STA261 - Module 5

Asymptotic Extensions

Rob Zimmerman

University of Toronto

July 30 - August 1, 2024

Limitations of Finite Sample Sizes

- In almost everything we've done so far, we've assumed a sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ of fixed size n
- We've needed to know the distributions of various statistics of $X_1, X_2, \ldots, X_n$
- This requirement has been very limiting, as the distributions of most statistics don't have closed forms (or are unknown entirely)
- Even the exact distribution of the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is only available for a few parametric families

even though we use $\bar{X}_n$, like, everywhere!

On the other hand, $\bar{X}_n \xrightarrow{p} \mathbb{E}[X_i]$ (assuming the X_i's are iid, $\mathbb{E}[X_i] < \infty$, etc.)

Driving Up the Sample Size

- On the other hand, we have plenty of *limiting* distributions as $n \to \infty$
- Example 5.1: If $X_1, X_2, \ldots \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\bar{X}_n \xrightarrow{p} \mu$ by WLLN and $\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0,1)$ by CLT
- Example 5.2: If $X_n \sim Bin(n, p_n)$ $\forall n$ and $n \cdot p_n \xrightarrow{n\to\infty} \lambda > 0$, then $X_n \xrightarrow{d} \text{Poisson}(\lambda)$ (STA257 or EXERCISE!)
- Of course, we never have $n = \infty$ in real life
- But if we have the luxury of a very large sample size, the "difference" between the exact distribution and the limiting distribution should (hopefully) be tolerable
- Since the normal distribution is particularly nice, we will milk the CLT for all it's worth

A Review of Standard Limiting Results

Note: g need not be defined only on $\mathbb{R}$! For example, $g(\vec{x}) = \sum_{i=1}^{n} x_i$ totally works (and hence we can use the CMT to prove Slutsky)

- In the following, let $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ be sequences of random variables, let X be another random variable, let $x, y \in \mathbb{R}$ be constants, and let $g(\cdot)$ be a continuous function

- Theorem 5.1: If $\overbrace{X_n \xrightarrow{p} X, \text{ then } X_n \xrightarrow{d} X}^{\text{the converse is not true in general; only when } X=x \text{ is constant}}$. If $X_n \xrightarrow{d} x$, then $X_n \xrightarrow{p} x$.

- Theorem 5.2 (**Slutsky's theorem**): If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} y$, then $Y_n \cdot X_n \xrightarrow{d} y \cdot X$ and $X_n + Y_n \xrightarrow{d} X + y$.

("CMT")

- Theorem 5.3 (**Continuous mapping theorem**): If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$. FYI: also true for a.s. convergence

Proofs: STA347 (maybe)

* $X_n \xrightarrow{d} X$ means that $F_{X_n}(x) \xrightarrow{n\to\infty} F_X(x)$ whenever x is a continuity point of $F_X(\cdot)$

* $X_n \xrightarrow{p} X$ means that $\forall \varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n\to\infty} 0$

* $X_n \xrightarrow{a.s.} X$ means that $\forall \varepsilon > 0$, $\mathbb{P}(\lim_{n\to\infty} |X_n - X| > \varepsilon) = 0$ ← FYI; not used in our course

"a.s." = "almost surely"

Notation Update

- For the rest of this module, we will accentuate statistics of finite samples with the subscript n (so $\mathbf{X}$ is now $\mathbf{X}_n$, etc.)

- For a generic statistic, we'll write $T_n = T_n(\mathbf{X}_n)$

- If we're talking about a limiting property of a sequence $\{T_n\}_{n\geq 1}$, we'll abuse notation and just write that T_n has that limiting property, when the meaning is clear from context

- Example 5.3: Instead of writing "the sequence of sample means $\{\overline{X}_n\}_{n\geq 1}$ converges in probability to μ", we'll just write "$\overline{X}_n$ converges in probability to μ" or simply "$\overline{X}_n \xrightarrow{P} \mu$"

Two Big Ones

- Theorem 5.4 (**Weak law of large numbers (WLLN)**): Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$. Then

$$\bar{X}_n \xrightarrow{p} \mu.$$

- Theorem 5.5 (**Central limit theorem (CLT)**): Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Then

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

- The CLT is equivalent to $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, which is the form we'll be using most often

by Slutsky's theorem (EXERCISE!)

Poll Time!

On Quercus: Module 5 - Poll 1

Asymptotic Unbiasedness

- As in Module 2, we're interested in estimators of $\tau(\theta)$
- But now we're concerned with their limiting behaviour as $n \to \infty$
- For finite n, we insisted that our "best" estimators be unbiased
- In the asymptotic setup, we can relax that slightly
- Definition 5.1: Suppose that $\{W_n\}_{n\geq 1}$ is a sequence of estimators for $\tau(\theta)$. If $\text{Bias}_\theta(W_n) \xrightarrow{n\to\infty} 0$ for all $\theta \in \Theta$, then $\{W_n\}_{n\geq 1}$ is said to be **asymptotically unbiased** for $\tau(\theta)$.
- Example 5.4: In the $N(\mu,\sigma^2)$ setup, $\frac{1}{n+1}\sum_{i=1}^n X_i$ is asymptotically unbiased for μ.

  Why? $\mathbb{E}_\mu\left[\frac{1}{n+1}\sum_{i=1}^n X_i\right] = \frac{n}{n+1}\mu$. So $\text{Bias}_\mu\left(\frac{1}{n+1}\sum_{i=1}^n X_i\right) = \mu\left(\frac{n}{n+1} - 1\right) \xrightarrow{n\to\infty} 0$.

Consistency

- $\overline{X}_n \xrightarrow{p} \mu$ is the prototypical example of an estimator converging in probability to the "right thing"

  by the WLLN

- We have a special name for this

- Definition 5.2: A sequence of estimators W_n of $\tau(\theta)$ is said to be **consistent** for $\tau(\theta)$ if $W_n \xrightarrow{p} \tau(\theta)$ for every $\theta \in \Theta$.

- Example 5.5: $X_1, X_2, \ldots \overset{iid}{\sim} Exp(\lambda)$. Then $1/\overline{X}_n^2$ is consistent for λ^2.

  Why? $\overline{X}_n \xrightarrow{p} 1/\lambda$ by WLLN.

  If $g(x) = 1/x^2$, $x \neq 0$, then $g(\overline{X}_n) \xrightarrow{p} g(1/\lambda)$ by CMT

  $= \lambda^2$

  $\Rightarrow 1/\overline{X}_n^2 \xrightarrow{p} \lambda^2$.

