Outline

1) Convergence in Probability and Distribution

2) Continuous Mapping, Slutsky's Theorem

3) Delta method

Example

Logistic Regression (fixed design)

(x_i, y_i) pairs $i = 1, \ldots, n$

- $x_i \in \mathbb{R}^d$ Continuous feature vector, fixed ($x_{i,1} = 1$ for intercept)
- $y_i \overset{ind.}{\sim} \text{Bern.}(\pi_\beta(x_i))$
- $\text{logit}(\pi_\beta(x_i)) = \log \frac{\pi_\beta}{1-\pi_\beta} = \beta' x_i$

$$P_\beta(y \mid x) = \prod_{i=1}^n \pi_\beta(x_i)^{y_i} (1 - \pi_\beta(x_i))^{1-y_i}$$

$$= \prod_{i=1}^n e^{(\beta' x_i) y_i + \log(1 - \pi_\beta(x_i))}$$

$$= e^{\beta' X' y + A(\beta; x_i)}$$

$$X = \begin{pmatrix} - x_1' - \\ \vdots \\ - x_n' - \end{pmatrix} \in \mathbb{R}^{n \times d}$$

Sufficient statistics: $T(y) = X'y$

Natural parameter: β

Idea to test $H_0 : \beta_1 = 0$: Condition on $X_{-1}' y$

... but that would condition on Y

Ideas to estimate β: UMVU? generically doesn't exist

Bayes? need prior on $\beta \in \mathbb{R}^d$

Software packages use general purpose asymptotic methods

$$\hat{\beta}_{MLE}(x,y) = \underset{\beta \in \mathbb{R}}{\arg\max}\ p_\beta(y \mid x)$$

$$= \underset{\beta \in \mathbb{R}^d}{\arg\max}\ \underbrace{\beta' x' y - A(\beta; x_i)}_{\text{(concave)}} = \ell(\beta; x, y)$$

Asymptotically, (large n)

$$\hat{\beta}_{MLE} \approx N(\underset{\text{unbiased}}{\beta},\ \underset{\text{efficient}}{J(\beta)^{-1}})$$

(Hessian) $\nabla^2 \ell(\hat{\beta}; x, y) \approx \mathbb{E}_\beta[\nabla^2 \ell(\beta; x, y)] = J(\beta)$

$$\hat{\Sigma} = (-\nabla^2 \ell(\hat{\beta}))^{-1} \approx \Sigma(\beta) = J(\beta)^{-1}$$

Test/interval: $Z_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \approx N(0,1)$ $\quad (\sigma_j^2 = \Sigma_{jj})$

Test $H_0: \beta_j = 0$: reject if $\hat{\beta}_j / \hat{\sigma}_j$ large/small/extreme

Invert: $|z_j| < z_{\alpha/2} \iff \beta_j \in \hat{\beta}_j \pm z_{\alpha/2}\hat{\sigma}_j$

Asymptotics

[So far, everything has been finite-sample, often using special properties of model $\mathcal{P}$ (e.g. exp. fam.) to do exact calculations.]

[For "generic" models, exact calculations may be intractable or impossible. But we may be able to approximate our problem with a simpler problem in which calculations are easy]

[Typically approximate by Gaussian, by taking limit as # observations $\to \infty$. But this is only interesting if approx. is good for "reasonable" sample size.]

Convergence

Let $X_1, X_2, \ldots \in \mathbb{R}^d$ sequence of random vectors

We care about 2 kinds of convergence:

1) cvg. in probability $(X_n \approx \text{constant})$
2) cvg. in distribution $(X_n \approx N_d(0, I_d)$, usually$)$

We say the sequence <u>converges in probability</u> to $c \in \mathbb{R}^d$ $(X_n \xrightarrow{P} c)$ if

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

(could really be any distance on any $\mathcal{X}$)

[Can converge to a r.v. X too, but we don't need this]

We say the sequence <u>converges in distribution</u> to random variable X $(X_n \Rightarrow X,\ X_n \xrightarrow{d} X)$ if

$$\mathbb{E} f(X_n) \to \mathbb{E} f(X) \quad \text{for all bdd, cts } f: \mathcal{X} \to \mathbb{R}$$

