Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Интервальные оценки.

ПМИ ФКН ВШЭ, 14 сентября 2019 г.

Денис Деркач[1], Алексей Артёмов[1,2]

[1] ФКН ВШЭ [2] Сколтех

Оглавление

Интервальные оценки

Интервальные оценки: мотивация

- Обычно мы пытаемся измерить параметр на конечной выборке.
- Было бы интересно понять не только точечную оценку из имеющейся выборки, но и предположение о том, где лежит настоящее значение параметра (классический подход) или насколько мы уверены в полученном значении параметра (байесовкий подход).

Как всегда, у нас возникают разные подходы к решению задачи, в зависимости от интерпретации вероятностей.

Байесовские доверительные интервалы

Байесовский доверительный интервал (credibility interval)

Определение

Байесовский p-доверительный интервал — это интервал $[L, U]$, к котором значение параметра θ принадлежит с апостериорной вероятностью p:

$$\mathbb{P}(L \leqslant \theta \leqslant U | X) = p.$$

NB: уровень доверия сокращают Cr.L. (часто используют C.L.).

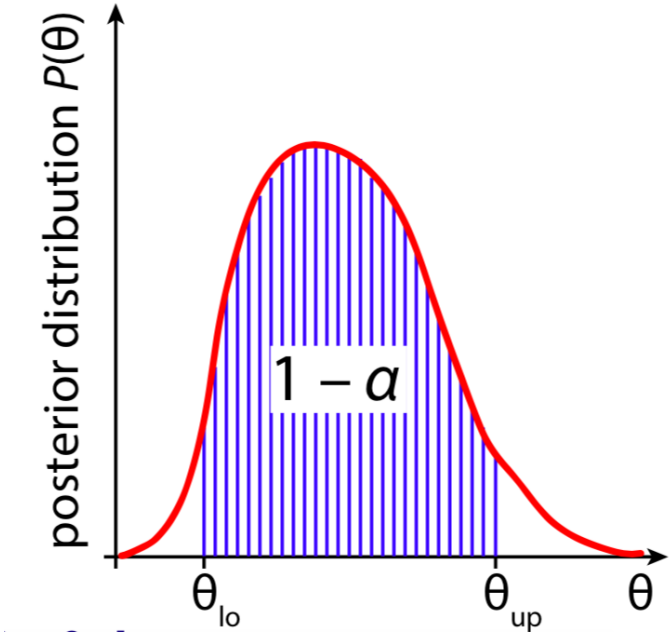
Байесовский доверительный интервал

Байесовский подход:

$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|X)d\theta$$

Подходы к выбору θ_{lo} и θ_{hi}:

- HPD (highest probability density) — брать только наиболее высокие вероятности.
- Центральный интервал - интегрировать от пика.
- Односторонний интервал — интегрировать от бесконечности.



Мешающие параметры

Определение

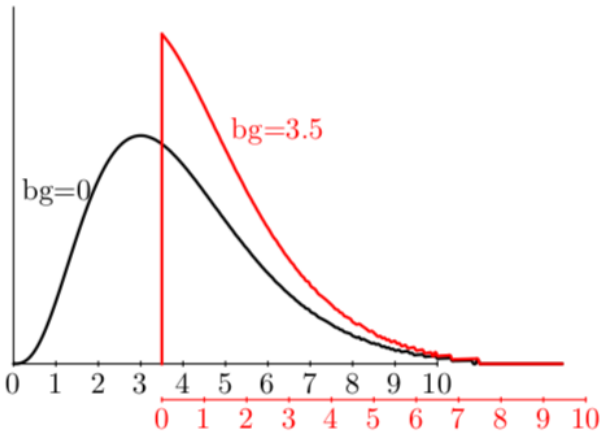
Мешающий параметр (Nuisance parameters) — любой неизвестный параметр вероятностного распределения в статистической задаче, связанной с изучением других параметров данного распределения.

В байесовском подходе, включение мешающих параметров также происходит простым способом (если нам известно его распределение $P(b)$, просто проинетгрируем по нему):

$$\mathbb{P}(\theta|\text{data}) = \int_b \frac{\mathbb{P}(\text{data}|\theta, b)\mathbb{P}(\theta)}{\mathbb{P}(\text{data})}\mathbb{P}(b)db.$$

Такой подход называется маргинализация функции распределения.

Поведение вблизи границ



Bayesian 90% Upper Limits (Uniform Prior)

observed =	0	1	2	3
background = 0.0	2.30	3.89	5.32	6.68
0.5	2.30	3.50	4.83	6.17
1.0	2.30	3.26	4.44	5.71
2.0	2.30	3.00	3.87	4.92
3.0	2.30	2.83	3.52	4.37

Поведение вблизи границ получается в байесовском подходе очень просто — мы используем априорное распределение с информацией о физической границе.

Полученные результаты очень логичны, если использовать плоское априорное распределение с чёткой левой границей.

Комбинирование измерений

Ещё одним хорошим свойством байесовского подхода является простая комбинация нескольких измерений:

$$\mathbb{P}(\theta|\mathrm{data}) = \frac{\mathbb{P}_1(\mathrm{data}|\theta)\ldots\mathbb{P}_N(\mathrm{data}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathrm{data})}.$$

При этом достаточно использовать только одно априорное распределение.
NB: иногда бывает полезно считать произведение в несколько заходов.

Априорная вероятность

В процессе определения границ эксперимента, мы заботились только о левой границе. А что происходит с правой? Должна ли она быть на бесконечности? в таком случае:

$$\int_{b}^{a} \mathrm{Uniform}(x)dx = 0, \forall a, b$$

То есть мы должны также ограничивать правую сторону, причём о ней у нас нет (или почти нет информации).

Кроме того, использование плоского априорного распределения довольно случайно ;-)

Априорная вероятность Джеффриса

Изначально вводилась так, чтобы быть инвариантной относительно некоторых преобразований координат. Для каждого семейства кривых, вероятность Джеффриса может быть подсчитана из этого условия. Например, для распределения Пуассона Джеффрис предлагал её сделать инвариантной к маштабу $\sim 1/\mu$.

