18.650: Fundamentals of Statistics September 18,20, 2023

Lecture 6,7 —Models, point estimation, & confidence intervals

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Overview

Let's add to the statistics pipeline from Lecture 1:

$$\begin{matrix}\text{i.i.d. data} \\ X_i \sim \mathbb{P}\end{matrix} \longrightarrow \boxed{\begin{matrix}\text{statistical} \\ \text{method}\end{matrix}} \longrightarrow \hat{\mathbb{P}} \longrightarrow \text{statistical inference} \tag{1}$$

The last step (what we do with $\hat{\mathbb{P}}$) is statistical inference.

There are three main tasks in statistical inference.

1. Estimation: find one estimate of $\mathbb{P}$, or a parameter that characterizes $\mathbb{P}$. Example: $\mathbb{P} = \mathbb{P}_\mu = \mathcal{N}(\mu, 1)$. The sample mean $\hat{\mu} = \bar{X}_n$ is an estimator of μ.

2. Confidence intervals (or confidence sets in higher dimensions): a random interval such that the parameter falls inside of it with high probability, e.g

$$\mathbb{P}_\mu \left(\mu \in \left[\bar{X}_n - \frac{2}{\sqrt{n}}, \bar{X}_n + \frac{2}{\sqrt{n}} \right] \right) \geq 0.95.$$

3. Tests: $\mu = 0$ or $\mu \neq 0$? The "null hypothesis" $\mu = 0$ typically has some special meaning, e.g. does a drug act differently from the placebo ($\mu \neq 0$) or not ($\mu = 0$)?

2 Models

Definition 2.1: Statistical model

A model is a set of probability distributions, which is typically a strict subset of the set of *all* probability distributions.

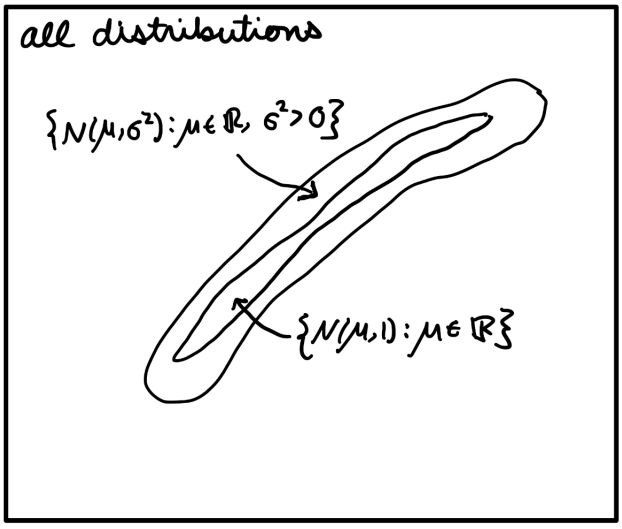

Figure 1: By imposing a model, we restrict our study to a certain subset of the set of all possible probability distributions. For example, we might consider all normal distributions, or all normal distributions with variance 1.

There are many ways to specify a model.

Example.

- using a common family of distributions, e.g.
  - $\{\text{Ber}(p) : p \in (0,1)\}$.
  - $\{\text{Exp}(\lambda) : \lambda \in [0, 22]\}$
- in terms of pdfs/pmfs, e.g. $\left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$.
- in terms of cdfs, e.g. $\{F(x) : F \text{ is a continuous CDF}\}$, which rules out cdf's that have jumps.

To make statements about general statistical models, we will refer to the model as

$$\{\mathbb{P}_\theta : \theta \in \Theta\}.$$

Here, Θ is the parameter space that θ lives in.

Definition 2.2: Parametric vs nonparametric model

If Θ has finite dimension, we call it a parametric model. If Θ has infinite dimension, we call it a nonparametric model.

Remark.

- Note that in the nonparametric case, there is still a "parameter", it's just infinite-dimensional. Usually, the parameter space is some sort of function class.

Example.

1. $\{\text{Exp}(\lambda) : \lambda \in (0, \infty)\}$ is parametric
2. $\{\text{pdf } f \text{ s.t. } |f(x) - f(y)| \leq L|x - y| \text{ for all } x, y\}$ (Lipschitz functions). This is nonparametric.
3. $\{\text{pdf } f \text{ is a polynomial}\}$. Nonparametric.
4. $\{\text{pdf } f \text{ is a polynomial with degree at most } d\}$. Parametric.

We'll focus mostly on parametric statistics in this class.

Definition 2.3: Functional

A functional of a probability distribution $\mathbb{P}$ is a function T which maps $\mathbb{P}$ to a number.

A functional is just like a function, but it takes as input a more general object than just a single number or several numbers.

Example.

