sampling research using a simulated zulily customer frame

John P. D. Martin

Friday, December 5, 2014

Executive Summary

The paper investigates the differences in sampling estimates and A/B testing sample sizes expected through use of a simulated customer purchase history frame. It is found that monthly stochastic purchasing behaviour and different yearly purchasing frequencies by customers can sometimes impact on the robustness of the sampling method and statistical inferences from A/B testing.

Finding 1: For a model customer frame with stochastic monthly purchasing behaviour. The sampling variance may be significantly greater than assumed from simply downscaling the yearly data.

As such, lower than expected monthly population variance would be computed from downscaling the yearly data and result in insufficient sample sizes being proposed for A/B testing compared to the expected 90% statistical power performance.

Recommendation 1: Use of simulated repeated sampling from historical monthly data should be used as a nonparametric method, to establish the monthly population variance rather than analysing scaled yearly data.

Finding 2: For a model customer frame with a wide range of yearly purchase frequencies, some reduction in the sampling variance can be achieved with stratified sampling compared to simple random sampling SRS of the whole frame.

Explicit use of the frame information via stratified sampling helps ensure more of the "one off" test samples contain a representative sample of the wide purchase frequency behaviour.

For the model investigated, it was found that stratified sampling (proportional to strata size with 7 strata) resulted in ~15% lower sampling variance. The actual results for a real customer frame would be different from this model result depending on the level of stochastic monthly purchasing behaviour, purchasing growth and seasonality but simulation using recent historical data would be internally consistent and insightful for any A/B testing experimental design. Based on historical data, stratified experimental design (sample per strata) can also be optimised to minimise sampling variance.

Recommendation 2: The combined use of stratified sampling experimental designs and "conservative" SRS A/B testing design effects, would make the statistical inferences identified from testing more reliable.

Introduction

Sampling from populations with important subgroups because of

- output requirements or
- significantly different variance behaviour

may require stratified experimental designs to achieve optimal confidence intervals for sample estimates and A/B testing. For zulily, important population subgroups are

(i) women customers by age groups,
(ii) women customers grouped by family demographics, (eg. grown up, teenagers, young, zero children),

(iii) old and new customers and
(iv) low, medium, high frequency purchasers.

In this paper, the impact of new/old customers and different purchase frequency customers on sampling estimate confidence intervals are investigated by use of customer purchase model based on the Poisson distribution and some aggregate zulily data.

Firstly, the Model Population frame is created at the yearly level based on published purchase frequency data. Then a monthly customer level model is derived by introducing monthly stochastic Poisson behaviour for each (yearly) purchase frequency.

Next the A/B testing sample size is determined (based on the variance of the monthly stochastic purchase model) for a proposed minimum detectable effect size of 0.1246, corresponding to a increase/decrease of 1 additional purchase transaction per customer per year.

Finally, the variance properties of repeated sampling from the frame is assessed comparatively for simple random sampling (SRS), stratified SRS and stratified systematic random sampling.

Any conclusions drawn on differences between the sampling methods strictly applies to the Model frame. However, the empirical mean and variance properties for the zulily frame could be used along with repeated sampling from the zulily frame to see if similar conclusions are obtained.

Model Population frame

A simulated yearly customer purchase history is created initially on zulily press releases

- A current customer base of 4.1 Million based on Second Quarter 2014 Results (1)
- Average of 6 purchase transactions per year for new customers (2)
- Average of 12 purchase transactions per year for old customers (2)
- The customer base grew by 86% in the last year based on Second Quarter 2014 Results (1)

From

(1) http://investor.zulily.com/releasedetail.cfm?ReleaseID=864744

(2) http://www.bizjournals.com/seattle/blog/techflash/2014/10/zulilys-big-hard-challenge-a-different-website-for.html

Assumptions in the customer base model are

- The yearly customer purchase history is modelled as the sum of two independent Poisson distributions for the new and old customer populations
- The new customer population comprises the customer base growth (86%) in the last year (4.1M(0.86/1.86)=) 1.9M
- The old customer population comprises the customer base last year, (4.1M/1.86=) 2.2M
- The forecast monthly purchase transaction behaviour is assumed to be a stochastic Poisson modification of the average monthly behaviour for each yearly purchase frequency. For example, each customer may have varying numbers 0, 1, 2 .. etc purchase transactions each month based on their yearly purchase average.

To determine the required sample size for A/B testing the following is assumed

- A detectable change in the average number of purchase transactions (per customer) by one additional (or less) transaction per year across the customer base is specified as the designated threshold ("effect size") for hypothesis testing of significant differences between two samples (based on control and treatment versions of zulily personalised customer webpages).
- For old customers, this means a change from an average of 12 purchase transactions per year to 13 (or 11).
- For new customers, this means a change from an average 6 purchase transactions per year to 7 (or 5).
- For the whole customer base model history (described below) this means an increase from an average of 9.125 purchase transactions per year to 10.125 (or decrease to 8.125) purchase transactions per year is regarded as a threshold change in the customer base purchase behaviour.
- The nominal effect size of this designated threshold (using the customer base model) is 0.1296.

