6 Jan 2025 – Introduction to Statistics – Week 01

Why? – Statistics helps us process data
– no time to collect data for entire population

1 group activity per week (30%)

→ sample size (e.g.: mess, weight calc.)

parameter — quantitative measure for the population (mean, median, mode, deviation, etc.)

population

target population
e.g.: → youth
→ diabetic people
→ everyone in this class

sampling scheme

theory of estimation

statistic — sample measures

sample

Should be a good representation of the population

$\boxed{\text{randomized sample}}$ → good representation → good results

good: close measures but exactly equal to

$\boxed{\text{if not}}$ → results will be biased (not acceptable)

Why sampling some of the population?

→ Resources are less
→ practical considerations

* destructive sampling

e.g.: how many in the target population like your product
what is the avg. lifespan of the bulbs you manufacture

randomized
statistics → sample measures → will always be known
estimate
parameters → population measures → will be **unknown**

$\boxed{\text{randomized sample of adequate size}}$

→ we may not always get adequate size (accuracy ↓)

* uncertainity

Estimation : using sample statistics to predict population parameters

randomized

rational guessing

* fishing net and fishing rod

higher probability/chance

* range and single value } estimation

↓

intervals

point estimation

→ try to be accurate

Doubts

① Where does all this fit in Computer Science?

- AI & ML, NLP
- Data Science & big data
- Computer vision
- Algorithm design & analysis: randomized algorithms ?? , performance analysis.
- Cryptography and security
- Software Engineering
- Database systems: query optimization, indexing and sampling.
- Networking
- Game development, simulation, robotics.

8 Jan 2025

→ population data will always be unknown.

→ | <u>Population Parameters</u> | <u>Estimators</u> |
|---|---|
| → mean, μ | $\bar{x}$ |
| → variance, σ^2 | s^2 |
| → proportion, p | $\hat{p}$ |

→ Statistics are used as estimators.

μ

$X_1, X_2, \ldots, X_n \rightarrow$ $\boxed{\text{random sample}}$ $\rightarrow$ all sampled units are independent

$$E(X_i) = \mu$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$E(\bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n} \sum E(X_i) = \frac{1}{n} \sum \mu = \mu$$

$$\boxed{E(\bar{x}) = \mu}$$

Sample mean is an unbiased estimator

imp criteria for an estimator

$$\underbrace{E(\bar{x} - \mu)}_{\text{bias}} = 0$$

$$\text{Bias} = E(T - \theta)$$

estimator (statistic)

parameter

$$\sigma^2 = V(X_i) = E(X_i^2) - (E(X_i))^2 = E(X_i^2) - \mu^2$$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

$$= \frac{1}{n} \sum X_i^2 - \bar{x}^2$$

$$E(S^2) = \frac{1}{n} \sum E(X_i^2) - E(\bar{x}^2)$$

$$= \frac{1}{n} \left(\sum (\sigma^2 + \mu^2) \right) - E(\bar{x}^2)$$

$$= \sigma^2 + \mu^2 - E(\bar{x}^2)$$

$$V(\bar{x}) = E(\bar{x}^2) - (E(\bar{x}))^2$$

$$= E(\bar{x}^2) - \mu^2$$

$$V(\bar{x}) = V\left(\frac{1}{n}\sum x_i\right)$$

$$= \frac{1}{n^2} V\left(\sum x_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} V(x_i)$$

$$= \frac{1}{n^2} n\sigma^2 = \sigma^2/n$$

$$V(aX_i) = a^2 V(X_i)$$

$$V(X_1 + X_2) = V(X_1) + V(X_2) + 2\,\mathrm{Cov}(X_1, X_2)$$

$$(a+b)^2 \quad (a-b)^2$$

random $\Rightarrow$ independent

$$\Rightarrow \underbrace{\mathrm{Cov}(X_i, X_j)}_{\text{linear dependence}} = 0$$

$$\therefore \quad sd(\bar{x}) = \frac{\sigma}{\sqrt{n}} = se(\bar{x})$$

se: standard error
$\underbrace{\hphantom{standard error}}$
avg. amount of error of an estimate

→ difference between statistics and estimators

$$\therefore \quad E(\bar{x}^2) = \sigma^2/n + \mu^2$$

$$\Rightarrow \quad E(s^2) = \sigma^2 + \cancel{\mu^2} - \left(\sigma^2/n + \cancel{\mu^2}\right)$$

$$= \sigma^2\left(1 - \frac{1}{n}\right)$$

$$E(s^2) \neq \sigma^2$$

$$\underbrace{s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2}_{\text{not an unbiased estimator}}$$

What is an unbiased estimator for σ^2?

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$= \frac{1}{n-1} \left[\sum x_i^2 - n\bar{x}^2 \right]$$

$$E(s^2) = \frac{1}{n-1} \sum \left(E(x_i^2) \right) - n \, E(\bar{x}^2)$$

$$= \frac{1}{n-1} \left(\sum \sigma^2 + \mu^2 \right) - n \left(\frac{\sigma^2}{n} + \mu^2 \right)$$

$$= \frac{1}{n-1} \quad n \left(\sigma^2 + \mu^2 \right) - \sigma^2 - n\mu^2$$

$$= \frac{1}{\cancel{(n-1)}} \cancel{(n-1)} \sigma^2$$

$$= \sigma^2$$

$$E(s^2) = \sigma^2 \qquad s^2 = \frac{1}{n-1} \sum \left(x_i - \bar{x} \right)^2$$

Doubts

① What is the difference between an estimator and a statistic?

From stats.stackexchange.com:

* A statistic is a function of a sample.
* An estimator is a function of a sample related to some quantity of the distribution.

  some property, usually unknown

* A statistic is not an estimator: An estimator is a statistic with something added. To turn a statistic into an estimator, you simply spell out which target quantity you want to estimate.

* Different estimators based on the same statistic:
  1) sample mean as an estimator for distribution / population mean: zero bias
  2) sample mean as an estimator for distribution variance: usually biased.

<u>13 Jan 2025</u>

① What is expectation?
What is random sample? Is it the same as random variable?
What are variance and covariance? → Physical significance...