Surveys, Sampling and Observational Data

Derek Li

Contents

1 Review

1.1 Basic Definition

Definition 1.1. ***Random experiment*** is the process of observing the outcome of a chance event.

Definition 1.2. ***Elementary outcomes*** are all possible results of the random experiment.

Definition 1.3. ***Sample space*** (Ω) is the set of all the elementary outcomes.

Definition 1.4. ***Random variable*** Y is a real-valued function defined over a sample space.

Definition 1.5. ***Variable*** is the function defined on population elements, characteristic of population elements. Variable can be quantitative (numerical) or qualitative (categorical).

Definition 1.6. ***Distribution*** or ***frequency distribution*** is the proportion of elements with value in an interval $[a, b], \forall a, b$.

1.2 Basic Notations

- Population: $E = \{e_1, e_2, \cdots, e_N\}$ with population size N, where e_i's are elements.
- Variable: $y, x, z, t, \cdots$.
- Range: $\{y(e), e \in E\}$.
- Probability: In discrete case,

$$P(y_i) = \frac{|\{e, y(e) = y_i\}|}{N} = \frac{N_i}{N}.$$

  In continuous case,

$$P(a, b) = P(a < y < b) = \int_a^b f(y)\mathrm{d}y,$$

  where $f(y)$ is the density function s.t.

$$f(y) \geqslant 0, \forall y \text{ and } \int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1.$$

1.3 Population Parameters

1.3.1 Population Mean (μ_y)

- Using distribution:

$$\mu_y = \sum_{i=1}^{k} y_i P(y_i) = \frac{1}{N}\sum_{i=1}^{k} N_i y_i.$$

- Using population values:

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

1.3.2 Population Variance (σ_y^2)

- Using distribution:

$$\sigma_y^2 = \sum_{i=1}^{k}(y_i - \mu_y)^2 P(y_i) = \frac{1}{N}\sum_{i=1}^{k} N_i(y_i - \mu_y)^2.$$

- Using population values:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \mu_y^2.$$

- Population standard deviation is $\sigma_y = \sqrt{\sigma_y^2}$.

1.3.3 Population Total (τ_y)

$$\tau_y = \sum_{i=1}^{N} y(e_i) = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} N_i y_i = N\mu_y.$$

1.3.4 Population Proportion

Define

$$y(e) = \begin{cases} 0, & e \text{ does not have the property} \\ 1, & e \text{ has the property} \end{cases},$$

then

$$p = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{M}{N} = \mu,$$

where M is the number of elements with the property.

1.3.5 Population Ratio

Ratio of two variables' means or totals:

$$R_{y/x} = \frac{\mu_y}{\mu_x} = \frac{N\mu_y}{N\mu_x} = \frac{\tau_y}{\tau_x}.$$

1.4 Basic Rules from Probability

In probability, the covariance of X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

In statistics, the covariance of x and y is

$$\begin{aligned}\mathrm{Cov}(x, y) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \mu_x\mu_y \\ &= \frac{1}{N}\sum_{i,j} N_{ij}(x_i - \mu_x)(y_j - \mu_y) = \frac{1}{N}\sum_{i,j} N_{ij} x_i y_j - \mu_x\mu_y.\end{aligned}$$

In probability, the correlation of X and Y is

$$\rho_{X,Y} = \rho_{Y,X} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

In statistics, the correlation of x and y is

$$\rho_{x,y} = \rho_{y,x} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}.$$

1.5 Sample

Definition 1.7. ***Sample*** is a subset of the population.

Definition 1.8. ***Random sample*** is a sequence of random variables (independent or dependent)

$$Y_1 = y_1, \cdots, Y_n = y_n,$$

where Y_i is the random variable and y_i is the obtained value.

Definition 1.9. ***Statistic*** is the function of sample values, such sample mean (average), sample variance, etc.

Definition 1.10. ***Sampling distribution*** is the distribution of the sample function. This distribution depends on the population distribution of y and function f.

Theorem 1.1 (Central Limit Theorem)**.** If one takes random sample from a population of mean μ and standard deviation σ, then as n gets large, $\overline{X}$ approaches the normal distribution with mean μ and variance $\frac{\sigma}{\sqrt{n}}$.

1.6 Estimation

Definition 1.11. An ***estimator*** $\widehat{\theta} = \phi(y_1, \cdots, y_n)$ is a sample function used to estimate population parameter θ.

