Visualizing multiple quantile plots

Marko A.A. Boon
Eindhoven University of Technology

John H.J. Einmahl
Tilburg University

Ian W. McKeague
Columbia University

January 18, 2012

Abstract. Multiple quantile plots provide a powerful graphical method for comparing the distributions of two or more populations. This paper develops a method of visualizing triple quantile plots and their associated confidence tubes, thus extending the notion of a QQ plot to three dimensions. More specifically, we consider three independent one-dimensional random samples with corresponding quantile functions Q_1, Q_2 and Q_3, respectively. The triple quantile (QQQ) plot is then defined as the three-dimensional curve $\mathbf{Q}(p) = (Q_1(p), Q_2(p), Q_3(p))$, where $0 < p < 1$. The empirical likelihood method is used to derive simultaneous distribution-free confidence tubes for $\mathbf{Q}$. We apply our method to an economic case study of strike durations, and to an epidemiological study involving the comparison of cholesterol levels among three populations. These data as well as the Mathematica code for computation of the tubes are available online.

Key words: Confidence region, empirical likelihood, quantile plot, three-sample comparison.

1 Introduction

The quantile-quantile (QQ) plot is a well-known and attractive graphical method for comparing two distributions, especially when confidence bands are included. Frequently in applications, however, there is a need to simultaneously compare more than two distributions, and it would be useful to have a readily available graphical method to do this. In the present paper we develop a way of visualizing triple quantile plots and their associated confidence tubes, thus extending the notion of a QQ plot to three dimensions.

Our approach is based on the nonparametric empirical likelihood method. There exists a large literature on empirical likelihood indicating that it is widely viewed as a desirable and natural approach to statistical inference in a variety of settings. Moreover, there is considerable evidence that procedures based on the method outperform competing procedures in terms of accuracy; see the monograph of Owen (2001) for numerous examples. Empirical likelihood based confidence bands for individual quantile functions have been derived in Li et al. (1996). Confidence tubes for multiple quantile plots under random censoring have been studied in Einmahl and McKeague (1999). In the present paper we employ a direct approach (that is only feasible in the non-censored situation), and we focus on the problem of how to provide a 3D-visualization of the empirical QQQ plots and the corresponding confidence tubes. The confidence tubes are presented in Section 2; they are valid under minimal conditions. The procedure is applied to data on strike durations and cholesterol levels in Sections 3 and 4, respectively.

QQ plots have been studied in detail using classical methods in Doksum (1974, 1977), Switzer (1976), and Doksum and Sievers (1976). The k-sample problem is studied in Nair (1978, 1982), but there essentially only pairwise comparisons are made. A review of graphical methods in nonparametric statistics with extensive

coverage of QQ plots can be found in Fisher (1983). Some refined approximation results for normalized QQ plots with statistical applications have been established in Beirlant and Deheuvels (1990). More recently, QQ plots for univariate and multivariate data have been studied in Marden (2004) and refined QQ plots in the generalized linear model have been considered in García Ben and Yohai (2004).

2 The confidence tubes

It is convenient first to set the notation in the one-sample case. For the corresponding notation in the three-sample case, we *add a further subscript j to refer to the j-th sample*. The distribution function of the $X_i, i = 1, \ldots, n$, is denoted by F and the corresponding (right-continuous) quantile function is denoted by Q. We write

$$L(\tilde{F}) = \prod_{i=1}^{n} (\tilde{F}(X_i) - \tilde{F}(X_i-))$$

for the likelihood, where $\tilde{F}$ belongs to $\mathcal{F}$, the space of all probability distribution functions on $\mathbb{R}$. The empirical likelihood ratio for $\tilde{F}(t) = p$ (for a given $p \in (0,1)$) is defined by

$$R_p(t) = \frac{\sup\{L(\tilde{F}) : \tilde{F}(t) = p, \tilde{F} \in \mathcal{F}\}}{\sup\{L(\tilde{F}) : \tilde{F} \in \mathcal{F}\}}.$$

Note that the supremum in the denominator is attained by the empirical distribution function

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,t]}(X_i);$$

hence the value of this supremum is n^{-n}. It easily follows by putting total probability mass p on the data less than or equal to t (for the numerator) that

$$R_p(t) = \left(\frac{p}{F_n(t)}\right)^{nF_n(t)} \left(\frac{1-p}{1-F_n(t)}\right)^{n(1-F_n(t))}. \tag{2.1}$$

Let Q_n be the empirical quantile function.

