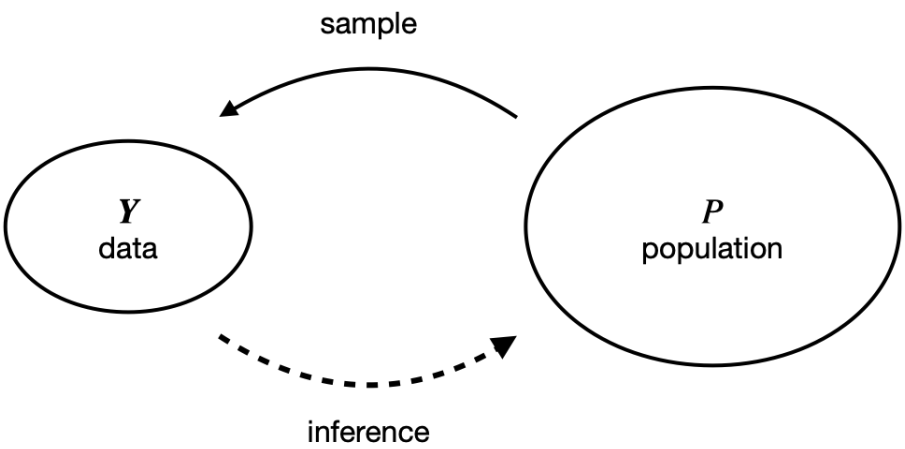
52309: Statistical Theory for CS

1. Introduction

The audience sitting in this class should be familiar by now with the general paradigm in statistics: basically, we use *data* $\boldsymbol{Y}$ to inform decisions about an unknown *population* P, from which $\boldsymbol{Y}$ is assumed to be a random sample. For example, estimating the 90th percentile of the distribution of number of calls that a call center receives per minute; or estimating the average decrease in blood pressure of a 1mg increase in the doze of a designated medication.



This course is basically an introduction to classic mathematical theory of statistics. We will learn first how to *formulate* a statistical problem mathematically, and then see general principles for designing "good solutions", i.e., constructing statistical procedures that make efficient use of the data when providing inference about the population.
Formulating a statistical problem begins with proposing a *model* for the data. We will generally assume that the observed data is a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ taking values in some sample space $\boldsymbol{\mathscr{Y}} := \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_n$. As usual, we

distinguish notationally between the random vector $\boldsymbol{Y}$ and a particular realization $\boldsymbol{y}$.

Definition. A *statistical model is* a family

$$\mathscr{F} = \{f_\theta : \theta \in \Theta\}$$

of distributions on $\mathscr{Y}$, where Θ is the index set of the family, i.e., labels $\theta \in \Theta$ index the different members in the family.

Comment. By referring to f_θ as a *distribution* we mean that it determines the probability measure (=probability law) on $\mathscr{Y}$. As we know, there are different ways to represent this probability measure. For example, if $Y_i \sim \text{Ber}(\theta), \quad \theta \in \Theta = (0,1)$, independently for $i = 1,...,n$, then, formally, the members f_θ of $\mathscr{F}$ can be taken as the p.m.f.'s $f_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$,

and if $Y_i \sim \mathscr{N}(\theta,1)$, independently for $i = 1,...,n$, then the members f_θ of $\mathscr{F}$ can be taken as the p.d.f.'s $f_\theta(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i-\theta)^2}{2\sigma^2}}$. These will generally

be our "natural" choices to describe members of the model $\mathscr{F}$. But keep in mind that an equivalent way to describe a distribution is by its c.d.f., or its m.g.f. (whenever it exists). The model $\mathscr{F}$ is of course the same no matter what representation we choose for f_θ.

The basic paradigm assumes that

$$\boldsymbol{Y} \sim f_\theta(\boldsymbol{y}),$$

for some unknown ("true") value θ. In other words, we assume that the observed data is a random realization from some *unknown* member in $\mathscr{F}$.

