Sampling Concepts

IUFRO-SPDC

Snowbird, UT September 29 – Oct 3, 2014

Drs. Rolfe Leary and John A. Kershaw, Jr.

Sampling Concepts

Simple Sampling Strategies:

- Random Sampling
- Systematic Sampling
- Stratified Sampling

"Blow-up" Estimation

Why Sample?

- $$$$
- Sample may actually provide a better "estimate" than a census
  - Bias
  - Measurement Error

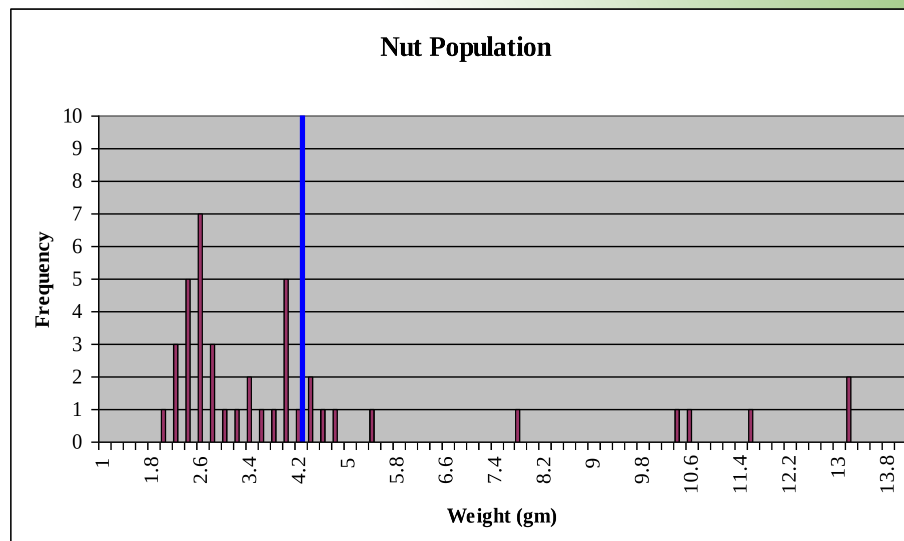
Sampling Implications for Experimental Design

- Selection of Experimental Units
- Assignment of Treatments to Experimental Units

Population versus Sample

- Population
  - The set of individuals we are interested in quantifying
- Sample
  - The set of individuals, selected according to some rules of probability, that we use to represent the population

