Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Статистические вопросы

ПМИ ФКН ВШЭ, 15 декабря 2018 г.

Денис Деркач[1]

[1] ФКН ВШЭ

Оглавление

Вероятность

Что такое вероятность?

- Величина, характеризующая ”степень возможности” некоторого события, которое может как произойти, так и не произойти (Философия: Энциклопедический словарь.)
- В принципе, может определяться без математических терминов.

Аксиоматика Колмогорова

- Каждому событию A, принадлежащему $\mathcal{F}$, ставится в соответствие неотрицательное число $\mathbb{P}(A)$, которое называется вероятностью события .
- Вероятность достоверного события $\mathbb{P}(\mathcal{F}) = 1$.
- Вероятность невозможного события $\mathbb{P}(\emptyset) = 0$.
- Если события X_1 и X_2 не пересекаются, то $\mathbb{P}(X_1 + X_2) = \mathbb{P}(X_1) + \mathbb{P}(X_2)$ (также и со счётным количеством событий).

Интерпретации вероятности

Априорные и апостериорные суждения

- Предположим, мы хотим узнать значение некоторой неизвестной величины.
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания.
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$.

Два подхода к вероятности

В прикладной статистической науке наиболее популярны два типа интерепретаций вероятностных процессов:

- классический (или частотный, frequentist) подход считает, что вероятность события X определяется:

$$P(X) = \lim_{N \to \infty} \frac{n}{N},$$

  где N - количество тестирований, n - количество выпадений X.

- байесовский подход считает $P(X)$ степенью уверенности, что X — верная гипотеза.

Оба подхода удовлетворяют всем требованиям на вероятность.
NB: существует также информационный подход (AIC) и прочие.

Частотный подход

- В частотном подходе предполагается, что случайность есть объективная неопределенность.
- При интерпретации случайности как объективной неопределеннсти единственным возможным средством анализа является проведение серии испытаний ($n \to \infty$!). При этом, мы не знаем, когда n является достаточно большой.

NB: Обычно мы говорим о вероятностях неповторяемых событий (P(дождь|завтра)).

Байесовский подход

- В байесовском подходе предполагается, что случайность есть мера нашего незнания.
- Все величины и параметры считаются случайными. Точное значение параметров распределения нам неизвестно, значит, они случайны с точки зрения нашего незнания.
- В качестве оценок неизвестных параметров выступают апостериорные распределения.

NB: Построение априорного знания субъективно.

Bayesian vs. Frequnetist

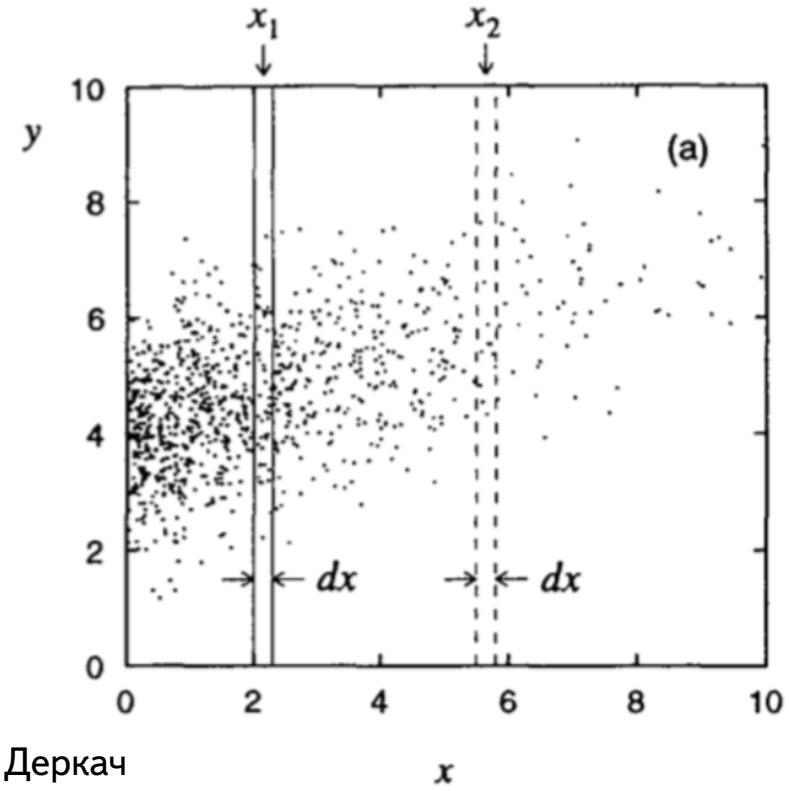
”Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone.” — Louis Lyons