- A two sided significance test is implemented as it is not confirmed whether the treatment personalised webpage will increase purchases
- No sample loss, nonresponse or growth in the customer base is assumed (whereas in reality some attrition and significant growth is occurring)
- The significance level for type I errors is set at the standard 95% level (corresponding to a 5% chance of false positives occurring, whereby confidence interval estimates of the purchase behaviour may not include the true population value)
- The significance level for type II errors is set at 90% level similar to clincal trials (corresponding to a 10% chance of false negative conclusions, whereby the estimate of difference between two samples (control and treatment) is not statistically significant when the true population values are actually different by more than the designated threshold value between the two groups)
- the control (monthly) purchase history rate is the average behaviour of ~ 0.768 (=9.125/12) purchase transactions per customer per month but based on a sample estimate. In real life, this A/B testing design element helps adjust for changes in the purchase behaviour arising from seasonality in purchase volume and the current monthly growth in zulily customer purchases.

Table 1 and Figure 1 below, show the modelled new, old and combined customer purchase behaviour based on aggregate data for purchase transactions in 2014. The distributions are assumed to be Poisson distributions using the published zulily data for new (mean= 6 purchases per year) and old customers (mean= 12 purchases per year). The variances of these two components of the zulily customer base are modelled as 6 & 12 respectively consistent with the Poisson distribution assumption.

The combined distribution is then a sum of two independent Poisson distributions. With the relative population sizes of the two components the estimated mean is ~ 9.216 purchases transactions per year and the estimated variance is 18.1235 (close to the theoretical value for the yearly variance of 18, calculated as the sum (6+12) of two independent components).

Table 1: Modelled customer counts by purchase transactions per year

	purch_trans_per_year	old_customers	new_customers	all_customers
0	0	13.52	4710	4723
1	1	162.21	28258	28420
2	2	973.24	84773	85747
3	3	3892.97	169547	173440
4	4	11678.92	254320	265999
5	5	28029.41	305184	333213
6	6	56058.81	305184	361243
7	7	96100.82	261586	357687
8	8	144151.23	196190	340341
9	9	192201.64	130793	322995
10	10	230641.96	78476	309118
11	11	251609.41	42805	294414

```
## 12                  12      251609.41         21403        273012
```

