Статистические методы и анализ данных (2)

Игорь Бойко

Три интерпретации вероятности

- Комбинаторная интерпретация
  - Исходы эксперимента равновероятны в силу симметрии (например, бросание монеты или игральной кости)
- Частотная (frequentist) интерпретация
  - Отношение числа успехов к числу попыток, в пределе бесконечного числа попыток
- Что если повторить эксперимент принципиально невозможно?
  - Например, какова вероятность, что завтра пойдёт дождь?

Байесовская (Bayesian) вероятность

- Степень уверенности кого-либо (человека, экспертной комиссии, человечества) в том, что событие случится
- Численно задаётся через (мысленное) пари
  - Готовы ли вы поставить 1000 рублей против 100, что завтра пойдёт дождь? А если 100 против 1000?
- Базируется на всей сумме знаний, которой обладает тот, кто оценивает вероятность.
  - Поэтому байесовская вероятность различна для каждого человека

Примеры применения

- Допустим, мы поставили эксперимент на коллайдере, и обнаружили 10 событий, похожих на суперсимметрию.
  - Какова вероятность, что суперсимметрия существует?
- Более драматичный пример. Мы провели измерение, и обнаружили нарушение закона сохранения энергии.
  - Стоит ли публиковать статью о создании вечного двигателя? С риском опозориться?

Теорема Байеса

- Для независимых событий
- P(A&B) = P(A|B)P(B) = P(B|A)P(A)
- P(A|B) – условная вероятность: вероятность случиться B, если уже случилось A
- Формулировка теоремы Байеса:
- P(A|B) = P(B|A)P(A)/P(B)

P(A|B) = P(B|A)P(A)/P(B)

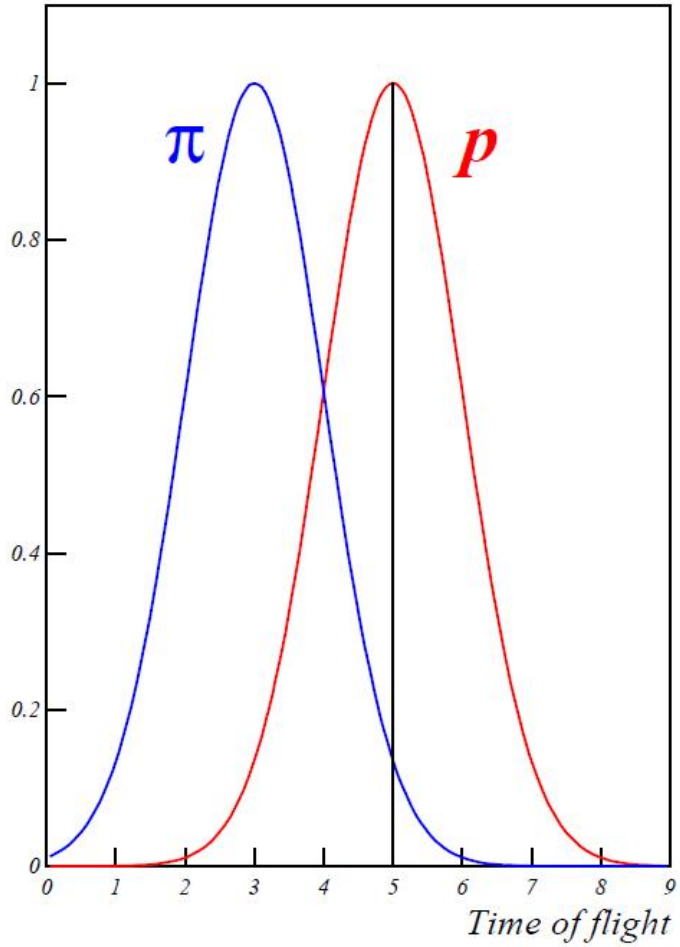
- Пусть A – некая теория, B – эксперимент по её проверке.
  - P(A) – вероятность (наша уверенность) до проведения эксперимента, что теория A верна.
  - P(B|A) – вероятность наблюдённого исхода эксперимента при условии, что теория A верна.
  - P(A|B) – наша уверенность в теории A после проведения эксперимента.
- После проведения эксперимента вероятность правильности теории пропорциональна этой же вероятности до эксперимента (априорная вероятность) и вероятности наблюдённого исхода при условии что теория верна

Итак…

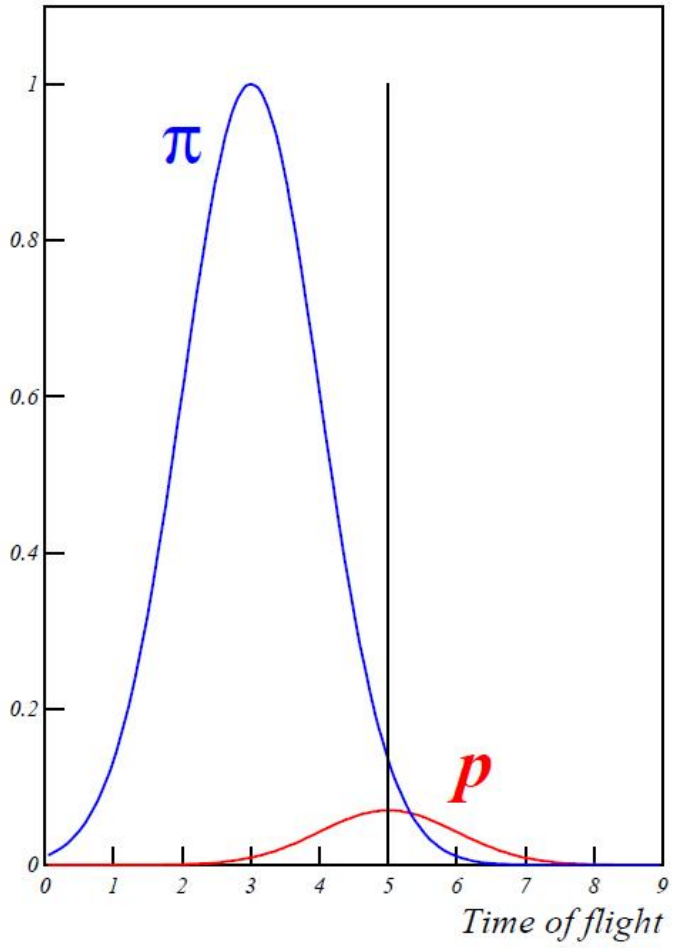
- …если вы открыли, что закон сохранения энергии нарушается – проверьте результат трижды!
- Даже если ваш эксперимент чрезвычайно точный и надёжный – велика вероятность экспериментальной ошибки, коль скоро опровергается теория, в которой мы [были] уверены.