Bayesian 90% Upper Limits ($1/\mu$ Jeffreys Prior)

observed =	0	1	2	3
background = 0.0	0.00	2.30	3.89	5.32
0.5	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00
2.0	0.00	0.00	0.00	0.00
3.0	0.00	0.00	0.00	0.00

Априорная вероятность Джеффриса

Сейчас предпочитают использовать распределения, которые минимизируют информацию Фишера. Для Пуассоновского распределения:

$$P(\mu) = \frac{1}{\sqrt{\mu}}.$$

Что не даёт правильных интервалов в присутствии шума. Исправим:

$$P(\mu) = \frac{1}{\sqrt{\mu + b}}.$$

То есть, наше априорное знание о сигнале зависит от знания о шуме :-(

Доверительные интервалы

Доверительные интервалы (confidence intervals)

Определение

Доверительный интервал — это интервал, построенный с помощью случайной выборки из распределения с неизвестным параметром, такой, что он содержит данный параметр с заданной вероятностью. То есть

$$\mathbb{P}(L \leqslant \theta \leqslant U) = p.$$

Заметим, что в байесовском подходе мы оцениваем

$$\mathbb{P}(L \leqslant \theta \leqslant U | X)$$

Покрытие (coverage, capture)

С точки зрения классической статистики.

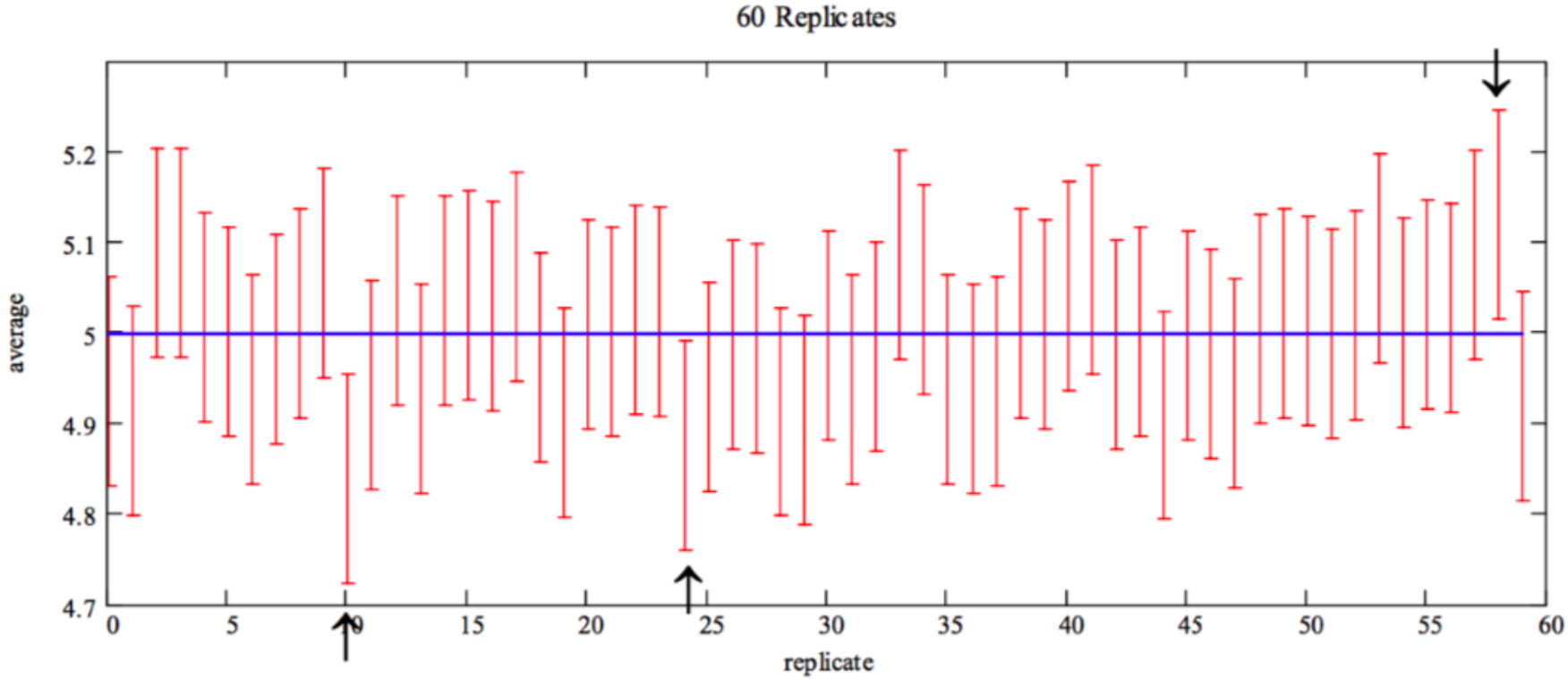
Определение
Вероятность покрытия интервальной оценки представляет собой долю случаев, в которых содержится выборочная статистика, полученная из бесконечных независимых и идентичных повторений эксперимента.

NB: наличие ”покрытия” у байесовского подхода под вопросом.

Доверительные интервалы на практике

Пример

60 экспериментов $X \sim \mathcal{N}(5; 0.1)$. 95% доверительные интервалы задаются $\mu = \text{avg} \pm 0.116$



Получилось, что $3/60 \approx 0.05$ интервалов не содержат настоящее значение.

Покрытие частотных интервалов

На практике, в основном, используются методы, обладающие ассимптотическим покрытием. Каждый метод для данной задачи имеет наблюдаемое покрытие (observed coverage). Если наблюдаемое покрытие $\mathbb{P} \leq \beta$, говорят о "недопокрытии" (undercoverage), если $\mathbb{P} \geq \beta$ о "перепокрытии" (overcoverage).

В принципе, overcoverage — меньшая проблема (но с точки зрения экспериментатора это ухудшает качества эксперимента).

Нормальная теория

Пусть мы берём $X \sim N(\mu; \sigma^2)$. Для известных μ и σ^2:

$$\beta = \mathbb{P}(a \leq X \leq b) = \int_a^b N(\mu, \sigma^2) dX'.$$

При этом, если μ неизвестна, мы больше не сможем подсчитать этот интеграл, вместо этого мы можем оценить вероятность $[\mu + c, \mu + d]$:

$$\beta = \mathbb{P}(\mu + c < X < \mu + d) = \int_{\mu+c}^{\mu+d} N(\mu, \sigma^2) dX' =$$

