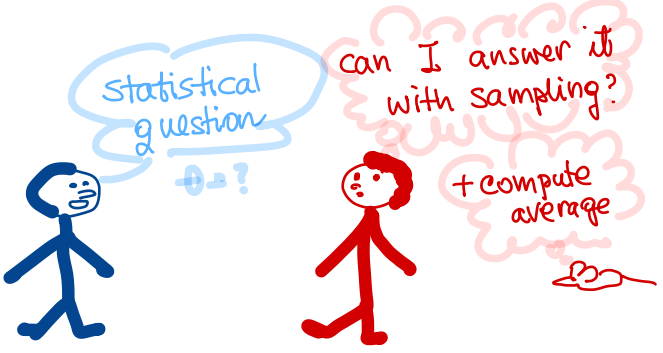
Lecture 1

Resampling Inference = Embracing Randomness



What is (resampling) inference?

Statistical question ← method to answer

depends on unknown P

Q1: Avg number people riding light rail? $\mu = ?$ ← parameter of interest

Data $\{X_1, .. X_n\}$ $j = 1:n$

$X_1, .. X_n \sim iid\ P$

X_i = # people rode on day i

$X_1, .. X_n$: r.v.'s

estimator for μ

$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ← Estimator: avg of data ← Data

$\hat{\mu}$: random variable

Q2: $\theta = Pr[X \leq 10] = ?$

↑ param of interest

$\hat{\theta} = \frac{\#\{\text{days } i \mid X_i \leq 10\}}{n}$ ← depends on P

$\hat{\theta}$: estimator

Q3: influence of alcohol consumption on cancer?

$X \in \mathbb{R}^m$ $X \sim$ { alcohol, history, smoking, age, . }

$\mathcal{D} = \{(x_i, y_i)\}$ subjects $i = 1:n$

$\mathcal{D}$: data; x_i: alcohol; y_i: cancer

depends on P_{xy} and Model

$\mathcal{D} = \{(x_i, y_i)\}$ — data; x_i: alcohol; y_i: cancer

Parameter of interest: $\beta_{alc} \gtrless 0$?

Model: $y = \beta_0 + \beta^T X + \varepsilon$

$= \beta_0 + \beta_{alc} \cdot x_{alc} + \beta_{sm} \cdot x_{sm} + \cdots + \varepsilon$

→ Linear regression algorithm → estimator $\hat{\beta}_{alc}$

I. Inference about a real, unknown data distribution P (Sampling from real P)

II. Inference about a complicated model P (by sampling from model P)

Ex1: Generative models (neural network/AI)

Ex 2: DNA Model (build models from 1st principles) X

← m long

$\{A, C, G, T\}^m$

Pr $X_j = A$

$X_{j+1} = C$

but inference is intractable

III. Bayesian inference $\subseteq$ II.