SurvMeth/Surv 625: Applied Sampling

Yajuan Si

Research Associate Professor
Survey Research Center
Institute for Social Research
University of Michigan, Ann Arbor

1/8/25

Section 1

Course introduction

SurvMeth/Surv 625: Applied sampling

- *Lohr: The design is by far the most important aspect of any survey: No amount of statistical analysis can compensate for a badly designed survey.*

SurvMeth/Surv 625: Applied sampling

- *Lohr: The design is by far the most important aspect of any survey: No amount of statistical analysis can compensate for a badly designed survey.*
- *Deming: Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability.*

SurvMeth/Surv 625: Applied sampling

- *Lohr: The design is by far the most important aspect of any survey: No amount of statistical analysis can compensate for a badly designed survey.*
- *Deming: Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability.*
- Applied Sampling is an applied statistical methods course concerned almost exclusively with the design of data collection. This course will cover the main techniques used in sampling practice.

SurvMeth/Surv 625: Applied sampling

- *Lohr: The design is by far the most important aspect of any survey: No amount of statistical analysis can compensate for a badly designed survey.*
- *Deming: Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability.*
- Applied Sampling is an applied statistical methods course concerned almost exclusively with the design of data collection. This course will cover the main techniques used in sampling practice.
- The course syllabus:
  Syllabus_SurvMeth625_Winter2025_YajuanSi.pdf

Design

- Research design
  - Experiments: Control of confounders or randomization of intervention
  - Quasi-experimental: Observational
  - Survey samples: Observational

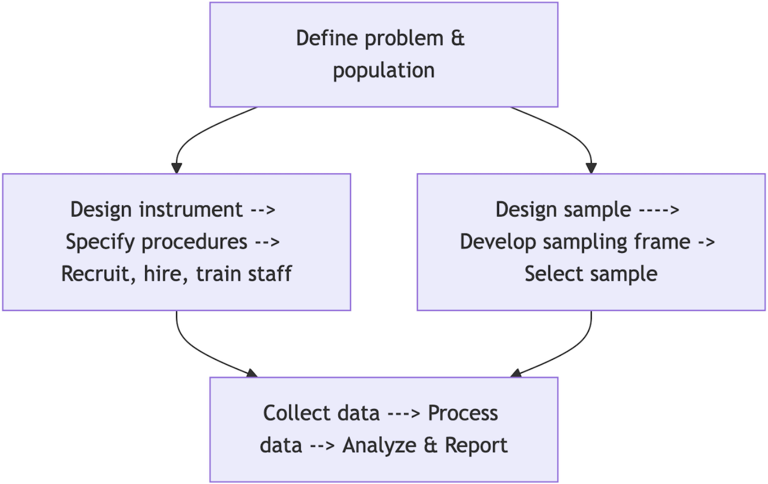
Design

- Research design
  - Experiments: Control of confounders or randomization of intervention
  - Quasi-experimental: Observational
  - Survey samples: Observational
- Design characteristics
  - Realism
  - Randomization
  - Representation
- Example studies with good or bad designs

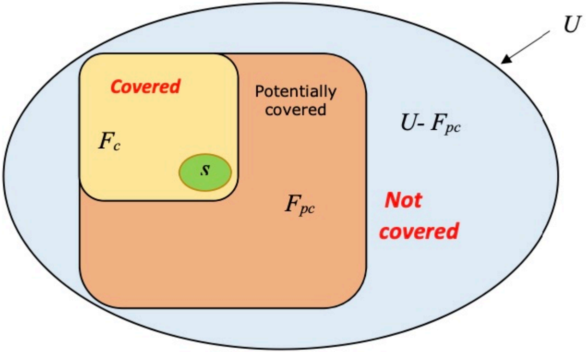
Design vs. analysis

- Sampling techniques (simple random sample, stratified, cluster sampling. etc.)
- Complex design: Multistage, area sampling
- Sampling error: sampling variance estimation
- Nonsampling errors
- Sampling **implementation, inference, and projection**

Sample surveys



Finite population: Valliant (2020)