Now we turn to the three-sample setup. The three random samples are assumed to be independent with sample sizes denoted n_1, n_2, n_3; write $m = n_1 + n_2 + n_3$. Set $\mathbb{F} = (F_1, F_2, F_3)$ and define the QQQ plot to be

$$\{(Q_1(p), Q_2(p), Q_3(p)) : p \in (0,1)\}.$$

This plot can be estimated with

$$\{(Q_{1n_1}(p), Q_{2n_2}(p), Q_{3n_3}(p)) : p \in (0,1)\},$$

the empirical QQQ plot. Observe that these are extensions of the classical two-sample QQ plots. In the sequel we consider the following more convenient version of the QQQ plot: the graph $\mathbf{Q}$ of the function

$$t_1 \mapsto (Q_2(F_1(t_1)), Q_3(F_1(t_1))),$$

for $t_1 \in \mathbb{R}$. Denote the joint likelihood by

$$L(\tilde{\mathbb{F}}) = L_1(\tilde{F}_1)L_2(\tilde{F}_2)L_3(\tilde{F}_3),$$

and the empirical likelihood ratio at $\mathbf{t} = (t_1, t_2, t_3)$ by

$$R(\mathbf{t}) = \frac{\sup\{L(\tilde{\mathbb{F}}) : \tilde{F}_2(t_2) = \tilde{F}_1(t_1), \tilde{F}_3(t_3) = \tilde{F}_1(t_1),\ \tilde{\mathbb{F}} \in \mathcal{F}^3\}}{\sup\{L(\tilde{\mathbb{F}}) : \tilde{\mathbb{F}} \in \mathcal{F}^3\}}.$$

Write $p_j = F_{jn_j}(t_j)$, $j = 1, 2, 3$, and $p = (n_1p_1 + n_2p_2 + n_3p_3)/m$. It easily follows, cf. (2.1), that

$$R(\mathbf{t}) = \left(\frac{p}{p_1}\right)^{n_1p_1} \left(\frac{1-p}{1-p_1}\right)^{n_1(1-p_1)} \cdot \left(\frac{p}{p_2}\right)^{n_2p_2} \left(\frac{1-p}{1-p_2}\right)^{n_2(1-p_2)} \left(\frac{p}{p_3}\right)^{n_3p_3} \left(\frac{1-p}{1-p_3}\right)^{n_3(1-p_3)}. \tag{2.2}$$

The confidence tubes we will propose are of the form $\{\mathbf{t} : R(\mathbf{t}) > c\}$, for some c.

We assume that $n_j/m \to \pi_j > 0$, as $m \to \infty$, and that the F_j are continuous, for $j = 1, 2, 3$. Let τ_1 be such that $F_1(\tau_1) > 0$ and let $\tau_2 \geq \tau_1$ be such that

$F_1(\tau_2) < 1$. Write $\mathbf{Q}[\tau_1, \tau_2]$ for the restriction of $\mathbf{Q}$ to $t_1 \in [\tau_1, \tau_2]$. Let W_1 and W_2 be two independent standard Wiener processes. Define, for $\alpha \in (0,1)$ and $0 < s_1 < s_2$, $C_\alpha[s_1, s_2]$ by

$$\mathbb{P}\left(\sup_{s \in [s_1, s_2]} \frac{W_1^2(s) + W_2^2(s)}{s} < C_\alpha[s_1, s_2] \right) = 1 - \alpha.$$

Set $\hat{C}_\alpha = C_\alpha[\hat{\sigma}_1^2(\tau_1), \hat{\sigma}_1^2(\tau_2)]$, where

$$\hat{\sigma}_1^2(t_1) = \frac{F_{1n_1}(t_1)}{1 - F_{1n_1}(t_1)}\,. \tag{2.3}$$

Define the confidence tube for $\mathbf{Q}[\tau_1, \tau_2]$ by

$$\mathcal{T} = \left\{ \mathbf{t} \in [\tau_1, \tau_2] \times \mathbb{R}^2 : R(\mathbf{t}) > e^{-\hat{C}_\alpha/2} \right\}.$$

Note that the confidence tubes are essentially invariant under permutations of the order of the three samples involved. We also note an interval property for the confidence tube $\mathcal{T}$ which is useful for computing purposes: one-dimensional cross-sections parallel to a given axis are intervals.