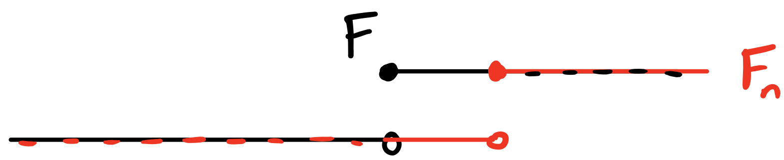
<u>Thm</u> $X_1, X_2, \ldots \in \mathbb{R}$, $F_n(x) = \mathbb{P}(X_n \le x)$, $F(x) = \mathbb{P}(X \le x)$

Then $X_n \Rightarrow X$ iff $F_n(x) \to F(x)$ $\forall x$: F cts at x

Also known as <u>weak convergence</u>

Ex: If $X_n \sim \delta_{1/n}$ $\left(X_n \overset{a.s.}{=} \frac{1}{n}\right)$, $X \sim \delta_0$, then $X_n \Rightarrow X$

$$F_n(x) = 1\{\tfrac{1}{n} \le x\} \longrightarrow 1\{0 \le x\} \quad \underline{\text{except}} \quad x = 0$$



Prop $\quad X_n \xrightarrow{P} c \quad$ iff $\quad X_n \Rightarrow \delta_c$

Proof $(\Leftarrow)$ Let $f_\varepsilon(x) = \max\left(1, \|x-c\|/\varepsilon\right) \ge 1\{\|x-c\| > \varepsilon\}$

$$\mathbb{P}(\|X_n - c\| > \varepsilon) \le \mathbb{E} f_\varepsilon(X_n) \longrightarrow 0$$

$(\Rightarrow)$ f bdd, cts, note $\mathbb{E} f(X) = f(c)$

$\forall \varepsilon > 0$, $\exists d(\varepsilon) > 0$ s.t. $\|x - c\| \le d(\varepsilon) \Rightarrow |f(x) - f(c)| \le \varepsilon$

$$\mathbb{E} f(X_n) - f(c) \le \mathbb{E}\left[|f(X_n) - f(c)| \cdot \left(1\{\|X_n - c\| \le d(\varepsilon)\} + 1\{\|\cdot\| > d\}\right)\right]$$

$$\le \varepsilon + \mathbb{P}(\|X_n - c\| > d(\varepsilon)) \cdot \sup_x |f(x) - f(c)|$$

$$\le 2\varepsilon \cdot \sup|f| \quad \text{for suff. large } n \quad \boxtimes$$

In a sequence of statistical models $\mathcal{P}_n = \{P_{n,\theta} : \theta \in \Theta\}$ with $X_n \sim P_{n,\theta}$, we say $\delta_n(X_n)$ is <u>consistent</u> for $g(\theta)$ if $\delta_n(X_n) \xrightarrow{P_\theta} g(\theta)$, meaning

$$\mathbb{P}_\theta\left(\|\delta_n(X_n) - g(\theta)\| > \varepsilon\right) \longrightarrow 0$$

Usually we omit the index n; sequence is implicit.

Limit Theorems

Let $X_1, X_2, \ldots$ iid random vectors

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Law of large numbers (LLN)

If $\mathbb{E}|X_i| < \infty$, $\mathbb{E}X_i = \mu$, then $\bar{X}_n \xrightarrow{p} \mu$ $\left(\bar{X}_n \xrightarrow{a.s.} \mu\right)$

Central limit theorem (CLT)

If $\mathbb{E}X = \mu \in \mathbb{R}^d$, $\mathrm{Var}(X_n) = \Sigma$ (finite)

Then $\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0, \Sigma)$

[There are stronger versions of both the LLN & CLT, but this will generally be enough for us]

Continuous Mapping

Theorem (Cts Mapping) g cts; $X_1, X_2, \dots$ r.v.s

If $X_n \Rightarrow X$ then $g(X_n) \Rightarrow g(X)$

If $X_n \xrightarrow{p} c$ then $g(X_n) \xrightarrow{p} g(c)$

Proof f bdd, cts $\Rightarrow$ $f \circ g$ bdd, cts

If $X_n \Rightarrow X$ then $\mathbb{E} f(g(X_n)) \longrightarrow \mathbb{E} f(g(X))$

$X_n \xrightarrow{p} c$ special case with $X \sim \delta_c$ $\boxtimes$

Theorem (Slutsky) Assume $X_n \Rightarrow X$, $Y_n \xrightarrow{p} c$

Then: $X_n + Y_n \Rightarrow X + c$

$X_n \cdot Y_n \Rightarrow cX$

$X_n / Y_n \Rightarrow X/c$ if $c \neq 0$

Proof Show $(X_n, Y_n) \Rightarrow (X, c)$, apply cts mapping.