$$= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp[\frac{1}{2} Y^2] dY.$$

То есть, мы можем переписать как $\beta = \mathbb{P}(X - d \leq \mu \leq X - c)$

Нормальная теория для интервальных оценок

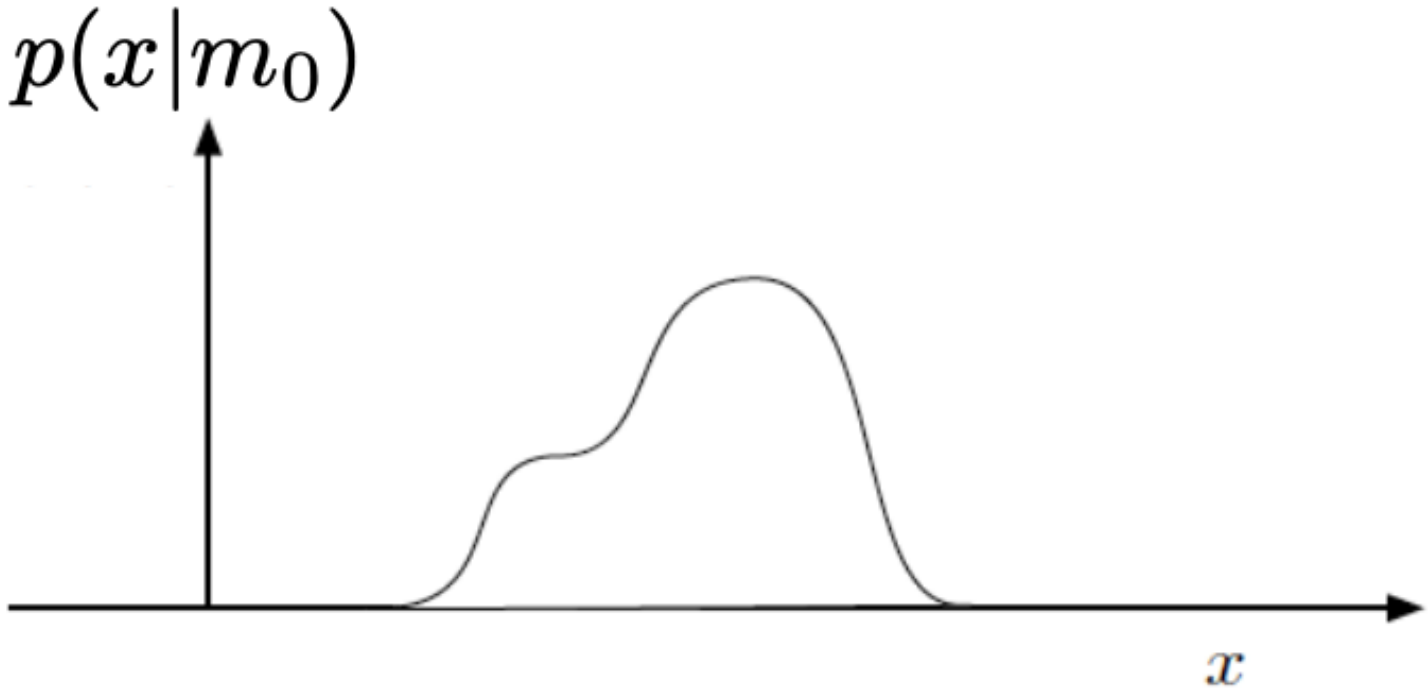
Такого рода оценка сработала так как:

- была получена функция от $(X - \mu)^2$;
- мы подразумевали, что функция интегрируема и область интегрирования не имеет границ.

Если вспомнить свойства оценки правдоподобия, то асимптотически эти пункты будут выполнены.

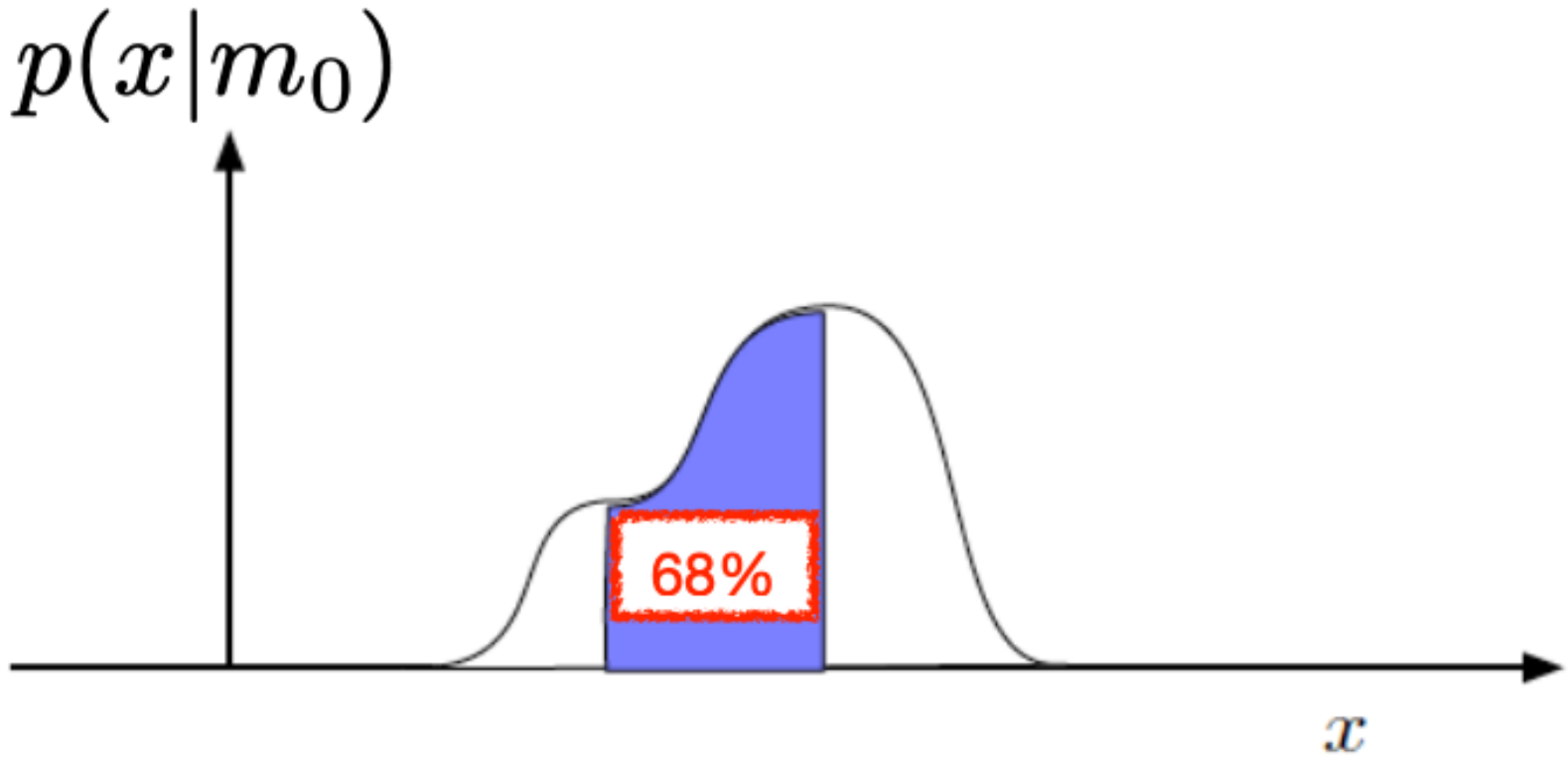
NB: для этого необходимо большое количество событий. NB2: все выводы очень просто распространяются на многомерные модели.