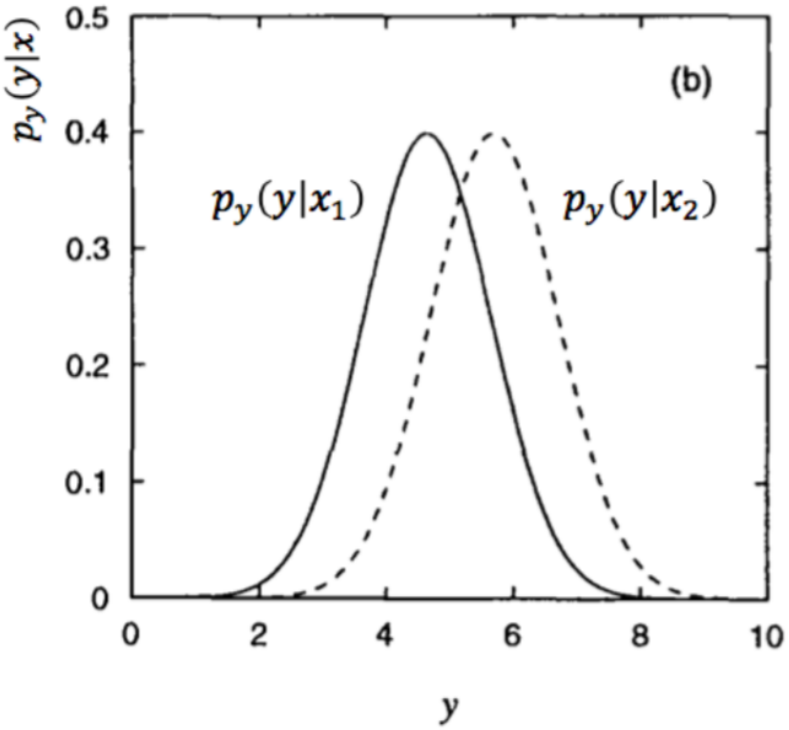
Как это влияет на результат?

- Каждая наука рассматривает свой лидирующий подход.
- Методы, которые мы описывали, в основном разработаны в классическом подходе.
- В случае достаточно большой выборки разницы (почти) нет :-)

Как это влияет на результат?

- В частотном подходе предполагается, что случайность есть объективная неопределенность.
- В байесовском подходе предполагается, что случайность есть мера нашего незнания.





Пример на броски монетки

Пример

Мы бросили монетку 14 раз, 10 раз выпал орёл. Какие шансы на то, что два следующих броска выпадет орёл?

Фрекентистский подход:

Оценим вероятность успеха: $\widehat{p}_{14} = 10/14 \approx 0.71$. Вероятность двух успехов: $\widehat{p}^2 \approx 0.51$.

Байесовский подход:

Перепишем теорему Байеса:

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)}$$

Пример на броски монетки: байесовское решение

Найдём правую часть:

$$\mathbb{P}(data|p) = \binom{14}{10} p^{10}(1-p)^4,$$

Вероятность данных не зависит от p:

$$\mathbb{P}(data) = \text{const},$$

Мы ничего не знаем о p:

$$\mathbb{P}(p) \sim \text{Uniform}(0,1) \equiv Beta(p,1,1).$$

Тогда

$$\mathbb{P}(p|data) = \frac{\mathbb{P}(data|p)\mathbb{P}(p)}{\mathbb{P}(data)} \sim p^{10}(1-p)^4.$$

Пример на броски монетки: байесовское решение

Найдём ответ:

$$\mathbb{P}(HH|data) = \int\limits_0^1 \mathbb{P}(HH|p)\mathbb{P}(p|data)dp = \text{const} \int\limits_0^1 p^2 p^{10}(1-p)^4 dp.$$

Точный подсчёт даст $\mathbb{P}(HH|data) \approx 49\%$.

Отличается от 51% в классическом подходе! Какой правильный?

https://bit.ly/1m54WgZ

Байесовское точечное оценивание

Фрекентистские подходы

- Оценка метода моментов (Method of moments, MOM):

$$\widehat{\alpha}_n = \alpha_n(\widehat{\theta}).$$

- Оценка максимально правдоподобия (ОМП, Maximum Likelihood Estimate, MLE):

$$\widehat{\theta} : \mathcal{L}_n(\theta) \to \max .$$

Оценка апостериорного максимума

Оценка апостериорного максимума, MAP

Формально, ОМП определяет значения параметров, при которых наши данные наиболее вероятны:

$$f(X;\theta) \sim f(X|\theta).$$

На самом деле, мы обычно задаёмся вопросом, какие значения параметров наиболее вероятны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)},$$

где f, g и h — соответствующие функции распределения.

