Statistics 210A Lecture 20 Notes

Daniel Raban

November 2, 2021

1 Convergence, Consistency, and Limit Theorems

1.1 A note about linear regression

Last time, we discussed linear regression, where we have $x_i \in \mathbb{R}^d$ and $y_i = x_i^\top \beta + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then we can write the density as

$$\begin{aligned} p(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^\top \beta)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}y^\top y + \frac{1}{\sigma^2}(X\beta)^\top y - \frac{\beta^\top X^\top X\beta}{2\sigma}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y\|^2 + \frac{\beta^\top}{\sigma^2}(X^\top y) - A(\beta)\right). \end{aligned}$$

Then $X^\top y, \|y\|^2$ are sufficient iff $(X^\top X)^{-1}X^\top y$ and $\|y\|^2$ are sufficient. This is equivalent to the OLS estimator $\widehat{\beta}$ and RSS $= \|y\|^2 - \|X\widehat{\beta}\|^2$ being sufficient. So we can make a suffiicencty reduction to $\widehat{\beta}$, $\widehat{\sigma}^2$. Here, one can show that $\widehat{\beta} = (X^\top X)^{-1}X^\top y \sim N_d(\beta, \sigma^2(X^\top X)^{-1})$with $\widehat{\beta} \amalg \widehat{\sigma}^2$. Note that this is d-dimensional, rather than n-dimensional, so we have a dimensionality reduction.

1.2 Convergence and consistency

Let $X_1, X_2, \cdots \in \mathbb{R}^d$ be random variables.

Definition 1.1. X_n **converges in probability** to c, written $X_n \xrightarrow{p} c$, if

$$\mathbb{P}(\|X_n - c\| > \varepsilon) \to 0 \qquad \forall \varepsilon > 0.$$

This says that X_n becomes roughly constant.

Definition 1.2. X_n **converges in distribution** to X, written $X_n \xrightarrow{D} X$ or $X_n \implies X$, if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous functions f.

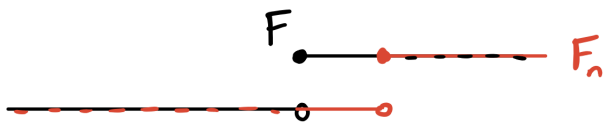
This says that when n is large, the distribution of X_n looks a lot like the distribution of X.

Theorem 1.1. *If $X_1, X_2, \cdots \in \mathbb{R}$, let the CDFs be $F_n(x) = \mathbb{P}(X_n \leq x)$ and $F(x) = \mathbb{P}(X \leq x)$. Then $X_n \implies X$ iff $F_n(x) \to F(x)$ for all x such that F is continuous at x.*

This is a weaker version of pointwise convergence, and convergence in distribution is sometimes called **weak convergence**. Here is why we want to only consider continuity points:

Example 1.1. Let $X_n \sim \delta_{1/n}$ and $X \sim \delta_0$. We want our definition to say $X_n \implies X$. The CDFs are

$$F_x(x) = \mathbb{1}_{\{1/n \leq x\}}, \qquad F(x) = \mathbb{1}_{\{0 \leq x\}}.$$



This example suggests that convergence in probability and in distribution are related.

Proposition 1.1. *$X_n \xrightarrow{p} c$ if and only if $X_n \implies \delta_c$.*

The kind of convergence we care most about in statistics is consistency:

Definition 1.3. If $\mathcal{P}_n = \{P_{\theta,n} : \theta \in \Theta\}$ with $X_n \sim P_{n,\theta}$, then we say that $\delta_n(X_n)$ is **consistent** for $g(\theta)$ if $\delta_n(X_n) \xrightarrow{p} g(\theta)$ for all θ, i.e.

$$\mathbb{P}_\theta(\|\delta_n(X_n) - g(\theta)\| > \varepsilon) \to 0.$$

1.3 Limit theorems

1.3.1 The law of large numbers and the central limit theorem

Theorem 1.2 ((Weak) law of large numbers)**.** *Let $X_1, X_2, \ldots$ be iid random vectors, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. If $\mathbb{E}[\|X_i\|] < \infty$ and $\mathbb{E}[X_i] = \mu$, then $\overline{X}_n \xrightarrow{p} \mu$.*

Remark 1.1. You may have seen a stronger version of this theorem, in which we can prove that $\overline{X}_n \to \mu$ **almost surely**. In statistics, we are interested in convergence in probability, so this will suffice for our purposes.

