

# Frequentist Inference

Muchang Bahng

Winter 2022

## Contents

<b>1</b>	<b>Statistical Decision Theory</b>	<b>2</b>
1.1	Statistical Models . . . . .	2
1.2	Statistics and Sufficiency . . . . .	6
1.3	Model Equivalence . . . . .	9
1.4	Exponential Families . . . . .	9
1.5	Decision Problems . . . . .	11

In statistics, we are given some data  $\mathcal{D} = \{x_i\}_{i=1}^n$ . The simplest thing we can do is summarize this data by extracting some nice characteristics—for example, the mean. This is known as **descriptive statistics**. In **inferential statistics**, we have much stronger assumptions. We assume that that data are realizations of random variables following a joint probability distribution. Sometimes, we may assume that these **samples** are iid coming from  $\mathbb{P}^*$ , known as the *true data generating distribution* (and sometimes known as the *population* in survey statistics or causal inference). As the name suggests, we must infer from  $\mathcal{D}$  what  $\mathbb{P}$  is. This immediately raises some questions: How should we interpret the population? What are we inferring? And how does this process work? Let's establish this confusion with an example.

### Example 0.1 (Measurement Problem)

Say we have a dataset consisting of real-valued measurements  $x_1, \dots, x_n$  to estimate some quantity  $\theta$ . We may try to summarize the mean of this data by computing

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (1)$$

This seems so common and intuitive that we might forget why this specific formula works. Two nice properties are:

1. It minimizes the sum of least squares

$$\bar{x} = \underset{a}{\operatorname{argmin}} \sum_{i=1}^n (x_i - a)^2 \quad (2)$$

2. The value  $\bar{x}$  makes the sum of the residuals to be 0.

These two properties land on the level of descriptive statistics. They describe the mean as a reasonable descriptive measure of the center of the observations, but they cannot justify  $\bar{x}$  as an estimate of the true value  $\theta$  since no explicit assumption has been made connecting the observations  $x_i$  with  $\theta$ .

To do inference, we can furthermore assume that the  $x_i$  are observed values of  $n$  independent random variables which have a common distribution depending on  $\theta$ . Which assumptions we make will determine which estimators are reasonable. Here are two cases in which means are not a reasonable estimate.

1. We assume that  $x_i = \theta + \epsilon_i$  where  $\epsilon_i$  satisfies  $\mathbb{P}(\epsilon_i < 0) = \mathbb{P}(\epsilon_i > 0)$ .
2. *Larger samples may not improve estimate.* If the  $x_i$  turns out to have finite variance the variance of the mean is  $\sigma^2/n$ . However, if the  $x_i$ 's have a Cauchy distribution, then the distribution of  $\bar{x}$  is the same as  $x$ , so nothing is gained by taking more measurements.

To answer the first question, the population is usually introduced as some finite true distribution of some quantity, but more often it is treated as an abstract data generating distribution. For example, say that we have a large barrel of grains, and we take a random sample of 100 grains and measure their weight. Though we can spend much more effort and time weighing every single grain in the barrel, for practical reasons we want to work with the sample. On the other hand, think of the distribution of facial features of humanity. We may assume that every time a human is born, we can think of it being sampled from some abstract distribution (specified by “God”), and so even taking all humans in the world is still a sample of this population.

## 1 Statistical Decision Theory

### 1.1 Statistical Models

In probability, we implicitly define a probability space  $(\Omega, \mathcal{F}, \mathbf{P})$  and work explicitly with the random variable  $X : \Omega \rightarrow \mathcal{X}$ , which represents the “data before it is observed.” Therefore, when introducing the definition of the statistical model, the random variable  $X$  is always implicitly defined.

**Definition 1.1 (Statistical Model)**

Let  $(\mathcal{X}, \mathcal{X})$  be a measurable space.

