Published in final edited form as:

Epidemiology. 2011 September ; 22(5): 750–751. doi:10.1097/EDE.0b013e318225c1de.

Estimation of population percentiles

Frank Schoonjans^{1,2}, Dirk De Bacquer³, and Pirmin Schmid⁴

¹Department of Endocrinology, Ghent University Hospital, Belgium

²MedCalc Software bvba, Mariakerke, Belgium

³Department of Public Health, Ghent University, Belgium

⁴Department of Transfusion Medicine, Clinical Center, National Institutes of Health, Bethesda, MD

Percentiles play an important part in descriptive statistics of continuous data, and their use is recommended for reference interval estimation.¹ We have selected various methods for the calculation of percentiles based on recommendations in the literature or use in popular software, and evaluated the accuracy of the percentile calculated in the sample as an estimate of the true population percentile² using Monte-Carlo techniques.

All selected methods calculate a rank or an index that points to a number in the sorted array of sample data, and linear interpolation is applied when the index does not correspond to an integer value. One method (method A^{1,3}) calculates a rank or an index $p(n+1)$ with p representing the centile (which is the percentile divided by 100) and n the sample size. Method B⁴ calculates an index $0.5+pn$. Method C⁵ (commonly used in spreadsheets) uses $p(n-1)+1$ and method D⁶ uses $p(n+1/3)+1/3$. Details of the use of these 4 methods are given in the eAppendix (<http://links.lww.com/EDE/A488>).

Experimental population data were obtained using a normal distribution pseudo-random number generator, programmed to generate a data set of 10^6 numbers with mean 0 and standard deviation 1.

From our population data, 6 sets of 100 000 random samples each were drawn using a pseudo-random number generator with uniform distribution. Each of these sets consisted of 100 000 random samples with sample size 20, 120, 500 and 1000. The average of the 5th and 95th percentiles obtained with the 4 methods in these sample sets were calculated and compared with the population values. For each sample, the relative difference with the population values was expressed as a percentage, and the mean and standard deviation of these percentages were calculated.

Next, the population data were transformed exponentially (base 10) to obtain a log-normal distribution and the experiments as described above were repeated.

The results for 95th percentile in the normally distributed data are represented in table 1 (more comprehensive tables with figures are available in the eAppendix, <http://links.lww.com/EDE/A488>). The results for the 5th percentile were symmetrical to the

Correspondence: Frank Schoonjans, University Hospital Ghent, Department of Endocrinology, De Pintelaan 185, 9000 Gent, Belgium, Phone: +32 9 3324516, Fax: +32 9 3323886, frank.schoonjans@ugent.be.

SDC: Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidemn.com)

results for the 95th percentile and are not shown. Method B presents the highest accuracy, followed by method D, A and C.

The results for 5th and 95th percentile in the log-normally distributed data are represented in the Table. For the 5th percentile, method A has a higher accuracy than the methods D, B and C, especially in small sample sizes, whereas for the 95th percentile method C presents the highest accuracy, followed by method B, D and A.

We find that, for the calculation of percentiles, it may still be advantageous to transform log-normally distributed data. For example, the 95th percentile in the log-normal data should be about 44.16 ($= 10^{1.65}$). With method B and $n = 20$ we find an average of 114.69. But if we first log-transform the data we find, on average, 1.64, which back-transforms to $10^{1.64}$ or 43.65, which is much closer to the true population value of 44.16. The effect of the log-transformation may be explained by the fact that linear interpolation is applied in the calculations of percentiles, and the transformation changes the distribution model within the interpolated interval.

We conclude that method B is the preferred method in general for continuous data, taking into account the recommendation to transform the data to a normal distribution if necessary.

Finally the large standard deviations of the observed differences illustrate the large statistical uncertainty associated with the estimated percentiles in small sample sizes. Therefore we stress the importance of reporting percentiles with their 95% confidence interval.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support: Pirmin Schmid is supported by a fellowship of the National Institutes of Health, Bethesda, MD.

References

1. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory: approved guideline - third edition. CLSI Document C28-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2008.
2. Walter S. Problems with percentiles. *Int J Epidemiol.* 1986; 15:431–532. [PubMed: 3771088]
3. Altman, DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
4. Armitage, P.; Berry, G.; Matthews, JNS. Statistical methods in medical research. 4th. Oxford: Blackwell Science; 2002.
5. Gumbel EJ. La Probabilité des Hypothèses. *Comptes Rendus de l'Académie des Sciences (Paris).* 1939; 209:645–647.
6. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. *Am Stat.* 1996; 50:361–365.

Table 1

Accuracy of percentiles calculated in samples with various sizes from normally and log-normally distributed population data. Results are presented as average percentile, the average of the relative differences (%) and their standard deviation (SD).

Method	Sample size	Normal distribution		Log-normal distribution			
		95 th percentile		5 th percentile		95 th percentile	
		Average	% Difference (SD)	Average	% Difference (SD)	Average	% Difference (SD)
A: p(n+1)	Population (n=10 ⁶)	1.64		0.02		44.16	
	20	1.85	12.20 (31.06)	0.03	18.92 (141.23)	180.43	308.61 (1766.28)
	120	1.68	1.85 (11.88)	0.02	2.97 (46.85)	52.76	19.49 (59.86)
	500	1.65	0.47 (5.78)	0.02	0.82 (22.28)	46.01	4.20 (23.12)
	1000	1.65	0.24 (4.07)	0.02	0.42 (15.69)	45.13	2.21 (16.00)
B: pn+0.5	Population (n=10 ⁶)	1.64		0.02		44.16	
	20	1.64	-0.56 (25.48)	0.04	83.31 (179.54)	114.69	159.73 (973.07)
	120	1.64	-0.33 (11.37)	0.03	11.74 (48.61)	48.43	9.69 (52.26)
	500	1.64	-0.07 (5.72)	0.02	2.77 (22.48)	45.16	2.27 (22.52)
	1000	1.64	-0.02 (4.05)	0.02	1.44 (15.76)	44.64	1.09 (15.70)
C: p(n-1)+1	Population (n=10 ⁶)	1.64		0.02		44.16	
	20	1.43	-12.97 (24.17)	0.06	147.81 (244.30)	48.07	8.87 (174.14)
	120	1.60	-2.50 (11.39)	0.03	20.77 (52.38)	44.18	0.05 (47.04)
	500	1.64	-0.59 (5.71)	0.02	4.82 (22.90)	44.24	0.19 (22.16)
	1000	1.64	-0.28 (4.06)	0.02	2.46 (15.92)	44.21	0.12 (15.60)
D: p(n+1/3)+1/3	Population (n=10 ⁶)	1.64		0.02		44.16	
	20	1.71	3.80 (27.05)	0.04	61.98 (163.85)	138.42	213.47 (1367.82)
	120	1.65	0.44 (11.52)	0.02	8.55 (47.85)	49.96	13.15 (54.94)
	500	1.65	0.12 (5.69)	0.02	2.15 (22.37)	45.44	2.91 (22.72)
	1000	1.65	0.07 (4.07)	0.02	1.07 (15.72)	44.79	1.43 (15.79)