Пример из практики

- Вы зарегистрировали частицу, и пытаетесь её идентифицировать: пион это или протон?
- Измеренное время пролёта идеально совпадает с ожидаемым для протона. Для пиона же такое время имеет плотность вероятности всего 10%.
- Казалось бы – это наверняка протон!



А что если пионов просто МНОГО?



- Следует учесть априорную вероятность (изначальные доли π и p) и результат эксперимента (измерение времени пролёта)
- Постериорная вероятность:

$$P_T(\pi) = \frac{N_\pi \cdot P_\pi(T)}{N_\pi \cdot P_\pi(T) \; + \; N_p \cdot P_p(T)}$$

Ещё раз

- Вероятность гипотезы после проведения эксперимента пропорциональна двум множителям:
  - её же вероятности до эксперимента
  - и вероятности данного исхода эксперимента, при условии если бы гипотеза была верна
- Обратите внимание: $P(A|B) \neq P(B|A)$
- $P(\text{беременна}|\text{женщина}) << P(\text{женщина}|\text{беременна}) \approx 1$

Оценка параметров

Напоминание

- Следует различать:
  - (истинные) параметры распределения $(\mu, \sigma, \ldots)$
  - (измеренные) характеристики выборки $(\overline{\mathrm{x}}, \mathrm{S}, \ldots)$
  - оценки параметров распределения по измеренным характеристикам выборки $(\hat{\mu}, \hat{\sigma}, \ldots)$

Измерение и оценка параметра

- Иногда результат измерения совпадает с оценкой параметра.
  - Например, за минуту проехало 10 машин. Наша оценка: поток машин составляет 10 в минуту.
- Чаще ситуация сложнее. Например, зарегистрировано 1000 событий. Но после вычета фона получаем: количество Хиггс-бозонов равно 100
- В общем случае задача ставится так: какова будет наша оценка параметра при условии, что измерение дало некоторый результат?

Метод максимального правдоподобия

(maximum likelihood method)

- Любое предположение о значении параметра можно назвать «гипотеза A». Из всех гипотез (значений параметра) мы хотим найти наиболее вероятную.
- Теорема Байеса: $P(A|B) = P(B|A)P(A)/P(B)$
- Допустим, априори все значения параметров равновероятны: $P(A)=const$.
- Тогда $P(A|B) \sim P(B|A)$, то есть вместо поиска наиболее вероятной гипотезы можно искать гипотезу, обеспечивающую максимальную вероятность измерить именно тот результат, который и был фактически измерен

Ещё раз:

- Проводим измерение, получаем результат x^*
- Для каждого из возможных значений параметра θ_i вычисляем вероятность p_i намерить именно x^*
- Нетривиальный шаг (теорема Байеса!): параметр, обеспечивающий наибольшую вероятность измерить x^* объявляем самым вероятным параметром

Конструирование правдоподобия

- Как правило, одновременно измеряются несколько величин x_i (набор замеров, массив событий, гистограмма со многими бинами)
- Если измерения независимы, то вероятность ансамбля измерений равна произведению вероятностей: $L(\theta) = \Pi\, p(x_i|\theta)$
- Удобнее использовать логарифм: $\ln L = \Sigma \ln(p(x_i|\theta))$
- (Многомерный) параметр θ находим покоординатно из уравнения правдоподобия: $\partial(\ln L)/\partial\theta_i = 0$

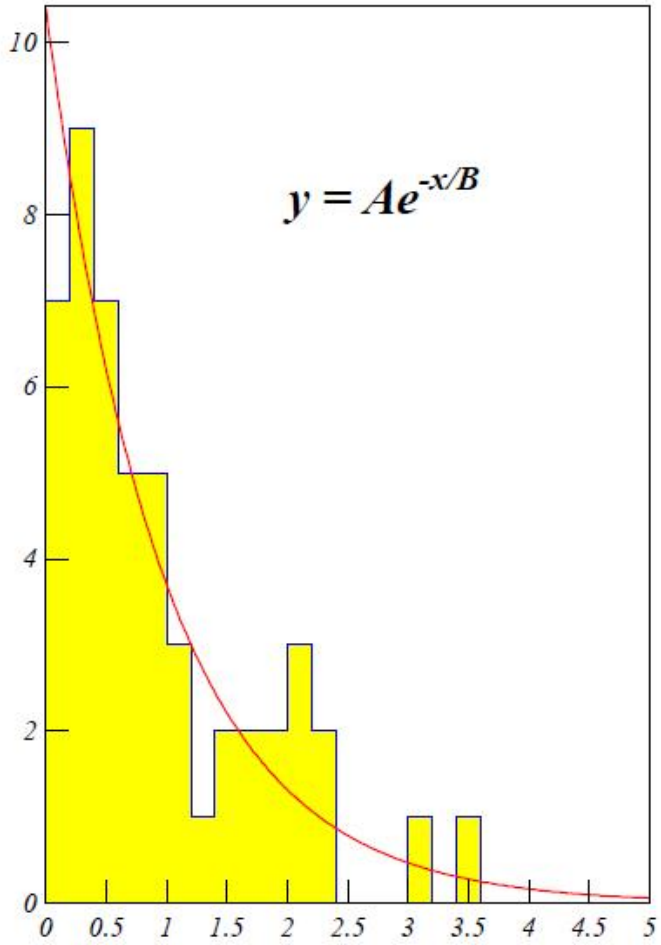
Фитирование

- Нахождение оптимального θ называют подгонкой (фитом) параметра к данным.
- «Фитирование данных» - неправильно!
  - Фитирование – это подгонка. Данные неизменны раз и навсегда (на то они и данные). Под данные подгоняется параметр, а не наоборот.
- Правильно: «фитирование параметра».
- Или: «подгонка К данным»
- Для фитирования используют «программу минимизации», которая умеет находить глобальный экстремум сложной функции от многомерного аргумента

Биновый и безбиновый фит

- В физике частиц обычно регистрируют множество событий (иногда миллионы)
- В каждом событии измеряют одну или несколько величин, которыми можно заполнить гистограмму, либо использовать весь массив значений во всей полноте
- Соответственно, фит называют binned fit или unbinned fit

