This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
Session 3
Sampling Design Alternatives

William A. Reinke, Ph.D.
Professor

Department of International Health
Johns Hopkins University
School of Hygiene and Public Health
Principles To Be Developed

• Sample Statistics Differ from but Are Related to Population Parameters

• Difference Can Be Reduced by Obtaining Larger Sample of Data

• Some Sampling Designs for Obtaining These Data Are
  – More Informative
  – Less Costly
  – More Efficient
# Main Measures of Interest

<table>
<thead>
<tr>
<th>Continuous Variables</th>
<th>Population Parameter</th>
<th>Sample Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Average: Arithmetic Mean</td>
<td>$\mu$</td>
<td>$\bar{X}$</td>
</tr>
<tr>
<td>- Dispersion: Standard Deviation</td>
<td>$\sigma$</td>
<td>$S$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discrete Variables</th>
<th>Population Parameter</th>
<th>Sample Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Relative Frequency: Proportion</td>
<td>$\pi$</td>
<td>$P$</td>
</tr>
</tbody>
</table>
UNIVERSE

PARAMETERS

\[ \mu \]
\[ \sigma \]
\[ \pi \]

SAMPLE

STATISTICS

\[ X \]
\[ S \]
\[ P \]

ESTIMATES

Diagram showing the relationship between parameters from a universe and statistics from a sample, with estimates of the mean (\( \mu \)) and standard deviation (\( \sigma \)).
Hypothetical Sample Results
Three Populations

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>75</td>
<td>73</td>
<td>79</td>
</tr>
<tr>
<td>μ</td>
<td>75</td>
<td>76</td>
<td>79</td>
</tr>
<tr>
<td>η</td>
<td>75</td>
<td>74</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>78</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>74</td>
<td>66</td>
</tr>
</tbody>
</table>

\[ \bar{X} \approx 75 \approx \mu \approx ? \]
Precise Estimates are Possible If -

- There is Little Variation Among Sample Results
- The Sample Size is Sufficiently Large

The Mathematical Relationship is -

\[
\text{Standard Error} = \sqrt{\frac{\text{Variance}}{\text{Sample Size}}}
\]

\[
= \sqrt{\frac{\sigma^2}{n}} \text{ or } \sqrt{\frac{\pi(1-\pi)}{n}}
\]
Daily Attendance (X)

$\sigma_{\bar{X}} = 15$
\[ \sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{15^2}{25}} = 3 \]

Monthly Average of Daily Attendance (X)
$\sigma \bar{x} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{15^2}{36}} = 2.5$

Daily Average of 36 Days (X)
Type II Error of Omission: 10%
Type I Error of Commission: 5%
## Determination of Sample Size

**Simple Random Sample**

<table>
<thead>
<tr>
<th>Purpose of analysis</th>
<th>Sources of Error</th>
<th>Type of Error</th>
<th>General formula for $n$</th>
<th>Special case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate universe mean</td>
<td>1</td>
<td>1</td>
<td>[ \left( \frac{ZS}{D} \right)^2 ]</td>
<td>[ \frac{4S^2}{D^2} ]</td>
</tr>
<tr>
<td>Decide whether Universe Mean Conforms to Defined Standard</td>
<td>1</td>
<td>2</td>
<td>[ \left( \frac{Z_1+Z_2}{S} \right)^2 ]</td>
<td>[ \frac{10.9S^2}{D^2} ]</td>
</tr>
<tr>
<td>Estimate Differences between Two Universe Means</td>
<td>2</td>
<td>1</td>
<td>[ 2 \left( \frac{ZS}{D} \right)^2 ]</td>
<td>[ \frac{8S^2}{D^2} ]</td>
</tr>
<tr>
<td>Decide Whether Real (non-zero) Differences Exists between Two Universe Means</td>
<td>2</td>
<td>2</td>
<td>[ 2 \left( \frac{Z_1+Z_2}{S} \right)^2 ]</td>
<td>[ \frac{21.8S^2}{D^2} ]</td>
</tr>
</tbody>
</table>

**Assumptions:**
- $Z = 2.0$ (95% confidence)
- $Z_1 = 2.0$ (5% Risk Type I Error)
- $Z_2 = 1.3$ (10% Risk Type II Error)
Sample Size

Sampling Error

Error Reduction

Error Increase

10 20 30 40 50 60 70 80 90

Sample Size
Rules of Stratification for Separate Analysis of Population Subgroups

- Select Subgroups as *Homogenous* as Possible
- Equalize Subgroup Sample Sizes as Much as Possible
# Population Situation

<table>
<thead>
<tr>
<th>Subgroup Village</th>
<th>Members per Subgroup (Households)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>400</td>
</tr>
<tr>
<td>B</td>
<td>800</td>
</tr>
<tr>
<td>C</td>
<td>200</td>
</tr>
<tr>
<td>D</td>
<td>500</td>
</tr>
<tr>
<td>E</td>
<td>100</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,000</strong></td>
</tr>
</tbody>
</table>

## Sampling Requirement

- Sample of 20 Households from Each of 3 Villages
- At Start Each of Household Has 60 Chances in 2,000 (p=.03 to Be Selected)
Sampling Requirement

- Sample of 20 Households from Each of 3 Villages
- At Start Each of Household Has 60 Chances in 2,000 (p=.03) to Be Selected

Example

Probability that a Specific Household in Village D is Selected:

\[
3 \times \frac{500}{2,000} \times \frac{20}{500} = \frac{60}{2,000}
\]

<table>
<thead>
<tr>
<th>Village Chosen</th>
<th>Probability Proportional to Size (PPS)</th>
<th>Probability in Selected Village</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within Subgroups Means

\[ \sigma^2_W \]

Between Subgroups Means

\[ \sigma^2_b \]
Rules of Multistage Sampling for Combining Subgroup Information to Obtain Aggregate Estimates

- Select Subgroups as *Heterogeneous* as Possible
- Select Subgroups with Probability Proportional to Size (PPS)
- Obtain Equal Number of Observations per Subgroups