Earth Sciences
2023; 12(3): 65-73
http://www.sciencepublishinggroup.com/j/earth
doi: 10.11648/j.earth.20231203.12
ISSN: 2328-5974 (Print); ISSN: 2328-5982 (Online)

Summarizing Standpipe Water Data for Meknes Small Rural Communities (Morocco): An Extended Table for Testing for the Uniform Distribution

Mohamed Saffi

Ecole Supérieure de Technologie, Salé, Morocco

Email address:
saffimo@hotmail.com

To cite this article:
Mohamed Saffi. Summarizing Standpipe Water Data for Meknes Small Rural Communities (Morocco): An Extended Table for Testing for the Uniform Distribution. *Earth Sciences*. Vol. 12, No. 3, 2023, pp. 65-73. doi: 10.11648/j.earth.20231203.12

Received: May 8, 2023; **Accepted**: June 1, 2023; **Published**: June 20, 2023

Abstract: This paper is twofold. On the one hand, standpipe water data for small rural communities in Meknes region are studied. 89 villages with population less than 1200 have been probed. In total, about 37000 inhabitants are covered by this study. From the conducted analysis, it appears that the daily per capita water-use rate for the considered population is uniformly distributed with an estimated average of about 10 liter. The lower and the upper 95% confidence bounds placed on this average are 4.3 and 15.9; respectively. Such data can, for example, be useful while planning *regional* water conveyance systems. On the other hand, the statistical test used in this work is based on the closed form of the Shapiro-Wilk statistic, *Wu*, made explicit by Cheng and Spiring in the case of uniformly distributed random variables. Even though *Wu* is computationally simple, its sampling distribution seems to be intractable for arbitrarily sample-size values. On that account, Monte Carlo simulations are run to generate custom quantiles frequently needed in a typical hypothesis testing problem. Then, the study extends and improves the Cheng-Spiring quantile Table 1 for testing for the uniform distribution. It also proposes a simplified and user-friendly graphical support which serves the same task.

Keywords: Water-Use, Standpipe, Rural, Uniform Distribution, Shapiro-Wilk Statistic, Méknes, INDH

1. Introduction

Providing Moroccan rural population with safe drinking water is one of the several goals of the National Initiative for Human Resources Program, popularly known under the acronym 'INDH' after its French name. The availability of potable water in rural districts can contribute to deep positive social ramifications e.g. to promote opportunities for girls to attend school and to help with reducing rural exodus as well as risks of waterborne diseases.

Water-use rate for rural communities, denoted hereinafter by q, is among essential data required to carry out reasonable allocations of drinking water resource at a *regional* scale. This could be ascribed to the fact that a rural population inclines to organize itself spatially as village clusters, [1] (p. 145). This paper makes an attempt to investigate the pattern of standpipe water consumption in Meknes rural areas. It focuses on small communities having a population of fewer than 1200. In this case, standpipe water is generally used to satisfy vital needs such as drinking and cooking while the demand of water intensive activities, like livestock keeping, is usually fulfilled making use of raw-water sources e.g. nearby streams.

The next two sections present the survey and suggest an intuitive exploration of the collected data set. Then a formal hypothesis test is conducted in section 4. Particularly an extended and improved Table to test for the uniform distribution family is provided. It has resulted in an amendment to Cheng-Spiring Theorem 5 that is presented in section 5. Some heuristics, which compare bias and variance magnitudes in the context of the present analysis, are detailed in section 6. Finally, the main conclusions are summed up.

2. Data and Survey

89 villages (small-communities) among Meknes provinces have been probed; all of them are supplied with standpipe water. Their population ranges from 90 to 1143 inhabitants and add up to 37652 inhabitants. Billed-water data for these

villages over four successive trimesters (i.e. one year) have been collected. Let V_{it} denotes the volume of water consumed in the village number i during trimester t and P_i is the corresponding population. Then, the daily per-capita water use rate, for the village i, is defined by:

$$q_i = \frac{V_{i1} + V_{i2} + V_{i3} + V_{i4}}{365 P_i}, \quad 1 \le i \le 89 \tag{1}$$

The series of observed rates $q_1, q_2, \ldots, q_{89}$ is interpreted as 89 random outcomes of the rate q. The scatter plot of Figure 1 displays the observed data-points $(\log_{10}(P_i), q_i)$. In principle, positive correlation implies that data tend to crowd in large numbers in quadrants I and III and become less dense in quadrants II and IV; and vice versa with negative correlation. It is clear from the graph that data points swarm quite uniformly over the four quadrants which is indicative of a very weak correlation. In fact, an estimate R of the population correlation coefficient ρ is numerically equal to $9.53\,10^{-3}$. And the associated 95% confidence interval is (-0.20;+0.22). This means that the null hypothesis according to which ρ=0 couldn't be rejected in a 5% level two-sided test, [2] (p. 43). Similar conclusions are drawn by running the conventional significance test of the hypothesis that the slope β is equal to zero, in the one-dimensional linear regression model $q=\alpha+\beta\log_{10}(P) + \varepsilon$; ε is normally distributed with zero mean and α is the intercept. The relationship between the t-statistic associated with the least squares estimate $\hat{\beta}$ of β and the sample correlation coefficient R is, [3] (p. 85):

$$t = \pm R\sqrt{\frac{n-2}{1-R^2}} \tag{2}$$

Numerically, t = 0.089. The standard error, $SE(\hat{\beta})$, of $\hat{\beta}$ is equal to 1.1962 and the usual 95% confidence interval for β is (-2.271, +2.484).