In general, Θ can be an arbitrary set. If Θ has finite dimension, for example when $\Theta = \mathbb{R}^d$ for some integer d, the family is called *parametric*, otherwise the family is said to be *nonparametric*. An example of a parametric family is all p-dimensional normal distributions with known covariance, $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^p$, or unknown covariance, $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times S_+^p$, where S_+^n denotes the set of all $p \times p$ positive-definite matrices. An example of a nonparametric family of distributions is the set of all distributions on the interval $[0,1]$ with finite variance, or (in a regression problem with fixed X_i's) the set of all distributions corresponding to $Y_i = f(X_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0,1)$ where $f : \mathbb{R} \to \mathbb{R}$ has at least K derivatives ("smooth").
In this course we concentrate on *parametric* families $\mathcal{F}$, in which case θ is the *parameter* indexing the family, and Θ is the corresponding *parameter space*.

Example. I'm observing the number of successes in n trials of the same experiment. I'm interested in the true probability of success.
Here it often makes sense to model the trials as *independent* (and identically distributed), in which case the model would be
$\mathcal{F} = \{\Pi_{i=1}^n f_p(y_i) : p \in [0,1]\}$,
where $f_p(y_i) = y_i^p(1-y_i)^{1-p}$, $y = 0,1$, is the p.m.f. of a Bernoulli r.v. with probability $p$for "success" ($y_i = 1$). The parameter of the family is p, and the parameter space is $\Theta = [0,1]$. We usually abbreviate this as

$$Y_i \overset{iid}{\sim} \text{Bern}(p), \qquad i = 1,\ldots,n.$$

Example. Bob goes to the supermarket to buy cheese. He sees on the shelf several packages of the same product, each indicating weight of 200g. Bob

wants to check if 200g indeed reflects the “true weight” of the product, or the brand is cheating. He picks out 5 random packages of the product to purchase, returns home and weighs the 5 items. The results were: 200.3, 195.0, 192.4, 205.2, 190.1g. What should Bob conclude?
Of course, it makes no sense to expect that any particular product weighs exactly 200g, but instead that each outcome is an independent noisy version of some “true” weight. In other words, to model the 5 outcomes as realizations of i.i.d. random variables

$$Y_i = \mu + \epsilon_i, \qquad i = 1,...,5,$$

with μ representing the mean (“true” weight) and ϵ_i being i.i.d. random variables with mean zero. In many cases it makes sense to further assume that ϵ_i are (approximately) normally distributed with (known or unknown) variance σ^2, hence $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, i.i.d. for $i = 1,...,5$. Now that we have a model on the data, we can formulate Bob’s question as a problem of testing the null hypothesis $H_0 : \mu = 200$ against $H_1 : \mu < 200$.

<u>Likelihood</u>

Assume a parametric model, $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, so that

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n) \sim f_\theta(\boldsymbol{y})$$

for some unknown $\theta \in \Theta$.

Definition (likelihood function). Denote the observed realization of $\boldsymbol{Y}$ by $\boldsymbol{y}$. The *likelihood function* is the function L defined on Θ and given by

$$L(\theta; \boldsymbol{y}) = f_\theta(\boldsymbol{y}) \quad \text{for each } \theta \in \Theta. \qquad (1)$$

For simplicity we will sometimes write just $L(\theta)$.

Comment. We need to clarify what we mean by $f_\theta(\boldsymbol{y})$, because we said before that, in general, f_θ can be any representation of the distribution of $\boldsymbol{Y}$ under θ, so $f_\theta(\boldsymbol{y})$ might not even be well defined if f_θ is not a function of $\boldsymbol{y} \in \mathscr{Y}$ (e.g., if f_θ is the m.g.f. of $\boldsymbol{Y}$ under θ). For our purposes adopting the following convention will be general enough: if f_θ is a continuous distribution for all $\theta \in \Theta$, then $f_\theta(\boldsymbol{y})$ in the definition (1) of the likelihood is understood as the p.d.f.; if f_θ is a discrete distribution for all $\theta \in \Theta$, then $f_\theta(\boldsymbol{y})$ in the definition (1) is understood as the p.m.f. And if f_θ is a mixture of continuous and discrete distributions (i.e. f_θ is comprised of a discrete part and a continuous part) with $\mathscr{Y}_d$ representing the support of the discrete part and $\mathscr{Y}_c$ the support of the continuous part, then $f_\theta(\boldsymbol{y})$ in the definition (1) is understood as the p.d.f. if $\boldsymbol{y} \in \mathscr{Y}_c$ and as the p.m.f. if $\boldsymbol{y} \in \mathscr{Y}_d$.