Fig. 1: 2014 bimodal purchase transaction zulily consumer mod

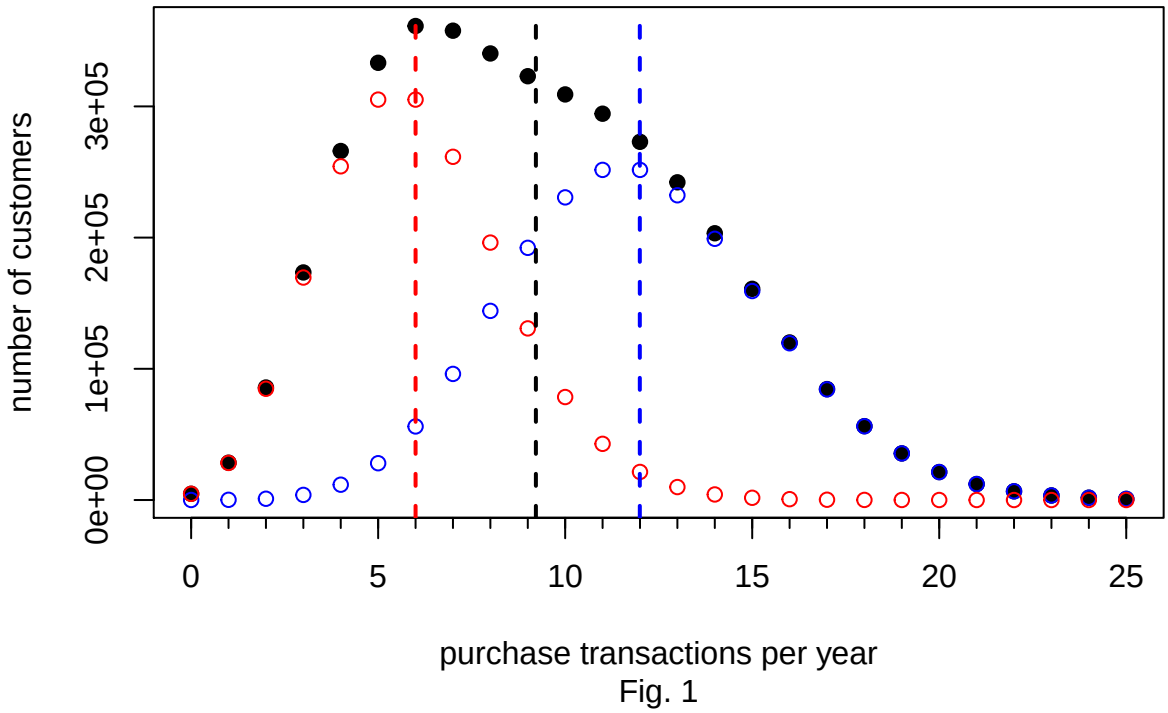

Fig. 1

customer model.pdf

In practice, the A/B testing will be conducted on a monthly basis. The "yearly" model (shown above) would create a "monthly"" model which is too smooth in behaviour, if the purchase transaction numbers are simply divided by 12. This is because most customers will not be uniform across each month in their buying habits. (This behaviour is expected even ignoring seasonality). Customers will make 0, 1, 2 .. etc purchase transactions per month and this volatility will contribute to the variance of the sampling results and hence the required sampling size for statistical power of 90%.

To model the monthly behaviour more realistically, a unit level model (4.1M records) is derived where each customer has a monthly stochastic number of purchase transactions 0, 1, 2, ... etc. The mean of the Poisson distribution used in the stochastic monthly behaviour for each customer is based on the yearly number of purchase transactions (of the customer) divided by 12. The R chunk "monthly customer model" in the associated R markdown source file of this report contains explicit details.

Table 2 and Figure 2 below, gives one realisation of the random Poisson based (individual customer) monthly purchase. The estimated mean purchase transactions per customer per month is ~ 0.768 (Figure 2) which is equivalent to the yearly average (9.215/12) and the numbers at aggregate level (Table 2) would be in close agreement with a downscaled yearly model. However, the variance in purchase transactions at monthly level is now significantly higher than if the yearly behaviour was simple scaled by 1/12.

These model estimates should be compared to the real monthly zulily distribution (for 2014) when assessing the usefulness of the model including analysis by repeated sampling. For example, seasonality in purchase transactions would vary the (mean) and sampling variance across months.

Table 2: Stochastic model purchases compared to average purchases by yearly purchase frequency groups

```
##   purch_trans_per_year scaled_yearly_model stochastic_monthly_model
```

```
## 1          0-4          148685.7          148294
## 2          5-6          319460.2          319841
## 3          7-8          435544.8          436130
## 4         9-10          499844.6          499834
## 5        11-12          542891.5          542809
## 6        13-15          700695.8          701455
## 7        16-25          501368.3          501105
```

Fig. 2: Stochastic monthly purchase transaction model (2014)

