

Statistics

- Some questions in life that cannot be answered without data.
- When you infer something from data that was collected, some other questions arise.
 - How was the data collected?
 - By whom?
 - Why?
 - When?
 - Who funded the effort?
- Basic purpose of collecting data.
 - Enumerate: "How many supporters of democracy in the population?"
Census
↳ Salmon
↳ gorillas
↳ Population
- Cause and effect
 - Does smoking cause cancer?
 - Does vaccine prevent COVID19?

Terminology

Population: set of objects that is being enumerated.

N-value: size of population

Sample: subset of population.

Census: enumeration of population.

Survey: enumeration of sample.

Statistic: numerical value of some characteristic of the population obtained from a sample.
diseased animal
defective fan
3% diseased. ← Statistic

Parameter: true numerical value of some characteristic of population.
3.13% diseased ← parameter.

Ex.1

You are quality control manager at a factory that produces fans.

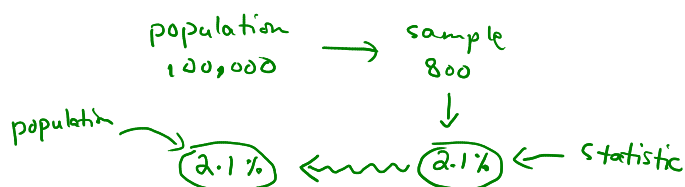
100,000 fans in a batch.

↳ If more than 2.5% fans are defective, then don't send the batch.

Take a sample of 800 fans from this batch of 100,000 fans.

↳ Number of defective fans was 17.

$$\% \text{ defective fans in the sample} = \frac{17}{800} \times 100 = 2.1\%$$



$$\frac{17}{800} \approx \frac{N}{100,000}$$

sample population

$$N \approx \frac{17}{800} \times 100,000$$
$$\approx 2125$$

Ex. 2

You are a wildlife biologist studying the ecosystem in a small lake.

Pike — eats the children of other fish.

N-value ?

Honey bees population
↓
Flower — Flower
Pollination

1. Capture

Count # of Pike

$$n_1 = 200$$

Tag them and put them back in the lake.

2. Re-capture

$$n_2 = 150$$

k = 21 had tags.

⇒ Ratio tagged Pikes in the sample is $\frac{21}{150}$

Ratio of tagged Pike in Sample $\approx$ Ratio of tagged Pike in Lake ← Population

$$\frac{21}{150} \approx \frac{200}{N}$$

$$\Rightarrow N \approx \frac{150 \times 200}{21} = 1428.57$$

After rounding,

$$N \approx 1429$$

Capture Re-Capture Method

Ratio of PTI supporters?

Sampling

How do you choose data points?

↳ Bias

Biased sample

vs.

Unbiased sample

— If people from only some segments of society are selected, then other segments are left out. ← Selection bias

— Convenience sampling

You choose people/data points that were most easily accessible.

↳ Outside some shop.

↳ time
↳ effort
↳ expense

Whenever you see some statistic, always ask "How was the data collected?"