Hence, the likelihood function looks at $f_\theta(\boldsymbol{y})$ as a function of θ, where $\boldsymbol{y}$ is fixed to the values we happened to observe for $\boldsymbol{Y}$. If $\boldsymbol{Y}$ is discrete, for example, then for any fixed θ, we know that $f_\theta(\boldsymbol{y})$ sums to 1 over all possible $\boldsymbol{y}$'s; we cannot say the same when treating $f_\theta(\boldsymbol{y})$ as a function of θ for fixed $\boldsymbol{y}$ —there is no claim that the sum (or integral, in the continuous case) over $\theta \in \Theta$ equals 1. Still, for any candidate θ, the likelihood $L(\theta; \boldsymbol{y})$ is a measure of *agreement/fidelity* between f_θ and the observed data. Note that specifying the likelihood function is equivalent to specifying the (parametric) statistical model.

We will often assume an *i.i.d. model*, i.e., that the coordinates Y_i of $\boldsymbol{Y}$ are independent and identically distributed random variables. To be consistent

with the above the model should be a family of distributions f_θ on $\boldsymbol{\mathcal{Y}} = \mathcal{Y}^n$, but in the iid case this is equivalent to modeling a *single* coordinate Y_i.
Hence,

$$Y_i \overset{iid}{\sim} f_\theta(y_i), \quad i = 1,...,n,$$

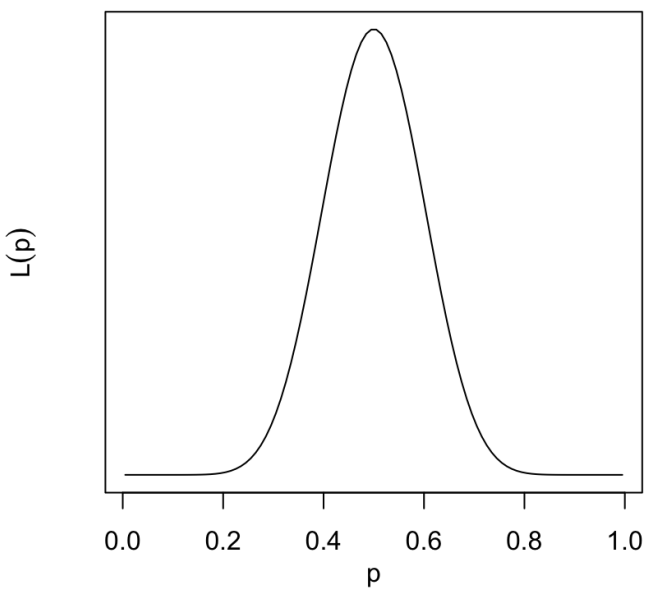
where $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is a family of distributions on $\mathcal{Y}$ (note that, with some abuse of notation, we still write f_θ for the *marginal* distribution of Y_i).
Under an iid model, the likelihood function is

$$L(\theta; \boldsymbol{y}) = f_\theta(\boldsymbol{y}) = \Pi_{i=1}^n f_\theta(y_i).$$

Example. A hospital had $n = 20$ babies delivered on a certain day. The gender of each of the babies was recorded, giving the sequence:
M, F, M, M, M, F, F, F, F, M, F, M, F, F, F, F, F, M, F, M.
In total, 12 female babies and 8 male babies. We are interested in the chance (probability) that a newborn baby is a girl
If we let $Y_i = 1$if the ith baby is a girl and $Y_i = 0$ otherwise, then it makes sense to model

