PROBABILITY DISTRIBUTION OF THE LENGTH OF THE SHORTEST TOUR BETWEEN A FEW RANDOM POINTS: A SIMULATION STUDY

Alexander Vinel
Daniel F. Silva

Department of Industrial and Systems Engineering
Auburn University
3301 Shelby Center
Auburn, AL 36830, USA

ABSTRACT

Inspired by an application in the field of on-demand public transportation, we perform a Monte Carlo simulation study on the probability distribution of the length of Traveling-Salesman-Problem (TSP) tours between small numbers of random locations. We consider a fixed convex region, where we generate a fixed number of random locations from a known probability distribution and find the corresponding euclidean TSP tour for them. We simulate this process extensively and perform both quantitative and qualitative analyses of the resulting experimental distribution for the TSP tour length. We show that, under certain assumptions on the shape of the region and the probability distribution of locations, the length of the TSP tour is well-approximated by a normal distribution, even for as few as five locations. Furthermore, we propose experimental models for estimating the mean and standard deviation of the tour length.

1 INTRODUCTION

This research is motivated by work we are currently advancing with a US-based public transportation operator to explore on-demand alternatives to traditional, fixed-route buses. Our industry partner recently piloted a service in a city in Florida, which works as follows. There are several bus routes that serve a point of interest, which is surrounded by low-density residential neighborhoods. In order to serve residents of those neighborhoods, the operator offers a shuttle that, for a fee, transports the user from the bus stop to any point within a small service area (about a 1-mile radius). This service is provided in small vehicles, shared by several passengers. The shuttle's route is determined by solving a Traveling Salesman Problem (TSP) between the bus stop and the assigned customers. This is similar to existing dial-a-ride systems, except it is integrated with the bus system, and coordinated using mobile technology.

The goal of our work is to determine the appropriate operation parameters such as fleet size, vehicle size, etc. to maximize profit. To do this, we are developing analytical and simulation models to evaluate the performance of the system, which require knowing the distribution of the duration of each shuttle trip as a function of the number of customers in the trip. We can approximate this value by using a similar random variable, X_n, defined as the length of the euclidean TSP tour between n points drawn at random from a known distribution in the given region. If we knew the distribution of X_n for each n, we could use it in the analytical models. For example, suppose there is only one shuttle available and if it is full, customers walk home. To estimate the distribution of the number of customers who will get shuttle service (vs. walk) we must estimate the number of customer arrivals to the bus stop during each shuttle tour; for this, we need to know both the bus arrival process (which we do) and also the distribution of the tour length. Of course, we could simply use the mean, which has been studied in the literature, but having distributional information about the tour length allows us to calculate distributions of the measures of performance, not just their

averages. Furthermore, having the distribution of X_n also allows us to simplify the simulation models by generating the corresponding random variates directly, as opposed to going through the cumbersome process of generating and solving thousands of TSP instances to optimality.

With that in mind, in this paper we explore the behavior of X_n. It is known that as $n \to \infty$, $X_n/\sqrt{n}$ converges to a constant for any convex region and continuous density function (Beardwood et al. 1959). Experimental results have suggested that for large n, X_n follows a normal distribution. However, because of the application we are considering, we are especially interested in small values of n, namely $n \le 15$, but little is known about the distribution of X_n in this case. We hypothesize that even for small values of n, X_n can be closely approximated by a normal distribution. We perform an extensive Monte Carlo simulation study and apply several fit tests. Our results suggest that the hypothesis holds for problems as small as $n = 5$ on the unit square, with points drawn from a uniform distribution. For a unit circle we make similar observations, with slightly larger discrepancies from normality. However, for anisotropic regions or for locations drawn from different distributions our hypothesis only appears to hold for $n \ge 10$ or larger (depending on the specific case). We also provide empirical models to estimate the mean and standard deviation of X_n as functions of n for the case of uniformly distributed locations in the unit square.

Ultimately, our goal is to use these distributions and the empirical estimates to model the length of each trip and incorporate these results into our analytical and simulation models to evaluate the performance of the current system and optimize the operation parameters. We hope these results will be useful to other researchers facing problems that can benefit from better understanding the behavior of small TSP instances between random points. We believe this is a common problem, appearing in dispatching for last-mile logistics, dial-a-ride systems, ATM replenishment, and many other applications.

The remainder of this paper is organized as follows. In Section 2 we review the relevant literature on the distribution of X_n, including convergence and concentration results. In Section 3 we describe our experimental approach. In Section 4 we present the results for the base case of the unit square with points drawn from a uniform distribution; we also provide experimental models for estimating the mean and standard deviation of X_n as a function of n, for the base case. In Section 5 we show how the distribution of X_n changes for different shapes of the service region and different probability distributions of locations. Finally, in Section 6 we summarize our results and discuss possible future research directions.

2 LITERATURE REVIEW

The Traveling Salesman Problem is perhaps the best known problem in combinatorial optimization (Applegate et al. 2006). The specific problem of determining the length of a TSP tour between points drawn independently and uniformly at random in the unit square (using euclidean distances) has been studied since at least the 1940s (Fejes 1940). However, Beardwood et al. (1959) are the first to achieve a major breakthrough for this problem, when they showed that

$$\lim_{n \to \infty} \frac{X_n}{\sqrt{n}} = \beta \quad \text{with probability 1.} \tag{1}$$

This classic result is well known and has been included in several books covering the TSP, combinatorial optimization, and probability theory; see, for example, Steele (1997). In their original work, Beardwood et al. (1959) showed that $0.625 \le \beta \le 0.9212$ (the upper bound is rounded up here). These bounds have been recently improved slightly by Steinerberger (2015). However, the exact value of β remains unknown.