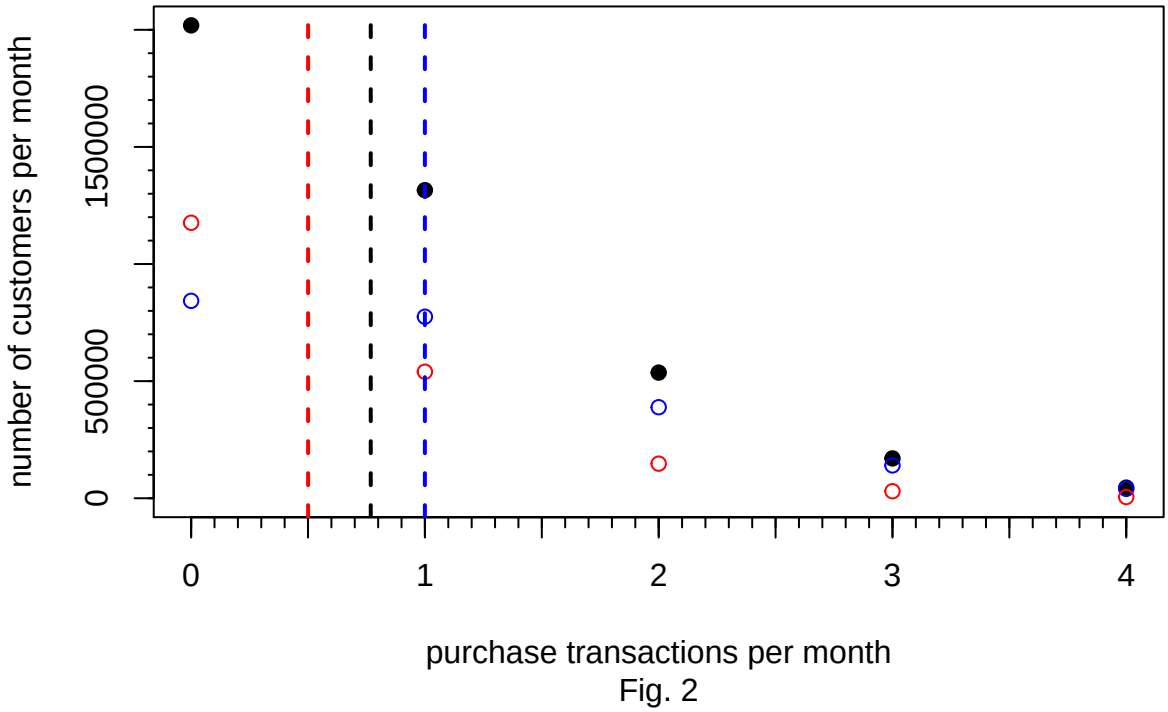

Fig. 2

customer model.pdf

Under the stochastic model, the monthly average purchase transactions is 0.768 (black line) for the whole population. While the new and old customers exhibit average purchase transactions (red and blue lines) of 0.5 & 1. These mean estimates are all consistent with a simple average of the yearly behaviour (9.215/12, 6/12 & 12/12). What is different with the stochastic model is that there is additional variance at customer level in the number of purchase transactions per month which increases the sampling variance occurring with A/B testing.

A/B testing

Under consideration with A/B testing, is the testing of modified zulily personalised customer webpages to better match the range of items that each customer may be interested in purchasing for herself and her friends/family.

As part of the experimental design for A/B testing, a suitable threshold value for change in purchase behaviour needs to firstly determined.

Given this agreed value for a noticeable change in sales performance, the required sample size is then determined using standard statistical sampling formulae based on

(i) expected mean purchase behaviour,
(ii) expected variance purchase behaviour and
(iii) requiring the sample size to achieve statistical power >= 90% for changes in the mean at least as large as the threshold value (as risk mitigation against false negative conclusions on the efficacy of a particular test modification compared to existing personalised webpage method)

Establishing a minimum detectable purchase change level for A/B testing

Since the intention of the webpage modification under test is to induce a fresh customer purchase in the month of the test. A suitable threshold value in observing a change of purchase behaviour is proposed as an increase (decrease) of one additional purchase per year.

In monthly terms, this threshold is

a minimum change in purchase behaviour of 0.0833 (=1/12)

Using the modelled zulily data

The monthly mean purchase transaction rate is 0.768 transactions per customer (=9.125/12)

The scaled yearly variance in the purchase transaction rate is 0.3548 (=sqrt(18.1235)/12). This value is expected to underestimate the real monthly variance for A/B testing use, since customers will vary the number of purchase transactions per month rather than have constant behaviour.

A more realistic monthly variance in the purchase transaction rate would be 0.8948 based on the monthly model derived in the previous section (containing stochastically generated numbers of monthly purchase transactions per customer). For each real frame, repeated sampling of historical data can be used to establish the most accurate sampling variance estimate.

Using these three values

- (mean~0.768, variance~0.8948, threshold change~1/12) for zulily customers and
- imposing a statistical power of 90% performance on the sample size (for two-sided significance test) to mitigate against false negative conclusions from A/B testing on the difference in results between the normal webpage (control sample) and test webpage (treatment sample),

gives the following required effect size to be observed for two sample (A/B) testing

(based on Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Hillsdale, NJ: Lawrence Earlbaum Associates.)

effect size = (mean(treatment) - mean(control))/ sqrt(population variance) * sqrt(2)

```
## [1] "effect size calculation"
```

```
## [1] 0.1246
```

```
##
##      Two-sample t test power calculation
##
##               n = 1355
##               d = 0.1246
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
## [1] "A double check on the 'pwr' R package estimate of statistical power:"

## [1] "the alternate estimate of statistical power using n = 1355 "

## [1] "should also be close to 90% power, expressed as decimal (0.90)"

## [1] 0.8993
```

Other small levels of purchase changes and their required sample sizes for A/B testing

To provide some perspective on the relative impact of 1 additional purchase per year on sales and the required sample size (enforcing 90% statistical power), Table 3 below, lists the required sample size (per group for two-sided significance test) and effect size for changes in sales performance by the nominal amounts of 2%, 5%, 1/12, 10% & 20%.

Table 3: Required sample sizes for A/B testing (sig level 5%, power 90%) under different thresholds

	threshold	effect size	A/B sample size
2% more/less purchases per month	0.02000	0.0299	23506.9
5% more/less purchases per month	0.05000	0.0748	3761.9
1 more/less purchase per cust per yr	0.08333	0.1246	1355.0
10% more/less purchases per month	0.10000	0.1495	941.2
20% more/less purchases per month	0.20000	0.2990	236.0

Under typical clinical trial conditions, due to high sample cost and the ethical impact of imposing placebo treatments on patients in control groups, the control & treatment groups are the same size.

With A/B testing of webpages, the sample cost may be considered to be lower and the ethics of imposing placebo (or status quo) treatment to the control group are not relevant, hence the control group can in principle be alternately set to equal the sum of the size of the treatment groups. This altered experimental design while potentially lowering the accuracy possible for each treatment group (if the available sample is limited) allows (i) interactions between simultaneous treatments to be separately measured and (ii) multiple thresholds of change in performance to be analysed.