Оценка MAP

Определение

Оценка апостериорного максимума (MAP) определяется как такое значение $\widehat{\theta}_n$ параметра θ, которое максимизирует $f(\theta|X)$.

Связь с MLE

MAP и MLE очевидно связаны:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)} = \frac{\prod_{i=1}^{n} f(X_i;\theta)g(\theta)}{h(X)} \sim \text{const} \prod_{i=1}^{n} f(X_i;\theta)g(\theta)$$

Логарифмируем:

$$\log f(\theta|X) = \log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Получается, что значение MAP оценки и значение оценки MLE совпадают с точностью до априорной оценки $\log(g(\theta))$.

Сопряжённые априорные оценки

Какую $g(\theta)$ выбрать?

- любую;
- но лучше выбирать сопряжённое априорное распределение, для которого функциональная форма совпадает с апостериорным.

Значения параметров сопряжённых распределений имеют смысл предыдущих измерений.

Список из Википедии.

Комментарий о MAP

- позволяет учесть предыдущие знания;
- выдаёт точечную оценку (не совсем байесовский);
- зависит от параметризации;
- при относительно больших n совпадает с MLE (а также в случае $g(X) = \text{const}$!).

О переходе к байесовскому представлению

Для перехода к Байесовскому методу нам надо оценить знаменатель выражения:

$$f(\theta|X) = \frac{f(X|\theta)g(\theta)}{h(X)}$$

Так как $h(X)$ — распределение данных при любых значениях параметров, можем записать:

$$h(X) = \int_{\Theta} f(X|\theta)g(\theta)d\theta.$$

И оценивать интервалы.

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).

Для этого мы вводим интервальные оценки.

Требования к интервальным оценкам

- как можно более объективно рассказать о результатах эксперимента;
- предъявить интервал, который покрывает настоящее значение параметра, с выбранной вероятностью;
- предоставить информацию, необходимую для принятия решения;
- сделать вывод о параметре, который включает в себя предыдущие знания.

NB: в случае большой выборки скорее всего достаточно будет использовать точечную оценку параметра и его стандартное отклонение.

Bayes vs. Frequentist

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: сокращение Cr.L. (часто используют C.L.).

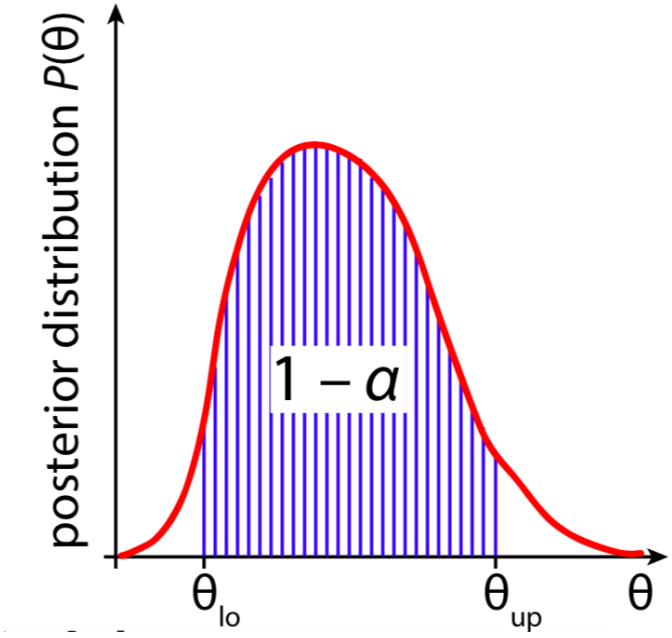
Байесовский доверительный интервал

Байесовский подход:

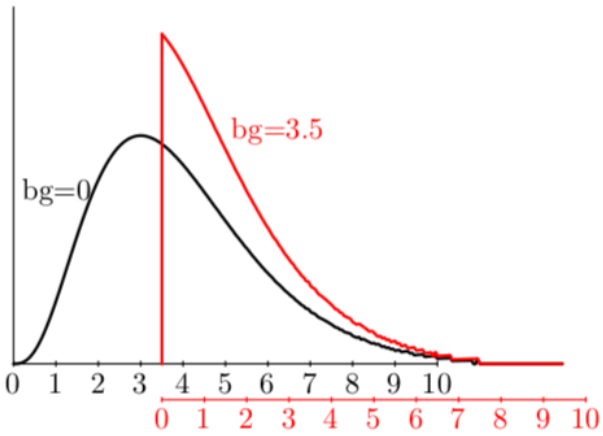
$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|X)d\theta$$

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.