1. A family of distributions  $\mathcal{P}$  over  $\mathcal{X}$  is called a **statistical model**.<sup>a</sup> It is sometimes written  $(\mathcal{X}, \mathcal{X}, \mathcal{P})$  to make the sample space more explicit.
2. Usually, we accompany  $\mathcal{P}$  with a **parameter space**  $\Theta$ , along with a surjective map  $\theta \rightarrow \mathbf{P}_\theta$ .

<sup>a</sup>Note that this is *not* a probability space! For each  $\mathbf{P} \in \mathcal{P}$ , the triplet  $(\mathcal{X}, \mathcal{X}, \mathbf{P})$  is a probability space.

**Example 1.1**

Suppose we want to model the net worth of American adults. We can model it as  $\mathcal{X} = \mathbb{R}$ . Conventionally, we can just take the Borel  $\sigma$ -algebra  $\mathcal{B}(\mathcal{X})$ . As for the family of distributions, we have a few options.

1. *All*. We set  $\mathcal{P}$  to be *all* probability measures on  $\mathbb{R}$ . This is a huge family and has no obvious parameter space.
2. *Gaussian with Fixed Variance*. We set  $\mathcal{P}$  to be all Gaussian distributions with variance 1. This is written

$$X \sim N(\theta, 1) \text{ for } \theta \in \mathbb{R} = \Theta \quad (3)$$

3. *Gaussian*. We set  $\mathcal{P}$  to be all Gaussian distributions. Then, we can write

$$X \sim N(\mu, \sigma^2) \text{ for } (\mu, \sigma^2) \in \mathbb{R} \times [0, +\infty) = \Theta \quad (4)$$

The reason we want a parameter space is that we want to introduce some extra structure on  $\mathcal{P}$ , such as notions of orderings, metrics, or operations. Common choices of  $\Theta$  are subsets of vector spaces  $\mathbb{R}^n$  or function spaces such as  $L^p$ . Ideally, we want  $\Theta$  to be a set that *indexes* the set  $\mathcal{P}$ . Such models are called *identifiable*.

**Definition 1.2 (Identifiability)**

A statistical model is **identifiable by  $\Theta$**  if there exists a bijection between  $\mathcal{P}$  and  $\Theta$ . In this case, we call it a **statistical experiment** and write as shorthand  $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ .<sup>a</sup>

<sup>a</sup>Note that you can always create a trivial identification by setting  $\Theta = \mathcal{P}$ . However, this doesn't give us anything, and we would like to ideally exploit the structure of  $\Theta$ .

Let's take a look at some models.

**Example 1.2**

We write some common notations for shorthand.

1. *Random Variable*. Say  $X$  is real-valued. Then,  $X \sim N(\theta, 1)$  for  $\theta \in \Theta$  is shorthand for

$$(\mathcal{X} = \mathbb{R}, \mathcal{X} = \mathcal{B}(\mathbb{R}), \mathcal{P} = \{N(\theta, 1) : \theta \in \Theta\}) \quad (5)$$

2. *Joint Independent Random Variables*.  $X_1 \dots X_n \stackrel{\text{iid}}{\sim} N(\theta, 1)$ ,  $\theta \in \Theta$  is shorthand for

$$(\mathcal{X} = \mathbb{R}^n, \mathcal{X} = \mathcal{B}(\mathbb{R}^n), \mathcal{P} = \{N(\theta, 1)^{\otimes n} : \theta \in \Theta\}) \quad (6)$$

3. *Regression*. Let  $X \in [-1, 1]$  and  $Y \in \mathbb{R}$ .  $Y = f(X) + \epsilon$ ,  $\epsilon \sim N(0, 1)$  for  $f \in \mathcal{F}$ , where  $\mathcal{F}$  is some class of functions. This is shorthand for

$$(\mathcal{X} = [-1, 1] \times \mathbb{R}, \mathcal{X} = \mathcal{B}([-1, 1] \times \mathbb{R}), \mathcal{P} = \{\mathbf{P}_{XY} : \mathbf{P}_X = U[-1, 1], P_{Y|X, f} = N(f(X), 1), f \in \mathcal{F}\}) \quad (7)$$

In other words, it is the set of probability measures such that the  $X$ -marginal is uniform and the conditional distribution of  $Y$  given  $x$  is  $N(f(x), 1)$ .