Definition 1.12. An ***estimate*** of θ : $\widehat{\theta}_n = \phi_n(y_1, \cdots, y_n)$ is the value of the sample function $\widehat{\theta}$ obtained from sample of size n.

Example 1.1. Suppose $\eta = \psi(\theta)$, then $\widehat{\eta} = \widehat{\psi}(\theta) = \psi(\widehat{\theta})$.

1.6.1 Variance of Estimator

Suppose $\mathrm{Var}[\widehat{\theta}] = \psi(\theta)$, then $\widehat{\mathrm{Var}}[\widehat{\theta}] = \psi(\widehat{\theta})$.

Example 1.2. In sample random sampling, $\widehat{\mu} = \overline{y}, \widehat{\sigma}^2 = S^2$. We have

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}[\overline{y}] = \frac{\sigma^2}{n},$$

then

$$\widehat{\mathrm{Var}}(\widehat{\mu}) = \frac{S^2}{n}.$$

1.6.2 Properties of Estimators

Definition 1.13. If $\mathbb{E}[\widehat{\theta}] = \theta$, then the estimator is called ***unbiased***.

Definition 1.14. ***Bias*** of $\widehat{\theta}$ is defined as

$$B(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta.$$

Definition 1.15. If $\mathbb{E}[\widehat{\theta}] \to \theta$ as $n \to \infty$, then the estimator is called ***asymptotically unbiased***.

Definition 1.16. If $\widehat{\theta}_1, \widehat{\theta}_2$ are both unbiased estimators of θ and $\mathrm{Var}[\widehat{\theta}_1] < \mathrm{Var}[\widehat{\theta}_2]$, then $\widehat{\theta}_1$ is more ***efficient*** than $\widehat{\theta}_2$.

Note that estimator should be at least asymptotically unbiased and has small variance at least if sample size n is large. Also, an unbiased estimator may not have small variance.

Definition 1.17. The ***mean squared error*** of θ is

$$\mathrm{MSE} = \mathrm{Var}[\widehat{\theta}] + B(\widehat{\theta}) = \mathrm{Var}[\widehat{\theta}] + (\mathbb{E}[\widehat{\theta}] - \theta)^2.$$

Definition 1.18. $\widehat{\theta}$ is consistent for θ if

$$\lim_{n\to\infty} \mathrm{Var}[\widehat{\theta}] = 0.$$

We have some rules for comparing estimators:

- If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators for θ and $\mathrm{Var}[\widehat{\theta}_1] < \mathrm{Var}[\widehat{\theta}_2]$, then we prefer $\widehat{\theta}_1$.
- If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are biased estimators and $\mathrm{MSE}(\widehat{\theta}_1) < \mathrm{MSE}(\widehat{\theta}_2)$, then we prefer $\widehat{\theta}_1$.

1.6.3 Error of Estimation

The error of estimation $|\widehat{\theta} - \theta|$ is not known in practice.

Definition 1.19. The ***error bound*** B_α at level $1 - \alpha$ is a value s.t.

$$P(|\widehat{\theta} - \theta| < B_\alpha) = 1 - \alpha.$$

We say B_α is an error bound with confidence $1 - \alpha$ and the ***confidence interval*** is

$$[\widehat{\theta} - B_\alpha, \widehat{\theta} + B_\alpha].$$

Example 1.3. Suppose $\widehat{\theta} \sin \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}})$, then

$$\begin{aligned} P(|\widehat{\theta} - \theta| < B_\alpha) &= P(-B_\alpha < \widehat{\theta} - \theta < B_\alpha) \\ &= P\left(-\frac{B_\alpha}{\sigma_{\widehat{\theta}}} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < \frac{B_\alpha}{\sigma_{\widehat{\theta}}}\right) = \left(-\frac{B_\alpha}{\sigma_{\widehat{\theta}}} < Z < \frac{B_\alpha}{\sigma_{\widehat{\theta}}}\right) = 1 - \alpha. \end{aligned}$$

Suppose $Z_{\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$ are the critical values of Z, then $P(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}) = 1 - \alpha$.

Therefore,

$$B_\alpha = Z_{\frac{\alpha}{2}} \sigma_{\widehat{\theta}} = Z_{\frac{\alpha}{2}} \sqrt{\mathrm{Var}[\widehat{\theta}]}.$$

2 Elements of the Sampling Problem

2.1 Definition

Definition 2.1. ***Element*** is the object on which measurement is taken.