In lieu of better theoretical bounds, over the years there have been several attempts at using simulation to provide an accurate estimate for β. In their original paper, Beardwood et al. (1959) generated tours by hand for 202-city and 400-city instances and estimated $\beta \approx 0.53\sqrt{2}$, a result which is often cited as 0.749, with arbitrary precision. Stein (1978) used Monte Carlo simulation over a (presumably small) set of instances and obtained $\beta \approx 0.765$. Because the TSP is NP-hard and this approach requires solving many large instances of the problem, some researchers have relied on heuristics to solve the necessary instances, yielding an estimated upper bound $\bar{\beta} \ge \beta$. For example, with this approach, Ong and Huang (1989) used

a 3-opt heuristic to get $\bar{\beta} \approx 0.74$. An alternate approach has been to use the Held-Karp bound (Held and Karp 1970) for the TSP to estimate a lower bound, then use it to calculate β. For example, Valenzuela and Jones (1997) approximated $\beta \approx 0.712 \pm 0.0002$ from two different Held-Karp estimates.

The above research covers the asymptotic behavior of X_n in detail, but does not address how it behaves for finite n. Applegate et al. (2006) discussed that for 10,000 randomly generated instances of 1,000 cities, the tour-length appears to follow a normal distribution. Beardwood et al. (1959) mentioned that they expected the tour length to be normally distributed, due to the central-limit theorem, but were unable to prove it. To the best of our knowledge, the exact distribution of X_n remains an open question, however several concentration results exist. Steele (1981) observed that the variance of the tour length is bounded independently of n. Dubhashi and Panconesi (2009) collect several bounds on the probability that $|X_n - \mathsf{E}[X_n]| > t$, which show that it decays exponentially as a function of n. These bounds strengthen the argument that for large n, X_n is well-approximated by a normal distribution. The tightest known bound on the tail probabilities for X_n is due to Rhee and Talagrand (1989), which states that there exists a unique constant K, such that for every n, $\mathsf{P}\{|X_n - \mathsf{E}[X_n]| > t\} \leq Ke^{-t^2/4K}$. This bound does not depend on n and can be useful for large n, but for small n it may be too loose. For example, for $n = 2$ it is obvious that for $t > 1/\sqrt{2}$ the above probability is zero, but the bound is greater. In this paper, we explore the distribution of X_n for n as small as three and observe that even in cases with $n = 4$ a normal distribution can be a reasonable approximation.

The above results are specifically for uniform-distributed locations in the unit square. Beardwood et al. (1959) proved a statement analogous to (1) for any bounded convex region, locations drawn independently from any probability measure with an almost-everywhere continuous density and in any dimension greater than two. The case of non-independent locations was shown not to converge, even for uniformly distributed ergodic sequences of locations, by Arlotto and Steele (2016).

The problem of estimating $\mathsf{E}[X_n]$ for differently shaped regions, with finite n, was addressed by Daganzo (1984) using regression models. Since then, several so-called continuous approximation models for estimating $\mathsf{E}[X_n]$ under different assumptions on the shape of the region or the distribution of locations have been proposed; for a review see Franceschetti et al. (2017). However, the distribution of X_n for general regions and location distributions has not really been studied in the literature.

Short TSP instances are a common feature or sub-problem in operations research models. In the literature, authors often cite and apply either asymptotic results or an estimate of $\mathsf{E}[X_n]$ that is adequate for the application at hand. However, in many applications the number of locations n is quite small, making it harder to use asymptotic or large-n approximations. The distribution of X_n is often ignored or assumed to follow a given form. We expect that the results presented in this paper will give researchers a more-precise way to estimate the distribution of X_n, particularly in applications with small n. Recent applications where this could be useful include ATM replenishment (Zhang and Kulkarni 2018), last-mile logistics (Brown and Guiffrida 2017), and dispensing medical supplies (Hudgeons 2018), not to mention traditional supply chain applications that use deterministic continuous approximations (Franceschetti et al. 2017).

In this paper, we consider the distribution of X_n for regions of different shapes, as well as different location distributions, for several values of n and note that the normality assumption holds even for small n. Furthermore, we consider the mean and standard deviation of X_n for uniform-distributed locations on the unit square. It is common in the literature to assume that $\beta \sqrt{n}$ is a good estimate for the $\mathsf{E}[X_n]$ (Eilon et al. 1971). However, it is not clear if this is a good estimate for small values of n. To the best of our knowledge, no attempts have been made at estimating the standard deviation $\sigma(X_n)$. We will present experimental results for estimating $\mathsf{E}[X_n]$ and $\sigma(X_n)$ as a function of n for small to medium values of n.

3 EXPERIMENTAL STUDY DESCRIPTION

The experiments are organized as a Monte Carlo simulation study. For a selected configuration of parameters, we generate a collection of TSP instances and solve them, recording the length of the optimal tour length. The locations for the TSP are selected by sampling from a pre-determined probability distribution over a selected region with euclidean distances between them. The types of regions used are: unit square (the

basic case, a square with side of length 1), unit circle, and rectangles with different ratios between the sides. We pick a value for the number of locations, *n*, in the TSP, varying from 3 (which is the first nontrivial case) to 300, and then generate 10,000 instances of TSP.

These 10,000 instances are then solved to optimality with Concorde TSP solver (Applegate et al. 2006). Concorde is accessed through R package *TSP* (Hahsler and Hornik 2007). The value 10,000 is selected to enable us to make significant conclusions, yet let us solve the problems to optimality in a reasonable time. The computations are run on a Windows 7 workstation with Intel Xeon E3-1241 processor and 32 GB of RAM. All statistical analysis was performed in R version 3.2.2. Empirical probability density functions (pdf), cumulative distribution functions (cdf), PP plots, QQ plots, and Cullen-Frey plots are created with the package *fitdistrplus* (Delignette-Muller and Dutang 2015).

4 BASE CASE: UNIFORM UNIT SQUARE

4.1 Distributional Results

As the base case, we consider a unit square with *n* random points distributed uniformly over it. As surveyed above, this case has received the most attention and there exist a considerable amount of asymptotic results that can be directly applied here. *We hypothesize that the observed empirical distribution of X_n will be well-described by normal distribution for modest values of n.* The purpose of this study is to test the hypothesis and outline the values of *n* for which we can make such a claim. We expect that while the distribution for $n = 3$ may display significant deviations from the perfect normal, as *n* increases, these deviations subside, so for $n = 300$ the normality of the empirical distribution should be more evident.