4. *Regression with Fixed Design.* Say that we have data  $\{(x_i, y_i)\}_{i=1}^n$ , and we are interested in the *conditional distribution*  $Y \mid X = x$ . We can treat  $x_i$ 's as fixed and model the  $Y_i$ 's as being generated by the following:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (8)$$

where  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1)$  and  $f \in \mathcal{F}$  for some function class  $\mathcal{F}$ . Therefore, to model the joint distribution, we have  $\mathcal{X} = \mathbb{R}^n$  and  $\mathcal{Y} = \mathcal{B}(\mathbb{R}^n)$ . Our statistical model is

$$\mathcal{P} := \left\{ \mathbf{P}_f = \bigotimes_{i=1}^n N(f(x_i), 1) : f \in \mathcal{F} \right\} \quad (9)$$

Naturally, we want to work with densities for each measure in the model, but this requires the measure to be absolutely continuous w.r.t. another measure in order for us to take the Radon-Nikodym derivative. Therefore, the following definition is natural and sets up things nicely.

**Definition 1.3 (Dominated Families of Measures)**

When all distributions  $\mathbf{P} \in \mathcal{P}$  are absolutely continuous w.r.t. measure  $\mu$ , then we say that the family  $\mathcal{P}$  is **dominated (by  $\mu$ )**.

In general, there are two types of models that we consider.

1. Consider the model  $(\mathcal{X}, \mathcal{Y}, \mathcal{P})$ . In a discrete statistical model, we consider at most countable  $X$  and  $\mathcal{P}$  dominated by the counting measure  $c$ .
2. Consider the model  $(\mathcal{X}, \mathcal{Y}, \mathcal{P})$  for some subset  $X \subset \mathbb{R}^n$ . In a *continuous statistical model*,  $\mathcal{P}$  is dominated by the Lebesgue measure over  $(\mathcal{X}, \mathcal{Y})$ .

If a model is dominated by  $\sigma$ -finite measure  $\mu$ , then by the Radon-Nikodym theorem, we have a nonnegative measurable function  $p = \frac{d\mathbf{P}}{d\mu}$  s.t. for all  $A \in \mathcal{Y}$ ,

$$\mathbf{P}(A) = \int_A p d\mu \quad (10)$$

Since we will often work with parameterized families  $\mathbf{P}_\theta$ , their density will be denoted  $p(\cdot \mid \theta)$  or  $p_\theta(\cdot)$ .

**Example 1.3 (Discrete Models)**

Consider when  $X = \{1, \dots, 6\}$ ,  $\mathcal{Y} = 2^X$ , and let's look at the probability measure, which is completely defined by

$$\mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = \mathbf{P}(\{3\}) = \frac{1}{3} \quad (11)$$

Then, the density is

$$p_\theta(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1, 2, 3 \\ 0 & \text{if } x = 4, 5, 6 \end{cases} \quad (12)$$

and we can verify for example that

$$\frac{2}{3} = \mathbf{P}(\{1, 2\}) = \int_{\{1, 2\}} p_\theta(x) dc(x) \quad (13)$$

$$= p_\theta(1)c(\{1\}) + p_\theta(2)c(\{2\}) \quad (14)$$

$$= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{2}{3} \quad (15)$$

**Example 1.4 (Absolutely Continuous Models)**

Note that we can also change the dominating measure  $\mu$ .