  $X_1, X_2, \ldots \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\dfrac{\overline{X}_n^2}{\overline{X^2_n}}$ is consistent for $\dfrac{\mu^2}{\mu^2+\sigma^2}$

  (EXERCISE!)

Showing Consistency

- Sometimes it's easy to show consistency directly from the definition

- Example 5.6: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that the sample mean $\overline{X}_n$ is consistent for μ. Let $\theta := (\mu, \sigma^2)$.

Let $\varepsilon > 0$. Then $\mathbb{P}_\theta(|\overline{X}_n - \mu| < \varepsilon)$

$$= \mathbb{P}_\theta(-\varepsilon < \overline{X}_n - \mu < \varepsilon)$$

$$= \mathbb{P}_\theta\left(\frac{-\varepsilon}{\sqrt{\sigma^2/n}} < \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} < \frac{\varepsilon}{\sqrt{\sigma^2/n}}\right)$$

$$= \mathbb{P}_\theta\left(\frac{-\varepsilon}{\sqrt{\sigma^2/n}} < Z < \frac{\varepsilon}{\sqrt{\sigma^2/n}}\right) \text{ where } Z \sim N(0,1)$$

$$= \Phi\left(\frac{\varepsilon}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\frac{-\varepsilon}{\sqrt{\sigma^2/n}}\right)$$

$$\xrightarrow{n \to \infty} \Phi(\infty) - \Phi(-\infty) = 1.$$

$$\Rightarrow \forall \varepsilon > 0,\ \mathbb{P}_\theta(|\overline{X}_n - \mu| < \varepsilon) \xrightarrow{n \to \infty} 1 \quad \Rightarrow \overline{X}_n \xrightarrow{p} \mu.$$

Showing Consistency

- It's usually easier to use standard limiting results (Slutsky, continuous mapping, etc.) than to go directly from the definition

- Example 5.7: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that the sample variance S_n^2 is consistent for σ^2.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right]$$

$$= \underbrace{\frac{n}{n-1}}_{1}\left[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}_{2} - \underbrace{(\bar{X}_n-\mu)^2}_{3}\right]$$

1 $\xrightarrow{p} 1$

2 $= \overline{(X_i-\mu)^2} \xrightarrow{p} \mathbb{E}[(X_i-\mu)^2] = \sigma^2$

by WLLN

3 $= \overline{X_i-\mu}^{\,2} \xrightarrow{p} \mathbb{E}[X_i-\mu]^2 = 0$

by WLLN + CMT

$$\xrightarrow{d} 1\cdot(\sigma^2 + 0) \quad \text{by Slutsky } (\times 2)$$

$$= \sigma^2$$

$$\Rightarrow S_n^2 \xrightarrow{p} \sigma^2 \quad \text{by Theorem 5.1}$$

Actually, we didn't use the $N(\mu,\sigma^2)$ distribution anywhere here! We showed that S_n^2 is always consistent for $Var(X_i)$ (assuming $Var(X_i) < \infty$)

Bringing Back the MSE

- In Module 2, we compared estimators by their MSEs
- To extend that idea to the asymptotic setup, we need a new mode of convergence
- Definition 5.3: Suppose that W_n is a sequence of estimators for $\tau(\theta)$. If $\text{MSE}_\theta(W_n) \xrightarrow{n\to\infty} 0$ for all $\theta \in \Theta$, then W_n is said to **converge in MSE** to $\tau(\theta)$. "$W_n \xrightarrow{MSE} \tau(\theta)$"
- Example 5.8: $X_1, X_2, \ldots \overset{iid}{\sim} \text{Bin}(K, p)$. Then $\bar{X}_n \xrightarrow{MSE} \underbrace{K \cdot p}_{=\mathbb{E}[X_i]}$.

Why? $\text{MSE}_p(\bar{X}_n) = \underbrace{\text{Bias}_p(\bar{X}_n)^2}_{=0 \text{ since } \bar{X}_n \text{ is always unbiased for } \mathbb{E}[X_i]} + \text{Var}_p(\bar{X}_n)$

$= \text{Var}_p(\bar{X}_n)$

$= \frac{1}{n} Kp(1-p) \xrightarrow{n\to\infty} 0$. So $\bar{X}_n \xrightarrow{MSE} Kp$.

Poll Time!

On Quercus: Module 5 - Poll 2

Convergence in MSE is Already Good Enough

- It turns out that convergence in MSE is strong enough to guarantee consistency

- Theorem 5.6: If W_n is a sequence of estimators for $\tau(\theta)$ that converges in MSE for all $\theta \in \Theta$, then W_n is consistent for $\tau(\theta)$.

Proof. EXERCISE! Hint: use Chebyshev!

(<u>Always</u> remember Chebyshev...)

A Criterion for Consistency

- If we know $\mathbb{E}_\theta[W_n]$ and $\mathrm{Var}_\theta(W_n)$, this next theorem often makes short work out of checking for consistency
- Theorem 5.7: If W_n is a sequence of estimators for $\tau(\theta)$ such that $\mathrm{Bias}_\theta(W_n) \xrightarrow{n\to\infty} 0$ and $\mathrm{Var}_\theta(W_n) \xrightarrow{n\to\infty} 0$ for all $\theta \in \Theta$, then W_n is consistent for $\tau(\theta)$.

Proof. For any $\theta \in \Theta$, $MSE_\theta(W_n) = \underbrace{Bias_\theta(W_n)^2}_{\downarrow n\to\infty \; 0} + \underbrace{Var_\theta(W_n)}_{\downarrow n\to\infty \; 0}$.

$$\Rightarrow MSE_\theta(W_n) \xrightarrow{n\to\infty} 0$$

By Theorem 5.6, W_n is consistent for $\tau(\theta)$. $\square$

The Sample Mean is Always Consistent

- Example 5.9: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} f_\theta$, where $\mathbb{E}[X_i] = \mu$. Show that $\overline{X}_n$ is consistent for μ.

$$\text{Bias}_\theta(\overline{X}_n) = \mathbb{E}_\theta[\overline{X}_n - \mu] = 0.$$

$$\text{Var}_\theta(\overline{X}_n) = \frac{1}{n} \cdot \text{Var}_\theta(X_i) \xrightarrow{n \to \infty} 0.$$

By Theorem 5.7, $\overline{X}_n$ is consistent for μ.

(Also $\overline{X}_n \xrightarrow{P} \mu$ is exactly what the WLLN says)

The Sample Variance is Always Consistent

- One can (very tediously) show that if $X_1, X_2, \ldots, X_n$ are a random sample from a distribution with a finite fourth moment, then

$$\text{Var}\left(S_n^2\right) = \frac{\mathbb{E}\left[(X_i - \mathbb{E}[X_i])^4\right]}{n} - \frac{\text{Var}(X_i)^2 (n-3)}{n(n-1)}$$

(Don't need to memorize!)