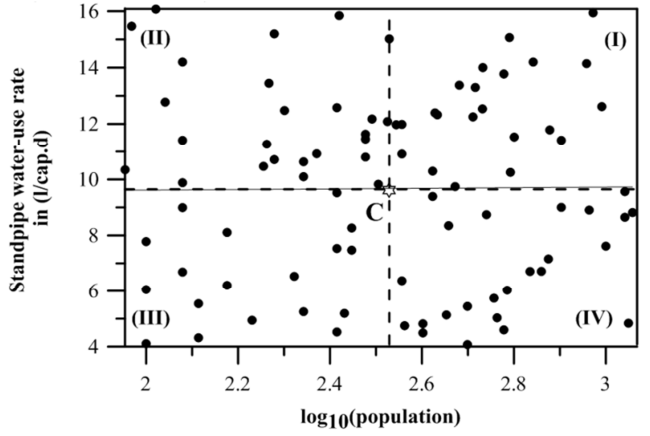


Figure 1. *Observed standpipe water-use rate versus the decimal logarithm of the population for Meknes villages: The solid line (nearly horizontal) is the regression line.*

So, practically the observed data carry no evidence against the assumption that β=0. Explained and residual sum of squares are 0.0918 and 1011.7931; respectively. That is, the proportion of variance in the observed water-use rates explained by the population is practically insubstantial. And the population doesn't seem to be an effective explanatory variable to be included in the analysis, in this case.

3. Data Informal Processing

The observed realizations of the daily water-use rate range from $q_{(1)}$=4.078 to $q_{(89)}$=16.079 liter per capita. The boxplot and the histogram of Figure 2 display the familiar appearance of the uniform distribution family. That visual perception is corroborated by the agreement between the sample summary statistics, computed via EXCEL (column 2, Table 1) and their population counterparts (column 3). The latter are derived under the conjecture of a uniform distribution and using the generalized least square estimates, $\hat{\mu}$ and $\hat{\sigma}$ of the population mean μ and standard deviation σ; namely [4] (p. 92):

$$\begin{cases} \hat{\mu} = \dfrac{q_{(n)} + q_{(1)}}{2} \\ \hat{\sigma} = \dfrac{1}{2\sqrt{3}} \dfrac{(n+1)}{(n-1)} \left(q_{(n)} - q_{(1)} \right) \end{cases} \tag{3}$$

$q_{(1)}$ and $q_{(n)}$ denote the first and the last order statistics in the series of observed water use-rates; n stands for the sample size. A more detailed description of the observed rate q_i for i=1, 2, …, 89 is provided by the empirical cumulative distribution function (*eCDF*).

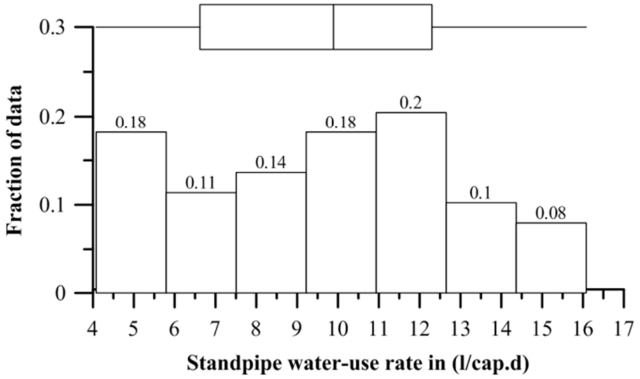


Figure 2. *The histogram and the boxplot for the standpipe water-use data of the probed Meknes villages. The labels stand for the fraction of data falling under each bin.*

Table 1. *Some basic summary statistics of the standpipe water-use data for the probed Meknes villages.*

	Sample statistic	Population statistic
Average (l/cap.d)	9.674	10.077
Standard deviation	3.391	3.542
Skewness	0.002	0
Excess kurtosis	-1.062	-1.200

The latter is believed to be a reliable and computable substitute for the true but inaccessible CDF. Expressly:

$$CDF(q) \approx eCDF(q) = \frac{N(q)}{89} \tag{4}$$

$N(q)$ denotes the number of observed rates q_i less than or equal to q. And $CDF(q)$ is the true (unknown) cumulative

distribution function of rate q. The legitimacy of the foregoing approximation is formally predicated upon the Glivenko-Cantelli Theorem, [5] (p. 100). Based on (4), Figure 3 shows that the *eCDF* (black bullets) is globally linear (particularly at the tails), which is another distinctive feature of the uniform distribution family. So, a plausible approximation to the true cumulative distribution function of q is denoted by $aCDF(q)$ and could be conjectured as:

$$aCDF(q) = \frac{q-\lambda_1}{\lambda_2-\lambda_1} \text{ for } \lambda_1 \le q \le \lambda_2 \tag{5}$$

λ_1 and λ_2 are the distribution range extremities. It is essential to bear in mind that the choice made in (5) is based on a *free* interpretation of the data; another less trivial choice may be possible as well. However, this 'arbitrariness' is bound by at least two major restrictions, both of them are in the mainstream; namely (i) the bias/variance tradeoff, and (ii) the (almost sure) convergence property of the $eCDF(q)$ to the true $CDF(q)$. Regarding the former restriction, a model that contains more parameters than required tends to overuse the data by learning not only from the signal carried in these data but also from the noise they are corrupted by. In such models, where excessive bias reduction is sought after, large variance becomes an inevitable side effect. So the simpler is the model the better. However, the model should incorporate a certain degree of complexity in order to capture the main features of the observed data. This is exactly what the latter restriction is about. That is, a reasonable approximating function, to the cumulative distribution function, need to be consistent with the $eCDF(q)$. The usual down-to-earth procedure to verify this requirement is to construct the 95% confidence band around the$eCDF(q)$. If it happens that the selected $aCDF(q)$ is not fully covered by this band then it ought to be a poor approximation to the true CDF. The lower and the upper bounds of the confidence band, $L(q;n,\alpha)$ and $U(q;n,\alpha)$, are found via DKWM inequality, [5] (p. 117).

$$\begin{cases} L(q;n,\alpha) = \sup(eCDF(q) - BHW,\ 0) \\ U(q;n,\alpha) = \inf(eCDF(q) + BHW,\ 1) \end{cases} \tag{6}$$

$BHW = (-0.5n^{-1}\log(0.5\alpha))^{1/2}$ represents the band half-width and $(1-\alpha)$ stands for the prescribed confidence level; α=0.05. The shaded area in Figure 3 highlights the confidence band for Meknes data. Remarkably, the band occurs to be straight enough to lodge the whole line that represents $a\hat{CDF}(q)$ (solid line in Figure 3). $a\hat{CDF}(q)$ is the estimate of $aCDF(q)$ based on the generalized least squares estimates, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, of the parameters λ_1 and λ_2. They are given by Lloyd, [4] (p. 92):

$$\begin{cases} \hat{\lambda}_1 = q_{(1)} - \dfrac{q_{(n)}-q_{(1)}}{(n-1)} \\ \hat{\lambda}_2 = q_{(n)} + \dfrac{q_{(n)}-q_{(1)}}{(n-1)} \end{cases} \tag{7}$$

The symbols in (7) are defined in (3). Again, this fact is more of another piece of evidence showing that the observed 89 data points lend themselves well to the uniform distribution family, than an outcome of pure happenstance.

