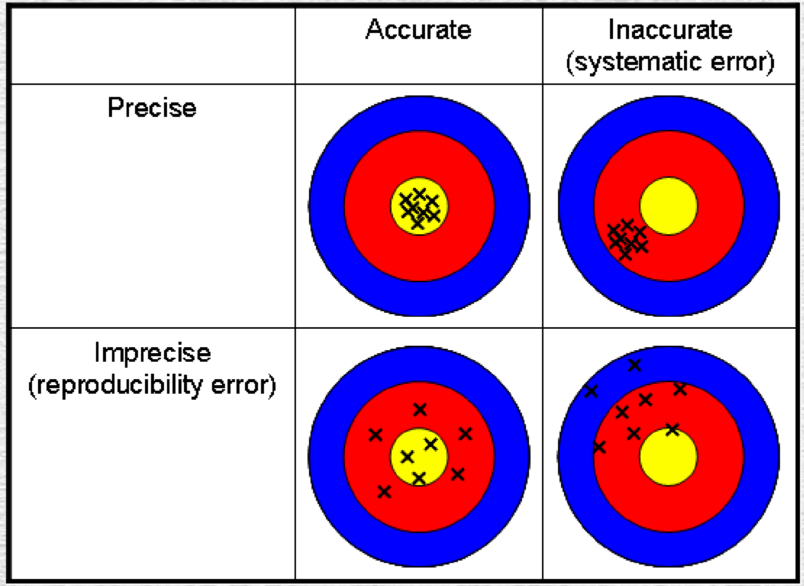
Sampling designs

Sampling

- Complete information is nearly always either impossible, impractical, or not advisable to obtain
- We must base our monitoring on samples of the quantities we're interested in
  - A sample = a subset of the population of interest
- We are interested in population-level parameters, so we estimate these from our samples
  - The **point estimate** is our estimate of the true value
  - The point estimate should be accompanied by a measure of sampling variation (the standard error), and an **interval estimate** (a confidence interval)

Minimizing bias, maximizing precision

- The two ways estimates can be bad:
  - They can be inaccurate = wrong on average (a.k.a. biased)
  - They can be imprecise = low repeatability, large differences between repeated estimates (big standard error)
- Neither is good, but one is not worse than the other
- The sampling design affects both



Sampling design

- Refers to the method by which a sample is selected
- There are many, but the best ones are a type of probability sample = one in which the probability of inclusion in the sample is known for each sampling unit
  - If the probability is the same for every unit, then we are using Simple Random Sampling (SRS)
  - If the probability the same for every unit within identified groups (strata), but different between the groups, we are using Stratified Random Sampling (StRS)
- Probability samples have good properties
  - Unbiased estimates of parameters
  - Possible to know the sampling error from a single sample
- Compare these good properties to a bad alternative, convenience sampling

Samples of convenience

- Collecting data that is easy to get
  - Not probability sampling!
  - Probability of inclusion is not known (but is presumably high for convenient locations, close to 0 for locations that are not convenient)
- May be very precise! Locations that are easy to reach may be homogeneous
- The problem is, areas that are easy to access may not be representative of an entire area → bias
  - More motorized vehicle traffic
  - More horses
  - More hikers
  - Topographically non-random
- There is no way to know from just the sample that is collected how un-representative the sample is

Roads, trails

Simple random sampling

- The simplest, most commonly used probability sampling design
- Each sampling unit has an equal chance of being selected
- Unbiased estimates = the average of all possible estimates is the population parameter
- It's the assumed sampling design for our common estimators of mean, variance, and standard error
- Example: estimating the density of chamise plants in the SDR watershed

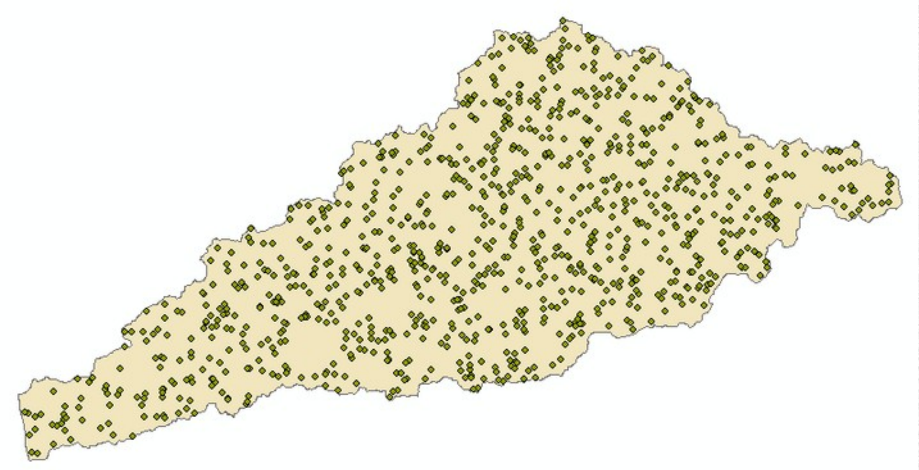
Estimators for SRS – the old, familiar formulas!