Theorem 1.3 (Central limit theorem)**.** *Let $X_1, X_2, \cdots \in \mathbb{R}^d$ be iid random vectors, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Assume that $\mathbb{E}[X_i = \mu$ and $\mathrm{Var}(X_i) = \Sigma < \infty$. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \implies N_d(0, \Sigma).$$

1.3.2 The continuous mapping theorem

Here are three tools for how we propagate convergence to other kinds of random variables:

Theorem 1.4 (Continuous mapping)**.** *Let $X_1, X_2, \ldots$ be random variables, and let g be a continuous function. If $X_n \implies X$, then $g(X_n) \implies g(X)$. In particular, if $X_n \xrightarrow{p} c$, then $g(X_n) \xrightarrow{p} g(c)$.*

Proof. If f is bounded and continuous, then $f \circ g$ is bounded and continuous, so

$$\mathbb{E}[f(g(X_n))] = \mathbb{E}[f \circ g(X_n)] \to \mathbb{E}[f \circ g(X_n)] = \mathbb{E}[f(g(X))]. \qquad \square$$

1.3.3 Slutsky's theorem

Theorem 1.5 (Slutsky)**.** *Assume $X_n \implies X$ and $Y_n \xrightarrow{p} c$. Then*

$$X_n + Y_n \implies X + c, \qquad X_n Y_n \implies X \cdot c, \qquad \frac{X_n}{Y_n} \implies \frac{X}{c}$$

(where we assume $c \neq 0$ for the last one).

Proof. Here is a sketch: The first step is to show that $(X_n, Y_n) \implies (X, c)$. Then apply the continuous mapping theorem. $\square$

1.3.4 The delta method

Last is the delta method, which informally says that if $X_n \approx (\mu, \sigma^2)$ with σ^2 small and f is differentiable, then $f(X_n) \approx N(f(\mu), \sigma^2 \dot{f}(\mu)^2)$.

Theorem 1.6 (Delta method)**.** *If $\sqrt{n}(X_n - \mu) \implies N(0, \sigma^2)$ and $f(x)$ is continuously differentiable at μ, then*

$$\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \sigma^2 \dot{f}(\mu)^2).$$

Proof. Here is the idea: Write $f(X_n) = f(\mu) + \dot{f}(\zeta_n)(X_n - \mu)$, where ζ_n is between μ and X_n (by the mean value theorem). $X_n \xrightarrow{p} \mu$ because $X_n - \mu \xrightarrow{p} 0$. Then $\zeta_n \xrightarrow{p} 0$, as well, because

$$\mathbb{P}(\|\zeta_n - \mu\| > \varepsilon) \leq \mathbb{P}(\|X_n - \mu\| > \varepsilon) \to 0.$$

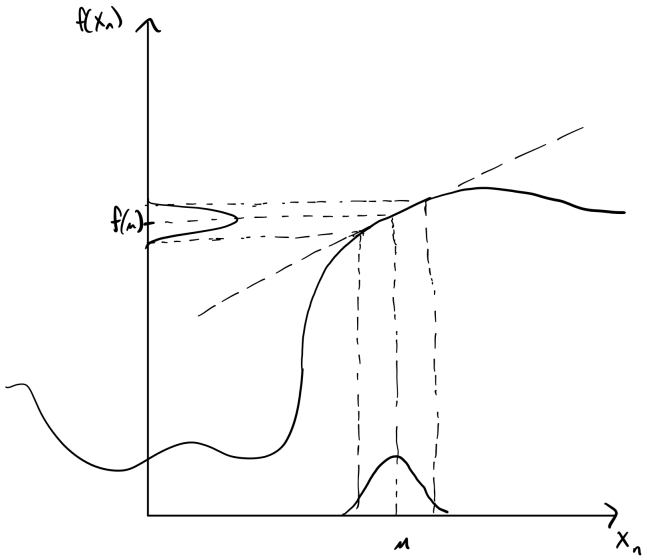
So, by the continuous mapping theorem applied to $\dot{f}$,

