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, *m* = number sucesses in population (m+n=N=population size), and *k* = number of balls drawn from urn (see Eq. (??) 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, and 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)

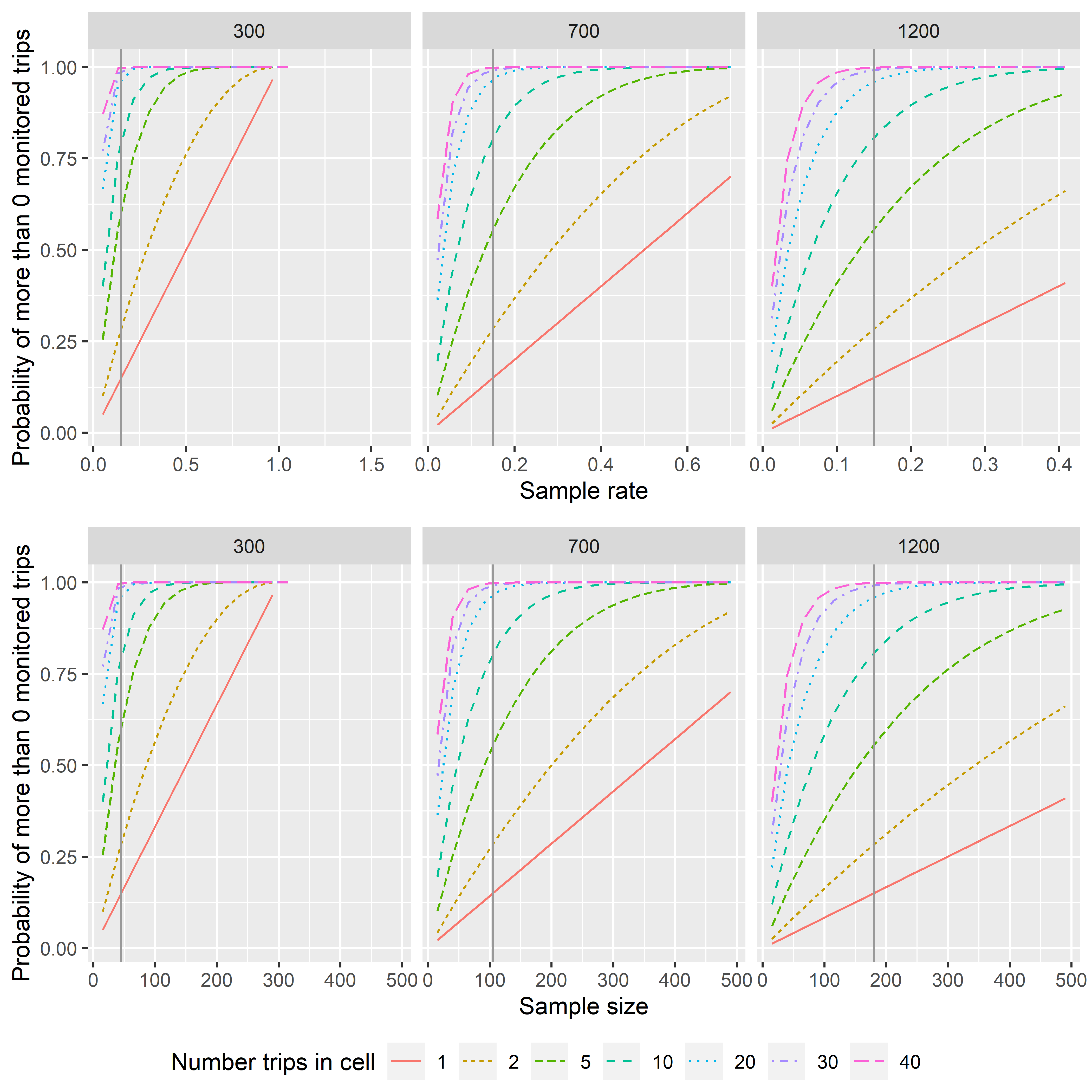


Figure 1.1: Probability of a sample containing no monitored trips as a function of sample size for a range of cell sizes. Vertical reference line on each plot indicates 15% sample rate.