- Example 5.10: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$, where $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$ and $\mathbb{E}\left[X_i^4\right] < \infty$. Show that S_n^2 is consistent for σ^2.

$\text{Bias}_{\sigma^2}(S_n^2) = 0$ from Assignment 0.

$$\text{Var}_{\sigma^2}(S_n^2) = \underbrace{\frac{\mathbb{E}_{\sigma^2}[(X_i-\mu)^4]}{n}}_{\xrightarrow{n\to\infty} 0} - \underbrace{\frac{\sigma^4(n-3)}{n(n-1)}}_{\xrightarrow{n\to\infty} 0} \xrightarrow{n\to\infty} 0$$

By Theorem 5.7, S_n^2 is consistent for σ^2.

Choosing Among Consistent Estimators

- Consistency is practically the bare minimum we can ask for from a sequence of estimators
- There are usually plenty of sequences that are consistent for $\tau(\theta)$

  Assignment 5: TONS of examples to play with!
- Which one should we use?
- It's tempting to go with whichever has the lowest variance for fixed n, but that would rule out a lot of fine estimators
- Example 5.11: $X_1, X_2, \ldots \overset{iid}{\sim} \text{Poisson}(\lambda)$, $\lambda > 0$.

$\bar{X}_n$ and S_n^2 are both consistent for λ, by previous stuff. For fixed n, we know (Module 2) that $\bar{X}_n$ is the UMVUE of λ, but does that mean we should just completely ignore S_n^2?

- $X_1, X_2, \ldots \overset{iid}{\sim} N(\mu, \sigma^2)$. S_n^2 and $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ are both consistent for σ^2. Which one should we use?

Asymptotic Normality

- There's a much more useful criterion, but first we need an important CLT-inspired definition

$T_n = T_n(\vec{X}_n)$

- Definition 5.4: Let T_n be a sequence of estimators for $\tau(\theta)$. If there exists some $\sigma^2 > 0$ such that

FYI: the definition extends to where $\sqrt{n}$ and $\tau(\theta)$ are replaced by sequences of constants $\{b_n\}_{n\geq 1}$ and $\{c_n\}_{n\geq 1}$: $b_n(T_n - c_n) \xrightarrow{d} N(0, \sigma^2)$

$$\sqrt{n}[T_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$

then T_n is said to be **asymptotically normal** with mean $\tau(\theta)$ and **asymptotic variance** σ^2.

Note: $\tau(\theta)$ is not necessarily the mean of T_n

- By virtue of the CLT, most unbiased estimators are asymptotically normal

Why not just talk about the distribution of T_n itself as $n \to \infty$?

Usually it's some degenerate distribution (ie., a constant). $\bar{X}_n \xrightarrow{d} \mu$, for example.

The distribution of $\sqrt{n}(\bar{X}_n - \mu)$ as $n \to \infty$ is "more interesting"

Asymptotic Normality: Examples

- Example 5.12: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \text{Bin}\,(k, p)$. Show that the sample mean $\overline{X}_n$ is asymptotically normal.

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}[\overline{X}_n]\right) \xrightarrow{d} N(0, \text{Var}_\theta(X_i)) \text{ by the CLT}$$

$$\Rightarrow \sqrt{n}\left(\overline{X}_n - kp\right) \xrightarrow{d} N(0, kp(1-p))$$

So $\overline{X}_n$ is asymptotically normal with mean kp and asymptotic variance $kp(1-p)$.

Asymptotic Normality: Examples

- Example 5.13: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Show that the second sample moment $\overline{X^2}_n$ is asymptotically normal.

$$\mathbb{E}_\lambda[X_i^2] = \text{Var}_\lambda(X_i) + \mathbb{E}_\lambda[X_i]^2 = 2/\lambda^2$$

$$\text{Var}_\lambda(X_i^2) = \mathbb{E}_\lambda[X_i^4] - \mathbb{E}[X_i^2]^2$$
$$= \frac{4!}{\lambda^4} - \left(\frac{2}{\lambda^2}\right)^2$$
$$= \frac{20}{\lambda^4}.$$

EXERCISE: prove $\mathbb{E}_\lambda[X_i^k] = \frac{k!}{\lambda^k}$.

By the CLT, $\sqrt{n}\left(\overline{X_n^2} - 2/\lambda^2\right) \xrightarrow{d} N\left(0, \frac{20}{\lambda^4}\right)$.

So $\overline{X_n^2}$ is asymptotically normal with mean $2/\lambda^2$ and asymptotic variance $\frac{20}{\lambda^4}$.

Asymptotic Distributions

- More generally, we can talk about the limiting distribution of $\sqrt{n}[T_n - \tau(\theta)]$ even when it's not normal

- Definition 5.5: Suppose that T_n is a sequence of estimators for $\tau(\theta)$. When it exists, the distribution of $\lim_{n\to\infty} \sqrt{n}[T_n - \tau(\theta)]$ is called the **asymptotic distribution** (or **limiting distribution**) of T_n.

In other words, if $\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{d} Y$ for some r.v. Y, then the asymptotic distribution of T_n is exactly the distribution of Y

- So if T_n is an asymptotically normal sequence of estimators for $\tau(\theta)$ with asymptotic variance σ^2, then its asymptotic distribution is $\mathcal{N}\left(0, \sigma^2\right)$

- Example 5.14: $X_1, X_2, \ldots \overset{iid}{\sim} Bin(k, \theta) \Rightarrow \bar{X}_n$ has asymptotic distribution $N(0, k\theta(1-\theta))$ by Example 5.12.

- We might prefer to speak of the distribution of T_n itself when n is large

We can say "for large n, the distribution of $\bar{X}_n$ approaches $N\left(k\theta, \frac{k\theta(1-\theta)}{n}\right)$", $\sqrt{n}(\bar{X}_n - k\theta) \approx N(0, k\theta(1-\theta))$

"$\approx$" = "approximately distributed as"

but we **CANNOT** say "for large n, the distribution of $\bar{X}_n$ <u>is</u> $N\left(k\theta, \frac{k\theta(1-\theta)}{n}\right)$"

... because it's <u>not</u>!

Poll Time!