Биновый фит

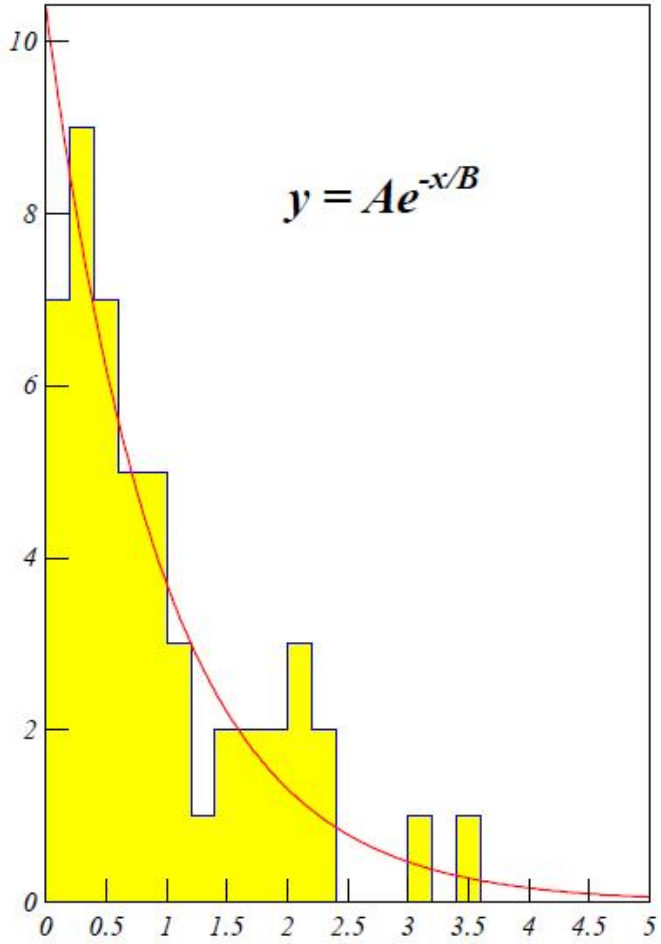


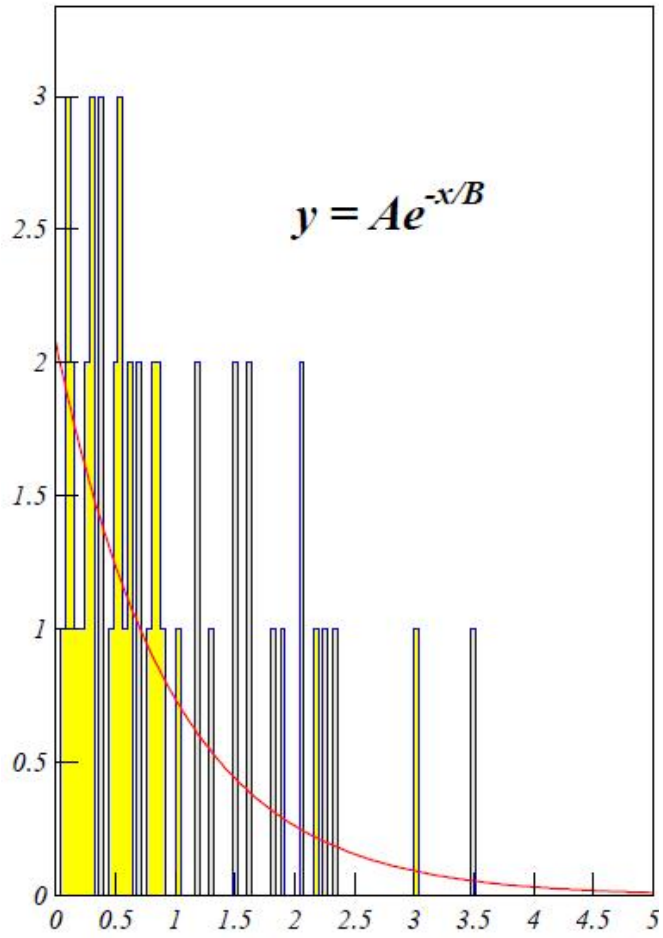
- Для каждого i-го бина вычисляется ожидаемое число событий $\mu_i(\theta)$.
- Фактическое число событий в бине n подчиняется распределению Пуассона: $P(n)=e^{-\mu}\ \mu^n/n!$
- Логарифм правдоподобия: $\ln L(\theta) = \Sigma n_i \ln \mu_i - \Sigma \mu_i - \Sigma \ln(n!)$

Недостатки бинового фита

- Загрубляется точность измерений: в пределах каждого бина все события идентичны, теряется информация о точном значении величины
- Паллиативное решение: уменьшить ширину бина!
- Всплывает ещё один недостаток: если измеряемая характеристика события x – многомерна, то число бинов (многомерной гистограммы) начинает расти в геометрической прогрессии при уменьшении бина
- Полное и окончательное решение: безбиновый фит!

Те же данные, но бин поменьше





Безбиновый фит

- Конструируем правдоподобие, пробегая по каждому событию – перемножаем (суммируем) плотность вероятности увидеть именно такое событие с измеренными характеристиками x_i:
- $\ln L = \Sigma\, p(x_i|\theta)$
- Обычно вероятность p – это произведение сечения рождения на вероятность (эффективность) регистрации события: $p(x_i|\theta) \sim \sigma(x_i|\theta)\, \varepsilon(x_i)$
- Внимание!! Плотность вероятности $p(x_i|\theta)$ – функция, нормированная на единицу! То есть, $\int p(x)dx = 1$
- Если $\int p(x)dx$ зависит от θ, то результат фита получится абсолютно неверный!

