Maximum Likelihood Estimator:

Find "the most likely" estimate of a parameter θ, $\hat{\theta} = \underbrace{T(X_1, \dots, X_n)}_{\text{some function of the data, called a statistic}}$.

Question: Does T use as much information as possible?

Idea of sufficiency:

Def Let the joint density function of $X_1, \dots, X_n$, i.e. the likelihood, be $f = f(x_1, \dots, x_n ; \theta)$.

Lets write $\vec{x} \Leftrightarrow \vec{y}$ if $f(\vec{x} ; \theta) = f(x_1 \dots x_n ; \theta)$

$= C(\vec{x}, \vec{y}) \cdot f(\vec{y} ; \theta)$.

C might depend on $\vec{x}, \vec{y}$, but not θ.

A statistic $T = T(\vec{x})$ is sufficient for θ

if $T(\vec{x}) = T(\vec{y}) \Rightarrow \vec{x} \Leftrightarrow \vec{y}$.

To rephrase: "A statistic is sufficient if the likelihood function can be evaluated knowing only that statistic."

Ex: Let $X_1, \dots, X_n \sim \text{Bernoulli}(p)$ random variables. IID.

$\Rightarrow f(x ; p) = p^x (1-p)^{1-x}$.

$\Rightarrow \mathcal{L}(p) = \prod_i p^{x_i} (1-p)^{1-x_i} = p^{\Sigma x_i} (1-p)^{n - \Sigma x_i}$.

$\Rightarrow \underbrace{S = \sum X_i}_{\text{a statistic}}$ is sufficient since $\mathcal{L}(p) = p^S (1-p)^{n-S}$.

<u>Ex:</u> $X_1, \dots, X_n \sim$ IID $N(\mu, \sigma^2)$

$$f(\vec{x}; \mu, \sigma^2) = \mathcal{L}(\mu, \sigma^2)$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-(x_1-\mu)^2/2\sigma^2} \cdots e^{-(x_n-\mu)^2/2\sigma^2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2}$$

$\frac{1}{n}\sum X_i$

$$\text{Expand } \sum (X_i-\mu)^2 = \sum (\underline{X_i - \bar{X}} + \underline{\bar{X} - \mu})^2$$

$$= \sum\left[(X_i-\bar{X})^2 + 2(X_i-\bar{X})(\bar{X}-\mu) + (\bar{X}-\mu)^2\right]$$

$$= \sum (X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum(X_i-\bar{X}) + n(\bar{X}-\mu)^2$$

where $S^2 = \frac{1}{n}\sum (X_i-\bar{X})^2$.

$$= nS^2 + 2(\bar{X}-\mu)\cancel{\left(\sum X_i - \sum X_i\right)} + n(\bar{X}-\mu)^2$$

$$= nS^2 + n(\bar{X}-\mu)^2$$

$$\Rightarrow \mathcal{L}(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^2 e^{-\frac{1}{2\sigma^2}\left(nS^2 + n(\bar{X}-\mu)^2\right)}$$

let $T = T(X_1, \dots, X_n) = (\bar{X}, S^2)$.

$\Rightarrow T = (\bar{X}, S^2)$ is a sufficient statistic.

Other sufficient statistics include:

$T_1 = (\bar{X}, \sqrt{S^2})$

$T_2 = (X_1, \ldots, X_n)$ ← always a sufficient statistic

$T_3 = (\bar{X}, S^2, X_3)$ ← sufficient, but redundant.

<u>Minimal sufficient</u> A statistic T is minimal sufficient if (1) it is sufficient
(2) it is a function of <u>every</u> other sufficient statistic

<u>Note</u> Ex: $X_1, \ldots, X_n$ is not a function of $\bar{X}$, and therefore <u>not</u> minimal sufficient in the previous example.

<u>Thm</u> T is minimal sufficient if $T(\vec{x}) = T(\vec{y})$ if and only if $\vec{x} \Leftrightarrow \vec{y}$.

<u>Thm</u> (Factorization Theorem) T is sufficient for θ if and only if there exist two functions $g = g(t;\theta)$ and $h = h(\vec{x})$ such that

$$f(\vec{x};\theta) = L(\theta) = g(T(\vec{x}), \theta)\, h(\vec{x}).$$

Idea for why the Factorization Theorem is true, and important:

$$\text{If} \quad f(\vec{x};\theta) = \mathcal{L}(\theta) = g(T,\theta)\, h(\vec{x}),$$

then $\ell(\theta) = \log g(T,\theta) + \log h(\vec{x})$

$$\Rightarrow \ell'(\theta) = \underbrace{\frac{1}{g(T,\theta)} \frac{\partial g}{\partial \theta}}_{\text{only depends on } T,\, \theta}$$

$\Rightarrow$ This means that the solution to $\ell'(\theta) = 0$ only depends on T $\Rightarrow$ $\hat{\theta}$ can be written only in terms of T.

<u>Alternative definition</u> (but equivalent) from Casella & Berger:

T is sufficient for θ if the conditional distribution of $X_1, \dots, X_n$ given T does <u>not</u> depend on θ.

To describe roughly what this means in the discrete case:

$$\mathbb{P}_\theta(\vec{X} = \vec{x} \mid T = T(\vec{x})) = \frac{\mathbb{P}_\theta(\vec{X} = \vec{x} \text{ and } T = T(\vec{x}))}{\mathbb{P}_\theta(T = T(\vec{x}))}$$

$$= \frac{\mathbb{P}_\theta(\vec{X} = \vec{x})}{\mathbb{P}_\theta(T = T(\vec{x}))} = \frac{f(\vec{x};\theta)}{q(T(\vec{x});\theta)}$$

is the pdf of T

$\Rightarrow$ If T is sufficient according to this definition, the Factorization Theorem is basically implied.
(Detailed discussion and proof of the Factorization Thm in Casella and Berger.)