Mean: $\bar{x} = \frac{\sum x_i}{n}$

Variance: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Standard deviation: $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Standard error of the mean: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$



This will be an estimate per ha for the entire watershed

Confidence intervals

- Once you have an estimate of the mean, you know it's likely to be wrong
- Given how much sampling variation you expect, what interval is likely to contain the mean?

  Confidence interval: $\bar{x} \pm t_{\alpha,\nu} s_{\bar{x}}$

- Specify the **confidence level**, i.e. 95%
- This specifies the **alpha level**, i.e. 5%, so $\alpha = 0.05$
- Sample size for both is $n = 100$
- Lower-case Greek nu (v) is degrees of freedom
  - For SRS, df = n-1
  - For StRS, df = n-h

SRS 95% CI calculation: density of Chamise per ha

$\bar{x} = 5{,}000$

$s = 500$

$s_{\bar{x}} = \frac{500}{\sqrt{100}} = 50$

$t_{0.05,99} = 1.98$

Confidence interval: $\bar{x} \pm t_{\alpha,\nu} s_{\bar{x}}$

Lower: $5{,}000 - 1.98 \times 50 = 4{,}901$

Upper: $5{,}000 + 1.98 \times 50 = 5{,}099$

Problems with SRS

- Doesn’t account for strata = groupings in the data, such as cover types
  - Points fall into cover types in proportion to their areal coverage – may not be the best allocation
  - Rare strata may not receive any sampling at all by chance
  - Some strata may be more variable than others
- Will not always give you the smallest possible standard errors for the sample size used

Is the density of chamise the same in all of these cover types?

Stratified random sampling

- Takes into account qualitative groupings of units
  - Categorical grouping variable defines the “strata”
  - Within strata, sampling is SRS – use the SRS estimators
- Units (plots, individuals) are measured within all strata
  - Get strata statistics (means, s, se)
  - From these, estimate mean and se for the entire region
- Need different estimators for mean and standard error when we want an overall estimate

Estimators for StRS

Mean: $\bar{x} = \sum W_h \bar{x}_h$

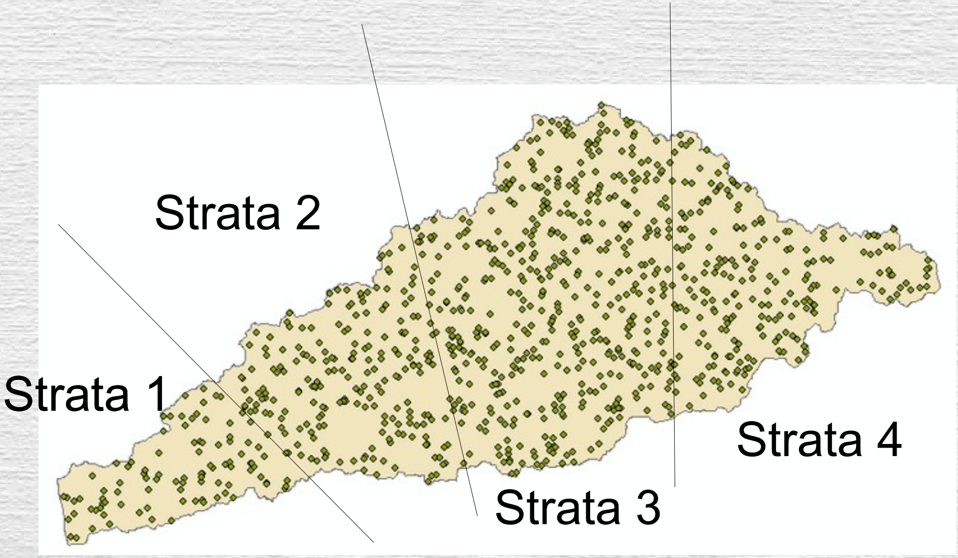
Strata weights: $W_h = \dfrac{n_h}{n}$

Weights = probability of inclusion for a unit in strata h

If samples are allocated proportionate to size of strata, this is the proportion of the area that is strata h