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.

Мешающие параметры (Nuisance parameters)

Определение

Мешающий параметр — 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 принетгрируем по нему).:

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

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 вводились так, чтобы быть инвариантны 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 Байесовской вероятности мы оцениваем

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

Покрытие (coverage)

Метод, который позволяет построить интервал $(\theta_a; \theta_b)$ такой, что $\mathbb{P}(\theta_a \leq \theta_0 \leq \theta_b) = \beta$, где θ_0 - настоящее значение параметра, обладает свойством покрытия.

Частотные интервалы будут флуктуировать вместе с новыми выборками. Потому, покрытие определяют как доля интервалов, которая содержит настоящее значение θ_0.

NB: наличие покрытия у байесовкого подхода под вопросом.

Покрытие частотных интервалов

На практике, в основном, используются методы, обладающие ассимптотическим покрытием. Если покрытие $\mathbb{P} \leq \beta$, говорят о "недопокрытии" (undercoverage), если $\mathbb{P} \geq \beta$ (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

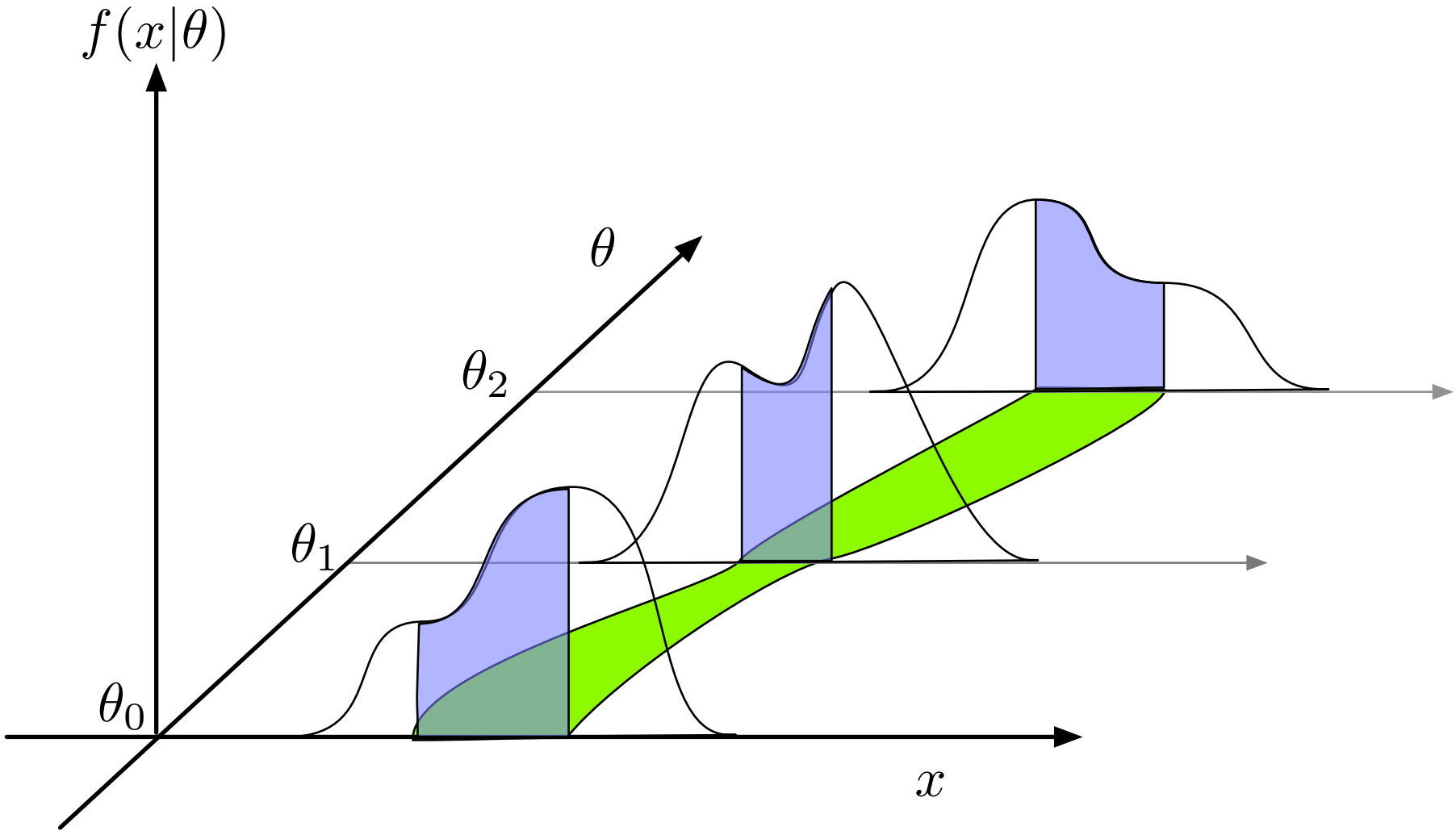
необходимо большое количество событий. NB2: все выводы очень просто распространяются на многомерные модели.