1. $T(\mathbb{P}) = \mathbb{P}(X \leq 1) = \mathbb{E}[\mathbb{1}(X \leq 1)]$. [The second formulation will be useful later on. To see why it holds, note that $\mathbb{E}[\mathbb{1}(X \leq 1)] = 1 \times \mathbb{P}(X \leq 1) + 0 \times \mathbb{P}(X > 1) = \mathbb{P}(X \leq 1)$].
2. $T(\mathbb{P}) = \mathbb{E}[X]$ where $X \sim P$.
3. $T(\mathbb{P}) = \int (f''(x))^2 dx$, where f is the pdf of $\mathbb{P}$.
4. $T(\mathbb{P}) = |\text{supp}(\mathbb{P})| = \#\{x : \mathbb{P}(X = x) \neq 0\}$ for a discrete distribution (supp for "support").

For a model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, we indicate that some statistic is evaluated under the distribution $\mathbb{P}_\theta$ with a subscript θ, e.g

$$\mathbb{P}_\theta(X \geq 1), \qquad \mathbb{E}_\theta[X], \qquad \mathbb{V}_\theta[X].$$

So for a normal family, we would write $\mathbb{P}_{\mu,\sigma^2}(X \geq 1)$ or $\mathbb{E}_{\mu,\sigma^2}[X]$ (which equals μ).

3 Fundamental concepts in inference

3.1 Point estimation.

A point estimate $\hat{\theta}$ or $\hat{\theta}_n$ (to emphasize its dependence on n data points) is a single best guess for a parameter θ.

We'll use the notation $\theta \rightsquigarrow \hat{\theta}$ to mean θ is estimated by $\hat{\theta}$. For example, $\mu \rightsquigarrow \bar{X}_n$. (Recall $\rightsquigarrow$ also stands for weak convergence, but the difference between the two meanings will be clear in context.)

Definition 3.1: Estimator

An estimator $\hat{\theta}$ of θ is a function of the data: $\hat{\theta} = g(X_1, \ldots, X_n)$ for a measurable function g.

Example.

> $\bar{X}_n$, $\max_i X_i$, $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ can all be used as estimators.
> X_1 and 4 are also valid estimators.

It's clear that X_1 and 4 are bad estimators (the former throws out $n-1$ data points, while the latter doesn't look at the data at all). How do we formalize this?

We'll consider two properties of an estimator: its **bias** and its **variance**.

Definition 3.2: Bias

The bias of an estimator $\hat{\theta}$ of θ is defined as $\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$. We say $\hat{\theta}$ is *unbiased* if $\text{bias}(\hat{\theta}) = 0$. We say $\hat{\theta}_n$ is *asymptotically unbiased* if $\text{bias}(\hat{\theta}_n) \to 0$ as $n \to \infty$.

Example.

> $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. Consider the following three estimators of μ.
>
> 1. $\hat{\mu}_1 = \bar{X}_n$. $\text{bias}(\hat{\mu}_1) = \mathbb{E}_\mu[\bar{X}_n] - \mu = 0$. Unbiased.
> 2. $\hat{\mu}_2 = X_1$. $\text{bias}(\hat{\mu}_2) = \mathbb{E}_\mu[X_1] - \mu = 0$. Unbiased.
> 3. $\hat{\mu}_3 = 0$. $\text{bias}(\hat{\mu}_3) = \mathbb{E}_\mu[0] - \mu = -\mu$. Biased unless $\mu = 0$.

The second property of an estimator is how much it fluctuates, measured by its variance.

Definition 3.3: Standard error

The standard error of an estimator $\hat{\theta}$ is $\text{se}(\hat{\theta}) = \sqrt{\mathbb{V}[\hat{\theta}]}$.

It turns out that there is a third quantity which simultaneously captures both the bias and the variance:

Definition 3.4: Mean squared error (MSE)

The mean squared error of an estimator $\hat{\theta}$ of θ is $\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$.

We now show

$$\text{MSE}(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + \mathbb{V}[\hat{\theta}].$$

Indeed,

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \text{bias})^2] = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)\text{bias}\right] + \text{bias}^2 \\
&= \mathbb{V}[\hat{\theta}] + 2\mathbb{E}\left[\hat{\theta} - \mathbb{E}[\hat{\theta}]\right]\text{bias} + \text{bias}^2 = \mathbb{V}[\hat{\theta}] + \text{bias}^2.
\end{aligned} \tag{2}$$

Example.

1. $\hat{\mu}_1 = \bar{X}_n$. $\text{MSE}(\hat{\mu}_1) = 0^2 + \mathbb{V}_\mu[\hat{\mu}_1] = 0^2 + \frac{1}{n} = \frac{1}{n}$.
2. $\hat{\mu}_2 = X_1$. $\text{MSE}(\hat{\mu}_2) = 0^2 + \mathbb{V}_\mu[\hat{\mu}_2] = 0^2 + 1 = 1$.
3. $\hat{\mu}_3 = 0$. $\text{MSE}(\hat{\mu}_3) = (\mu)^2 + \mathbb{V}_\mu[0] = \mu^2 + 0 = \mu^2$. (Zero variance, but possibly large bias!)