For example, using the current zulily example. With four samples (and three subsamples) comprising

1. Control group, sample size ~4065 (broken into 3 subgroups of 1355)
2. treatment 1, sample size 1355
3. treatment 2, sample size 1355
4. treatment 1 + treatment 2, sample size 1355

The 3 control subgroups provide some empirical cross validation of the performance of the sample size of 1355. Each treatment group can be assessed against the control group for increase of 1 additional purchase per year. The interaction impact between treatment 1 and treatment 2 is independently measured. Finally, the combined treatment sample size 4065, can be compared to the combined control group to assess if a lower rate of improvement (~5% more purchases per year, see Table 2) may be identified even if the A/B test threshold change $> 1/12$ did not occur.

In terms of monthly transactions, the total average monthly transaction expected is 3 Million per month, so the above experiment comprises ~0.27% (=8130/3M) of the customer transactions.

Assessing the impact of any internal structure within the sample frame, on the randomisation method

Three different sampling methods

- simple random sampling (SRS),
- stratified SRS and
- stratified systematic random sampling

will be assessed comparatively using repeated sampling to see if the expected confidence intervals used in the above A/B testing sample size performance can be reproduced.

As previously described, the Model population frame consists of 4.1 Million customers. Each customer will have a unique identifier and the three sampling methods will be applied to this frame to obtain repeated samples. The variance across the repeated samples is then calculated to see if the sampling method produces different variances because of the internal structure of the population frame.

In Table 4, four different estimates of mean and sampling variance are presented.

The first (reference) value was obtained by calculating the **population mean and population variance** for the monthly customer model (including monthly stochastic number of purchases). The **sampling variance** for sample size of 1355 was then obtained from the population variance and sample size using standard formula.

The second estimate was obtained by **repeated simple random sampling** across the entire unit level monthly customer model, treating the frame as one strata. For each sample of 1355, the mean was calculated. To determine the sampling variance, the variance of the distribution of the means was calculated. In addition, to ensure enough repeated samples were taken the accumulated estimates of sampling variance was monitored for convergence. The data on the distribution and convergence of the estimated sampling variance is shown in the Appendix.

The third estimate was obtained by **stratified repeated simple random sampling** where the monthly customer model frame was split into 7 roughly equal (by population) strata. The sample per strata was allocated proportional to population. The weighted mean was calculated for each repeated sample. To determine the sampling variance, the variance of the distribution of the means was calculated. Again the accumulated estimates of sampling variance was monitored for convergence. The data on the distribution and convergence of the estimated sampling variance is also shown in the Appendix. There are more sophisticated versions of stratified SRS where the sample size per strata is optimised to mimimise the overall sample variance (given inhomogeniety in variance behaviour across strata).

Finally, the fourth estimate was obtained by **stratified random systematic sampling** where again the monthly customer model frame was split into 7 roughly equal (by population) strata. The sample per strata was also allocated proportional to population. The difference with random systematic sampling was that the sample was created by choosing the first sample unit in each strata randomly then the rest of the sample per strata was determined as a regular skip within the strata. It is expected to be similar in performance to stratified SRS but is of some benefit in ordered frames if there is a requirement for subgroups to be adequately sampled. For this paper, the fourth method acts as cross validation on the third method (as the first and second methods are also performing cross validation with respect to each other). As for methods two and three, the weighted mean was calculated for each repeated sample and the sampling varaince from the distribution of the means. The data on the distribution and convergence of the estimated sampling variance is again shown in the Appendix.

Table 4: Comparison of estimated means and sampling variance for different sampling methods

Sampling Method	Estimated Mean Purchase Transaction per customer per month	Estimated sampl
population calculation	0.7683	.0006604

Sampling Method	Estimated Mean Purchase Transaction per customer per month	Estimated sampl
SRS	0.7685	.0006505
stratified SRS	0.7676	.0005586
random stratified systematic sampling	0.7679	.0005713

It can be seen from Table 4 and the Appendix, that stratified SRS is producing some reduction in sampling variance compared to SRS. This efficiency gain is understood to occur because stratified SRS restricts the range of possible samples to reflect more closely the proportion of the population for each stratified range of purchase frequencies. If the population behaviour was uniform in density then stratified SRS would not show any difference to SRS.

The difference in variance performance between repeated SRS and the population calculation results .0006505 & .0006604 (rows 2 & 1) arise from residual sampling error expected with repeated sampling. That is, .0006505 is an estimate and should have a confidence interval including .0006604 the true value, if it is unbiassed.

Conclusions

The stochastic nature of Monthly purchase transaction behaviour may contribute significantly to the sampling variance for A/B testing. Stratified SRS should provide some reduction in sampling variance compared to SRS due to customer peaks in the yearly purchase transation behaviour. Historical analysis including simulated repeated sampling of monthly purchase data should be used to establish the population variance (and the contribution of stochasticity). Some information on the impact of seasonality on sampling variance should be possible by comparison of historical analysis across months.