- U = target population
- F_{pc} = potentially covered; F_c = actually covered
- $U - F_{pc}$ = not covered at all
- s = sample

Vacabulary

- Observation unit/element
- Target population
- Census
- Sample
- Sampled population
- Sampling unit
- Sampling frame
- Coverage: under/over-coverage
- Response: nonresponse

Vacabulary

- Observation unit/element
- Target population
- Census
- Sample
- Sampled population
- Sampling unit
- Sampling frame
- Coverage: under/over-coverage
- Response: nonresponse
- Inclusion: selected and responded
- Non-probability sampling
- Sampling error: resulting from taking one random sample instead of examining the whole population
    - Margin of error
- Nonsampling error: any errors that cannot be attributed to the sample-to-sample variability
    - Selection bias
    - Measurement error

Statistics

- Statistics: population quantity of interest t, e.g., population total, mean, percentiles, regression coefficient, etc.

Statistics

- Statistics: population quantity of interest t, e.g., population total, mean, percentiles, regression coefficient, etc.
- **Sampling distribution**: the distribution of different values of the statistics $\hat{t}$ obtained by the process of taking all possible samples from the population, i.e., repeated sampling

Statistics

- Statistics: population quantity of interest t, e.g., population total, mean, percentiles, regression coefficient, etc.
- **Sampling distribution**: the distribution of different values of the statistics $\hat{t}$ obtained by the process of taking all possible samples from the population, i.e., repeated sampling
- (Design-based) Randomization theory: *the population data are fixed, and the sampling inclusion indicators are random variables.*

Statistics

- Statistics: population quantity of interest t, e.g., population total, mean, percentiles, regression coefficient, etc.
- **Sampling distribution**: the distribution of different values of the statistics $\hat{t}$ obtained by the process of taking all possible samples from the population, i.e., repeated sampling
- (Design-based) Randomization theory: *the population data are fixed, and the sampling inclusion indicators are random variables.*
  - The probabilities of selection for units in the samples give the sampling distribution of the statistics $\hat{t}$

Statistics

- Statistics: population quantity of interest t, e.g., population total, mean, percentiles, regression coefficient, etc.
- **Sampling distribution**: the distribution of different values of the statistics $\hat{t}$ obtained by the process of taking all possible samples from the population, i.e., repeated sampling
- (Design-based) Randomization theory: *the population data are fixed, and the sampling inclusion indicators are random variables.*
    - The probabilities of selection for units in the samples give the sampling distribution of the statistics $\hat{t}$
- Model-based sampling inference: the population is a set of random variables following some probability distribution, and the actual sample values are realizations of these random variables. *The sample data are fixed, and the population distribution is unknown.*.

Probability: Review

1. A random variable Y takes on different values in different samples

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y
3. The expectation of Y is defined as $E(Y) = \sum_y y * P(Y = y)$

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y
3. The expectation of Y is defined as $E(Y) = \sum_y y * P(Y = y)$
4. The variance of Y is: $V(Y) = E[\{Y - E(Y)\}^2] = E(Y^2) - [E(Y)]^2$

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y
3. The expectation of Y is defined as $E(Y) = \sum_y y * P(Y = y)$
4. The variance of Y is: $V(Y) = E[\{Y - E(Y)\}^2] = E(Y^2) - [E(Y)]^2$
5. The covariance of two random variables X and Y:
$Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y
3. The expectation of Y is defined as $E(Y) = \sum_y y * P(Y = y)$
4. The variance of Y is: $V(Y) = E[\{Y - E(Y)\}^2] = E(Y^2) - [E(Y)]^2$
5. The covariance of two random variables X and Y: $Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$
6. The correlation of two random variables X and Y: $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$

Probability: Review

1. A random variable Y takes on different values in different samples
2. The probability distribution of Y depicts the set of possible values y and the probability of each value occurring $P(Y = y)$, where the quantity y is called a realization of Y
3. The expectation of Y is defined as $E(Y) = \sum_y y * P(Y = y)$
4. The variance of Y is: $V(Y) = E[\{Y - E(Y)\}^2] = E(Y^2) - [E(Y)]^2$
5. The covariance of two random variables X and Y:
   $Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$
6. The correlation of two random variables X and Y:
   $Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$
7. The coefficient of variation is defined as $CV(Y) = \frac{\sqrt{V(Y)}}{E(Y)}$, for $E(Y) \neq 0$.

