

Statistical Considerations for Agroforestry Studies¹

James A. Baldwin²

Abstract: Statistical topics that related to agroforestry studies are discussed. These included study objectives, populations of interest, sampling schemes, sample sizes, estimation vs. hypothesis testing, and P-values. In addition, a relatively new and very much improved histogram display is described.

As the title implies, I would like to discuss various statistical topics that relate to agroforestry studies. I will cover a few points on study objectives, then move on to sampling and analysis, and finally describe a new data display technique.

Study Objectives

Study objectives are crucial to any study, but I have found that in many studies the objectives are only written down when the final report or manuscript is being prepared. These objectives need to be examined by peers in your field along with the rest of the study plan. After such review, the study objectives should be capable of being realized, specific, and a fixed—not moving—target. You will get the credit for good work, and your reviewers can share the blame if something is amiss with the objectives and design.

Population of Interest

After the objectives have been decided upon, the population of interest needs to be defined; for example:

- All farms on Pohnpei
- 23 farms on Pohnpei that introduced a new agroproduct since 1988
- One particular farm
- One particular area of a particular farm
- All farms with mango trees

All of the above examples are legitimate populations of interest. The important point is that the population needs to be defined before any of the sampling begins. All of your inferences will be directed to this population.

Unfortunately, one is not always able to sample the population of interest. Typical reasons for this are timing, not having permission granted, and lack of accessibility. These problems lead to differentiating between the “target” population and the “sampled” population.

Inferences about the sampled population are based on appropriately collected data. Inferences about the target population are based on how well you can convince someone about the

similarity of the target and sampled population. After reflecting upon these two populations, you usually need to reconsider your study objectives.

Sampling Schemes and Estimators

Three basic types of sampling schemes are available: Purposive sampling, Systematic sampling, and Probability sampling.

Purposive sampling is sometimes called “convenience” sampling. Statisticians also use even less flattering terms for it. An example is “That tree looks typical. Let's sample it.” The obvious problem is that this type of sampling introduces the biases of the person sampling (not necessarily the researcher). In addition, your inferences from such collected data will be suspect at best. Because with little additional effort one can use a sampling scheme with known properties, I cannot recommend purposive sampling for any scientific inquiry.

Systematic sampling is sometimes used if it is convenient to take a sample in some regular order. For example, every fifth tree could be chosen rather than a simple random sample of trees. A sample mean from such a sampling scheme can be more precise than that of a simple random sample. Unfortunately, the estimate of the precision of a systematic sample can require stringent assumptions to be accurate.

Within probability sampling, we have simple random sampling, stratified random sampling, PPS (Probability Proportional to Size), and SALT (Sampling At List Time). Only simple random sampling and PPS sampling are described below.

For a simple random sample of plot centers on an island, just overlay a rectangle on a map of the island. Sample points are selected by choosing uniform random numbers on each of the horizontal and vertical scales. Ignore any points that fall in the ocean. Continue until you meet the required sample size. Unfortunately, this scheme will not get you a simple random sample of farms.

If you are selecting farms, one method is to choose each farm with a probability proportional to its size. If you do not know its size, then the “uniform grid” method described earlier will result in such a sampling scheme (PPS sampling).

To fix ideas, suppose we have the following data on five farms:

Farm:	A	B	C	D	E
Acres:	10	20	30	50	100
Tons of mangoes:	9	23	35	43	105

Suppose we want to sample two farms and estimate the total mango production (from this example we know that the total is 215 tons). (Any resemblance to actual mango production is purely coincidental and extremely unlikely.)

¹ An abbreviated version of this paper was presented at the Workshop on Research Methodologies and Applications for Pacific Island Agroforestry, July 16-20, 1990, Kolonia, Pohnpei, Federated States of Micronesia.

² Mathematical Statistician, Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, P.O. Box 245, Berkeley, CA 94701.

Simple Random Sampling

We can choose two farms for a simple random sample in two ways. In the first method we randomly select one farm and determine the mango production on that farm. For the second farm we randomly choose one farm from the remaining farms and determine its mango production. This is called “simple random sampling without replacement” because each farm can only be chosen once.

