Finite Population Sampling

Fernando TUSELL

May 2012

Outline

Sampling of independent observations

- We have been assuming samples

$$X_1, X_2, \ldots, X_n$$

made of independent observations.

Sampling of independent observations

- We have been assuming samples

$$X_1, X_2, \ldots, X_n$$

made of independent observations.
- This makes sense:

Sampling of independent observations

- We have been assuming samples

$$X_1, X_2, \ldots, X_n$$

  made of independent observations.
- This makes sense:
  - When we sample an infinite population: seeing one value does not affect the probability of seeing the same or another value.

Sampling of independent observations

- We have been assuming samples

$$X_1, X_2, \ldots, X_n$$

made of independent observations.
- This makes sense:
  - When we sample an infinite population: seeing one value does not affect the probability of seeing the same or another value.
  - When we sample with replacement.

Sampling of independent observations

- We have been assuming samples

$$X_1, X_2, \ldots, X_n$$

  made of independent observations.
- This makes sense:
  - When we sample an infinite population: seeing one value does not affect the probability of seeing the same or another value.
  - When we sample with replacement.
- With finite populations without replacement, what we see affects the probability of what is yet to be seen.

Finite versus infinite populations (I)

- With infinite populations, precision depends only on sample size.

Finite versus infinite populations (I)

- With infinite populations, precision depends only on sample size.
- Usually, standard error of estimation is $\frac{\sigma}{n}$ where n is sample size and σ^2 the population variance.

Finite versus infinite populations (I)

- With infinite populations, precision depends only on sample size.
- Usually, standard error of estimation is $\frac{\sigma}{n}$ where n is sample size and σ^2 the population variance.
- If estimator is **consistent** we approach (but never quite hit with certainty) the true value of the parameter.

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

$$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

would verify $m = \hat{m}$ if $n = N$.

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

$$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

  would verify $m = \hat{m}$ if $n = N$.
- All parameters can, in principle, be known with certainty!

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

$$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

would verify $m = \hat{m}$ if $n = N$.
- All parameters can, in principle, be known with certainty!
- With $n \neq N$,

.5cm

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

$$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

  would verify $m = \hat{m}$ if $n = N$.
- All parameters can, in principle, be known with certainty!
- With $n \neq N$,
  - If $n/N \approx 0$, independent sampling good approximation.

.5cm

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

$$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

would verify $m = \hat{m}$ if $n = N$.

- All parameters can, in principle, be known with certainty!
- With $n \neq N$,
  - If $n/N \approx 0$, independent sampling good approximation.
  - If $n/N \gg 0$, we have to take into account that we are looking at a substantial portion of the population.

.5cm

An overview of things to come

We will see:

- What makes sampling without replacement more complex.

An overview of things to come

We will see:

- What makes sampling without replacement more complex.
- What relationship there is among independent and non-independent sampling.

An overview of things to come

We will see:

- What makes sampling without replacement more complex.
- What relationship there is among independent and non-independent sampling.
- What other types of sampling exist.

The central approximation

- Requirement: replacement of "large" population size N.

The central approximation

- Requirement: replacement of "large" population size N.
- If n is "large" and $X_1, \dots, X_n$ "near" independent,

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(m, \sigma^2/n)$$

The central approximation

- Requirement: replacement of "large" population size N.
- If n is "large" and $X_1, \ldots, X_n$ "near" independent,

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(m, \sigma^2/n)$$

- Then,

$$\text{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Estimation of the population total

- Since $T = Nm$, we just have multiply by N the extremes of the interval for m.

Estimation of the population total

- Since $T = Nm$, we just have multiply by N the extremes of the interval for m.
- Hence,

$$\mathrm{Prob}\left(N\overline{X} - Nz_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq T \leq N\overline{X} + Nz_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1-\alpha$$

Estimation of a proportion

- If X_i is a binary variable, $\overline{X}$ is the sample proportion.

Estimation of a proportion

- If X_i is a binary variable, $\overline{X}$ is the sample proportion.
- We have $\overline{X} \sim N(p, pq/n)$

Estimation of a proportion

- If X_i is a binary variable, $\overline{X}$ is the sample proportion.
- We have $\overline{X} \sim N(p, pq/n)$
- Usual estimate of variance is $\hat{p}(1 - \hat{p})/n$.