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$
9. Conditional probability adjusts the event probability given a related event occurred: $P(Y = y \mid X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$
9. Conditional probability adjusts the event probability given a related event occurred: $P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)}$
10. The conditional expectation of Y given $X = x$ is $E(Y \mid X = x) = \sum_y y * P(Y = y \mid X = x)$

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$
9. Conditional probability adjusts the event probability given a related event occurred: $P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)}$
10. The conditional expectation of Y given $X = x$ is $E(Y \mid X = x) = \sum_y y * P(Y = y \mid X = x)$
11. The conditional variance of Y given $X = x$ is $V(Y \mid X = x) = \sum_y [y - E(Y \mid X = x)]^2 * P(Y = y \mid X = x)$

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$
9. Conditional probability adjusts the event probability given a related event occurred: $P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)}$
10. The conditional expectation of Y given $X = x$ is $E(Y \mid X = x) = \sum_y y * P(Y = y \mid X = x)$
11. The conditional variance of Y given $X = x$ is $V(Y \mid X = x) = \sum_y [y - E(Y \mid X = x)]^2 * P(Y = y \mid X = x)$
12. Successive conditioning: $E(Y) = E[E(Y \mid X)]$

Probability: Review cont.

8. The joint probability of two occurring event: $P(X = x, Y = y)$
9. Conditional probability adjusts the event probability given a related event occurred: $P(Y = y \mid X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$
10. The conditional expectation of Y given $X = x$ is $E(Y \mid X = x) = \sum_y y * P(Y = y \mid X = x)$
11. The conditional variance of Y given $X = x$ is $V(Y \mid X = x) = \sum_y [y - E(Y \mid X = x)]^2 * P(Y = y \mid X = x)$
12. Successive conditioning: $E(Y) = E[E(Y \mid X)]$
13. Total variability $V(Y) = V[E(Y \mid X)] + E[V(Y \mid X)] =$ between.ave + ave.within

Notation: Population

- A population of N elements: $Y_1, \ldots, Y_N$

Notation: Population

- A population of N elements: $Y_1, \ldots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$

Notation: Population

- A population of N elements: $Y_1, \ldots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$

Notation: Population

- A population of N elements: $Y_1, \dots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \tag{1}$$

Notation: Population

- A population of N elements: $Y_1, \dots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \quad (1)$$

- Standard deviation: S

Notation: Population

- A population of N elements: $Y_1, \ldots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \tag{1}$$

- Standard deviation: S
- For a binary Y_i

Notation: Population

- A population of N elements: $Y_1, \dots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \tag{1}$$

- Standard deviation: S
- For a binary Y_i
  - Population proportion: $P = t/N$

Notation: Population

- A population of N elements: $Y_1, \ldots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \tag{1}$$

- Standard deviation: S
- For a binary Y_i
  - Population proportion: $P = t/N$
  - Population element variance

$$S^2 = \frac{N}{N-1}P(1-P) \tag{2}$$

Notation: Population

- A population of N elements: $Y_1, \dots, Y_N$
- Population total: $t = Y = Y_1 + Y_2 + \cdots + Y_N$
- Population mean: $\bar{Y} = t/N$
- Population element variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}(\sum_{i=1}^{N} Y_i^2 - t^2/N) \tag{1}$$

- Standard deviation: S
- For a binary Y_i
  - Population proportion: $P = t/N$
  - Population element variance

$$S^2 = \frac{N}{N-1}P(1-P) \tag{2}$$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \dots, y_n$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \ldots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \ldots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$
- Sample mean: $\bar{y} = t/n$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \dots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$
- Sample mean: $\bar{y} = t/n$
- Sample element variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \doteq \hat{S}^2$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \ldots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$
- Sample mean: $\bar{y} = t/n$
- Sample element variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \doteq \hat{S}^2$
- Sample fraction: $f = n/N$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \ldots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$
- Sample mean: $\bar{y} = t/n$
- Sample element variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \doteq \hat{S}^2$
- Sample fraction: $f = n/N$
- Finite population correction factor: $fpc = 1 - f$