Theorem *Let $\alpha \in (0,1)$. Under the above assumptions,*

$$\lim_{m \to \infty} \mathbb{P}(\mathbf{Q}[\tau_1, \tau_2] \subset \mathcal{T}) = 1 - \alpha.$$

In order to assess the accuracy of the proposed confidence tube (calibrated on the basis of the above limit theorem), we carry out a small simulation study. This accuracy does not depend on F_1, F_2, and F_3 (distribution-freeness). Therefore we can and will restrict ourselves to the case where all these three distribution functions are equal to the standard normal distribution function. We consider the case of a 95% confidence tube and choose τ_1 and τ_2 such that $F_{1n_1}(\tau_1) = 1 - F_{1n_1}(\tau_2) = 0.05$. First, we simulate 20,000 replicates of the Wiener processes W_1 and W_2 over a very fine grid (more than 10^8 equidistant points), which provides

an accurate approximation to $\hat{C}_{0.05} = C_{0.05}[1/19, 19]$, see (2.3). We consider the cases $n_1 = n_2 = n_3 = 100$ and $n_1 = n_2 = n_3 = 200$, and for both cases we compute $5 \cdot 10^6$ confidence tubes. This yields empirical confidence levels 94.4% and 95.1%, respectively. These results show that the confidence tubes are highly accurate for moderate sample sizes.

The proof of the Theorem can be obtained from the proofs in Einmahl and McKeague (1999). In that paper the observations are subject to random censoring, which makes the calculation of $R(\mathbf{t})$ and hence the proofs much more difficult. Here we provide a direct and easier proof, using the explicit expression for $R(\mathbf{t})$ in (2.2).

Proof Write $\sigma_1^2 = F_1/(1 - F_1)$. First we show that

$$(2.4) \qquad -2 \log R(\,\cdot\,, Q_2(F_1(\cdot)),\, Q_3(F_1(\cdot))) \xrightarrow{d} \frac{W_1^2 \circ \sigma_1^2 + W_2^2 \circ \sigma_1^2}{\sigma_1^2} \quad \text{on } D[\tau_1, \tau_2],$$

with W_1 and W_2 as above.

From (2.2) we obtain, with the notation there,

$$\begin{aligned} -2 \log R(\mathbf{t}) = -2 \Bigg(& n_1 p_1 \log \frac{p}{p_1} + n_1(1 - p_1) \log \frac{1 - p}{1 - p_1} \\ & + n_2 p_2 \log \frac{p}{p_2} + n_2(1 - p_2) \log \frac{1 - p}{1 - p_2} + n_3 p_3 \log \frac{p}{p_3} + n_3(1 - p_3) \log \frac{1 - p}{1 - p_3} \Bigg). \end{aligned}$$

Denote the three empirical processes with

$$\alpha_{jn_j} = \sqrt{n_j}(F_{jn_j} - F_j), \quad j = 1, 2, 3.$$

It is well-known that the α_{jn_j} converge in distribution on $D[a, b]$ (for any $a < b$) to $B_j \circ F_j$, respectively, where B_1, B_2 and B_3 are independent, standard Brownian bridges. Hence, we obtain, using a Taylor approximation of $\log(1 + x)$, that,

$$-2 \log R(t_1, Q_2(F_1(t_1)), Q_3(F_1(t_1))) = \frac{n_1(p - p_1)^2}{p_1(1 - p_1)} + \frac{n_2(p - p_2)^2}{p_2(1 - p_2)} + \frac{n_3(p - p_3)^2}{p_3(1 - p_3)} + o_{\mathbb{P}}(1),$$

uniformly on $[\tau_1, \tau_2]$, and this expression converges in distribution on $D[\tau_1, \tau_2]$ to