Today’s Population



Population and Sample Parameters:
The Mean

- Population Mean

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Population Mean

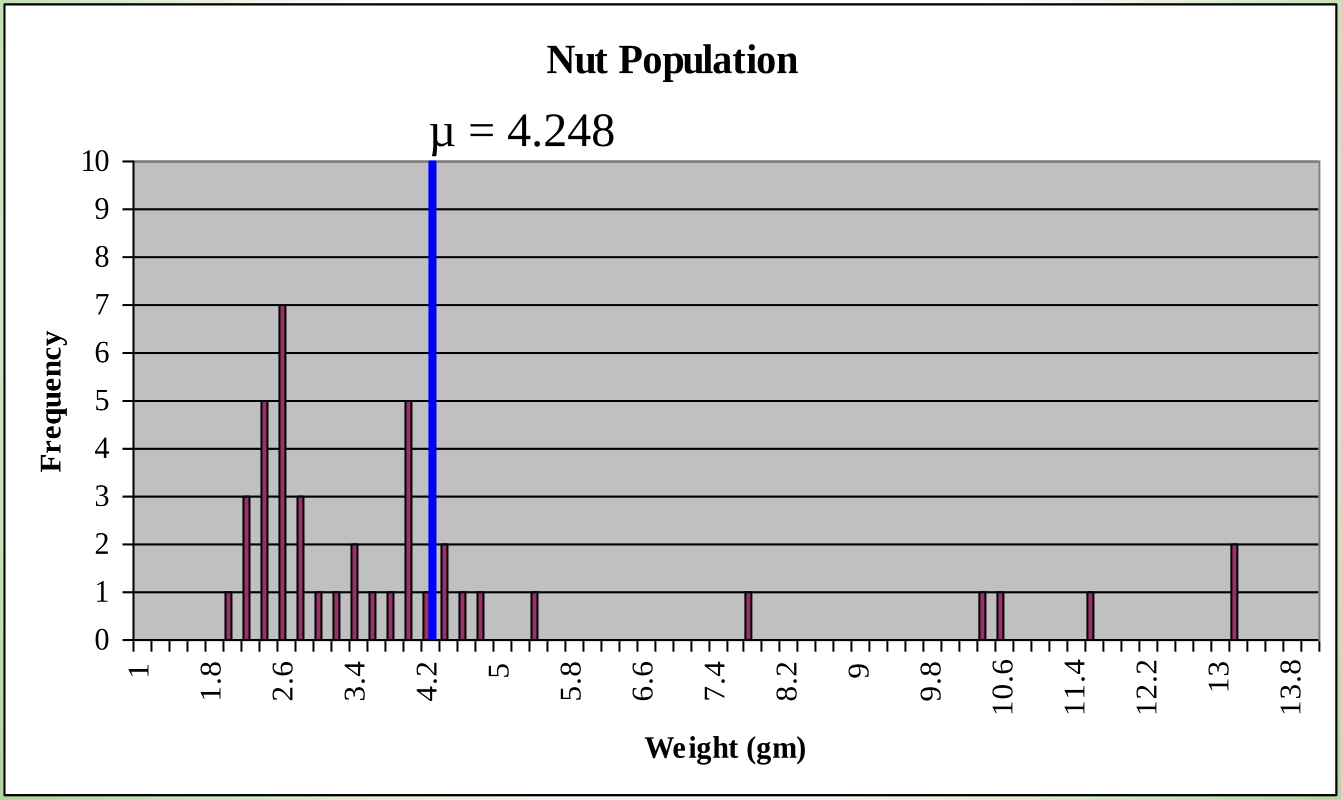
$$\mu = \frac{\sum_{i=1}^{42} X_i}{42}$$

$$= \frac{13.20 + 10.31 + 10.49 + \cdots + 2.54 + 2.06 + 2.25}{42}$$

$$= \frac{178.42}{42}$$

$$= 4.248$$

Our Population Distribution



Population and Sample Parameters: Standard Deviation

- Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

- Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{x})^2}{n-1}}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{42}(X_i - \mu)^2}{42}}$$

$$= \sqrt{\frac{\sum_{i=1}^{42}(X_i - 4.248)^2}{42}}$$

$$= \sqrt{\frac{(13.20 - 4.248)^2 + (10.31 - 4.248)^2 + \cdots + (2.06 - 4.248)^2 + (2.25 - 4.248)^2}{42}}$$