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).

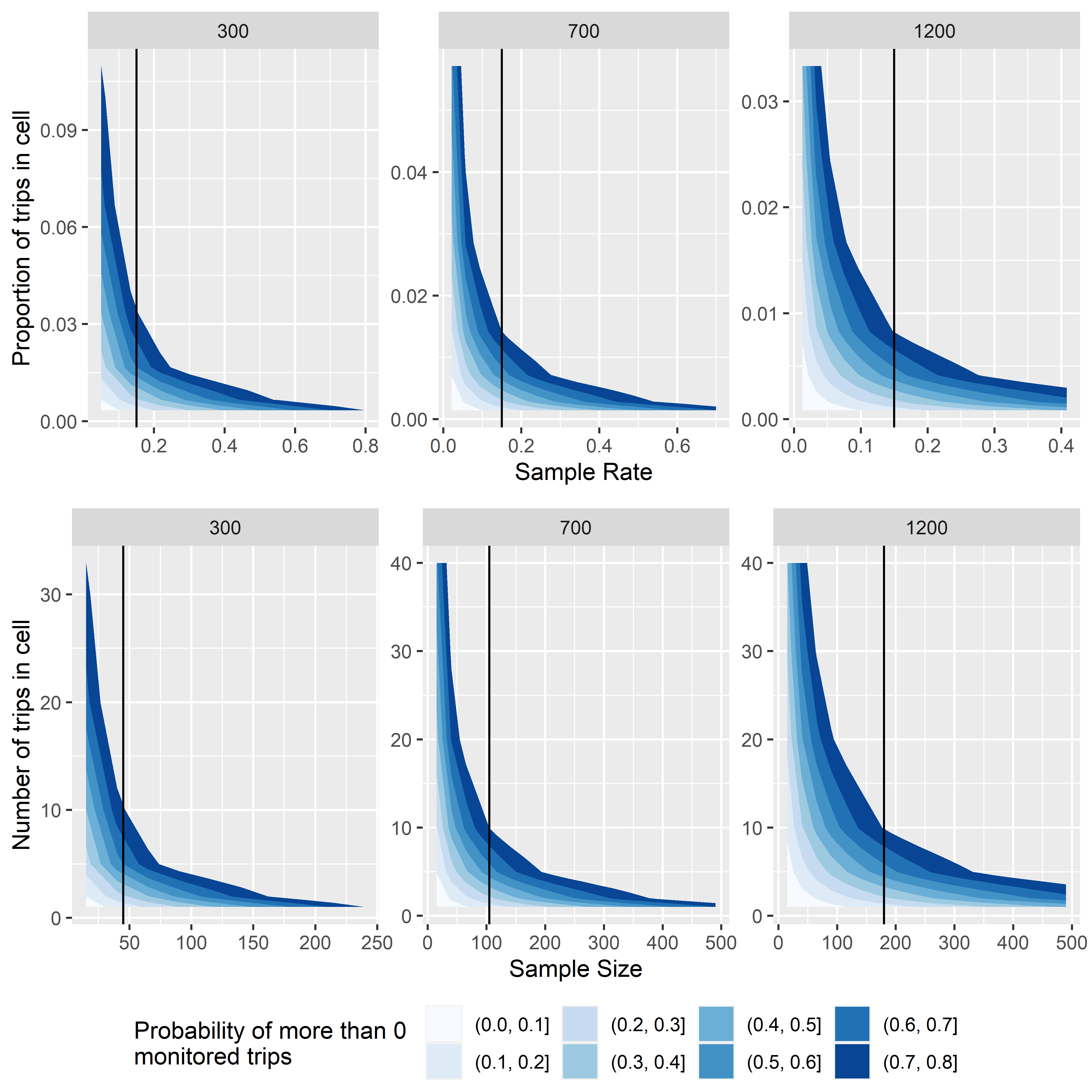


Figure 1.2: Probability of a sample containing no monitored trips as a function of sample size for a range of cell sizes. Vertical reference line on each plot indicates 15% sample rate.