On Quercus: Module 5 - Poll 3

The Delta Method

- If some sequence T_n is asymptotically normal for θ and some function $g(\cdot)$ is nice enough, then the next result gives a remarkably easy method of producing an asymptotically normal sequence of estimators of for $g(\theta)$

- **Theorem 5.8 (Delta method)**: Suppose that $\theta \in \Theta \subseteq \mathbb{R}$ and $\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$. If $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $g'(\theta) \neq 0$, then

$$\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2\sigma^2\right).$$

Assignment 3: a way to handle the case that $g'(\theta) = 0$.

Proof. Taylor expand $g(T_n)$ around θ to get $g(T_n) = g(\theta) + g'(\tilde{\theta}_n)\cdot(T_n - \theta)$ for some $\tilde{\theta}_n$ between T_n and θ.

$$\Rightarrow \sqrt{n}\left(g(T_n) - g(\theta)\right) = \underbrace{g'(\tilde{\theta}_n)}_{①} \cdot \underbrace{\sqrt{n}(T_n - \theta)}_{②}$$

①: Since $T_n \xrightarrow{P} \theta$ by Slutsky (check!) $\tilde{\theta}_n \xrightarrow{P} \theta$. By CMT, $g(\tilde{\theta}_n) \xrightarrow{P} g(\theta)$.

② $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ by assumption.

By Slutsky,

$$\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow{d} g'(\theta)\cdot N(0, \sigma^2) \stackrel{d}{=} N(0, [g'(\theta)]^2\sigma^2).$$

$\square$

The Delta Method: Examples

- Example 5.15: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R} \setminus \{0\}$ and $\sigma^2 > 0$. Find the limiting distribution of $1/\overline{X}_n$.

Let $g(x) = 1/x$, $x \neq 0$. Then $g'(x) = -1/x^2$, $x \neq 0$.

By the CLT, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

By the delta method, $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, (g'(\mu))^2 \sigma^2)$

$$\Longrightarrow \sqrt{n}\left(1/\bar{X}_n - 1/\mu\right) \xrightarrow{d} N(0, \sigma^2/\mu^4).$$

So $1/\bar{X}_n$ has asymptotic distribution $N(0, \sigma^2/\mu^4)$.

For large n, the distribution of $1/\bar{X}_n$ is approximately $N(1/\mu, \sigma^2/n\mu^4)$.

The Delta Method: Examples

- Example 5.16: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Bernoulli (θ) where $\theta \in (0,1)$. Find the limiting distribution of $\log\left(1 - \overline{X}_n\right)$.

Let $g(x) = \log(1-x)$ for $x \in (0,1)$ $\Rightarrow g'(x) = \frac{-1}{1-x} = \frac{1}{x-1}$, $x \in (0,1)$.

By the CLT, $\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} N(0, \theta(1-\theta))$.

By the delta method, $\sqrt{n}\left(\log(1-\overline{X}_n) - \log(1-\theta)\right) \xrightarrow{d} N\left(0, \left(\frac{1}{\theta-1}\right)^2 \theta(1-\theta)\right)$

$$= N\left(0, \frac{\theta}{1-\theta}\right).$$

So $\log(1-\overline{X}_n)$ has asymptotic distribution $N\left(0, \frac{\theta}{1-\theta}\right)$.

For large n, the distribution of $\log(1-\overline{X}_n)$ is approximately $N\left(\log(1-\theta), \frac{\theta}{n(1-\theta)}\right)$

The Delta Method: Examples

- Example 5.17: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ where $\mathbb{E}_\theta\left[X_i\right] = \theta$ and $\text{Var}_\theta\left(X_i\right) = \sigma^2$. If $\tau : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $\tau'(\theta) \neq 0$, describe the distribution of $\tau(\overline{X}_n)$ as n becomes large.

By the CLT, $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

By the delta method, $\sqrt{n}\left(\tau(\overline{X}_n) - \tau(\mu)\right) \xrightarrow{d} N\left(0, [\tau'(\mu)]^2\sigma^2\right)$.

So the asymptotic distribution of $\tau(\overline{X}_n)$ is $N\left(0, [\tau'(\mu)]^2\sigma^2\right)$.

i.e., the distribution of $\sqrt{n}\left(\tau(\overline{X}_n) - \tau(\mu)\right)$ as $n \to \infty$

For large n, the distribution of $\tau(\overline{X}_n)$ is approximately $N\left(\tau(\mu), \frac{[\tau'(\mu)]^2\sigma^2}{n}\right)$

Back to Choosing Estimators

- We know that when $T_n = \overline{X}_n$, the CLT says that

$$\frac{T_n - \mathbb{E}_\theta\left[T_n\right]}{\sqrt{\mathsf{Var}_\theta\left(T_n\right)}} \xrightarrow{d} \mathcal{N}\left(0, 1\right)$$

- Recall the Fisher information $I_n(\theta) = \mathsf{Var}_\theta\left(S(\theta \mid \mathbf{X}_n)\right)$

- In Module 2, we said that an unbiased estimator W_n of $\tau(\theta)$ was efficient if its variance attained the Cramér-Rao Lower Bound $[\tau'(\theta)]^2/I_n(\theta)$

- We also noticed that if the X_i's were iid, then $I_n(\theta) = nI_1(\theta)$

... by Theorem 2.10, under the same conditions as the CRLB itself

Asymptotic Efficiency

- So if we could replace the T_n in the CLT statement with a general unbiased and efficient W_n, it would look like

$$\frac{W_n - \tau(\theta)}{\sqrt{[\tau'(\theta)]^2/nI_1(\theta)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- Or equivalently

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right)$$

- This is not a *result*, but a *condition* that we can demand of our estimators

- Definition 5.6: A sequence of estimators W_n is **asymptotically efficient** for $\tau(\theta)$ if

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right)$$

Asymptotic Efficiency: Examples

- Example 5.18: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Exp (λ), where $\lambda > 0$. Show that $1/\overline{X}_n$ is asymptotically efficient for λ.

By the CLT, $\sqrt{n}(\overline{X}_n - 1/\lambda) \xrightarrow{d} N(0, 1/\lambda^2)$.

Let $g(x) = 1/x$, $x \neq 0$ $\Rightarrow g'(x) = -1/x^2 \Rightarrow g'(1/\lambda) = -\lambda^2$.

By the delta method, $\sqrt{n}\left(1/\overline{X}_n - \lambda\right) \xrightarrow{d} N(0, \lambda^2)$.

Now, what's $I_1(\lambda)$? $\ell(\lambda|x) = \log(\lambda) - \lambda x$

$\Rightarrow S(\lambda|x) = 1/\lambda - x$

$\Rightarrow -\frac{\partial}{\partial\lambda} S(\lambda|x) = 1/\lambda^2$

$\Rightarrow I_1(\lambda) = \mathbb{E}_\lambda\left[-\frac{\partial}{\partial\lambda} S(\lambda|X)\right] = 1/\lambda^2$.