Notation: Sample

- **One** sample of n elements: $y_1, y_2, \ldots, y_n$
- Sample total: $t = y_1 + y_2 + \cdots + y_n$
- Sample mean: $\bar{y} = t/n$
- Sample element variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \doteq \hat{S}^2$
- Sample fraction: $f = n/N$
- Finite population correction factor: $fpc = 1 - f$
- These are the calculated/observed summary statistics based on the one sample dataset

Sampling inference

- Denote the selection probability for one sample s as $P(s)$ and obtain the sample estimate $\hat{t}_s$

Sampling inference

- Denote the selection probability for one sample s as $P(s)$ and obtain the sample estimate $\hat{t}_s$
- Expectation: the mean of the sampling distribution of $\hat{t}$: $E(\hat{t}) = \sum_{all\ possible\ samples\ s} \hat{t}_s P(s)$.

Sampling inference

- Denote the selection probability for one sample s as $P(s)$ and obtain the sample estimate $\hat{t}_s$
- Expectation: the mean of the sampling distribution of $\hat{t}$: $E(\hat{t}) = \sum_{all\ possible\ samples\ s} \hat{t}_s P(s)$.
- Variance:$V(\hat{t}) = E[(\hat{t} - E(\hat{t}))^2] = \sum_s (\hat{t}_s - E(\hat{t}))^2 P(s)$

Sampling inference

- Denote the selection probability for one sample s as $P(s)$ and obtain the sample estimate $\hat{t}_s$
- Expectation: the mean of the sampling distribution of $\hat{t}$: $E(\hat{t}) = \sum_{all\ possible\ samples\ s} \hat{t}_s P(s)$.
- Variance:$V(\hat{t}) = E[(\hat{t} - E(\hat{t}))^2] = \sum_s (\hat{t}_s - E(\hat{t}))^2 P(s)$
- Standard error: $SE(\hat{t}) = \sqrt{V(\hat{t})}$, i.e., the standard deviation of the sampling distribution of $\hat{t}$

Sampling inference

- Denote the selection probability for one sample s as $P(s)$ and obtain the sample estimate $\hat{t}_s$
- Expectation: the mean of the sampling distribution of $\hat{t}$: $E(\hat{t}) = \sum_{all\ possible\ samples\ s} \hat{t}_s P(s)$.
- Variance:$V(\hat{t}) = E[(\hat{t} - E(\hat{t}))^2] = \sum_s (\hat{t}_s - E(\hat{t}))^2 P(s)$
- Standard error: $SE(\hat{t}) = \sqrt{V(\hat{t})}$, i.e., the standard deviation of the sampling distribution of $\hat{t}$
- Confidence interval: $CI(s) = [low_s, up_s]$, if we repeatedly take samples from the population, construct a 95% CI for each possible sample, we expect 95% of the resulting intervals to include the true value, i.e., a 95% chance that the sample containing the true value

Quality measure

- Bias: the difference between the expectation and the true value $Bias(\hat{t}) = E(\hat{t}) - t$.

Quality measure

- Bias: the difference between the expectation and the true value
  $Bias(\hat{t}) = E(\hat{t}) - t$.
- Mean squared error:
  $MSE(\hat{t}) = E[(\hat{t} - t)^2] = E[(\hat{t} - E(\hat{t}) + E(\hat{t}) - t)^2] = V(\hat{t}) + Bias^2(\hat{t})$

Quality measure

- Bias: the difference between the expectation and the true value $Bias(\hat{t}) = E(\hat{t}) - t$.
- Mean squared error: $MSE(\hat{t}) = E[(\hat{t} - t)^2] = E[(\hat{t} - E(\hat{t}) + E(\hat{t}) - t)^2] = V(\hat{t}) + Bias^2(\hat{t})$
- An estimator $\hat{t}$ of t is **unbiased** if $E(\hat{t}) = t$, **precise** if $V(\hat{t})$ is small, and **accurate** if $MSE(\hat{t})$ is small and the CI coverage probability is close to the nominal level.