Variance of the mean: $s_{\bar{x}}^2 = \sum W_h^2 \dfrac{\sigma_h^2}{n_h}$

Standard error of the mean: $s_{\bar{x}} = \sqrt{s_{\bar{x}}^2}$



Calculation of mean for stratified samples

Strata	Mean	s	n	$s_{\bar{x}}$
Strata 1	4,800	100	20	22.36
Strata 2	5,000	110	30	20.08
Strata 3	5,800	105	30	19.17
Strata 4	4,000	80	20	17.88

Strata	Weights	W $\bar{x}$
Strata 1	0.2	960
Strata 2	0.3	1500
Strata 3	0.3	1740
Strata 4	0.2	800
	Mean:	5000

Within strata estimates

Estimate for entire watershed

StRS 95% CI calculation: density of Chamise per ha

Strata	Mean	s	n	$s_{\bar{x}}$
Strata 1	4,800	100	20	22.36
Strata 2	5,000	110	30	20.08
Strata 3	5,800	105	30	19.17
Strata 4	4,000	80	20	17.88

Strata	Weights	$W \bar{x}$
Strata 1	0.2	960
Strata 2	0.3	1500
Strata 3	0.3	1740
Strata 4	0.2	800
	Mean:	5000

Strata	Weights	$W^2 s^2/n$
Strata 1	0.2	20.0
Strata 2	0.3	36.3
Strata 3	0.3	33.1
Strata 4	0.2	12.8
	$s^2_{\bar{x}}$	102.2
	$s_{\bar{x}}$	10.1

StRS 95% CI calculation: density of Chamise per ha

$\bar{x} = 5{,}000$

$s_{\bar{x}} = 10.1$

$t_{0.05,96} = 1.98$

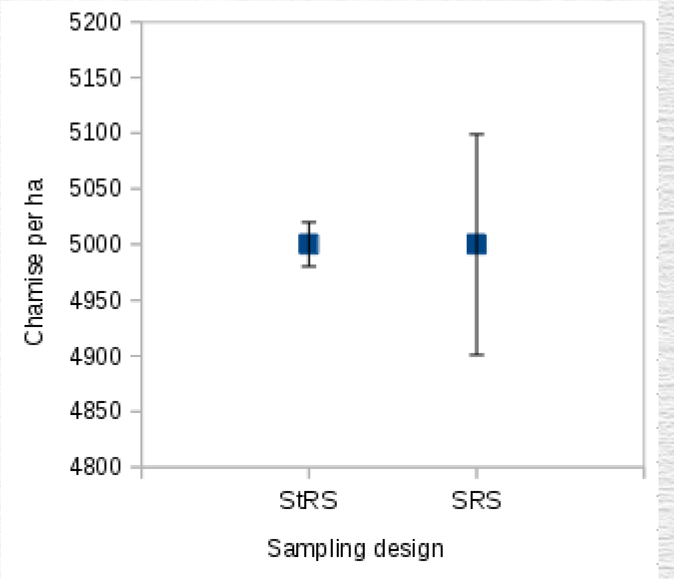
Confidence interval: $\bar{x} \pm t_{\alpha,\nu} s_{\bar{x}}$

Lower: $5{,}000 - 1.98 \times 10.1 = 4{,}980$

Upper: $5{,}000 + 1.98 \times 10.1 = 5{,}020$

95% CI’s for SRS and StRS

- The size of the CI depends on:
  - How variable the data are (s)
  - How much data is collected (n)
  - The sampling design (SRS or StRS)
- StRS is better if:
  - The amount of difference between strata means is big compared to the amount of variation within the strata
  - How big? Big enough to compensate for the lower df