Важный нюанс: нормировка

- Итак, $p(x_i|\theta) \sim \sigma(x_i|\theta)\,\varepsilon(x_i)$
- Разумеется, полное сечение рождения ($\int\sigma$) зависит от параметров теории (θ). Поэтому нужно использовать нормированную плотность вероятности:
- $p(x_i|\theta) = \sigma(x_i|\theta)\varepsilon(x_i) \,/\, \int\sigma(x|\theta)\varepsilon(x)dx$

Важный нюанс: нормировка

- Итак, $p(x_i|\theta) \sim \sigma(x_i|\theta)\,\varepsilon(x_i)$
- Разумеется, полное сечение рождения ($\int\sigma$) зависит от параметров теории (θ). Поэтому нужно использовать нормированную плотность вероятности:
- $p(x_i|\theta) = \sigma(x_i|\theta)\varepsilon(x_i) \,/\, \int\sigma(x|\theta)\varepsilon(x)dx$

$$lnL(\theta) = \sum_{i=1}^{N} ln \frac{\sigma(x_i|\theta)\ \varepsilon(x_i)}{\int \sigma(x|\theta)\ \varepsilon(x)dx} =$$

$$= \sum ln\ \sigma(x_i|\theta)\ -\ N\ ln \int \sigma(x|\theta)\ \varepsilon(x)dx\ +\ \sum ln\ \varepsilon(x_i)$$

Недостатки безбинового фита

- Суммирование сечений пробегает через весь миллион событий (CPU!!) – и так для каждой итерации (каждого предположения о параметре θ)
- Необходимо находить (как правило, численно) интеграл $\int \sigma \varepsilon dx$ – хорошо хоть единожды на итерацию.
- Если измерение x одномерно или «маломерно», то компромиссное решение – сделать бины малой ширины

Включение в фит априорной информации

- Метод максимального правдоподобия основывался на предположении: «если все значения параметра равновероятны…». А если нет?
- Следует включить в фит (в функцию правдоподобия) априорную информацию!
- Например, до нас уже было кем-то измерено, что параметр θ равен $\theta_0 \pm \sigma$
- Значит, вместо $L(\theta) = \Pi\, p(x_i|\theta)$ используем $L' = L^* \exp(-(\theta-\theta_0)^2/2\sigma^2)$

Метод наименьших квадратов

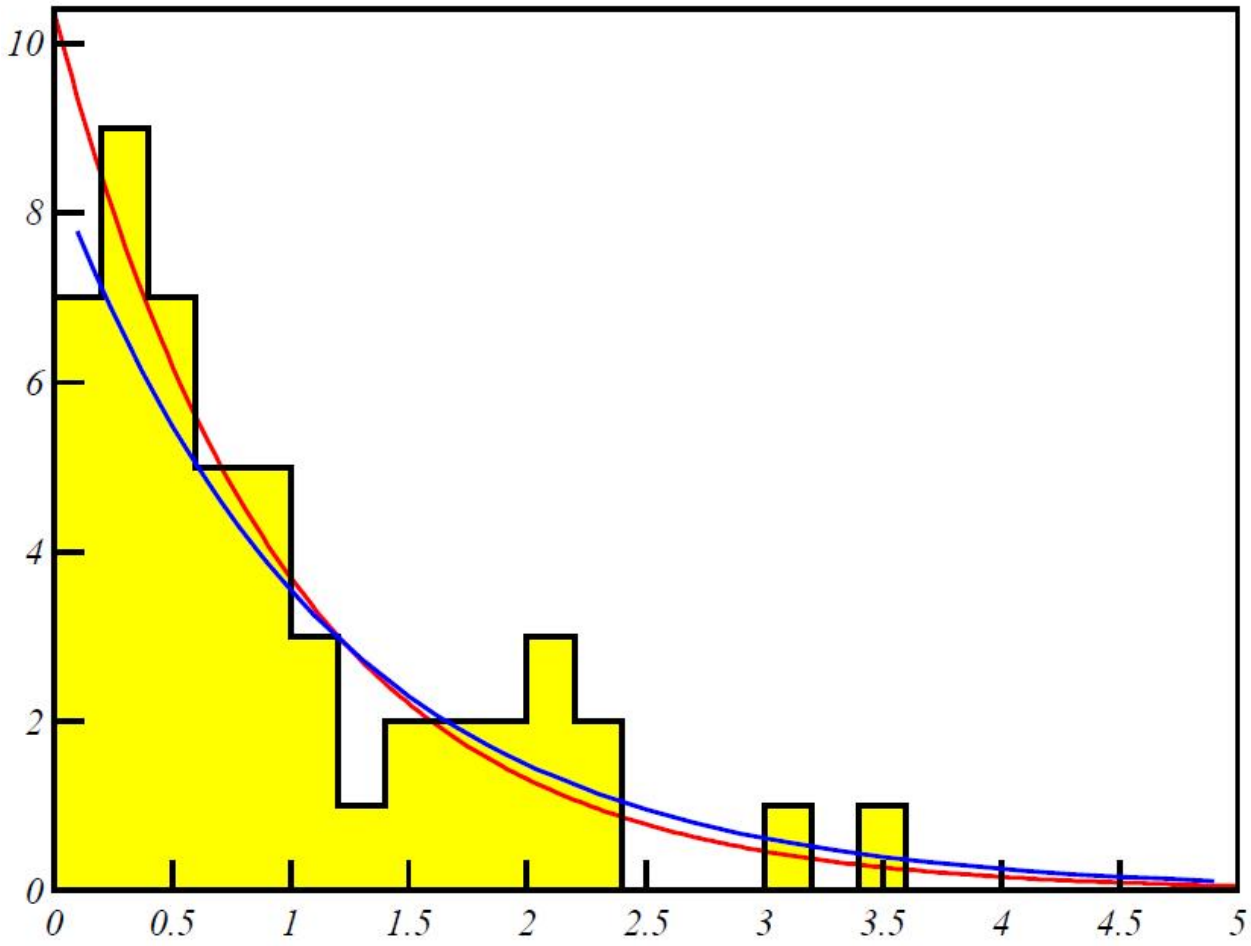
- Является частным случаем метода максимального правдоподобия
- Пусть погрешности измерений y_i независимы и каждое имеет нормальное распределение $N(0,\sigma_i^2)$ вокруг «предсказания теории» $f_i(\theta)$.
- Вычислим правдоподобие такого набора измерений:

$$-2\ lnL = -2\ ln\ \Pi\left(e^{-\frac{(y_i-f_i)^2}{2\sigma_i^2}}\right) = \sum \frac{(y_i - f_i(\theta))^2}{\sigma_i^2} = \chi^2(\theta)$$