$Y_i \overset{iid}{\sim}$ Binom$(1,p)$, $\quad i = 1,...,n,$

where p is the chance of giving birth to a girl. We can now compute the likelihood function:



$$
\begin{aligned}
L(p;\boldsymbol{y}) &= \Pr(Y_1 = y_1) \times \Pr(Y_2 = y_2) \times \Pr(Y_3 = y_3) \times \cdots \times \Pr(Y_{20} = y_{20}) \\
&= \Pr(Y_1 = 0) \times \Pr(Y_2 = 1) \times \Pr(Y_3 = 0) \times \cdots \times \Pr(Y_{20} = 0) \\
&= (1-p) \times p \times (1-p) \times \cdots \times (1-p) \\
&= p^{\sum y_i}(1-p)^{n-\sum y_i} \\
&= p^{12}(1-p)^8
\end{aligned}
$$

Example (previous example continued). For Bob's cheeses experiment, it made sense to model the data as i.i.d. continuous random variables,

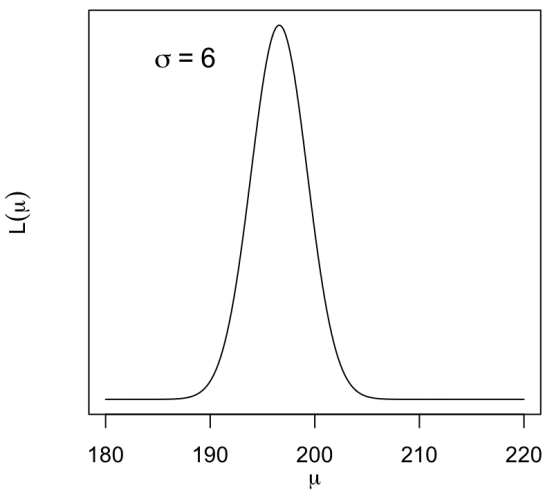
$$Y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad i = 1,\ldots,n = 5,$$

and suppose for now that σ^2 is known. Then $\mathcal{F} = \{\varphi_{\mu,\sigma^2} : \mu \in \mathbb{R}\}$ where φ_{μ,σ^2} is the density of a $\mathcal{N}(\mu, \sigma^2)$ r.v., and the likelihood is

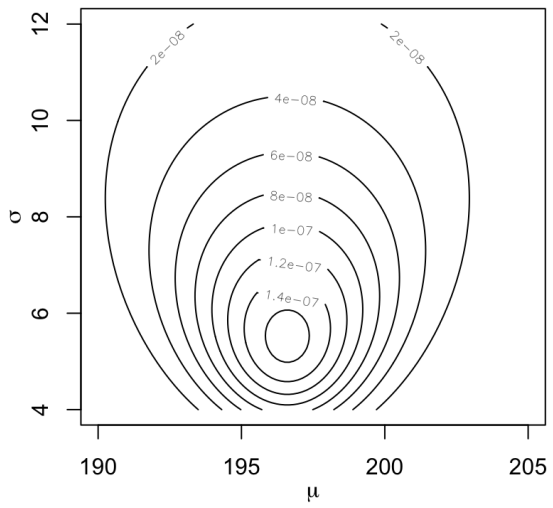
$$L(\mu;\boldsymbol{y}) = \Pi_{i=1}^n \phi_{\mu,\sigma^2}(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp^{-\frac{(y_i-\mu)^2}{2\sigma^2}} =$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^5 \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^5 (y_i-\mu)^2}$$

The following figure plots $L(\mu;\boldsymbol{y})$ for $\sigma = 6$:



If σ is unknown, the likelihood in (2) will be viewed as a function of (μ, σ):



Example (linear regression with fixed design). $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times p}$ is a fixed matrix, and where $\theta = (\boldsymbol{\beta}, \sigma^2)$ unknown. The likelihood is

$$L\big(\boldsymbol{\beta}, \sigma^2; (\boldsymbol{y})\big) = f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2\right\}$$

Note that here the likelihood corresponding to each of the (independent) y_i's depends on (σ^2 and) the *same* p-dimensional vector $\boldsymbol{\beta}$, although these "individual" likelihoods are not the same because of the $\mathbf{x}_i$'s.