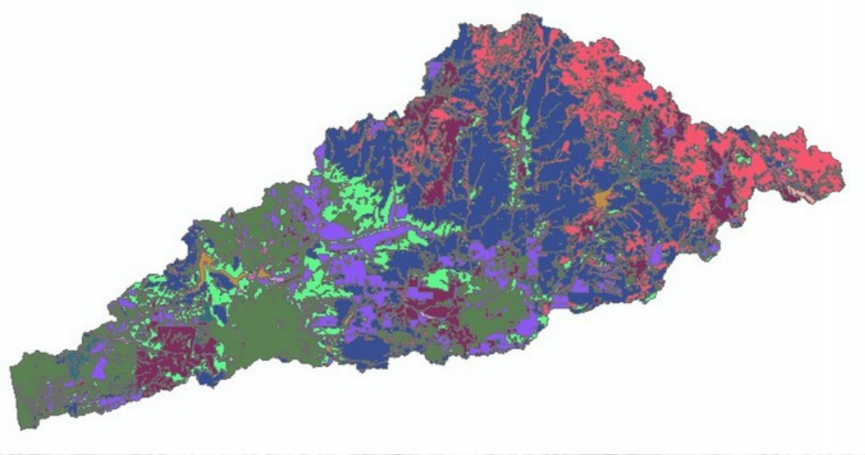
- Note that you only get the benefit of StRS if you use the StRS estimators

Stratified sampling vs. ANOVA

- A stratified sampling design is very similar to an ANOVA experimental design
- But, the purposes are different
  - ANOVA = compare means between groups
  - Stratified sampling = estimate an overall mean, using strata to minimize the standard error
- This difference in purpose can lead to different advice relative to design
  - ANOVA = assumes equal variances among groups, works best with balanced designs (equal n per group)
  - Stratified sampling = does not assume equal variances, more samples should be allocated to the most variable strata to reduce the standard error of the overall estimate

Ways to allocate samples in StRS

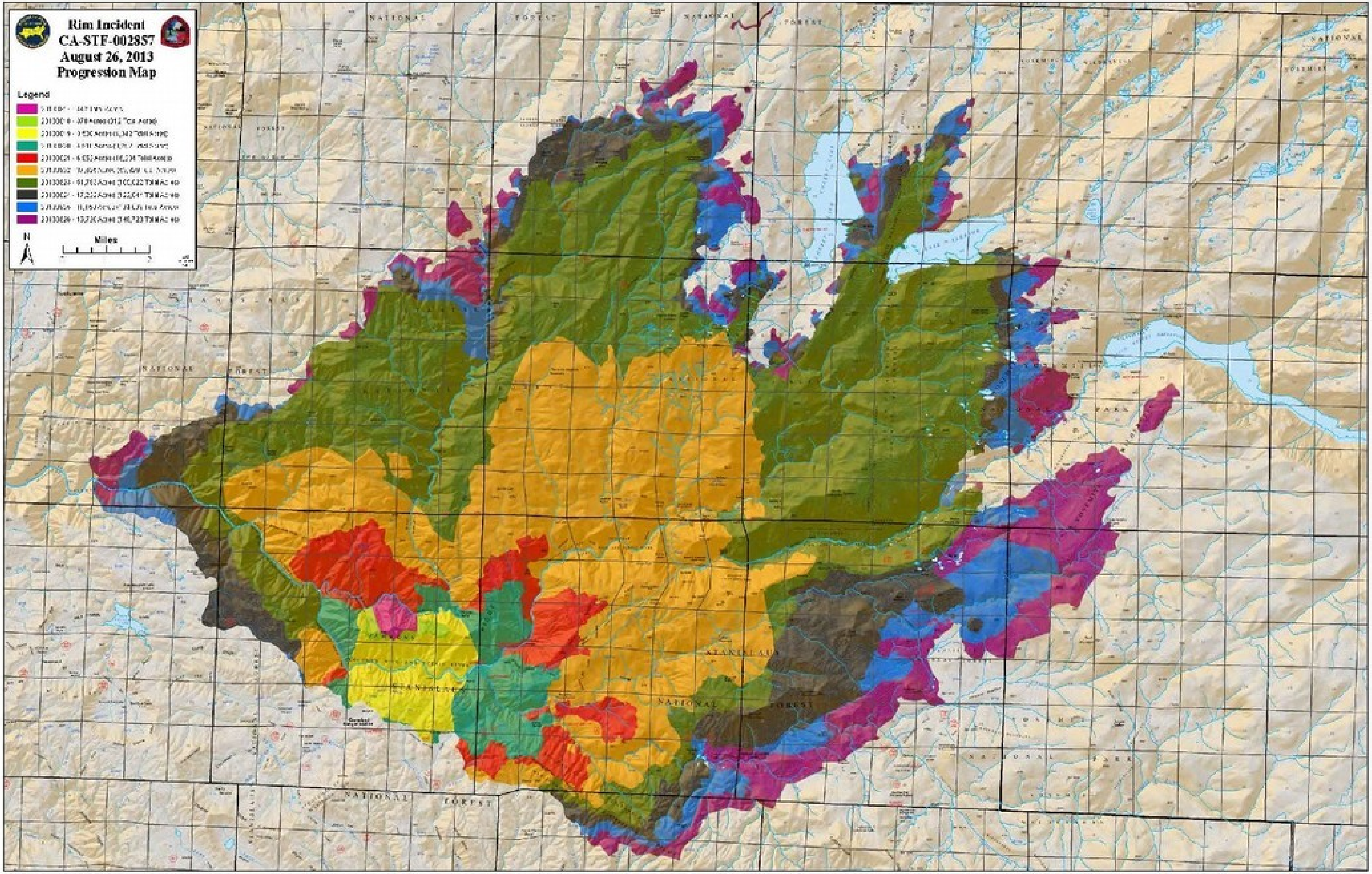
- Stratified sampling does not require a particular allocation of samples to strata
- Some possible approaches:
  - Equal numbers in each strata
  - Numbers proportionate to strata size
  - Numbers proportionate to strata standard deviations



