CS-BIGS 1(1): 15-25
http://www.bentley.edu/csbigs/vol1-1/cogdill.pdf

Analysis of Teller Service Times in Retail Banks

Travis Cogdill
University of North Texas, USA

Michael Monticino
University of North Texas, USA

Abstract. Retail banks typically staff teller lines as if all tellers had the same service capabilities. Likewise, queuing models to recommend staffing levels often assume that all servers have the same service time distribution and that this distribution is exponential. These assumptions are motivated more by operational and analytical convenience than supported by data. Until recently there have been little available data to test these assumptions and their impact on customer service. This article presents results from analyzing a large dataset of teller service times from a retail bank. The validity of common teller staffing assumptions are explored along with the impact that deviations from the assumptions have on staffing recommendations obtained from queuing models.

Introduction

Retail banks typically staff their teller lines as if all tellers had the same service capabilities. Assuming that each teller in a bank can cash a check, make a deposit, or fulfill a money order request with equal efficiency is more the result of insufficient teller performance data and the inability to effectively act upon data if available, than the reality of teller performance. Similarly, most applications of queuing models to recommend staffing levels assume that service times are exponentially distributed and that each staff member has the same service time distribution. This assumption allows analytic and numerical formulations that significantly simplify the process of finding (theoretically) optimal staffing levels. Previously, there has not been strong motivation for abandoning these assumptions given the relative scarcity of data. Times have changed. Banks now have sophisticated teller transaction platforms and IT architectures that allow tracking of all teller transaction activities. Service time data can, at least in principle, be routinely generated and integrated with staffing systems. Moreover, many classical queuing results have been extended to wider classes of distributions (Nozaki and Ross, 1978) and computer simulation has emerged as a viable tool in the analysis of queuing processes (Atlason et al., 2004). This paper has two main objectives.

- First, to investigate the validity of two common teller staffing assumptions: whether the class of exponential distributions provides useful approximations of teller service times; and whether tellers across an enterprise or within a branch have effectively equal service time statistics.

- Second, to determine what impact observed violations of these assumptions have on staffing recommendations obtained from queuing models.

We are in a fairly unique position to be able to address these objectives. A large dataset of teller service times (within branch lobbies) was obtained to validate staffing algorithms as part of a project with ARGO Data Resources, Inc. to develop an integrated Teller Performance and Workforce Scheduling product for retail banks. The dataset comprises teller service times for multiple transaction types in over 900 branches of a regional bank over the month of June 2005. Most service time analyses available in the literature (and few are available) are concerned with call-center data, as opposed to the face-to-face service engagements studied here. For example, recent work by Brown et al. (2005) analyzes telephone service times from a large call center to determine the adequacy of distributional assumptions.

This article is accessible to upper division undergraduate mathematics, statistics and operations research students (and their instructors) interested in data analysis that can have a real impact on business operations. The results in the article should also be of interest to operations research practitioners and banking industry management. The level of statistics necessary to follow and reproduce the analysis is basic to intermediate. Some knowledge of queuing theory is needed to appreciate the technical applications to staffing models, although the results themselves are readily accessible.

The next section discusses the data and investigates the appropriateness of standard distributional approximations of service times. Section 3 analyzes the assumption that tellers have equivalent service time capabilities. The impacts of using equal service time statistics and distributional assumptions to develop staffing recommendations are examined in section 4.

TELLER SERVICE TIMES

Data description. Teller service time data were recorded through the teller transaction platform system used by each teller to process customer transactions in a consumer retail bank with over 900 branches across thirteen states in the United States. The bank will be referred to as Regional Bank. Service time is defined as the elapsed time from when a teller initializes customer service through the teller transaction platform to when that customer engagement is ended on the platform. This may exclude some customer engagement time before and after a transaction. Although friendly greetings and small-talk are important for maintaining a positive customer relationship, including them within the service time can be misleading. For instance, tellers may be more likely to have extended conversations with customers during less busy periods in the branch, skewing service time statistics. One goal in analyzing service times is to recommend staffing levels. So it is better to focus on the actual transaction portion of the customer-teller engagement to get an accurate matching of teller performance to staffing needs. If a bank wanted to formally account for non-transaction customer-teller interaction they could add a "customer relations" period to the service times and incorporate this into staffing models (this can be done in the Workforce Scheduling product developed with ARGO – see www.argodata.com/StaffPerfrm.html for an overview of the product).

Service times are categorized by branch, teller performing the transaction, and transaction type. Regional Bank classifies transactions into forty-four (44) types. Transaction types include deposits, cashing checks, purchasing money orders, buying a savings bond, and making a credit card payment. The total dataset analyzed included over 5.5 million transactions processed across 956 bank branches over the month of June 2005. A total of 5403 tellers were represented, with the number of tellers recorded in branches ranging from two to twenty-eight. The three most common transaction types were cash check (not on bank), deposits and cash check (on bank), accounting for nearly 79% of all transactions. These three transaction types – in particular, deposits – will be the focus of our analysis. Cash check (not on bank) indicates a transaction such as cashing a payroll check, while cash check (on bank) is a check cashed against a customer account. We will abbreviate these latter two types of transactions by "cash check-NOB" and "cash check-OB." The data included with this article arise from a subset of the total dataset, giving cash check-OB, cash check-NOB and deposit transaction times (in seconds) for five branches with the number of tellers recorded in these branches ranging from four to eleven.

Table 1. Service time statistics (in seconds) aggregated across all branches of Regional Bank.

