Background information related to allocation

# 1 Background

three things to consider that could design allocation around 1) sample size - number of units selected - drives variance and probability of getting into a box 2) front matter - variance savings that one gets with sample size and sample rate (proportion of population sampled)  
3) expansion of trips - how many trips will each sampled trip expand to on average

To consider - sample size, not sample rate, is the primary driver of both the probability of getting data and the variance.

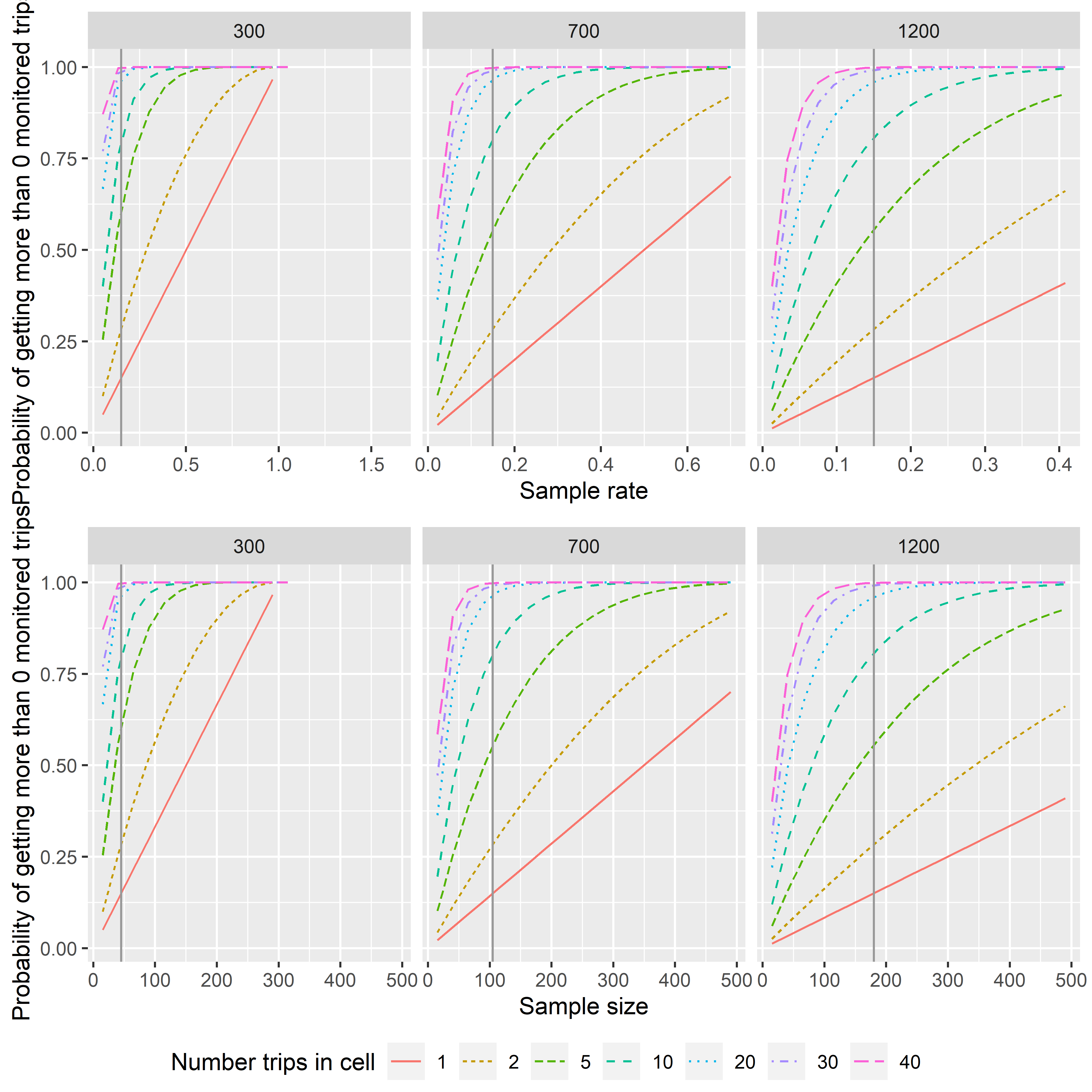
## 1.1 Probability of selecting a cell

Setting some basic population parameters: *N* is the population size, *K* is the number of successes in population (the number of fishing trips in the area of interest), *k\_sample* is the number of successes in the sample, and *n\_sample* the sample size.