Definition 3.5: Consistency

An estimator $\hat{\theta}_n$ of θ is consistent if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$.

Example.

1. $\hat{\mu}_1 = \bar{X}_n \xrightarrow{\mathbb{P}} \theta$ by the LLN.
2. $\hat{\mu}_2 = X_1$ does *not* converge to θ
3. $\hat{\mu}_3 = 0$ does *not* converge to θ.

Theorem 3.6

If $\text{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent.

3.2 Asymptotic Normality

If $\hat{\theta}_n$ is consistent, then we know $\hat{\theta}_n - \theta$ converges to zero. But often we want to know *how fast* $\hat{\theta}_n - \theta$ converges to zero. This can be measured by the standard error of $\hat{\theta}_n$, which will go to zero at some rate. If we normalize $\hat{\theta}_n - \theta$ by $\text{se}(\hat{\theta}_n)$, then we can expect to get a random variable in the limit with variance 1. Typically, the CLT and/or delta method ensures this limit is normally distributed.

Definition 3.7: asymptotic normality

An estimator is asymptotically normal if

$$\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \rightsquigarrow \mathcal{N}(0, 1).$$

Example.

$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. Then $\text{se}(\bar{X}_n) = 1/\sqrt{n}$, and by the CLT we know that indeed, $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, 1)$.

3.3 Confidence intervals

A confidence interval is an interval in which θ lies with some probability.

Definition 3.8: confidence interval

A $1-\alpha$ confidence interval (CI) for θ is a *random* interval of the form $C_n = (A_n, B_n)$ such that

$$\mathbb{P}_\theta(\theta \in (A_n, B_n)) = \mathbb{P}_\theta(A_n < \theta < B_n) \geq 1-\alpha \quad \forall \theta.$$

Here, $1-\alpha$ is called the *coverage* of the CI.

Note that the randomness comes from A_n and B_n, *not* from θ.

Interpretation. If for a given dataset we get $A_n = 2, B_n = 4$ for our 95% confidence interval, then it is *not* true that $\mathbb{P}_\theta(\theta \in (2,4)) \geq 0.95$. That probability is either 0 or 1, because either θ lies in $(2,4)$ or it doesn't. The way to interpret a confidence interval is that if we repeat an experiment a 100 times (collecting new data each time), and each time we construct a confidence interval based on the data, then we can expect that 5 out of those confidence intervals "miss" (don't contain θ), but the other 95 succeed at trapping θ inside.

Remark.

If (A_n, B_n) is a $1-\alpha$ confidence interval, then so is $(A_n - 10, B_n + 10)$, because widening the interval will only increase the probability that θ lies in it. But this is giving us less precise information. So we shoot for the narrowest interval at the given confidence level.

Remark.

In higher dimensions, we construct confidence *sets* rather than confidence intervals. The shape of the confidence set can be pretty much anything.

3.3.1 Constructing a CI

We build a CI using asymptotic normality. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. Then

$$Z = \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \approx \mathcal{N}(0,1)$$

for large n by the CLT. For now on, assume $Z \sim \mathcal{N}(0,1)$ exactly.

Let $z_{\alpha/2}$ be the point such that $\mathbb{P}(|Z| \leq z_{\alpha/2}) = 1-\alpha$, i.e. the point such that

$\mathbb{P}(Z \leq -z_{\alpha/2}) = \mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$. Then

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(|Z| \leq z_{\alpha/2}\right) = \mathbb{P}\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2}\right) \\ &= \mathbb{P}\left(\bar{X}_n - \sqrt{\frac{p(1-p)}{n}}z_{\alpha/2} \leq p \leq \bar{X}_n + \sqrt{\frac{p(1-p)}{n}}z_{\alpha/2}\right) \end{aligned} \tag{3}$$

So is

$$\left(\bar{X}_n - \sqrt{\frac{p(1-p)}{n}}z_{\alpha/2},\ \bar{X}_n + \sqrt{\frac{p(1-p)}{n}}z_{\alpha/2}\right)$$

our desired CI? No, because it depends on the unknown p! Instead we just replace p by $\bar{X}_n$, since $\bar{X}_n \approx p$ by the LLN. We get

$$95\%\text{CI} = \left(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}z_{\alpha/2},\ \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}z_{\alpha/2}\right).$$

Alternative solution: note that $\sqrt{p(1-p)} \leq 1/2$, regardless of what p is. So $\bar{X}_n \pm \frac{1}{2\sqrt{n}}z_{\alpha/2}$ is also a valid CI. This is a more conservative confidence interval.

Formally, the CI we have constructed is only valid *asymptotically*.

Definition 3.9: asymptotic coverage

If

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$$

then we say that C_n has asymptotic coverage $1-\alpha$.