Estimation of a proportion

- If X_i is a binary variable, $\overline{X}$ is the sample proportion.
- We have $\overline{X} \sim N(p, pq/n)$
- Usual estimate of variance is $\hat{p}(1-\hat{p})/n$.
- Sometimes we use a (conservative) estimate: $pq \leq 0.5$, hence a bound for σ^2 is $0.5/n$.

Sampling error with confidence $1-\alpha$.

- From

$$\mathrm{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1-\alpha$$

we see that we will be off the true value m by less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with probability $1-\alpha$.

Sampling error with confidence $1 - \alpha$.

- From

$$\mathrm{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

  we see that we will be off the true value m by less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with probability $1 - \alpha$.
- This is called the "$1 - \alpha$ (sampling) error".

Sampling error with confidence $1 - \alpha$.

- From

$$\mathrm{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

  we see that we will be off the true value m by less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with probability $1 - \alpha$.
- This is called the "$1 - \alpha$ (sampling) error".
- "Sampling error" also used to mean standard deviation of the estimate.

Finding the required sample size *n*

- **Example:**
  What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?

Finding the required sample size n

- **Example:**
  What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?
- **Solution:**
  Error is less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with confidence $1-\alpha$.

Finding the required sample size n

- **Example:**
  What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?
- **Solution:**
  Error is less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with confidence $1-\alpha$.
- Confidence 0.95 means $z_{\alpha/2} = 1.96$

Finding the required sample size n

- **Example:**
  What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?
- **Solution:**
  Error is less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with confidence $1-\alpha$.
- Confidence 0.95 means $z_{\alpha/2} = 1.96$
- Want $0.03 > 1.96\sqrt{\frac{\sigma^2}{n}}$. Worst case scenario is $\sigma^2 = 0.25$.

Finding the required sample size n

- **Example:**
  What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?
- **Solution:**
  Error is less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with confidence $1-\alpha$.
- Confidence 0.95 means $z_{\alpha/2} = 1.96$
- Want $0.03 > 1.96\sqrt{\frac{\sigma^2}{n}}$. Worst case scenario is $\sigma^2 = 0.25$.
- Therefore, $n > \frac{(1.96)^2 \times 0.25}{0.03^2} = 1067.11$ will do. Will take $n = 1068$.

Interesting facts (I)

- Under independent sampling (infite population or sampling with replacement), required sample size depends only on variance and precsion required.

Interesting facts (I)

- Under independent sampling (infite population or sampling with replacement), required sample size depends only on variance and precsion required.
- Questions like: "Is a sample of 4% enough?" are badly posed.

Interesting facts (I)

- Under independent sampling (infite population or sampling with replacement), required sample size depends only on variance and precsion required.
- Questions like: "Is a sample of 4% enough?" are badly posed.
- $n = 4\%$ of a population with $N = 10000$ insufficient to give a precision of 0.03 with confidence 0.95.

Interesting facts (I)

- Under independent sampling (infite population or sampling with replacement), required sample size depends only on variance and precsion required.
- Questions like: "Is a sample of 4% enough?" are badly posed.
- $n = 4\%$ of a population with $N = 10000$ insufficient to give a precision of 0.03 with confidence 0.95.
- ... but $n = 0.3\%$ of a population with $N = 1000000$ will be more than enough!

Interesting facts (II)

- As long as populations are large detail is expensive!

Interesting facts (II)

- As long as populations are large detail is expensive!
- To estimate a proportion in the CAPV with the precision stated requires about $n = 1068$.

Interesting facts (II)

- As long as populations are large detail is expensive!
- To estimate a proportion in the CAPV with the precision stated requires about $n = 1068$.
- To estimate the same proportion for each of the three Territories with the same precision, requires three times as large a sample!

Interesting facts (II)

- As long as populations are large detail is expensive!
- To estimate a proportion in the CAPV with the precision stated requires about $n = 1068$.
- To estimate the same proportion for each of the three Territories with the same precision, requires three times as large a sample!
- Subpopulation estimates have much lower precision than those for the whole population.

Estimation of the mean (I)

- In independent sampling,

$$
\begin{aligned}
E[\overline{x}] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{m + m + \ldots + m}{n} = \frac{nm}{n} = m
\end{aligned}
$$

Estimation of the mean (I)

- In independent sampling,

$$
\begin{aligned}
E[\overline{x}] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{m + m + \ldots + m}{n} = \frac{nm}{n} = m
\end{aligned}
$$

- $E[X_i] = m$ irrespective of what other values are in the sample.

