Random Samples and Statistics

Collecting data → important aspect of doing statistics.

We want to use a subset of the population to draw conclusions/inference about the entire population.

This subset is called a "Sample".

The way a sample is obtained will affect our confidence in the conclusions we derive using it.

Example Suppose we have a vaccine for Covid-19

We would be interested in knowing its "efficacy".

Consider the following experiment:

① Get two samples of people, group A and group B.

② Give vaccine to group A → treatment arm

placebo to group B → control arm.

③ Expose both groups to the virus, and measure the appropriate response variable.

④ Apply statistical tests to check if there is statistically significant difference between the two responses.

To get correct conclusions about the vaccine → group A and B need to be carefully selected.

(i) Is it possible that certain people can never make it to the samples?

(ii) Do your sample contain only young individuals?

(iii) Are your samples representative of the population?

(iv) Are the sample sizes large enough?

Statistical tools → usually based on principles of probability

Need a precise notion of "randomness" of a sample.

Random Sample

We say $\{X_1, X_2, \dots, X_n\}$ a set of random variables is a "random sample" → if:

(i) X_i has the same probability distribution for $i = 1, 2, 3, \dots, n$.

(ii) the set $\{X_1, X_2, \dots, X_n\}$ is an independent set of random variables.

Note ① Sometimes a random sample → also reffered to as a set of independent and identically distributed random variables.

"iid" random variables.

② The common distribution of all the X_i's

"population distribution".

③ $\{X_1, X_2, X_3, \dots, X_n\}$ —sampling procedure→ $\{x_1, x_2, \dots, x_n\}$

Random sample: a collection of random variables

Sample data: Changes everytime we run the sampling procedure

Example Suppose we are getting a random sample of size 3

from the population: $\{1, 2, 3, 4, 5\}$

To get a random sample → need to sample with replacement.

Let $\{Num1, Num2, Num3\}$ → random sample → collection of r.v's

- samples first number in the sample
- samples 2nd number in the sample
- samples third number in the sample.

∴ Sampling with replacement → there are $5 \times 5 \times 5$ ways of getting sample data.

ie we have 5^3 choices of sample datasets → Each run of sampling procedure gets one of the elements in

$$\{(1,1,1), (1,1,2), (1,2,1), (2,1,1), \dots, (5,5,4), (5,5,5)\}$$

Statistic

We are often interested in some features/characteristics of the random sample.

A statistic → is a quantity calculated using a random sample (say $\{X_1, X_2, \dots, X_n\}$)

Since the calculation of the statistic involves random variables → the statistic will also be a random variable.

Note: ① The value of the statistic changes as the sample data changes.

② A statistic is a random variable defined on sample data corresponding to a random sample.

If $X :=$ values of the r.v's in random sample

sample data → $\{(x_1, x_2, \dots, x_n) \mid x_i \in X \;\; x_{i=1,2,\dots,n}\}$

Since $\{X_1, X_2, \dots, X_n\}$ is an independent set, the "joint distribution" for $X_1, X_2, \dots, X_n$ is given as

$$p(x_1, x_2, x_3, \dots, x_n) = p_X(x_1) \cdot p_X(x_2) \cdots p_X(x_n)$$

where $p_X(x)$ → pmf of the common distribution X.

We are assuming that X is discrete.

If X is continuous use pdf $f_X(x)$ to define the joint pdf of $X_1, X_2, \dots, X_n$.

③ Can use the joint pmf/pdf of $X_1, X_2, \dots, X_n$ to calculate the probability distribution of the statistic of interest

"Sampling distribution" of the statistic.

Example Revisited:

Recall: Population $= \{1, 2, 3, 4, 5\}$

Random sample $= \{Num1, Num2, Num3\}$

Examples of statistics defined using $\{Num1, Num2, Num3\}$

Name of the statistic	Definition of the statistic	Examples of evaluation on sample data
① Sample mean	$\bar{X} = \frac{Num1 + Num2 + Num3}{3}$	$\bar{X}(1,1,1) = \frac{1+1+1}{3} = 1$ $\bar{X}(2,1,2) = \frac{2+1+2}{3} = \frac{5}{3}$ $\bar{X}(3,1,1) = \frac{3+1+1}{3} = \frac{5}{3}$ $\bar{X}(4,5,1) = \frac{4+5+1}{3} = \frac{10}{3}$
② Sample Median	$\tilde{X} = \text{median}\{Num1, Num2, Num3\}$	$\tilde{X}(1,1,1) = 1$ $\tilde{X}(2,1,4) = 2$ $\tilde{X}(3,3,1) = 3$ $\tilde{X}(1,5,4) = 4$
③ Sample Maximum	$Max = \max\{Num1, Num2, Num3\}$	$Max(1,1,1) = 1$ $Max(1,2,5) = 5$ $Max(4,1,3) = 4$ $Max(3,1,1) = 3$
④ Sample Total	$T_0 = Num1 + Num2 + Num3$	$T_0(1,1,1) = 1+1+1 = 3$ $T_0(1,3,5) = 1+3+5 = 9$ $T_0(2,2,4) = 2+2+4 = 8$ $T_0(3,1,2) = 3+1+2 = 6$
⑤ Sample Variance	$S^2 = \frac{(Num1-\bar{X})^2 + (Num2-\bar{X})^2 + (Num3-\bar{X})^2}{2}$	$S^2(1,1,1) = \frac{(1-\bar{X}(1,1,1))^2 + (1-\bar{X}(1,1,1))^2 + (1-\bar{X}(1,1,1))^2}{2} = 0$ $S^2(1,2,3) = \frac{(1-\bar{X}(1,2,3))^2 + (2-\bar{X}(1,2,3))^2 + (3-\bar{X}(1,2,3))^2}{2} = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{2} = \frac{1^2 + 0^2 + 1^2}{2} = \frac{2}{2} = 1$