So the CRLB is $\dfrac{(\tau'(\lambda))^2}{I_1(\lambda)} = \dfrac{1}{1/\lambda^2} = \lambda^2$

Same thing! So we conclude that $1/\overline{X}_n$ is asymptotically efficient for λ.

Asymptotic Efficiency: Examples

- Example 5.19: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Show that $\overline{X}_n$ is asymptotically efficient for λ.

By the CLT, $\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{d} N(0, \lambda)$.

$$L(\lambda|x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$\Rightarrow \ell(\lambda|x) = -\lambda + x \cdot \log(\lambda) + c$, where c is free of λ

$\Rightarrow S(\lambda|x) = -1 + \frac{x}{\lambda}$

$\Rightarrow -\frac{\partial}{\partial\lambda} S(\lambda|x) = \frac{x}{\lambda^2}$

$\Rightarrow I_1(\lambda) = \mathbb{E}_\lambda[-\frac{\partial}{\partial\lambda} S(\lambda|X)] = \frac{1}{\lambda^2} \cdot \mathbb{E}_\lambda[X] = 1/\lambda$

So the asymptotic variance $\overline{X}_n$ is $\frac{(\tau'(\lambda))^2}{I_1(\lambda)} = \lambda$

$\tau(\lambda) = \lambda$

Since the asymptotic variance of $\overline{X}_n$ is equal to the CRLB for unbiased estimators of λ, we conclude that $\overline{X}_n$ is asymptotically efficient for λ.

Large Sample Behaviour for the MLE

- We're ready to see why the MLE is almost always the point estimator of choice when n is large

- To understand this, we need to distinguish between an arbitrary parameter $\theta \in \Theta$ and the true parameter that generated the data, which we will call θ_0

- We'll show that the MLE is asymptotically efficient, under certain "regularity conditions"

Regularity Conditions

- Recall how the Cramér-Rao Lower Bound required some conditions:

① $\text{Var}_\theta(T(\vec{X}_n)) < \infty \ \forall \theta \in \Theta$ ② $\frac{d}{d\theta}\mathbb{E}_\theta[T(\vec{X}_n)] = \int \frac{\partial}{\partial\theta}[T(\vec{x}) \cdot f_\theta(\vec{x})]\, d\vec{x}$

(i.e., we can push the derivative inside the integral)

- Such conditions are generically referred to as *regularity conditions*, and they're used to rule out various pathological counterexamples and edge cases

- The exact regularity conditions for our next result are quite technical and not worth getting involved with in this course

- Instead, we will go with four *sufficient* regularity conditions that are relatively easy to check, and which are satisfied by many common parametric models

Poll Time!

On Quercus: Module 5 - Poll 4

Unif$(0,\theta)$ does <u>not</u> satisfy $\frac{d}{d\theta}\int_{\mathcal{X}} \cdots = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} \cdots$

because the support $\mathcal{X} = (0,\theta)$ depends on θ.

The MLE is Often Asymptotically Normal

- Theorem 5.9: Let $X_1, X_2, \ldots \overset{iid}{\sim} f_{\theta_0}$, and let $\hat{\theta}_n(\mathbf{X}_n)$ be the MLE of θ_0 based on a sample of size n. Suppose the following regularity conditions hold:
  - Θ is an open interval (not necessarily finite) in $\mathbb{R}$
  - The log-likelihood $\ell(\theta \mid \mathbf{x}_n)$ is three times continuously differentiable in θ
  - The support of f_θ does not depend on θ
  - $I_1(\theta) < \infty$ for all $\theta \in \Theta$

  Then

$$\sqrt{n}[\hat{\theta}_n(\mathbf{X}_n) - \theta_0] \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I_1(\theta_0)}\right).$$

  That is, $\hat{\theta}_n(\mathbf{X}_n)$ is a consistent and asymptotically efficient estimator of θ_0.

Write $\hat{\theta}_n = \hat{\theta}_n(\vec{X}_n)$ for simplicity.

Proof (sketch). Take a Taylor series of $\ell'(\hat{\theta}_n|\vec{x})$ around θ_0. For large n, we get

$\ell'(\hat{\theta}_n|\vec{x}) \simeq \ell'(\theta_0|\vec{x}) + (\hat{\theta}_n - \theta_0)\cdot \ell''(\theta_0|\vec{x})$ with equality as $n \to \infty$ (this is where those regularity conditions are needed — trust me on this!)

$\Rightarrow \quad 0 \simeq \ell'(\theta_0|\vec{x}) + (\hat{\theta}_n - \theta_0)\cdot \ell''(\theta_0|\vec{x})$

$\Rightarrow \hat{\theta}_n - \theta_0 \simeq \dfrac{-\ell'(\theta_0|\vec{x})}{\ell''(\theta_0|\vec{x})}$

$$\rightarrow \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \simeq \frac{-\frac{1}{\sqrt{n}}\, l'(\theta_0|\vec{X}) \quad ①}{\frac{1}{n}\, l''(\theta_0|\vec{X}) \quad ②}$$

① $-\frac{1}{\sqrt{n}}\, l'(\theta_0|\vec{X}) = -\frac{1}{\sqrt{n}} S(\theta_0|\vec{X})$

$$= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(\theta_0|X_i)$$

$$= \sqrt{n}\left(-\frac{1}{n}\sum_{i=1}^{n} S(\theta_0|X_i) - 0\right)$$

$$= \sqrt{n}\left(-\frac{1}{n}\sum_{i=1}^{n} S(\theta_0|X_i) - \mathbb{E}_{\theta_0}[S(\theta_0|X_i)]\right)$$

$$= -\sqrt{n}\left(\overline{S(\theta_0|X)} - \mathbb{E}_{\theta_0}[S(\theta_0|X_i)]\right)$$

$$\xrightarrow{d} -N\left(0, \mathrm{Var}_{\theta_0}(S(\theta_0|X))\right) \text{ by the CLT}$$

$$= -N(0, I_1(\theta_0))$$

②

$$\frac{1}{n}\, l''(\theta_0|\vec{X})$$

$$= \frac{1}{n}\frac{\partial}{\partial\theta} S(\theta|\vec{X})\Big|_{\theta=\theta_0}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta} S(\theta|X_i)\Big|_{\theta=\theta_0}$$

$$= \overline{\frac{\partial}{\partial\theta} S(\theta|X)\Big|_{\theta=\theta_0}}$$

$$\xrightarrow{P} \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta} S(\theta|X)\Big|_{\theta=\theta_0}\right]$$

$$= -I_1(\theta_0)$$

By Slutsky's theorem, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \frac{1}{I_1(\theta_0)} \cdot N(0, I_1(\theta_0)) = N(0, 1/I_1(\theta_0))$.