# 2 Proximity design (awaiting new name?)

## 2.1 from the 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.)

## 2.2 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 (binomial approximation to the hypergeometric)

## 2.3 Proximity (or Interspersion) Index

This method is designed to spread sampled trips throughout the fisheries so that the proportion of trips that were sampled or near a sampled neighbor is consistent between strata. The premise being that is sampled trips are evenly distributed throughout all fishing activities, then data users will have a representative sample regarless of the type of analysis they are conducting. This allocation method was initially developed as a version of the methods used in the SEA analysis (citation), however the final version differs substantially.

By superimposing a regular spatio-temporal grid onto the fishing areas, we can quantify how trips are distributed in apace and time, both monitored and unmonitored. Then for each grid cell, the number of unmonitored trips with a sampled neighboring trip (either in teh cell or adjacent to the cell) is determined. The proportion of unmonitored trips that are expected to have a monitored neightbor is estimated as a measure of proximity; as sampling rates increase, the proximity index also increases. The proximity index is a function of the available budget, each stratum’s monitoring cost, sample rate, and spatiotemporal distribution of fishing effort. Strata with clustered fishing effort will achieve a specified proximity index at a lower sample rate than strata with more diffuse fishing effort; more samples are allocated towards strata with trips that are more spread out in space and time.

Although for a wide range of cell sizes (both temporal and spatial) the size of the cell does not impact the index, for this analysis cells were defined as a 200k km-wide hexagons and a 1-week period (see Apendix X). Neighbors are defined as any trip within the same cell or an adjacent cell (i.e. the 6 neighboring hexagons and the week before and after the trip). Trips that cross multiple cells are included in each cell, weighted by the inverse of the total number of cells the trip crosses (e.g. a trip that crosses two cells is included in each cell as a half trip in each).

The estimate of the number of trips that are expected to have neighbors is calculated as the sum of number of trips in the hexagon(s) multiplied by the probability that one or more of those trips are sampled. The index for a given stratum and sampling rate is the expected proportion of trips with neighbors (eq. ).

For a predetermined sample rate, different strata will have different sample rates. Therefore, for a given proximity index, the sample rate and associated cost of sampling will vary across strata. The final sample rates will be the rates that generate the maximal proximity index (common across all strata) that can be afforded for a total overall budget. The overall cost is the product of the number of trips in the stratum, the stratum-specific sampling rate, and the cost per trip for that stratum:

where is the stratum specific rate for the final proximity index value, is the cost per trip for stratum *h*, and is the number of trips needed to acheive the final proximity index value for stratum *h*.

Proximity index values were calculated for each stratum over a range of sampling rates. Based on these estimates, we can identify the sample rates for each stratum to acheive a proximity index value, and for each proximity value estimate the cost of sampling (Fig ).

## 2.4 CV Scaling

### 2.4.1 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. While these are related quantities, they have different effects on the variance estimate.

All populations have a base variance, Eq (); the variability in a measured parameter (e.g., length) between all sample units (both in the sample and unsampled). 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*). For the ADP, the population variance is the between trip variance over all trips in a stratum and will be different for different species (years, gear types, *etc.*)

This variance is used to estimate the variance associated with estimates; the variance of the mean is the population variance divided by the sample size, Eq. (). The square root of the variance of the estimate is the standard deviation and, in some situations, can be used to estimate confidence intervals. As the sample rate increases, variance will decrease by until all sample units are included in the sample (sample rate = 100%), at which point the standard deviation will be zero.

Unfortunaly, we rarely know the population variance and must rely variance estimates from 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. , note the substitution of *n* for *N* in the equation below).

We can see the impact of sample size on variance by plotting against the sample rate, , left panel of Fig. 2.1.