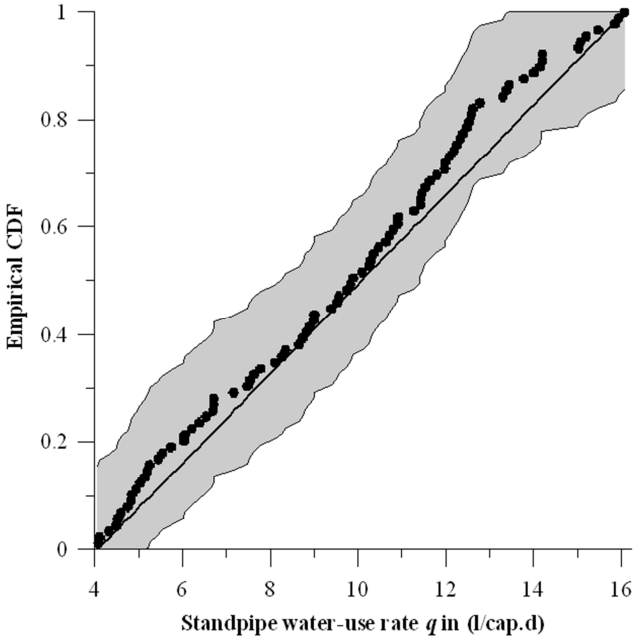


Figure 3. *The empirical CDF (black bullet) corresponding to the standpipe water-use rate data for Meknes provinces along with the associated 95% confidence band (grey area). The solid line is the generalized least squares estimate of the aCDF(q) in (5).*

4. Testing Meknes Data for the Uniform Distribution

The informal data exploration presented in the previous section seems to support the claim that "water consumption rate$q$ is uniformly distributed." This statement would be taken for the null hypothesis H0, against the alternative H1: "rate q is not uniformly distributed." For the time being no specific type of departure from H0 is made explicit, just in line with the material presented in chapter 3 of Cox and Hinkley book, [6]. That is, the aim is to decide whether the 89 data-point sample is typical or atypical compared to the parent population, assuming that H0 is true. The typicality of data is highlighted by Emmert-Streib and Dehmer, [7] (p. 948). In this work, it is quantified by the Shapiro-Wilk statistic *Wu* and its sampling distribution. Under H0, Cheng and Spiring give the closed form of statistic *Wu* in terms of the first and the last order statistics $q_{(1)}$ and $q_{(n)}$ and the corrected sum of squares, [8]:

$$Wu = \frac{\left[\left(q_{(n)}-q_{(1)}\right)\dfrac{n+1}{n-1}\right]^2}{\sum_{i=1}^{n}\left(q_i-\bar{q}\right)^2} \tag{8}$$

Where $\bar{q}$ and n are the sample mean and size. The statistic *Wu* is computationally simple and stays invariant under linear transformations. But from Cheng and Spiring Theorems 4 and 5 it appears that its exact distribution under H0, $CDF(Wu|\text{H0},n)$, shows some propensity to develop a cumbersome algebraic form when the sample size is increased.

So, it would be convenient to generate the quantiles serving the task of hypothesis testing directly via numerical simulation. The present survey, based on 89 villages, is not covered by Cheng and Spiring Table 1, [8]. The latter is therefore extended to include large values of n as well as the 0.025- and 0.975-quantiles. The question is brought down to solving the following basic equation in the quantile $\kappa(p,n)$ for a target percentile p pre-specified between 0 and 1, [9]:

$$p = CDF(\kappa|n,\mathrm{H0}) \tag{9}$$

The standard Monte Carlo method [10] is used to relax the foregoing equation by replacing the exact sampling distribution $CDF(Wu|n,\mathrm{H0})$ with a tractable Monte Carlo estimate, $MC\_CDF(Wu|n,\mathrm{H0})$. The latter could be, for instance, a piecewise linear analog of the empirical cumulative distribution function constructed from the generated samples, [11]. The other steps are as per usual:

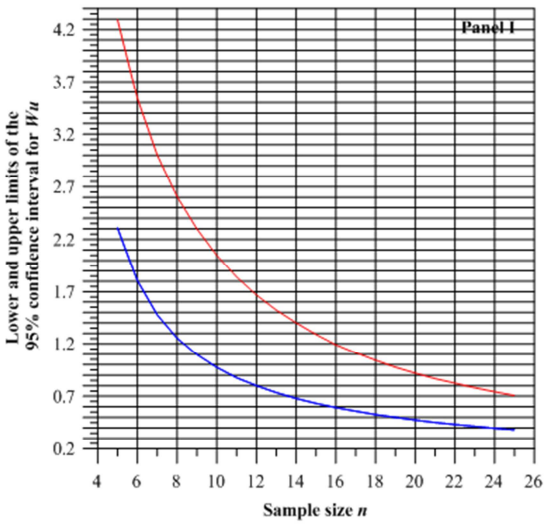
(i) Generate M sets of uniformly distributed and independent n random numbers: $S_1=(r_{11},r_{21},\ldots,r_{n1})$, $S_2=(r_{12},r_{22},\ldots,r_{n2})$, …, $S_M=(r_{1M},r_{2M},\ldots,r_{nM})$.

