STA304

Neil Montgomery
2016-04-07

systematic random sampling

repeated systematic sampling

With a single systematic sample from a population, there is no good data-driven way to estimate $V(\bar{y}_{sy})$. So either the population had better be random, or it's possible systematic sampling may perform very poorly.

Another approach can be to select multiple systematic samples, n_s in number, resulting in n_s estimates of μ we can call $\bar{y}_i$.

We average these estimates to obtain:

$$\hat{\mu} = \frac{1}{n_s} \sum_{i=1}^{n_s} \bar{y}_i$$

repeated systematic sampling - variance estimator

Just use the sample variance of the $\{\bar{y}_1, \ldots \bar{y}_{n_s}\} \ldots$

$$\hat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{s^2_{\bar{y}}}{n_s}$$

with $s^2_{\bar{y}}$ is just the sample variances of the $\bar{y}_i$.

how to organize a repeated systematic sample

The idea is to keep the same desired *overall* sample size, but acquire the samples using multiple systematic samples.

For example, a population has size $N = 26000$ and we would like a sample size of $n = 260$. A single systematic sample would be a "1 in k" with $k = 100$ selected in the usual way.

Or, $n_s = 10$ systematic samples of size 26 could be selected, with starting positions chosen randomly between 1 and $kn_s = 1000$.

repeated systematic sampling example

A processed food company wants to evaluate the quality of its final packaging machinery…are they putting the same amount by weight in each package? 20 packages are produced each minute, 24 hours per day, 7 days per week. Consider a population of one week's worth of production with $N = 201600$.

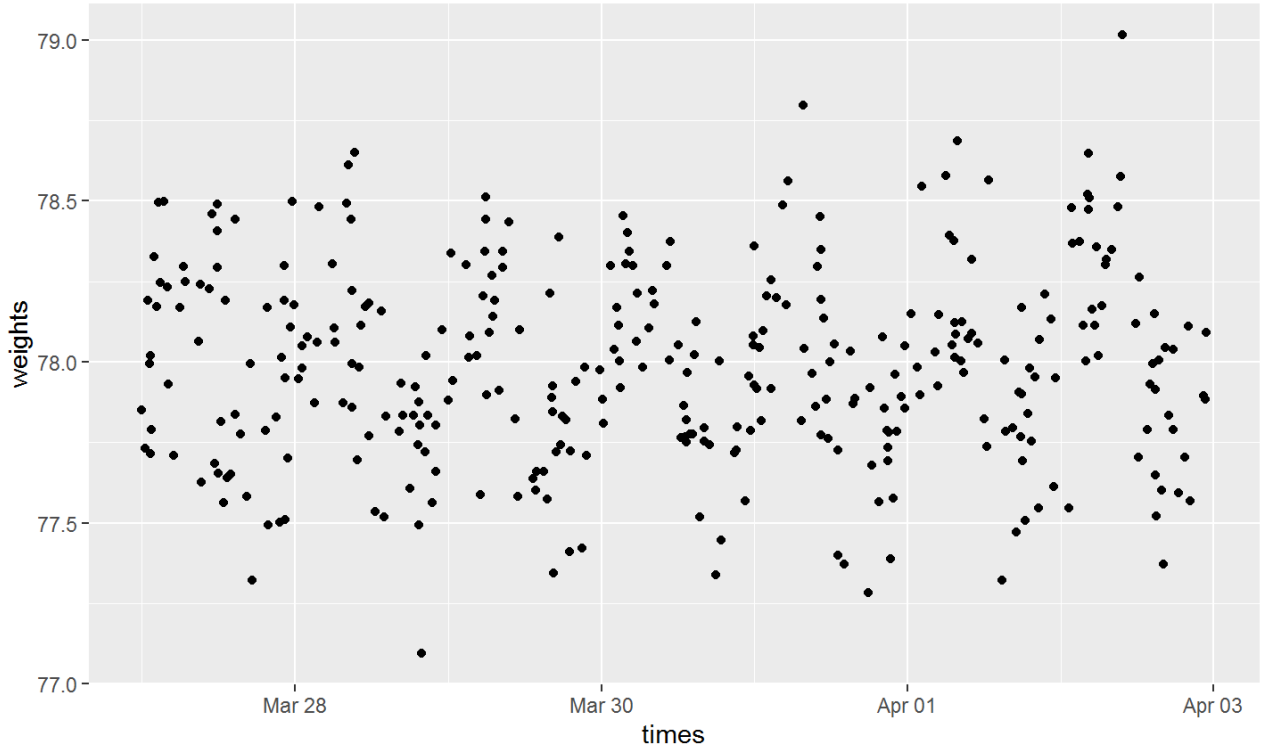
The desired sample size is $n = 336$. One could do a 1 in $k = 600$ systematic sample, but temperature and humidity are known factors that can affect both the machinery and the raw materials. So it's possible the population may not be random.