Estimation of the mean (I)

- In independent sampling,

$$\begin{aligned} E[\overline{x}] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\ &= \frac{m + m + \ldots + m}{n} = \frac{nm}{n} = m \end{aligned}$$

- $E[X_i] = m$ irrespective of what other values are in the sample.
- Without replacement, distribution of X_i depends on what other values are already present in the sample.

Estimation of the mean (I)

- In independent sampling,

$$
\begin{aligned}
E[\overline{x}] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{m + m + \ldots + m}{n} = \frac{nm}{n} = m
\end{aligned}
$$

- $E[X_i] = m$ irrespective of what other values are in the sample.
- Without replacement, distribution of X_i depends on what other values are already present in the sample.
- The same result as for independent sampling is true!

Estimation of the mean (II)

- **Theorem 1**
  In a finite population of size N with $m = \sum_{i=1}^{N} y_i / N$, for samples $Y_1, \ldots, Y_n$ without replacement of size $n < N$ we have:
  $$E[\overline{Y}] = m$$

Estimation of the mean (II)

- **Theorem 1**
  In a finite population of size N with $m = \sum_{i=1}^{N} y_i/N$, for samples $Y_1, \ldots, Y_n$ without replacement of size $n < N$ we have:
  $$E[\overline{Y}] = m$$
- **Proof**

Estimation of the mean (II)

- **Theorem 1**
  In a finite population of size N with $m = \sum_{i=1}^{N} y_i/N$, for samples $Y_1, \ldots, Y_n$ without replacement of size $n < N$ we have:
  $$E[\overline{Y}] = m$$
- **Proof**
  - $Y_1, Y_2, \ldots, Y_n$ are the elements of the sample.

Estimation of the mean (II)

- **Theorem 1**
  In a finite population of size N with $m = \sum_{i=1}^{N} y_i / N$, for samples $Y_1, \ldots, Y_n$ without replacement of size $n < N$ we have:
  $$E[\overline{Y}] = m$$
- **Proof**
  - $Y_1, Y_2, \ldots, Y_n$ are the elements of the sample.
  - $y_1, y_2, \ldots, y_N$ are the elements of the population.

Estimation of the mean (III)

- There are $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ different samples.

Estimation of the mean (III)

- There are $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ different samples.
- Of those, $\binom{N-1}{n-1}$ contain each of the values $y_1, y_2, \ldots, y_N$.

Estimation of the mean (III)

- There are $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ different samples.
- Of those, $\binom{N-1}{n-1}$ contain each of the values $y_1, y_2, \ldots, y_N$.
- Clearly,

$$\sum (Y_1 + Y_2 + \ldots + Y_n) = \binom{N-1}{n-1}(y_1 + y_2 + \ldots + y_N)$$

where the sum in the left is taken over all $\binom{N}{n}$ different samples. Dividing by $\binom{N}{n}$ finishes the proof.

Estimation of the mean (IV)

- Indeed,

$$
\begin{aligned}
\frac{\sum(Y_1 + Y_2 + \ldots + Y_n)}{\binom{N}{n}} &= \frac{\binom{N-1}{n-1}(y_1 + y_2 + \ldots + y_N)}{\binom{N}{n}} \\
&= \frac{n}{N}(y_1 + y_2 + \ldots + y_N)
\end{aligned}
$$

Estimation of the mean (IV)

- Indeed,

$$\begin{aligned}\frac{\sum(Y_1+Y_2+\ldots+Y_n)}{\binom{N}{n}} &= \frac{\binom{N-1}{n-1}(y_1+y_2+\ldots+y_N)}{\binom{N}{n}}\\ &= \frac{n}{N}(y_1+y_2+\ldots+y_N)\end{aligned}$$

- Therefore,

$$E[\overline{Y}] = \frac{\sum(Y_1+\ldots+Y_n)/n}{\binom{N}{n}} = \frac{(y_1+\ldots+y_N)}{N} = E[\overline{y}] = m$$

The indicator variable method

- We have

$$(Y_1 + Y_2 + \ldots + Y_n) = (y_1 Z_1 + y_2 Z_2 + \ldots y_N Z_N)$$

where Z_i is a binary variable which takes value 1 if y_i belongs to a given sample.