There are many ways to test for normality, usually characterized as either graphical methods or goodness-of-fit tests. Graphical methods include various approaches to visualize empirical distributions, highlighting properties that are specific to normal distributions. These methods are not as quantitative as the goodness-of-fit tests, but offer some significant advantages, such as being more intuitive and allowing the reader to judge the distribution assumption by themselves (Peat and Barton 2008). Further, graphical methods are usually robust to outliers and can accommodate large sample sizes (Field 2009).

The most widely used goodness-of-fit test is the Shapiro-Wilk test, as it has been shown (empirically) to be the most powerful normality test (Yap and Sim 2011). At the same time, it has been pointed out by several authors that goodness-of-fit tests are often ill-suited for testing very large samples, as even the smallest deviations from normality may be significant (Field 2009). In fact, it has been argued that those methods miss the point, which is not whether the data follows a true normal distribution, but rather whether the data is close enough to a normal distribution to make statistical inference in the specific application (Rochon et al. 2012). In our experiments, the Shapiro-Wilk test gave very inconsistent results. We observed that for the sample sizes considered (order of thousands) the *p*-values reported by the test were extremely sensitive to the specific sample. Furthermore, even on samples generated from a true normal distribution, we would reject the normality hypothesis on a regular basis as long as the sample size was large enough. Based on these observations and following the advice from the literature outlined above, we conclude that for the purpose of this study the Shapiro-Wilk test is too sensitive and hence we do not report goodness-of-fit statistics here. Instead, we rely on the following graphical methods.

Empirical pdf and cdf of the distributions, compared with their fitted normal counterparts. These allow us to verify that the most general properties of the empirical distribution (number of modes, symmetry, etc.) are in accordance with normality. These are not suitable to analyze the tails.

PP plots. These graph the theoretical cdf on the x-axis of a plot, against the empirical cdf on the y-axis. If the data are normally distributed, the points should be very close to the $x = y$ line. In our experiments, PP plots confirm the results from the empirical pdf and cdf, but don't provide any information about the tails, so we do not report them here.

QQ plots. These are constructed by plotting the inverse normal (with the sample mean and variance) of each quantile on the x-axis of a plot, against the standardized (with the sample mean and variance)

values of the empirical observations on the y-axis. If the data are normally distributed, the points should be very close to the $x = y$ line. Because the intervals for the quantiles are evenly distributed, QQ plots reveal discrepancies in the tails more clearly than other methods.

Higher moment analysis. A normally distributed random variable has zero skewness and kurtosis equal to three. Both can serve as measures of tail heaviness for an empirical distribution, so if an observed distribution has both measures close to zero and three respectively, one can conclude that a normality assumption is valid for a lot of practical purposes. To visualize this analysis we employ Cullen Fray plots (Cullen and Frey 1999), see Figures 3 and 7. Here, distribution families are projected on a plane with kurtosis and the square of skewness as coordinates. Given an empirical distribution, we can calculate sample skewness and sample kurtosis projecting it on the same plane. In the figures used in this study we further use bootstrap methods to account for variability of the estimators.

4.1.1 Empirical Density and Cumulative Distribution

Figure 1 gives the empirical densities and cdfs of X_n for $n = 3, 5, 10, 300$ along with corresponding fitted theoretical normal densities and cdfs. As noted above, these graphs are not very well-suited to make conclusive arguments. On one hand, we can observe that for all cases the distributions are unimodal, with mean, mode and median that are close to each other. Further, compared to the fitted normal pdf and cdf in all cases the discrepancy is small. On the other hand, as will be clear from a more careful analysis of QQ plots, these observation by themselves are insufficient to make strong conclusions.
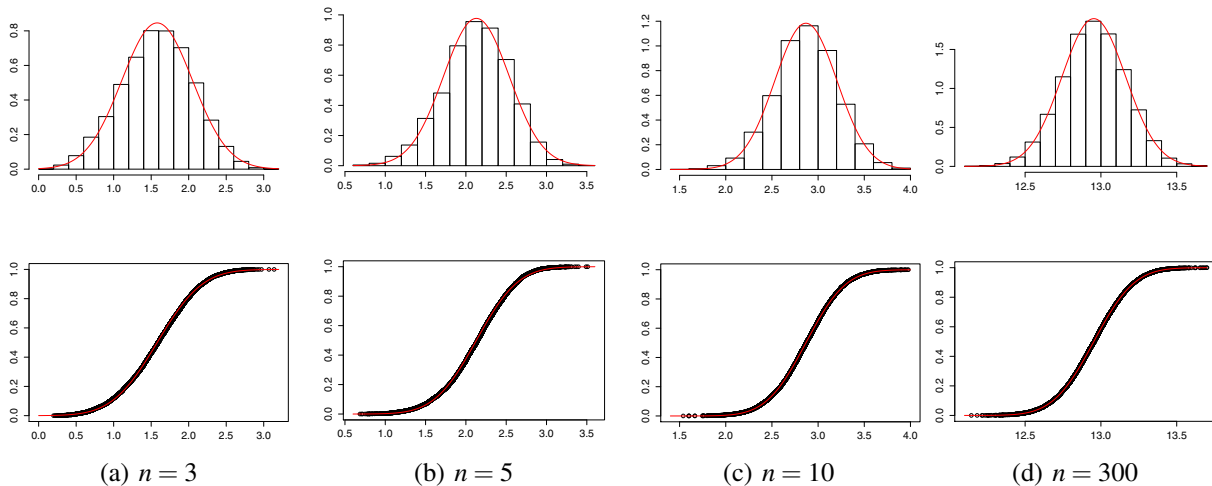

Figure 1: Empirical (*x*-axis) and fitted (*y*-axis) pdfs (top) and cdfs (bottom) for tours over a unit square.