$$\sqrt{n}(f(X_n) - f(\mu)) = \underbrace{\dot{f}(\zeta_n)}_{\xrightarrow{p} \dot{f}(\mu)} \underbrace{\sqrt{n}(X_n - \mu)}_{\implies N(0, \sigma^2)},$$

So by Slutsky's theorem, $\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \sigma^2 \dot{f}(\mu)^2)$. $\square$

Remark 1.2. We don't need to have $\sqrt{n}$ in the front. The theorem is still true if we replace $\sqrt{n}$ with a_n, as long as $a_n \to \infty$. Where in the proof did we use that $\sqrt{n} \to \infty$? This was necessary for the fact that $X_n \xrightarrow{p} \mu$.

Here is a picture of the delta method:



There is also a multivariate version:

Theorem 1.7 (Delta method, multivariate)**.** *If* $\sqrt{n}(X_n - \mu) \implies N(0, \Sigma)$ *and* $f : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable at* μ*, then*

$$\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \nabla^\top \Sigma \nabla f).$$

The proof is the same as the univariate case.

Example 1.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, and let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\nu, \tau^2)$ be independent of the X_i. Suppose we estimate $(\mu + \nu)^2$ with $(\overline{X} + \overline{Y})^2$. We can say a few things:

1. By the law of large numbers, $\overline{X} \xrightarrow{p} \mu$ and $\overline{Y} \xrightarrow{p} \nu$ as $n \to \infty$. The function $f(x, y) = (x + y)^2$ is continuous, so $f(\overline{X}, \overline{Y}) \xrightarrow{p} f(\mu, \nu)$. In other words,

   $$(\overline{X} + \overline{Y})^2 \xrightarrow{p} (\mu + \nu)^2,$$

   so $(\overline{X} + \overline{Y})^2$ is consistent for $(\mu + \nu)^2$.

2. The central limit theorem says that $\sqrt{n}(\overline{X} - n) \implies N(0, \sigma^2)$ and $\sqrt{n}(\overline{Y} - n) \implies N(0, \tau^2)$. Here,

$$\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 2(x + y).$$

So the delta method tells us that

$$\begin{aligned} f(\overline{X}, \overline{Y}) &\approx N\left(f(\mu, \nu), \frac{1}{n}\nabla f(\mu, \nu)^\top \begin{bmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{bmatrix} \nabla f\right) \\ &= N\left((\mu + \nu)^2, 4(\mu + \nu)^2(\sigma^2 + \tau^2)/n\right). \end{aligned}$$

More rigorously,

$$\sqrt{n}((\overline{X} + \overline{Y})^2 - (\mu + \nu)^2) \implies N(0, 4(\mu + \nu)^2(\sigma^2 + \tau^2)).$$

3. What if $(\mu + \nu)^2 = 0$? Then

$$\sqrt{n}((\overline{X} - \overline{Y})^2 - (\mu + \nu)^2) \xrightarrow{p} 0.$$

We also know

$$\sqrt{n\overline{X}} + \sqrt{n\overline{Y}} \implies N(0, \sigma^2 + \tau^2),$$

so if we square this sum,

$$n(\overline{X} + \overline{Y})^2 \implies (\sigma^2 + \tau^2)\chi_1^2.$$

If we keep getting things converging to 0, we can keep blowing up the error to find what the distribution of the error rate is in this way.