The indicator variable method

- We have

$$(Y_1 + Y_2 + \ldots + Y_n) = (y_1 Z_1 + y_2 Z_2 + \ldots y_N Z_N)$$

where Z_i is a binary variable which takes value 1 if y_i belongs to a given sample.

- The probability of that happening is n/N. Then,

$$E[(Y_1 + Y_2 + \ldots + Y_n)] = \frac{n}{N}(y_1 + y_2 + \ldots y_N),$$

which again gives the previous result $E[\overline{Y}] = \overline{y} = m$.

Population variance an quasi-variance

- They are defined as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N} \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N-1}$$

Population variance an quasi-variance

- They are defined as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N} \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N-1}$$

- Similarly for sample analogues:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n} \qquad \tilde{s}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

Population variance an quasi-variance

- They are defined as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N} \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N-1}$$

- Similarly for sample analogues:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n} \qquad \tilde{s}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

- Turns out some formulae are simpler in terms of quasi-variances.

Variance of $\overline{Y}$ (I)

- **Theorem 2**
  In a finite population of size N, the estimator $\overline{Y}$ of $m = \sum_{i=1}^{N} y_i / N$ based on a sample of size $n < N$ without replacement $Y_1, \ldots, Y_n$ has variance:

$$\mathrm{Var}[\overline{Y}] = \frac{\tilde{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)$$

Variance of $\overline{Y}$ (I)

- **Theorem 2**
  In a finite population of size N, the estimator $\overline{Y}$ of $m = \sum_{i=1}^{N} y_i/N$ based on a sample of size $n < N$ without replacement $Y_1, \ldots, Y_n$ has variance:

$$\mathrm{Var}[\overline{Y}] = \frac{\tilde{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)$$

- Factor

$$\left(1 - \frac{n}{N}\right)$$

  usually called "finite population correction factor" or "correction factor".

Variance of $\overline{Y}$ (II)

- **Remarks:**

Variance of $\overline{Y}$ (II)

- **Remarks:**
- It is the same expression as in independent random sampling with i) σ^2 replaced by $\tilde{\sigma}^2$, and ii) corrected with the factor $(1 - n/N)$.

Variance of $\overline{Y}$ (II)

- **Remarks:**
- It is the same expression as in independent random sampling with i) σ^2 replaced by $\tilde{\sigma}^2$, and ii) corrected with the factor $(1 - n/N)$.
- If $n = N$, the variance $\mathrm{Var}(\overline{Y})$ is 0 (why?).

Variance of $\overline{Y}$ (II)

- **Remarks:**
- It is the same expression as in independent random sampling with i) σ^2 replaced by $\tilde{\sigma}^2$, and ii) corrected with the factor $(1 - n/N)$.
- If $n = N$, the variance $\text{Var}(\overline{Y})$ is 0 (why?).
- Formula covers middle ground between infinite populations ($n/N = 0$) and census sampling ($n/N = 1$).

Variance of $\overline{Y}$ (III)

- **Proof**

$$
\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \mathrm{Var}\left(\frac{y_1 Z_i + \ldots + y_N Z_N}{n}\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \mathrm{Var}(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right]
\end{aligned}
$$

Variance of $\overline{Y}$ (III)

- **Proof**

$$
\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \mathrm{Var}\left(\frac{y_1 Z_i + \ldots + y_N Z_N}{n}\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \mathrm{Var}(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right]
\end{aligned}
$$

- We only need expressions for $\mathrm{Var}(Z_i)$ and $\mathrm{Cov}(Z_i, Z_j)$.

Variance of $\overline{Y}$ (IV)

- Since Z_i is binary with probability n/N,

$$\mathrm{Var}(Z_i) = (n/N)(1 - n/N).$$

Variance of $\overline{Y}$ (IV)

- Since Z_i is binary with probability n/N,

$$\mathrm{Var}(Z_i) = (n/N)(1 - n/N).$$

- But $\mathrm{E}[Z_i Z_j] = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}$, so

$$\mathrm{Cov}(Z_i, Z_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n(1-n/N)}{N(N-1)}$$

Variance of $\overline{Y}$ (IV)

- Since Z_i is binary with probability n/N,

$$\mathrm{Var}(Z_i) = (n/N)(1 - n/N).$$

- But $\mathrm{E}[Z_i Z_j] = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}$, so

$$\mathrm{Cov}(Z_i, Z_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n(1 - n/N)}{N(N-1)}$$

- Replacing in expression for $\mathrm{Var}(\overline{Y})$ will lead to result.