Transaction Type	Transactions	Minimum	Maximum	Mean	Median	Std. Deviation
Cash check (not on bank)	600952	22	299	52.13	47.0	20.165
Deposits	2904215	25	299	72.90	61.0	35.829
Cash check (on bank)	906905	23	299	54.10	49.0	19.339
All Transactions Types	5589777	8	299	67.60		33.795

Data cleaning. The data initially contained instances of transaction times of several hours. These were almost surely due to system or user error – for instance, a teller neglecting to log-off the teller platform system. Such values can significantly skew a teller's service time statistics. Similarly, a teller may erroneously begin a session as one transaction type, quickly realize the mistake and then end the session. This will result in a very short transaction time entry. Both these types of data values should be discarded before service time statistics are analyzed. There are several ways to automate the process of addressing erroneous outlier values for large datasets. A simple method is to set upper and lower bounds on possible service times and eliminate all data greater or less than the respective bounds. Other methods include trimming out the highest and lowest x% of service time observations for each teller, or discarding any data that is more than K standard deviation units from the average service time (see Rice, 1995, section 10.4.3 for a discussion of data trimming). We opted for the simplest approach since some branches had evidently configured the teller platforms to "time-out" after 300 seconds. That is, any service time greater than 300 seconds, whether because the teller neglected to end the session or because the transaction actually took longer than 300 seconds, was recorded by the branch as a 300 second transaction. We eliminated this censored data – any recorded service time of more than 299 seconds – since it involved only a small percentage (less that 0.5%) of the transactions. We set a conservative lower bound of 5 seconds that did not eliminate any transaction service times.

Aggregated service time statistics. We first examine service times aggregated across all branches of Regional Bank to gain context before drilling down to the branch and teller level. Table 1 gives service time statistics for the three most common transaction types. Note that the average service times for deposits differs significantly from cash check-OB or cash check-NOB. (The large number of transactions recorded easily ensures that this difference in mean service times is statistically significant.) The variation in the time required to complete a deposit transaction is also significantly greater than for cash checks, with a standard deviation 75% greater. This agrees with experience. Bank customers typically cash a single check, while deposits are more likely to vary in the quantity of checks and/or cash involved.

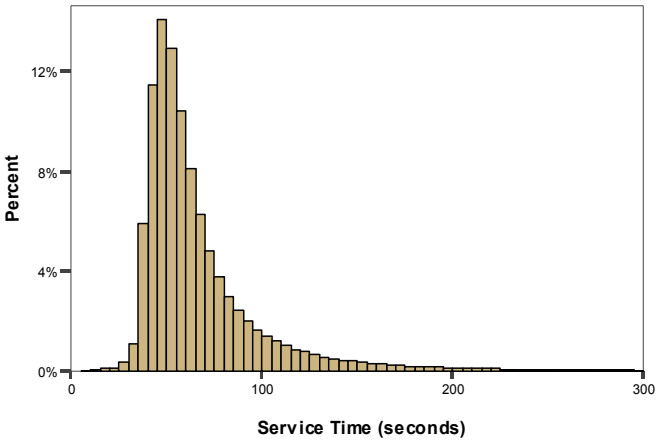

Figure 1. Distribution of service times aggregated across all branches of Regional Bank and transaction types.

Typically, a bank would use these aggregate statistics to characterize tellers in any of its branches. This is not unreasonable given employee training programs designed to produce consistent teller performance. It would also be common to aggregate service time data across transaction types. Figure 1 shows the distribution of service times aggregated across all Regional Bank branches and transaction types. The right-skewed, unimodal distribution has an inter-quartile range from 48 to 75 seconds, with nearly all service times greater than 30 seconds. The latter represents a fairly consistent minimum amount of time required to complete a transaction.

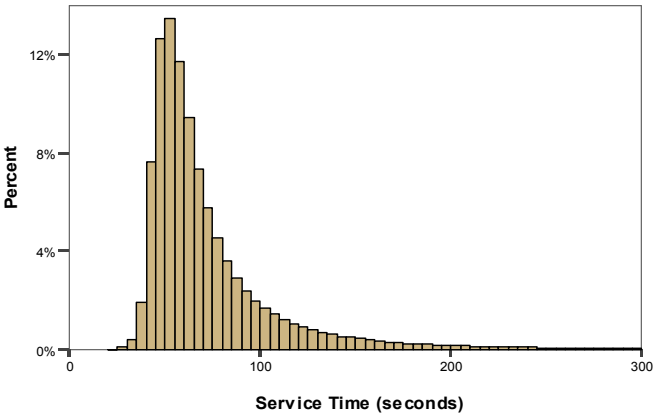

Figure 2. Distribution of deposit service times aggregated across all branches of Regional Bank.

Similarly shaped distributions are seen in Figures 2-4 for each of the three most common transaction types (aggregated across Regional Bank). The service time distributions for each transaction type are unimodal and right-skewed, reflecting a typical time to complete a transaction, and the fact that very long transaction times are more likely to occur than a very quick transactions. The service time distribution for deposits is more skewed and has a noticeably longer right tail than for either of the cash check transaction types. There is nearly a ten second difference between mean and median for deposit

services times and only a five second difference for cash check-OB (or NOB). Moreover, the spread between the 75th to 95th percentiles for deposits is from 80 to 146 seconds, while the spread between these percentiles for checks cashed is less than half as large – from 57 to 87 for cash check-NOB and from 57 to 86 for cash check-OB. Again, the data confirm experience that deposits tend to be more varied and complicated transactions to complete than cashing checks.

