Frequentist Inference

Muchang Bahng

Winter 2022

Contents

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 θ. We may try to summarize the mean of this data by computing

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{1}$$

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

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

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 θ since no explicit assumption has been made connecting the observations x_i with θ.

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 θ. 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 ϵ_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 σ^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 $X : \Omega \to \mathcal{X}$, 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)

Given a measurable space $(\mathcal{X}, \mathscr{X})$, a **statistical model** is a collection of probability measures $\mathcal{P}$. The triple $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ is called a **statistical experiment**.[a]

[a] Note that this is *not* a probability space! For each $\mathbf{P} \in \mathcal{P}$, the triplet $(\mathcal{X}, \mathscr{X}, \mathbf{P})$ is a probability space.

Therefore, we abuse notation and talk about the following in shorthand:

1. $X \sim N(\theta, 1)$ for $\theta \in [-1, 1]$, means the statistical model $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1) : \theta \in [-1, 1]\})$.
2. $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ with $\theta \in \Theta$ is shorthand for the statistical model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{N(\theta, 1)^{\otimes n} : \theta \in \Theta\})$.

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 1.2 (Dominated Families of Measures)

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

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

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

If a model is dominated by μ, then by the Radon-Nikodym theorem, we have a nonnegative measurable function $p = \frac{d\mathbf{P}}{d\mu}$ s.t. for all $F \in \mathscr{X}$,

$$\mathbf{P}(A) = \int_A p \, d\mu \tag{3}$$

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 1.1 (Discrete Models)

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

$$\mathbf{P}(\{1\}) = \mathbf{P}(\{2\}) = \mathbf{P}(\{3\}) = \frac{1}{3} \tag{4}$$

Then, the density is

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

and we can verify for example that

$$\frac{2}{3} = \mathbf{P}(\{1, 2\}) = \int_{\{1,2\}} p_\theta(x) \, dc(x) \tag{6}$$

$$= p_\theta(1)c(\{1\}) + p_\theta(2)c(\{2\}) \tag{7}$$

$$= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{2}{3} \tag{8}$$

Note that we can also change the dominating measure μ.

Example 1.2

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

$$\frac{d\mathbf{P}}{dc'}(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3 \\ 0 & \text{if } x = 4, 5, 6 \end{cases} \tag{9}$$

And there is nothing wrong with this since

$$\frac{2}{3} = \mathbf{P}(\{1, 2\}) = \int_{\{1,2\}} p_\theta(x)\, dc(x) \tag{10}$$

$$= p_\theta(1)c(\{1\}) + p_\theta(2)c(\{2\}) \tag{11}$$

$$= \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 2 = \frac{2}{3} \tag{12}$$

Question 1.1

What happens if we have a half-discrete and half-continuous measure? What is the dominating measure? This isn't as obvious, and then we're really in trouble if there is no canonical form.

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.

Example 1.3 (Absolutely Continuous Models)

Often, the family $\mathcal{P}$ may be too abstract for us to work with, and we want to parameterize it by some functionals $\theta : \mathcal{P} \to \Theta$. This new space Θ has some extra structure that makes it easier to work with.

Definition 1.3 (Parameter Space)

A **parameter space** for the model $\mathcal{P}$ is a set Θ, together with a map $\theta : \mathcal{P} \to \Theta$.

Question 1.2

So is θ a function or an element of θ here? Is it really necessary to make it this complicated?

Example 1.4

Definition 1.4 (Identifiability)

A statistical model is **identifiable** by parameter space Θ if for all $\mathbf{P}, \mathbf{P}' \in \mathcal{P}$,

$$\theta(\mathbf{P}) = \theta(\mathbf{P}') \implies \mathbf{P} = \mathbf{P}' \tag{13}$$

If a model is identifiable, then we can index the distributions by the parameter value $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$, and write our experiment as the tuple $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta)$.

Question 1.3

When do we ever work with statistical models that are not identifiable?

1.2 Statistics and Sufficiency

Definition 1.5 (Statistic)

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

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 1.6 (Sufficient Statistic)

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

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

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

$$\mathbb{E}_\theta[\mathbb{1}_A \mid \sigma(T)] = \mathbb{1}_A \tag{14}$$

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 1.7 (Likelihood Function)

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

$$L : \Theta \to [0, +\infty), \quad L(\theta) = p(x \mid \theta) \tag{15}$$

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

$$p(x \mid \theta) = g(T(x), \theta)\, h(x) \tag{16}$$

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

Proof. In text.