Построение Неймана

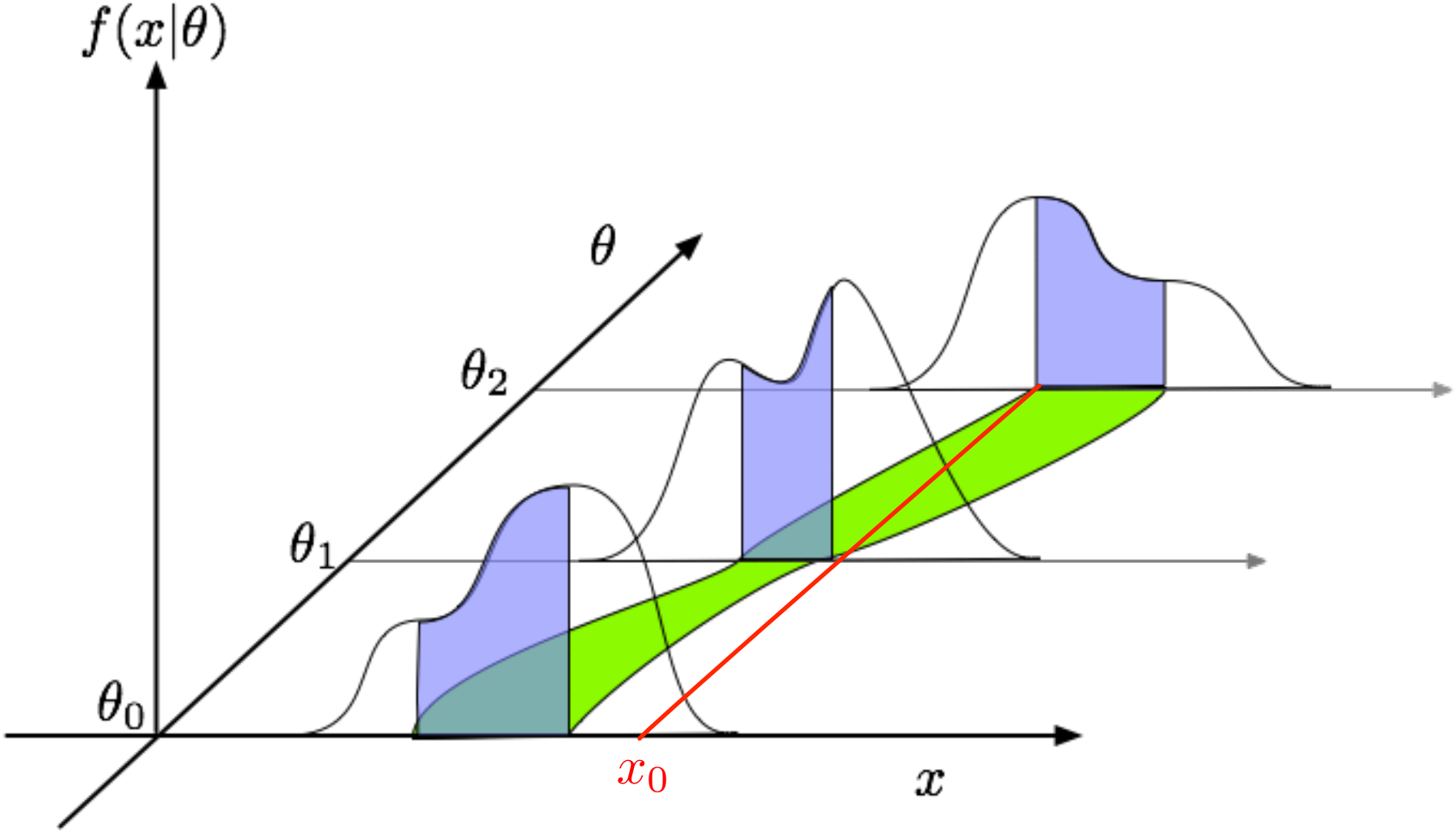
The Neyman construction for constructing frequentist confidence intervals involves the following steps:

- Given a true value of the parameter θ, determine a p.d.f. $f(x;\theta)$ for the outcome of the experiment. Often x is an estimator for the θ.
- Using some procedure, define an interval in x that has a specified probability (say, 90%) of occurring
- Do this for all possible true values of θ, and build a confidence belt of these intervals
- Compute the confidence belt given the value of x.