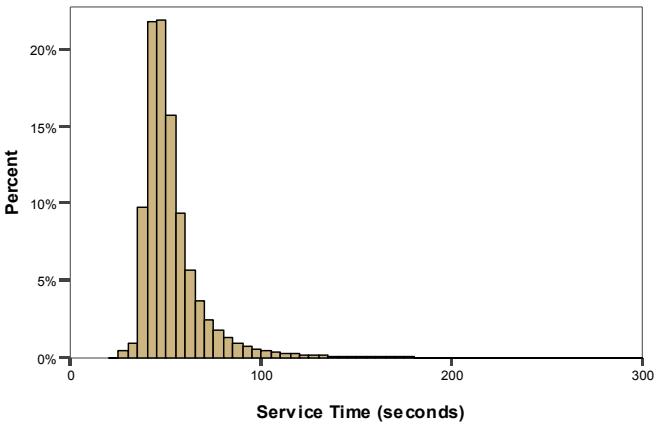

Figure 3. Distribution of cash check (on bank) service times aggregated across all branches of Regional Bank.

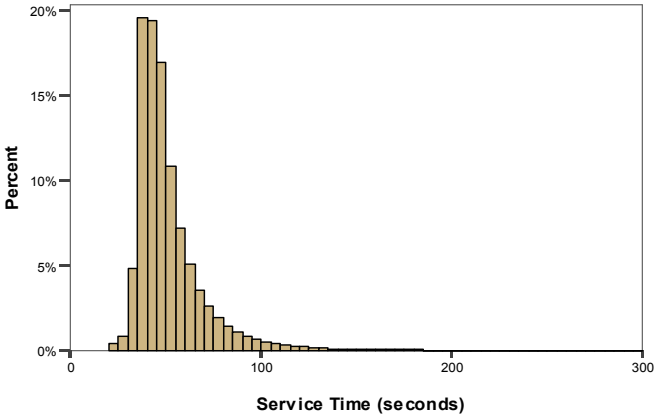

Figure 4. Distribution of cash check (not on bank) service times aggregated across all branches of Regional Bank.

Recall that an exponential distribution places non-zero probability on intervals [0,x), for all $x \in \mathbf{R}^+$, and has a standard deviation equal to its mean. Thus, an exponential distribution will place a higher probability on short service times and have a larger standard deviation, and hence a heavier tail, than observed in the data (see Table 1). Moreover, an exponential distribution is unable to represent the modal characteristics of the data. The mismatched properties of the exponential model and data are evident in the Q-Q plot of the empirical distribution versus the fitted exponential for deposits given in Figure 5 (the fitted exponential distribution is that which has the same mean as the data). The other transaction types have similar Q-Q plots. Kolmogorov-Smirnov tests can also be performed to test the hypotheses that the service times for transaction types come from exponential distributions. These tests easily reject the hypothesis. (In general, care must be taken when performing hypothesis tests with the very large samples sizes applied here – even reasonable models for data may be rejected, more the result of the stringent requirements imposed under large sample sizes than of the operational inappropriateness of the models.)

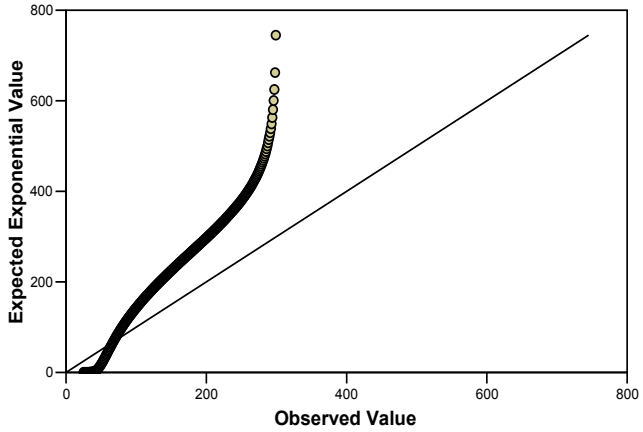

Figure 5. Q-Q plot comparing service times for deposits (aggregated across all Regional Bank branches) to fitted exponential distribution.

The reader can perform analogous descriptive and inferential analysis of service times for each transaction type with the attached data. Aggregating the data over all branches/tellers, just over tellers within a branch, or for individual tellers produces similar results that exponential distributions do not provide good models for teller service times.

The primary problem with the exponential model is its inability to capture the modal character of the service time data. A standard distribution often used to model processes exhibiting skewness and a mode is the lognormal distribution. For example, Brown et al. (2005) give data indicating that the lognormal provides a good model for call center service times. Similarly, Bolotin (1994) presents results suggesting that call duration for individual telephone customers are lognormal. Ulrich and Miller (1993) and Breukelen (1995) provide theoretical arguments, based on mathematical psychology, supporting reaction times being lognormal.