<u>Identifiability</u>

Before we address how to carry out statistical inference for θ, we need to make sure that the problem is well-defined.

*Definition (identifiability)***.** A model $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ is identifiable if

$$f_\theta = f_{\theta'} \implies \theta = \theta' \quad \text{for all } \theta, \theta' \in \Theta$$

In other words, if the map $\theta \mapsto f_\theta$ is one-to-one.

Informally, identifiability ensures that, if we were able to observe an infinite number of realizations of the experiment (the entire vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$), then we would be able to tell what θ is without ambiguity. Note: non-identifiability is a degeneracy of the *model itself*; you can think of it as the case where, even if we observed infinite number of samples, we wouldn't be able to tell the true value of the parameter).

We sometimes speak of *identifiability of* θ rather than identifiability of the model $\mathscr{F} = \{f_\theta(\boldsymbol{y}) : \theta \in \Theta\}$; the meaning is the same.

Example. $Y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, $i = 1,\ldots,n$, σ^2 known or unknown. Then μ is identifiable. It is enough to consider $n = 1$ (prove this!). Let θ represent the unknown parameters, ie $\theta = \mu$ or $\theta = (\mu, \sigma)$. Then

$$f_\theta = f_{\theta'} \;\Rightarrow\; \underbrace{E_{f_\theta}[Y]}_{=\mu} = \underbrace{E_{f_{\theta'}}[Y]}_{\mu'},$$

so μ is identifiable. Also,

$$f_\theta = f_{\theta'} \;\Rightarrow\; \underbrace{E_{f_\theta}[Y^2]}_{=\sigma^2+\mu^2} = \underbrace{E_{f_{\theta'}}[Y^2]}_{=(\sigma')^2+(\mu')^2} \;,$$

so σ is also identifiable (note that this is the same as saying that σ^2 is identifiable, because $\sigma \mapsto \sigma^2$ is one-to-one on $[0,\infty)$).

In fact, the last example is an application of a more general result:

Lemma. The model is identifiable if there exists a function $t(\boldsymbol{y})$ (of the data $\boldsymbol{y}$ only) such that the mapping $\theta \mapsto E_\theta[T(\boldsymbol{Y})]$ is one-to-one.

Proof. Take any $\theta, \theta' \in \Theta$, and assume that $f_\theta = f_{\theta'}$. Then, in particular, it must be that $E_\theta[T(\boldsymbol{Y})] = E_{\theta'}[T(\boldsymbol{Y})]$. But by assumption this implies $\theta = \theta'$.

■

Example. Let $Z_i \overset{iid}{\sim} \mathcal{N}(\mu,1)$, but we observe only $Y_i = Z_i^2$, $i = 1,...,n$. Then μ is not identifiable, because $\mu' = -\mu$ gives the same likelihood (we can again reason for a single observation Y_i).

Example. $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. Two values $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^{n\times p}$ give the same distribution iff $\boldsymbol{X\beta} = \boldsymbol{X\beta}'$. Hence, for $\boldsymbol{\beta}$ to be identifiable, we need that for any $\boldsymbol{\beta}, \boldsymbol{\beta}'$, it holds that $\boldsymbol{X\beta} = \boldsymbol{X\beta}' \;\Rightarrow\; \boldsymbol{\beta} = \boldsymbol{\beta}'$, equivalently, $\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}') = \boldsymbol{0} \;\Rightarrow\; \boldsymbol{\beta} - \boldsymbol{\beta}' = \boldsymbol{0}$. This is the case iff $\boldsymbol{X}$ has full column rank (kernel is trivial). In particular, $\boldsymbol{\beta}$ cannot be identifiable if $p > n$.

<u>Sufficiency</u>

*Definition***.** A function $T(\boldsymbol{y})$ of the data only (i.e., it involves no unknown parameters) is called a *statistic.*

A statistic can be of any dimension. Examples: sample mean $\overline{Y}_n$; sample mean and standard error $(\overline{Y}_n, S_n)$. In the example with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, i.i.d., $T = Y_i - \mu$ is not a statistic, and $T = \sqrt{n}\overline{Y}_n/\sigma$ is a statistic only if σ is known.