So $\hat{\theta}_n$ is asymptotically efficient! Consistency follows from Slutsky's theorem (again). "$\square$"

A Useful Corollary

- Theorem 5.10: Suppose the hypotheses of Theorem 5.9 hold, and that $\tau : \Theta \to \mathbb{R}$ is continuously differentiable with $\tau'(\theta_0) \neq 0$. Then

$$\sqrt{n}[\tau(\hat{\theta}_n(\mathbf{X}_n)) - \tau(\theta_0)] \xrightarrow{d} \mathcal{N}\left(0, \frac{[\tau'(\theta_0)]^2}{I_1(\theta_0)}\right).$$

  That is, $\tau(\hat{\theta}_n(\mathbf{X}_n))$ is a consistent and asymptotically efficient estimator of $\tau(\theta_0)$.

Proof: EXERCISE!

Asymptotically Efficient MLEs: Examples

- Example 5.20: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Find the asymptotic distribution of the MLE of μ. $(\hat{\mu}_n = \overline{X}_n)$.

Check the conditions of Theorem 5.9:

$\ell(\mu \mid x) = c - \frac{(x-\mu)^2}{2\sigma^2}$, where c is free of μ

$\ell'(\mu \mid x) = \frac{x-\mu}{\sigma^2}$

$\ell''(\mu \mid x) = -\frac{1}{\sigma^2} \Rightarrow I_1(\mu) = E_\mu\left[-\left(-\frac{1}{\sigma^2}\right)\right] = 1/\sigma^2$

$\ell'''(\mu \mid x) = 0$

(i) $\Theta = (-\infty, \infty) \subseteq \mathbb{R}$ is open in $\mathbb{R}$ ✓

(2) $\ell'''(\mu \mid x) = 0$, which is continuous in μ ✓

(3) $f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} > 0 \quad \forall x \in \mathbb{R}$

so the support of f_μ is $\mathbb{R}$, which doesn't depend on μ ✓

(4) $I_1(\mu) = 1/\sigma^2 < \infty \quad \forall \mu \in \mathbb{R}$ ✓

By Theorem 5.9, $\hat{\mu}_n = \bar{X}_n$ is asymptotically efficient, with asymptotic distribution $N(0, \sigma^2)$

Asymptotically Efficient MLEs: Examples

- Example 5.21: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p), where $p \in (0,1)$. Find the asymptotic distribution of the MLE of p, and then that of $1/p$.

EXERCISE!

Use the delta method for $1/p$.

The MLE Isn't Always Asymptotically Normal

- Example 5.22: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$, where $\theta > 0$. Show that the MLE of θ is not asymptotically normal.

$\hat{\theta}_n = X_{(n)}$.

If $\sqrt{n}(X_{(n)} - \theta) \to N(0, ?)$, then $Y_n := \sqrt{n}(\theta - X_{(n)}) \to N(0, ?)$ too.

But... $\mathbb{P}_\theta(Y_n \le y)$

$$= \mathbb{P}_\theta(\theta - X_{(n)} \le y/\sqrt{n})$$

$$= \mathbb{P}_\theta(X_{(n)} \ge \theta - y/\sqrt{n})$$

$$= 1 - \left(\frac{\theta - y/\sqrt{n}}{\theta}\right)^n$$

$$= 1 - \left(1 - \frac{y}{\sqrt{n}\theta}\right)^n$$

$$\xrightarrow{n \to \infty} \begin{cases} 1, & y/\theta \ge 0 \\ 0, & y/\theta < 0 \end{cases} = \mathbb{1}_{y \ge 0}$$

(i.e., the asymptotic distribution of Y_n is degenerate at 0, so it's not a normal random variable!)

EXERCISE: sometimes different scalings of $T_n - \theta$ give us interesting results (e.g., if $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $1 \cdot (\bar{X}_n - \mu) \xrightarrow{d} \mathbb{1}_{x \le 0}$ but $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.)

In the $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$ case, what - if anything - does $n\left(\frac{n+1}{n} \cdot X_{(n)} - \theta\right)$ converge in distribution to?

Approximate Tests and Intervals

- We've seen that a lot of statistics are asymptotically normal
- What about test statistics?
- If we're willing to approximate a test statistic (whose exact distribution we might not know for fixed n) by one with a normal distribution, we can perform tests and create intervals that we couldn't have before
- As in Modules 3 and 4, we'll start off with tests and then use the test statistics from those to construct confidence intervals

Wilks' Theorem

- Recall the LRT statistic for testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ was given by $\lambda(\mathbf{X}_n) = \frac{L(\theta_0|\mathbf{X}_n)}{L(\hat{\theta}_n|\mathbf{X}_n)}$, where $\hat{\theta}_n = \hat{\theta}(\mathbf{X}_n)$ is the unrestricted MLE of θ based on $\mathbf{X}_n$

- Amazingly, the LRT statistic always converges in distribution to a known distribution, regardless of the statistical model (assuming it's nice enough)

- Theorem 5.11 (**Wilks' theorem**): Let $X_1, X_2, \ldots \overset{iid}{\sim} f_\theta$, where the model satisfies the same regularity conditions as in Theorem 5.9. If we test $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^c$ using $\lambda(\mathbf{X}_n)$, then

$$-2 \log (\lambda(\mathbf{X}_n)) \xrightarrow{d} \chi^2_{(1)}$$

under H_0.

Proof: EXERCISE!

Poll Time!

On Quercus: Module 5 - Poll 5

Approximate LRTs: Examples

- Example 5.23: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p), where $p \in (0,1)$. Construct an approximate size-α LRT of $H_0 : p = p_0$ versus $H_A : p \neq p_0$.