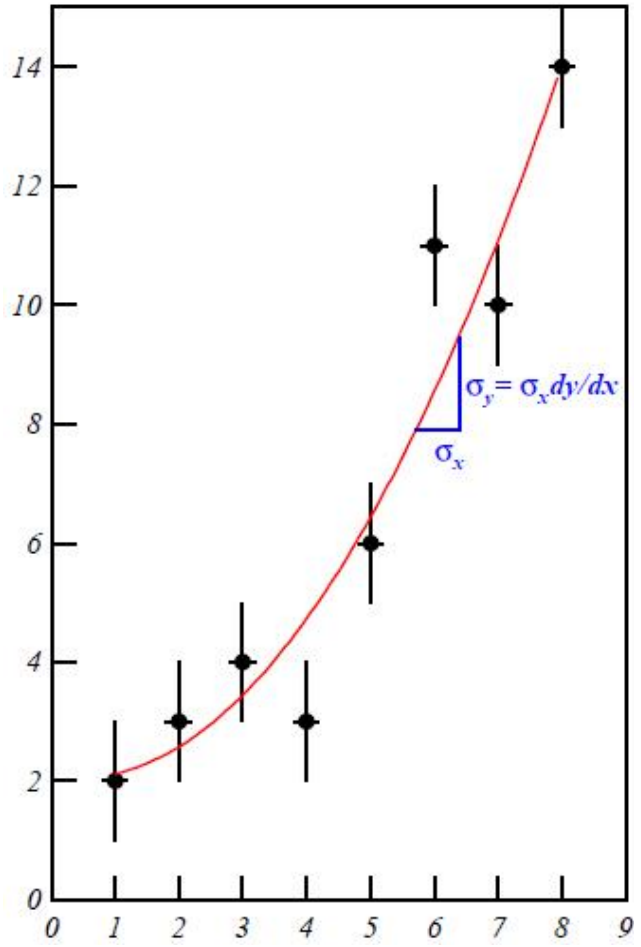
Замечание

- Метод χ^2 работает только для нормально распределённых ошибок
- Ни в коем случае не фитируйте гистограмму методом χ^2
  - При большом числе событий распределение Пуассона близко к нормальному, но всё же оно асимметрично, а значит метод будет систематически «предпочитать» отклонения вверх, а не вниз
  - При малом числе событий распределение Пуассона вовсе не похоже на нормальное
  - Пустые бины (ноль событий) приходится игнорировать – то есть систематически игнорируются «флуктуации вниз»
- Но соблазн такой остаётся, так как метод χ^2 (в отличие от метода максимального правдоподобия) предоставляет меру согласия фита с данными – об этом ниже

Всё та же гистограмма: фит правильный и неправильный



Перенос ошибок

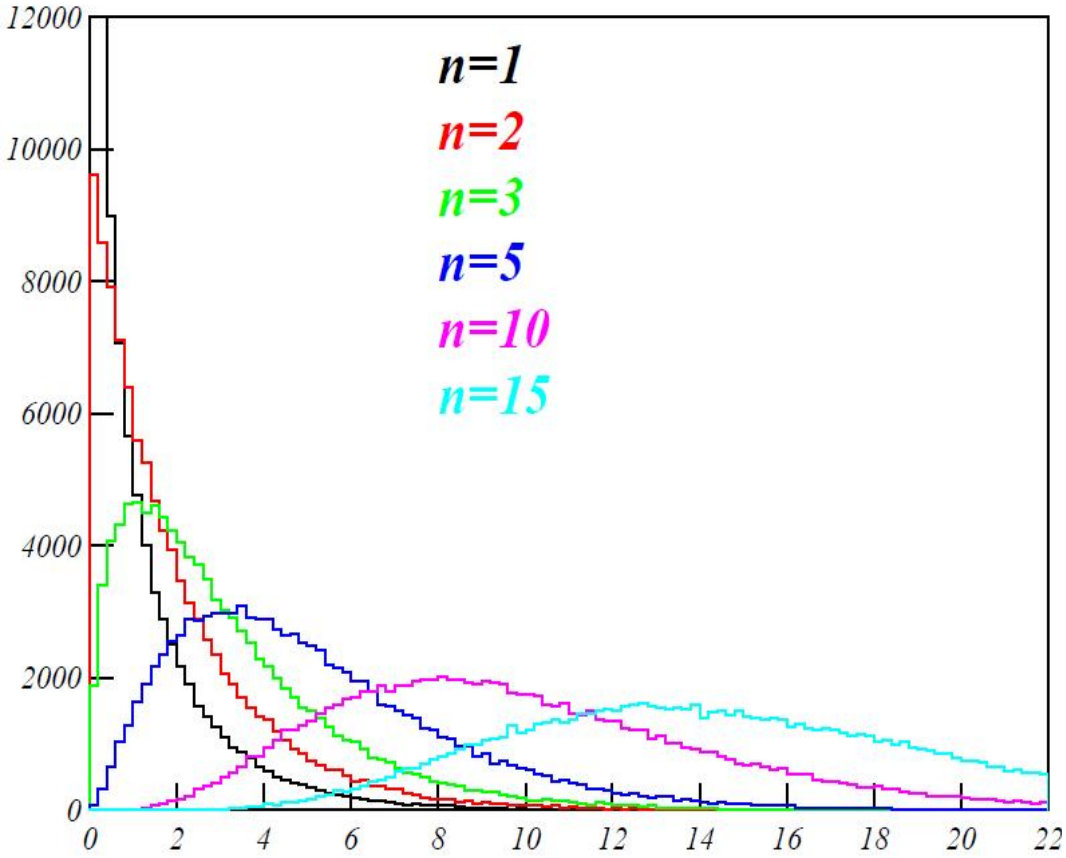


- Пусть измерения $y_i \pm \sigma_i^y$ сделаны «в точках» x_i, которые имеют свои ошибки σ_i^x (независимые от σ_i^y)
- Фитируется кривая $y=f(x)$
- Ошибки по x можно учесть, перенеся их в ошибку y:
- $(\sigma'_y)^2 = \sigma_y^2 + \sigma_x^2 (df/dx)^2$
- Способ применим, если 2-я производная мала: 1-я производная мало меняется в пределах σ_i^x

Качество фита

- Функционал $\chi^2(\theta)$. подчиняется распределению χ^2 с числом степеней свободы N-k
  - N – число измерений, k – число параметров.
  - Ровно N степеней свободы было бы, если бы отклонения отсчитывались от истинных значений, а не от «предсказаний», полученных из самого же фита
- Качество фита можно оценить по квантилям распределения χ^2_{N-k}

Распределение χ^2 (хи-квадрат)



- Сумма квадратов n нормально распределенных величин x_i=N(0,1)
- $\chi^2 = \Sigma x_i^2$
- n – число степеней свободы (n.d.f.)
- Среднее значение (матожидание): n
- Мода (максимум): при χ^2 = n-2

Распределение χ^2

$$f(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\,(\chi^2)^{n/2-1}\,\mathrm{e}^{-\chi^2/2}$$