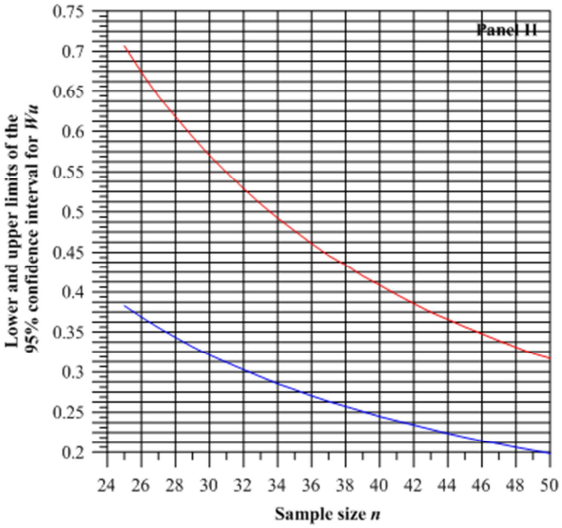
(ii) Compute the corresponding realizations of the Shapiro Wilk statistic, namely: $Wu_1=Wu(S_1)$, $Wu_2=Wu(S_2)$,…, $Wu_M=Wu(S_M)$ using (8).

(iii) Sort the values obtained in step (ii) into order statistics $Wu_{(1)}$, $Wu_{(2)}$, …, $Wu_{(M)}$ and take them for (0.5/M), (1.5/M) (M-0.5)/M quantiles; respectively.

(iv) Then compute an estimate $\hat{\kappa}(p,n,M)$, for the p-quantile $\kappa(p,n)$ using linear interpolation.

The MATLAB command *rand*(.) is used to generate the sets of random numbers required in step (i). The quantile estimator in steps (iii) and (iv) could be performed in many ways, herein the MATLAB implementation is opted for and carried out via the command *quantile*(.,.); it also corresponds to relations (4) and (5) in Avramidis and Wilson paper, [12]. The latter show (in proposition 2, relation (20)) that the solution $\hat{\kappa}(p,n,M)$ delivered by the procedure above is biased. And, under some regularity conditions, the bias is of order 1/M. This property meets the classical condition mentioned by Miller, [13] (equality (1.2)) that allows for bias reduction by the jackknife technique. However, when the number of replications M is made larger, the bias correction becomes negligible compared with the statistical error, [12] (section 2.2). This fact is confirmed by some heuristics performed in the last section of this work. In light of this, the previous four-step procedure is directly run with 120000 replications, and it results in Table 2. Calculation is carried out using MATLAB. Table 2 and Cheng and Spiring's Table 1 [8] (originally based on 10000 replications) are in good agreement. Yet, it is worthy of mention that the quantiles delivered by Cheng and Spiring's Theorem 5 are inconsistent with the values they have tabulated; details are deferred to the next section to keep this exposition streamlined. Table 2 can also be useful while checking the goodness-of-fit for other distributions that are fully specified, [8]. A user friendly representation of this Table, specialized to the 95% non-rejection region in a two-sided test, is suggested in the four panels of Figure 4 for the sample size n between (5 and 25); (25 and 50); (50 and 75); and finally (75 and 100). For example, with n = 89 the size of the sample used in this work; panel IV delivers graphically the 95% non-rejection region of about (0.115, 0.164) (i.e. section [L,U]). A similar range can also be read from Table 2 by linear interpolation. The observed value of the test statistic, corresponding to the probed Meknes villages, is 0.146. So, formally we conclude that the observed (standpipe) water-use rate data bear no substantial evidence against the null H0. And, that the uniform distribution can be considered as a suitable approximation, possibly among others, to the true but inaccessible distribution. Auxiliary information such as the feasible range for Wu and its central value can be straightforwardly obtained from Cheng and Spiring's Theorems 6 and 1, [8]. For n = 89, they are (0.047, 2.092) and 0.1379; respectively. In brief, an estimate of the average daily standpipe water-use rate for Meknes villages is numerically equal to 10.077 liter per capita with the 95% confidence interval of (4.249, 15.906). It is noteworthy that consumer oriented surveys for several rural communities and desirably by more than two independent studies are highly recommended, [14].





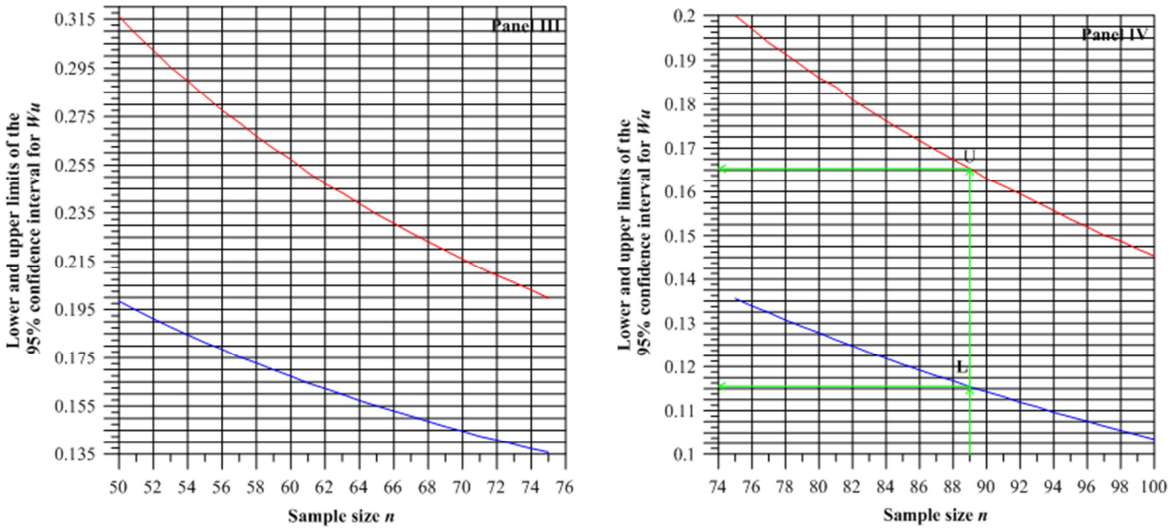


Figure 4. *The lower (blue line) and the upper (red line) 95% confidence limits for the Shapiro-Wilk statistic Wu defined in (8). For example, with a sample size* $n = 89$ *, the non-rejection region is given by the Panel IV. It corresponds to section [L, U] (green line).*