Построение Неймана

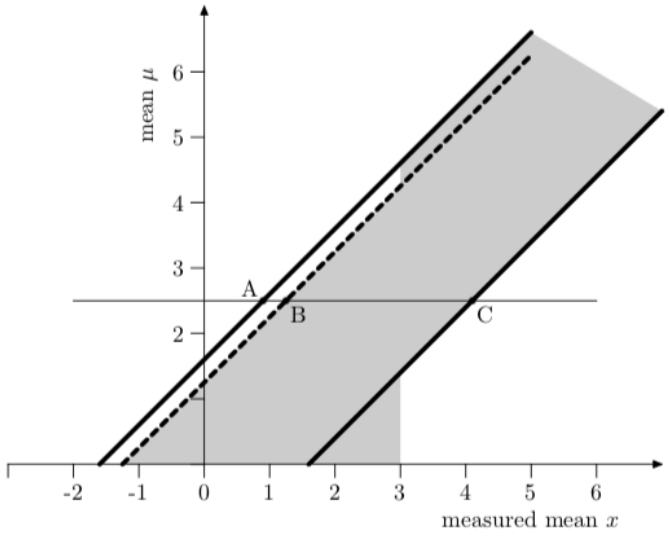


Построение Неймана



При таком построении покрытие получается всегда 100%.

Построение Неймана: проблемы



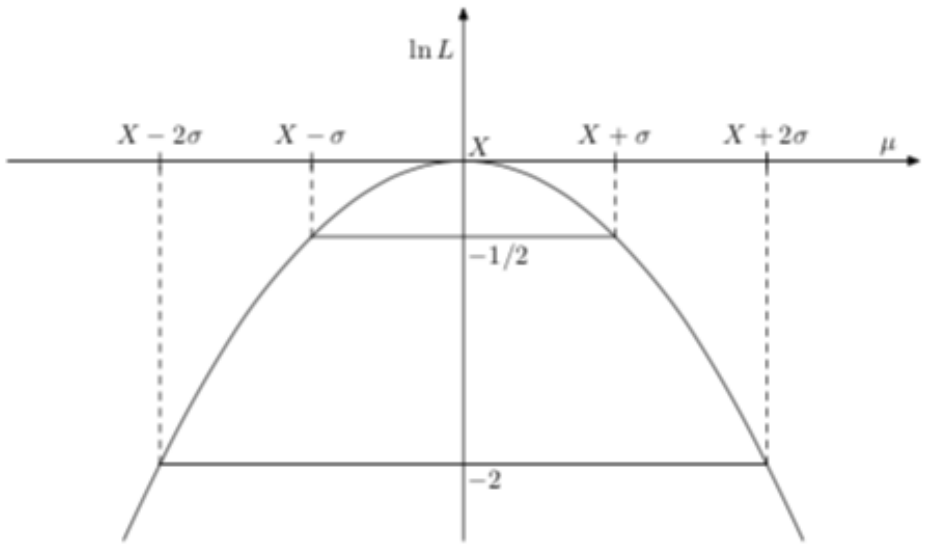
При таком построении появляются проблемы, связанные с поведением вблизи границ:

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

Эти проблемы решаются дополнительными построениями, например, унифицированный подход предлагает дополнить запрещённые регионы, анализируя относительное правдоподобие.