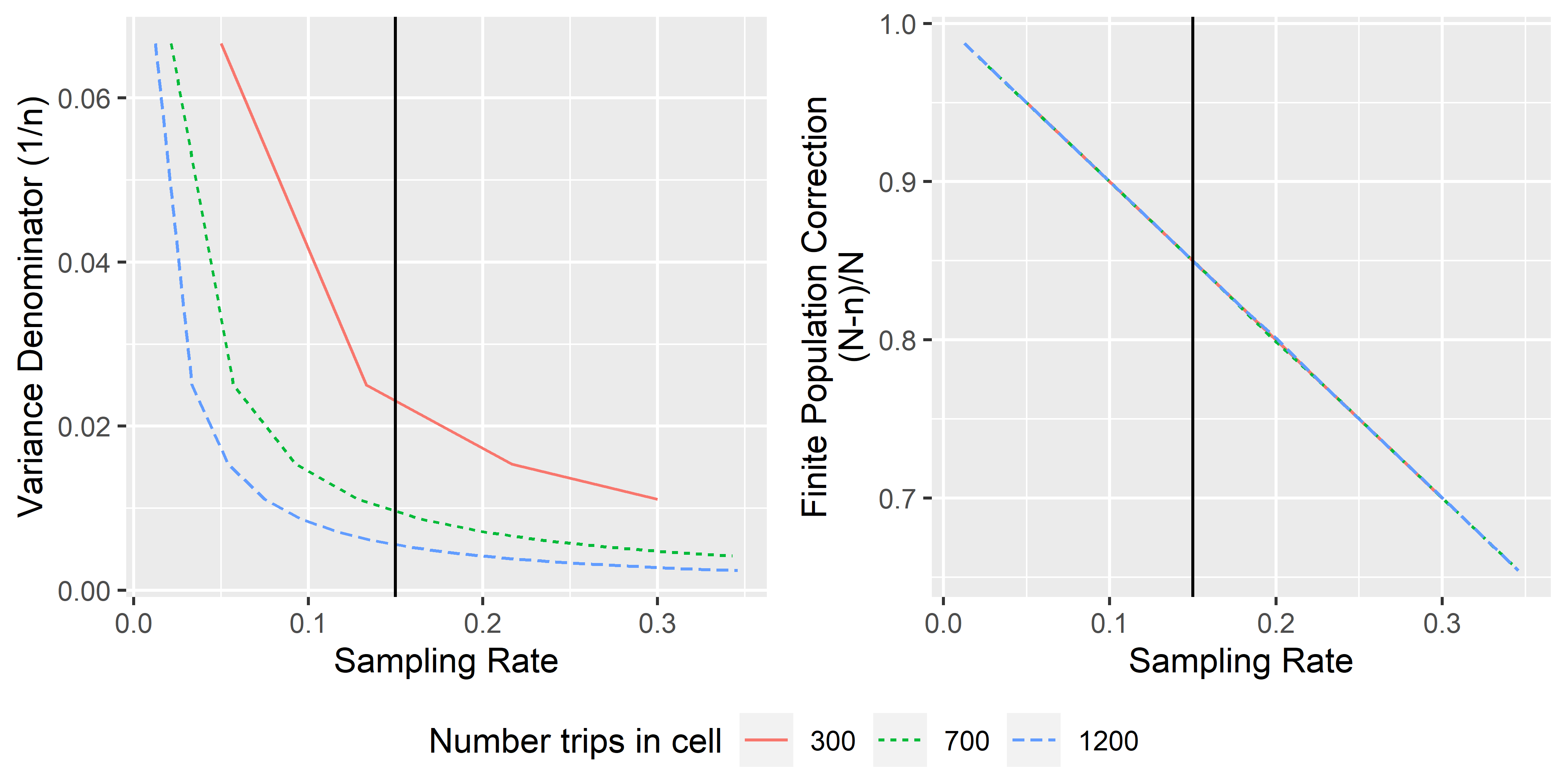


Figure 2.1: something

In addition to the variance savings we gain by increasing sample size, as we sample a larger and larger portion of the population, our uncertainty about the estimate decereases, right panel of Fig. 2.1. This savings, the finite population correction factor (FPC), is a linear function of the population size (). Generally, the FPC is not applied at sample rates less than 5% since the savings is minimal. This gives teh traditional variance formula for the estimated mean (eq X).