- All will give unbiased estimates
- Allocating proportionate to standard deviation size will give the smallest standard errors

Independence of units

- In sampling, we are often observing data as it is found – no experimental manipulation
- In a designed experiment, we assume responses are independent of one another – lack of independence is a violation of an assumption
  - Estimates of the effects of a treatment will be biased
  - Will often set a minimum distance between samples to ensure independence
  - May try to characterize spatial dependence and extract its effect from our study
- In sampling, if units are not independent that's just a feature of the population we are studying
  - Estimates of the parameter are still unbiased even if the units are dependent



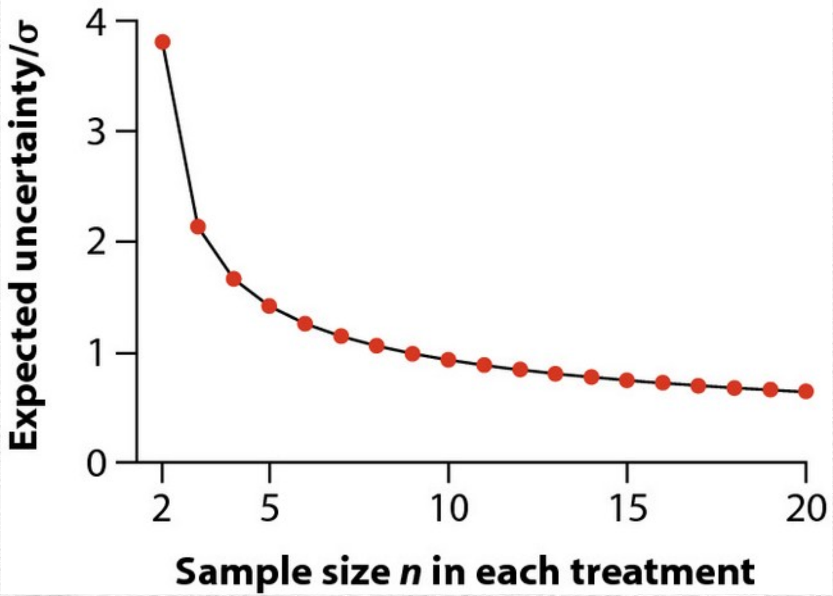
Rim fire map – spread over time

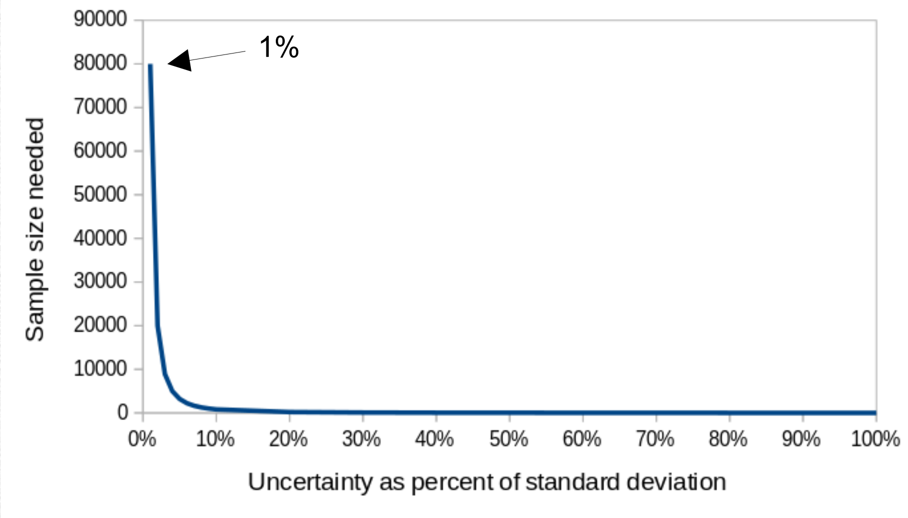
Some additional considerations

- Sample size issues
  - How many total points should be measured?
  - How many points should you measure in each strata?