$$
\begin{aligned}
(2.5) \qquad & \frac{\big((\pi_2+\pi_3)B_1(F_1(t_1)) - \sqrt{\pi_1\pi_2}B_2(F_1(t_1)) - \sqrt{\pi_1\pi_3}B_3(F_1(t_1))\big)^2}{F_1(t_1)(1-F_1(t_1))} \\
& + \frac{\big((\pi_1+\pi_3)B_2(F_1(t_1)) - \sqrt{\pi_1\pi_2}B_1(F_1(t_1)) - \sqrt{\pi_2\pi_3}B_3(F_1(t_1))\big)^2}{F_1(t_1)(1-F_1(t_1))} \\
& + \frac{\big((\pi_1+\pi_2)B_3(F_1(t_1)) - \sqrt{\pi_1\pi_3}B_1(F_1(t_1)) - \sqrt{\pi_2\pi_3}B_2(F_1(t_1))\big)^2}{F_1(t_1)(1-F_1(t_1))}, \text{ as } m \to \infty.
\end{aligned}
$$

Noting that a standard Brownian bridge $B(x)$ divided by $1-x$ is, as a process, equal in distribution to $W(x/(1-x))$ (with W as standard Wiener process), it follows as in Einmahl and McKeague (1999, p. 1361) that on $[\tau_1, \tau_2]$ the process in (2.5) is equal in distribution to

$$
\frac{W_1^2(\sigma_1^2(t_1)) + W_2^2(\sigma_1^2(t_1))}{\sigma_1^2(t_1)}.
$$

Hence we have (2.4).

This yields, as $m \to \infty$,

$$
\begin{aligned}
& \mathbb{P}(\mathbf{Q}[\tau_1, \tau_2] \subset \mathcal{T}) = \mathbb{P}(-2\log R(t_1, Q_2(F_1(t_1)), Q_3(F_1(t_1))) < \hat{C}_\alpha \text{ for all } t_1 \in [\tau_1, \tau_2]) \\
& \to \mathbb{P}\left(\sup_{t_1 \in [\tau_1, \tau_2]} \frac{W_1^2(\sigma_1^2(t_1)) + W_2^2(\sigma_1^2(t_1))}{\sigma_1^2(t_1)} < C_\alpha[\sigma_1^2(\tau_1), \sigma_1^2(\tau_2)] \right) \\
& = \mathbb{P}\left(\sup_{s \in [\sigma_1^2(\tau_1), \sigma_1^2(\tau_2)]} \frac{W_1^2(s) + W_2^2(s)}{s} < C_\alpha[\sigma_1^2(\tau_1), \sigma_1^2(\tau_2)] \right) = 1 - \alpha. \qquad \square
\end{aligned}
$$

3 Application to strike duration data

In this section we apply our procedure to contract strike duration data for U.S. manufacturing industries for the period 1968–1976, see Kennan (1985). The strike durations are measured in days. In this period there were 566 strikes involving at least 1000 workers and lasting at least one day; the durations range from 1–235 days. In order to investigate the influence of industrial production on strike duration, we split the strikes into three groups according to the level of

a monthly industrial production index (a residual value from a regression that removes seasonal and trend components), with each strike assigned the index of the month in which the strike began. The three groups are specified by the index being "close to zero" (that is, between $-c$ and c with $c = 0.022$), below $-c$, or above c, meaning average, low, or high production level, respectively. This leads to three samples with sizes 216, 156 and 194.

The 95% confidence tube for the QQQ plot is displayed in Figure 1, where we have chosen τ_1 and τ_2 so that $F_{1n_1}(\tau_1)$ and $1 - F_{1n_1}(\tau_2)$ are approximately 0.05. Note that the diagonal line $t_1 \mapsto (t_1, t_1)$ stays everywhere inside the tube, so there is no evidence that length of strike depends on production level. The tube thus gives a formal, global, testing procedure to distinguish between the effects of different production levels, but it also allows simultaneous comparison of strike duration quantiles over the three production levels. Note that the tube is narrow for short strikes and much wider for the long strikes, since there are many short strikes but relatively fewer long strikes.