The probability of getting no monitored trips in the area of interest varies with *K* and sample size. We can compute this using base R and the function dhyper(x, m, n, k) where (from the R documentation) x = number of successes drawn, n= number failures in population, =number sucesses in population (m+n=N=population size), and k=number of balls drawn from urn (see Eq. (2.4) below).

In our application this would be P(*k\_sample* = x) = dhyper(x, *K*, *N* - *K*, *n\_sample*) = *k\_sample*. Note that to be consistent with other texts, this is often noted as phyper(k\_lc,N-K,K,n\_sample), see <https://www.vrcbuzz.com/hypergeometric-distribution-probabilities-using-r/> As an example, if the number of success is m <- 5, the number of failures is n <- 11, the sample size is k <- 4, then the probability P(X=0) should be dhyper(0,m,n,k) = 0.1813187.

If we define a population of 1500 trips where the number of trips in a cell is 1, 2, 5, 10, 20, 30, 40 and sample sizes range from 15 to 490, we can see how the probability of a sample containing no trips from a given area changes with sample size (or rate) and the size of the cell (Figure 1.1)

 Similarly, the probability of a cell containing no data increases with both decreasing sample rate and the number of trips in a cell (Figure 1.2).

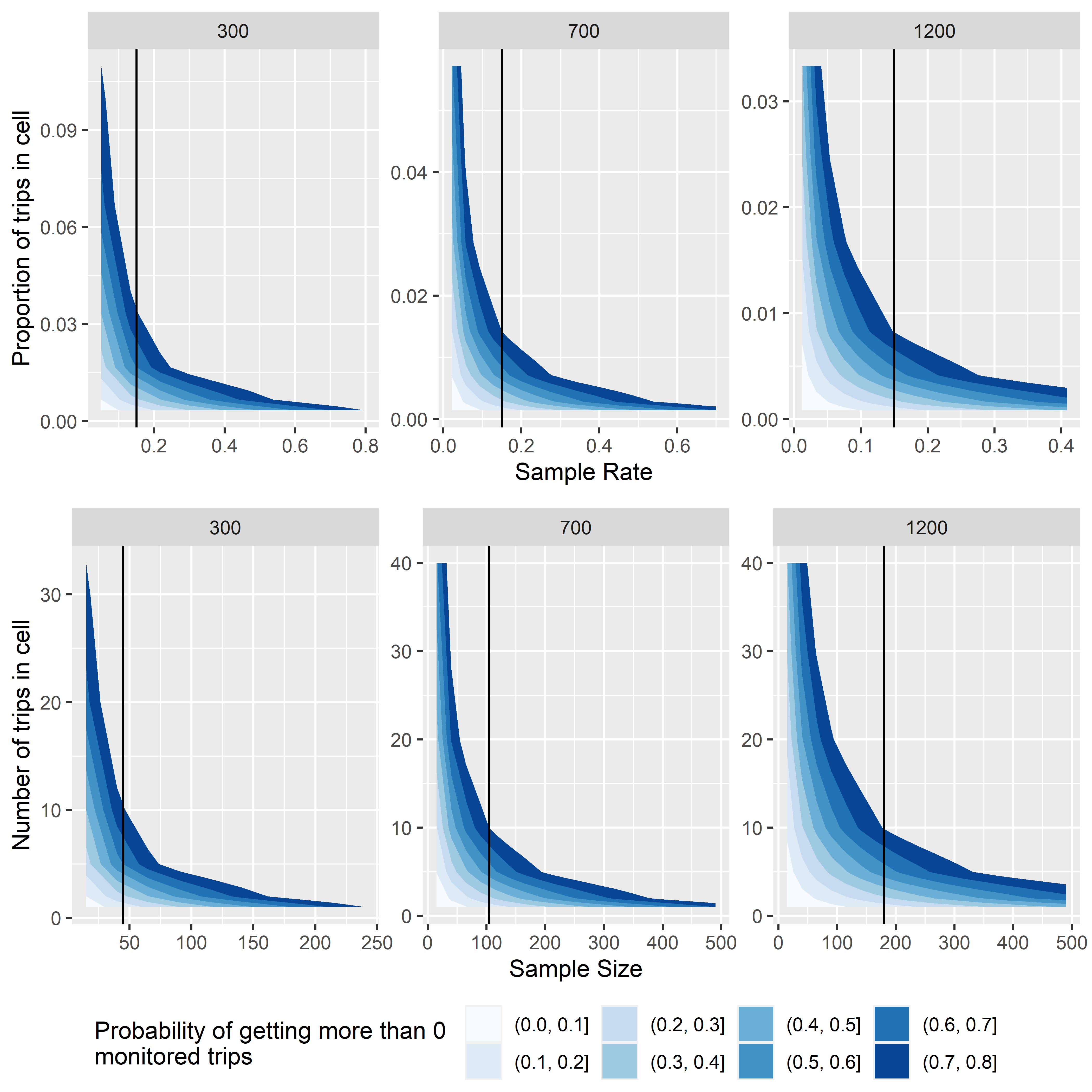


Figure 1.2: Probability of a sample containing no monitored trips as a function of sample size for a range of cell sizes. Vertical reference line on each plot indicates 15% sample rate.

## 1.2 Impacts of sample size on variance

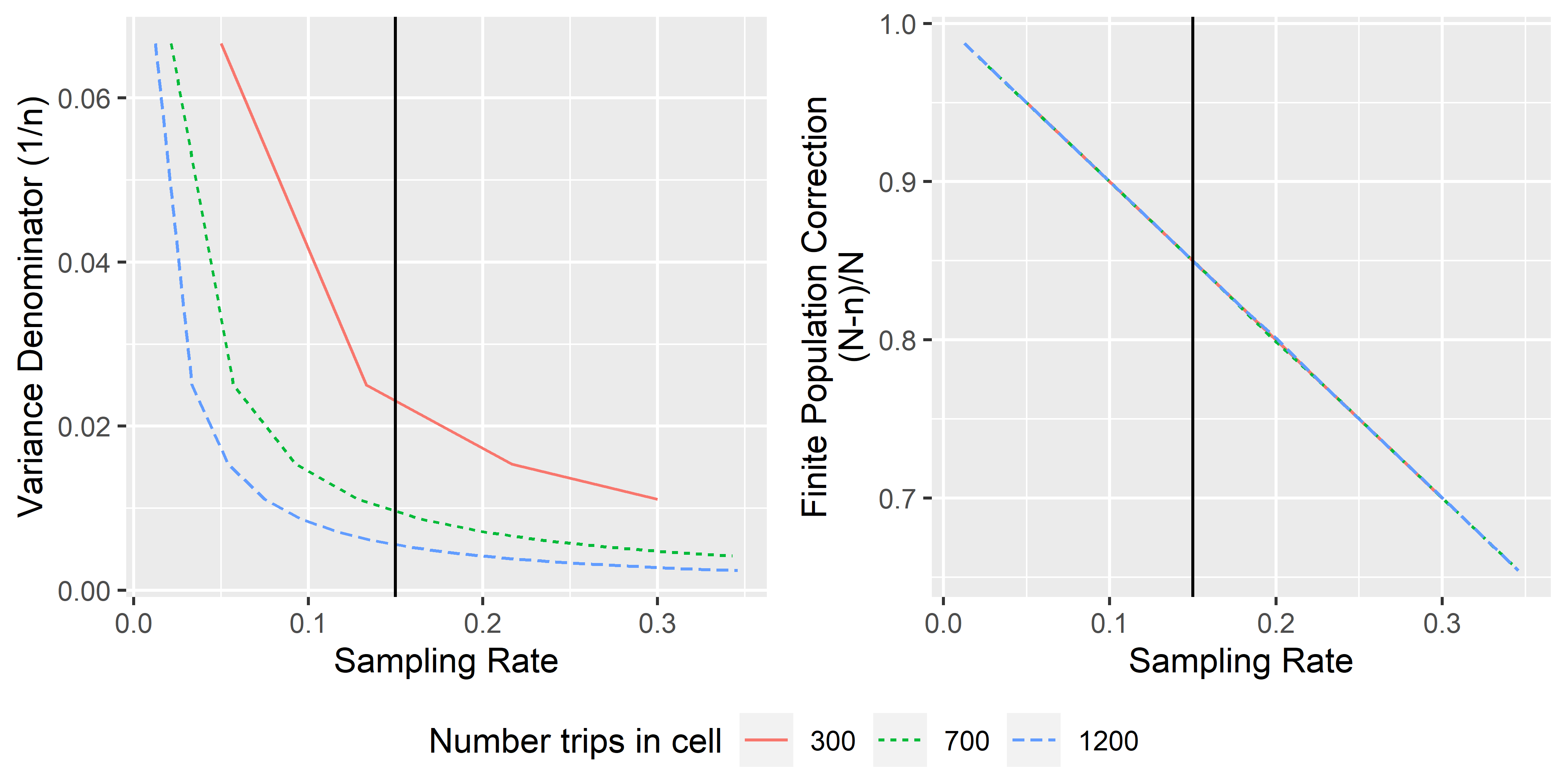
The variance in estimates from a population is a function of both the sample size and the overall proportion of the population that is sampled. Thile these are related quantities, they have different effects on the variance estimate.