**Example 1.5**

Consider the previous example but now consider the dominating measure  $c' = 2c$ . Then, the Radon-Nikodym derivative is

$$\frac{d\mathbf{P}}{dc'}(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3 \\ 0 & \text{if } x = 4, 5, 6 \end{cases} \quad (16)$$

And there is nothing wrong with this since

$$\frac{2}{3} = \mathbf{P}(\{1, 2\}) = \int_{\{1, 2\}} p_{\theta}(x) dc(x) \quad (17)$$

$$= p_{\theta}(1)c(\{1\}) + p_{\theta}(2)c(\{2\}) \quad (18)$$

$$= \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 2 = \frac{2}{3} \quad (19)$$

So the choice of the dominating measure matters and actually influences our density function. However, there is clearly a canonical one: the counting measure and the Lebesgue measure. Furthermore, we can talk about half-discrete and half-continuous measure, where it isn't as obvious which dominating measure to choose. But note that since the sum of two  $\sigma$ -finite measures is  $\sigma$ -finite, such a construction is not that difficult. These measures do come up in practice, but you can worry about them when you actually do encounter them.

**Example 1.6 (Non-Identifiable Models)**

Most of the time, we will work with identifiable models, and it may seem like defining this surjective map  $\theta$  is overcomplicating things. However, in machine learning, it is often the case that we work with non-identifiable models. Sometimes, we may have  $\theta \neq \theta'$  yet  $\mathbf{P}_{\theta} = \mathbf{P}_{\theta'}$ .

**Example 1.7 (Linear Redundancy)**

Consider the regression statistical model again, but now we consider the class of linear functions  $\mathcal{F} = \{f(x) = (\theta_1 + \theta_2)x : \theta_1, \theta_2 \in \mathbb{R}\}$ , which leads to our model

$$Y_i = (\theta_1 + \theta_2)x_i + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, 1) \quad i = 1, \dots, n \quad (20)$$

Our statistical model is not identifiable since  $(\theta_1, \theta_2) = (1, 3)$  and  $(\theta_1, \theta_2) = (2, 2)$  both give  $(\theta_1 + \theta_2)x_i = 4x_i$ , and hence gives the same product measure

$$(\theta_1, \theta_2) \mapsto \mathbf{P}_f = \bigotimes_{i=1}^n N(4x_i, 1) \quad (21)$$

**Example 1.8 (Neural Networks)**

Let's look at something that is a bit less obvious.

## 1.2 Statistics and Sufficiency

### Definition 1.4 (Statistic)

A **statistic** is a measurable function  $T : (\mathcal{X}, \mathcal{X}) \rightarrow (\mathcal{T}, \mathcal{T})$ .

In particular, estimators are statistics.

Any statistic generates a new model under the pushforward measure. But we may have broken identifiability. The idea of sufficiency is when does  $T$  preserve all information on data? The key idea is that a statistic  $T$  is sufficient if—once we know  $T(X)$ —the remaining randomness in  $X$  tells us nothing further about which  $\mathbf{P}_\theta$  generated the data.

### Definition 1.5 (Sufficient Statistic)

Let  $(\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$  be an identifiable model. A statistic  $T$  is **sufficient for  $\mathcal{P}$**  if for all  $A \in \mathcal{X}$ ,  $\mathbf{P}_\theta(A \mid T)$  admits a version that does not depend on  $\theta$ .<sup>a</sup>

<sup>a</sup>Some technical measure theory stuff: If  $\mathcal{T}$  is nice (for example, Borel), then there exists a function  $h_{A,\theta} : T \rightarrow \mathbb{R}$  s.t.  $\mathbf{P}_\theta(A) = \mathbf{P}_\theta(A \cap T^{-1}(B)) = \int_B h_{A,\theta}(t) d\mathbf{P}_\theta^T(t)$ ,  $t \mapsto \mathbf{P}_\theta(A \mid T = t)$ .

Sufficiency is a property of how the parameter enters the distribution of the data. The same statistic may be sufficient for one model but another, and crucially depends on how the model is parameterized.

### Theorem 1.1

If  $T$  is invertible, with  $T^{-1}$  measurable, then  $T$  is sufficient.

