

Random sample/observations and observed sample/data¹

Suppose there is a population Z , say normally distributed $N(\mu, \sigma^2)$.

- “random sample”: Z_1, \dots, Z_n are from the population
- “Observed sample” (after we did the sampling): still denote by Z_1, \dots, Z_n , but the values are fixed (and not random). For example 1.7, 1.8, 1.67, 1.71

Why do we need to discuss two cases?

- Statistics tells people “a way / a method” to do an “experiment” *before people do it*. People want to know how good the method is, or statisticians need to investigate the efficiency of the method *before people adopt the method*. In other word, we need to evaluate the method.

Since there are many possibilities when people do the experiment, statisticians must consider all the possibilities in order to evaluate method. Thus, we must treat Z_1, \dots, Z_n as random in order to *include all possibilities* and to find its efficiency under all possibilities.

Thus, when we say “a statistical method is good”, we mean “it is good if we consider all possibilities”, or it is “good” on “average”. It does not mean the method is good in all cases. It is not surprised if a statistical method “fails” in one experiment.

- Another reason to investigate the random sample is for statistical inference. If a Hypothesis is correct, statisticians can tell you what your experiment will look like (Based on the random sample, i.e. by considering *all possibilities*.) Your experiment result should be “within some range” (i.e. statistical distributions)
- Simply speaking, “random sample” is for theoretical analysis. Any value you can calculate can be treated as RANDOM or A (NONRANDOM) VALUE

Example If people can get samples Z_1, Z_2, Z_3, Z_4, Z_5 (say, 7, 8, 8.5, 9, 10), how to estimate the population mean.

¹This materials will not be included directly in the test or examination

method 1: “sample mean” = $(Z_1 + Z_2 + Z_3 + Z_4 + Z_5)/5 = (7 + 8 + 8.5 + 9 + 10)/5$

method 2: delete the largest and smallest and then take average = $(8 + 8.5 + 9)/3$ (used in Gymnastics grading for players)

Which method is better? We can say one method is better than the other. But it is hard to tell which final result is better.

Example Similarly, in linear regression model, we consider a random model (usually called theoretical model)

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1,$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2,$$

⋮

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n.$$

where

- X_i is known, observable, and non-random,
- ε_i is called random error (unobservable).
- thus Y_i is random.

These are “random samples”. We use it to investigate the overall statistical performance of a method (such as LSE)

After you observed the data, for example, the data in Part 1 of Chapter 1

Obs.	X	Y
1	1.0	0.60
2	1.5	2.00
3	2.1	1.06
4	2.9	3.44
5	3.2	1.17
6	3.9	3.54

You can write the estimated model as

$$\begin{aligned}\hat{Y}_i &= 0.0949 + 0.7699X_i \\ 0.8648 &= 0.0949 + 0.7699 * 1.0 \\ 1.2497 &= 0.0949 + 0.7699 * 1.5 \\ 1.7117 &= 0.0949 + 0.7699 * 2.1 \\ 2.3276 &= 0.0949 + 0.7699 * 2.9 \\ 2.5586 &= 0.0949 + 0.7699 * 3.2 \\ 3.0975 &= 0.0949 + 0.7699 * 3.9\end{aligned}$$

Or you can also (but not used often) write the estimated model as

$$\begin{aligned}Y_i &= 0.0949 + 0.7699X_i + e_i \\ 0.60 &= 0.0949 + 0.7699 * 1.0 - 0.2648 \\ 2.00 &= 0.0949 + 0.7699 * 1.5 + 0.7503 \\ 1.06 &= 0.0949 + 0.7699 * 2.1 - 0.6517 \\ 3.44 &= 0.0949 + 0.7699 * 2.9 + 1.1124 \\ 1.17 &= 0.0949 + 0.7699 * 3.2 - 1.3886 \\ 3.54 &= 0.0949 + 0.7699 * 3.9 + 0.4425\end{aligned}$$

BUT, again, e_i is sometime treated as random, WHY?