Variance of $\overline{Y}$ (V)

$$
\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \underbrace{\mathrm{Var}(Z_i)}_{(n/N)(1-n/N)} + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \underbrace{\mathrm{Cov}(Z_i, Z_j)}_{-\frac{n(1-n/N)}{N(N-1)}}\right] \\
&= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j\right]
\end{aligned}
$$

- Will rewrite expression in brackets.

Variance of $\overline{Y}$ (VI)

- Remark that,

$$
\begin{aligned}
\sum_{i=1}^{N}(y_i - m)^2 &= \sum_{i=1}^{N} y_i^2 - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N} \\
&= \frac{N-1}{N}\left[\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}\frac{y_i y_j}{N-1}\right]
\end{aligned}
$$

Variance of $\overline{Y}$ (VI)

- Remark that,

$$\begin{aligned}\sum_{i=1}^{N}(y_i - m)^2 &= \sum_{i=1}^{N} y_i^2 - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N} \\ &= \frac{N-1}{N}\left[\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}\frac{y_i y_j}{N-1}\right]\end{aligned}$$

- The expression in square brackets in th r.h.s is therefore $\frac{N}{N-1}\sum_{i=1}^{N}(y_i - m)^2$.

Variance of $\overline{Y}$ (VII)

- We are now done!

$$\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\underbrace{\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j\right]}_{\frac{N}{N-1}\sum_{i=1}^{N}(y_i-m)^2} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{\sum_{i=1}^{N}(y_i-m)^2}{N-1} \\
&= \left(1-\frac{n}{N}\right)\frac{\tilde{\sigma}^2}{n}
\end{aligned}$$

Sample size for given precision (I)

- The $(1-\alpha)$ error is

$$\delta = z_{\alpha/2}\sqrt{\frac{\tilde{\sigma}^2}{n}(1-n/N)}$$

Sample size for given precision (I)

- The $(1-\alpha)$ error is

$$\delta = z_{\alpha/2}\sqrt{\frac{\tilde{\sigma}^2}{n}(1 - n/N)}$$

- Solving for n we obtain

$$n = \frac{N z_{\alpha/2}^2 \tilde{\sigma}^2}{N\delta^2 + \tilde{\sigma}^2 z_{\alpha/2}^2}$$

Sample size for given precision (I)

- The $(1-\alpha)$ error is

$$\delta = z_{\alpha/2}\sqrt{\frac{\tilde{\sigma}^2}{n}(1-n/N)}$$

- Solving for n we obtain

$$n = \frac{Nz^2_{\alpha/2}\tilde{\sigma}^2}{N\delta^2 + \tilde{\sigma}^2 z^2_{\alpha/2}}$$

- In terms of the variance, it can be written as:

$$n = \frac{Nz^2_{\alpha/2}\sigma^2}{(N-1)\delta^2 + \sigma^2 z^2_{\alpha/2}}$$

Sample size for given precision (II)

- $\tilde{\sigma}^2$ or σ^2 are required.

Sample size for given precision (II)

- $\tilde{\sigma}^2$ or σ^2 are required.
- We either replace an upper bound or conservative estimation for σ^2.

Sample size for given precision (II)

- $\tilde{\sigma}^2$ or σ^2 are required.
- We either replace an upper bound or conservative estimation for σ^2.
- Failing that, we estimate σ^2 or $\tilde{\sigma}^2$.

Why strata?

- Sometimes we know something about the composition of the population, knowledge that can be put to use.

Why strata?

- Sometimes we know something about the composition of the population, knowledge that can be put to use.
- **Example:** We might know that males and females have different spending in e.g. tobacco or cosmetics.