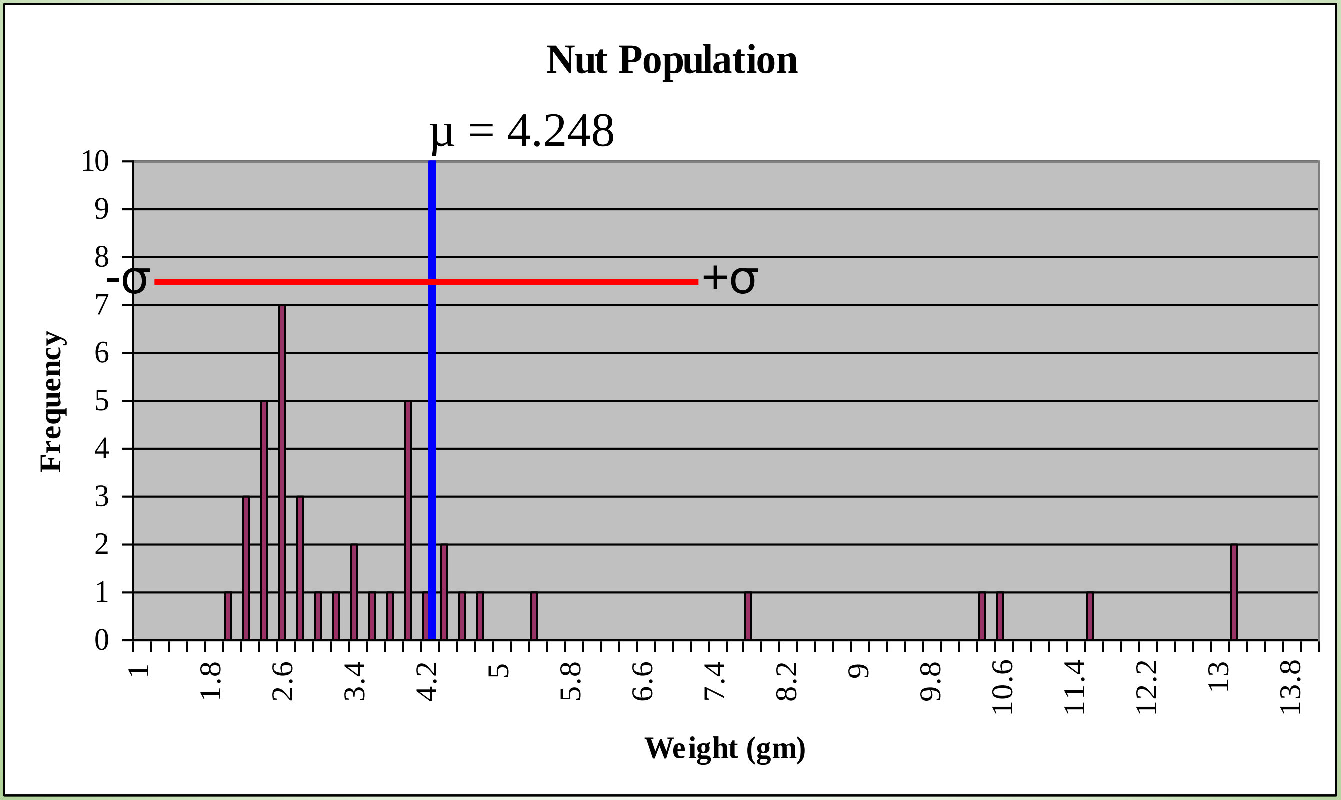
$$= \sqrt{\frac{80.1366 + 54.0505 + \cdots + 4.7878 + 3.9924}{42}}$$

$$= \sqrt{\frac{376.0102}{42}} = \sqrt{8.9526} = 2.9921$$

Mean and Standard Deviation

- Mean measures where the population is “located”
  - The average value
- Standard Deviation measures dispersion of individuals about the mean
  - The average distance individuals are from the mean

Our Population Distribution



Elements of a Sample

- Sample Frame
- Individual Sample Observations (Individuals selected for quantification)
- Error (Individuals not selected)

Sample Frame for Random Sampling

{1}, {2}, {3}
{4}, {5}, {6}
{7}, {8}, {9}
{10}, {11}, {12}
{13}, {14}, {15}
{16}, {17}, {18}
{19}, {20}, {21}

{22}, {23}, {24}
{25}, {26}, {27}
{28}, {29}, {30}
{31}, {32}, {33}
{34}, {35}, {36}
{37}, {38}, {39}
{40}, {41}, {42}

Probability of Selection under Random Sampling

- Each individual represents 1/42nd of the sampling frame (population)
  - Pr(selection) = 1/42 = 0.02381
- Five individuals are being selected
  - Pr(sampled) = 5*(1/42) = 0.11905

Example: sample of size 5

- Select 5 elements from our sample frame
- Randomly sort the sample frame:
  - Generate random numbers for sampling unit
  - Rank and select the first 5