Hypothesis Testing

"Does medicine X effectively treat condition Y?"
Idea: Compare treated and untreated populations.

A hypothesis is a belief that something is true.

Philosophical side: Hypotheses are easy to disprove, and hard (if not impossible) to prove.

Null hypothesis : "background hypothesis", ex: medicine X doesn't help or hurt condition Y.

Alternative hypothesis : complement of the Null hypothesis
$\Rightarrow$ Ex: medicine X effectively treats condition Y.

Parametric hypothesis testing : partition parameter space according to hypotheses

Null Hypothesis : $H_0 : \theta \in \Theta_0 \subset \Theta$

Alternative hypothesis : $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$

Idea: Observe data $\vec{X}$, if $\vec{X} \subset R \subset \mathcal{X}$

then reject H_0.

R: rejection region

$\mathcal{X}$: space of all possible outcomes.

If $\vec{X}$ is <u>inconsistent</u> with probable outcomes assuming H_0 is true, then we reject H_0.

	Retain H_0	Reject H_0
H_0 true	✓	Type I error
H_1 true	Type II error	✓

Functionally, in one dimension, compute some statistic $T = T(\vec{X})$. Then if $T \in R \underbrace{\{|t| > c\}}_{\text{for example}}$, reject H_0.

Chief challenges: ① Find an appropriate test statistic T

② How to choose R, the rejection region?

or equivalently c, sometimes called "the critical value".

<u>Side remark</u> Often it is much more useful to instead compute point estimates or confidence intervals than to do hypothesis testing.

<u>Ex</u>: H_0: mean height of NYU students is $6'$

H_1: H_0^c

Compute $\hat{\mu} = \frac{1}{n} \sum X_i$. Reject H_0 if $|\hat{\mu} - 6'| > c$.

Things to be concerned with when doing hypothesis testing:

<u>Power function</u> of a hypothesis test

$$\beta(\theta) = \mathbb{P}_\theta(X \in R) \quad \leftarrow \text{function of } \theta.$$

$$= \int_R f(x;\theta)\,dx.$$

The <u>size</u> of a test is

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

A test has <u>level α</u> if its size is less than or equal to α.

<u>Types of hypotheses</u>

$\theta = \theta_0$: simple hypothesis

$\theta > \theta_0$: composite hypothesis

<u>Types of tests</u>

$$\left.\begin{array}{l} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{array}\right\} \text{two-sided test}$$

$$\left.\begin{array}{l} H_0: \theta \leq \theta_0 \\ H_0: \theta > \theta_0 \end{array}\right\} \text{one-sided test.}$$

Example: $X_1, \dots, X_n$ IID $N(\mu, \sigma^2)$ ← σ^2 known.

$H_0 : \mu \leq 0 \qquad \Theta_0 = (-\infty, 0]$

$H_1 : \mu > 0 \qquad \Theta_1 = (0, \infty)$

Pick a test statistic: $T = T(\vec{X}) = \frac{1}{n} \sum X_i = \bar{X}$.

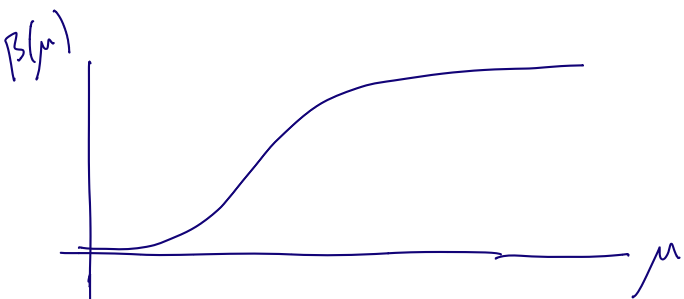
Reject H_0 if $T > c$, i.e. R = rejection region

$= \{ (X_1, \dots, X_n) : T(X_1, \dots, X_n) > c \}$

$$\text{Power } \beta(\mu) = \mathbb{P}_\mu(\bar{X} > c)$$

$$= \mathbb{P}_\mu \left(\underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{N(0,1)} > \frac{c - \mu}{\sigma/\sqrt{n}} \right)$$

$$= \mathbb{P}_\mu \left(Z > \frac{c-\mu}{\sigma/\sqrt{n}} \right) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma} \right)$$



$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0)$$

$$= 1 - \Phi\left(\frac{\sqrt{n}\, c}{\sigma} \right) = \alpha$$

Determine c based on your choice of α.

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}\, c}{\sigma} \right) \implies c = \frac{\sigma}{\sqrt{n}} \Phi(1 - \alpha).$$

$\implies$ Reject H_0 when $T = \bar{X} > c$, i.e. when the data are unlikely if H_0 is actually true.

The idea of a "most powerful" test (for fixed α):

We want to reject H_0 whenever H_1 is true.

$\Rightarrow$ so, if $\theta \in \Theta_1$, we want to reject H_0 as often as possible:

$\Rightarrow$ We want to maximize $\beta(\theta) = P_\theta(\vec{X} \in R)$ when $\theta \in \Theta_1$.

Note This is in contrast to the size, which is the maximum power under the null hypothesis. (which should be small.)

"Most powerful" tests often do not exist, or are virtually impossible to determine.