4.1.2 QQ Plots

To consider deviations in the tails, Figure 2 provides QQ plots for a number of values of *n* studied. For reference we also include a QQ plot generated based on sampling 10000 realizations from true normal distribution (with R function *rnorm*) given in Figure 2a. Consider Figure 2b, which corresponds to $n = 3$. Here we observe a typical violation of normality, that is often seen in practice when considering QQ plots. Specifically, the left tail of the empirical distribution is consistently heavier than predicted by a normal distribution. The right tail, exhibits a similar behavior, but to a smaller extent. This observation conforms to our expectations, since TSP tours over three random points on a unit square are always triangles, and it is natural to expect that "smaller" triangles are more prevalent, compared to "larger" ones. At the same time, starting with $n = 4$ the discrepancies from a straight line on the QQ plots are less pronounced.
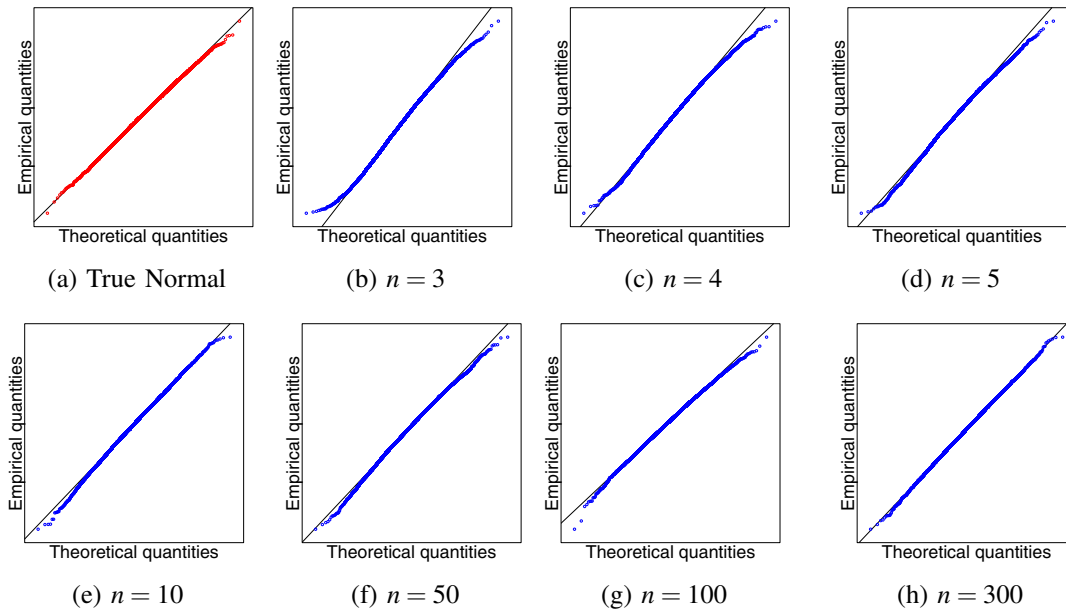
Figure 2: QQ plots for tours over a unit square.

4.1.3 Higher-Moment Analysis

Figure 3 gives Cullen-Frey plots for unit square and $n = 3$ and $n = 5$. We omit the plots for other values of n, as they are essentially indistinguishable from Figure 3. We can observe that, while for $n = 3$ the sample kurtosis is slightly less that 3, it is significantly closer to normal in skewness-kurtosis coordinates compared to most other theoretical distributions. Further, as soon as we consider $n \geq 5$ the empirical distribution coincides with the normal in these coordinates. Note that bootstrapping sample skewness and kurtosis does not have a significant effect. Compare this case with Figure 7 which we analyze in detail in Section 5.3, where the empirical distribution is clearly distinct from normal.
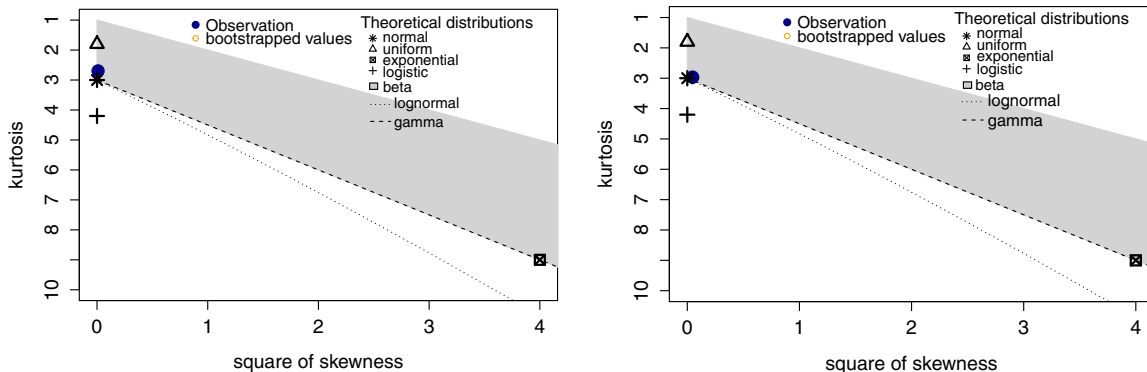

Figure 3: Cullen Frey plots for tours over a unit square for $n = 3$ (left) and $n = 5$ (right).

4.2 Parameter Estimation

The second goal of the study is to estimate the sample mean and sample variance of X_n as functions of the number of locations n, and verifying whether such functions can be used as reliable estimators for small values of n. As discussed above, asymptotic behavior of the expected value is well established in the

literature. Namely, we can conclude that the sample mean should be estimated as $E[X_n] \simeq \beta\sqrt{Sn}$, where $\beta \approx 0.712$ and S is the area of the region, i.e., $S = 1$ for a unit square. While there are some claims that this estimation is tight for moderate values of n (Eilon et al. 1971), no studies, to the best of our knowledge, have established how well it holds for very small n. Further, while Rhee and Talagrand (1989) suggested that variance of X_n in the unit square is bounded as $n \to \infty$, we are not aware of any claims regarding its non-asymptotic behavior.