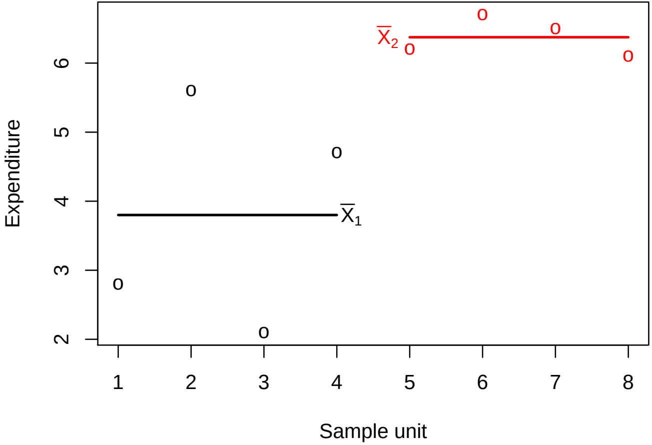
Why strata?

- Sometimes we know something about the composition of the population, knowledge that can be put to use.
- **Example:** We might know that males and females have different spending in e.g. tobacco or cosmetics.
- To estimate average spending, it makes sense to sample males and females, and combine the estimations.

Why strata?

- Sometimes we know something about the composition of the population, knowledge that can be put to use.
- **Example:** We might know that males and females have different spending in e.g. tobacco or cosmetics.
- To estimate average spending, it makes sense to sample males and females, and combine the estimations.
- Sometimes, the target quantity might be similar, but the variance quite different. Also makes sense to differentiate.

Example 1



- Makes sense to estimate mean in each subpopulation

Definitions and notation

- We assume the population is divided in h strata. Total size is $N = N_1 + N_2 + \ldots + N_h$.

Definitions and notation

- We assume the population is divided in h strata. Total size is $N = N_1 + N_2 + \ldots + N_h$.
- The i-th stratum has a mean $m_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$ and variance $\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(y_{ij} - m_i)^2$.

Definitions and notation

- We assume the population is divided in h strata. Total size is $N = N_1 + N_2 + \ldots + N_h$.
- The i-th stratum has a mean $m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$ and variance $\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (y_{ij} - m_i)^2$.
- Clearly,

$$
\begin{aligned}
m &= \sum_{i=1}^{h} \left(\frac{N_i}{N} \right) m_i \\
\sigma^2 &= \sum_{i=1}^{h} \frac{N_i}{N} \sigma_i^2 + \sum_{i=1}^{h} \frac{N_i}{N} (m_i - m)^2
\end{aligned}
$$

Estimation of the mean

- The estimation of the mean sampling without replacement the whole population has variance $\frac{\tilde{\sigma}^2}{n}(1 - n/N)$.

Estimation of the mean

- The estimation of the mean sampling without replacement the whole population has variance $\frac{\tilde{\sigma}^2}{n}(1 - n/N)$.
- Similarly, the estimation of the mean of each stratum has variance $\sigma_i^2 = \frac{\tilde{\sigma}_i^2}{n}(1 - n_i/N_i)$.

Estimation of the mean

- The estimation of the mean sampling without replacement the whole population has variance $\frac{\tilde{\sigma}^2}{n}(1 - n/N)$.
- Similarly, the estimation of the mean of each stratum has variance $\sigma_i^2 = \frac{\tilde{\sigma}_i^2}{n}(1 - n_i/N_i)$.
- The variance of the global mean reconstituted from the estimated means of the strata is

$$\sigma_*^2 = \sum_{i=1}^{h} \left(\frac{N_i}{N}\right)^2 \frac{\tilde{\sigma}_i^2}{n_i}(1 - n_i/N_i)$$

Does the estimation of m improve?

- Yes. If we sample each stratum in proportion to its size (i.e., $n_i/N_i = n/N$ for all i), then:

$$\frac{\tilde{\sigma}^2}{n}(1 - n/N) - \sigma_*^2 =$$

$$\left(1 - \frac{n}{N}\right) \sum_{i=1}^{h} \left(\frac{N_i}{N}\right) \left[\frac{N_i - 1}{N - 1} - \frac{N_i}{N}\right] \frac{\tilde{\sigma}_i^2}{n_i} +$$

$$\left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{h} \frac{N_i}{N - 1} (m_i - m)^2$$

Does the estimation of m improve?

- Yes. If we sample each stratum in proportion to its size (i.e., $n_i/N_i = n/N$ for all i), then:

$$\frac{\tilde{\sigma}^2}{n}(1 - n/N) - \sigma_*^2 = \\ \left(1 - \frac{n}{N}\right) \sum_{i=1}^{h} \left(\frac{N_i}{N}\right) \left[\frac{N_i - 1}{N - 1} - \frac{N_i}{N}\right] \frac{\tilde{\sigma}_i^2}{n_i} + \\ \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{h} \frac{N_i}{N - 1}(m_i - m)^2$$

- Marked Improvement when the m_i's very different.