There are several difficulties with setting sample rates for monitoring programs in order to decease variance of the estimates. First we don’t have an estimate of variance for the parpameters being estimated (e.g., species specific catch). In the CAS, variance estimation incorporates several analysis steps including within trip sampling variance, post-stratification of estimates, use of ratio estimates. All of these steps impact the variance estimates differently for different species. Second, we have many priorities (e.g., many species-specific estimates) and each has a different undrelying population variance. As a result of these issues, traditional sample size setting methods, such as Neyman and optimal allocation, are not suitable to multi-objective monitoring program design. However, the FPC and sample size do not depend on the parameter being estimated, and although the savings “rate” will not necessarily be the same for all species, any variance savings that we acheive will be relevant for all estimates.

The FPC and sample rate combine to form a single scaling factor, *F*, Fig. 2.2, eq X.

In more complex sampling designs, this scaling facor is also more complex reflecting differences in sample size within trips, within post-strata, and withing strata. However, for our purposes, we can use the simple version of the CV scaling factor to buffer teh proximity index and decrease the potential for strata to have very low sample sizes. *(need to include description of what happens with low sample size and why we want adequate smaple everywhere)*

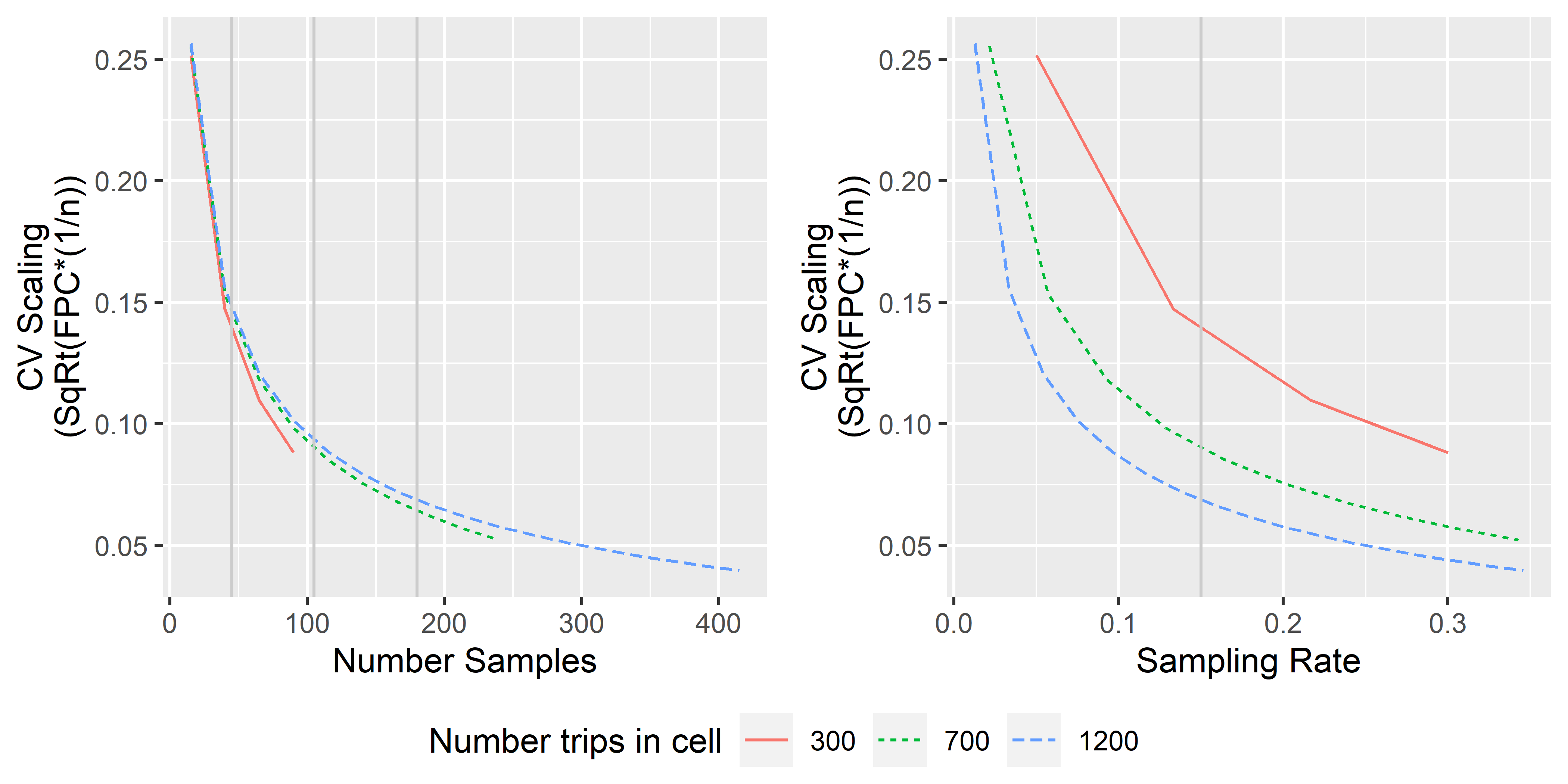


Figure 2.2: something

## 2.5 Proximity Index Allocation

Recall that for a given sample rate, higher proximity index values were associated with strata with spatiotemporally clumped fishing effort, and thus were allocated a smaller portion of the total sample amount. While the proximity index indicates that a large porpotion of unmonitored trips are located near monitored trips, allocation based soley on this index can result in very small sample sizes for some strata. Since variance is a function of sample size, these small sample sizes can lead to catch estimates with high variability. In addition, estimated length and age composition data that drive some stock assessments will be sparse, leading to stock assessment harvest recomendations with higher uncertainty.

To buffer against low sample sizes, the CV scaling parameter was incorporated into the allowcation process such that the proportion of trips with neighbors () is multiplied by the CV scaling factor (eq X).

As before, an iterative approach is used to find the values of for all strata so that the index is consistent across strata, maximizing the value that can be acheived for a predetermined cost.

# 3 Cost-weighted boxes

## 3.1 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 unmonitored cells, weighted by sampling costs, under 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. (??)):

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.(??), below.