The complete list of potential samples (ignoring the order of selection) of size 2 (without replacement) is

AB,AC,AD,AE,BC,BD,CD,CE,DE

If we sample “with replacement,” then that means that a farm could be selected on the first draw and again on the second draw. The complete list of potential samples (again ignoring order) of size 2 with replacement is

AA,AB,AC,AD,AE,BB,BC,BD,BE,CC,CD,CE,DD,DE,EE

If we chose farms A and C by either method, we would take the average mango production and multiply by 5 to estimate the total mango production:

$$\text{estimate} = 5 \cdot (9+35)/2 = 110 \text{ tons}$$

This formula is just the total number of farms multiplied by the estimate of the average production per farm. Again, we know that the “true” total is 215 tons.

PPS-with Replacement

The PPS-with replacement sampling scheme needs more explicit formulas to describe how it works. To generalize, suppose our example consists of a sample of size n with replacement and probability proportional to a farm’s area is taken from a population of N farms. For farm i , the area is labeled a_i and the measurement of interest (tons of mangoes) is labeled y_i . We want to estimate the sum of all of the y_i ’s, namely,

$$Y = \sum_{i=1}^N y_i$$

One estimate of the total is the following

$$\hat{Y}_{ppz} = \frac{1}{n} \sum_{i=1}^N \frac{y_i}{z_i}$$

where z_i is the probability of selecting farm i on any one draw.

$$\text{Usually } z_i = a_i / \sum_{j=1}^N a_j$$

An estimate of the variance of $\hat{Y}_{ppz}$ is given by

$$v(\hat{Y}_{ppz}) = \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{ppz} \right)^2 / n(n-1)$$

If $z_i = \frac{1}{N}$, then each farm has an equal chance of being selected and we have a simple random sample with replacement.

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{1/N} = N\bar{y}$$

We also end-up with the usual variance formula.

PPS-Without Replacement

We use the same notation as before. The only difference now is that we sample without replacement, i.e., no farm can be chosen more than once.

One estimate of the total is the Horvitz-Thompson (HT) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}$$

where π_i is the probability of selecting farm i in the sample. An estimate of the variance of $\hat{Y}$ is given by

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n-1} \sum_{j>i}^n \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

assuming that all $\pi_{ij} > 0$ where π_{ij} is the probability that both farms i and j are included in the sample.

If we call the probability of selecting farm i on the first draw p_i , then $p_i = a_i / \sum_{j=1}^N a_j$. In other words, the probability of selection (on the first draw, at least) is proportional to the size of the farm.

When $n=1$, then $\pi_i = p_i$. When $n=2$, then

$$\pi_i = p_i \left(1 + \sum_{j \neq i}^N \frac{p_j}{1 - p_j} \right)$$

When n is much bigger than 2 the formulas become increasingly complicated and the π_i ’s need to be estimated from simulations.

An alternative for larger sample sizes is Murthy’s estimator

$$\hat{Y}_M = \frac{1}{P(s)} \sum_{i=1}^N y_i P(s|i)$$

where

$P(s|i)$ = conditional probability of getting the set of farms that was drawn, given that the i th farm was drawn *first*

$P(s)$ = unconditional probability of getting the set of farms that was drawn

Even this estimator becomes nearly impossible to calculate without simulations when n is much bigger than 11 or 12.

The estimate of the variance of $\hat{Y}_M$ is given by

$$v(\hat{Y}_M) = \frac{1}{P(s)^2} \sum_{i=1}^n \sum_{j>i}^n [P(s)P(s|ij) - P(s|i)P(s|j)]$$

$$\cdot p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2$$

where $P\langle s|ij \rangle$ is the conditional probability of getting the observed sample farms given that farms i and j were selected in the first two draws.

Comparing the Sampling Schemes

The percentage of time that any two particular farms would be selected under the four sampling schemes can vary (table 1):

Simple random sampling with and without replacement and PPS sampling with and without replacement. For example, under PPS sampling without replacement, we expect to obtain farms D and E in our sample 36 percent of the time.

Each combination of farms for each sampling scheme yields varying values (table 2). Notice that all sampling methods are unbiased: all have a mean of 215 tons. But the standard deviations differ. The estimator for PPS with replacement has a standard error only one-seventh the size as that of the simple random with replacement estimator. Apparently the sampling scheme can make a large difference in the precision of the summary statistics.

Sample Size

“What sample size should I take?” is one of the most frequently asked questions a statistician helps to answer. And the answer depends on several facts that you need to supply the statistician.

If you are estimating a population statistic (such as total farm production of mangoes), then you need to tell the statistician how close you need to be to the true value. The statistician will translate this into a statement something like “95 percent of the time we want to be within 2.5 tons of the true total production.”