In order to address this goal we have estimated sample average and sample standard deviation based on 10,000 samples collected for $n = 3, 5, \ldots, 300$, and then fit a nonlinear least-squares model. We use $E[X_n] \sim b\sqrt{n}$ and obtain a least-squares estimate for the value of b. In our experiment, the best fit is achieved for $b = 0.768$ with residual sum-of-squares of 0.9679. In the absence of a theoretical result for the nonlinear model for the standard deviation, we consider $\sigma(X_n) \sim c + d/\sqrt{n}$. The parameters are estimated as $c = 0.4985$ and $d = 0.1826$ with residual sum-of-squares: 0.00016. The observed and fitted values for all n are given in Table 1 and Figure 4.

First, note that the value $b = 0.768$ is significantly higher than $\beta \approx 0.712$ which follows from the literature. However, it is consistent with the early estimate of Stein (1978), which probably used small n, as we have here. Further, with this value, the fitted curve underestimates the sample average for $n < 100$, suggesting that an even higher value would be required for better accuracy if n is small. This is not surprising, since the best estimates for β are derived based on experiments on very large TSP instances. This means that while $\beta = 0.712$ is the true (or close to true) value of the scaling constant, it is not necessarily the best estimate to use in practice. The proposed model, $\sigma(X_n) \sim c + \frac{d}{\sqrt{n}}$, seems to be very well supported by the observed experiments. Namely, the fitting error is small and there is no consistent under- or over-estimation for smaller values of n. It is also in accordance with asymptotically bounded variance, as reported in the literature.

Table 1: Empirical and fitted values for mean and standard deviation of X_n for various n.

		3	5	7	10	30	50	100	200	300
$E[X_n]$	empirical	1.58	2.13	2.47	2.87	4.56	5.69	7.76	10.70	12.95
	fitted	1.33	1.72	2.03	2.43	4.21	5.43	7.68	10.86	13.31
$\sigma(X_n)$	empirical	0.47	0.41	0.36	0.34	0.28	0.26	0.23	0.22	0.21
	fitted	0.47	0.41	0.37	0.34	0.27	0.25	0.23	0.22	0.21

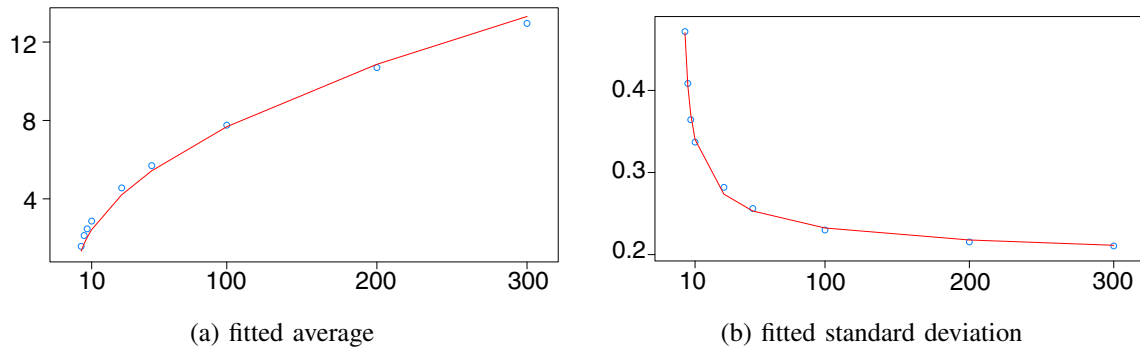(a) fitted average

(b) fitted standard deviation

Figure 4: Fitted average and standard deviation of X_n (y-axis) as a fuction of n (x-axis) in the unit square.

5 EFFECTS OF REGION SHAPE AND PROBABILITY DISTRIBUTION

Beardwood et al. (1959) showed that as n grows, there is a very limited effect of size and shape of the region on the distribution of tour length. Namely, $E[X_n] \sim \sqrt{Sn}$, where S is the area of the region, while

the shape is irrelevant as long as it is convex. In this section, we study whether similar conclusions can be made in the non-asymptotic sense.

5.1 Effect of Shape: The Unit Circle

A unit circle (a circle of radius one) is a natural counterpart to the unit square. Both regions are symmetric around center and are approximately isotropic. Hence, we expect that the conclusions in Section 4 will carry through to this case. Figure 5 gives the QQ plots for $n = 3, 5, 10$, and 100 for locations on unit circle. Surprisingly, the QQ plot for $n = 5$ shows a slightly greater deviation from normality in the unit circle than in the unit square. However, for all $n \geq 10$ the behavior in the circle and the square appears practically identical. We do not report empirical pdfs and cdfs or Cullen-Frey plots for the circle, because they are virtually indistinguishable from the ones obtained for the unit square. Table 2 reports the empirical data for the estimated average and standard deviation. We report the ratio between the values obtained for unit circle and unit square. Since the area of a unit circle is π, the theoretical result in Beardwood et al. (1959) suggests that for the sample average this ratio should be $\sqrt{\pi} \approx 1.77$. Our experiments with squares of varying sizes suggest that a similar relationship exists for the sample standard deviation.