Table 2. *P-quantiles computed via Monte Carlo Method based on 120000 replications; n is the sample size.*

n	percentile							
	0.010	**0.025**	**0.050**	**0.100**	**0.900**	**0.950**	**0.975**	**0.990**
3	6.0295	6.0752	6.1497	6.3022	7.9738	7.9933	7.9984	7.9997
4	3.0666	3.2410	3.4387	3.7359	5.3164	5.4331	5.4934	5.5304
5	2.1998	2.3097	2.4202	2.5785	4.0074	4.1770	4.2918	4.3855
6	1.7087	1.8049	1.8927	2.0093	3.2236	3.4011	3.5390	3.6720
7	1.3979	1.4780	1.5535	1.6503	2.6967	2.8617	3.0041	3.1563
8	1.1882	1.2552	1.3196	1.4021	2.3158	2.4699	2.6050	2.7568
9	1.0368	1.0954	1.1517	1.2242	2.0250	2.1719	2.2994	2.4460
10	0.9223	0.9732	1.0205	1.0830	1.7927	1.9305	2.0504	2.1944
11	0.8289	0.8752	0.9185	0.9746	1.6018	1.7298	1.8408	1.9813
12	0.7577	0.7995	0.8374	0.8872	1.4455	1.5611	1.6669	1.7976
13	0.6976	0.7344	0.7686	0.8142	1.3180	1.4230	1.5236	1.6465
14	0.6460	0.6797	0.7126	0.7535	1.2107	1.3082	1.4012	1.5150
15	0.6013	0.6323	0.6623	0.6999	1.1167	1.2068	1.2911	1.3995
16	0.5639	0.5931	0.6200	0.6552	1.0343	1.1154	1.1916	1.2899
17	0.5319	0.5593	0.5842	0.6163	0.9672	1.0426	1.1125	1.1996
18	0.5028	0.5273	0.5515	0.5812	0.9060	0.9753	1.0410	1.1251
19	0.4760	0.5007	0.5224	0.5507	0.8512	0.9155	0.9745	1.0529
20	0.4528	0.4749	0.4959	0.5219	0.8026	0.8617	0.9187	0.9934
21	0.4310	0.4527	0.4721	0.4973	0.7591	0.8150	0.8663	0.9349
22	0.4134	0.4335	0.4521	0.4747	0.7208	0.7725	0.8232	0.8865
23	0.3968	0.4157	0.4331	0.4547	0.6851	0.7335	0.7802	0.8386
24	0.3803	0.3977	0.4147	0.4354	0.6529	0.6980	0.7418	0.7952
25	0.3665	0.3831	0.3988	0.4183	0.6226	0.6650	0.7068	0.7589
26	0.3526	0.3688	0.3836	0.4023	0.5961	0.6356	0.6742	0.7218
27	0.3405	0.3555	0.3697	0.3873	0.5706	0.6082	0.6433	0.6890
28	0.3287	0.3434	0.3571	0.3739	0.5485	0.5840	0.6183	0.6609
29	0.3178	0.3316	0.3447	0.3610	0.5278	0.5620	0.5935	0.6332
30	0.3084	0.3209	0.3337	0.3491	0.5068	0.5387	0.5702	0.6086
31	0.2988	0.3116	0.3236	0.3386	0.4898	0.5202	0.5495	0.5863
32	0.2902	0.3026	0.3138	0.3282	0.4721	0.5018	0.5287	0.5638
33	0.2821	0.2939	0.3047	0.3185	0.4559	0.4836	0.5098	0.5424
34	0.2743	0.2853	0.2963	0.3092	0.4416	0.4681	0.4923	0.5249
35	0.2670	0.2779	0.2880	0.3007	0.4276	0.4532	0.4766	0.5056
36	0.2594	0.2703	0.2803	0.2925	0.4135	0.4375	0.4609	0.4910
37	0.2529	0.2634	0.2729	0.2848	0.4012	0.4244	0.4460	0.4737
38	0.2474	0.2573	0.2664	0.2777	0.3902	0.4129	0.4340	0.4598
39	0.2419	0.2511	0.2596	0.2705	0.3791	0.4001	0.4200	0.4449

n	percentile							
	0.010	**0.025**	**0.050**	**0.100**	**0.900**	**0.950**	**0.975**	**0.990**
40	0.2357	0.2448	0.2535	0.2641	0.3690	0.3893	0.4087	0.4330
42	0.2252	0.2343	0.2422	0.2519	0.3491	0.3680	0.3856	0.4077
44	0.2152	0.2238	0.2316	0.2409	0.3316	0.3490	0.3657	0.3857
46	0.2068	0.2144	0.2217	0.2305	0.3158	0.3322	0.3479	0.3671
48	0.1988	0.2062	0.2130	0.2213	0.3012	0.3166	0.3310	0.3485
50	0.1916	0.1986	0.2049	0.2128	0.2884	0.3029	0.3166	0.3335
54	0.1782	0.1846	0.1905	0.1976	0.2648	0.2777	0.2898	0.3045
58	0.1671	0.1728	0.1781	0.1845	0.2449	0.2563	0.2669	0.2798
62	0.1570	0.1623	0.1670	0.1730	0.2276	0.2379	0.2471	0.2590
66	0.1481	0.1530	0.1574	0.1628	0.2130	0.2222	0.2308	0.2414
70	0.1401	0.1447	0.1488	0.1538	0.1999	0.2083	0.2160	0.2259
75	0.1317	0.1358	0.1396	0.1441	0.1855	0.1930	0.1998	0.2089
80	0.1239	0.1278	0.1311	0.1353	0.1729	0.1798	0.1860	0.1939
85	0.1170	0.1205	0.1238	0.1276	0.1621	0.1683	0.1739	0.1808
90	0.1109	0.1143	0.1173	0.1208	0.1524	0.1580	0.1630	0.1697
95	0.1056	0.1086	0.1113	0.1146	0.1438	0.1489	0.1537	0.1594
100	0.1006	0.1035	0.1060	0.1091	0.1362	0.1410	0.1454	0.1509