Definition 2.2. ***Population*** is a set of elements defined according to the aims and objects of the survey.

Definition 2.3. ***Target population*** is the population intended to be investigated (sampled).

Definition 2.4. ***Sampling population*** is the population effectively sampled.

Definition 2.5. ***Sampling units*** is the non-overlapping collections of elements that cover the population effectively sampled.

Definition 2.6. ***Frame*** is the list or any technical device which provides sampling units, or access to sampling units.

Definition 2.7. ***Sample*** is the collection of sampling units selected from a frame.

2.2 Design of Sample Survey

2.2.1 Preparation and Execution

The general procedure is

- Identify target survey group.
- Develop questions.
- Pilot or test the questions/surveys.
- Determine the method of conducting the survey.
- Conduct the survey.
- Use an appropriate analysis technique to analyze the information collected.

Here are the steps in sampling process:

- Define the population.
- Identify the sampling frame.
- Select a sampling design or procedure (probability sampling or nonprobability sampling)
- Determine the sample size.
- Draw the sample.

2.2.2 Methods to Select a Sample

- ***Census***: Complete survey of a population.
- ***Probability/Random sampling***: Every element in the population has a known nonzero probability (not necessarily equal) of being sampled and involves random selection at some point.
- ***Nonprobability/Nonrandom sampling***: Accidental sampling, quota sampling and purposive/judgmental sampling.
    - ***Quota sampling***: The criteria for selection of elements are based on quotas - assumptions regarding the population of interest. After the quotas are decided, the choice of actual sampling units to fit into quotas is mostly left to the interviewer.

The key problem with nonrandom sampling is the selection of elements does not allow proper estimation of sampling errors.

2.3 Random Sampling

- ***Simple random sampling***: Sample is selected completely at random. No special constraints on the sample are imposed. Note that it is an unbiased sample.
- ***Stratified sampling***: The population is divided into sub-populations (strata) and random sample is selected from every stratum. The constraint imposed is stratification.
- ***Cluster sampling*** (***one stage***): The population is divided into large number of (small) clusters, equal or nonequal and clusters are selected at random. The sample is all elements from selected clusters and the constraint imposed is clustering.
    - For stratified sampling, population divided into **few** subgroups; for cluster sampling, population divided into **many** subgroups.
    - For stratified sampling, **homogeneity within subgroups** and **heterogeneity between subgroups**; for cluster sampling, **heterogeneity within subgroups** and **homogeneity between subgroups**.
    - For stratified sampling, choosing elements from within each subgroup; for cluster sampling, random choosing subgroups.
- ***Cluster sampling*** (***two stage***): The population is divided into large number of (bigger) clusters. Sample of clusters is selected and then sample comes from each selected cluster. The sample is all selected elements and the constraint imposed is clustering.
- ***Cluster sampling*** (***multi-stage***): The population is divided into clusters on several levels and sampling is performed at every stage. Sampling is performed at every stage. The sample is all sampling units selected at the last stage and the constraint imposed is multi-clustering.
- ***Double sampling*** (***two-phase sampling***): Sample from sample. Take a bigger sample with some basic measurements and then take a subsample from previously selected sample with more detailed measurements. The sample is the selected elements from the second phase.

    Example 2.1. Phase I: Large sample and stratification variable. Phase II: Subsample and variable of interest.

- ***Systematic sampling***: Elements are selected from an ordered sampling frame. First element is selected at random, and subsequent elements follow a predetermined pattern, usually an interval.
- ***Composite designs***: Most large scale surveys are done using cluster sampling combined with stratification.

  Example 2.2. Clusters are selected from each stratum and then elements are selected from each cluster.

2.4 Errors in Surveys

Definition 2.8. ***Sampling errors*** are due to random sampling, controlled by sample design, sample size and error bound.

Definition 2.9. ***Non-Sampling errors*** are caused by factors other than those related to sample selection. They are not easily identified or quantified.

Non-Sampling errors may come from:

- Imperfect sampling population (inadequate frame-coverage error)
- Poorly designed questionnaire.
- Selection bias, sampling bias.
- Non-Response problem.
- Response error (inaccurate response)
- Systematic error (interviewer bias)
- Processing, editing, entering error.