Table 2: The ratio of average and standard deviation of tour length over unit circle to unit square.

ratio	3	5	7	10	30	50	100	200	300	theoretical value
average	1.71	1.73	1.73	1.72	1.73	1.74	1.75	1.76	1.76	1.77
std. dev.	1.70	1.60	1.58	1.56	1.59	1.64	1.70	1.72	1.72	–

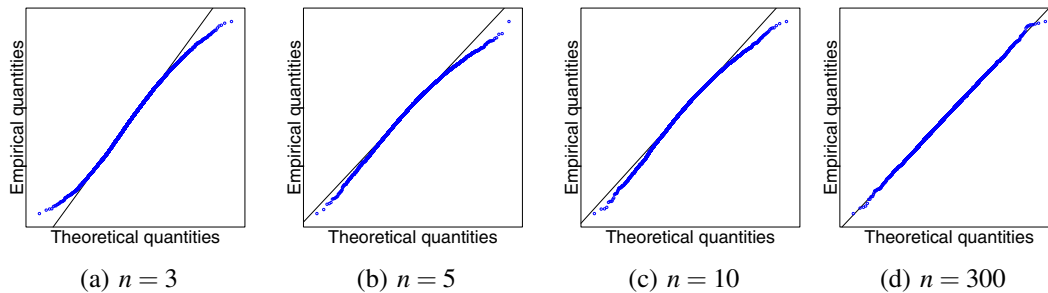(a) $n = 3$ (b) $n = 5$ (c) $n = 10$ (d) $n = 300$

Figure 5: QQ plots for tours over unit circle.

5.2 Effect of Shape: Rectangular Regions

Now let us compare this performance with the case of a rectangular region. Namely, we consider rectangular regions with length 1 and width a, and set the value of a to 2, 10, and 100. Naturally, as $n \to \infty$ locations uniformly distributed on a rectangular region will fill it entirely, explaining why the shape of the region is asymptotically irrelevant. On the other hand, a small value of n locations, which is the focus of this study, is not sufficient to fully erase the shape effects. Hence, we expect that we will see significant deviations from the conclusions established earlier when a is large. On the other hand, we expect that these deviations will disappear for sufficiently large n.

We report QQ plots in Figure 6 for various values of n and a and the sample average (sample standard deviation) ratios between the rectangular region and unit square in Table 3. As with the circle, theoretical results suggest that the ratio for the sample average should approach $\sqrt{S}$ (where S is the area), and in this case $S = a$. Our experiments exhibit the expected behavior. For $n = 3$ the empirical distributions are significantly not normal for all values of a and the corresponding ratios are very far from the theoretical predictions. On the other hand, for $n = 100$ with $a = 2$ and $a = 10$ the distributions are essentially normal

and the theoretical predictions for sample average and sample standard deviation are well satisfied. For $a = 100$ (QQ plot not reported here), even for $n = 100$ the underlying distribution is significantly not normal.
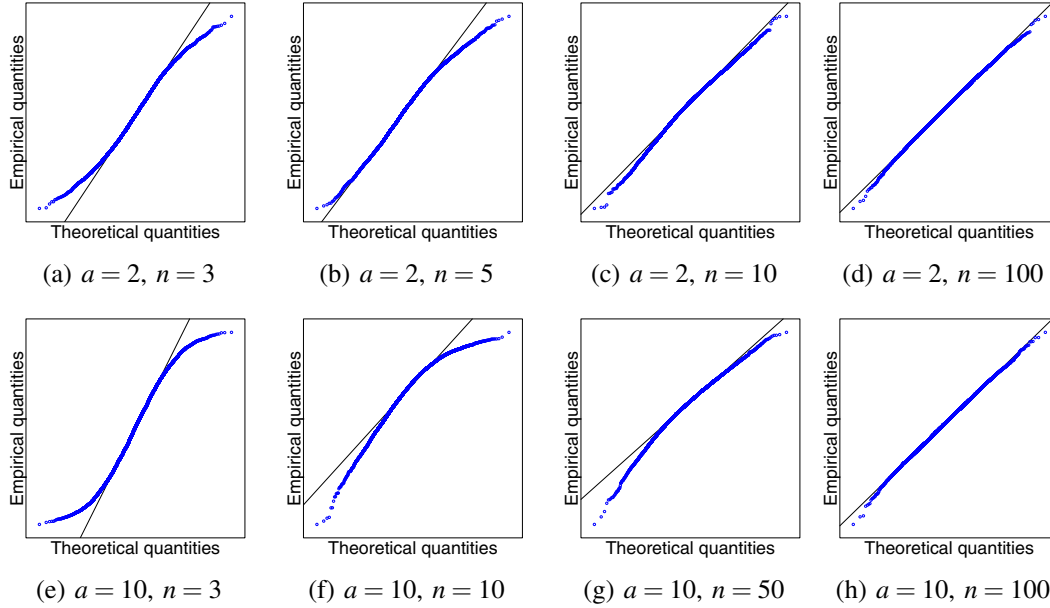

Figure 6: QQ plots for tours over rectangular regions.

Table 3: The ratio of average and standard deviation of tour length over rectangular region to unit square.

ratio	a	3	5	7	10	30	50	100	theoretical value
average	2	1.53	1.54	1.53	1.50	1.43	1.42	1.42	1.41
	10	6.51	6.38	6.20	5.87	4.47	3.94	3.50	3.16
	100	63.77	62.71	60.71	57.17	41.13	33.86	25.56	10.0
st. dev.	2	1.74	1.66	1.59	1.47	1.38	1.42	1.43	–
	10	9.34	8.57	7.88	6.46	3.17	2.68	2.87	–
	100	94.65	86.77	79.74	65.61	30.54	21.19	11.94	–

5.3 Effect of Probability Distribution

Beardwood et al. (1959) established that an expression similar to (1) holds for general probability distributions over bounded regions. To explore how this translates to the non-asymptotic case, we consider two types of non-uniform distributions: locations distributed isotropically and normally with variance 1 around the origin (referred to below as *centered*); and isotropic locations with distance from origin distributed normally with average 10 and variance 1 (referred to below as *disk*). In principle, both cases correspond to unbounded regions, i.e., asymptotic results due to Beardwood et al. (1959) do not apply. On the other hand, for small n the unboundedness of the region should not be significant, so we use these simple examples to explore the effect of location distribution. A more thorough study exploring a wider variety of distributional assumptions is required, but is beyond the scope of the current work.