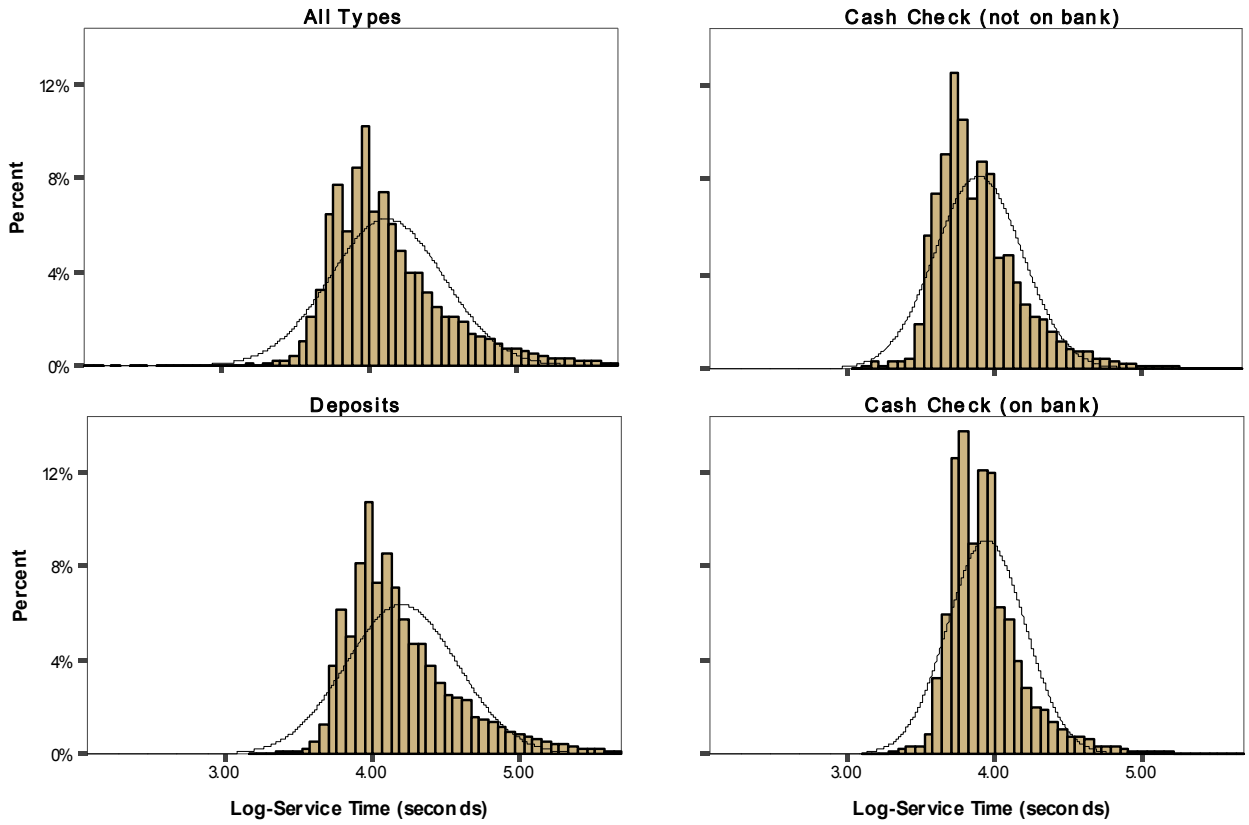

Figure 6. Distribution of the logarithm of service time for transaction types aggregated across Regional Bank. The top left panel shows log service times over all transaction types. Fitted normal densities are shown for each distribution.

Figure 6 shows the distribution of the logarithm of service times overlaid with a fitted normal distribution. Notice that the log service times are still somewhat skewed and that the fitted normal distributions are not able to fully approximate the peaked modes of the log service times.

Thus, neither of the distributions typically used to model service times – exponential nor lognormal – provide good approximations for the teller data given here. This holds true in the cases explicitly examined above for which the service times were aggregated across all tellers in all branches, as well as when data is aggregated only at the branch level and when service times for individual tellers are analyzed (assuming sufficient data for the specific teller). The operational effect of assuming these poorly fitting distributions when staffing tellers is examined in section 4.

BRANCH AND TELLER SERVICE TIMES

Now we examine service times within branches and for individual tellers and compare them to aggregated service time statistics. Banks commonly use benchmark service time statistics. These benchmarks may be derived by aggregating data across all tellers in all branches or more typically through some industry-wide analysis. The benchmarks are then assumed to characterize all tellers in all branches with staffing decisions made accordingly. Most of the analysis presented here will be focused on deposits. Similar results are obtained for the other transaction types.

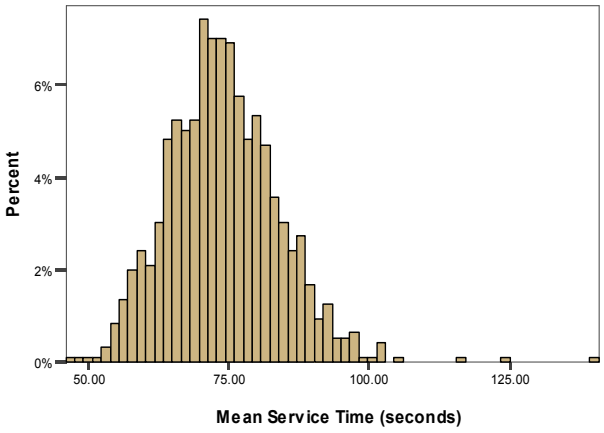

Figure 7. Branch-average service times for deposits.

There is a relatively large range in average service times across the branches as well as among tellers. The average time that it takes tellers in specific branches to complete a deposit ranges from 46 seconds to 141 seconds. Figure 7 gives the distribution on branch-average service times for deposits across Regional Bank branches (branch-average service time is the average of all deposits completed within that branch). Aggregating data across all tellers in all branches would provide a benchmark for deposit service times for Regional Bank of 72.9 seconds (see Table 1). The 10th percentile of the branch-average deposit service time across branches is 62.1 and the 90th percentile is 86.3 – so at least 20% of branches have branch-average deposit service times more than 10 seconds off the benchmark, and some markedly greater. While the difference from the benchmark may not be large for a single transaction, the difference can have a significant operational impact for the number of transactions handled during typical busy periods at a branch. Average service time differences between branches are not necessarily an indication that tellers are poorly trained. Some branches may have many more small business customers making complicated deposits. Similarly, some branches may have drive-up teller service and only customers requiring more attention may use the lobby tellers. Table 2 shows the deposit service time statistics for the branches represented in the attached dataset. Note the 20 second difference between average service times between Branch 85 and 267, and the fact that the standard deviation for Branch 586 is 20 seconds greater than Branch 85.

Table 2: Deposit service time statistics for branches in attached dataset.