If A/B testing sample sizes were determined by SRS formula but the sample were selected by stratified SRS, then more reliable statistical inferences from any significant results identified should be obtained as a result of more robust samples (stratified SRS) and tougher A/B test critical values (SRS sample variance higher than stratfied SRS).

Appendix - results of repeated sampling

Presented below are graphs of the distribution of repeated sampling estimates of (i) means and (ii) sample variance for three different sampling methods

- SRS
- stratified SRS
- stratified random systematic sampling

With large sample sizes >>30, the expectation is that the distribution of repeated mean estimates will approximate a normal distribution. The three sampling methods (with sample size 1355 and 2000 repeated samples), shown in Figures 3a, 4a & 5a exhibit this behaviour.

With large sample sizes, it is also expected that the estimated sample variance from repeated samples will converge to a stable value. The three sampling methods, Figures 3b, 4b & 5b exhibit this behaviour.

repeated sampling SRS

Fig. 3: Repeated sample estimates using SRS

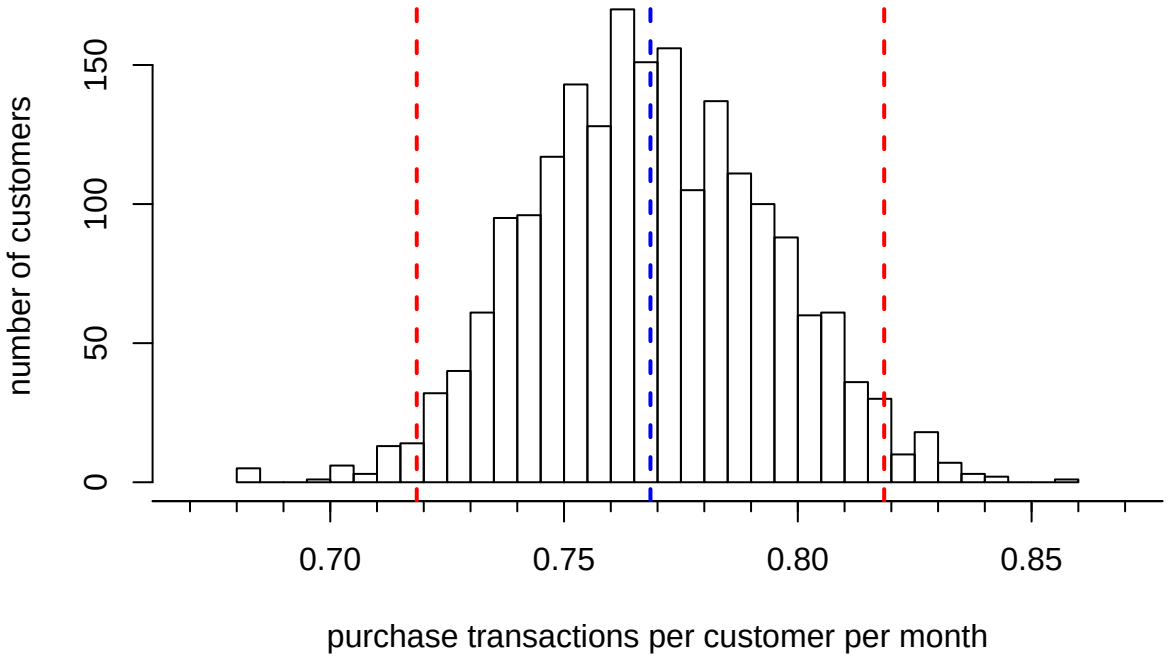

Fig. 3

sampling SRS.pdf

Fig. 3b: SRS repeated sampling convergence