Figures 7 and 8 report Cullen-Frey and QQ plots respectively. We can observe that the effects due to normally distributed locations are very similar to the ones observed for changes in shape. Specifically, a different underlying probability distribution significantly changes the behavior of the empirical distribution of X_n. However, as n grows, this effect subsides and X_n approaches normality.
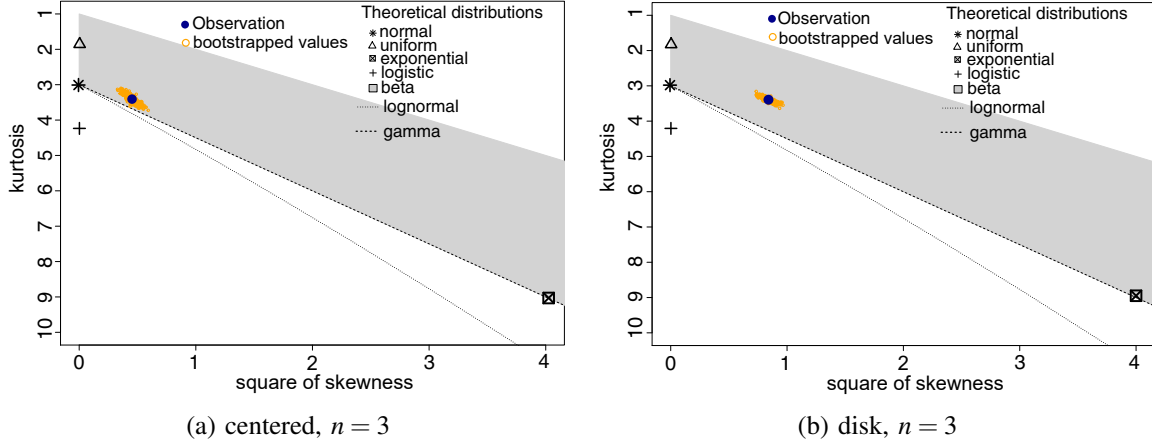
(a) centered, $n = 3$ (b) disk, $n = 3$

Figure 7: Cullen Frey plots for tours over regions with normal distribution of locations.

(a) centered, $n = 3$ (b) centered, $n = 10$ (c) centered, $n = 30$ (d) centered, $n = 100$

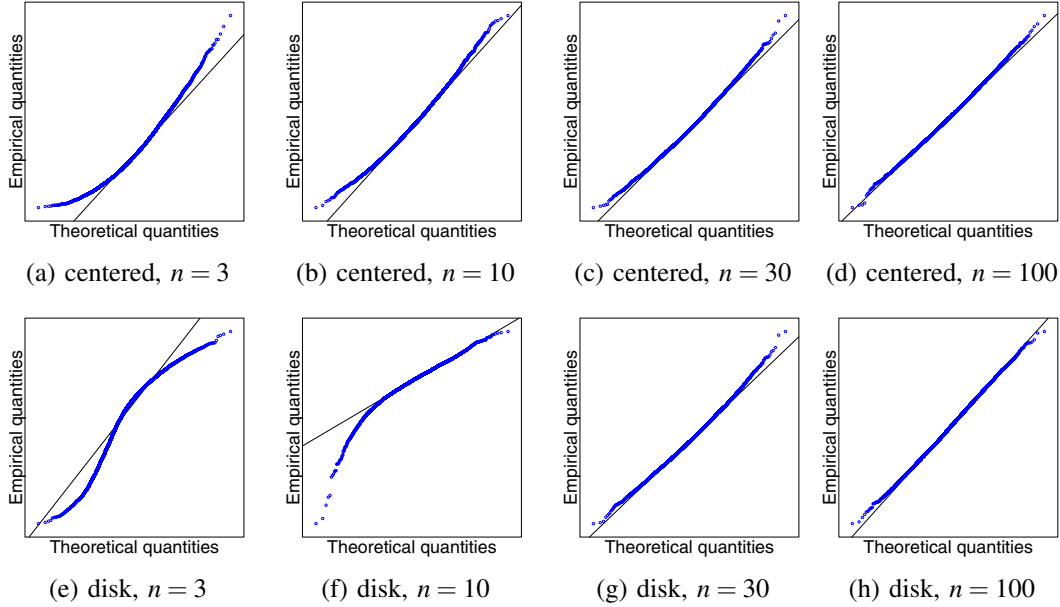(e) disk, $n = 3$ (f) disk, $n = 10$ (g) disk, $n = 30$ (h) disk, $n = 100$

Figure 8: QQ plots for tours over regions with normal distribution of locations.

6 CONCLUSIONS AND FUTURE WORK

It is clear from the aggregate of the analysis performed, that the experiments support our hypothesis that X_n is well approximated by a normal distribution, for an isotropic region (unit square and circle) with uniform distribution for values of n as low as at least 10. On the other hand, we can conclude that for $n = 3$ there exists a consistent deviation from normality, especially in the left tail of the distribution. For $n = 2$, the TSP tour length reduces to the distance between two random points in the region, and hence, one can derive an exact analytical expression for its distribution. A definitive conclusion is harder to make for $n \in [4, 10]$. Our experiments suggest that while there exist some deviations in tail behavior, as follows from the QQ plots, these are not necessarily substantial for the unit square. Further, the higher-moment analysis, as illustrated by Cullen-Frey plots does not find any significant deviation from normality. It is worth noting here, that when deciding on whether a normality assumption is valid, as with goodness-of-fit

tests, one must consider the specific application and whether the outcomes are sensitive to the tail of the distribution in question. Hence, we surmise that normality can be a valid assumption for some test cases for TSPs with as few as 4 locations, while it will be inappropriate in others, and hence sensitivity to tail behavior must be evaluated more closely. At this point we do not have a satisfactory explanation for the fact that deviations from normality are more prevalent in the unit circle, compared to unit square.

For a custom anisotropic region or different probability distribution, the hypothesis fails for small values of n. However, depending on the degree of anisotropy (or deviation from the uniform distribution), for values as small as $n = 10$, X_n can be well-approximated by a normal distribution. For highly anisotropic regions the critical value of n is larger. This observation can be explained through intuitive geometric arguments. Indeed, for a small n the anisotropy of the region can play a significant role (consider an extremely thin rectangle), yet as n grows a uniform sample tends to evenly fill the whole region reducing the effect of shape on the optimal TSP path.