4 Application to cholesterol data

In this section we analyze some data collected as part of the Diverse Populations Collaboration, a study of the relationships between risk factors for various chronic diseases across several countries and cultures, see McGee et al. (2005) for detailed background. Specifically, we consider total serum cholesterol level (in mg/dl) at baseline on men aged 45–65 years who were living in either Massachusetts, Honolulu, or Puerto Rico at the time of their entry into the study; the sizes of the samples available from the three populations are 675, 4602 and 4887, respectively.

We are interested in comparing the distributions of cholesterol levels in the three populations using our confidence tubes. Again, we have computed a 95%

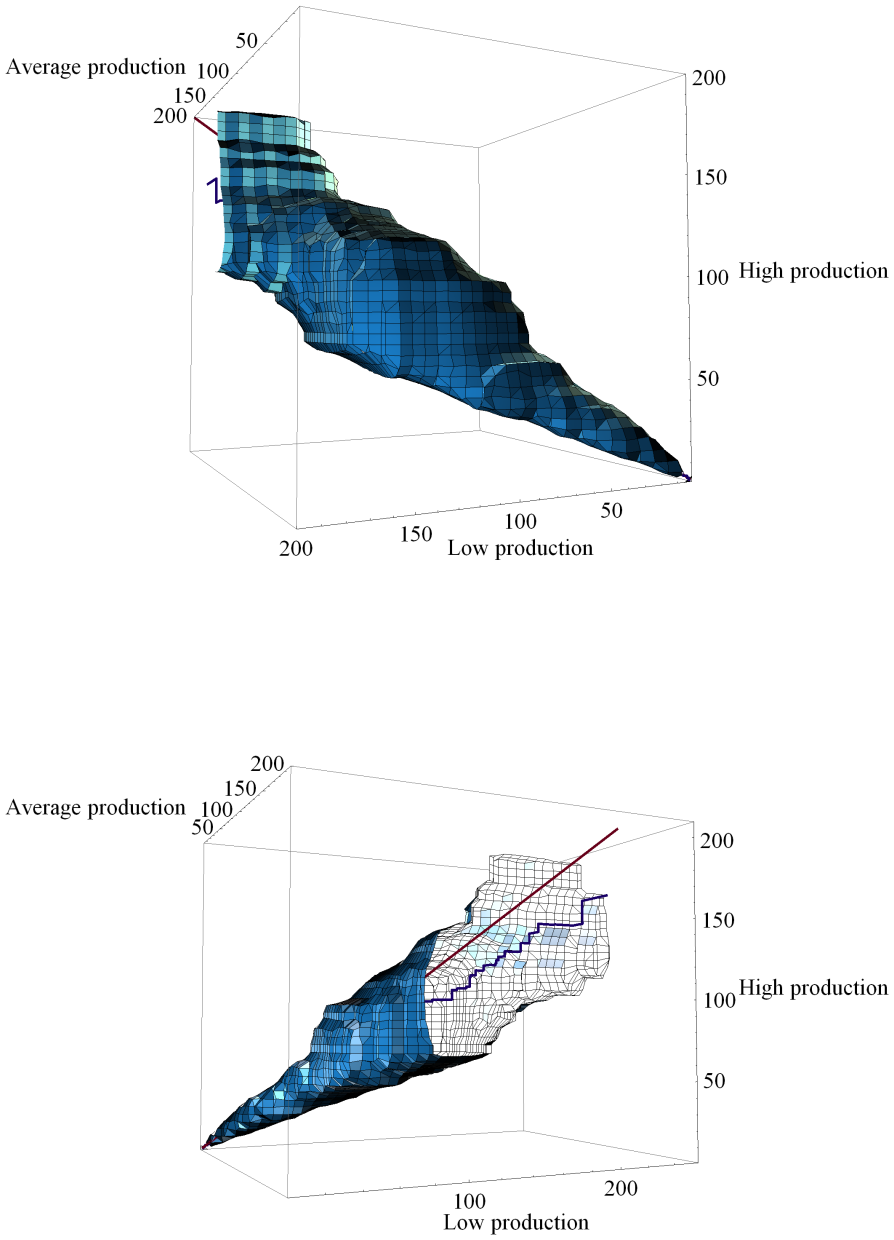

Figure 1: 95% confidence tube for the QQQ plot of the strike durations in average, low, and high productivity periods. The empirical QQQ plot and the diagonal are also depicted.