Our Sample

Observation	Nut #	Wt
1	9	4.03
2	41	2.06
3	42	2.25
4	5	13.2
5	6	7.64

Estimation of Population Mean: The Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Note: Formula assumes equal probability of sampling

Our Sample Mean

$$\bar{X} = \frac{\sum_{i=1}^{5} X_i}{5}$$

$$= \frac{(4.03 + 2.06 + 2.25 + 13.2 + 7.64)}{5}$$

$$= \frac{29.18}{5}$$

$$= 5.836$$

Estimation of Population Standard Deviation: Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

Our Sample Standard Deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

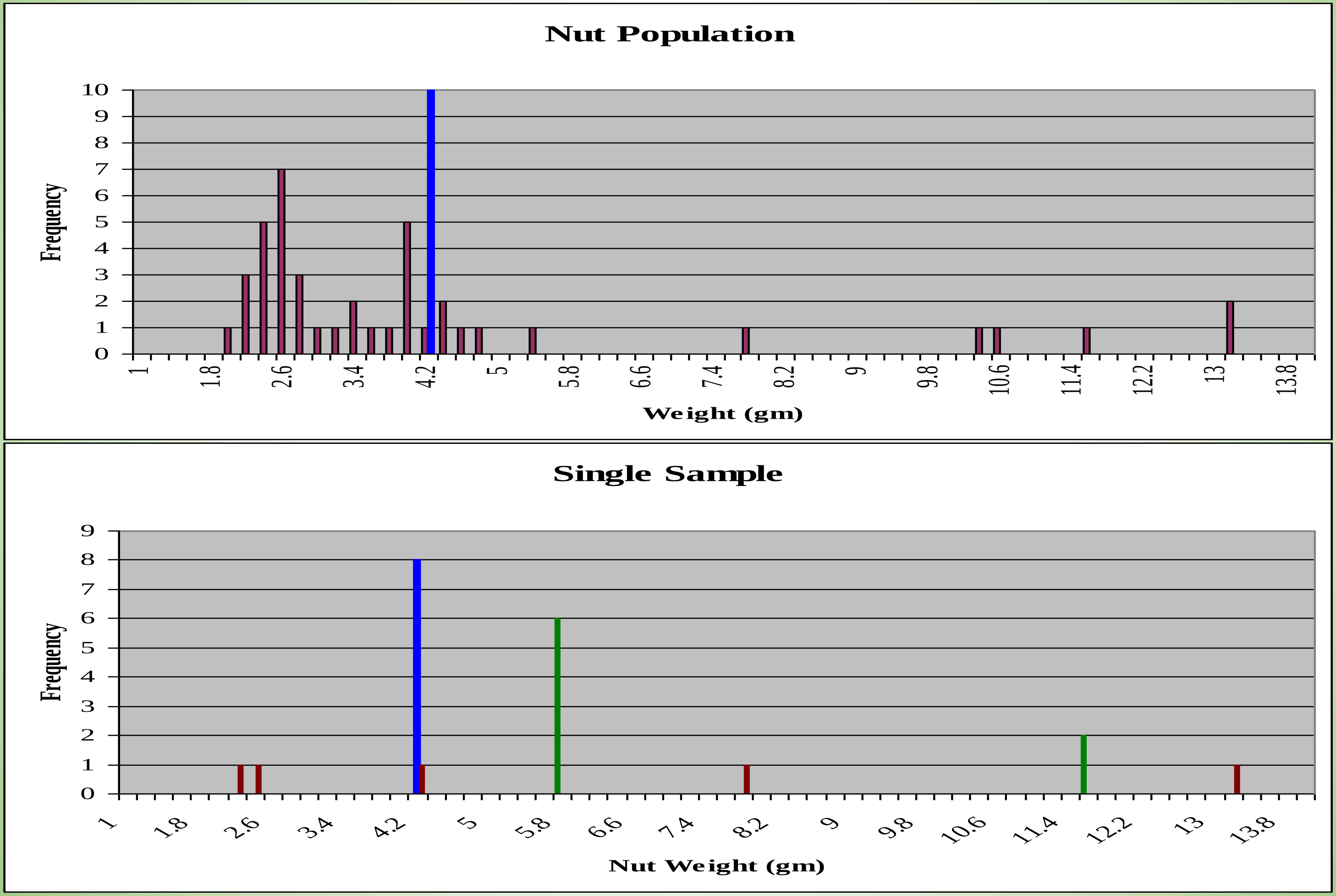
$$= \sqrt{\frac{\left(4.03^2 + 2.06^2 + 2.25^2 + 13.2^2 + 7.64^2\right) - \frac{(4.03 + 2.06 + 2.25 + 13.2 + 7.64)^2}{5}}{5-1}}$$