Доверительные интервалы на основе функции правдоподобия

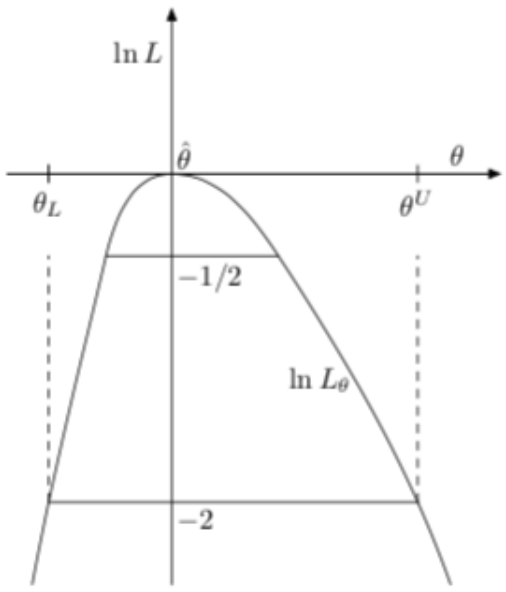
Мотивация



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

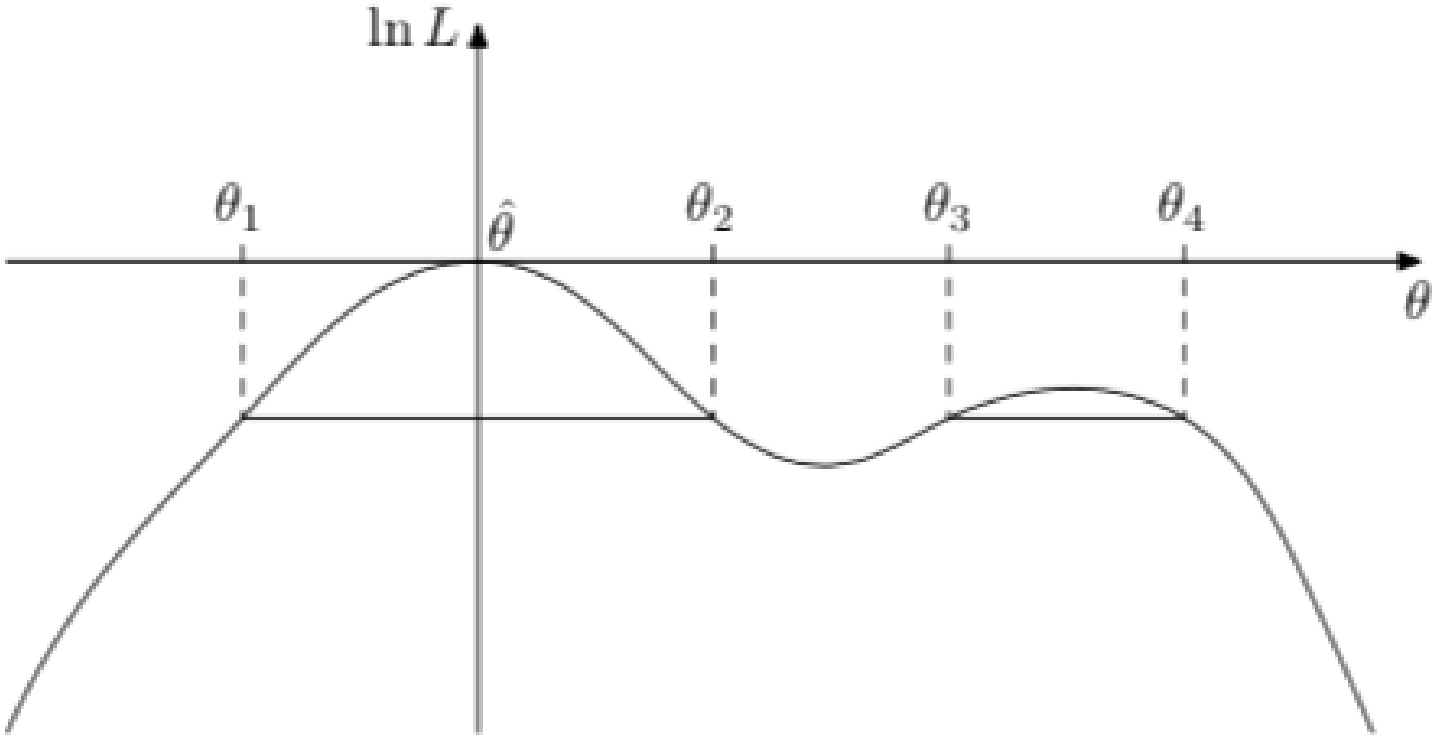
На предыдущих слайдах мы видели, что нормальная теория позволяет честно получать доверительные интервалы для величин, распределённых по Гауссу. Этот результат можно читать по-другому: если правдобие представляет собой параболу, то мы можем честно подсчитать доверительные интервалы.

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.

- 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.
- Использовать объединённый метод, эквивалент бутсрепа, но без фиксированных мешающих параметров.