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

### 3.1.1 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 (??).

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

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. (??).

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 (??).

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. (??)), 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 (??) or (??).

*If our goal is to minimize the total number of cells without data(across strata), Eq. (??), 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.*

#### 3.1.1.1 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.

### 3.1.2 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*,

## 3.2 New

This allocation distributes sampling effort to partial coverage strata (i.e., both observed and electronic monitoring) so that, for a given amount of funding, the proportion of predefined spatio-temporal boxes that we expect to be unmonitored is consistent across strata. Note that it does not allocate sample in order to get data from all boxes and hence, it is not allocating sample effort to ensure we have data from any particular box (not chasing small boxes). It is also important to note that the proportion of boxes that are expected to be unmonitored is not minimized per se, but rather sample effort is added to each stratum iteratively to maintain a consistent expected proportion of unmonitored boxes across all strata.

To achieve this, spatial-temporal hexagons are defined the same way as in the Proximity allocation scheme: spatial hexagonal cells 250km across and temporal block 1-week in length. Trips are allowed to span multiple boxes in both time and space (based on the landing reports), and trips are considered neighboring if they are in adjacent spatial or temporal cells.

The probability that we do not have any data is estimated using the binomial approximation of the hypergeometric distribution where *e* is the number of trips in a cell (), 1 minus the probability that all trips in the cell are monitored. Using this, we calculated the expected proportion of boxes that will not be monitored given a sampling rate, , as the average across all cells of the probability of having no data. **(this will need clarification - the probabilty is the number of monitored boxes over the total number of boxes; I think this might just be the sum, not the average???)**

where indexes the number of hexagons in stratum , is the total number of hexagons in the stratum, is the initial (assumed) sample rate used to estimate the probability that a hexagon is unmonitored, and is the number of trips in hexagon and stratum .

Since the number of trips in each stratum (and each year) differs, the rate that decreases with increasing sample rate also differs with stratum and year ().

{r logo-apps, fig.align = 'center', out.width = "25%", fig.cap = "The APPS logo for our 50th year!"} knitr::include\_graphics(here::here("figs", "logo-apps.png"))