For the base case, we also derived least-squares estimators for the sample average and sample standard deviation of X_n that are consistent with the best results available in the literature and provide a way to characterize the empirical distributions for very small values of n. To the best of our knowledge these are the first such estimates in the literature for small n. In particular, our estimate for the standard deviation is very well behaved and approaches a constant at rate $\sqrt{Sn}$. These estimates carry over well to square or circular regions of any size, but not to rectangular regions or other probability distributions.

Further work is required to gain a better understanding of the distribution of X_n with different underlying distributions of locations. There are myriad alternatives to consider here, and we have shown that the normality hypothesis does not necessarily hold for small n with non-uniform distributions. While not relevant to the motivating application, it may be of interest to test the hypothesis in higher dimensions, as the Beardwood et al. (1959) results suggest it might hold.

ACKNOWLEDGMENTS

REFERENCES

Applegate, D. L., R. E. Bixby, V. Chvatal, and W. J. Cook. 2006. *The Traveling Salesman Problem: A Computational Study*. Princeton, NJ: Princeton University Press.

Arlotto, A., and J. M. Steele. 2016. "Beardwood–Halton–Hammersley Theorem for Stationary Ergodic Sequences: A Counterexample". *The Annals of Applied Probability* 26(4):2141–2168.

Beardwood, J., J. H. Halton, and J. M. Hammersley. 1959. "The Shortest Path through many Points". *Mathematical Proceedings of the Cambridge Philosophical Society* 55(4):299–327.

Brown, J. R., and A. L. Guiffrida. 2017. "Stochastic Modeling of the Last Mile Problem for Delivery Fleet Planning". *Journal of the Transportation Research Forum* 56(2):93–108.

Cullen, A. C., and H. C. Frey. 1999. *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum Press.

Daganzo, C. F. 1984. "The Length of Tours in Zones of Different Shapes". *Transportation Research Part B: Methodological* 18(2):135–145.

Delignette-Muller, M. L., and C. Dutang. 2015. "fitdistrplus: An R Package for Fitting Distributions". *Journal of Statistical Software* 64(4):1–34.

Dubhashi, D. P., and A. Panconesi. 2009. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press.

Eilon, S., C. D. T. Watson-Gandy, and N. Christofides. 1971. *Distribution Management: Mathematical Modelling and Practical Analysis*. New York: Hafner.

Fejes, L. 1940. "Über einen geometrischen Satz". *Mathematische Zeitschrift* 46(1):83–85.

Field, A. 2009. *Discovering Statistics Using SPSS*. London: Sage publications.

Franceschetti, A., O. Jabali, and G. Laporte. 2017. "Continuous Approximation Models in Freight Distribution Management". *TOP* 25(3):413–433.

Hahsler, M., and K. Hornik. 2007. "TSP – Infrastructure for the Traveling Salesperson Problem". *Journal of Statistical Software* 23(2):1–21.

Held, M., and R. M. Karp. 1970. "The Traveling-Salesman Problem and Minimum Spanning Trees". *Operations Research* 18(6):1138–1162.

Hudgeons, A. 2018. *Dispensing Medical Countermeasures in Public Health Emergencies via Home Health Agencies and Points of Distribution*. Undergraduate honors thesis, Department of Industrial Engineering, University of Arkansas, Fayetteville, AR. http://scholarworks.uark.edu/ineguht/59/.

Ong, H., and H. Huang. 1989. "Asymptotic Expected Performance of some TSP Heuristics: An Empirical Evaluation". *European Journal of Operational Research* 43(2):231–238.

Peat, J., and B. Barton. 2008. *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. Malden, MA: Blackwell Publishing.

Rhee, W. T., and M. Talagrand. 1989. "A Sharp Deviation Inequality for the Stochastic Traveling Salesman Problem". *The Annals of Probability* 17(1):1–8.

Rochon, J., M. Gondan, and M. Kieser. 2012. "To Test or not to Test: Preliminary Assessment of Normality when Comparing Two Independent Samples". *BMC Medical Research Methodology* 12(1):81.

Steele, J. M. 1981. "Complete Convergence of Short Paths and Karp's Algorithm for the TSP". *Mathematics of Operations Research* 6(3):374–378.

Steele, J. M. 1997. *Probability Theory and Combinatorial Optimization*. Philadelphia, PA: SIAM.

Stein, D. M. 1978. "Scheduling Dial-a-Ride Transportation Systems". *Transportation Science* 12(3):232–249.

Steinerberger, S. 2015. "New Bounds for the Traveling Salesman Constant". *Advances in Applied Probability* 47(1):27–36.

Valenzuela, C. L., and A. J. Jones. 1997. "Estimating the Held-Karp Lower Bound for the Geometric TSP". *European Journal of Operational Research* 102(1):157–175.

Yap, B. W., and C. H. Sim. 2011. "Comparisons of Various Types of Normality Tests". *Journal of Statistical Computation and Simulation* 81(12):2141–2155.

Zhang, Y., and V. Kulkarni. 2018. "Automated Teller Machine Replenishment Policies with Submodular Costs". *Manufacturing & Service Operations Management* Forthcoming.

AUTHOR BIOGRAPHIES

ALEXANDER VINEL is an Assistant Professor in the Industrial and Systems Engineering Department of Auburn University. He holds a doctorate degree in industrial engineering from the University of Iowa. His research interests are in the areas of stochastic optimization and risk-averse decision making, with applications in transportation systems and data analytics. His email address is alexander.vinel@auburn.edu

DANIEL F. SILVA is an Assistant Professor in the Industrial and Systems Engineering Department of Auburn University. He holds a doctoral degree in Operations Research from the Georgia Institute of Technology. His research interests include stochastic modeling, queueing systems and stochastic optimization, with applications to transportation, logistics, and manufacturing. His email address is silva@auburn.edu