$$= \sqrt{\frac{258.1566 - \frac{(29.18)^2}{5}}{4}}$$

$$= \sqrt{\frac{258.1566 - 170.2945}{4}}$$

$$= \sqrt{\frac{87.8621}{4}} = \sqrt{21.9655} = 4.6867$$

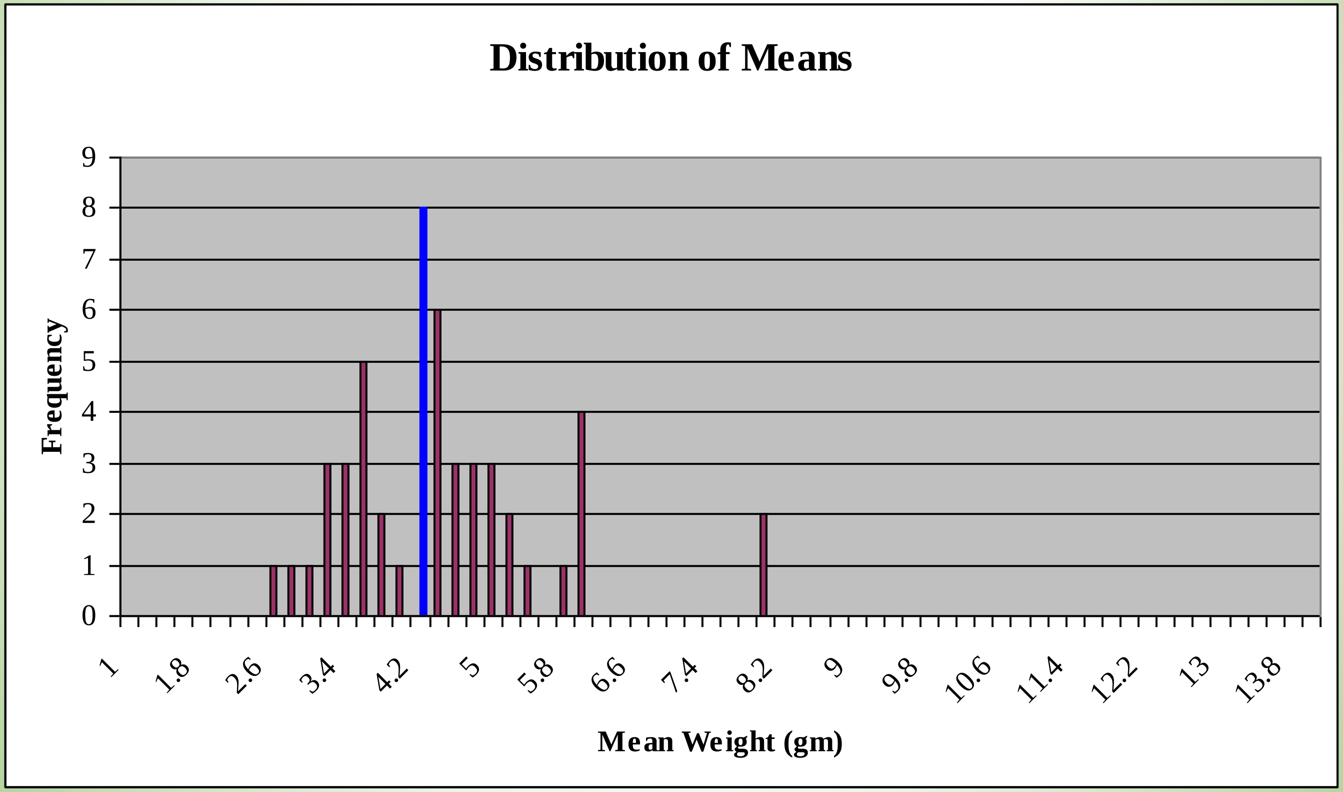
Population and Sample



Sampling Error

- Now we only selected 5 of 42 elements from the population
- Our estimates have error associated with them
- If we sample another 5 elements, we most likely will get different answers
- We need to assess that error

Distribution of Means of size 5



Standard Error

- Distribution of means is known (approximately)
- Estimate sampling error from a single sample:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Standard Error of Estimate (Mean)

Our Sample of Size 5

- Uncorrected

$$s_{\bar{x}} = \frac{4.6867}{\sqrt{5}} = \frac{4.6867}{2.236} = 2.0960$$

- Corrected

$$s_{\bar{x}} = \frac{4.6867}{\sqrt{5}} \cdot \sqrt{\frac{42-5}{42}} = 2.0960 \cdot \sqrt{\frac{37}{42}} = 2.0960 \cdot 0.938 = 1.967$$

Confidence Interval

$$C.I. = \bar{x} \pm t \cdot s_{\bar{x}}$$

$$C.I. = \bar{x} \pm t \cdot \left(\frac{s}{\sqrt{n}} \right)$$

Using Student's t-table

Our probability is
1 - Confidence

Degrees of	Two-tailed probability of obtaining a large value								
freedom	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1	1.3764	1.9626	3.0777	6.3137	12.7062	31.821	63.6559	636.5776
2	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9645	9.925	31.5998
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	12.9244