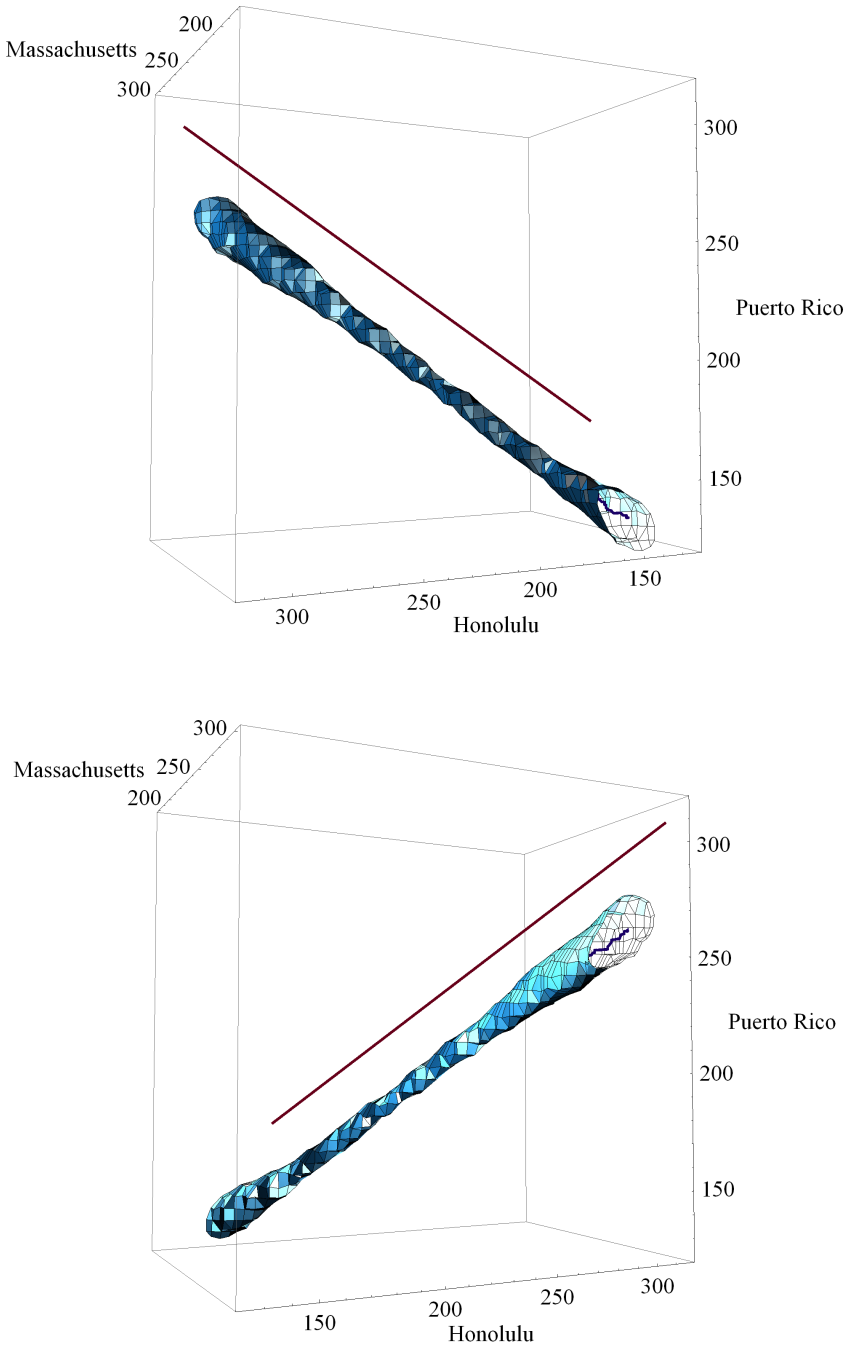

Figure 2: 95% confidence tube for the QQQ plot of the cholesterol levels for men aged 45–65 in Massachusetts, Honolulu, and Puerto Rico. The empirical QQQ plot and the diagonal are also depicted.