Example 1.5 (Sums of Two Dice)

Consider two dice rolls where $(\mathcal{X} = \{1, \dots, 6\}^2, \mathscr{X} = 2^{\mathcal{X}})$ and statistic $T : X \to \mathbb{N}$ 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 $\Theta = \{\theta = (\theta_1, \dots, \theta_6) \in \mathbb{R}^6 \colon \theta_1 + \ldots + \theta_6 = 1\}$, and let $\mathbf{P}_\theta$ be the multinomial measure assigning $\mathbf{P}_\theta(\{k\}) = \theta_k$. Then, our model is the family $\mathcal{P} = \{\mathbf{P}_\theta \otimes \mathbf{P}_\theta \colon \theta \in \Theta\}$, and T is not a sufficient statistic. Consider the conditional probability

$$\mathbf{P}_\theta^{\otimes 2}(\{(1,6)\} \mid S = 7) = \frac{\mathbf{P}_\theta(\{1\})\mathbf{P}_\theta(\{6\})}{\sum_{k=1}^6 \mathbf{P}_\theta(\{k\})\mathbf{P}_\theta(\{7-k\})} = \frac{\theta_1\theta_6}{\sum_{k=1}^6 \theta_k\theta_{7-k}} \tag{17}$$

   This is clearly dependent on θ.

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

$$p_\theta(x) = \frac{e^{\theta x}}{\sum_{k=1}^6 e^{\theta k}} \tag{18}$$

   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 $\mathcal{P} = \{\mathbf{P}_\theta \otimes \mathbf{P}_\theta \colon \theta \in \Theta\}$, and T is a sufficient statistic. The joint density is

$$p(x, y \mid \theta) = p(x \mid \theta)p(y \mid \theta) = \frac{e^{\theta x}}{\sum_{k=1}^6 e^{\theta k}} \frac{e^{\theta y}}{\sum_{k=1}^6 e^{\theta k}} = \underbrace{\frac{e^{\theta(x+y)}}{\left(\sum_{k=1}^6 e^{\theta k}\right)^2}}_{g(x+y,\theta)} \cdot \underbrace{1}_{h(x,y)} \tag{19}$$

   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 1.4

What does it mean for Model 1 to be nonparameteric?

Question 1.5

Organization wise, why did you put the experiment and simulation equivalence? I feel like the flow from sufficient stats to minimal sufficient, complete and ancillary is more natural.

Definition 1.8 (Minimal Sufficient Statistic)

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

$$\sigma(T) \subset \sigma(\sigma(S) \cup N), \quad N = \{A \in \mathscr{X} : \mathbf{P}_\theta(A) = 0 \ \forall \theta \in \Theta\} \tag{20}$$

That is, minimal sufficient statistics partition the sample space into the coarsest equivalence classes that preserve all information about θ. 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 $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta)$ be a dominated model. A statistic T is minimal sufficient iff

$$T(x) = T(x') \iff \theta \mapsto \frac{p(x \mid \theta)}{p(x' \mid \theta)} \text{ is constant} \tag{21}$$

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

Proof.

Question 1.6

Can we verify this theorem statement with Prop 1.18 in the notes?

Example 1.6 (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

$$x \mapsto \frac{p(x \mid \theta)}{y \mid \theta} = \frac{\mathbb{1}\{x_{(n)} \leq \theta\}}{\mathbb{1}\{y_{(n)} \leq \theta\}} \tag{22}$$

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

Question 1.7

Is the "independent" term in Basu's Theorem mean independent RVs?

Question 1.8

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?

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 1.9 (Observationally Equivalent)

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

$$\mathbf{P}_{\theta,T}(A) = \mathbf{Q}_{\theta,S}(A) \quad \text{for all } A \in \mathcal{T}, \mathbf{P}_\theta \in \mathcal{P}, \mathbf{Q}_\theta \in \mathcal{Q} \tag{23}$$

Example 1.7

Let $X = (\mathcal{X}_1, \dots, X_n) \sim N(\mu, 1)$ iid. Let $Y \sim N(\mu, \frac{1}{n})$. Then,

$$T(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{24}$$

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 Decision Problems

Definition 1.10 (Statistical Experiment)

1. The **decision rule** is a measurable function $\delta : X \to (D, \mathcal{D})$
2. The **loss function** is $L : \Theta \times D \to \mathbb{R}$. $\theta : \mathbf{P} \to \Theta$ satisfies the property that $\theta(\mathcal{P}) = \theta(\mathcal{Q}) \implies \mathbf{P} = \mathbf{Q}$.

The **risk** is defined

$$R(P, \delta) = \int L(\theta(\mathbf{P}), \delta(x))\, d\mathbf{P}(x), \quad R(\theta, \delta) = \int L(\theta, \delta(x))\, d\mathbf{P}_{\theta(x)} \text{ for } \theta \in \Theta \tag{25}$$

The two main questions are:

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

Example 1.8 (Point Estimation)

Example 1.9 (Hypothesis Testing)

Example 1.10 (Confidence Sets)

Example 1.11 (Density Estimation)