Branch ID	Number of Deposits	Mean	Median	Std. Deviation
62	1746	67.14	60.00	27.73
85	8561	57.27	50.00	25.24
267	1763	77.89	68.00	32.68
443	2461	60.70	54.00	23.76
586	1356	75.25	58.00	45.45
Total	15887	62.71	54.00	29.45

The higher average service times and larger standard deviations of Branches 267 and 586 are largely driven by a single teller in each branch. Teller 4796 in Branch 267 accounted for 32% of deposits processed. This teller had an average service time of 93.9 seconds, more than 20 seconds slower than the other tellers in the branch. While, in Branch 586, teller 56 processed 45% of deposits and had an average service time of 90.9 seconds with a standard deviation of 52.4 (teller 2623 also underperformed, but handled fewer transactions) – this was 40 seconds slower than the teller processing the next greatest number of deposits (teller 299).

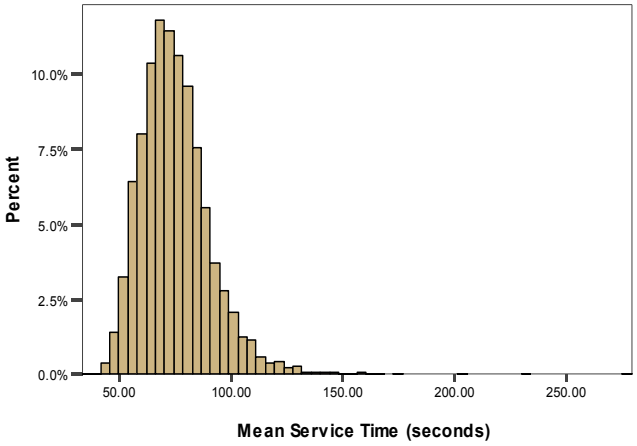

Figure 8. Average teller service times for deposits.

Table 3 shows summary statistics for individual tellers in each sample branch. As already indicated, there are notable differences among tellers. Average deposit service times range over almost 18 seconds among tellers in Branch 85. A branch manager would be able to determine whether these differences are due to poor performance or because a particular teller regularly handles more complicated transactions – for example, a branch may have a designated commercial customer line. Average teller service times across Regional Bank are shown in Figure 8. The interquartile range for average service times (deposits) across all Regional Bank tellers is from 64.5 to 83.3 seconds, and the 10th to 90th percentiles range from 57 to 94 seconds. As seen in the next section, staffing as if all tellers performed at the benchmark level would lead to long customer waits if tellers were actually at the 90th percentile and would overstaff the lobby if the tellers performed at the 10th percentile.

STAFFING AND QUEUING APPLICATIONS

The analysis above indicates that common assumptions do not hold for the teller service data examined here. Service time distributions are not well approximated by exponential models, and while a (two-parameter) lognormal model fits better, it is still not able to completely capture all data features. There are also significant performance differences among tellers within a branch and across branches. The latter observation calls into question the standard bank practice of treating and scheduling tellers as if they had the same service

capabilities. The statistician George Box remarked that "all models are wrong, just some are useful." So, the relevant question is whether the deviations from standard assumptions have any important effect either on the evaluation of service performance or on the staffing level needed to meet customer service goals.

Impact of distribution assumptions. Three M/G/1 queuing systems are simulated to explore the impact of making exponential or lognormal assumptions about service times. The queuing systems differ only in the service time distributions assumed. One system uses the empirical service time distribution for deposits aggregated across Regional Bank. The other two systems assume an exponential distribution and lognormal distribution fitted to the deposit data, respectively. Each system is simulated over sixty minutes with a customer arrival rate

Table 3: Deposit service time statistics for tellers in attached dataset.

Branch ID	Teller ID	N	Median	Mean	Std. Deviation
	4820	671	60	65.5	16.8
62	5174	28	73	94.2	52.4
	5211	400	69	83.1	42.9
	5286	647	54	57.8	15.9
	Total	**1746**	**60**	**67.1**	**27.7**
	70	414	48	53.1	19.6
	357	1237	53	57.7	17.9
85	472	1718	54	62.0	26.2
	3678	1411	52	63.5	32.1
	3983	3405	43	51.3	21.8
	4837	376	60	69.2	31.4
	Total	**8561**	**50**	**57.3**	**25.2**
	2503	850	65	71.7	25.1
267	2556	341	58	66.3	26.3
	4796	572	80	93.9	39.4
	Total	**1763**	**68**	**77.9**	**32.7**
	2230	32	55	63.9	22.3
443	2958	782	49	51.5	12.3
	3732	854	66	74.2	31.3
	4208	793	51	55.1	14.6
	Total	**2461**	**54**	**60.7**	**23.8**
	56	607	72	90.9	52.4
	63	58	64	72.1	23.8
	87	5	43	45.6	11.8
586	299	416	47	50.5	14.0
	321	1	51	51.0	.
	2623	145	68	92.2	53.6
	3023	33	51	51.3	8.6
	3327	20	60	71.1	36.5
	4353	4	59	59.3	2.1
	4424	67	54	69.6	40.1
	Total	**1356**	**58**	**75.2**	**45.4**

It might be expected that an exponential model would underestimate the average waiting time since it places a higher probability on short service times. However, as of 30 per hour (mean inter-arrival time of 2 minutes) and no customers in queue at the start of the simulation. The customer service measure evaluated is the average time that a customer spends waiting in line. shown in Figure 9, the exponential model actually leads to longer average waiting times than for the empirical distribution. This is due to the larger variance of the exponential distribution. On the other hand, the fitted lognormal model provides a much closer approximation to the average waiting time given by the empirical distribution. The horizontal asymptotes in Figure 9 are the average waiting times once the system has reached stationarity. It is interesting to note that the queuing processes with the empirical distribution and the lognormal model converge faster to their respective stationary average waiting times than does the exponential model – this is again because of the larger variance of the exponential service time model.