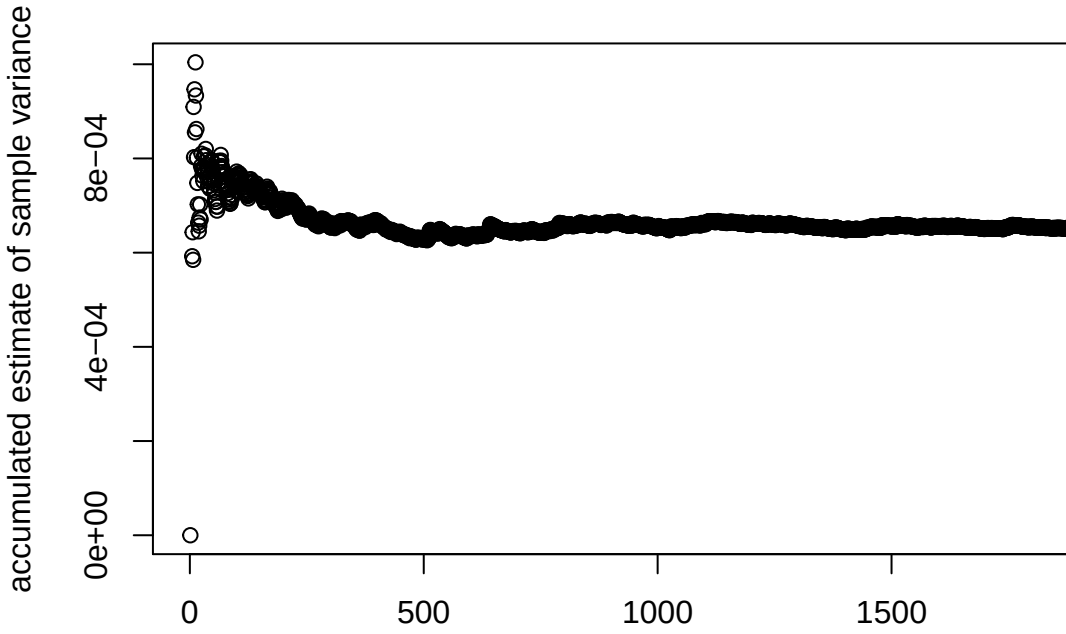

Fig. 3b

repeated sampling SRS stats.pdf

repeated stratified sampling SRS

Fig. 4: Repeated sample estimates using stratified SR

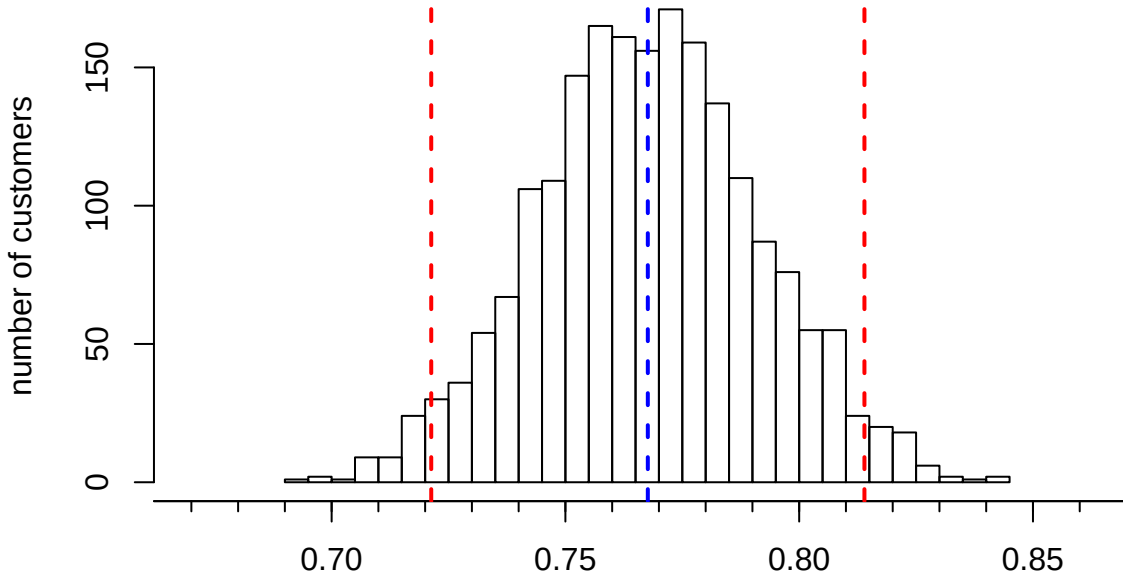

Fig. 4

stratified sampling SRS.pdf

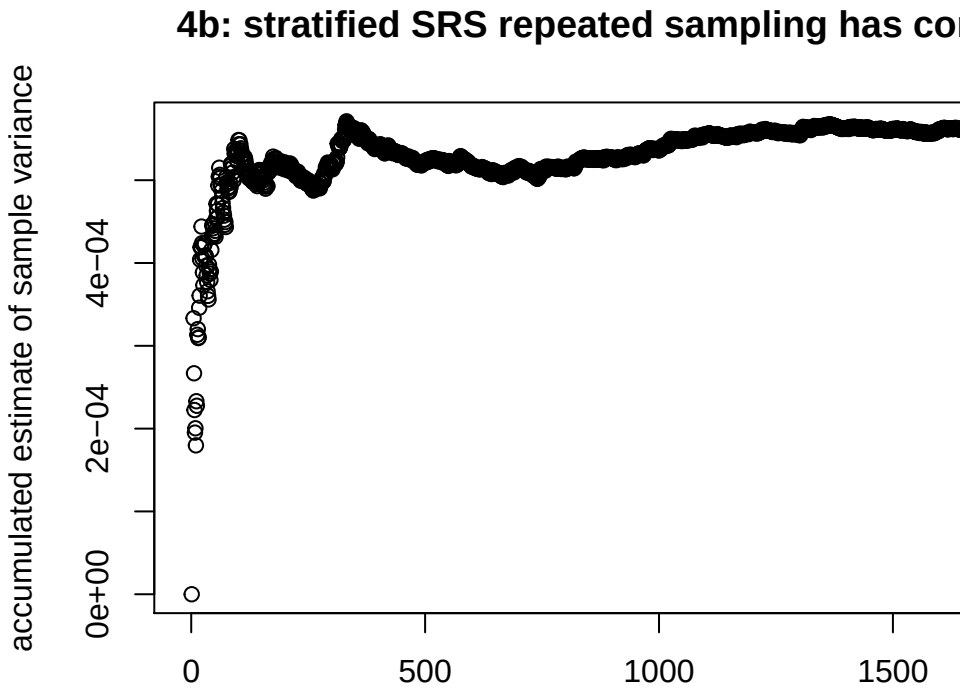

Fig. 4b

repeated sampling stratified SRS stats.pdf

repeated systematic sampling SRS

Fig. 5: Repeated sampling using stratified systematic sa

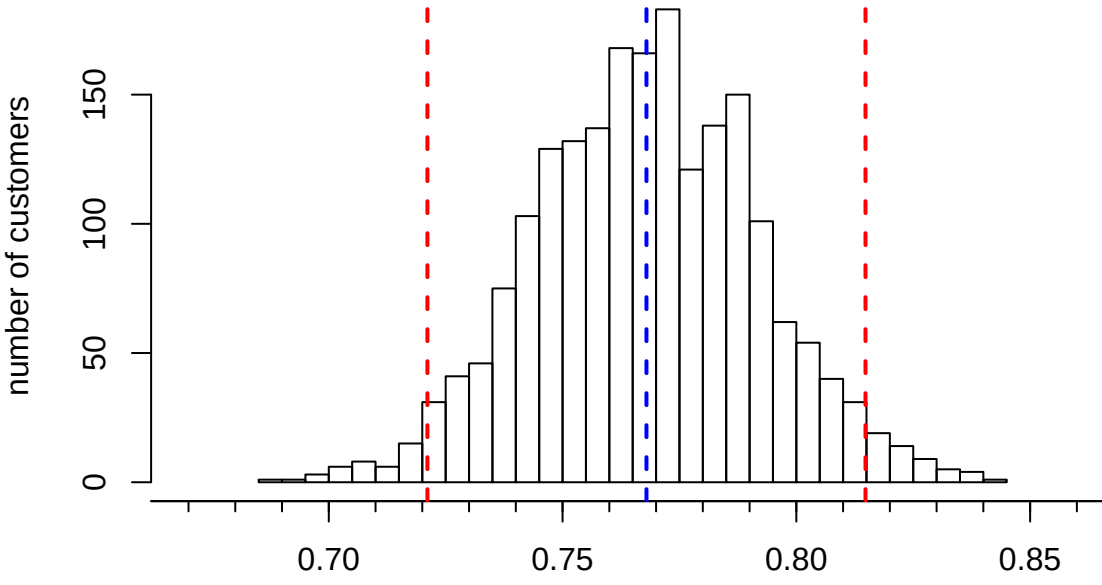

Fig. 5

systematic sampling SRS.pdf

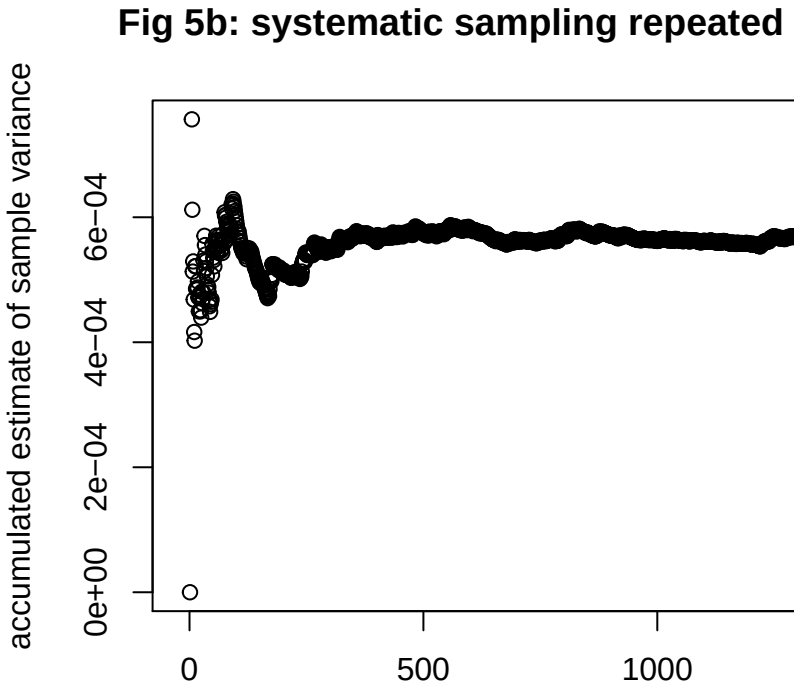

Fig. 5b

repeated sampling random stratified systematic stats.pdf