Построение Неймана



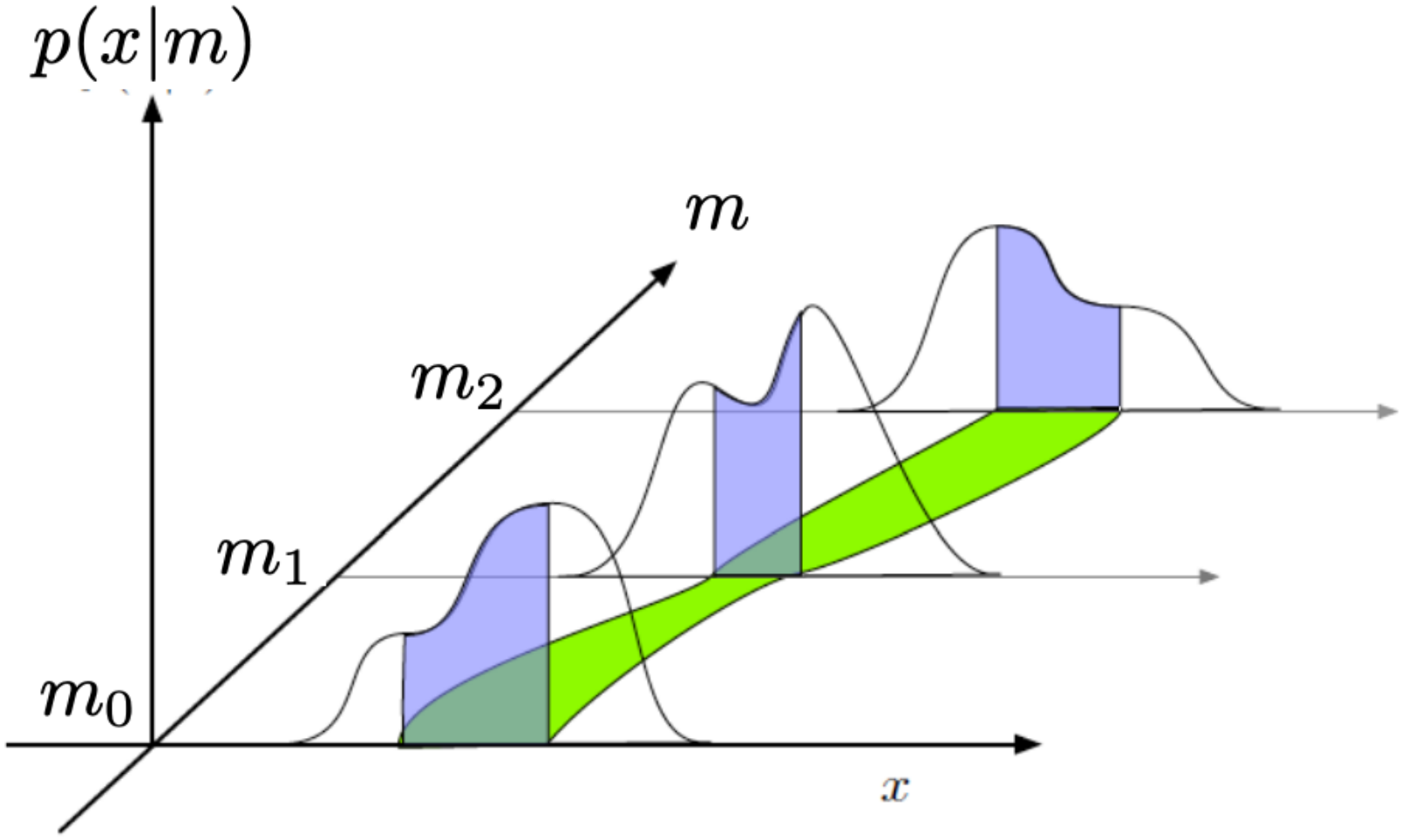
Для некоторого значения параметра m, построим PDF $p(x|m)$, где x — возможный исход эксперимента.

Построение Неймана



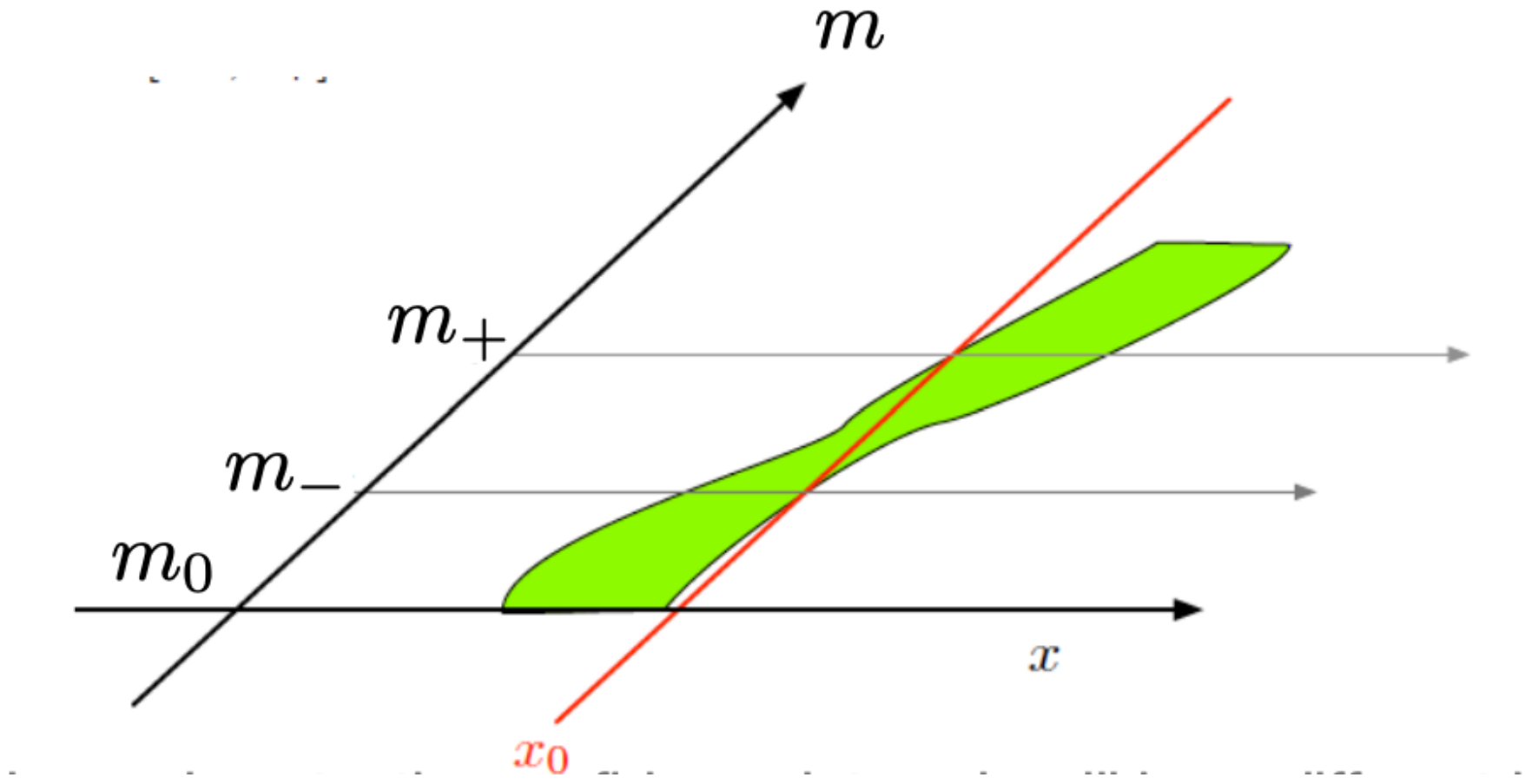
С помощью какой-то процедуры получим доверительный интервал на x с заданным покрытием (например, 68%).

