ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/222105754

Sample Quantiles in Statistical Packages

Article *in* The American Statistician · November 1996

DOI: 10.1080/00031305.1996.10473566

CITATIONS
883

READS
10,811

2 authors:

Rob J Hyndman
Monash University (Australia)
343 PUBLICATIONS **36,483** CITATIONS
SEE PROFILE

Yanan Fan
The Commonwealth Scientific and Industrial Research Organisation
81 PUBLICATIONS **3,594** CITATIONS
SEE PROFILE

All content following this page was uploaded by Rob J Hyndman on 25 February 2014.

The user has requested enhancement of the downloaded file.

Sample Quantiles in Statistical Packages

Rob J. Hyndman; Yanan Fan

The American Statistician, Vol. 50, No. 4. (Nov., 1996), pp. 361-365.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199611%2950%3A4%3C361%3ASQISP%3E2.0.CO%3B2-G

The American Statistician is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Mon Aug 20 16:35:04 2007

STATISTICAL COMPUTING

This department includes the two sections *New Development in Statistical Computing* and *Statistical Computing Software Reviews*; suitable contents for each of these sections are described under the respective section heading. Articles submitted for the department, outside the two sections, should not be highly technical and should be relevant to the teaching or practice of statistical computing.

Sample Quantiles in Statistical Packages

Rob J. HYNDMAN and Yanan FAN

There are a large number of different definitions used for sample quantiles in statistical computer packages. Often within the same package one definition will be used to compute a quantile explicitly, while other definitions may be used when producing a boxplot, a probability plot, or a QQ plot. We compare the most commonly implemented sample quantile definitions by writing them in a common notation and investigating their motivation and some of their properties. We argue that there is a need to adopt a standard definition for sample quantiles so that the same answers are produced by different packages and within each package. We conclude by recommending that the median-unbiased estimator be used because it has most of the desirable properties of a quantile estimator and can be defined independently of the underlying distribution.

KEY WORDS: Percentiles; Quartiles; Sample quantiles; Statistical computer packages.

1. INTRODUCTION

The quantile of a distribution is defined as

$$Q(p) = F^{-1}(p) = \inf\{x:\ F(x) \geq p\}, \qquad 0 < p < 1,$$

where $F(x)$ is the distribution function. Sample quantiles provide nonparametric estimators of their population counterparts based on a set of independent observations $\{X_1, \ldots, X_n\}$ from the distribution F. Let $\{X_{(1)}, \ldots, X_{(n)}\}$ denote the order statistics of $\{X_1, \ldots, X_n\}$, and let $\hat{Q}_i(p)$ denote the ith sample quantile definition.

One difficulty in comparing quantile definitions is that there is a number of equivalent ways of defining them. However, the sample quantiles that are used in statistical packages are all based on one or two order statistics, and can be written as

$$\hat{Q}_i(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)} \quad \text{where} \quad \frac{j-m}{n} \leq p < \frac{j-m+1}{n} \tag{1}$$

for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$. The value of γ is a function of $j = \lfloor pn+m \rfloor$ and $g = pn+m-j$. Here, $\lfloor u \rfloor$ denotes the largest integer not greater than u; later we shall use $\lceil u \rceil$ to denote the smallest integer not less than u.

We consider estimators of the form (1), including some that are not found in statistical packages. There have been several other nonparametric quantile estimators proposed that are not of the form (1) (e.g., Harrell and Davis 1982; Sheather and Marron 1990), but these are not implemented in widely available packages and so are not considered here. We also exclude sample quantiles that are not defined for all p including hinges and other letter values (Hoaglin 1983) and related methods (Freund and Perles 1987).

A closely related problem is the selection of plotting position in a quantile plot in which $X_{(k)}$ is plotted against p_k or in a quantile–quantile plot in which $X_{(k)}$ is plotted against $G^{-1}(p_k)$ where G is a distribution function. Various rules for p_k have been suggested (see Cunnane 1978; Harter 1984; Kimball 1960; Mage 1982). Each plotting rule corresponds to a sample quantile definition by defining $\hat{Q}_i(p_k) = X_{(k)}$ and using linear interpolation for $p \neq p_k$. However, the criteria by which a plotting position is chosen (e.g., the five postulates of Gumbel 1958, pp. 32–34 or the three purposes of Kimball 1960) may be quite different from the criteria for choosing a good sample quantile definition.