[Wouldn't normally be true that $X_n \Rightarrow X$, $Y_n \Rightarrow Y$ implies $(X_n, Y_n) \Rightarrow (X, Y)$ without specifying joint dist.]

Theorem (Delta Method)

If • $\sqrt{n}\,(X_n - \mu) \Rightarrow N(0, \sigma^2)$

• $f(x)$ differentiable at $x = \mu$

Then $\sqrt{n}\,\big(f(X_n) - f(\mu)\big) \Rightarrow N(0, \dot{f}(\mu)^2 \sigma^2)$

Informal statement:

$$X_n \approx N(\mu, \sigma^2/n) \Rightarrow f(X_n) \approx N(f(\mu), \dot{f}(\mu)\, \sigma^2/n)$$

Proof $f(X_n) = f(\mu) + \dot{f}(\mu)(X_n - \mu) + o(X_n - \mu)$

$$\sqrt{n}\,\big(f(X_n) - f(\mu)\big) = \dot{f}(\mu) \cdot \sqrt{n}\,(X_n - \mu) + \underbrace{\sqrt{n} \cdot o(X_n - \mu)}_{\xrightarrow{P} 0}$$

$$= N(0, \dot{f}(\mu)^2 \sigma^2)$$

Multivariate: $\sqrt{n}\,(X_n - \mu) \Rightarrow N_d(0, \Sigma)$, $f: \mathbb{R}^d \to \mathbb{R}^k$

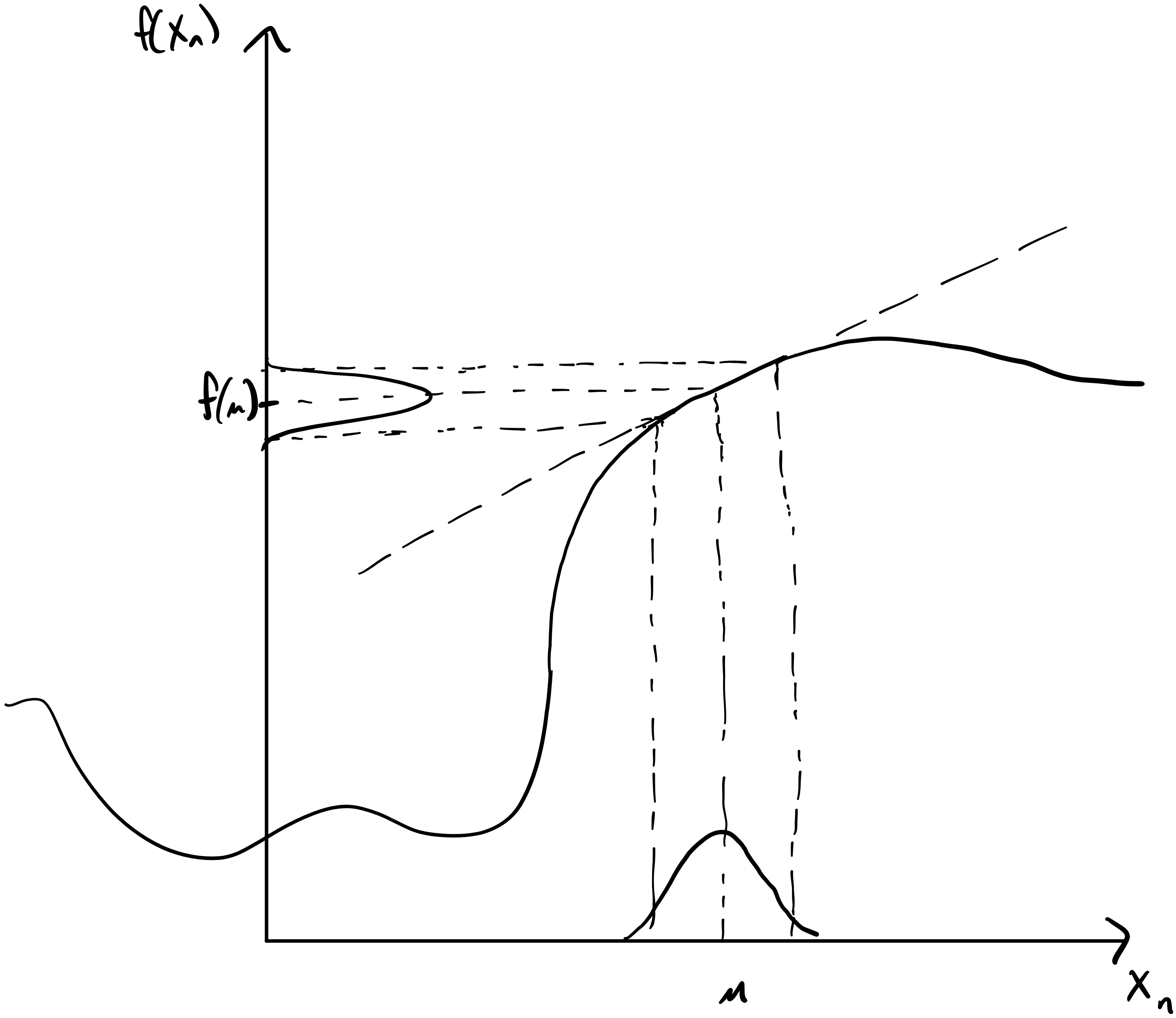
Derivative $Df(x) = \begin{pmatrix} - \nabla f_1(x) - \\ \vdots \\ - \nabla f_k(x) - \end{pmatrix}$ exists at μ