Построение Неймана



Сделаем это для всех m, и построим "доверительный пояс".

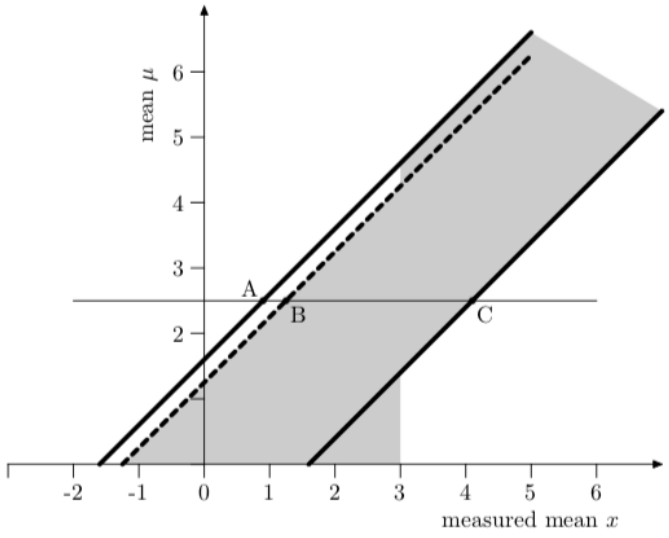
Построение Неймана



Провести эксперимент, получить x_0 Найти значения границ доверительного пояса для x_0.

В итоге m лежит в интервале $[m_-; m_+]$.

Построение Неймана: проблемы

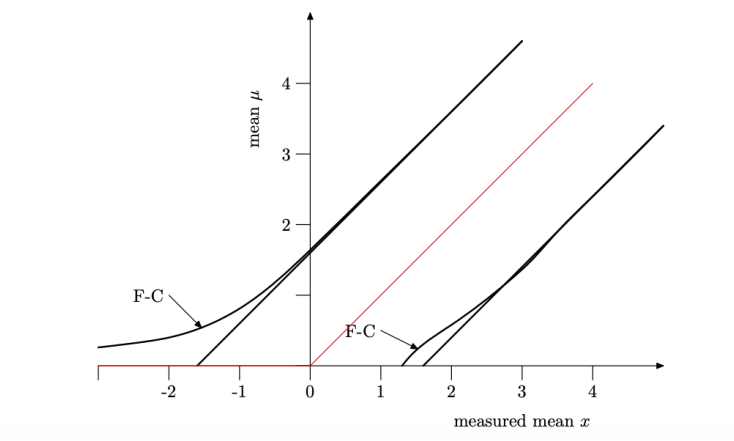


При таком построении появляются проблемы, связанные с поведением вблизи границ:

- пустые интервалы;
- ”флип-флоп” в районе перехода к отделяемому от физической границы пределу.

Эти проблемы решаются дополнительными построениями, например, Фельдман и Казинс предлагают дополнить запрещённые регионы, анализируя относительное правдоподобие.

Пару слов об универсальном подходе (Фельдман-Казинс)



- предложим ранкинг:

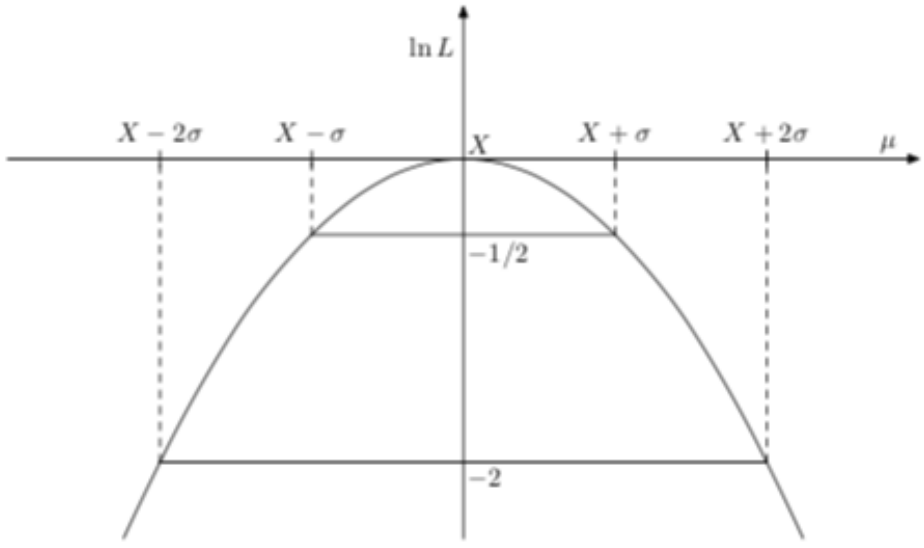
$$\frac{\mathcal{L}(x;m)}{\mathcal{L}(x;m_{best})}$$

;

- решает большинство проблем с краевыми эффектами;
- в многомерном случае есть сложности.