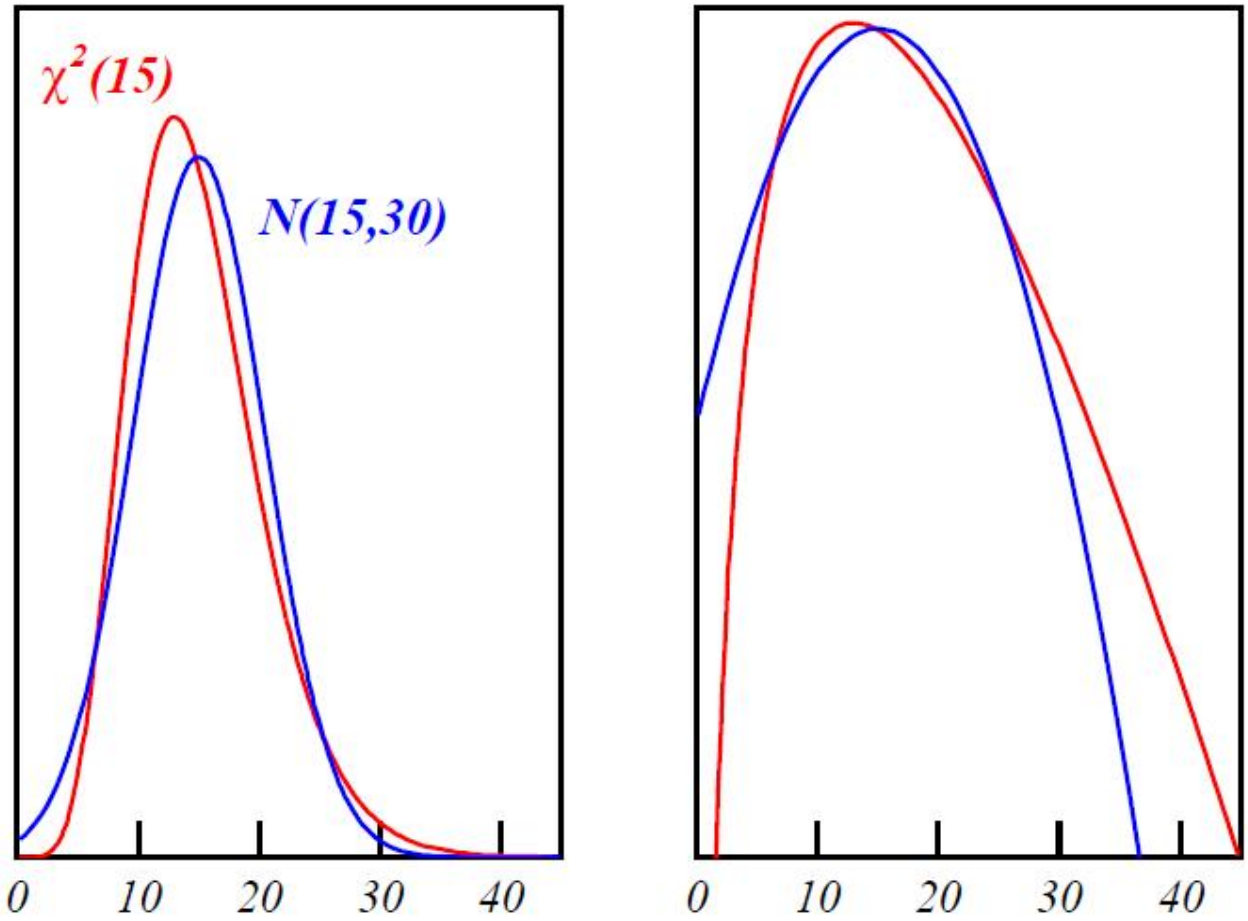
$$E(\chi^2) = n \qquad D(\chi^2) = 2n \quad \sigma(\chi^2) = \sqrt{2n}$$

- При $n \to \infty$ распределение стремится к нормальному $N(n,2n)$

Квантили χ^2

Уровень квантили →	90%	95%	99%	99.9%
n.d.f.= 1	2.71	3.84	6.64	10.83
n.d.f.= 2	4.60	5.99	9.21	13.82
n.d.f.= 5	9.24	11.07	15.09	20.52
n.d.f.= 10	15.99	18.31	23.21	29.59

Сравнение распределений χ^2(ndf=15) и нормального



Квантили χ^2

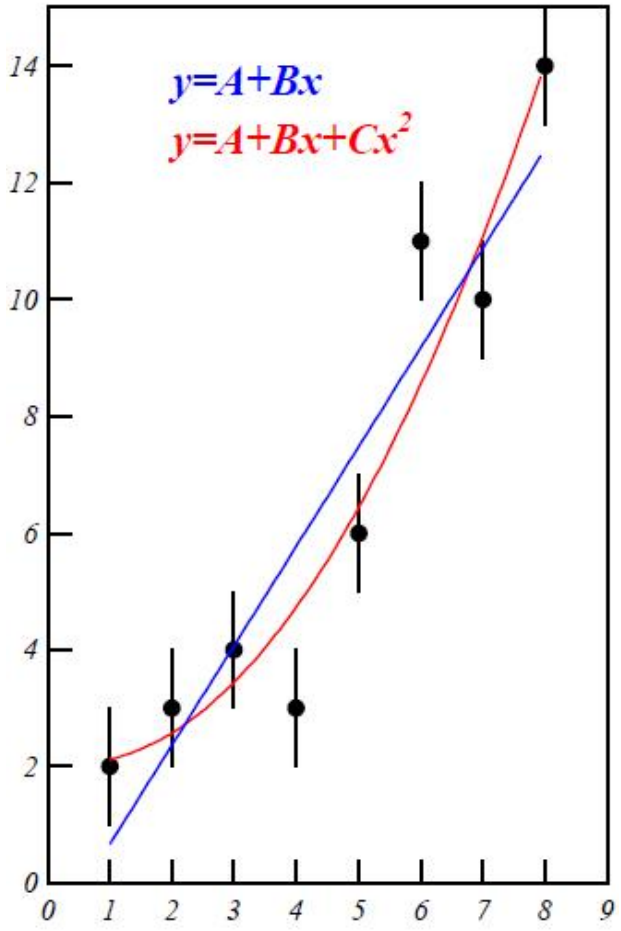
	90%	95%	99%
n.d.f.= 1	2.71	3.84	6.64
n.d.f.= 10	15.99	18.31	23.21

- Если при n.d.f.=10 фит дал χ^2 =23.21 – значит, согласие фита с данными довольно плохое
- Вероятность обнаружить столь же плохое согласие или ещё хуже (p-value) составляет 1%
- Это впрочем не значит, что фитируемая модель исключена с вероятностью 99%! Об этом ниже.