Definition. A statistic $T(\boldsymbol{y})$ is *sufficient* (for θ) if the conditional distribution of $\boldsymbol{Y}$ given $T(\boldsymbol{Y})$ does not depend on θ.

Informally, a sufficient statistic captures all of the information that the raw data $\boldsymbol{y}$ holds about θ. In other words, I lose no information on θ if you tell me $T(\boldsymbol{y})$ instead of the entire vector $\boldsymbol{y}$.

Example. *In the baby births example, we calculated before:*

$$L(p; \boldsymbol{y}) = p^{\sum y_i}(1-p)^{n-\sum y_i},$$

i.e., the likelihood depends only on the number of female (alternatively, the number of male) births. Therefore, any changes to $\boldsymbol{y}$ that keep $T(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$ constant, will have the same likelihood. It makes sense to expect that this statistic is sufficient; let's check with the definition: for any $\boldsymbol{y} = (y_1, \ldots, y_n)$ and any $t = 0,1,...,20$:

$$\begin{aligned}
\Pr(\boldsymbol{Y} = \boldsymbol{y} \mid T(\boldsymbol{Y}) = t) &= \frac{\Pr(\boldsymbol{Y} = \boldsymbol{y}, T(\boldsymbol{Y}) = t)}{\Pr(T(\boldsymbol{Y}) = t)} \\
&= \begin{cases} \frac{\Pr(\boldsymbol{Y} = \boldsymbol{y})}{\Pr(T(\boldsymbol{Y}) = t)}, & \text{if } T(\boldsymbol{y}) = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{p^{T(\boldsymbol{y})}(1-p)^{n-T(\boldsymbol{y})}}{\Pr(T(\boldsymbol{Y}) = t)}, & \text{if } T(\boldsymbol{y}) = t \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}}, & \text{if } T(\boldsymbol{y}) = t \\ 0, & \text{otherwise} \end{cases} \qquad = \begin{cases} \frac{1}{\binom{n}{t}}, & \text{if } T(\boldsymbol{y}) = t \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}$$

which does not depend on p, hence $T(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$ is sufficient.

The following theorem gives a general result.

Theorem (Fisher–Neyman factorization theorem). Let $\boldsymbol{Y} \sim f_\theta(\boldsymbol{y})$ and denote by $L(\theta; \boldsymbol{y})$ the corresponding likelihood. Then:

$T(\boldsymbol{y})$ is sufficient $\iff L(\theta; \boldsymbol{y}) = g(\theta; T(\boldsymbol{y})) \cdot h(\boldsymbol{y})$ for some function g, h.

Proof. It will be convenient notationally to prove for the discrete case, i.e., when $f_\theta(\boldsymbol{y})$ is a probability mass function. The proof for the continuous case is the same.

(<=) Assume $L(\theta; \boldsymbol{y}) = g(\theta; T(\boldsymbol{y})) \cdot h(\boldsymbol{y})$. Then

$$
\begin{aligned}
\mathrm{Pr}_\theta(\boldsymbol{Y} = \boldsymbol{y} \mid T(\boldsymbol{Y}) = t) &= \frac{\mathrm{Pr}_\theta(\boldsymbol{Y} = \boldsymbol{y}, T(\boldsymbol{Y}) = t)}{\mathrm{Pr}_\theta(T(\boldsymbol{Y}) = t)} \\
&= \begin{cases} \dfrac{\mathrm{Pr}_\theta(\boldsymbol{Y} = \boldsymbol{y})}{\mathrm{Pr}_\theta(T(\boldsymbol{Y}) = t)}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{g(\theta; T(\boldsymbol{y}) h(\boldsymbol{y})}{\int g(\theta; T(\boldsymbol{y}') h(\boldsymbol{y}') 1(T(\boldsymbol{y}') = t) d\boldsymbol{y}'}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{\mathrm{Pr}_\theta(\boldsymbol{Y} = \boldsymbol{y})}{\mathrm{Pr}_\theta(T(\boldsymbol{Y}) = t)}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{g(\theta; T(\boldsymbol{y}) h(\boldsymbol{y})}{\int g(\theta; T(\boldsymbol{y}') h(\boldsymbol{y}') 1(T(\boldsymbol{y}') = t) d\boldsymbol{y}'}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{g(\theta; t) h(\boldsymbol{y})}{\int g(\theta; t) h(\boldsymbol{y}') 1(T(\boldsymbol{y}') = t) d\boldsymbol{y}'}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{h(\boldsymbol{y})}{\int h(\boldsymbol{y}') 1(T(\boldsymbol{y}') = t) d\boldsymbol{y}'}, & \text{if } t = T(\boldsymbol{y}) \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}
$$

