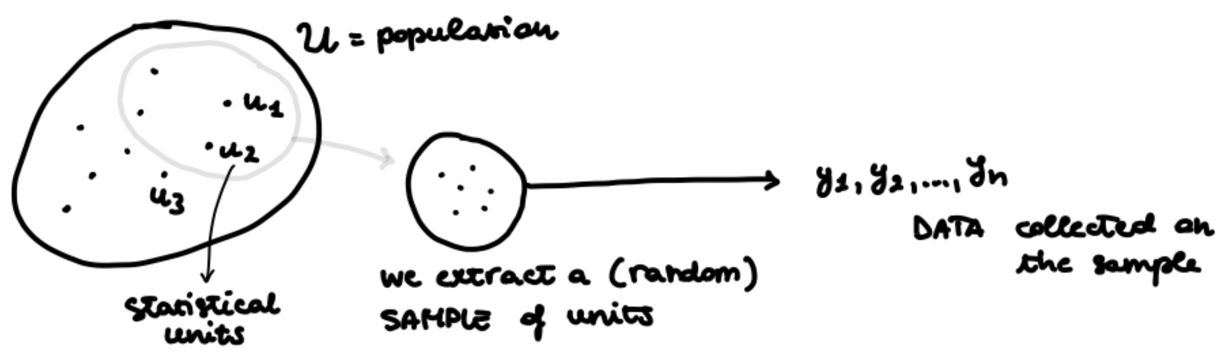
1 oct 2024 - LEC 1

PREREQUISITES OF STATISTICAL INFERENCE



The observations are $y_i = y(u_i)$

STATISTICAL INFERENCE: we try to understand the population starting from a sample

We use probabilistic tools

STATISTICAL MODEL: on the data, we specify a probabilistic model suited to describe a particular phenomenon.

probabilistic model: $Y \sim p_0(y)$ we draw $(y_1, \dots, y_n)$

statistical model: given $(y_1, \dots, y_n)$, we define a set of "reasonable" distributions $p(y)$ that could have generated it, and try to recover the particular $p_0(y)$ within this set

$y_i \rightsquigarrow Y \sim p(y; \theta) \quad \theta \in \Theta \subseteq \mathbb{R}^p$

known (prob. model)

unknown (parameter): identifies the particular element

e.g. $\underline{y} = (y_1, \dots, y_n)$ heights of a sample of students

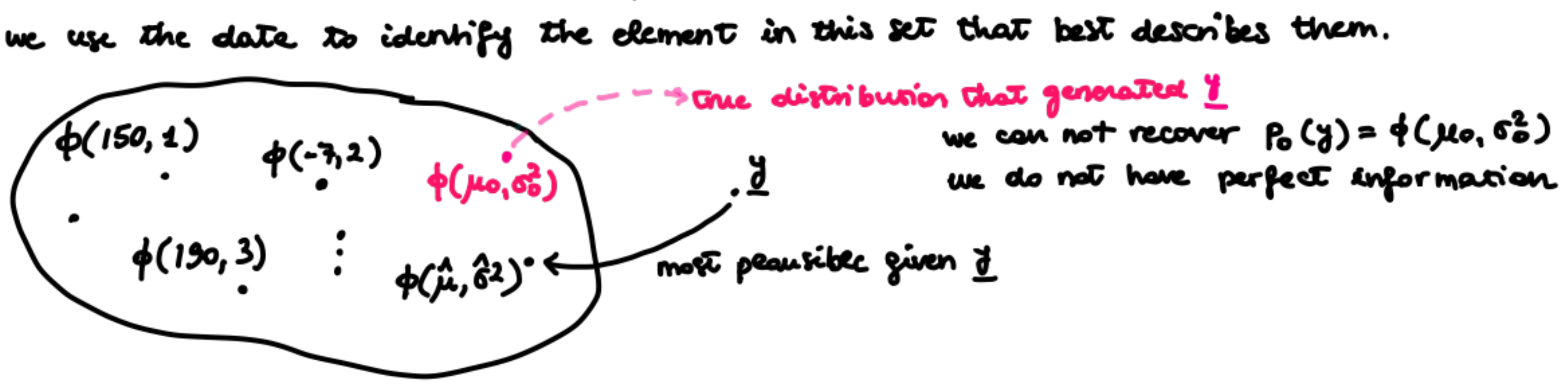
a Gaussian distribution is a reasonable model

random variable: <u>before</u> observing the data. The distribution of Y describes all possible realisations y_i

$\Rightarrow$ statistical model $\underbrace{Y_i \sim N(\mu, \sigma^2)}$ $\quad \mu \in \mathbb{R} \quad \sigma^2 \in (0, +\infty)$

MODEL: set of all Gaussian distributions

inference: we use the data to identify the element in this set that best describes them.



e.g. $\underline{y} = (y_1, \dots, y_n)$ counts of cars passing in a street

statistical model $\quad Y_i \sim \text{Pois}(\lambda) \quad \lambda \in (0, +\infty)$ indep.