4	0.7407	0.941	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	8.6101
5	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321	6.8685
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	5.9587
7	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	5.4081
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554	5.0414
9	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498	4.7809
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	4.5868
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058	4.4369
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.681	3.0545	4.3178

Our "degrees of freedom" are n-1

Where have we
seen this before?

95% Confidence Interval for our Sample (Uncorrected)

- n = 5, CI = .95
- T((1-.95),5-1) = 2.776
- CI:

$$\begin{aligned} C.I. &= 5.836 \pm t \cdot 2.096 \\ &= 5.836 \pm 2.776 \cdot 2.096 \\ &= 5.836 \pm 5.819 \end{aligned}$$

- CI: $Pr(0.017 \leq \mu \leq 11.655) = .95$

Some Exploration of Random Sampling Results

- Repeated Sampling
- Effects of Sample Size
- How many samples should I measure?

- Let’s Do Some Exploration

Sample Size

$$C.I. = \bar{x} \pm t \cdot \left(\frac{s}{\sqrt{n}} \right)$$

$$E = t \cdot \left(\frac{s}{\sqrt{n}} \right)$$

$$n = \frac{t^2 \cdot s^2}{E^2}$$

Sample Size

- Our sample of size 5 produced a sampling error of 5.819 at 95% confidence
- Error is almost 100% of the mean
- Not a very good sample
- We might want (or our boss might want) a sampling error of 2.5 grams with 95% confidence