One common misconception is thinking about an adequate sample size in terms of a proportion of the population size. We hear “we took a 5 percent sample” or even “we took *only* a 5

percent sample.” If there is one thing I would like to convince you about, it is thinking about sample size as an absolute number rather than as a percentage of the total population size.

For example, if we sampled 10 individuals from a population of 1,000 individuals, we would get almost exactly the same precision for our estimator as if we had 1,000,000 individuals in the population. This happens despite the wildly different relative sample sizes (10 out of 1,000 vs. 10 out of 1,000,000).

This can be seen from the formula of standard error. If N is the population size, n is the sample size, and σ is the standard deviation of the population, then the standard error is given by

$$s.e. = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$$

When n is small compared to N , the rightmost term, $\sqrt{(N-n)/N}$ is very close to 1 and, therefore, does not influence the standard error. It is the term $1/\sqrt{n}$ that has the most influence and it only depends on the absolute (and not the relative) sample size.

Estimation vs. Hypothesis Testing

Long before analyzing the data, the researcher needs to decide about which questions need to be placed in “Hypothesis Testing” terms and which in “Estimation” terms.

Estimation and hypothesis testing try to answer two different types of research questions. For example, estimation might try to answer the question “How much change in production occurred from the previous year?” A similar question for hypothesis testing might be “Is there a large change from the previous year?”

Table 1-Percentages for each potential sample for various sampling schemes¹

Farms selected	Simple random (wr)	Simple random (wor)	PPS (wr)	PPS (wor)
AA	4	0	0	0
AB	8	10	1	1
AC	8	10	1	2
AD	8	10	2	3
AE	8	10	4	7
BB	4	0	1	0
BC	8	10	3	3
BD	8	10	4	6
BE	8	10	9	14
CC	4	0	2	0
CD	8	10	7	8
CE	8	10	14	21
DD	4	0	6	0
DE	8	10	23	36
EE	4	0	23	0

¹ wr = with replacement
wor = without replacement.

Table 2-Estimates for each potential sample for various sampling schemes¹

Farms selected	Simple random (wr)	Simple random (wor)	PPS (wr)	PPS (wor)
AA	45	-	189	-
AB	80	80	212	175
AC	110	110	217	179
AD	130	130	185	157
AE	285	285	205	211
BB	115	-	242	-
BC	145	145	243	202
BD	165	165	211	180
BE	320	320	231	234
CC	175	-	245	-
CD	195	195	213	184
CE	350	350	233	238
DD	215	-	181	-
DE	370	370	201	216
EE	525	-	220	-
Mean	215	215	215	215
SE	116	101	16	21

¹ wr = with replacement
wor = without
- = that particular combination of farms is impossible to select under the sampling scheme.

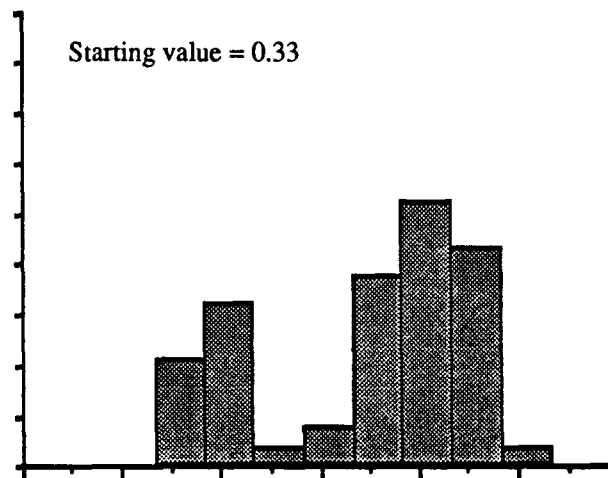
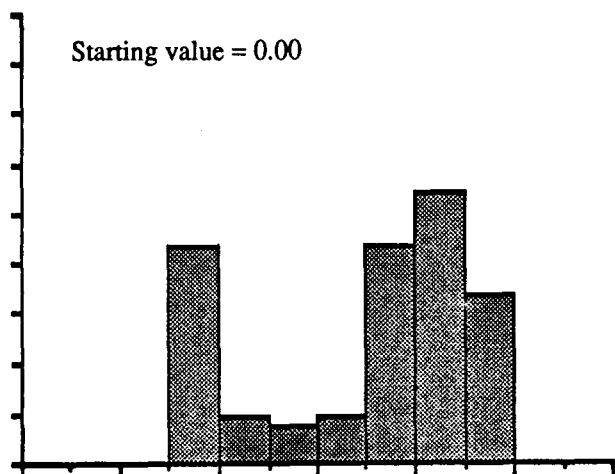


Figure 1-Histograms with same bin widths but different starting values

The hypothesis testing question requires more information than the estimation question: you must be able to supply a definition for how “large” is a large change. The definition of “large” cannot be answered by the statistician or by the data collected. But frequently it is difficult, if not impossible, to supply a definition either because it just is not known or there is extreme controversy as to what constitutes a large change.