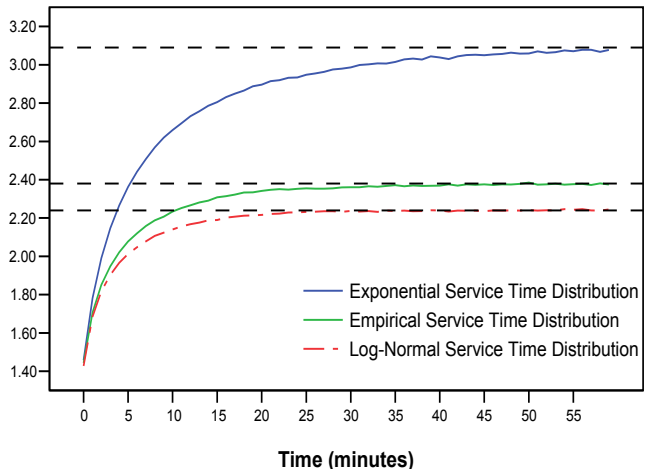

Figure 9. M/G/1 queue average waiting times in minutes (vertical axis) using empirical service time distribution for deposits, an exponential model and a lognormal model.

The ultimate impact of the distributional assumptions is upon the staffing decisions made under them. For example, suppose that Regional Bank had a customer service goal of keeping the average time that a customer waits in line at 2.5 minutes or less. Then, as indicated in Figure 9, a staffing system based on the exponential model would recommend that at least two tellers would be needed, while basing the system on the empirical (data) distribution would indicate that one teller is sufficient.

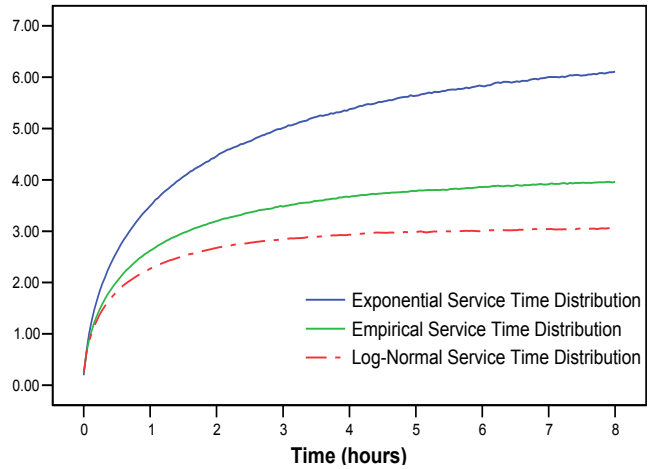

Figure 10. M/G/3 queue average waiting times in minutes (vertical axis) for empirical service time distribution for deposits, exponential model and lognormal model.

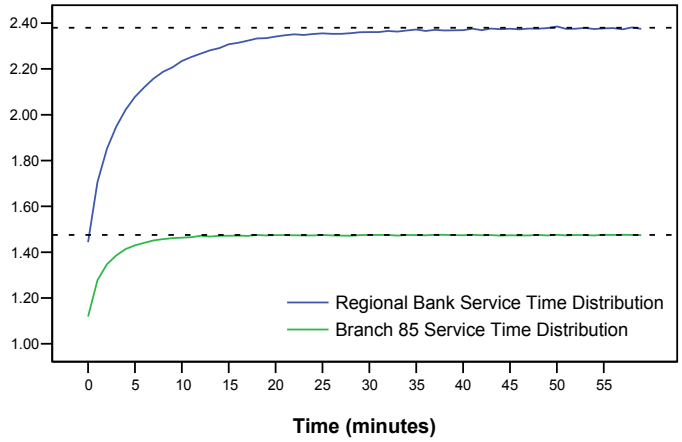

Figure 12. M/G/1 queue average customer waiting times (minutes) for the aggregated Regional Bank service time distribution for deposits, and the Branch 85 deposit service time distribution.

To illustrate what can happen in a multi-server system, three M/G/3 queuing systems are simulated with, as before, a fitted exponential, an empirical, and a lognormal model. The arrival rate for all three models is 140 customers per hour over an eight hour simulation window. Average waiting times are given in Figure 10. Behavior similar to the M/G/1 systems is observed, but with a more significant difference between the models. The exponential model has an average waiting time of more than six minutes by the end of the simulation window, compared to waiting times of approximately four minutes for the empirical model and three minutes for the lognormal.

These examples show that the exponential model can significantly overestimate the number of tellers needed to meet customer service goals, and that the lognormal model may somewhat underestimate the number of tellers needed. Both misestimation types can affect a bank's bottomline – the former unnecessarily increases a bank's labor costs and the latter may lead to higher customer attrition through poor customer service. Moreover, the relatively large misestimates given by the exponential model need to be weighed carefully against any analytic convenience afforded. As computational power continues to improve, the trade-offs between applying empirical distributions and more analytically tractable (e.g., exponential) models increasingly favors empirical distributions.

Staffing Based on Aggregate Data. As seen in section 3, there can be significant differences between service time distributions of branches and tellers. What is the impact