  - How many visits to each point? Is a single measurement enough, or do you need to account for season, detectability issues, etc.?
- Early detection vs. unbiased estimates
  - If you're trying to detect an invasive exotic, you are more worried about finding it early than about getting unbiased estimates of biomass
  - How does this change things?

Picking a sample size

- We know that...
  - More data is always better
  - More data is more expensive
- Question is: at what point do you have enough data that additional samples are not worth the expense?
- Couple of approaches:
  - Sample size equations
  - Empirical methods





Picking sample size to achieve a desired level of precision

- Uncertainty/σ = uncertainty ($ts_{\bar{x}}$) as number of standard deviations
- Bigger samples lead to less uncertainty
- We can specify the uncertainty we want to achieve, and calculate sample size needed

Using uncertainty to calculate a needed n

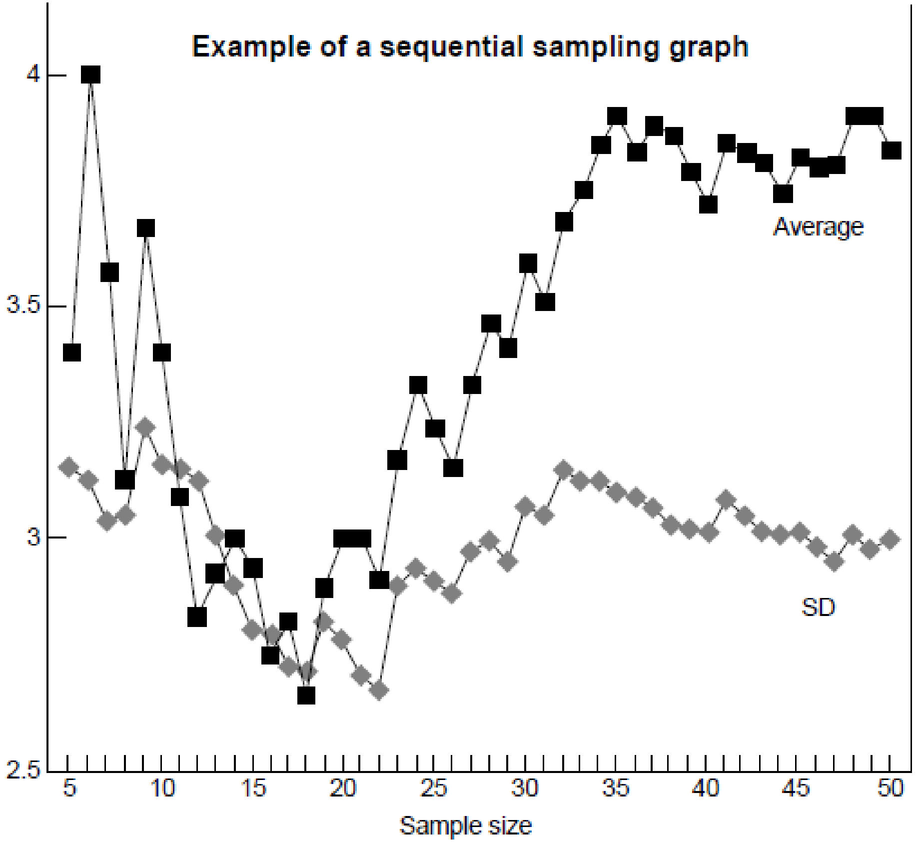
- Specify a desired uncertainty level, then plug into this equation:

$$n = 8\left(\frac{\sigma}{\text{uncertainty}}\right)^2 = 8\left(\frac{0.4}{0.1}\right)^2 = 128$$