When the definition of “large” is unknown, then usually confidence intervals (an estimation procedure) are constructed. But you must remember this about confidence intervals: The confidence percentage (usually 95 percent) is associated with the procedure and not any particular interval you might get. The confidence interval procedure guarantees that, in the long run, the procedure will result in an interval that covers the “true” parameter being estimated 95 percent of the time. There is not a 95 percent chance of your specific interval containing the true value.

P-Values

The P-value is the probability of obtaining a statistic at least as extreme as the observed statistic given that the null hypothesis is true. For example, if someone else has twice your budget for sampling, that someone will have smaller P-values even though there is no difference in the phenomenon that you are investigating. The P-value depends on the population’s variability, the study’s sample size, and the “biological size” of what’s being studied.

P-values are one of the most misused numbers in statistical analysis. A P-value is many times incorrectly used to imply the importance of a hypothesis, and it cannot do so. A P-value (by itself) does not indicate importance, lack of importance, likelihood of the alternative hypothesis being true, or whether you should publish your results.

Display of Data

Displaying your data is of obvious importance to show what your data suggests. One of the common displays, the lowly histogram that you have all had to construct at one time or another, has had several improvements lately.

First, the usual histogram is described. Each sample point is stacked in the bin it belongs to with the bins described by a bin width and a starting value. Figure 1 shows two histograms with the same bin width but different starting values. Would you draw the same conclusions from these two different representations of the same data?

Figure 2 shows two histograms now with the same starting values but different bin widths. Which bin width allows an adequate description of the data?

In constructing the histogram, we took “bricks” that represented the sample points and stacked them into the associated bin. Now consider two modifications: First, instead of placing the brick in the bin that contains the sample point, we center the brick directly on top of the sample point. Where the bricks overlap we break the bricks to fit flush with the horizontal axis (fig. 3).

Second, we change the shape of the brick from a rectangular shape to a smoother shape. These shapes are now called “kernels” and their widths are called band widths rather than bin widths. Naturally, we now call the method the kernel method.

Figure 4 shows two kernel estimates with different bandwidths.

There are several methods for choosing the bandwidth for the kernel method. One commonly used method is to choose the bandwidth that is optimal for the normal distribution:

$$\text{bandwidth} = 1.06 s n^{-1/5}$$

where s is the sample standard deviation and n is the sample size. If we stick with the usual histogram, the optimal bin width for the normal distribution is

$$\text{bin width} = 3.49 s n^{-1/3}$$

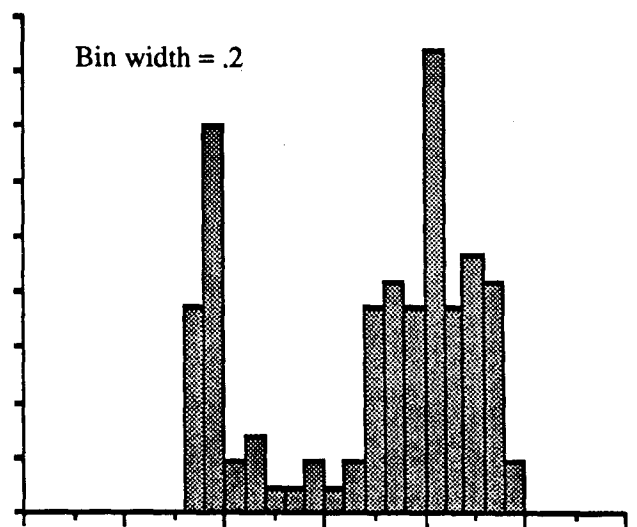
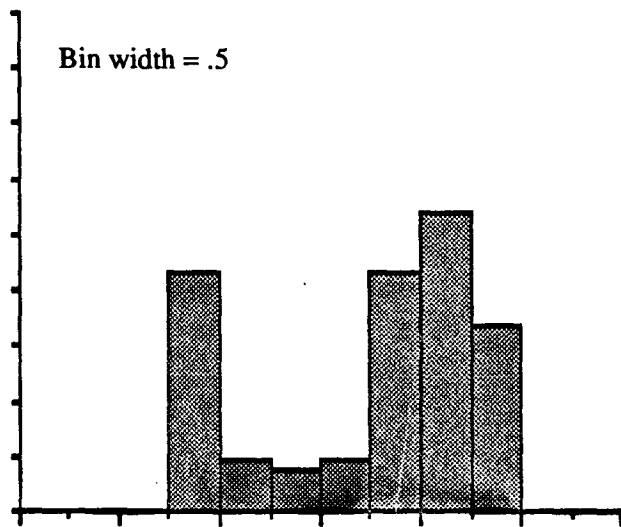
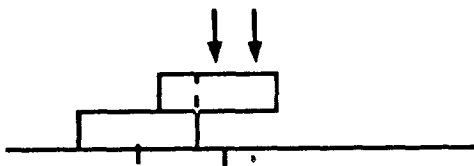