*Proof.*  $\sigma(T) \subset X$ . Take  $A \in \mathcal{X}$ . Then,  $A = T^{-1}(T(A)) \in \sigma(T)$ , which implies that  $X \subset \sigma(T)$ . Therefore,  $X = \sigma(T)$ . Therefore,

$$\mathbb{E}_\theta[\mathbb{1}_A \mid \sigma(T)] = \mathbb{1}_A \quad (22)$$

Therefore, sufficient statistics always exist, e.g. the identity map. This isn't interesting, but what is more interesting is whether  $T$  destroys some information yet still is sufficient. The following theorem relates this in terms of the likelihood function.

### Definition 1.6 (Likelihood Function)

Given a  $\mu$ -dominated identifiable model  $(\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$ , the function (for fixed  $x \in X$ )

$$L : \Theta \rightarrow [0, +\infty), \quad L(\theta) = p(x \mid \theta) \quad (23)$$

is called the **likelihood function**.

The theorem states that a statistic is sufficient if and only if the likelihood function can be factorized.

### Theorem 1.2 (Fisher-Neyman Factorization)

Consider a identifiable model  $(\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$  dominated by a  $\sigma$ -finite measure  $\mu$ , with density  $p(x \mid \theta) = \frac{d\mathbf{P}_\theta}{d\mu}$ . Then,  $T$  is sufficient iff

$$p(x \mid \theta) = g(T(x), \theta) h(x) \quad (24)$$

for measurable functions  $g : (T, \mathcal{T}) \times \Theta \rightarrow [0, +\infty)$  and  $h : X \rightarrow [0, +\infty)$ .

*Proof.* In text.

### Example 1.9 (Sums of Two Dice)

Consider two dice rolls where  $(\mathcal{X} = \{1, \dots, 6\}^2, \mathcal{Y} = 2^{\mathcal{X}})$  and statistic  $T : X \rightarrow \mathbb{N}$  defined  $T(x, y) = x + y$ . We can think of  $T$  as “extracting” information from the two dice rolls to their sum. We will consider two models.

1. Let  $\Theta = \{\theta = (\theta_1, \dots, \theta_6) \in \mathbb{R}^6 : \theta_1 + \dots + \theta_6 = 1\}$ , and let  $\mathbf{P}_\theta$  be the multinomial measure assigning  $\mathbf{P}_\theta(\{k\}) = \theta_k$ . Then, our model is the family  $\mathcal{P} = \{\mathbf{P}_\theta \otimes \mathbf{P}_\theta : \theta \in \Theta\}$ , and  $T$  is not a sufficient statistic. Consider the conditional probability

$$\mathbf{P}_\theta^{\otimes 2}(\{(1, 6)\} \mid S = 7) = \frac{\mathbf{P}_\theta(\{1\})\mathbf{P}_\theta(\{6\})}{\sum_{k=1}^6 \mathbf{P}_\theta(\{k\})\mathbf{P}_\theta(\{7-k\})} = \frac{\theta_1\theta_6}{\sum_{k=1}^6 \theta_k\theta_{7-k}} \quad (25)$$

This is clearly dependent on  $\theta$ .

2. Let  $\Theta = \mathbb{R}$  and  $\mathbf{P}_\theta$  be defined by the density

$$p_\theta(x) = \frac{e^{\theta x}}{\sum_{k=1}^6 e^{\theta k}} \quad (26)$$

Note that  $\theta = 0$  gives a fair dice,  $\theta > 0$  biases towards higher faces, and  $\theta < 0$  biases towards lower faces. Then, our model is the family  $\mathcal{P} = \{\mathbf{P}_\theta \otimes \mathbf{P}_\theta : \theta \in \Theta\}$ , and  $T$  is a sufficient statistic. The joint density is

$$p(x, y \mid \theta) = p(x \mid \theta)p(y \mid \theta) = \frac{e^{\theta x}}{\sum_{k=1}^6 e^{\theta k}} \frac{e^{\theta y}}{\sum_{k=1}^6 e^{\theta k}} = \frac{e^{\theta(x+y)}}{\underbrace{\left(\sum_{k=1}^6 e^{\theta k}\right)^2}_{g(x+y, \theta)}} \cdot \underbrace{1}_{h(x, y)} \quad (27)$$

and so by the Fisher-Neyman Factorization,  $T$  is sufficient.