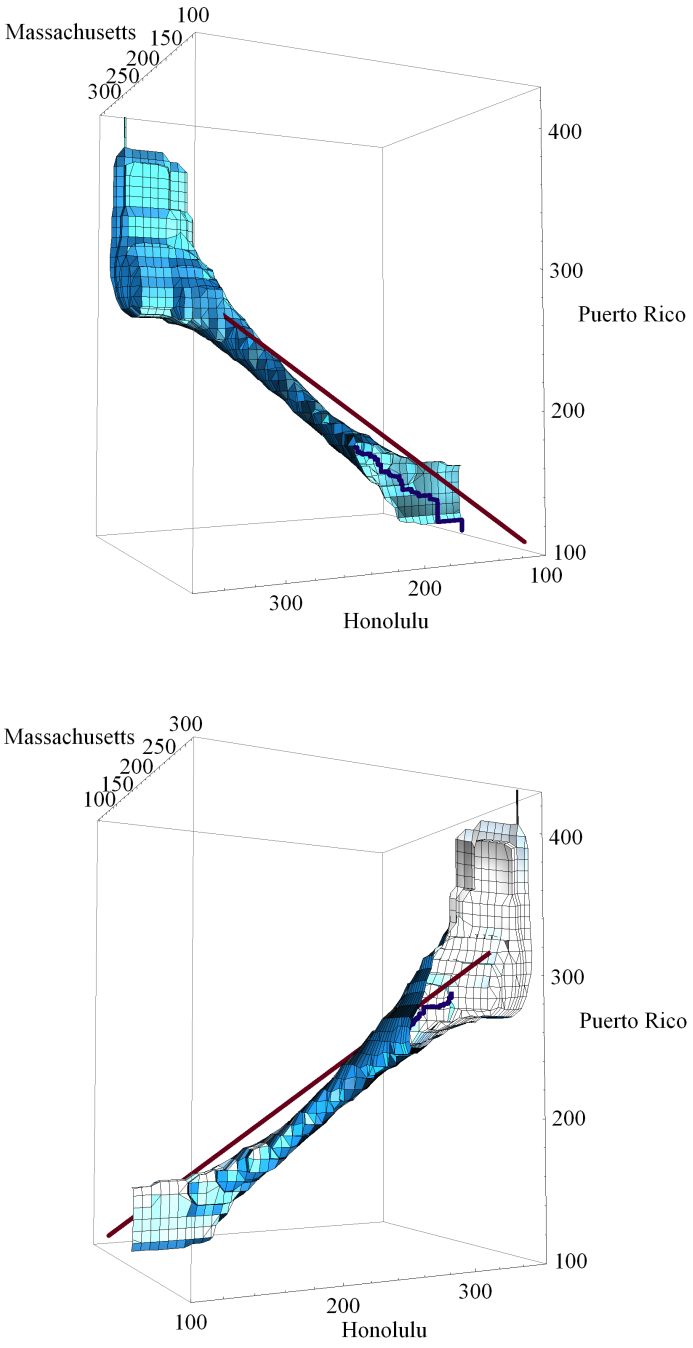

Figure 3: 95% confidence tube for the QQQ plot of the cholesterol levels for obese (BMI $>$ 30) men aged 45–65 in Massachusetts, Honolulu, and Puerto Rico. The empirical QQQ plot and the diagonal are also depicted.

confidence tube for the QQQ plot (see Figure 2) where we have chosen τ_1 and τ_2 such that $F_{1n_1}(\tau_1)$ and $1 - F_{1n_1}(\tau_2)$ are approximately 0.05. (For these plots we used data on all the subjects from Massachusetts, but only 1000 of the subjects from Honolulu or Puerto Rico.) Note that now the diagonal is entirely outside the tube. That is, across all cholesterol levels we see differences between the three populations. More specifically, the cholesterol level quantiles for Puerto Rico are throughout significantly smaller than those in the other two populations, and Honolulu has smaller quantiles than Massachusetts.