Figure 2-Histograms with same starting values but different bin widths

“Bricks” are centered over each data point



and then broken where there's overlap.

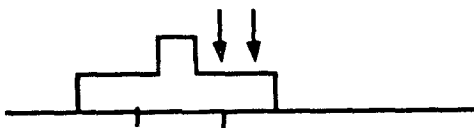


Figure 3-Constructing a “new” histogram with “bricks” centered over each data point

Conclusions

Statisticians can offer a wide variety of assistance for your studies throughout the planning, implementation, analysis, and writing stages. Please try to take advantage of their services.

References

- Cochran, W.G. 1977. Sampling techniques, 3rd ed. New York, NY: John Wiley & Sons; 428 p.
- Silverman, B.W. 1986. Density estimation for statistics and data analysis. London: Chapman and Hall; 175 p.
- Whorton, B.J. 1989. Kernel methods for estimating the utilization distribution in home range studies. Ecology 70 (1): 164-168.

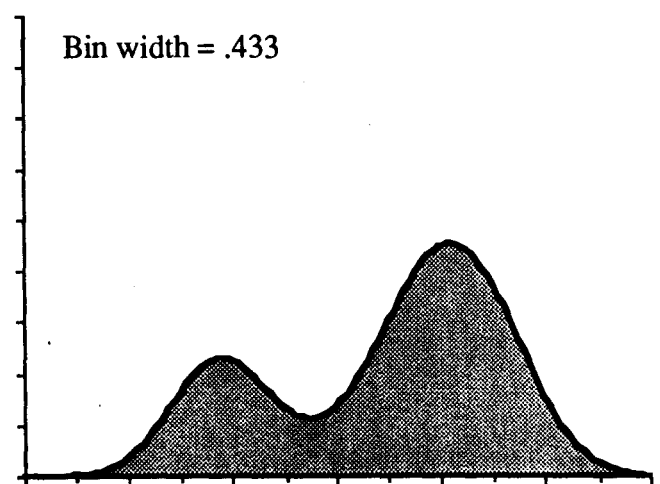
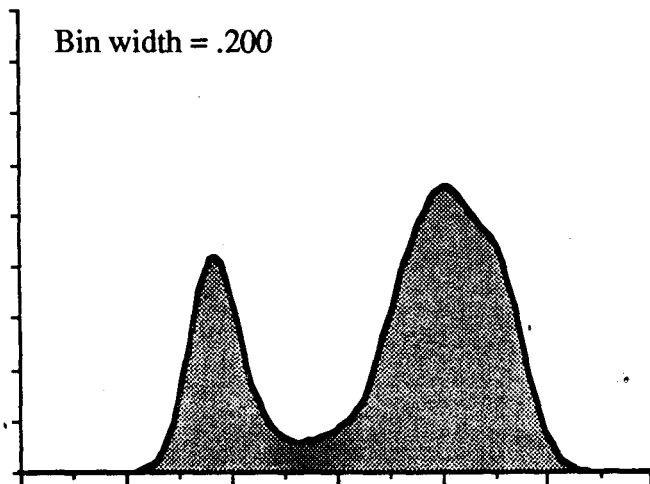


Figure 4-Display of data using the Kernel method with two different bandwidths