Both models are identifiable, but  $T$  behaves differently. Note that the first model is a vastly larger model in which no reduction beyond the full data is possible, and so  $T$  is not able to capture this well enough. On the other hand, the second model is indexed by a scalar parameter.

### Question 1.1

What does it mean for Model 1 to be nonparametric?

### Definition 1.7 (Minimal Sufficient Statistic)

A sufficient statistic  $T$  is **minimal sufficient** for an experiment  $(\mathcal{X}, \mathcal{Y}, \mathcal{P}, \Theta)$  if for any other sufficient  $S$ , it satisfies  $\sigma(T) \subset \sigma(S)$  modulo  $\mathbf{P}_\theta$ -null sets.

$$\sigma(T) \subset \sigma(\sigma(S) \cup N), \quad N = \{A \in \mathcal{Y} : \mathbf{P}_\theta(A) = 0 \ \forall \theta \in \Theta\} \quad (28)$$

That is, minimal sufficient statistics partition the sample space into the coarsest equivalence classes that preserve all information about  $\theta$ . Another way to think about it is that a sufficient statistic  $T$  is minimal sufficient if it is a function of every other sufficient statistic.

**Theorem 1.3**

Let  $(\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$  be a dominated model. A statistic  $T$  is minimal sufficient iff

$$T(x) = T(x') \iff \theta \mapsto \frac{p(x \mid \theta)}{p(x' \mid \theta)} \text{ is constant} \quad (29)$$

$\mathbf{P}_\theta$ -a.s. for all  $\theta \in \Theta$ .

*Proof.*

**Example 1.10 (Minimal Sufficiency for Uniform)**

Let  $X_1, \dots, X_n \sim U(0, \theta)$  for  $\theta > 0$ . Then,  $X_{(n)} = \max\{X_1, \dots, X_n\}$  is minimal since

$$x \mapsto \frac{p(x \mid \theta)}{y \mid \theta} = \frac{\mathbb{1}\{x_{(n)} \leq \theta\}}{\mathbb{1}\{y_{(n)} \leq \theta\}} \quad (30)$$

is constant a.e. iff  $y_{(n)} = x_{(n)}$ .

**Question 1.2**

Page 4 of Casella. How does estimator, estimand relate to statistic? What does observable actually mean? How are hypothesis testing, point estimation, density estimation, etc realized under this framework?

**Definition 1.8 (Complete Statistic)**

A **statistic**  $T : (\mathcal{T}, \mathcal{S})$  is **complete** for statistical model  $\mathcal{P}$  if for all  $\sigma(T)$ -measurable random variables  $U$ , it holds that

$$\forall \mathbf{P} \in \mathcal{P}, \mathbb{E}_{\mathbf{P}}[U] = 0 \implies \forall \mathbf{P} \in \mathcal{P}, U = 0 \text{ } \mathbf{P} - \text{a.s.} \quad (31)$$

This says that  $T$  is allowed to vary with the data, but while it varies, it doesn't carry additional information about what generated the data.

**Theorem 1.4 (Bahadur)**

If  $T : (\mathcal{X}, \mathcal{X}) \rightarrow (\mathcal{T}, \mathcal{S})$  is complete and sufficient for  $\mathbf{P}$ , then  $T$  is minimal sufficient.

*Proof.* Let  $T$  be such a statistic and let  $S$  be another sufficient statistic. We wish to show that  $\sigma(T) \subset \sigma(S) \text{ mod } \mathbf{P}_\theta$ -null sets for all  $\theta \in \Theta$ . Fix  $\text{Bin} \sigma(T)$  and define

$$H_B := \mathbb{E}_\theta[\mathbb{1} \mid \sigma(S)] \quad (32)$$

which clearly satisfies  $0 \leq H_B \leq 1$ . Let  $U = \mathbb{E}_\theta[H_B \mid \sigma(T)] - \mathbb{1}_B$ .