We compare sample quantile definitions of the form (1) by describing their motivation and whether or not they pos-

Table 1. Six Desirable Properties for a Sample Quantile

P1:	$\hat{Q}_i(p)$ is continuous.
P2:	$\text{Freq}(X_k \leq \hat{Q}_i(p)) \geq pn$.
P3:	$\text{Freq}(X_k \leq \hat{Q}_i(p)) = \text{Freq}(X_k \geq \hat{Q}_i(1-p))$.
P4:	Where $\hat{Q}_i^{-1}(x)$ is uniquely defined, $\hat{Q}_i^{-1}(X_{(k)}) + \hat{Q}_i^{-1}(X_{(n-k+1)}) = 1 \quad \text{for} \quad k = 1, \ldots, n.$
P5:	Where $\hat{Q}_i^{-1}(x)$ is uniquely defined, $\hat{Q}_i^{-1}(X_{(1)}) > 0$ and $\hat{Q}_i^{-1}(X_{(n)}) < 1$.
P6:	$\hat{Q}_i(.5)$ is equal to the sample median defined by $[X_{(l)} + X_{(l+1)}]/2$ if $n = 2l$; $X_{(l+1)}$ if $n = 2l+1$.

Rob J. Hyndman is Lecturer, Department of Mathematics, Monash University, Clayton, Vict., Australia 3168. Yanan Fan is Editor, World Scientific Publishing Co. Pte Ltd., 1022 Tai Seng Ave., #05-3520 Tai Seng Industrial Estate, Singapore 534415. The authors thank Dr. Jane Matthews, Kally Yuen, Vicky Ryan, and Tony Wohlers for letting them (and helping them) use their packages.

© 1996 American Statistical Association

sess the six properties shown in Table 1. (The notation $\text{Freq}(X_k \leq x)$ denotes the number of observations less than or equal to x.)

Property P1 is based on the common assumption that $Q(p)$ is a continuous function of p. Property P2 is the sample analog of the result $F(Q(u)) \geq u$ (with equality when F is continuous). Properties P3 and P4 are symmetry properties that require that the tails of the underlying distribution are treated equally. P3 is equivalent to Freund and Perles' (1987) criterion B for quartiles. Property P5 reflects the result that for a continuous distribution, we expect there to be positive probability for values beyond the range of the data. Property P6 is sensible given the widespread use of the sample median.

2. DISCONTINUOUS FUNCTIONS

Definition 1. The oldest and most studied definition is the inverse of the empirical distribution function obtained by setting $m = 0$ and

$$\gamma = \begin{cases} 1 & \text{if } g > 0 \\ 0 & \text{if } g = 0. \end{cases}$$

This is the step function shown schematically in Figure 1. The value of the function at each jump is shown as a solid point (•). For this definition

$$\text{Freq}(X_k \leq \hat{Q}_1(p)) = \lceil pn \rceil$$

and

$$\text{Freq}(X_k \geq \hat{Q}_1(1-p)) = \lfloor pn + 1 \rfloor.$$

Definition 2. $\hat{Q}_2(p)$ is similar to $\hat{Q}_1(p)$ except that averaging is used when $g = 0$. Hence $m = 0$,

$$\gamma = \begin{cases} \frac{1}{2} & g = 0 \\ 1 & g > 0 \end{cases}$$

and

$$\text{Freq}(X_k \leq \hat{Q}_2(p)) = \text{Freq}(X_k \geq \hat{Q}_2(1-p)) = \lceil pn \rceil.$$

$\hat{Q}_2(p)$ is shown in Figure 2.

Definition 3. $\hat{Q}_3(p)$ is defined as the order statistic $X_{(k)}$ where k is the nearest integer to np. So we set $m = -\frac{1}{2}$ and,

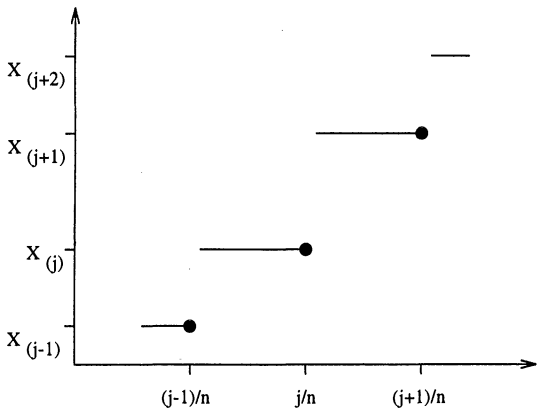

Figure 1. Schematic Representation of $\hat{Q}_1(p)$.

