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. ***Sample survey*** is a partial investigation of the finite population using samples.

The purpose of the sample survey is to obtain information about the population.

Definition 1.6. ***Population*** is a set of elements defined according to the aims and objects of the survey.

Definition 1.7. ***Variable*** is the function defined on population elements, characteristic of population elements. Variable can be quantitative (numerical) or qualitative (categorical).

Definition 1.8. ***Distribution*** or ***frequency distribution*** is the proportion of elements with value in an interval $[a, b], \forall a, b$.

Definition 1.9. ***Sampling*** is the selection of part of the population.

Definition 1.10. ***Sampling method*** is a scientific and objective procedure of selecting units from a population. It provides a sample that is expected to be representative of the population as a whole, and procedures for estimation of the population parameters.

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

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

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 1.11. ***Sample*** is a subset of the population.

Definition 1.12. ***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 1.13. ***Sample function*** is also called statistic, such sample mean (average), sample variance, etc.

Definition 1.14. ***Sampling distribution*** is the distribution of the sample function. This distribution depends on the population distribution of y and function f.