So in a sense, completeness is stronger than minimality.



**Definition 1.9 (Ancillary)**

A statistic  $V : (\mathcal{X}, \mathcal{X}) \rightarrow ()$

This connects to completeness in a very nice way, but first let's give an example.

**Example 1.11**

Consider uniform.

**Theorem 1.5 (Basu)**

Consider a statistical model  $(\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$ . If  $T : (\mathcal{X}, \mathcal{X}) \rightarrow (\mathcal{T}, \mathcal{T})$  is complete and sufficient and  $V$  is ancillary, then  $V \perp T$  under  $\mathbf{P}_\theta$  for all  $\theta \in \Theta$ .

*Proof.*

**Example 1.12**

Is a geometric intuition of being an orthogonal decomposition in  $L^2$  accurate?

**1.3 Model Equivalence**

While a sufficient statistic allows us to reduce our model to a simpler one, we may want to look for maximal compression.

**Definition 1.10 (Observationally Equivalent)**

If we have two models  $M_1 = (\mathcal{X}, \mathcal{X}, \mathcal{P}, \Theta)$ ,  $M_2 = (\mathcal{Y}, \mathcal{Y}, \mathcal{Q}, \Theta)$ , they are **observationally equivalent** if there exists sufficient statistics  $T : (\mathcal{X}, \mathcal{X}) \rightarrow (\mathcal{T}, \mathcal{T})$ ,  $S : (\mathcal{Y}, \mathcal{Y}) \rightarrow (\mathcal{T}, \mathcal{T})$  such that

$$\mathbf{P}_{\theta, T}(A) = \mathbf{Q}_{\theta, S}(A) \quad \text{for all } A \in \mathcal{T}, \mathbf{P}_\theta \in \mathcal{P}, \mathbf{Q}_\theta \in \mathcal{Q} \quad (33)$$

**Example 1.13**

Let  $X = (X_1, \dots, X_n) \sim N(\mu, 1)$  iid. Let  $Y \sim N(\mu, \frac{1}{n})$ . Then,

$$T(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^n X_i \quad (34)$$

But what if we wanted to go from  $Y$  back to  $X$ ? This leads us to the idea of simulation equivalence, which is informally observationally equivalent under sufficient statistics “with additional randomness.” e.g. if we set  $\tilde{X} = Y - Z_i$  for  $Z_i \sim N(0, 1)$  iid.

**1.4 Exponential Families**

A nice property is that basically a family of models where completeness is easy to verify.

**Definition 1.11 (Exponential Family)**

A collection of probability measures  $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$  is an **exponential family** if it is  $\mu$ -dominated for some  $\sigma$ -finite  $\mu$ , and the density is of the following form

$$p(x | \theta) := \frac{d\mathbf{P}_\theta}{d\mu}(x) = \exp(\eta^T(\theta)T(x) - B(\theta)) h(x) \quad (35)$$

where  $T : \mathcal{X} \rightarrow \mathbb{R}^k$  is measurable,  $h : \mathcal{X} \rightarrow [0, +\infty)$  is measurable,  $\eta : \Theta \rightarrow \mathbb{R}^k$  is any function (not necessarily measurable), and  $B : \Theta \rightarrow \mathbb{R}$  is any function.

Note that this can change if  $\mu$  changes. So really, this should be an exponential family *with respect to a dominating measure*. Then, every statement and example we talk about is always with respect to this measure  $\mu$ . Usually, we will talk about with respect to counting measure or Lebesgue measure.

Exponential families have a sufficient statistic that is “natural” in the sense that  $T$  is always a sufficient statistic. Second, if we have iid observations with model  $\mathcal{P}^n = \{\mathbf{P}^{\otimes n} : \theta \in \Theta\}$ , then  $\sum_{i=1}^n T(x_i)$  is sufficient. This makes sense intuitively since the product of these densities is still exponential (since the product becomes a sum in the exponent). Then use Fisher-Neyman factorization.