Доверительные интервалы на основе функции правдоподобия

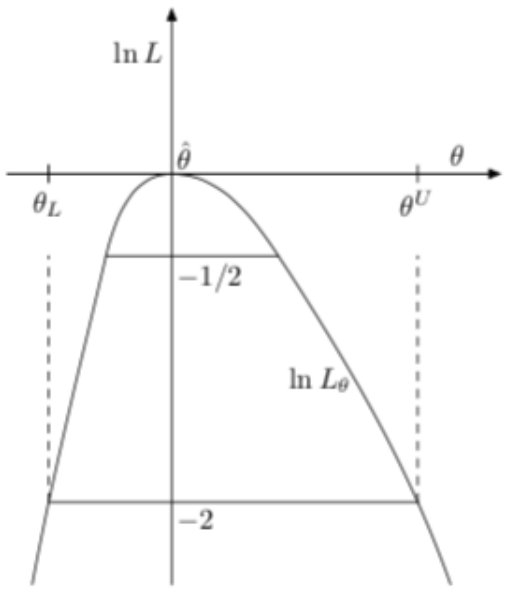
Мотивация



Log-likelihood function for Gaussian X, distributed $N(\mu, \sigma^2)$.

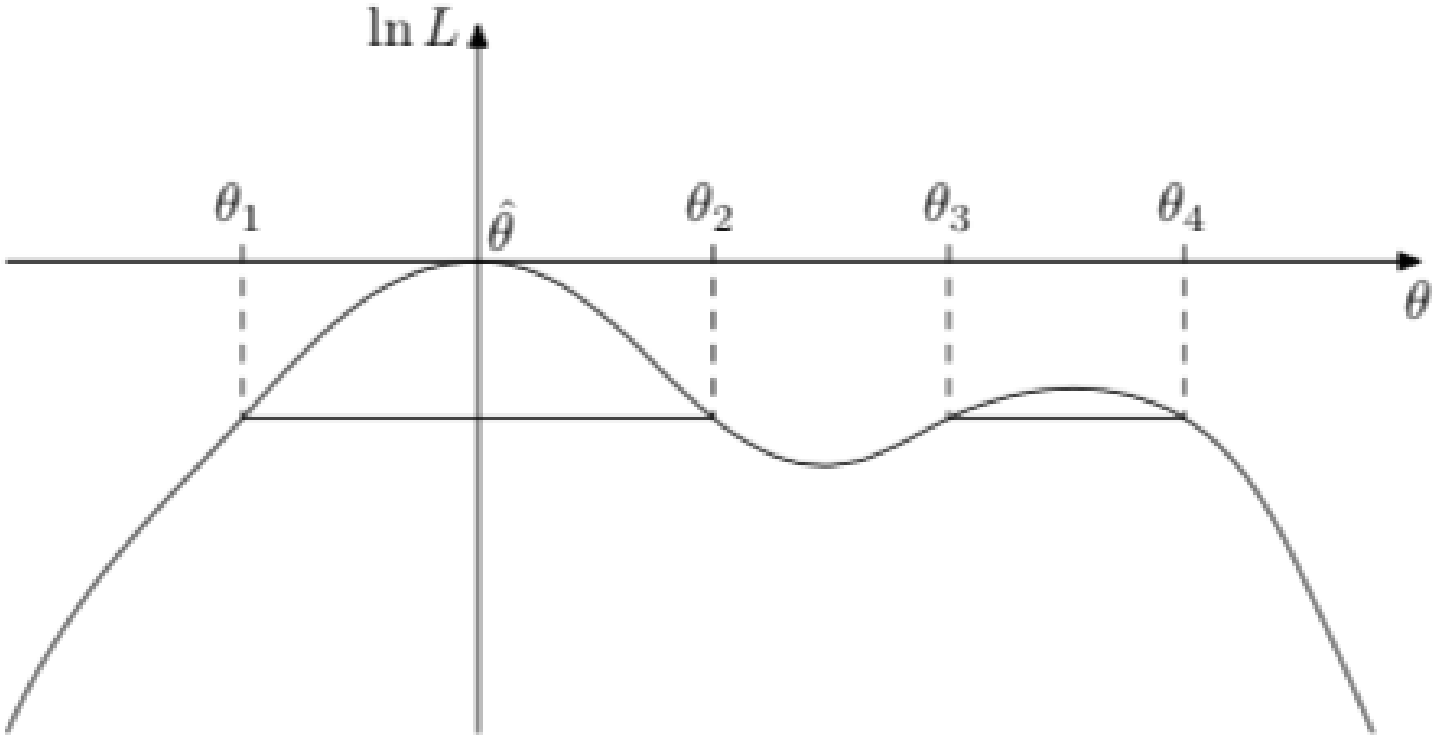
На предыдущих слайдах мы видели, что нормальная теория позволяет честно получать доверительные интервалы для величин, распределённых по Гауссу. Этот результат можно читать по-другому: если функция правдобия представляет собой параболу, то мы можем честно подсчитать доверительные интервалы.

Независимость правдоподбия от параметризации



В случае, если функция правдоподобия непараболическая, мы (почти) всегда можем привести её к параболическому виду некоторой трансформацией $g(\theta)$. При этом сама фукнция от параметризации независит, потому мы можем оценивать θ_L и θ_H через $\ln L = \ln L_{\max} - 1/2$ (для 68% интервала).

Сложные случаи



"Pathological" log-likelihood function.

В случае многомодальной функции правдоподобия при подобном построении есть шанс найти вторую моду.

Многомерные случаи

Самые большие проблемы начинаются в многомерном случае, так в классическом выводе мы должны гарантировать покрытие.

- Использовать нормальную теорию (если правдоподобие гаусово).
- Простой способ, использовать профильную функцию правдоподобия:

$$g(x_k) = \max_{x_i, i \neq k} \ln L(X).$$

  Этот способ даст возможность анализировать простые негаусовы правдоподобия, быстрый способ;
- Использовать объединённый метод, эквивалент бутстрапа, но без фиксированных мешающих параметров, медленный, но более надёжный способ.

Профильная функция правдоподобия

Предположим, что у нас есть параметры $\theta = (\psi, \lambda)$, для которых у нас известна функция правдоподобия $\mathcal{L}_n(\psi, \lambda)$.

