# Formalism of statistical modeling

### Definition 1.1

A statistical model is a triple $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$.

- $\mathcal{X}$ is a set called the **sample space**.
- $\Theta$ is a set called the **parameter space**.
- $\{f_\theta : \theta \in \Theta\}$ is a family of pdf's or pmf's on $\mathcal{X}$ indexed by $\Theta$, that is, $f_\theta : \mathcal{X} \to [0, \infty)$ with either

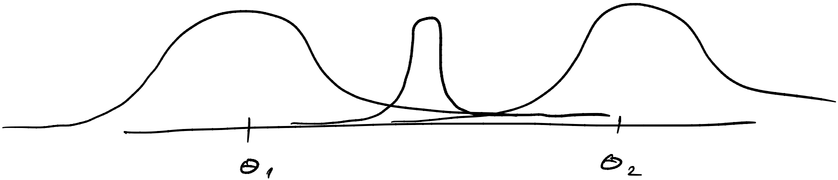
$$\int_{\mathcal{X}} f_\theta(x)dx = 1, \qquad \text{or} \qquad \sum_{x \in \mathcal{X}} f_\theta(x) = 1.$$

# Example: Normal location model

- $\mathcal{X} = \mathbb{R}^p$
- $\Theta = \mathbb{R}^p$
- $f_\theta(x) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\|x-\theta\|_2^2\right), x \in \mathcal{X}, \theta \in \Theta.$

$\Sigma = I_p$

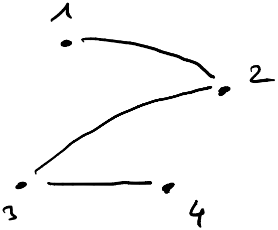
$p = 1$



# Example: Erdös-Rényi random graph model

- $\mathcal{X} = \{A \in \{0,1\}^{n \times n} : A' = A, A_{ii} = 0 \text{ for all } i \in [n]\}$
- $\Theta = [0,1]$
- For $A \in \mathcal{X}$, $\theta \in \Theta$,

$$f_\theta(A) = \prod_{\substack{i,j=1 \\ i<j}}^{n} \theta^{A_{ij}} (1-\theta)^{1-A_{ij}}.$$




$$A = \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 \\ 4 & 0 & 0 & 1 & 0 \end{array}$$

# STATISTICAL MODELS

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We assume the actual data have been generated from the distribution $f_\theta$ for some unknown $\theta \in \Theta$. To emphasize that the data are realizations of a **random process** and could have been also different, we describe them mathematically as **random variables**.

formally: $X$ is a random variable taking values in $\mathcal{X}$

$X \sim f_\theta$ for some $\theta \in \Theta$

observed

unobserved

# THE IID MODEL

iid ... **i**ndependent **i**dentically **d**istributed

Often it makes sense to assume a product form:

$$\mathcal{X} = \mathcal{X}_0^n \qquad \text{e.g. } \mathcal{X}_0 = \mathbb{R}$$

$$f_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i), \qquad \theta \in \Theta, x = (x_1, \dots, x_n)' \in \mathcal{X}_0^n,$$

$p_\theta$ a pdf or pmf on $\mathcal{X}_0$ $\qquad$ $n$ ... sample size.

For instance: (both our examples above)

- measuring the molecular weight of the same substance $n$ times with a mass spectrometer
- taking an fMRI of $n$ randomly selected individuals
- throwing $n$ darts at a target

# The IID model

$$\mathcal{X} = \mathcal{X}_0^n$$

$$f_\theta(x) = \prod_{i=1}^{n} p_\theta(x_i), \qquad \theta \in \Theta, x = (x_1, \dots, x_n)' \in \mathcal{X}_0^n,$$

$p_\theta$ a pdf or pmf on $\mathcal{X}_0$ $\qquad n \dots$ sample size.

In this case:

$$X = (X_1, \dots, X_n)' \sim f_\theta, \qquad \text{for some } \theta \in \Theta$$

in other words:

$$X_1, \dots, X_n \overset{iid}{\sim} p_\theta, \qquad \text{for some } \theta \in \Theta$$

$X_i$ takes values in $\mathcal{X}_0$

# Parameters of interest vs. nuisance parameters

Consider a statistical model

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We are often only interested in some components of $\theta \in \Theta$.

E.g.:

$$f_\theta(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$
$$\theta = (\mu, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+$$

But we want to estimate only $\mu \in \mathbb{R}$. Then we call $\sigma^2 > 0$ a **nuisance parameter**.

# Parameters of interest vs. nuisance parameters

Consider a statistical model

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We are often only interested in some components of $\theta \in \Theta$.

In general: Let $\psi : \Theta \to \Psi$ be a function. We may want to estimate and do inference on

$$\psi(\theta).$$

E.g.: $\psi : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, $\psi(\mu, \sigma^2) = \mu$.