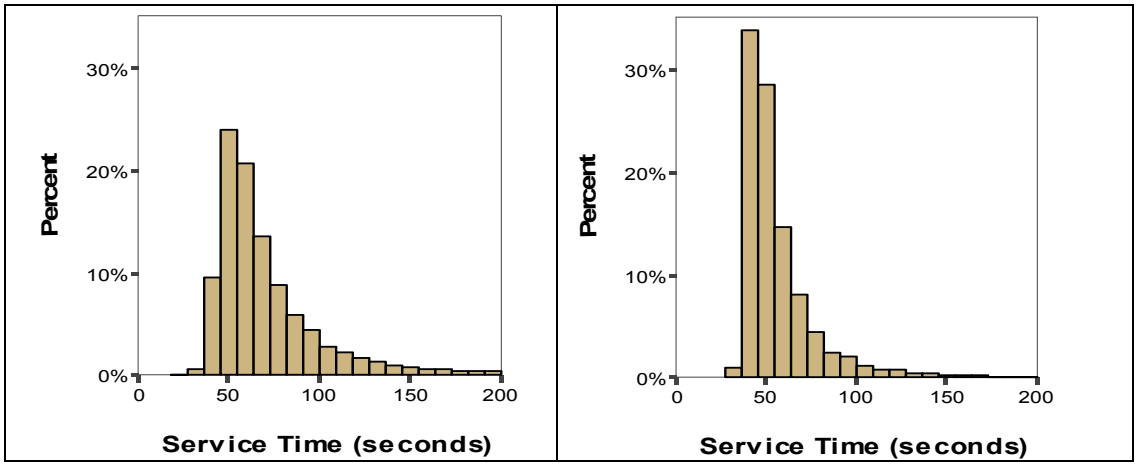

Figure 11. Service time distributions for deposits – left panel aggregated across all of Regional Bank, right panel gives deposit transactions in Branch 85.

of assuming equal service time distributions either across Regional Bank or within a branch? First consider the effect of treating all branches the same. Two M/G/1 queuing systems are simulated each with customer arrival rate of 30 per hour. One system assumes the empirical service time distribution for deposits aggregated across all of Regional Bank. The second system uses the empirical distribution for deposits from just Branch 85 (distributions shown in Figure 11). Recall that Branch 85 has a lower mean service time for deposits and lower standard deviation than Regional Bank aggregated across all branches. Even for a single server and relatively low customer traffic, there is nearly a one minute difference in average customer waiting time between Branch 85 and Regional Bank (Figure 12).

The effect is more pronounced for three servers, as is seen in simulations of two M/G/3 queuing systems over an 8 hour window with a customer arrival rate of 160 per hours. One system has the Regional Bank deposits service time distribution for each teller, and the other uses the Branch 85 deposits service time distribution. A dramatic difference in waiting times is observed (Figure 13). Average waiting times for the Regional Bank model are nearly ten times those for Branch 85. Even after just one hour (the system starts empty), the average waiting time for the Regional bank model is more than 2.5 minutes, five times the average wait of less than 30 seconds for Branch 85. Note that the arrival rate of 160 customers per hour is near the maximum effective service rate of the Regional Bank model, while safely within a

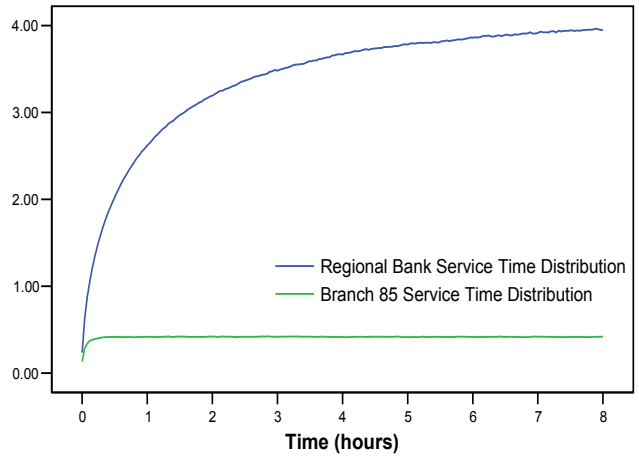

Figure 13. M/G/3 queue average customer waiting times (minutes) for the aggregated Regional Bank service time distribution for deposits, and the Branch 85 deposit service time distribution.

manageable range for the three tellers in the Branch 85 model. This is the case in many of the examples given here. It is precisely in these situations – when the system is at high utilization – that incorrect assumptions on the service time distribution have a significant impact on performance measures and staffing decisions. Such arrival rates are not unusual in branch banks such as Regional Bank – all the examples given here have been observed in data from our work with bank transaction volume forecasting and teller staffing. While an arrival rate of 160 customers per hour may not be sustained throughout an eight hour work day, a changing arrival rate that is consistently near the maximum service rate of a varying number of servers is the goal of setting efficient staffing levels. Thus, having an accurate model of service performance is essential. In particular, branches should be treated as a distinct service center and modeled with their specific branch data, and not as identical units modeled with aggregated data.

Assuming statistically identical servers. Perhaps the most common assumption in multiple server queuing models is that all of the servers in the system are statistically identical with respect to service times. At a bank branch level this assumption is reasonable at face value. Teller training within a branch should be fairly consistent, resulting (in theory) in similar service performance. However, there can be considerable variation among tellers within a branch. The impact on service time is illustrated by comparing tellers at the 10th and 90th percentiles. Teller 357 has a mean service time for deposits of 51.3 seconds (10th percentile of Regional Bank) and Teller 5124 has a mean service time for deposits of 93.9 (90th percentile). Two M/G/1 systems are simulated – one with Teller 357's empirical service time distribution, and the other with Teller 5124's distribution. Both systems have arrival rates of 30 customer per hours over a one hour simulation window. The large difference in average waiting times is shown in Figure 14.