Определение

Профильная функция правдоподобия определяется:

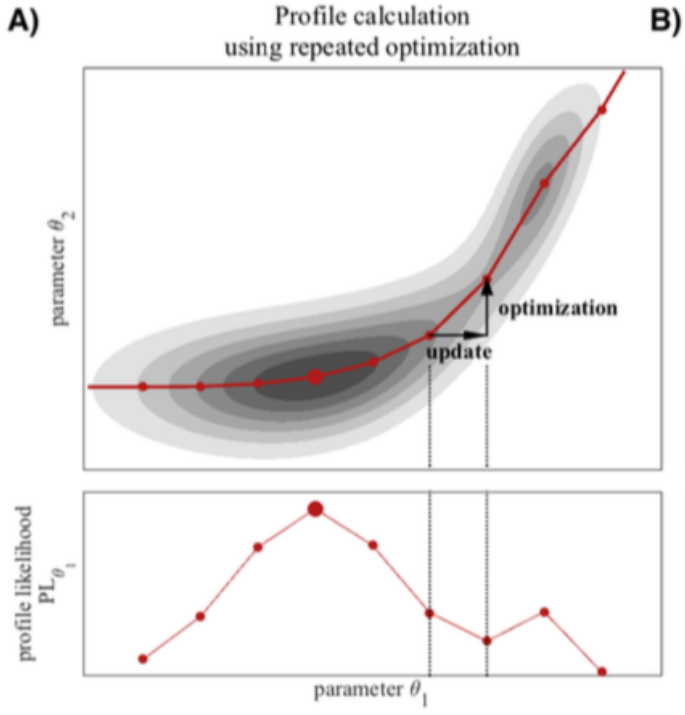
$$\mathcal{L}_n(\psi, \hat{\hat{\lambda}}) = \arg\max_{\lambda} \mathcal{L}_n(\psi, \lambda).$$

$\hat{\hat{\lambda}}$ - значение ОМП для фиксированного ψ.

Построение профильного правдоподобия

На практике, мы делаем профилирование:

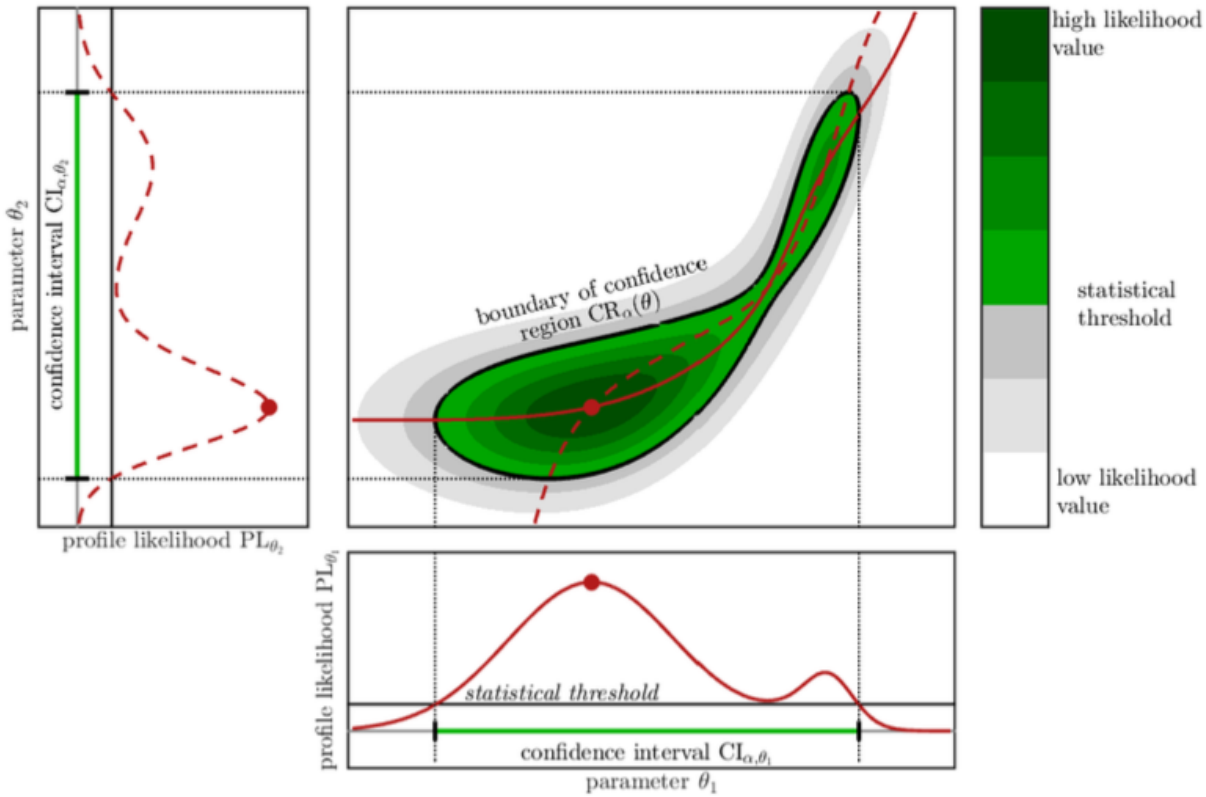
- фиксируем интересующий параметр;
- приближаем остальные;
- переходим к новому значению параметра.



R Boiger et al 2016 Inverse Problems 32 125009

Профильное правдоподобие

Пути профилирования могут отличаться для разных параметров



Точки начала доверительных интервалов для 1D и 2D не обязательно совпадают.

Теорема Вилкса

Теорема

Для достаточно хорошей функции правдоподобия $\mathcal{L}_n(\psi, \lambda)$, где $\psi \in \mathbb{R}^k$ и $\psi \in \mathbb{R}^l$ верно по распределению:

$$-2 \log \frac{\mathcal{L}_n(\psi, \hat{\hat{\lambda}})}{\mathcal{L}_n(\hat{\psi}, \hat{\lambda})} \to \chi^2_k$$

Это даёт возможность построить доверительный интервал (в k-мерном пространстве):

$$\beta = \mathbb{P}(-2 \log \frac{\mathcal{L}_n(\psi, \hat{\hat{\lambda}})}{\mathcal{L}_n(\hat{\psi}, \hat{\lambda})} \leq \chi^2_k)(\beta)$$

.

Резюме и наблюдения

- метод профильного правдоподобия даёт быстрый и хороший результат в случае, если многомерное правдоподобие не слишком плохое (например, в нём нет ”дырок”);
- свойства покрытия лучше соблюдаются для интервалов с малым доверительным значением;
- для сложных случаев необходимо использовать более продвинутые методы снижения размерности.