It is also of interest to examine whether the patterns noted above continue to hold when we stratify over three levels of BMI (body mass index, in units of kg/m^2): normal (18.5–25), overweight (25–30) and obese (> 30). We have computed the 95% confidence tubes for the QQQ plots of the three populations (using all the data in this case), for the normal, overweight and obese men separately. It turns out that the tubes for the normal and overweight men look very similar to the tube in Figure 2 for the unstratified situation. The tube for the obese men (based on sample sizes 87, 160 and 628, respectively), however, looks quite different, see Figure 3. It is interesting to note that in this case the diagonal is partly inside and partly outside the tube. The fact that the diagonal is not entirely inside the tube means that, although there is again a significant difference between the distributions of cholesterol levels in the three populations, the differences now only occur at *lower* cholesterol levels. The lowest cholesterol level quantiles are again found in Puerto Rico.

In Figures 2 and 3 we see that the tubes are narrower in the middle and wider at the ends. This is due to the fact that there are more data in the center of the distribution than in the tails. The tube in Figure 3 is wider than that in Figure 2 since the sample sizes for the obese group are much smaller.

Supplementary materials

Strike data Strike durations for U.S. manufacturing industries 1968–1976 (Section 3). (txt file)

Cholesterol data Cholesterol levels for men aged 45–65 at three locations. (Section 4). (dat file)

Mathematica code Computer code used for the computation of the confidence tubes. Code can be viewed with the Mathematica viewer at "www.wolfram.com". (three nb files)

Acknowledgements We are grateful to Jaap Abbring for pointing out the strike data and to Daniel McGee for providing the cholesterol data set.

References

Beirlant, J. and Deheuvels, P. (1990). On the approximation of P-P and Q-Q plot processes by Brownian bridges. *Statistics and Probability Letters* **9**, 241–251.

Doksum, K.A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics* **2**, 267–277.

Doksum, K.A. (1977). Some graphical methods in statistics. A review and some extensions. *Statistica Neerlandica* **31**, 53-68.

Doksum, K.A. and Sievers, G.L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63**, 421–434.

Einmahl, J.H.J. and McKeague, I.W. (1999). Confidence tubes for multiple quantile plots via empirical likelihood. *The Annals of Statistics* **27**, 1348–1367.

Fisher, N.I. (1983). Graphical methods in nonparametric statistics: A review and annotated bibliography. *International Statistical Review* **51**, 25–58.

García Ben, M. and Yohai, V.J. (2004). Quantile-quantile plot for deviance residuals in the generalized linear model. *Journal of Computational and Graphical Statistics* **13**, 36–47.

Kennan, J. (1985). The duration of contract strikes in U.S. manufacturing. *Journal of Econometrics* **28**, 5–28.

Li, G., Hollander, M., McKeague, I.W. and Yang, J. (1996). Nonparametric likelihood ratio confidence bands for quantile functions from incomplete survival data. *The Annals of Statistics* **24**, 628–640.

Marden, J.I. (2004). Positions and QQ plots. *Statistical Science* **19**, 606–614.

McGee, D.L. and the Diverse Populations Collaboration (2005). Body mass index and mortality: a meta-analysis based on person-level data from twenty-six observational studies. *Annals of Epidemiology* **15**, 87–97.

Nair, V.N. (1978). *Graphical Comparisons of Populations in some Non-linear Models*, Ph.D. thesis, University of California at Berkeley.

Nair, V.N. (1982). Q-Q plots with confidence bands for comparing several populations. *Scandinavian Journal of Statistics* **9**, 193–200.

Owen, A. (2001). *Empirical Likelihood*, Boca Raton, FL: Chapman & Hall/CRC.

Switzer, P. (1976). Confidence procedures for two-sample problems. *Biometrika* **63**, 13–25.

MAAB	JHJE	IWMcK
Dept. Mathematics and Comp. Sci.	Dept. of Econometrics	Dept. of Biostatistics
Eindhoven University of Technology	Tilburg University	Columbia University
P.O. Box 513	P.O. Box 90153	722 West 168th Street
5600 MB Eindhoven	5000 LE Tilburg	New York, NY 10032
The Netherlands	The Netherlands	USA
marko@win.tue.nl	j.h.j.einmahl@uvt.nl	im2131@columbia.edu