Another option would be to do $n_s = 21$ systematic samples of size 16 could be selected on a 1 in 12600 basis.

I have simulated a dataset that is available on github for this lecture.

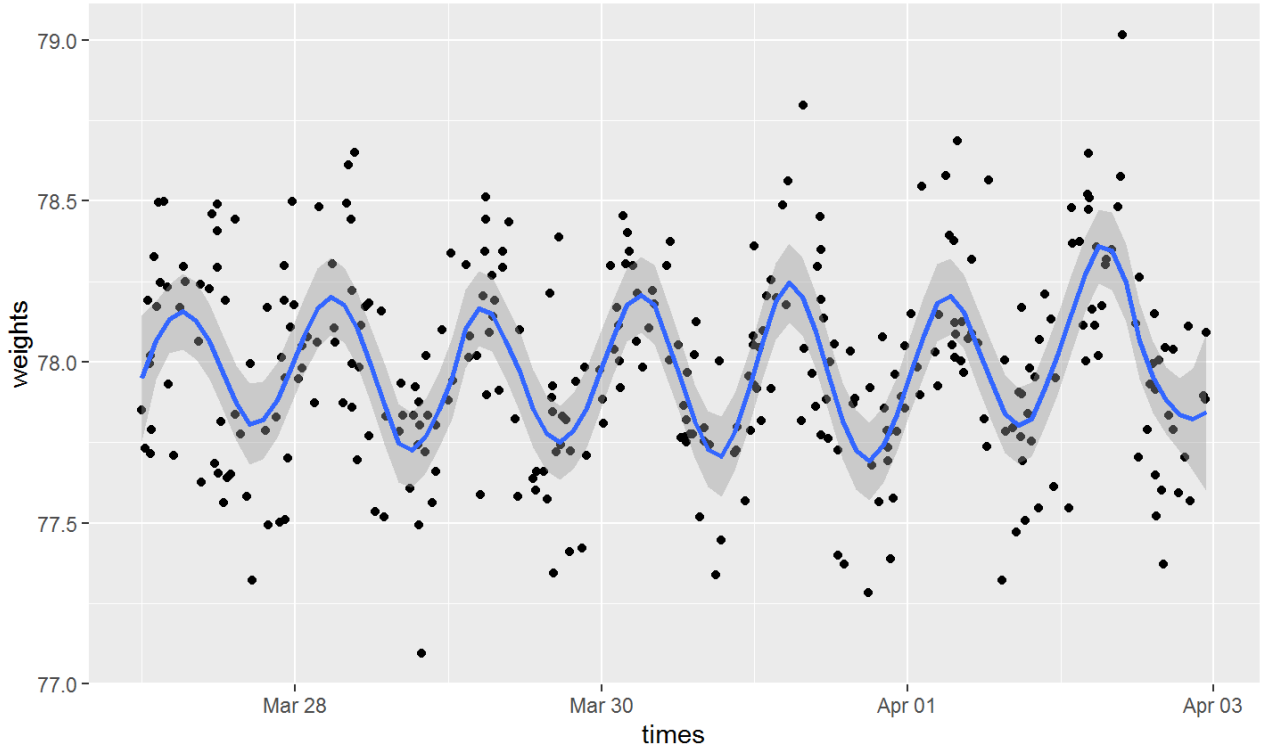
analysis of the 21 1-in-12600 samples - I

Here is a plot of all the samples together, ordered by time:



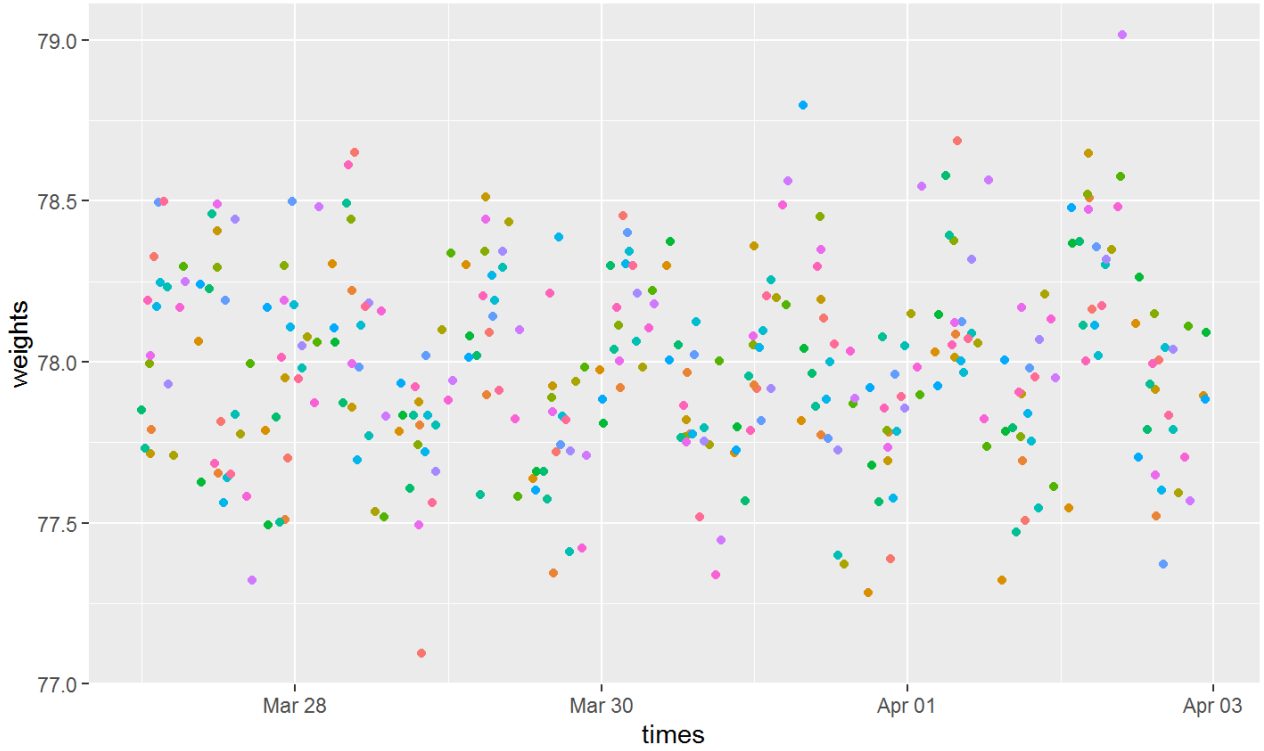
analysis of the 21 1-in-12600 samples - II

Here's the same plot with a smoother put over top:



analysis of the 21 1-in-12600 samples - III

Here's the same plot by sample:



analysis of the 21 1-in-12600 samples - IV

Estimating the mean and the variance is straightforward:

$$\hat{\mu} = 77.9845211 \qquad s^2_{\bar{y}} = 0.0052617$$

$$\hat{V}(\hat{\mu}) = \left(1 - \frac{336}{201600}\right) \frac{0.0052617}{21} = 0.0002501382$$

The usual bound on the error of estimation is $2\sqrt{\hat{V}} = 0.0316315$.

Better than SRS? What would you expect?

The theoretical "usual bound" for a SRS of size 336 turns out to be exactly 0.0333696

introduction to cluster sampling

a final sampling design

Sampling designs seen so far:

- SRS - the theoretical basis for all the others
- Stratified - good when population can be divided in advance
- Systematic - good when population has a special order or when no frame available

But sampling can still be very costly under any of these designs. The usual source of high costs is simple geography - travel time. Also, SRS and stratified still require a frame.

(single-stage) cluster sampling

In cluster sampling, the sampling unit is a collection of elements from the population. The population is divided into clusters. A simple random sample of clusters is selected. *All* elements of the cluster are measured.

Clusters are often determined geographically.

There is a basic trade-off in the composition of clusters.

There may be a large "intra-cluster correlation". So each additional element in a cluster might provide little marginal value. In this case large clusters could lead to poor population parameter estimates.

On the other hand, if clusters are too small, sampling costs may be too high.

We've already seen two examples of cluster sampling.

population mean estimate with cluster sampling

The setup is a bit involved:

N - number of clusters

n - sample size (number of clusters selected)

m_i - i^{th} cluster size (number of elements in cluster i)

$\overline{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$ - the *sample* average cluster size

$M = \sum_{i=1}^{N} m_i$ - the number of elements in the population

$\overline{M} = M \Big/ N$ - the average cluster size for the population

y_i - the total of the measurements in the i^{th} cluster.

In general, the old N and y_i now apply to *entire clusters*.

population mean estimate

The population mean is in this context equal to:

$$\mu = \frac{\sum_{i=1}^{N} y_i}{M}$$

The cluster sample estimator is:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

The denominator is random. This is actually identical to the ratio estimator $\hat{R} = r$ in the case where the "population" is:

$$\{(y_1, m_1), (y_2, m_2), \ldots, (y_N, m_N)\}$$

The estimator $\bar{y}$ is exactly r like from before.

estimated variance

Comes straight from ratio estimator theory:

$$\hat{V}(\overline{y}) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{\overline{M}^2}\right)\frac{s_r^2}{n}$$

where

$$s_r^2 = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}m_i\right)^2}{n-1}$$

and $\overline{M}$ can be estimated by $\overline{m}$ if required.

(For population total use $M\overline{y}$.)

cluster sampling example

We'll try question 8.2 from the text - to estimate the mean repair cost per month for a type of industrial equipment ("band saw").

Band saw dealer sells to $N = 96$ industries. A sample size of $n = 20$ is selected. Number of saws and total repair cost is recorded.

(Oddly enough, the total number of saws sold isn't available (?)).