Систематические погрешности

В принципе, каждый источник систематической погрешности характеризуется своей случайной величиной (вернее, почти каждый). Предположим, что мы знаем плотность этой случайной величины:

- байесовский способ: без проблем, просто маргинализируем правдоподобие;
- классический способ: задача становится очень многомерной, надо делать профилирование;
- смешанный способ: давайте сделаем вид, что мы байесовцы, маргинализируем, а потом используем как классический вывод.
- смешанный способ 2: возьмём профильный лайклихуд и анализируем байесовским способом, опасно.

Какой способ препочесть?

- Универсального способа не существует.
- В случае достаточно большой статитистики, нам всё равно.
- Для малой статистики или слишком близкой границы каждый метод имеет недостатки, которые надо учитывать.

Дельта-метод

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений $\text{Bernoulli}(p)$ и $\text{Bernoulli}(r)$?

Мотивирующий пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений $\text{Bernoulli}(p)$ и $\text{Bernoulli}(r)$?
- $\frac{p}{1-p} / \frac{r}{1-r}$

Пример

Пусть есть выборка $X_1, ..., X_N \sim \text{Bernoulli}(p)$. Какие вопросы о параметрах мы обычно хотим задавать?

- Чему равна вероятность успеха?
- p
- Каковы шансы на успех?
- $\frac{p}{1-p}$
- Сравнить шансы на успех в случае двух распределений $\text{Bernoulli}(p)$ и $\text{Bernoulli}(r)$.
- $\frac{p}{1-p} / \frac{r}{1-r}$

Обычно мы используем $\hat{p} = \sum_i X_i / N$ для оценки p. Кажется, что для других величин мы можем использовать похожие оценки: $\frac{\hat{p}}{1-\hat{p}}$.

Как при этом оценить дисперсию?

Ряд Тейлора

Пусть $\mathbf{T} = (\mathbf{T_1}, .., \mathbf{T_k})$ — случайные величины со средними $\theta = (\theta_\mathbf{1}, ..., \theta_\mathbf{k})$. Пусть задана дифференцируемая функция $g(\mathbf{T})$ (оценка какого-то параметра). Найти дисперсию этой оценки.

Будем называть $g'_i(\theta) = \frac{\partial}{\partial \mathbf{t_i}} \mathbf{g(t)} \Big|_{\mathbf{t_1}=\theta_\mathbf{1};...;\mathbf{t_k}=\theta_\mathbf{k}}$.

Разложим $g(t)$ в ряд Тейлора:

$$g(t) \approx g(\theta) + \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{g'_i}(\theta)(\mathbf{t_i} - \theta_\mathbf{i})$$

Из этого следует:

$$E_\theta g(T) = g(\theta).$$

Ряд Тейлора

Аналогично дисперсия:

$$\mathsf{Var}_\theta g(T) \approx E_\theta \left([g(\mathbf{T}) - \mathbf{g}(\theta)]^{\mathbf{2}} \right) \approx E_\theta \left(\left(\sum_{i=1}^{k} g_i'(\theta)(T_i - \theta_i) \right)^2 \right) =$$

$$= \sum_{i=1}^{k} [g_i'(\theta)]^{\mathbf{2}} \mathsf{Var}_\theta \mathbf{T_i} + \mathbf{2} \sum_{\mathbf{i>j}} \mathbf{g_i'}(\theta)\mathbf{g_j'}(\theta)\mathsf{Cov}_\theta(\mathbf{T_i}, \mathbf{T_j}).$$

Замечание: здесь мы не использовали почти никакой информации о функции $g(T)$.

Вернёмся к мотивирующему примеру, нас интересовала дисперсия оценки $\frac{\hat{p}}{1-\hat{p}}$. Здесь $g(p) = \frac{p}{1-p}$ Используя предыдущие выкладки несложно подсчитать, что:

$$\mathsf{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx [g'(p)]^2 \mathsf{Var}(\hat{p}) = \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}.$$

Пример

X - случайная величина, с ненулевым матожиданием μ.

Необходимо оценить матожидание и дисперсию для оценки:

$g(\mu) = 1/\mu$.

Используя:

$$E_\mu\left(g(X)\right) \approx g(\mu),$$

$$\mathsf{Var}_\mu(g(X)) \approx [g'(\mu)]^2 \mathsf{Var}_\mu(X).$$

Получаем:

$$E_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu},$$

$$\mathsf{Var}_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu^4}\mathsf{Var}_\mu(X).$$

(Продолжение следует)

Теорема (Теорема Слуцкого)

Если $X_n \to X$ по распределению и $Y_n \to a$ по вероятности, причём $a = \text{const}$, тогда:

- $Y_n X_n \to aX$ по распределению,
- $X_n + Y_n \to X + a$ по распределению.

Теорема (Дельта-метод)

Пусть Y_n — последовательность случайных величин для которых $\sqrt{n}[Y_n - \theta] \to \mathcal{N}(0, \sigma^2)$ по распределению. Тогда для дифференцииуемой в θ функции $g(.)$ с ненулевой производной, $\sqrt{n}[g(Y_n) - g(\theta)] \to \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$ по распределению.

Доказательство: Ряд Тейлора для $g(Y_n)$ около $Y_n = \theta$:

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + o(Y_n - \theta)$$

Третье слагаемое стремится к 0 по вероятности. Тогда мы сможем применить теорему Слуцкого:

$$[g(Y_n) - g(\theta)] = g'(\theta)(Y_n - \theta),$$

мы получили согласно условиям необходимую сходимость.

NB: В случае нулевой первой производной и ненулевой второй, оценка сходится к $\sigma^2 \frac{g''(\theta)}{2} \chi_1^2$ по распределению.

Пример

Продолжим предыдущий пример. Пусть есть выборка со средним $\bar{X}$, тогда:

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \mathcal{N}\left(0, \left(\frac{1}{\mu}\right)^4 \mathsf{Var}_\mu X_1\right) \text{ по распределению}$$

Если мы не знаем матожидание и дисперсию, можно ввести их оценку:

$$\widehat{\mathsf{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2,$$

то есть

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \mathcal{N}(0,1)$$

Пример

Второй раз применив теорему Слуцкого, получим:

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}}-\frac{1}{\mu}\right)}{\sigma/\mu^2} \to \mathcal{N}(0,1)$$

.

Вспомним центральную предельную теорему!

Заключение

- При нахождении значения параметров важны не только точечные оценки, но и интервальные.
- Интервальные оценки могут быть построены байесовским и частотным методом, их смысл отличается и они имеют разные проблемы.
- Простейший способ построения интервальной оценки для обратной задачи - Дельта-метод.