To calculate the sampling distribution of these statistics → need to know the distribution of the population $\{1, 2, 3, 4, 5\}$

Random Samples and Statistics

population.

sample-data.

$\{x_1, x_2, x_3, \ldots, x_n\}$

{student 1, student 2, student 3, student 4, student 5}

↓ measure their heights.

Sample 1: {John, Macallus, Mia, Neo, Asmin}

sample data: {5.9, 5.12, 5.6, 4.2, 6.0}.

Sample 2: {Kevin, Nina, Rhea, Soumath, Angela}

Sample 3: {Matt, Sofia, Luke, Sruthi, Sargeant}

Sampling without replacement vs Sampling with replacement.

↓ every subsequent sample choice depends on the other choices

↓ choices are going to be independent of each other.

Sample 4: {Denis, Dennis, Dennis, Dennis, Dennis}

Typically: If the sample is small relative to the population size

↓

we can say that "sampling without replacement will be approximately the same as sampling with replacement"

↓

If sample is ≤ 0.05 of the population.

Sampling with replacement: gives sample choices that are independent of each other. (good)

might not get a sample of "distinct" elements.

↓ test subjects.

(not so good!)

Sampling without replacement: gives sample choices that are dependent on each other (bad)

gives a sample of distinct individuals.

Random Sample → collection of n random variables

size n $\{X_1, X_2, X_3, \ldots, X_n\}$

(i) The probability dist of X_i for $i = 1, 2, \ldots, n$ is the same

↓

distribution of the population.

(ii) $\{X_1, X_2, \ldots, X_n\}$ is a independent set random variable.

Recall: $\{X_1, X_2, \ldots, X_n\}$ are ind r.vs if

↓

$\{X_{i_1}, X_{i_2}, \ldots, X_{i_k}\} \subseteq \{X_1, \ldots, X_n\}$

the joint probability dist of $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ satisfies

$$p(x_1, x_2, \ldots, x_n) = p_{X_{i_1}}(x_1) \cdot p_{X_{i_2}}(x_2) \cdots p_{X_{i_k}}(x_k)$$

Random sample $\{X_1, X_2, X_3, \ldots, X_n\}$ → (sampling procedure) → $(x_1, x_2, x_3, \ldots, x_n)$

↓ random variables ↓ sample data.

Example:

Population: $\{2, 5, 9, 10\}$

Want a random sample of size 2. (need to sample with replacement)

Sample = {Num1, Num2} → sample data.

$\{(x_1, x_2) \mid x_i \in \{2, 5, 9, 10\}\}$

∴ Choices of sample data → $4 \cdot 4$

$4^2 = 16$ choices of sample data.

$\{(2,2), (2,5), (5,2), (2,9), \ldots (10,10)\}$

Statistic ⟶ some number calculated using a random sample.

⇕

Statistic itself will be a random variable

⇕

value of the statistic changes everytime the sample data changes.

Num1 \ Num2	2	5	9	10
2	(2,2)	(2,5)	(2,9)	(2,10)
5	(5,2)	(5,5)	(5,9)	(5,10)
9	(9,2)	(9,5)	(9,9)	(9,10)
10	(10,2)	(10,5)	(10,9)	(10,10)

Example: Population: $\{2, 5, 9, 10\}$ ⟶

x	2	5	9	10
$P_X(x)$	0.1	0.1	0.6	0.2

Sample: $\{Num1, Num2\}$

Interesting Statistics

evaluating the statistic on sample data.

① sample mean. $\bar{X} = \frac{Num1 + Num2}{2}$

$\bar{X}(2,2) = \frac{2+2}{2} = 2$

$\bar{X}(5,9) = \frac{5+9}{2} = \frac{14}{2} = 7$

$\bar{X}(5,10) = \frac{5+10}{2} = \frac{15}{2} = 7.5$

② sample Min $Min = \min\{Num1, Num2\}$

$Min(2,2) = 2$

$Min(2,5) = 2$

③ Sample Max $Max = \max\{Num1, Num2\}$

$Max(2,2) = 2$

$Max(2,5) = 5$

④ Sample range $Range = Max(Num1, Num2) - Min(Num1, Num2)$.

$P(\text{sample mean} \leq 5)$?

⇕

$P(\bar{X} \leq 5)$