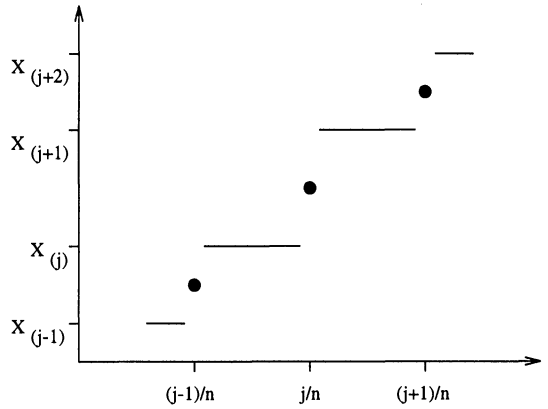

Figure 2. Schematic Representation of $\hat{Q}_2(p)$.

when $g > 0$, let $\gamma = 1$. At $g = 0$ there is more than one way to define "nearest." One approach, which is implemented in SAS, is to choose the nearest even order statistic at $g = 0$. Hence

$$\gamma = 0 \quad \text{if } g = 0 \text{ and } j \text{ even}$$

and

$$\gamma = 1 \quad \text{otherwise.}$$

For this definition

$$\text{Freq}(X_k \leq \hat{Q}_3(p)) = \begin{cases} \lfloor pn \rfloor & \text{if } g = 0 \text{ and } \lfloor pn \rfloor \text{ even} \\ \lfloor pn + \frac{1}{2} \rfloor & \text{otherwise} \end{cases}$$

and

$$\text{Freq}(X_k \geq \hat{Q}_3(1-p)) = \begin{cases} \lceil pn + \frac{1}{2} \rceil & \text{if } g = 0 \\ & \text{and } \lfloor (1-p)n \rfloor \text{ even} \\ \lceil pn + 1 \rceil & \text{otherwise.} \end{cases}$$

Figure 3 shows $\hat{Q}_3(p)$. We summarize the properties of these sample quantile definitions in Table 2.

3. PIECEWISE LINEAR CONTINUOUS FUNCTIONS

The related problem of selecting a plotting position when plotting quantiles leads to a number of sample quantile definitions constructed by linearly interpolating between plot-

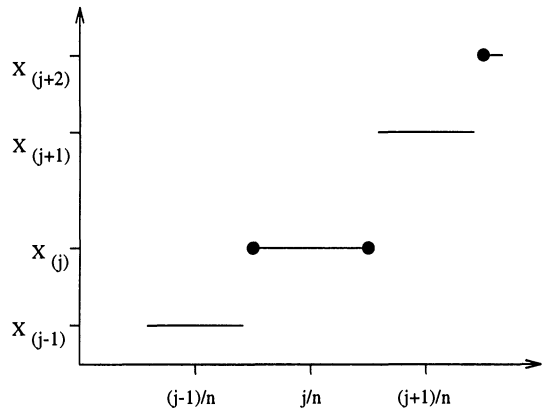

Figure 3. Schematic Representation of $\hat{Q}_3(p)$ (j Even).

Table 2. Summary of Properties of $\hat{Q}_1(p)$, $\hat{Q}_2(p)$, and $\hat{Q}_3(p)$

Definition	P1	P2	P3	P4	P5	P6
1		√				√ (n odd)
2		√	√			√
3		√				

ting positions. Blom (1958) considered the plotting position

$$p_k = \frac{k-\alpha}{n-\alpha-\beta+1},$$

where α and β are constants, which includes all the usual plotting positions that are advocated. Interpolating between the points $(p_k, X_{(k)})$ gives a sample quantile of the form (1) with $m = \alpha + p(1-\alpha-\beta)$ and $\gamma = g$. Harter (1984) provides a review of the various plotting positions that have been proposed. One example is shown in Figure 4 for $\alpha = 0$ and $\beta = 1$ so that $p_k = k/n$. This is an interpolation of the step function of Definition 1.