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;
- смешанный способ: давайте сделаем вид, что мы байесовцы, маргинализируем, а потом используем как классический вывод.

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.

Критерий Неймана-Пирсона

Критерий Неймана-Пирсона для случая двух простых гипотез

Лемма (Неймана-Пирсона)

$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1$

Статистика Неймана-Пирсона:

$$T = \frac{\mathfrak{L}(\theta_1)}{\mathfrak{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i; \theta_1)}{\prod_{i=1}^{n} f(x_i; \theta_0)}. \tag{1}$$

Допустим, что H_0 отвергается при $T > k$. Выберем k так, что $\mathbb{P}_{\theta_0}(T > k) = \alpha$.

Тогда, критерий Неймана-Пирсона (на основе статистики (1)) будет иметь наибольшую мощность $W(\theta_1)$ среди всех критериев размера α.

Проблемы статистики Неймана-Пирсона

- статистика должна быть полностью известна для любых x;
- мы можем оценивать только простые гипотезы.

В реальной жизни, обычно мы оцениваем многомерные с известными или неизвестными мешающими параметрами (сложные гиптезы). В общем случае, мы не можем построить равномерно наиболее мощный критерий, но для экспоненциального семейства распределений мы можем построить достаточные статистики.

Приблизительное решение

Вообще, в случае отстутвия возможности строить глобальный критерий, мы можем построить локальный:

- $H_0 : \theta = \theta_0$;
- $H_1 : \theta = \theta_1 = \theta_0 + \Delta$.

В этом случае, мы можем разложить лог-правдоподобие в ряд Тейлора:

$$\ln L(X; \theta_1) = \ln L(X; \theta_0) + \Delta \frac{\partial L(X; \theta)}{\partial \theta}|_{\theta=\theta_0} + \dots.$$

Но по лемме Неймана-Пирсона:

$$\ln L(X; \theta_1) - \ln L(X; \theta_0) \leq c_\alpha$$

Таким образом, мы построили локальный критерий, который зависит только от производной правдоподобия: $\frac{\partial L(X;\theta)}{\partial \theta}|_{\theta=\theta_0} > k_\alpha$

Локальные тесты

Можно вспомнить, что:

$$\mathbb{E}\left[\frac{\partial L(X;\theta)}{\partial \theta}|_{\theta=\theta_0}\right] = 0;$$

$$\mathbb{E}\left[\frac{\partial L(X;\theta)}{\partial \theta}\right] = +NI,$$

где N - количество событий, а I - информационная матрица Фишера.

Вспомним об ассимптотической нормальности правдоподобия, можем переписать локальный критерий так:

$$\frac{\partial L(X;\theta)}{\partial \theta}|_{\theta=\theta_0} \geq \lambda_\alpha \sqrt{NI},$$

Это будет самым мощным локальным критерием.

Байесовский подход

Нам необходимо найти:

$$\mathbb{P}(hyp|data) = \frac{\mathbb{P}(data|hyp)\mathbb{P}(hyp)}{\mathbb{P}(data)}$$

Нормализацию можно найти, интегрируя по всем возможным значениям параметров, что довольно сложно для некоторых типов гипотез. Мы можем изучать **байевсовский фактор**:

$$R = \frac{\mathbb{P}(H_0|data)\mathbb{P}(H_1)}{\mathbb{P}(H_1|data)\mathbb{P}(H_9)}$$

Полученное соотношение можно рассматривать как шансы на успех при ставке H_0 против H_1. При этом соотношение всё равно будет зависеть от априорного знания.

Парадокс Линдли (Lindley)

Тестирование точечной нулевой гипотезы против неточечной альтернативы. Например, броски монетки:

- $H_0 : p = 0.5$.
- $H_1 : p! = 0.5$.

В эксперименте Jahn, Dunne and Nelson (1987) говорится, что на 104490000 попытки, было получено 52263471 орлов и 52226529 решек. Что это значит с точки зрения статистик?

Парадокс Линдли

- Фрекентиская статистика:

$$z(x) = \sqrt{\frac{N}{\theta_0(1-\theta_0}}\left(\frac{1}{N}\sum x_i - \theta_0\right),$$

  то есть p-value: $p \ll 0.01$, нулевая гипотеза отвергается.
- Байесовский фактор:

$$R = \frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \approx 19.$$

  Нулевая гипотеза должна быть принята!

Решение: методы дают разные ответы

- фрекентистский подход говорит, что нулевая гипотеза плохо объясняет данные;
- байесовский подход говорит, что нулевая гипотеза описывает данные лучше, чем все альтернативные.