**Definition 1.12**

An exponential family is **canonical** if  $\Theta \subset \mathbb{R}^k$  and  $\eta(\theta) = \theta$ . We consider

$$\mathcal{H} := \{\eta \in \mathbb{R}^k : \int e^{\eta^T T(x)} h(x) d\mu(x) < \infty\} \quad (36)$$

the natural, or canonical, parameter space.

**Definition 1.13**

We call a canonical exponential family **full rank** if  $\mathcal{H}$  contains an open subset.

Note that being canonical really limits what you can consider as a dominating measure. So the  $\mathcal{H}$  really does a lot.

**Theorem 1.6**

Let  $\mathcal{P} = \{\mathbf{P}_\eta : \eta \in \mathcal{H}\}$  by a canonical exponential family of full rank. Then,  $T$  is complete and sufficient.

*Proof.* This is mysterious, but it's a very elegant proof. A few notes about this proof. Suppose that I can write (which is not always correct)

$$p(t | \eta) = e^{\eta^T t - B(\eta)} \quad (37)$$

Then,  $\mathbb{E}_\eta g(T) = 0$  for all  $\eta$  implies that  $g(T) = 0$  a.s. Then, we can look at the Laplace transform

$$\int g(t) e^{\eta^T t - B(\eta)} d\nu(t) = 0 \text{ for all } \eta \quad (38)$$

But since  $B$  is independent of the integral, it should mean that

$$\int e^{\eta^T t} \underbrace{g(t) d\nu(t)}_{d\nu'(t)} = 0 \quad (39)$$

Then  $g(t) = 0$  for all  $\nu$  almost all  $t$ .

Note that completeness and sufficiency are independent. Only when they come together, we get minimal statistic.

What if we have  $\{\mathbf{P}_\theta : \theta \in \Theta\}$  not in canonical form? If  $\eta$  is injective, then there is hope. Define  $\Xi = \eta(\Theta)$  and define the reparameterized family

$$\{\mathbf{Q}_\xi : \xi \in \Xi\}, \quad \mathbf{Q}_\xi := \mathbf{P}_{\eta^{-1}(\xi)} \quad (40)$$

#### Example 1.14

Make sure to verify the examples of exponential families: Poisson, Gaussian, etc.

## 1.5 Decision Problems

#### Definition 1.14

A **decision space** is a measurable space  $(\mathcal{D}, \mathcal{D})$ .

#### Example 1.15

A **decision rule** is a statistic taking values in a decision space, i.e. a measurable function  $\delta : X \rightarrow (D, \mathcal{D})$

#### Definition 1.15 (Loss Function)

Given identifiable model, the **loss function** is a function  $L : \Theta \times D \rightarrow \mathbb{R}$  where the map  $d \mapsto L(\theta, d)$  is measurable.  $\theta : \mathbf{P} \rightarrow \Theta$  satisfies the property that  $\theta(\mathcal{P}) = \theta(\mathcal{Q}) \implies \mathbf{P} = \mathbf{Q}$ .

#### Definition 1.16

Given identifiable model  $M$ , decision space  $(\mathcal{D}, \mathcal{D})$ , and a loss function  $L : \theta \times \mathcal{D} \rightarrow \mathbb{R}$ , the **risk** of the

$$R(P, \delta) = \int L(\theta(\mathbf{P}), \delta(x)) d\mathbf{P}(x), \quad R(\theta, \delta) = \int L(\theta, \delta(x)) d\mathbf{P}_{\theta(x)} \text{ for } \theta \in \Theta \quad (41)$$

The two main questions are:

1. Which decision functions  $\delta$  are good?
2. Which models are better? When do we prefer  $(\mathcal{Y}, \mathcal{Y}, \mathbf{Q}, \Theta)$  over  $(\mathcal{X}, \mathcal{X}, \mathbf{P}, \Theta)$ .

#### Example 1.16 (Point Estimation)

#### Example 1.17 (Hypothesis Testing)

#### Example 1.18 (Confidence Sets)

<b>Example 1.19 (Density Estimation)</b>