Of course, P1 is satisfied for all such definitions. Also,

$$\begin{aligned}\text{Freq}(X_k \le \hat{Q}_i(p)) &= \lfloor pn + m \rfloor \\ &= \lfloor pn + \alpha + p(1-\alpha-\beta) \rfloor\end{aligned}$$

and

$$\begin{aligned}\text{Freq}(X_k \ge \hat{Q}_i(1-p)) &= \lfloor pn - m + 1 \rfloor \\ &= \lfloor pn - \alpha - p(1-\alpha-\beta) + 1 \rfloor.\end{aligned}$$

Hence P2 is satisfied for all p if and only if $\alpha + \beta \le 1$ and $\alpha > 0$, and P3 is satisfied for all p if and only if $\alpha = \beta = \frac{1}{2}$, in which case $m = \frac{1}{2}$.

For P4 to hold we require $p_k + p_{n-k+1} = 1$, and so $\alpha = \beta$, and for P5 we need $\alpha < 1$ and $\beta < 1$.

If $n = 2l$,

$$\hat{Q}_i(.5) = (1-\gamma)X_{(\lfloor l+m \rfloor)} + \gamma X_{(\lfloor l+m+1 \rfloor)},$$

where $\gamma = m - \lfloor m \rfloor$. So for even n, P6 is satisfied if and only if $m = \frac{1}{2}$ when $p = \frac{1}{2}$, which occurs when $\alpha = \beta$. If $n = 2l + 1$,

$$\hat{Q}_i(.5) = (1-\gamma)X_{(\lfloor l+m+1/2 \rfloor)} + \gamma X_{(\lfloor l+m+3/2 \rfloor)}$$

where $\gamma = \frac{1}{2} + m - \lfloor \frac{1}{2} + m \rfloor$. So for odd n, P6 is satisfied if and only if $m = \frac{1}{2}$ when $p = \frac{1}{2}$. So again we need $\alpha = \beta$.

Definition 4. Parzen (1979) suggested defining a sample quantile by interpolating the step function of Definition 1 as shown in Figure 4. This amounts to $p_k = k/n$.

Definition 5. A very old definition, proposed by Hazen (1914) and popular among hydrologists, is based on $p_k = (k - \frac{1}{2})/n$. This is the value midway through each step of Definition 1.

The remaining definitions are derived on the basis of estimation arguments. Let L be some measure of location such as the mean, median, or mode. There are two classes of quantile definitions that are derived using estimation arguments. The first approach chooses $p_k = \mathsf{L}F(X_{(k)})$ and the second approach chooses $p_k = F(\mathsf{L}X_{(k)})$. If F is the uniform distribution, the two approaches are equivalent. Also, if L denotes the median and F is strictly monotonic, the two approaches are equivalent because the median is invariant under monotonic transformation.

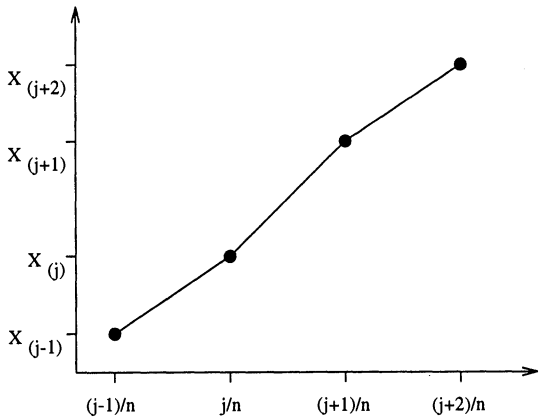

Figure 4. Schematic Representation of $\hat{Q}_4(p)$.

Following the first approach, note that $F(X_k)$ has a uniform distribution so $F(X_{(k)})$ has the same distribution as the kth-order statistic from a uniform distribution, namely the beta distribution $\beta(k, n-k+1)$. Hence this approach is distribution-free in the sense that the resulting plotting positions do not depend on the distribution F. Definitions $\hat{Q}_6(p)$, $\hat{Q}_7(p)$, and $\hat{Q}_8(p)$ can be derived in this way.

Definition $\hat{Q}_9(p)$ is derived following the second approach (and because $\hat{Q}_8(p)$ uses L = median, it can also be derived following the second approach). Note that definitions derived in this way are L-unbiased because

$$Q(p_k) = Q(F(\mathsf{L}X_{(k)})) = \mathsf{L}X_{(k)} = \mathsf{L}\hat{Q}_i(p_k).$$

However, these definitions are not distribution-free because different values of p_k result for different distributions F.

Clearly, only $\hat{Q}_8(p)$ is both L-unbiased and distribution-free.