Sample Size

- When we look at the sample size calculation we are calculating n (left side of equation)
- But……….t depends on n (right side of equation)
- Must be solved iteratively

$$n = \frac{t_{\alpha,n-1}^2 \cdot s^2}{E^2}$$

Sample Size

- Start with a guess
- n0 = 10
- t(.05,9) = 2.2622
- S = 4.6867
- N1 = 18 $\neq$ n0
- So repeat with n1
- Repeat until n(i) = n(i-1)

$$n_1 = \frac{2.2622^2 \cdot 2776^2}{2.5^2} = 17.98 = 18$$

$$t_{.05,17} = 2.1098$$

$$n_2 = \frac{2.1098^2 \cdot 2776^2}{2.5^2} = 15.64 = 16$$

$$t_{.05,15} = 2.1314$$

$$n_3 = \frac{2.1314^2 \cdot 2776^2}{2.5^2} = 15.97 = 16$$

Elements of a Sample

- Sample Frame
- Individual Sample Observations (Individuals selected for quantification)
- Error (Individuals not selected)

Sample Layout

- Random
  - Individuals selected independent of each other
  - No spatial or temporal pattern
- Systematic
  - Individuals selected relative to one another
  - Once first individual selected, all other sample units determined
  - Spatial or temporal pattern

Random Sampling

- Statistically the most efficient
  - Mean and standard error unbiased with single sample
- Can be logistically very hard to implement
- Hard to develop truly random samples

Systematic Sampling

- samples laid out along a systematic grid
- logistically the easiest
- unbiased mean
- biased standard deviation

Systematic Sample of our population

{1}, {2}, {3}

{4}, {5}, {6}

{7}, {8}, {9}

{10}, {11}, {12}

{13}, {14}, {15}

{16}, {17}, {18}

{19}, {20}, {21}

{22}, {23}, {24}

{25}, {26}, {27}

{28}, {29}, {30}

{31}, {32}, {33}

{34}, {35}, {36}

{37}, {38}, {39}

{40}, {41}, {42}

Questions

- How many independent simple random samples of size 5 are there?
- How many independent systematic samples of size 5 are there?
- For the sample illustrated, how many independent choices were made?

Properties of systematic samples

- Unbiased mean
- One sample selection (once 1 plot is located, all others are fixed)
- Sampling units are not independent
- Need at least two independent selections to calculate unbiased estimate of standard deviation (remember the n-1)
- s, as estimated from systematic sample, tends to be too large

Why is s too large?

- How many random samples of size 5?
- What is range of random samples of size 5?
- How many systematic samples of size 5?
- What is range of systematic samples of size 5?
- So what does systematic sampling do for us?

Systematic Sample of Size 5

- Only 9 samples possible
  - 3.222, 4.416, 3.168, 7.656, 5.476, 2.69, 3.018, 4.726, 5.882
- Clearly not as much variance possible
- "True" standard error would be the standard deviation of these 9 means

Let’s do some exploration

Stratified Sampling

- Our nut population is composed of 3 nut types:
  - Walnuts (6)
  - Filberts (15)
  - Peanuts (21)
- Var(between nut types) >> Var(within nut types)
- Could use stratified sampling to reduce variability and/or sample size requirements

Stratified Sampling with Proportional Allocation

- Samples are allocated to each Stratum proportional to some measure of abundance in population
  - Frequency
  - Area
  - Total Weight

$$n_j = N\left(\frac{A_j}{\sum_{j=1}^{k} A_j}\right)$$

Strata Summaries

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$s(x_j) = \sqrt{\frac{\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2}{n_j - 1}}$$

$$s(\bar{x}_j) = \frac{s(x_j)}{\sqrt{n_j}}$$

Population Summary

$$\overline{X}_{Pop} = \frac{\sum_{j=1}^{k} \left(A_j \bullet \overline{X}_j \right)}{\sum_{j=1}^{k} A_j}$$

$$s\left(\overline{X}_{Pop}\right) = \sqrt{\sum_{j=1}^{k} \left(\frac{A_j}{A_{Total}} \right)^2 s\left(\overline{X}_j\right)^2}$$

Let’s do some exploration

Implications of Sample Design

- “Blow-up” Estimation
- Want to sample 10 nuts from our current population of 42
- Use that sample to estimate total for a new population
  - 50 walnuts
  - 300 filberts
  - 2000 peanuts
  - “real” total weight = 6600 g