All populations have a base variance, Eq (1.1); the variability in a measured parameter (e.g., length) between sample units.

The variance of the mean is the base variance (the population variance or variance of the data) divided by the sample size, Eq. (1.2). As the sample rate increases, variance will decrease by . Note that we are not summing only over those sample units that were sampled, but all samples in the population (i.e., *i* = 1 to *N* rather than *i* = 1 to *n*)

Unfortunaly, we rarely know the population variance and must rely variance estimates fro previous studies or on the sample data only. In estimating the parameter mean based on a sample, we don’t have information from every sample unit and hence will also need to estimate the variance (Eq. (1.3).

$$ We can see the impact of sample size on variance by plotting against the sample rate, , Fig. 1.3.

 In addition, as we sample a larger and larger portion of the population, our uncertainty about the estimate decereases. HERE BE THE FPC discussion, Fig. 1.3.

some things to note - generally don’t apply FPC when sample rates below 5%, doesn’t make a difference don’t know actual variance - makes sample size analysis challenging, Fig. 1.4.

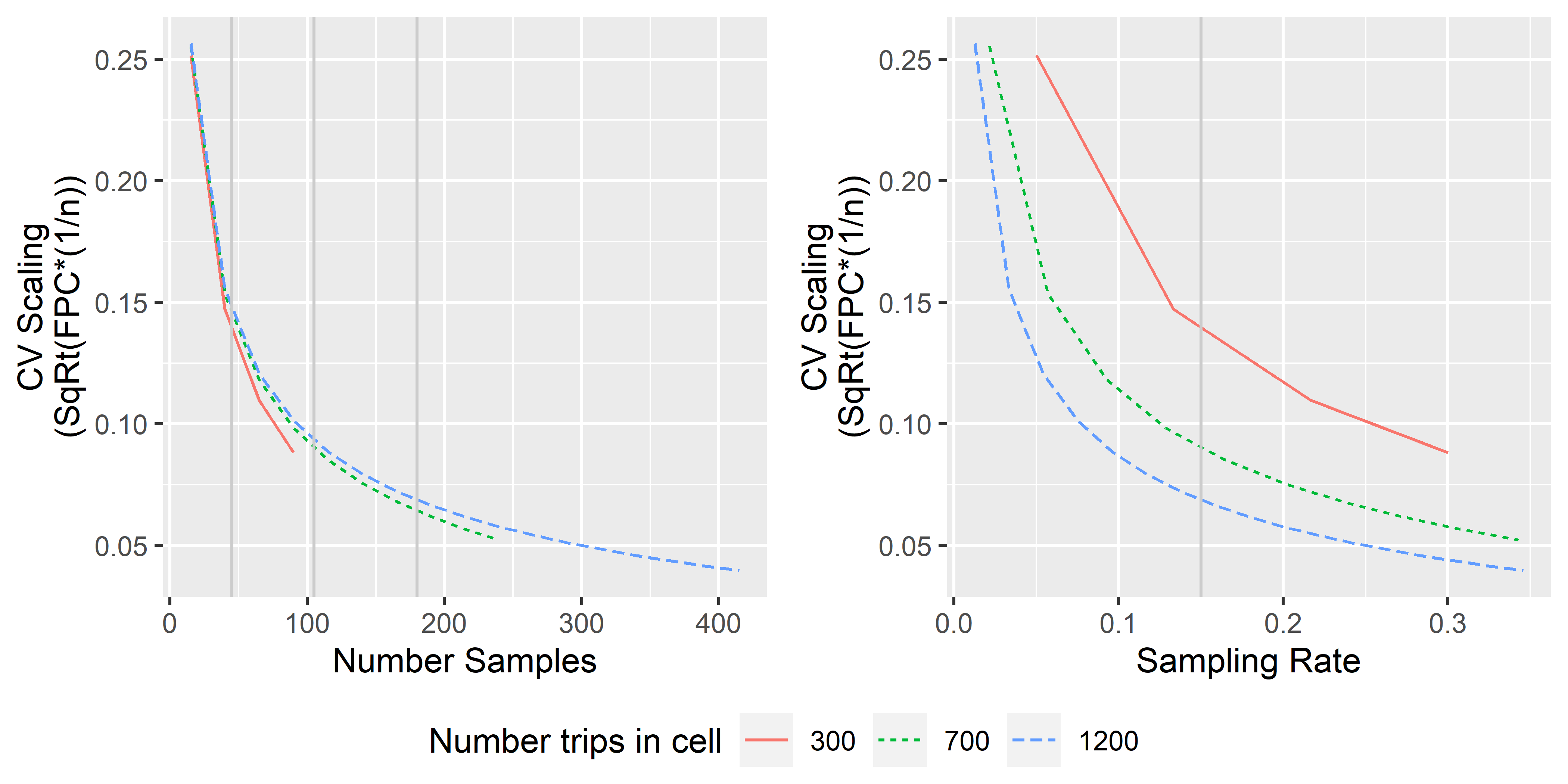


Figure 1.4: something

# 2 Cost-weighted boxes

## 2.1 Probability of monitoring as a function of sample size

In the [Cost-weighted Boxes Deployment Design](https://docs.google.com/document/d/1KxOkCigvOXgbxbzxSeJIGwBzLBWQgVX3iKs1wj4kdhI/edit#) (<https://docs.google.com/document/d/1KxOkCigvOXgbxbzxSeJIGwBzLBWQgVX3iKs1wj4kdhI/edit#>), the proposed allocation method is built on having spatial-temporal coverage that minimizes the proportion of post-strata likely to have no data, weighted by sampling costs, with the assumption that this will give us the best coverage per unit cost.

Starting with the simpler case that ignores costs and is based on the more traditional minimization of variance. The allocation scheme that minimizes variance of an estimate, for example the mean , based on a stratified sample with *h* = 1 to *H* strata and is the vector of sample sizes that minimizes the stratified variance estimate (Eq. (2.1)):

$$ where and

This can be minimized by setting the strata-specific sample sizes, , proportional to the standard deviation in each stratum. In otherwords, when where and *k* is a constant such that. More specifically, the overall variance will be minimized when

$$ Note that the relies on knowing , which is rarely the case. Generally, the estimated variance is used under the assumption that the estimate will accurately reflect the true variability in each stratum. This is the approach used above.

Allocation will be optimized when the sample rates in the strata are proportional to their standard deviations and inversely proportional to the stratum-specific cost of sampling, Eq.(2.3), below.

where is the cost associated with sampling stratum *h*.

### 2.1.1 One stage deployment

For the cost weighted box design, we would replace (variance) with proportion of post-strata within stratum *h* with a less than a pre-specified probability of having no monitored trips.

The probability of taking a sample of size *n* from a population *N* total trips where *K* trips are of interest (in a given cell), and getting *k* successes in the sample is given by (2.4).