- This says that to achieve an uncertainty of 0.1 when the standard deviation is 0.4 we need to collect a sample of n = 128
- Values for σ and uncertainty can come from:
  - A small “pilot study” (preliminary data)
  - A desired ratio - “uncertainty should be no more than 25% of s” - then use 1/0.25 = 4

Empirical methods

- Can do a pilot study
- Collect samples one at a time
- Update the estimate each time a new unit is sampled
- Plot the estimates against sample number
- When the estimate stops changing greatly with each new sample the sample size is adequate



N	Plants	Average	SD
1	1	1.00	0.00
2	2	1.50	0.50
3	0	1.00	0.79
4	9	3.00	3.52
5	5	3.40	3.15
6	7	4.00	3.12
7	1	3.57	3.04
8	0	3.13	3.05
9	8	3.67	3.24
10	1	3.40	3.16
11	0	3.09	3.15
12	0	2.83	3.12
13	4	2.92	3.01
14	4	3.00	2.90
15	2	2.93	2.81
16	0	2.75	2.80
17	4	2.82	2.73
18	0	2.67	2.73
19	7	2.89	2.82
20	5	3.00	2.78
21	3	3.00	2.71
22	1	2.91	2.68
23	9	3.17	2.90
24	7	3.33	2.94
25	1	3.24	2.91
26	1	3.15	2.88
27	8	3.33	2.97
28	7	3.46	3.00
29	2	3.41	2.95
30	9	3.60	3.07
31	1	3.52	3.05
32	9	3.69	3.15
33	6	3.76	3.13
34	7	3.85	3.13
35	6	3.91	3.10
36	1	3.83	3.09
37	6	3.89	3.07
38	3	3.87	3.03
39	1	3.79	3.03
40	1	3.73	3.02
41	9	3.85	3.09
42	3	3.83	3.06
43	3	3.81	3.02
44	1	3.75	3.02
45	7	3.82	3.02
46	3	3.80	2.99
47	4	3.81	2.96
48	9	3.92	3.02
49	4	3.92	2.99
50	0	3.84	3.00

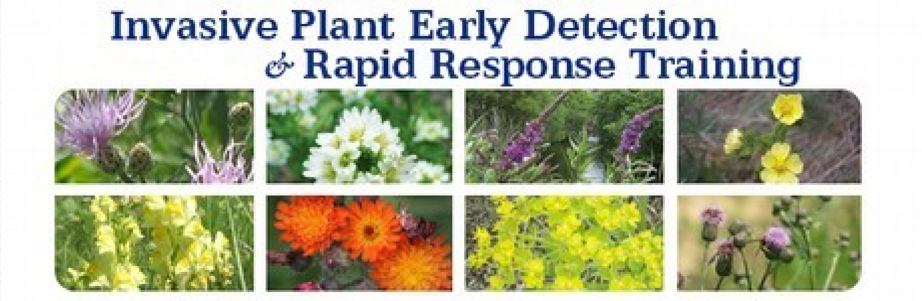
Figure 5. Example of a sequential sampling graph. The running average and standard deviation are plotted for sample sizes of n=5 up to n=50. Sampling was conducted in an area of 50 m x 100 m with a quadrat size of 1 m x 5 m. Actual values are shown on the right.

Sampling for early detection

- Sometimes we are not primarily concerned about estimates of parameters
- Example: perennial pepperweed
  - Invasive plant
  - Has been located in San Diego County, within the SDRP
  - When it's found, it's attacked and removed to avoid spread
- An unbiased estimate of the amount of cover, biomass, etc. is not needed – just need to find it as early as possible and kill it
- Sampling should be **extensive**, but less **intensive** (many sites surveyed, rapid assessment techniques at each site)

Perennial pepperweed

Caulerpa in Agua Hedionda



Rapid assessment

- This can mean sampling in the field
  - Driving roads during periods of high detectability
  - Aerial search
  - Sticky traps for arthropods
- Can mean remote sensing
- Can mean use of “citizen scientists”
  - Volunteers are cheap
  - Information is less reliable than from professionals
  - Consider this a “low resolution” method, subject to high error rates (false positives and false negatives)

Use models to guide early detection

- Invasives don't generally show up at random
- Some sites are more likely to support them
  - Environmental, habitat information
- Some sites are more likely for them to arrive
  - Spread from existing populations outside of the park
  - Higher risk near developments, along roads, trails, waterways