Definition 6. Weibull (1939) and Gumbel (1939) proposed $p_k = k/(n+1)$. In this case $p_k = \mathsf{E}F(X_{(k)})$ and the vertices divide the sample space into $n+1$ regions, each with probability $1/(n+1)$ on average. In particular, $\Pr(X < X_{(1)}) = \Pr(X > X_{(n)}) = 1/(n+1)$.

Definition 7. Gumbel (1939) also considered the modal position $p_k = \mathsf{mode}F(X_{(k)}) = (k-1)/(n-1)$. One nice property is that the vertices of $\hat{Q}_7(p)$ divide the range into $n-1$ intervals, and exactly $100p\%$ of the intervals lie to the left of $\hat{Q}_7(p)$ and $100(1-p)\%$ of the intervals lie to the right of $\hat{Q}_7(p)$.

Definition 8. The median position, $\mathsf{M}F(X_{(k)})$, where M denotes the median, is more difficult to obtain. Using an approximation to the incomplete beta function ratio (Johnson and Kotz 1970, p. 48) we find $\mathsf{M}F(X_{(k)}) \approx (k - \frac{1}{3})/(n + \frac{1}{3})$. Therefore, we define the sample quantile by setting $p_k = (k - \frac{1}{3})/(n + \frac{1}{3})$.

In fact, the resulting sample quantile is median unbiased of order $o(n^{-1/2})$ (Reiss 1989). Reiss also states that the resulting sample quantile is optimal in the class of all esti-

Table 3. Summary of Properties of $\hat{Q}_i(p)$, $i = 4, \ldots, 9$

Definition	α	β	m	P1	P2	P3	P4	P5	P6
4	0	1	0	√	√				
5	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	√	√	√	√	√	√
6	0	0	p	√	√		√	√	√
7	1	1	$1-p$	√	√		√		√
8	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}(p+1)$	√	√		√	√	√
9	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}p+\frac{3}{8}$	√	√		√	√	√

mators that are median unbiased of order $o(n^{-1/2})$ and equivariant under translations (shifting the observations amounts to shifting the distribution of $\hat{Q}_i(p)$).

Hoaglin (1983) shows that when p is an integer multiple of $(.5)^l$ where l is an integer, $\hat{Q}_8(p)$ gives approximately the same results as "letter values."

Benard and Bos-Levenbach (1953) also argue for $p_k = \mathrm{M}F(X_{(k)})$, but use the approximation $p_k = (k - .3)/(n + .4)$.

Definition 9. Blom (1958) shows that $p_k = (k - \frac{3}{8})/(n + \frac{1}{4})$ gives a better approximation to $F(\mathrm{E}X_{(k)})$ for the normal distribution. Therefore, $\hat{Q}_9(p_k)$ is an approximately unbiased estimate of $Q(p_k)$ when F is normal. Because this definition is distribution-dependent, it tends to be used for normal QQ plots rather than as a general sample quantile definition. Analogous p_k for other distributions are listed in Cunnane (1978).

We summarize the properties of these definitions in Table 3.

4. STATISTICAL PACKAGES

In this section we summarize the sample quantile definitions that are implemented in some major statistical packages. Note that we only consider commands that compute quantiles explicitly, and we ignore implicit quantile definitions that are used in probability plots, quantile–quantile plots, and boxplots. Often a package will use a different definition of sample quantile in a plot from what is used when explicitly computing quantiles.

BMDP: Since the 1990 release of BMDP, quartiles in BMDP 2D have been computed using $\hat{Q}_6(p)$. More general quantiles cannot be computed. Note that the manual (BMDP 1992) incorrectly describes the method of computing quartiles.

GLIM: The `$tab percentile` command of GLIM V3.77 gives $\hat{Q}_2(p)$, while `$tab interpolate` gives $\hat{Q}_5(p)$ (GLIM 1987).

Minitab: The `DESCRIBE` command computes quartiles using $\hat{Q}_6(p)$ (Minitab 1994). The quartiles produced by the experimental command `%DESCRIBE` are not documented, but numerical experiments suggest that $\hat{Q}_2(p)$ is used. Other quantiles are not available.

SAS: `PROC UNIVARIATE` allows five different quantile definitions (SAS 1990): $\hat{Q}_1(p), \hat{Q}_2(p), \hat{Q}_3(p), \hat{Q}_4(p)$, and $\hat{Q}_6(p)$.

Splus: The `quantile( )` command of Splus 3.1 uses $\hat{Q}_7(p)$ (although S-PLUS (1991) states that $\hat{Q}_5(p)$ is used).