5. Monte Carlo Versus Closed Form Sampling Distributions for $n = 3$ or 4

The function defined by Theorem 5 in reference [8] is not a feasible PDF because it doesn't integrate to 1 over its range; it integrates approximately to 1.084. As such, the quantiles corresponding to 0.99; 0.95; 0.90 and 0.10 percentiles are in disagreement with their counterparts computed directly by numerical simulation (columns 2 and 3, Table 3). The misfit is also clearly displayed on Figure 5; it starts to show up at Wu=200/54 (black arrow). The amended expression of the sampling distribution for n=4 is deduced using Theorem 2.1.5 in the Casella and Berger book, [15] (p. 51). It should be:

$$PDF(Wu)=\begin{cases} a\left(1-\frac{2}{\pi}\arcsin(b_1c_1)\right), & \frac{150}{54}\le Wu<\frac{200}{54} \\ a\left(1-\frac{2}{\pi}\arcsin(b_2c_2)\right), & \frac{200}{54}\le Wu<\frac{225}{54} \\ a, & \frac{225}{54}\le Wu<\frac{300}{54} \\ 0, & \text{elsewhere} \end{cases} \quad (10)$$

Where

$$a=\frac{25\pi 2^{1/2}}{9Wu^2},\quad b_1=\left[3\left(\frac{50}{9Wu}-1\right)^{1/2}\right]^{-1},\ b_2=\left[3\left(\frac{50}{9Wu}-1\right)\right]^{-1},$$

$$c_1=1+2\left(\frac{25}{3Wu}-2\right)^{1/2},\quad c_2=2\left(\frac{50}{3Wu}-4\right)^{1/2}$$

The amendment in (10) is brought by the factor b_2. Good agreement is observed between (10) and the outcome of the Monte Carlo code (Figure 5). It is also compiled in the last two columns of Table 3. Figure 6 confirms that the sampling distribution derived from Cheng and Spiring's Theorem 4 [8] and that simulated by Monte Carlo method agree well.

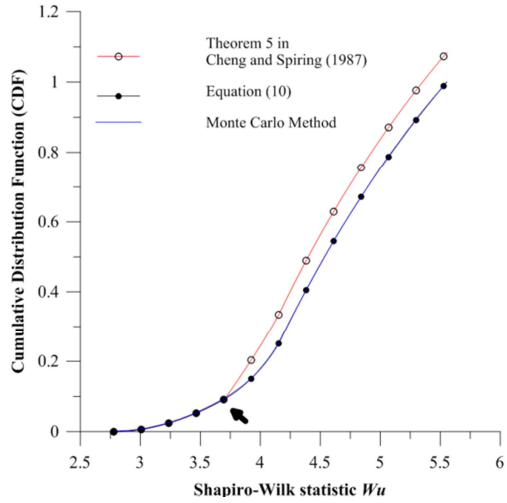


Figure 5. *The cumulative distribution function of the Shapiro-Wilk statistic computed via (i) Equation (10) (ii) Monte Carlo Method and (iii) Cheng-Spiring Theorem 5, for sample size* $n = 4$.

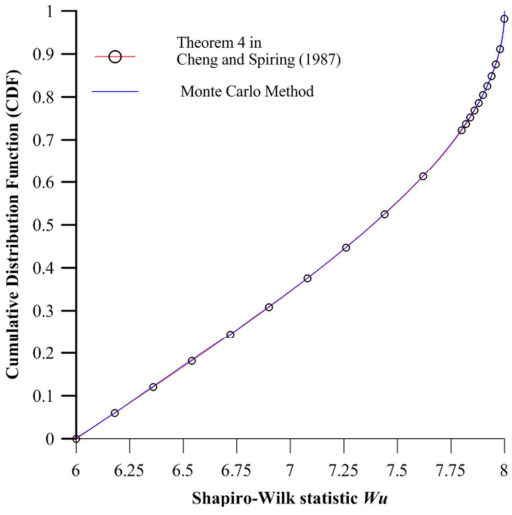


Figure 6. *The cumulative distribution function of the Shapiro-Wilk statistic computed via Cheng-Spiring Theorem 4 and the Monte Carlo Method, for sample size* $n = 3$.

Table 3. *Quantiles cross-checked for sample size* $n = 4$ *. The Monte Carlo Method (MCM) simulation is based on 120000 replications.*

Percentile	Cheng and Spiring Tabulated quantile	Quantile based on Cheng and Spiring Theorem 5	Quantile based on (10)	Quantile generated in this work by the MCM
0.01	3.080	3.067	3.067	3.067
0.05	3.440	3.447	3.447	3.439
0.10	3.740	3.718	3.738	3.736
0.90	5.310	5.130	5.316	5.316
0.95	5.430	5.239	5.433	5.433
0.99	5.530	5.329	5.531	5.530

6. Heuristic Investigations on the Accuracy of the Monte Carlo Quantile Estimate

Estimator, $\hat{\kappa}(p,n,M)$ of the root $\kappa(p,n)$ of (9) is by construction a random variable. That is, its realization changes randomly from a code run to the next. And, the mean squared error (*MSE*) is a measure widely accepted for the task of assessing how well $\hat{\kappa}(p,n,M)$ approximates the unknown p-quantile $\kappa(p,n)$. The heuristics presented in this section are intended to provide a straightforward verification (for n=3 and 4) of the fact that the bias becomes effectively negligible when the number of replications is increased. That is, the variance may be used as a suitable proxy for the $MSE(\hat{\kappa})$. The simulations are illustrated with the least favorable case of extreme quantiles notorious for being associated with severe bias as explained in the discussion following the proposition 2 in Avramidis and Wilson paper, [12]. Thereby, percentile p would be set equal to 0.990. The $MSE(\hat{\kappa})$ is conventionally broken into two parts, yet a bigger picture is given in reference [16] (p. 223): The first part is the variance $Var(\hat{\kappa})$; it refers to the expected squared deviation of the estimate $\hat{\kappa}(p,n,M)$ around its mean. $Var(\hat{\kappa})$ is numerically computed by running the Monte Carlo code T times (T=200, say), for specified p, n and M. The sample variance of the resulting series of the code outputs $\hat{\kappa}_1$, $\hat{\kappa}_2$,…, $\hat{\kappa}_T$ is taken for an approximate value of $Var(\hat{\kappa})$. The second part is the bias-squared, and it measures the gap between the average of the estimate $\hat{\kappa}(p,n,M)$ and the true mean $\kappa(p,n)$. The latter is, in principle, unknown for arbitrary values of n. Otherwise, it would be unnecessary to seek to solve (9) by approximation. But for n=3, thanks to Cheng and Spiring Theorem 4 [8] combined with elementary rules of integration [17], the closed form solution $\kappa(p,3)$ is as follows:

$$\kappa(p,3) = \frac{24}{3+(1-p)^2} \tag{11}$$

Particularly, $\kappa(0.990,3) \approx 7.999733$. As such, the bias could be explicitly computed. Similarly, for n=4 $\kappa(p,n)$, based on (10), is given by:

$$\kappa(p,4) = \frac{50\sqrt{2}\pi}{9(2+\pi\sqrt{2}-2p)} \quad \text{for} \quad p \geq \left[1-\frac{\pi\sqrt{2}}{6}\right] \tag{12}$$

and $\kappa(0.990,4) \approx 5.5306588$. Figure 7 and Figure 8 contrast the magnitudes of the variance $\hat{\kappa}(p,n,M)$ and the corresponding squared bias for M running between 10000 and 120000. It is observed that, globally, both the variance and the bias tend decrease with the number of replications as illustrated by the associated regression lines. Moreover, the squared bias is negligible in magnitude compared with the variance. For large values of n, (e.g. n = 89, the size of the sample used in this work) Figure 9 illustrates to what extent the variance $Var(\hat{\kappa})$ is impacted by the number of replications: For M=10000 (The value adopted by Cheng and Spiring, [8]), the 200 Monte Carlo code runs result in a series of realizations represented by the yellow histogram with a mean value of 0.171925 and a variance equals to 4.56510^{-7} . When M is increased to 120000 (The value adopted by this work), the mean value remains almost unchanged; it is equal to 0.171885 but the corresponding variance reduces substantially to 4.53210^{-8}. This is visualized in the data spread shown by the dotted histogram superimposed to the previous one in Figure 9. And it means that, the approximate solution to (9) $\hat{\kappa}(p,n,120000)$ compiled in Table 2, is likely not to disperse too much around its mean value. Therefore, it could be considered as accurate enough for most practical purposes. These conclusions hold true for all numerical experiments conducted with various values of $p$and n.

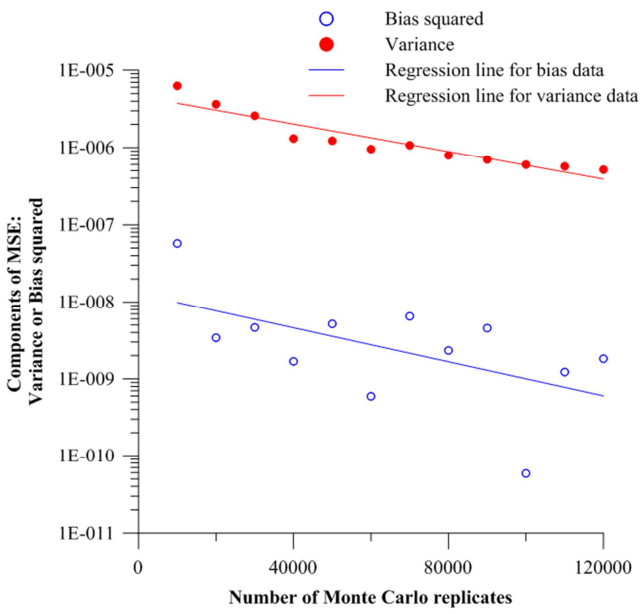


Figure 7. *Mean squared error components for the Monte Carlo estimate of the 0.99-quantile: Variance (red bullet) and bias squared (blue circle) versus the number of replications (sample size n=4).*

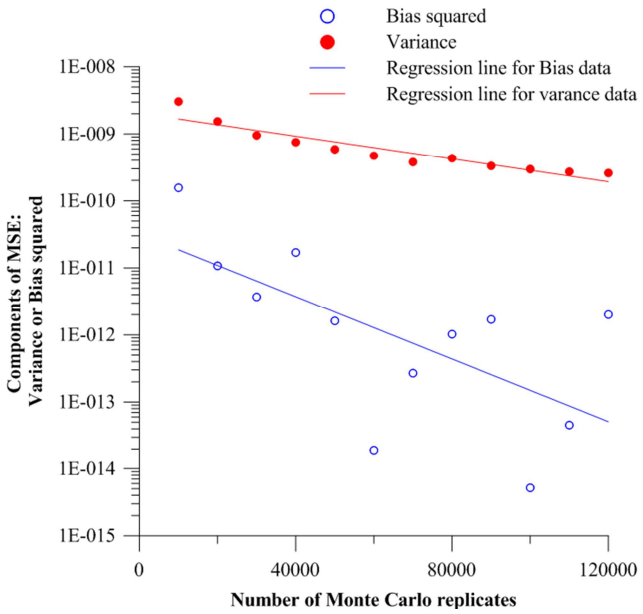


Figure 8. *Mean squared error components for the Monte Carlo estimate of the 0.99-quantile: Variance (red bullet) and bias squared (blue circle) versus the number of replications (sample size n=3).*