Quality measure

- Bias: the difference between the expectation and the true value $Bias(\hat{t}) = E(\hat{t}) - t$.
- Mean squared error: $MSE(\hat{t}) = E[(\hat{t}-t)^2] = E[(\hat{t}-E(\hat{t})+E(\hat{t})-t)^2] = V(\hat{t})+Bias^2(\hat{t})$
- An estimator $\hat{t}$ of t is **unbiased** if $E(\hat{t}) = t$, **precise** if $V(\hat{t})$ is small, and **accurate** if $MSE(\hat{t})$ is small and the CI coverage probability is close to the nominal level.
- Statistical inference validity assessment calculates bias, variance, MSE and CI.

Lohr Example 2.6

```
library(survey)
library(sampling)
library(SDAResources)
library(tidyverse)
data(agpop)  ## Load the data set agpop
N <- nrow(agpop)
N  ## 3078 observations
```

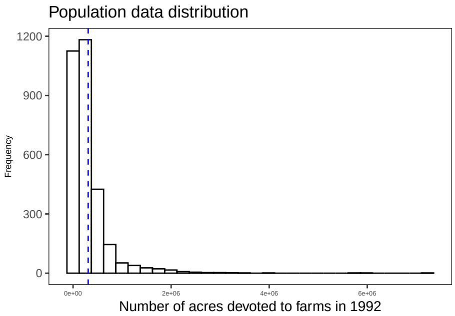
```
[1] 3078
```

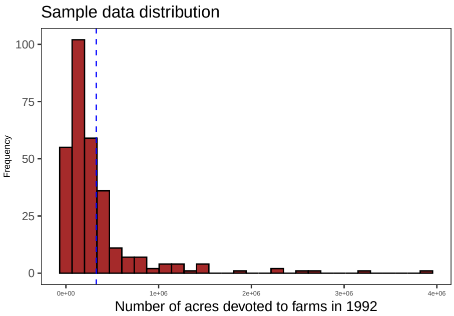
```
## Select an SRS of size n=300 from agpop
set.seed(8126834)
index <- srswor(300,N)
## each unit k is associated with index 1 or 0, with 1 indicating selectio
index[1:10]
```

```
 [1] 0 0 0 1 0 0 0 0 0 0
```

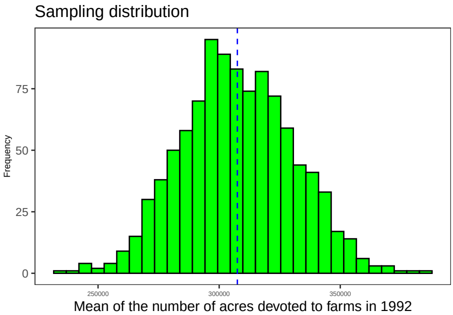
```
## agsrs is an SRS with size 300 selected from agpop
## extract the sampled units from the data frame containing the population
agsrs<- getdata(agpop,index)
```

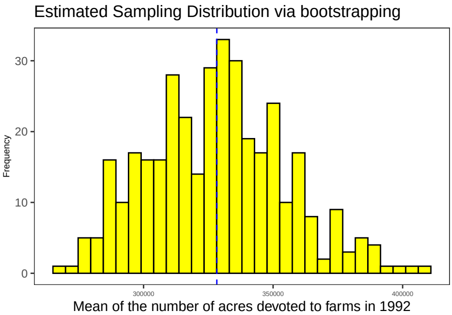
Population distribution





Sampling distribution





Central limit theorem!!!

References

Valliant, Richard. 2020. "Comparing Alternatives for Estimation from Nonprobability Samples." *Journal of Survey Statistics and Methodology* 8 (2): 231–63.