Again, the operational question is of what impact the identical server assumption might have on staffing recommendations. If a standard multiple teller queuing model is created under the assumption that all of the tellers are statistically identical, some mixing of the individual tellers' service time distributions has to take place. The most straight-forward way to accomplish this is to use aggregated service time data from the multiple tellers to derive an assumed common service time distribution. Indeed, if data is available at a branch level,

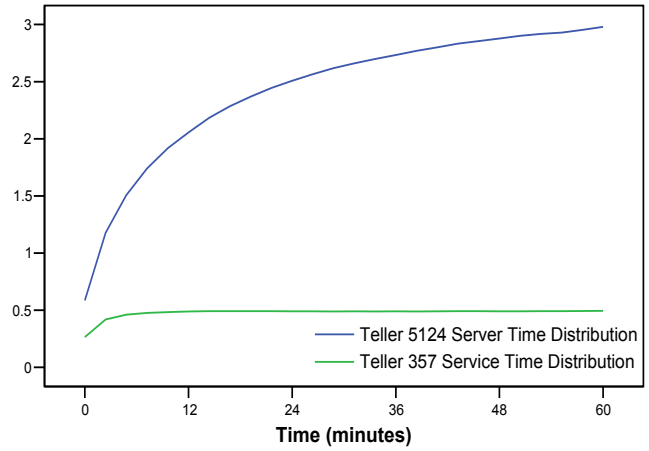

Figure 14. Average customer waiting times (minutes) for M/G/1 systems - one with the 90th percentile teller service time distribution, other with the 10th percentile teller service time distribution for deposits.

but does not distinguish between individual tellers, aggregated branch data may be the only way to construct a service time distribution model. This can markedly impact system performance measures, particularly when one teller is underrepresented in the data because of the relative number of transactions in the aggregated data.

Two queuing systems are simulated to illustrate this point. First, an M/G/2 model is constructed using aggregated data from tellers 3983 and 4827 of Branch 85 to construct a shared empirical service time distribution for the two tellers. Next a 2-teller model is constructed where one teller has a service time distribution derived from the data of Teller 3983 only, and the other teller has service time distribution derived from only the data of

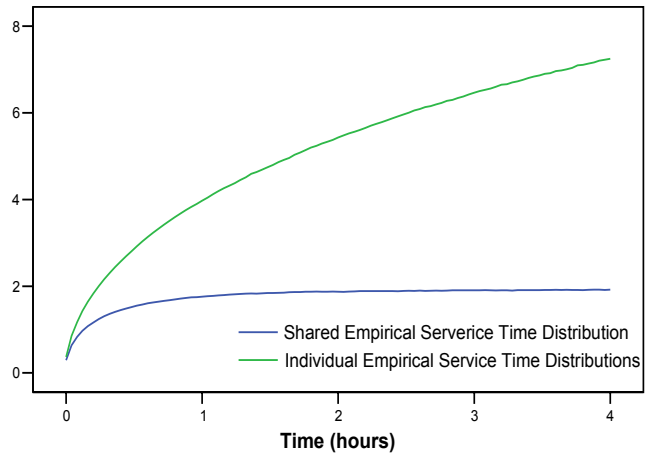

Figure 15. Average waiting times (minutes) for the shared empirical service time distribution and the individual service time distributions.

Teller 4827. The models are simulated over a four hour time window with a customer arrival rate of 120 customers per hour. Teller 4827's mean service time of 69.2 seconds is almost 18 seconds longer than the mean service time for Teller 3983 (Table 3). Teller 4827 also accounts for less than 10% of the recorded transactions for the two servers. This leads to the dramatic differences in average waiting time seen in Figure 15. The mixed two teller system maintains waiting times of less than 2 minutes, while the more realistic 2-server model shows waiting times growing unchecked, at more than 7 minutes by the end of the 4 hour window. Thus, the identical server assumption can yield much different system performance results than a more realistic individualized server model and thus lead to inappropriate staffing levels. When sufficient data is not available for individual tellers such errors may be unavoidable – for instance, when a new teller has just been hired. Nonetheless, it is important to realize the potential for sizeable misestimations in service performance before setting staffing levels.

SUMMARY

Standard distributional assumptions on service times have largely gone untested in the banking industry. Likewise the use of benchmark service goals derived from aggregated data have been assumed sufficient in deriving staffing levels. Now that service processes in banks have been computerized and integrated, there is little reason not to challenge these assumptions. In fact, as the results in this article indicate, it is important to do just that in order to determine efficient teller staffing levels. We have shown that two commonly made assumptions – exponential service times and statistically identical servers – are not supported by the lobby teller service time data for Regional bank. Moreover, applying these assumptions in queuing-based staffing models can lead to different estimated performance levels than when empirical, individualized models are used. These differences are often significant enough to impact teller staffing decisions, which ultimately impact a bank's financial and consumer relations goals.

ACKNOWLEDGEMENTS

We thank our colleagues at ARGO Data Resources. We especially thank Max Martin, ARGO CEO, for providing the opportunity to work with a talented team of people developing innovative tools for the banking industry.

REFERENCES

Atlason, Júlíus, Marina A. Epelman and Shane G. Henderson. 2004. Call Center Staffing with Simulation and Cutting Plane Methods, *Annals of Operations Research*, 127: 333-358.

Bolotin, V. 1994. Telephone circuit holding time distributions, 14th International Tele-traffic Conference (ITC-14), Elsevier.

Breukelen, G. 1995. Theoretical note: Parallel information processing models compatible with lognormally distributed response times, *Journal of Mathematical Psychology*, 39: 396-399.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing science perspective, *Journal of the American Statistical Association*, 100(469): 36-50.

Nozaki, Shirley A. and Sheldon M. Ross. 1978. Approximations in Finite-Capacity Multi-Server Queues with Poisson Arrivals, *Journal of Applied Probability*, 15(4): 826-834.

Mandelbaum, A., and R. Schwartz. 2002. Simulation Experiments with M/G/100 Queues in the Halfin-Whitt (Q.E.D) Regime, Technical Report, Technion. http://iew3.technion.ac.il/serveng/References/references.html

Rice, John. 1995. *Mathematical Statistics and Data Analysis,* 2nd ed. Duxbury Press, Belmont, CA.

Ulrich, R., and J. Miller. 1993. Information processing models generating lognormally distributed reaction times, *Journal of Mathematical Psychology*, 37: 513–525.

Correspondence: monticino@unt.edu