Then $\sqrt{n}\,\big(f(X_n) - f(\mu)\big) \approx \sqrt{n}\, Df(\mu)\,(X_n - \mu)$

$$\approx N_k(0, Df(\mu)\, \Sigma\, Df(\mu)')$$

$$= N(0, \nabla f(\mu)' \Sigma \nabla f(\mu)) \quad \text{if } k = 1$$

Delta Method



[Scaling factor doesn't need to be $\sqrt{n}$,
but need $X_n - \mu \xrightarrow{P} 0$]

Ex $X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$

$Y_1, \dots, Y_n \overset{iid}{\sim} (\nu, \tau^2)$ $\quad X, Y$ indep.

For large n, what is the distribution of $(\bar{X}+\bar{Y})^2$?

1) $\bar{X} \xrightarrow{p} \mu, \quad \bar{Y} \xrightarrow{p} \nu$ as $n \to \infty$

$\Rightarrow (\bar{X}+\bar{Y})^2 \xrightarrow{p} (\mu+\nu)^2$ ✓

2) $\sqrt{n}(\bar{X}-\mu) \Rightarrow N(0,\sigma^2) \qquad \sqrt{n}(\bar{Y}-\nu) \Rightarrow N(0,\tau^2)$

Let $f(x,y) = (x+y)^2$

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

$$f(\bar{X},\bar{Y}) \approx N\left(f(\mu,\nu), \nabla f' \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \nabla f / n \right)$$

$$= N\left((\mu+\nu)^2, 4(\mu+\nu)^2(\sigma^2+\tau^2)/n \right)$$

More accurate:

$$\sqrt{n}\left((\bar{X}+\bar{Y})^2 - (\mu+\nu)^2 \right) \Rightarrow N\left(0, 4(\mu+\nu)^2(\sigma^2+\tau^2)\right)$$

3) What if $(\mu+\nu)^2 = 0$? Conclusion still holds:

$$\sqrt{n}\,(\bar{X}+\bar{Y})^2 \xrightarrow{p} 0$$

Note $\sqrt{n}\bar{X} + \sqrt{n}\bar{Y} \Rightarrow N(0, \sigma^2+\tau^2)$ (cts mapping)

<u>not</u> Slutsky!!

So $n(\bar{X}+\bar{Y})^2 \Rightarrow (\sigma^2+\tau^2)\chi^2_1$ (cts mapping)

why not delta method?

In general, can do higher-order Taylor expansions for delta method if derivatives $= 0$:

$$f(X_n) \approx \underbrace{f(\mu)}_{O(1)} + \underbrace{\dot{f}(\mu)(X_n-\mu)}_{O_p(n^{-1/2})} + \underbrace{\frac{\ddot{f}(\mu)}{2}(X_n-\mu)^2}_{O_p(n^{-1})} + \cdots$$

If $\dot{f}(\mu) = 0$, use second-order term:

$$n\left(f(X_n) - f(\mu)\right) \approx \frac{\ddot{f}(\mu)}{2}\left(\sqrt{n}(X_n-\mu)\right)^2$$

$$\stackrel{\sim}{\approx} \frac{\ddot{f}(\mu)\sigma^2}{2}\chi_1^2$$