for $\lambda_{\alpha,m,r}$ while $\lambda^{**}_{\alpha,m,r}$ is good for m large, and $\lambda^{***}_{\alpha,m,r}$ only for r large and r/m small. Table I contains a comparison of the three approximations for $m = 10$ (sample size $2m + 1 = 21$), and $\alpha = 0.10$.

TABLE I

THREE APPROXIMATIONS TO VALUES $\lambda_{\alpha,m,r}$

For a sample of size $2m + 1$ from a normal population one has

$$P\left\{\frac{X_{(m+1)} - \mu}{X_{(m+1+r)} - X_{(m+1-r)}} > \lambda_{\alpha,m,r}\right\} = \alpha \quad \text{for} \quad m = 10.\ \alpha = 0.10.$$

r	λ^*	λ^{**}	λ^{***}
1	2.2045	1.5876	2.8657
2	.8537	.8745	.9552
3	.5223	.5335	.5731
4	.3765	.3937	.4094
5	.2904	.3113	.3184
6	.2381	.2563	.2605
7	.1919	.2173	.2204
8	.1658	.1887	.1910
9	.1386	.1668	.1686

REFERENCES

[1] Z. W. BIRNBAUM, "On a statistic similar to Student's t." *Nonparametric Techniques in Statistical Inferences*. London, Cambridge University Press, 1970, pp. 427–433.

[2] F. N. DAVID and N. L. JOHNSON, "Some tests of significance with ordered variables," *J. Roy. Statist. Soc. Ser. B*, Vol. 18 (1956), pp. 1–20.

[3] Z. W. BIRNBAUM and I. VINCZE, "Limiting distributions of statistics similar to Student's t." to be submitted for publication.

[4] R. A. ERICKSON, "On a Rényi type statistic for the median," unpublished M. S. thesis, University of Washington, 1970.

[5] JEAN TAGUE, "Monte Carlo tables for the S-statistic," unpublished, Memorial University of Newfoundland, 1969.