Example 3.23 $\Rightarrow \lambda(\vec{X}_n) = \left(\frac{p_0}{\bar{X}_n}\right)^{\Sigma X_i} \left(\frac{1-p_0}{1-\bar{X}_n}\right)^{n-\Sigma X_i}$

$\Rightarrow \log(\lambda(\vec{X}_n)) = n\left[\bar{X}_n \cdot \log\left(p_0/\bar{X}_n\right) + (1-\bar{X}_n)\cdot \log\left(\frac{1-p_0}{1-\bar{X}_n}\right)\right]$

$\Rightarrow -2\cdot\log(\lambda(\vec{X})) = -2n\left[\bar{X}_n \cdot \log\left(p_0/\bar{X}_n\right) + (1-\bar{X}_n)\cdot \log\left(\frac{1-p_0}{1-\bar{X}_n}\right)\right]$

By Wilks' theorem, $R = \left\{\vec{x} \in \mathcal{X}^n : -2n\left[\bar{x}\cdot\log\left(p_0/\bar{x}\right) + (1-\bar{x})\cdot\log\left(\frac{1-p_0}{1-\bar{x}}\right)\right] \geq \chi^2_{(1),\alpha}\right\}$

is the rejection region of an approximate size-α test of $H_0 : \theta = \theta_0$ vs $H_A : \theta \neq \theta_0$.

Approximate LRTs: Examples

- Example 5.24: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$, σ^2 known. Construct an approximate size-α LRT of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$.

Example 3.21 $\longrightarrow \lambda(\vec{X}_n) = \exp\left(-\frac{n}{2\sigma^2}(\bar{X}_n - \mu_0)^2\right)$

$$\Longrightarrow -2 \cdot \log(\lambda(\vec{X})) = \frac{n}{\sigma^2}(\bar{X}_n - \mu_0)^2$$

By Wilks' theorem, $R = \{\vec{x} \in \mathcal{X}^n : \frac{n}{\sigma^2}(\bar{x} - \mu_0)^2 \geq \chi^2_{(1),\alpha}\}$ is the rejection region of an approximate size-α test of $H_0: \mu = \mu_0$ vs. $H_A: \mu \neq \mu_0$.

In fact, it's an <u>exact</u> size-α test! Why? Compare to a Z-test...

Wald Tests

- Definition 5.7: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$. For testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$, a **Wald test** is a test based on the **Wald statistic**

$$W_n(\mathbf{X}_n) = (\hat{\theta}_n - \theta_0)^2 \underbrace{I_n(\hat{\theta}_n)}_{\text{"plug-in Fisher information"}},$$

where $\hat{\theta}_n = \hat{\theta}_{\text{MLE}}(\mathbf{X}_n)$ is the usual MLE.

- Theorem 5.12: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$, where the model satisfies the same regularity conditions as in Theorem 5.9. If we test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ using $W_n(\mathbf{X}_n)$, then

$$W_n(\mathbf{X}_n) \xrightarrow{d} \chi^2_{(1)}$$

under H_0.

Proof: EXERCISE!

Wald Tests: Examples

- Example 5.25: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p), where $p \in (0,1)$. Construct an approximate size-α Wald test of $H_0 : p = p_0$ versus $H_A : p \neq p_0$.

$W_n(\vec{X}_n) = (\hat{p}_n - p_0)^2 \cdot I_n(\hat{p}_n)$, where $\hat{p}_n = \bar{X}_n$.

See Slide 50

What's the Fisher information? $I_n(p) = \dfrac{n}{p(1-p)} \Rightarrow I_n(\hat{p}_n) = \dfrac{n}{\bar{X}_n(1-\bar{X}_n)}$.

So $W_n(\vec{X}_n) = \dfrac{(\bar{X}_n - p_0)^2 \cdot n}{\bar{X}_n(1-\bar{X}_n)} \xrightarrow{d} \chi^2_{(1)}$ under H_0, by Theorem 5.12.

So $R = \left\{ \vec{x} \in \mathcal{X}^n : \dfrac{(\bar{x}-p_0)^2 \cdot n}{\bar{x}(1-\bar{x})} > \chi^2_{(1),\alpha} \right\}$ is the rejection region of an approximate size-α test of $H_0: p = p_0$ vs $H_A: p \neq p_0$.

OR: $R' = \left\{ \vec{x} \in \mathcal{X}^n : \left| \dfrac{\bar{x} - p_0}{\sqrt{\bar{x}(1-\bar{x})/n}} \right| > z_{\alpha/2} \right\}$ is the rejection region of an approximate size-α test of $H_0: p = p_0$ vs $H_A: p \neq p_0$.

EXERCISE: does $R = R'$?

Wald Tests: Examples

σ^2 known!

- Example 5.26: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$. Construct an approximate size-α Wald test of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$. $\hat{\mu}_n = \bar{X}_n$

From Example 5.20, $I_n(\mu) = n/\sigma^2$, so $W_n(\vec{x}_n) = \frac{(\bar{x}_n - \mu_0)^2}{\sigma^2/n} = \left(\frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}}\right)^2$.

By Theorem 5.12, $R = \{\vec{x} \in \mathcal{X}^n : \frac{(\bar{x}-\mu_0)^2}{\sigma^2/n} > \chi^2_{(1),\alpha}\}$ is the rejection region of an approximate (exact, in this case!) size-α test of $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$.

OR: $R' = \{\vec{x} \in \mathcal{X}^n : \left|\frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}}\right| > z_{\alpha/2}\}$ is the rejection region of an approximate (exact) size-α test of $H_0: \mu = \mu_0$ vs $H_A: \mu \neq \mu_0$.

It's our old friend, the two-sided z-test!

Score Tests

- Definition 5.8: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$. For testing $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$, a **score test** (also called a **Rao test** or a **Lagrange multiplier test**) is a test based on the **score statistic**

$$R_n(\mathbf{X}_n) = \frac{[S_n(\theta_0 \mid \mathbf{X}_n)]^2}{I_n(\theta_0)}.$$

- Theorem 5.13: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$, where the model satisfies the same regularity conditions as in Theorem 5.9. If we test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ using $R_n(\mathbf{X}_n)$, then

$$R_n(\mathbf{X}_n) \xrightarrow{d} \chi^2_{(1)}$$

under H_0.

Equivalently, $\dfrac{S_n(\theta_0 \mid \vec{X}_n)}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} N(0,1).$

Score Tests: Examples

- Example 5.27: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (p), where $p \in (0,1)$. Construct an approximate size-α score test of $H_0 : p = p_0$ versus $H_A : p \neq p_0$.

$$R_n(\vec{X}_n) = \frac{S(p_0|\vec{X}_n)^2}{I_n(p_0)}$$

$$= n^2\left(\frac{\bar{X}_n}{p_0} - \frac{1-\bar{X}_n}{1-p_0}\right)^2 \cdot \frac{p_0(1-p_0)}{n}$$

$$= \frac{(\bar{X}_n - p_0)^2}{p_0(1-p_0)/n}$$

$$L(p|\vec{x}) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ell(p|\vec{x}) = \sum x_i \cdot \log(p) + (n-\sum x_i)\cdot \log(1-p)$$

$$S(p|\vec{x}) = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right)$$

$$S'(p|\vec{x}) = n\left(\frac{-\bar{x}}{p^2} - \frac{1-\bar{x}}{(1-p)^2}\right)$$

$$I_n(p) = -\mathbb{E}_p\left[n\left(\frac{\bar{X}_n}{p^2} - \frac{1-\bar{X}_n}{(1-p)^2}\right)\right] = \frac{n}{p(1-p)}$$

By Theorem 5.13, $R = \left\{\vec{x} \in \mathcal{X}^n : \frac{(\bar{x}-p_0)^2}{p_0(1-p_0)/n} > \chi^2_{(1),\alpha}\right\}$ is the rejection region of an approximate size-α test of $H_0: p = p_0$ vs $H_A: p \neq p_0$.

