Andrew Peters

Sample Design

Cranberries are a small portion of the agricultural industry in the United States. Making up $3-$4 billion dollars (citation Needed) net value added of the national $230 billion dollars of the agricultural sector of the United States economy. Cranberries are culturally and economically significant during the winter holidays such as Thanksgiving and Christmas. They have also impacted my life in a significant way, as I have grown up working on a cranberry farm in the small town of Bandon, Oregon.

Cranberries are unique in their optimal growing conditions. Grown in cool climates and in beds of sand. They require large amounts of water to grow, but contrary to popular belief, they do not grow in water. The famous pictures of cranberries floating atop a great pool of water are exclusive to harvesting. Where beds are flooded, and the vines are agitated in a specific way to cause the berries to break loose and float to the top of the water. This is not a report about how cranberries are cultivated and harvested. It is important, however, to get a baseline understanding of the fundamentals of cranberry farming before introducing the Sampling Design which will be the main topic of this paper.

While cranberry farming, and farming in general, is fundamentally unpredictable and highly variable from year to year. I will introduce the two factors of cranberry yield that will be important to this design. In general, farmers grow their cranberry vines in beds of sand we call “bogs”. It is convention to shape these fields into rectangles, of sizes between 1-3 acres, with sprinklers in squares that are 30ft by 30ft. This allows for even spread of fertilizer. With fertilizer usage being one of the factors considered in this sample. Second, we will be looking at different vine varieties. In the last seven years, agricultural schools in the Midwest have introduced hybrid cranberry vine varieties which have significantly increased average crop yield.

This paper will look at an experiment performed by me on different cranberry farms, which investigated the impact of fertilizer usage on crop yield. I performed this experiment prior to entering the master’s program at Purdue, so I believe it would be beneficial to design the sampling methods correctly; to have an outline of the correct methods, should I desire to run the experiment again.

The purpose of this study, and by extension, the purpose of this sample design, is to estimate the impact of fertilizer usage on cranberry yield, particularly when looking at the most popular vine varieties. This is important because the number one cost to a cranberry grower is fertilizer, followed by the price to plant new hybrid varieties of cranberry vines. While outside the scope of this paper, if one could ascertain the lowest level of fertilizer needed to show a nonsignificant decrease in yield, the cost reduction would directly help the family owned and run farms that are the backbone of small towns, such as Bandon, Oregon. With the background of cranberries understood and the purpose of the sample stated. We can now begin with the design itself.

The population of inference for this sample are the cranberry bogs in Coos County, Oregon, generally ranging between 1-3 acres per bog. The cranberry bogs are managed by local farmers and will be the primary population units being analyzed in this sample. The measured information available is historic yield data measured in barrels (100lbs per barrel), fertilizer usage (lbs./acre), and vine variety. This information will be acquired through personal connections. Cranberry growers in the area have regular meetings and keep detailed records of yield, fertilizer usage, and vine varieties of individual bogs.

The variables I will use for the sample will be fertilizer usage, stratified into low, medium, and high application rates. Which for the purpose of this paper will remain arbitrary. The vine varieties will be stratified into hybrid and non-hybrid. Due to the semi arbitrary nature of this paper, it will suffice to say from personal knowledge, that the popular new hybrid varieties yield approximately the same number of berries when compared to each other, and significantly higher yield when compared to non-hybrid varieties. This is an area where if a true experiment were conducted, it would be necessary to investigate further. This is also an area where individual Farms could serve as clusters if for example cost were a more prohibitive factor. I would have access to the necessary information easily, which is why I am leaning away from clusters, and using stratification.

The primary variable of estimation is yield (barrels/acre) for which we will use mean yield and total yield. The secondary variable is the impact of fertilizer usage, which will be estimated with a regression estimate. With an additional estimate looking at the yield difference by vine variety, in which we will use domain estimation, where the domains hybrid and non-hybrid act as subpopulations.

With the definitions of the population, population units, measured information, and descriptions of variables to be estimated and methods of estimation stated, we can focus on taking the sample itself.

The sampling frame will consist of cranberry bogs in Bandon, Oregon. This frame will be constructed with data from records detailing fertilizer usage, vine variety, and historical yield data. The list of bog data will be complete, meaning all farmers who give access to their data will have all the listed data points available for all production bogs. To construct this sampling frame, the volunteer farmers will supply individual data to be added to a master spreadsheet for analysis. This sampling frame will not be able to include all farmers in Oregon, but those included will have accurate and usable data.

Sample design will consist of stratification by fertilizer usage (low, medium, and high application rates), and vine variety (hybrid vs. non-hybrid). Assuming this experiment was conducted, Python would be the chosen tool to organize the data and run analysis. With this assumption, the randomization and selection process will be done with a Python random number generator.

The selection will randomly choose bogs within each stratum. Optimal allocation will be used to determine the sample size of each stratum, considering both the variance and cost of collecting the data. This will ensure that strata with higher variability or lower cost will be prioritized. The allocation of samples to high variability strata such as medium fertilizer hybrids will help improve precision in the estimates. Most cranberry bogs will fall into the medium fertilizer non-hybrid variety stratum due to different factors. For example, a single farmer fertilizes many other farms, which implies more consistent fertilizer application. Hybrid vine varieties were recently introduced to Oregon, meaning they are still in the early stages of adoption. They are also expensive to incorporate into a farm, as they must be purchased at premium prices, take up to five years to reach full production, and must replace older producing bogs. Hybrid varieties being as new as they are, there will be a large portion of bogs that are only partially producing. Thus, by using optimal allocation, the sample design will ensure representation of strata while addressing the differences in variability. Costs will be low in this sample as all the data will be provided by farmers for free, as the outcome of the experiment could relay useful information.