Качество фита «в уме»

- При большом числе степеней свободы распределение χ^2 близко к нормальному N(n,2n)
- Качество фита легко прикинуть без использования таблиц квантилей
- Пусть n.d.f.=100, и получен χ^2 =120.
- Отклонение от матожидания +20; среднеквадратичное отклонение $\sigma=\sqrt{(2*100)}=14$
- Итак, отклонение 20/14 = 1.4σ – качество фита удовлетворительное
- **Никогда** не приводите результат в виде χ^2/n.d.f.=1.20
- Обязательно указывайте χ^2/n.d.f.=120.0/100

Сколько нужно параметров фита?



- Усложняя фитируемую функцию (увеличивая число параметров), можно улучшить согласие фита с данными.
- Оправдано ли добавление новых параметров? Является ли улучшение статистически значимым?
- Если функция уже хорошо описывает данные, то при добавлении Δk новых параметров фита, χ^2 уменьшается в среднем на $\Delta\chi^2 = \Delta k$, причём $\sigma(\Delta\chi^2) = \sqrt{(2\Delta k)}$
- Итак, добавление параметров оправдано, если $\Delta\chi^2$ в несколько раз больше, чем $\sqrt{(2\Delta k)}$
- На рисунке: $\Delta k=1$, $\Delta\chi^2=7.4$. Превышение $6.4/\sqrt{(2=4.5\ \sigma}$
- Использование параболы оправдано!

Погрешность оценки параметра

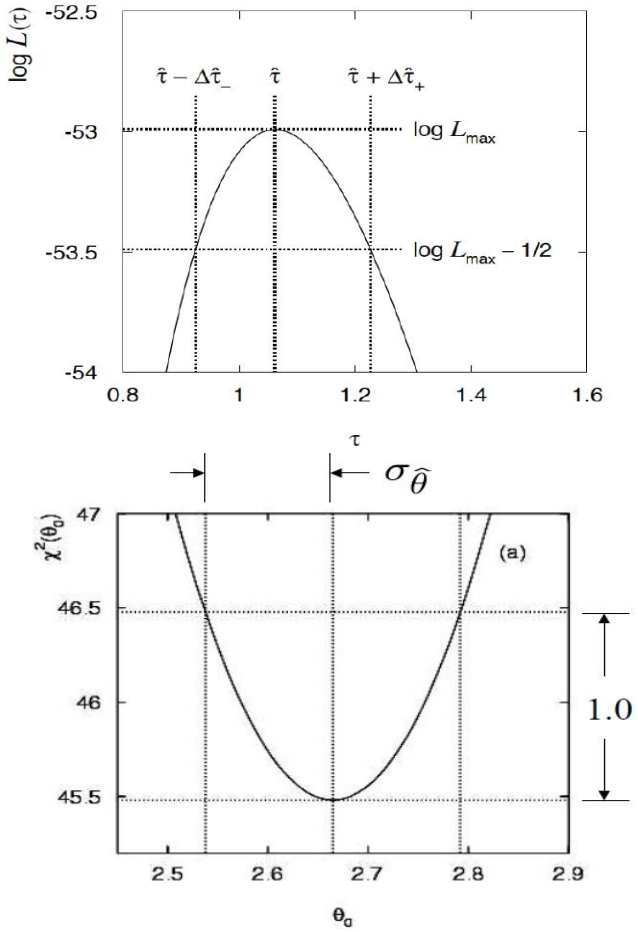
- Вспомним упавнение правдоподобия: $\partial(\ln L)/\partial\theta_i = 0$
- Разложим в окрестности максимума:

$$lnL(\theta) = lnL(\hat{\theta}) + \frac{1}{2}\frac{\partial^2 lnL}{\partial\theta^2}(\theta-\hat{\theta})^2 + \ldots = lnL_{\max} - \frac{(\theta-\hat{\theta})^2}{2\sigma_\theta^2}$$

- Вспомним, что правдоподобие – это вероятность: $p(\theta) \sim \mathrm{e}^{(\theta-\hat{\theta})^2/2\sigma_\theta^2}$
- Итак, оценка параметра имеет нормальное распределение, с погрешностью (стандартным отклонением) $\sigma_\theta^2 = -1 \,/\, \frac{\partial^2 lnL}{\partial\theta^2}$

Погрешность оценки параметра

- Погрешность оценки параметра $\sigma(\theta)$ – это интервал, на котором логарифм правдоподобия уменьшается на ½
- Аналогично для МНК, $\sigma(\theta)$ определяется увеличением χ^2 на 1.
- Иногда кривая несимметрична в районе экстремума. Тогда ошибка считается асимметричной: $\theta = 1.1^{+0.2}_{-0.3}$