Abraham Wald on sample selection

Abraham Wald on sample selection

Abraham Wald (1902-1950)

Abraham Wald on sample selection

Abraham Wald (1902-1950)

Abraham Wald on sample selection

Abraham Wald (1902-1950)

- Hungarian-born. Graduated (Ph.D. Mathematics) from University of Vienna, 1931.

Abraham Wald on sample selection

Abraham Wald (1902-1950)

- Hungarian-born. Graduated (Ph.D. Mathematics) from University of Vienna, 1931.
- Fled to the USA in 1938, as Nazi persecution intensified in Austria.

Abraham Wald on sample selection

Abraham Wald (1902-1950)

- Hungarian-born. Graduated (Ph.D. Mathematics) from University of Vienna, 1931.
- Fled to the USA in 1938, as Nazi persecution intensified in Austria.
- Important contributions to the war effort as statistician (notably sequential analysis)

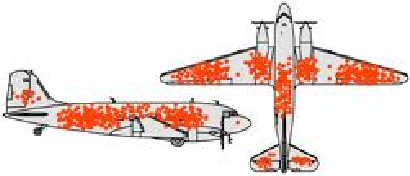
Abraham Wald on sample selection

Abraham Wald (1902-1950)

- Hungarian-born. Graduated (Ph.D. Mathematics) from University of Vienna, 1931.
- Fled to the USA in 1938, as Nazi persecution intensified in Austria.
- Important contributions to the war effort as statistician (notably sequential analysis)
- Was consulted about aircraft armoring.

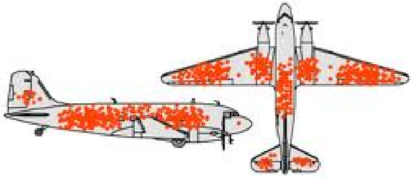
What Wald saw that the others did not

- *Mark hits in B-29 bombers as they come back.*



What Wald saw that the others did not

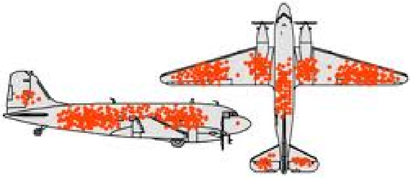
- *Mark hits in B-29 bombers as they come back.*



- Pretty obvious! Will armor the most beaten areas.

What Wald saw that the others did not

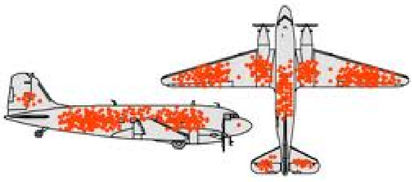
- *Mark hits in B-29 bombers as they come back.*



- Pretty obvious! Will armor the most beaten areas.
- *I didn't tell you to do that!*

What Wald saw that the others did not

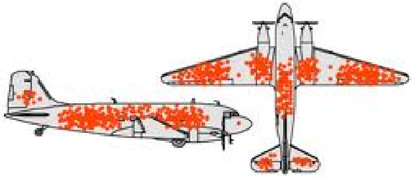
- *Mark hits in B-29 bombers as they come back.*



- Pretty obvious! Will armor the most beaten areas.
- *I didn't tell you to do that!*
- Do you want us to protect the areas with no hits?

What Wald saw that the others did not

- *Mark hits in B-29 bombers as they come back.*



- Pretty obvious! Will armor the most beaten areas.
- *I didn't tell you to do that!*
- Do you want us to protect the areas with no hits?
- *That's exactly what I suggest!*

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!
- Do not let the survey taker to choose the units.

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!
- Do not let the survey taker to choose the units.
- A random sample is not a “grab set”.

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!
- Do not let the survey taker to choose the units.
- A random sample is not a “grab set”.
- Build a census, randomize properly, address the chosen units and no others.

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!
- Do not let the survey taker to choose the units.
- A random sample is not a "grab set".
- Build a census, randomize properly, address the chosen units and no others.
- If you use systematic sampling (every n-th unit with random start), make sure no periodicities exist that will destroy randomness.