Using, $n_h = n \cdot \frac{N_h \cdot \frac{S_h}{\sqrt{c_h}}}{\Sigma\left(N_h \cdot \frac{Sh}{\sqrt{c_h}}\right)}$

- n_h: Sample size for stratum h
- n: Total Sample Size
- N_h: Population of Stratum h
- N: Total Population size
- S_h: Variance of Stratum h
- c_h: Cost observation in Stratum h

This method of optimal allocation ensures minimization of variance and cost in each stratum. With the sample sizes within strata determined, we know that the probability of selection is $\pi_i = \frac{n_h}{N_h}$.

Looking first at the primary variable of estimation, yield (barrels/acre).

Using Mean Yield across all strata:

$$\overline{y_{str}} = \sum_{h=1}^{H} W_h * \overline{y_h}$$

- H: Total Strata
- W_h: Weight for Stratum h, with $W_h = \frac{N_h}{n}$
- $\overline{y_h}$: Sample mean in Stratum h

And, estimated Total Yield:

$$\hat{t} = \sum_{h=1}^{H} N_h * \overline{y_h}$$

The variance for the mean yield:

$$\mathrm{Var}(\overline{y_{str}}) = \sum_{h=1}^{H} (1 - \pi_h) W_h^2 \cdot \frac{S_h^2}{n_h}$$

With an approximate confidence interval found by:

$$\bar{y} \pm z_{\alpha/2} \cdot SE(\bar{y})$$

$$SE(\bar{y}) = Var(\overline{y_{str}})$$

Our secondary variable of interest is the impact of fertilizer usage on yield. This can be estimated with a regression model with an auxiliary variable fertilizer usage of the form:

$$y_i = \beta_0 + \beta_1 x_i$$
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$
$$\widehat{\beta_1} = \frac{\sum_{i \in S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in S}(x_i - \bar{x})^2}$$

Where:

- y_i: Yield of bog i.
- x_i: fertilizer usage for bog i
- β_0: Intercept of regression line
- β_1: Represents change in yield per fertilizer.

This is another area that after changing the arbitrary levels of fertilizer to measurable units will be a far better estimate.

Our tertiary estimate of yield changes based on vine variety is an area that would also need refinement based on factors outside of the scope of this paper. For simplicity the vine varieties were split into hybrid and non-hybrid varieties. However, there are four or five varieties that dominate the space in Bandon. Therefore, while the specifics could be refined outside this project, the baseline estimation of differences between the vine varieties is warranted.

$$\overline{y_d} = \frac{\sum_{i \in S_d} y_i}{n_d} = \frac{yield\ for\ ith\ bog\ in\ domain\ d}{size\ of\ domain\ d}$$

Then to compare the difference in yields (more vine varieties possible in later iterations of experiment).

$$\Delta Y = \overline{y_{\text{hybrid}}} - \overline{y_{\text{non-hybrid}}}$$

Lastly, looking at the design effects, which indicate the factor of increase in the variance of our estimates. As or mean yield estimate is a key feature of the questions asked for our experiment, finding the effect on the variance will give an important understanding of the impact on our estimates. The unequal weighting effect will provide an estimate of the increase in variance due to the unequal weights caused by the proportional allocation of samples towards larger strata.

$$UWE = 1 + s_w^2/\overline{w^2}$$

The stratification effect provides an estimation of the change in variance of the mean due to stratification.

$$SU \approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{S^2} \approx \frac{\sum_{h=1}^{H} \frac{N_h}{N} S_h^2}{\sum_{h=1}^{H} \frac{N_h}{N} [S_h^2 + (\overline{y_{Uh}} - \overline{y_U})^2]}$$

The overall estimated design effect for our weighted stratified design is:

$$DEFF_{strat} = UWExSU$$

Now the ratio of the design effect of our stratified sample and the design effect of a simple random sample will give us an idea of the impact on the variance of the mean estimate.

$$DEFF_{SRS} = S^2/n$$

$$DEFF = DEFF_{strat}/DEFF_{SRS}$$

If the ratio is less than one, then stratification reduces variance compared to a simple random sample. Otherwise, there is no difference in variance in the different designs.

In conclusion, this sampling design aims to estimate the cranberry yield and the role that different fertilizer levels may play in growth of differing vine varieties in Bandon, Oregon. The utilization of a stratified random sample using optimal allocation to ensure proportional representation is a cost-effective approach that is representative of the population. The analysis uses point estimates for mean yield, total yield, and regression estimation to compare the impact fertilizer has on crop yield. Domain estimation studies the role vine variety has in the growth of cranberries and the end of year yield. While there are portions of this paper that may overreach the overall scope of the project, this sample design offers a robust framework from which experiments can be conducted. This design is also expandable to incorporate other key features of cranberry growth, such as environmental factors and other characteristics which impact the yield during a growing season.