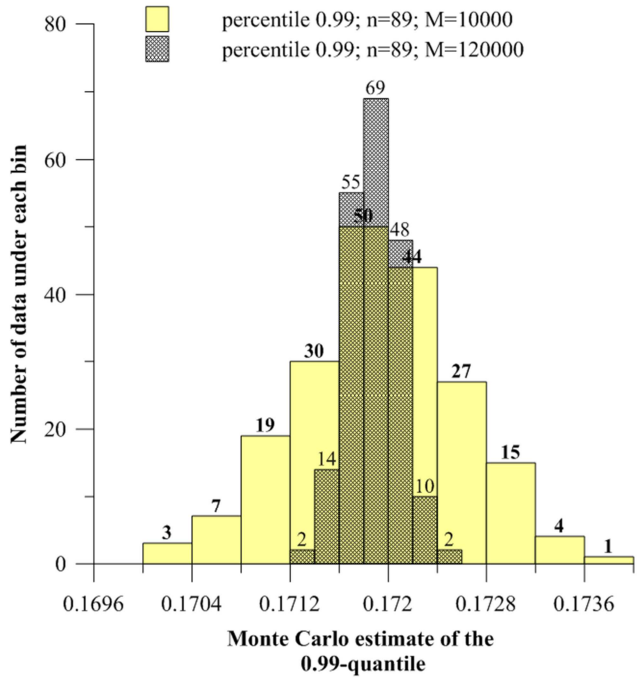


Figure 9. *The impact of the number of replications M on the spread/accuracy of the 0.99-quantile estimated by the Monte Carlo Method. The labels indicate the number of data points under each bin. The Monte Carlo code is run 200 times in each case.*

7. Conclusions

This work surveyed standpipe water consumption in 89 small villages (less than 1200 inhabitants) in the rural Meknes (Morocco). The involved population is about 37652 inhabitants. It can be concluded that:

(i) The observed data, being consistent with the uniform distribution, suggest that the population form one 'class' of consumers mainly concerned about covering their *basic* needs e.g. safe drinking water, cooking and ablutions, as it is substantiated by the facts on the ground. Actually, the other water intensive activities (e.g. livestock keeping) are generally satisfied by nearby raw-water sources.

(ii) In similar vein, the ANOVA presented in the second section shows that the population is not an effective explanatory variable of the variance in the observed standpipe water-use rates.

(iii) An estimate of the average daily standpipe water-use rate is 10.077 liter per capita with the 95% confidence interval of (4.249, 15.906). Such data are among essential inputs for *regional* water planning and management policies, especially under water scarcity in a deeply changing environment [18].

(iv) Also, the study extends and improves the Cheng-Spiring quantile Table 1 for testing hypotheses significance about the uniform distribution family. A simplified and user-friendly graphical support is proposed as well.

References

[1] Troin, J.-F. (2002). Maroc, Régions, pays, territoires. 1st Ed. Maisonneuve & Larose, Paris. ISBN: 2-7068-1630-9, pp: 502.

[2] Draper, N. R., & Smith, H. (1998). Applied Regression Analysis. New York. John Wiley & Sons, 3rd Ed. ISBN: 0-471-17082-8, pp: 705.

[3] Maddala, G. S. (1992). Introduction to Econometrics. New York. Macmillan, 2nd Ed. ISBN: 0-02-374545-2, pp: 637.

[4] Lloyd, E. H. (1952). Least squares estimation of location and scale parameters using order statistics, Biometrika, 39: 88-95. DOI: 10.2307/2332466.

[5] Wainwright, M. J. (2019). High-Dimensional Statistics, a Non-Asymptotic Viewpoint. Cambridge. Cambridge University Press, 1st Ed. ISBN: 978-1-108-49802-9, pp: 572.

[6] Cox, D. R. & Hinkley, D. V. (1974). Theoretical Statistics. London. Chapman and Hall, 1st Ed. ISBN: 978-0-412-12420-4, pp: 522.

[7] Emmert-Streib, F. & Dehmer, M. (2019). Understanding statistical hypothesis testing: The logic of statistical inference. Machine Learning & knowledge Extraction, 1, pp 945-961, DOI: 10.3390/make1030054.

[8] Cheng, S. W. & Spiring. F. A. (1987). A test to identify the uniform distribution, with applications to probability plotting and other distributions. IEEE Transactions on Reliability, R-36 (1): 98-105. DOI: 10.1016/0026-2714(88)90079-0.

[9] Goodman, J. (1983). Accuracy and efficiency of Monte Carlo method. Bechtel Power Corporation, 12400, E. Imperial Hwy. Norwalk, CA 90650, USA. (online paper) https://inis.iaea.org>-publication

[10] Rubinstein, R. Y. & Kroese, D. P. (2017). Simulation and Monte Carlo Method. Hoboken. John Wiley & Sons, 3rd Ed. ISBN: 9781118632208, pp: 396.

[11] Becker, R. A. & Chambers, J. M. (1984). S: An Interactive Environment for Data Analysis and Graphics. Belmont California. Chapman and Hall, 1st Ed. ISBN 10: 0-534-03313-X, pp: 580.

[12] Avramidis, A. N. & Wilson. J. R. (1998). Correlation-induction techniques for estimating quantiles in simulation experiments. Operation Research, 46 (4): 574-591. DOI: 10.1145/224401.224614.

[13] Miller, R. G. (1974). The jackknife – A review. Biometrika, 61 (1): 1-15. JSTOR, DOI: 10.2307/2334280.

[14] Ioannidis, J. P. A. (2005). Why most published research findings are false, PLoS Medicine, 2 (8) e124: 696-701. DOI: 101371/journal.pmed.0010124.

[15] Casella, G. & Berger, R. L. (2002), Statistical Inference. United States, Duxbury, 2nd Ed. ISBN-13: 978-0534243128, pp 498.

[16] Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. New York. Springer, 2nd Ed. ISBN-13: 978-0387848570, pp: 764.

[17] Gradshteyn, I. S., & Ryzhik, I. M. (2007). Tables of Integrals, Series and Products. San Diego. Elsevier, 7th Ed. ISBN-13: 978-0-12-373637-6, pp: 1220.

[18] Bolashvili, N., Karalashvili, T., Geladze, V., Machavariani, N., Karalashvili, A., Chikhradze, N., Giorgi, G. & Kartvelishvili. (2017). Sustainable management of water resources on the background of current climate change. Earth Sciences. Vol. 6. No. 5. pp 75-79. DOI. 10.11648/j.earth.20170605.13.