The probability of the sample having no monitored trips in a given cell is given by (2.5).

If we know these probabilities for each cell in the stratum, we can define a “not monitored cell” as any cell where the probability of no data is greater than *a*, (i.e., ) based on the the number of trips in the cell (*K*), total number of trips in the population (*N*), and the sample size (*n*), Eq. (2.5).

Letting be the indicator of no data, for a given stratum, we have the proportion of cells with no data as:

where *i* from 1 to *B* indexes the total number of cells in the stratum.

Intuitively, we want to allocate more sample to strata that have a higher proportion of cells with few trips (*i.e.*, strata where there are a higher number of cells where *K* is small relative to *N*, see Figure 1.1).

Refering back to our original minimization problem and substituting in for, we could find the set of that minimizes the sum over all strata of having no data, \*i.e., minimizes (2.7).

Intuitively, we want to allocate more sample to strata that have a higher proportion of cells with few trips (*i.e.*, strata where there are a higher number of cells where *K* is small relative to *N*, see Figure 1.1).

By using the mean of the proportions, (Eq. (2.7)), each stratum would be weighted equally while using the sum of the proportions (divided by the total number of boxes, *i.e.*, the weighted mean), sampling strata with larger numbers of cells will contribute more to the overall proportion (metric being minimized).

For a given set of strata, with population sizes for each stratum, , and numbers of trips in each box in each stratum, , we need to estimate the sample sizes for each stratum, , that minimizes under either (2.7) or (2.8).

*If our goal is to minimize the total number of cells without data(across strata), Eq. (2.8), weighted mean proportion is the more appropriate metric. If we want to ensure that we minimize the proprtion of cells without data in each stratum, the mean of the proportions metric is more appropriate.*

#### 2.1.1.1 Cost optimized

If we want to “optimize” this in terms of costs of sampling, then we need to incorporate the differential costs associated with each stratum, balancing costs with the probability of not getting data.

Where is the mean cost per unit of sampling in stratum *h* and we need to minimize .

It is not obvious how to solve for and so we use simulation.

### 2.1.2 Two-stage deployment

In this version,

is defined as the rate at which the proportion of post-strata that have a prespecified probability of having at least one trip (or delivery) monitored.

Rank areas by the number of trips in each; areas will add into the proportion in that order, causing stair steps in some cases.

for a given number of post-strata (areas), each with a given number of trips, the sample size needed to acheive a pre-specified is difficult to compute since it involves solving the probability for *n*:

In addition, this will vary for each post-strata based on the number of trips in the post-strata. However, after a lot of math, we notice that only some terms vary with n and that some factorials cancel,leaving only falling products.

For example,

equals the falling product from *N* to *N-K+1*,

# 3 SEA-ish design (awaiting new name?)

## 3.1 from teh analytic document

Sample rates such that we *maximize* the percentage of trips that occur in those boxes that have greater than a predetermined probability of 1 or more trips per post-stratum (inspired by SEA methods of Gasper et al. 2017). Final sample rates are likely to differ between strata, but an important feature of this design is that strata are set to be equal in terms of their data resolution outcome.

We would change the percentage of trips (currently 75%) contained within post-strata with greater than 50% probability of having one trip, not the probability of having one trip observed. The different strata are likely to “respond” differently as that proportion changes (could plot this). Use hypergeometric distribution to estimate the probability of having one trip in a post-strata (instead of conducting simulation analysis).

Post-strata definitions will impact expected gap sizes and data resolution; larger post-strata will result in lower data resolution but perhaps fewer gaps. Hence, in considering post-strata definition, we can use 2 or 3-weeks temporal bins instead of 1 week and/or could drop/bin trip target and compare results to assess advantages of different post-stratification definitions or impacts of differing sample rates.)

## 3.2 from geoff’s description

define a set of hexagons \* note that placement of hex centers will impact the numbers of trips in each, hence allow ‘overlapping’ neighborhoods.  
\* identify the 6 hexagons immediately adjactent to hex-of-interest \* determine the whether each trip in the center hex has a sampled neighbor in either the center hex or the surrounding hexes. \* computed as the sample rate^number of trips \*