SPSS: The `frequencies` command of SPSS appears to use $\hat{Q}_6(p)$, although this is nowhere documented.

5. SUMMARY AND CONCLUSIONS

Only $\hat{Q}_5(p)$ satisfies all six properties, P1–P6. However, it is a compromise definition in the sense that it is derived by interpolating between the midpoints of the inverse of the distribution function. It is not justified on the basis of an estimation argument. Definitions $\hat{Q}_6(p)$–$\hat{Q}_9(p)$ each satisfy five of the six properties, and their derivations are more easily justified. Of these, $\hat{Q}_8(p)$ seems the best because it gives (approximately) median-unbiased estimates of $Q(p)$ regardless of the distribution, F. Both $\hat{Q}_6(p)$ and $\hat{Q}_7(p)$ are also distribution-free, but they are not unbiased, whereas $\hat{Q}_9(p)$ is approximately unbiased for the normal distribution, but not for other distributions.

The current variation in sample quantile definitions causes confusion, and so there is a need to standardize the definition of sample quantile across packages and within packages. This is an analogous situation to the problem of defining sample variance. In that case the statistical community has adopted the unbiased definition (with denominator $n-1$) as the standard rather than the more intuitive average of squared deviations (with denominator n) or the minimum MSE definition (with denominator $n+1$ for a normal distribution). This avoids confusion and ensures comparable results on all software. We believe there is a similar need to adopt a standard sample quantile definition, and we propose that $\hat{Q}_8(p)$ is the best choice.

[Received March 1995. Revised March 1996.]

REFERENCES

Benard, A., and Bos-Levenbach, E. C. (1953), "Het Uitzetten van Waarnemingen op Waarschijnlijkheitspapier," *Statistica*, 7, 163–173.

Blom, G. (1958), *Statistical Estimates and Transformed Beta-Variables*, New York: John Wiley.

BMDP (1992), *BMDP Statistical Software Manual Release 7*, BMDP Statistical Software Inc.

Cunnane, C. (1978), "Unbiased Plotting Positions—A Review," *Journal of Hydrology*, 37, 205–222.

Freund, J. E., and Perles, B. M. (1987), "A New Look at Quartiles of Ungrouped Data," *The American Statistician*, 41, 200–203.

GLIM (1987), *The GLIM System Release 3.77 Manual* (2nd ed.).

Gumbel, E. J. (1939), "La Probabilité des Hypothèses," *Comptes Rendus de l'Académie des Sciences (Paris)*, 209, 645–647.

——— (1958), *Statistics of Extremes*, New York: Columbia University Press.

Harrell, F. E., and Davis, C. E. (1982), "A New Distribution-Free Quantile Estimator," *Biometrika*, 69, 635–640.

Harter, H. L. (1984), "Another Look at Plotting Positions," *Communications in Statistics, Theory and Methods*, 13, 1613–1633.

Hazen, A. (1914), "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply" (with discussion), *Transactions of the American Society of Civil Engineers*, 77, 1539–1669.

Hoaglin, D. C. (1983), "Letter Values: A Set of Selected Order Statistics," in *Understanding Robust and Exploratory Data Analysis*, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: John Wiley.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—2*, Boston: Houghton Mifflin.

Kimball, B. F. (1960), "On the Choice of Plotting Positions on Probability Paper," *Journal of the American Statistical Association*, 55, 546–560.

Mage, D. (1982), "An Objective Graphical Method for Testing Normal Distributional Assumptions Using Probability Plots," *The American Statistician*, 36, 116–120.

Minitab (July 1994), *Minitab Reference Manual: Release 10 for Windows*, Minitab Inc.

Parzen, E. (1979), "Nonparametric Statistical Data Modeling" (with discussion), *Journal of the American Statistical Association*, 74, 105–131.

Reiss, R. D. (1989), *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics*, New York: Springer-Verlag.

S-PLUS (1991), *S-PLUS Reference Manual version 3.0,* Statistical Sciences, Inc., Seattle, WA.

SAS (1990), *SAS Procedures Guide*, Version 6 (3rd ed.), Cary, NC: SAS Institute Inc.

Sheather, S. J., and Marron, J. S. (1990), "Kernel Quantile Estimators," *Journal of the American Statistical Association*, 85, 410–416.