which does not depend on θ.

(=>) Assume that for any $\boldsymbol{y}, t$, it holds that $\Pr_\theta(\boldsymbol{Y} = \boldsymbol{y} \mid T(\boldsymbol{Y}) = t)$ does not depend on θ. Then:

$$\begin{aligned} L(\theta; \boldsymbol{y}) = \Pr_\theta(\boldsymbol{Y} = \boldsymbol{y}) &\overset{(1)}{=} \Pr_\theta(\boldsymbol{Y} = \boldsymbol{y}, T(\boldsymbol{Y}) = T(\boldsymbol{y})) \\ &= \underbrace{\Pr_\theta(T(\boldsymbol{Y}) = T(\boldsymbol{y}))}_{g(\theta; T(\boldsymbol{y}))} \underbrace{\Pr_\theta(\boldsymbol{Y} = \boldsymbol{y} \mid T(\boldsymbol{Y}) = T(\boldsymbol{y})}_{h_\theta(\boldsymbol{y})} \\ &\overset{(2)}{=} g(\theta; T(\boldsymbol{y})) \cdot h(\boldsymbol{y}) \end{aligned}$$

where (1) is because $\boldsymbol{Y} = \boldsymbol{y}$ implies $T(\boldsymbol{Y}) = T(\boldsymbol{y})$, and where in (2) we used the assumption that $h_\theta(\boldsymbol{y})$ does not depend on θ. ■

Example. $Y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, $i = 1,...,n$, both μ and σ^2 unknown. We have

$$L(\mu, \sigma; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi} \cdot \sigma} \right)^n \cdot e^{-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}}.$$

We can write:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \mu)^2 = \sum_{i=1}^{n} \Big((y_i - \bar{y}) + (\bar{y} - \mu) \Big)^2 &= \sum_{i=1}^{n} (y_i - \bar{y})^2 + 2(\bar{y} - \mu) \underbrace{\sum_{i=1}^{n} (y_i - \bar{y})}_{= \sum_{i=1}^{n} y_i - n\bar{y} = 0} + n \cdot (\bar{y} - \mu)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 + n \cdot (\bar{y} - \mu)^2, \end{aligned}$$

so that

$$
\begin{aligned}
L(\mu, \sigma) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sigma^n} \cdot \exp\left(-\frac{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2 + n \cdot (\bar{y} - \mu)^2}{2\sigma^2}\right) \\
&= g\left(\mu, \sigma;\ \bar{y}, \sum_{i=1}^n \left(y_i - \bar{y}\right)^2\right).
\end{aligned}
$$

Hence, $T(\boldsymbol{y}) = \left(\bar{y}, \sum_{i=1}^n \left(y_i - \bar{y}\right)^2\right)$ is a sufficient statistic for $\theta = (\mu, \sigma^2)$.

Note that every one-to-one transformation of a sufficient statistic $T(\boldsymbol{y})$ is also sufficient. For instance, in the last example $\tilde{T}(\boldsymbol{y}) = \left(\bar{y}, \sum_{i=1}^n y_i^2\right)$ is also sufficient, as $T(\boldsymbol{y}) \mapsto \tilde{T}(\boldsymbol{y})$ is one-to-one.