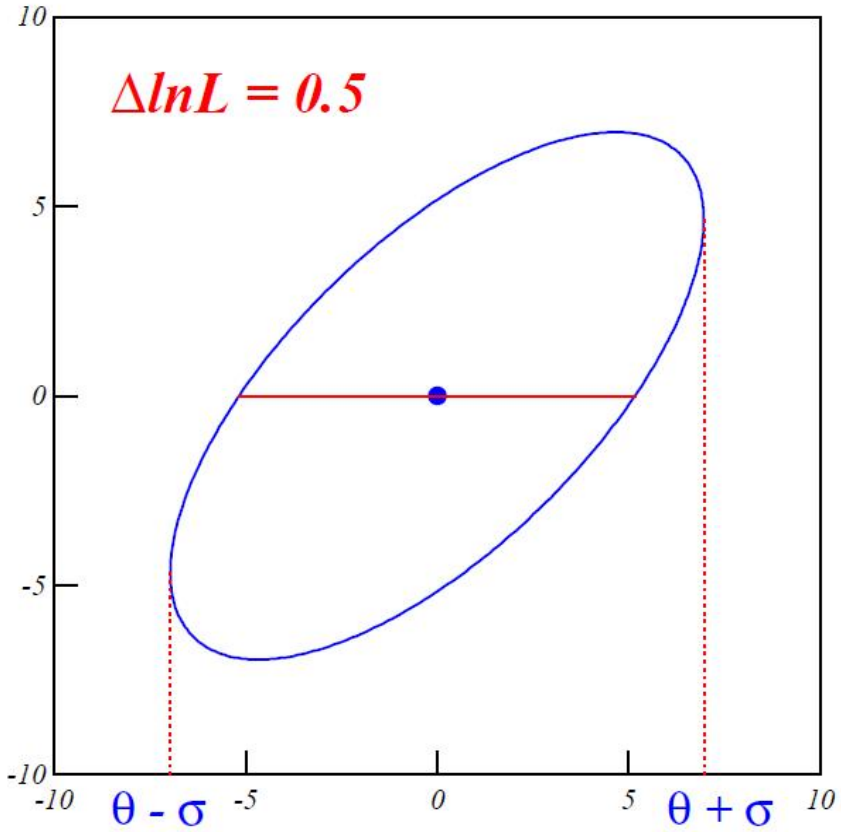
Оценка нескольких параметров

- Число параметров почти всегда больше, чем 1. Предыдущие рассуждения следует расширить на многомерный случай.
- Вместо $1/\sigma^2 = -(\ln L)''$ строим матрицу $U_{ij} = -\frac{\partial ln L}{\partial \theta_i \theta_j}$
- Погрешность σ заменяется на ковариационную матрицу $V=U^{-1}$:

$$V_{ij} = \overline{(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)} = \mathrm{cov}(\theta_i, \theta_j)$$

- Погрешность оценки параметра по прежнему задаётся приращением lnL на ½, или χ^2 на 1.

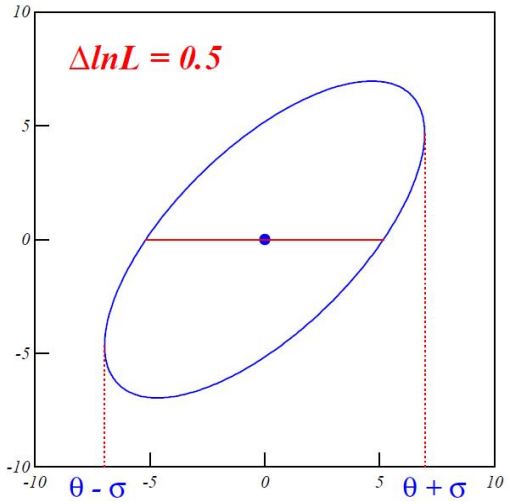
Эллипс ошибок



- Условие ΔlnL=1/2 задаёт эллипс (эллипсоид) ошибок
- В общем случае оценки параметров скоррелированы – эллипс «повёрнут»
- Ошибка оценки $\sigma(\theta)$ параметра задаётся касательными по «габаритам» эллипса
- Ошибка оценки параметра задаёт 68% вероятности именно этого параметра – по остальным параметрам подразумевается интегрирование

Сколько процентов вероятности лежит внутри эллипса ошибок?

- Для того, чтобы эллипс ошибок с определённой вероятностью содержал внутри себя все истинные параметры, эллипс следует прочертить на уровне $\Delta\chi^2 = \chi^2_{UP}$.
- Для каждой вероятности значение χ^2_{UP} совпадает с квантилями распределения χ^2, причём n.d.f. равно числу параметров



	50%	70%	90%	95%	99%
1	0.46	1.07	2.70	3.84	6.63
2	1.39	2.41	4.61	5.99	9.21
3	2.37	3.67	6.25	7.82	11.4
10	9.34	11.8	16.0	18.3	23.2

Оценка параметра по скоррелированным измерениям

- Корреляция возможна не только между оценками параметров θ_i, но и между исходными измерениями $x_i \pm \sigma_i$
- Ковариационная матрица (матрица ошибок)

$$V = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y & \rho_{xz}\sigma_x\sigma_z \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 & \rho_{yz}\sigma_y\sigma_z \\ \rho_{xz}\sigma_x\sigma_z & \rho_{yz}\sigma_y\sigma_z & \sigma_z^2 \end{pmatrix}$$

- Если два измерения включают независимую и общую погрешность, то общая погрешность является недиагональным элементом матрицы ошибок: $\begin{pmatrix} \sigma_x^2 + \sigma_0^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_y^2 + \sigma_0^2 \end{pmatrix}$
- При построении χ^2 заменяем $1/\sigma_i^2$ на V^{-1}.
- Вместо $\chi^2 = \Sigma(y-y_0)^2/\sigma^2$ имеем: $\chi^2(\theta) = (\vec{y} - \vec{y_0}(\theta))^T \; V^{-1} \; (\vec{y} - \vec{y_0}(\theta))$

Усреднение скоррелированных измерений (метод BLUE)

- По набору измерений $x_i \pm \sigma_i$ одной и той же величины требуется оценить параметр – эту самую измеряемую величину x
- Без корреляций процедура тривиальна:

$$\hat{\mu} = \left(\sum \frac{x_i}{\sigma_i^2}\right) / \left(\sum \frac{1}{\sigma_i^2}\right) \qquad \sigma^2 = 1 / \left(\sum \frac{1}{\sigma_i^2}\right)$$

- Требуется распространить процедуру на случай корреляций (матрица ошибок $V_{ij}=\mathrm{cov}(x_i,x_j)$)
- Ответ: $\hat{x} = \sum \alpha_i x_i = \vec{\alpha}\vec{x} \qquad \hat{\sigma^2} = \sum\sum V_{ij}\alpha_i\alpha_j = \vec{\alpha}^T V \vec{\alpha}$

$$\vec{\alpha} = V^{-1}\vec{E} / (\vec{E}^T V^{-1} \vec{E})$$

- E – вектор, составленный из единиц
- Ссылка: NIM A270 (1988) 110

Перенос ошибок (многомерный случай)

- Пусть имеется вектор величин $\theta = (\theta_1 .. \theta_n)$ и вектор функций $\eta(\theta) = (\eta_1(\theta) .. \eta_m(\theta))$. Имеется оценка параметров $\hat{\theta}_i$ с матрицей ошибок $V_{ij} = cov(\theta_i, \theta_j)$
- Требуется найти ошибки и корреляции (матрица ошибок U) для оценок $\hat{\eta} = \eta(\hat{\theta})$
- Искомая матрица $U = AVA^T$, где A – это матрица Якоби $A_{ij} = \partial\eta_i / \partial\theta_j$
- Результат точен для линейных функций η и не работает, если производные сильно меняются на интервалах порядка $\sigma(\theta)$