Weibull, W. (1939), "The Phenomenon of Rupture in Solids," *Ingeniörs Vetenskaps Akademien Handlingar*, 153, 17.

A Note on the Calculation of $\Pr\{X_1 < X_2 < \cdots < X_k\}$

A. J. HAYTER and W. LIU

Suppose that X_i are independent random variables, and that X_i has cdf $F_i(x), 1 \le i \le k$. Many statistical problems involve the probability $\Pr\{X_1 < X_2 < \cdots < X_k\}$. In this note a numerical method is proposed for computing this probability.

KEY WORDS: Multivariate probability; Statistical computing.

Suppose that X_i are independent random variables (discrete or continuous), and that X_i has cdf $F_i(x), 1 \le i \le k$. We want to compute the probability $P_k = \Pr\{X_1 < X_2 < \cdots < X_k\}$. In isotonic regression it is essential to find P_k in order to find the level probabilities $P(l, k)$. See, for example, Robertson, Wright, and Dykstra (1988 pp. 74–77). In ranking and selection problems the probability of correct ranking can be expressed in terms of P_k. See, for example, Bechhofer (1954). The size and power of a multiple comparison test proposed by Hochberg and Marcus (1978) also depend on P_k. Generally speaking, P_k can be computed exactly for small values of $k, k \le 5$ say, by using repeated numerical integrations or summations. For large values of k, however, this approach becomes infeasible, and analytic approximations tend to be used; see Gupta (1963) and the references therein when X_i are normal random variables.

Our numerical method relies on the following simple recursive relationship. Define

$$r_1(x) = \Pr\{X_1 < x\} = F_1(x)$$

$$r_l(x) = \Pr\{X_1 < X_2 < \cdots < X_l < x\}, \qquad l \ge 1.$$

Then it is easy to see that

$$r_l(x) = \int_{(-\infty, x)} r_{l-1}(y)\, dF_l(y), \qquad l \ge 2 \qquad (1)$$

and

$$P_k = \int_{-\infty}^{\infty} r_{k-1}(y)\, dF_k(y). \qquad (2)$$

The function $r_l(x)$ can thus be calculated recursively by (1). The basic method is to evaluate the right-hand side of (1) at points on a chosen grid; $r_l(x), x \in R$ is then approximated by proper interpolation and extrapolation. Continuing in this way until $r_{k-1}(x)$ is found, P_k can then be computed from (2). In this process most of the computing time is spent on the recursive calculations of $r_l(x)$, which involves only one-dimensional numerical integrations or summations. The computing intensity therefore increases about linearly in k. We have experimented with independent normal random variables X_i by using linear approximations, several different grids, and NAG library routine for one-dimensional numerical integrations; on a Silicon Graphics Indigo 2 it took less than one minute to compute one P_{50} accurate to the third decimal place.

In Hayter and Liu (1996) we have used a similar recursive method to compute the critical points and power of the one-sided studentised range test of Hayter (1990). In fact, such recursive computing methods are frequently used in sequential analysis. See, for example, Eales and Jennison (1992). It seems that some other probabilities that have defied exact calculations so far can be calculated similarly.

A. J. Hayter is Professor of Industrial Engineering, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205. W. Liu is Lecturer in Statistics, Department of Mathematics, University of Southampton, Southampton S017 1BJ, England.

REFERENCES

Bechhofer, R. E. (1954), "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances," *The Annals of Mathematical Statistics*, 25, 16–39.

Eales, J. D., and Jennison, C. (1992), "An Improved Method for Deriving Optimal One-Sided Group Sequential Tests," *Biometrika*, 79, 13–24.

Gupta, S. S. (1963), "Probability Integrals of Multivariate Normal and Multivariate *t*," *The Annals of Mathematical Statistics*, 34, 792–828.

Hayter, A. J. (1990), "A One-Sided Studentized Range Test for Testing Against a Simple Ordered Alternative," *Journal of the American Statistical Association*, 85, 778–785.

Hayter, A. J., and Liu, W. (1996), "On the Exact Calculation of a One-Sided Studentised Range Test Against a Simple Ordered Alternative," *Computational Statistics and Data Analysis*, 22, 17–25.

Hochberg, Y., and Marcus, R. (1978), "On Partitioning Successive Increments in Means or Ratios of Variances in a Chain of Normal Populations," *Communications in Statistics*, Ser. A, 7, 1501–1513.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: John Wiley.

© 1996 American Statistical Association

View publication stats