Methodologies:

- POINT ESTIMATION: we identify one element (the most plausible) within the set of distributions $\hat{p}(y)$
- INTERVAL ESTIMATION (confidence intervals): subset of reasonable elements, where the subset has a known degree of "confidence" that the true element (p_0) is contained in it.
- HYPOTHESIS TESTING: we ask if there is enough evidence in the sample to draw conclusions about a particular statement ("null hypothesis")

POINT ESTIMATE

given the set of elements $\{p(y; \theta) \; ; \; \theta \in \Theta \subseteq \mathbb{R}^p\}$ we want to identify the most plausible $\hat{p}(y)$

It is identified through a particular element of θ: $\hat{\theta}$ i.e. $\hat{p}(y) = p(y; \hat{\theta})$

ESTIMATE $\quad \hat{\theta} = \hat{\theta}(y)$ is a function of the observed values (realisations)

ESTIMATOR $\quad \hat{\Theta} = \hat{\Theta}(Y)$ is a function of the random variable

$\Rightarrow$ we study the distribution of the estimator, its expected value, variance, ...

HYPOTHESIS TESTING

$\begin{cases} H_0: \theta \in \Theta_0 \subset \Theta & \text{null hyp.} \\ H_1: \theta \in \Theta \setminus \Theta_0 & \text{alternative hyp.} \end{cases}$ e.g. $\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases}$

TEST: we partition the sample space into the REJECT ($\mathcal{R}$) and ACCEPTANCE ($\mathcal{A}$) regions

$\mathcal{A}$: values of y that suggest that H_0 is true

$\mathcal{R}$: values of y that suggest that H_0 is false $\rightarrow$ reject H_0

TEST STATISTIC: a function of the data that defines the two regions. $T(y)$

$\mathcal{A} = \{y \in \mathcal{Y} : T(y) \text{ suggests } H_0\}$

$\mathcal{R} = \{y \in \mathcal{Y} : T(y) \text{ suggests } H_1\}$

How do we draw conclusions?

I. FIXED SIGNIFICANCE LEVEL α

we guard against the 1st type error: we fix α to a small value (e.g. $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.10$)

and we derive $\mathcal{A}$ and $\mathcal{R}$ so that $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$ is equal to α (or at most α)

In other words, $\alpha = \mathbb{P}(1^{st} \text{ type error}) = \mathbb{P}(y \in \mathcal{R} \mid H_0 \text{ true})$

Of course, the smaller α is, the smaller $\mathcal{R}$ will be (I want to reject H_0 only if I am really really confident)

II. OBSERVED SIGNIFICANCE LEVEL (p-value) $\quad \alpha^{obs}$

it is the probability of observing "more extreme" values (i.e. more against H_0) than the ones we observed.

- if the reject region is of a one-tailed test

  $H_1: \theta > \theta_0$ (right tail) $\Rightarrow \alpha^{obs} = \mathbb{P}_{\theta_0}(T \geq t^{obs}) = \mathbb{P}(T(Y) \geq t(y^{obs}) \mid H_0 \text{ true})$

  $H_1: \theta < \theta_0$ (left tail) $\Rightarrow \alpha^{obs} = \mathbb{P}_{\theta_0}(T \leq t^{obs})$

- if the reject region is of a two-tailed test

  $H_1: \theta \neq \theta_0 \Rightarrow \alpha^{obs} = 2 \min\{\mathbb{P}_{\theta_0}(T \geq t^{obs}) \; ; \; \mathbb{P}_{\theta_0}(T \leq t^{obs})\}$

The two procedures are related: if $\alpha^{obs} < \alpha$, then I reject H_0 in a fixed-level test of level α

CONFIDENCE INTERVALS of confidence $(1-\alpha)$

it is a random interval $\hat{C}(Y)$ such that $\mathbb{P}(\theta \in \hat{C}(Y)) = 1 - \alpha$ for all $\theta \in \Theta$

With probability $(1-\alpha)$, the interval contains the true value of the parameter, whatever it is.

After we compute the interval with the data (hence, we get a fixed numeric interval), it either contains the true θ or not.

The probability must be interpreted regarding to the random quantity.

We build it through the identification of a PIVOTAL QUANTITY: a function $g(Y; \theta)$ of the r.v. Y and the parameter θ such that its distribution does not depend on θ (hence it is completely known).

Then we look for the interval (u, v) such that $1 - \alpha = \mathbb{P}(u < g(Y, \theta) < v)$.

With the data we compute $\quad \hat{c}(y^{obs}) = \{\theta \in \Theta : g(y^{obs}, \theta) \in (u, v)\}$

$\hookrightarrow$ all values of the parameter that, given the observed data, give a value of g within the (u, v) interval.