Score Tests: Examples

- Example 5.28: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$. Construct an approximate size-α score test of $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$.

EXERCISE!

The Trinity of Tests

- The LRT, the Wald test, and the score test form the backbone of classical hypothesis testing
- Observe that under H_0, all three tests are asymptotically equivalent (i.e., all three test statistics all converge in distribution to a $\chi^2_{(1)}$)
- For this reason, the three tests are sometimes collectively referred to as the **trinity of tests**
- Although asymptotically equivalent, the speed of convergence to $\chi^2_{(1)}$ can be quite different for each one – for small n, they can be quite different in terms of power and other "small-sample" properties
- One might tell you to reject H_0 while another might not!

FYI: if $\ell(\theta|x) = a\theta^2 + b\theta + c$ for some $a, b, c \in \mathbb{R}$, then all three tests are equivalent for finite n (proved in 1982)

Approximate Confidence Intervals

- Using any of the asymptotic tests to test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$, it's sometimes possible to invert any of the test statistics to obtain an approximate $(1 - \alpha)$-confidence interval for θ

- Out of the three, the LRT is usually the hardest to invert into an actual interval, and the Wald statistic is usually the easiest

- In practice, you can always try to use numerical solvers when the algebra doesn't work

- For Wald and score intervals, the standard recipe is to take the square root of the test statistic and compare it to $\mathcal{N}(0, 1)$

Approximate Confidence Intervals: Examples

- Example 5.29: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim}$ Bernoulli (p), where $p \in (0,1)$. Construct an approximate $(1-\alpha)$-confidence interval for p based on the Wald statistic.

Example 5.25, $1-\alpha \simeq \mathbb{P}_p\left(\frac{|\bar{X}_n - p|}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n}} < z_{\alpha/2} \right)$ when n is large

$$= \mathbb{P}_p\left(-z_{\alpha/2} < \frac{p-\bar{X}_n}{\sqrt{\bar{X}_n(1-\bar{X}_n)/n}} < z_{\alpha/2} \right)$$

$$= \mathbb{P}_p\left(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} < p < \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right)$$

$\Rightarrow \left(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right)$ is an approximate $(1-\alpha)$-CI for p.

- This confidence interval shows up everywhere in polling (and is a staple of introductory Statistics classes); its half-length is called the **margin of error**

In practice you almost always see $\alpha = 0.05$ (thanks, Fisher...), whence $z_{\alpha/2} \approx 1.96$

Approximate Confidence Intervals: Examples

- Example 5.30: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p), where $p \in (0,1)$. Construct an approximate $(1-\alpha)$-confidence interval for $\log\left(\frac{p}{1-p}\right)$ based on the Wald statistic.

From Example 5.29,

since $p \mapsto \log\left(\frac{p}{1-p}\right)$ is a monotone increasing bijection

$$1-\alpha \simeq \mathbb{P}_p\left(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} < p < \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right)$$

$$= \mathbb{P}_p\left(\log\left(\frac{\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1-\bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}\right) < \log\left(\frac{p}{1-p}\right) < \log\left(\frac{\bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1-\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}\right)\right)$$

So $\left(\log\left(\frac{\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1-\bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}\right), \log\left(\frac{\bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1-\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}}\right)\right)$ is an approximate $(1-\alpha)$-CI for $\log\left(\frac{p}{1-p}\right)$.

Approximate Confidence Intervals: Examples

- Example 5.31: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Construct an approximate $(1-\alpha)$-confidence interval for λ based on the Wald statistic.

$\hat{\lambda}_n = \overline{X}_n$.

$\ell(\lambda|\vec{x}) = -n\lambda + \sum x_i \cdot \log(\lambda) + c$, where...

$\Rightarrow S(\lambda|\vec{x}) = -n + \dfrac{\sum x_i}{\lambda}$

$\Rightarrow S'(\lambda|\vec{x}) = \dfrac{-\sum x_i}{\lambda^2}$

$\Rightarrow I_n(\lambda) = -\mathbb{E}_\lambda\left[\dfrac{-\sum x_i}{\lambda^2}\right] = n/\lambda$

$\Rightarrow I_n(\hat{\lambda}_n) = n/\overline{X}_n$

So $W_n(\vec{X}_n) = \dfrac{(\overline{X}_n - \lambda)^2 \cdot n}{\overline{X}_n} = \dfrac{(\lambda - \overline{X}_n)^2}{\overline{X}_n/n}$

$$\Rightarrow 1-\alpha \simeq \mathbb{P}_\lambda\left(-z_{\alpha/2} < \frac{\lambda - \overline{X}_n}{\sqrt{\overline{X}_n/n}} < z_{\alpha/2}\right)$$

$$= \mathbb{P}_\lambda\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}} < \lambda < \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}\right)$$

So $\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}}\right)$ is an approximate $(1-\alpha)$-CI for λ.

When the Fisher Information Causes Problems...

- When f_θ is too complicated to allow for exact $(1-\alpha)$-confidence intervals, it's standard practice to use Wald intervals and score intervals

- But there might be another problem: calculating the Fisher information!

- In real-life multiparameter models, $I_n(\boldsymbol{\theta})$ is a matrix and is often impossible to work out directly, which makes calculating $I_n(\hat{\boldsymbol{\theta}}_0)$ or $I_n(\hat{\boldsymbol{\theta}})$ futile

- When this happens, people like to swap $I_n(\cdot)$ with $J_n(\cdot)$ in the Wald and score statistics .. but is this actually justified???

- Yes! It can be shown that $J_n(\vec{X}_n)$ is a consistent estimator of $I_n(\theta_0)$

- Moreover, in a famous 1978 paper, Efron and Hinkley showed empirically that $\cancel{J_n(\hat{\theta})}$ is *superior* to $I_n(\hat{\theta})$

  $J_n(\vec{X}_n)$

  Optional reading, if you're curious....