Математическая статистика

Уильям Петти (1623-1682). 1662 – "Трактат о налогах и сборах".
1676 – "Политическая арифметика".
1662 г. Джон Граунт "Таблицы смертности Лондона".

Корреляция

n пар переменных (x_i, y_i). Средние значения

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Коэффициент корреляции

$$< x\,y >= \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}); \quad < x^2 >= \sum_{i=1}^{n}(x_i - \bar{x})^2; \quad < y^2 >= \sum_{i=1}^{n}(y_i -$$

$$K = \frac{< x\,y >}{\sqrt{< x^2 >}\sqrt{< y^2 >}}; \quad -1 \le K \le 1.$$

Скалярное произведение двух векторов. K – косинус угла.
K сдвигово и масштабно инвариантен.

Исследование операций

1. Статистическая теория запасов
2. Теория очередей
3. Задача коммивояжера
4. Линейное программирование
5. Сетевые графики
6. Динамические экономические модели

Теорема Байеса

Совместное распределение вероятностей двух переменных

$$dp(x,h) = \rho(x,h)\,dx\,dh; \qquad \int \rho(x,h)\,dx\,dh = 1 \tag{1}$$

определяет *частные вероятности* x и h:

$$dp(x) = \int_h dp(x,h) = dx \int \rho(x,h)\,dh;$$

$$dp(h) = \int_x dp(x,h) = dh \int \rho(x,h)\,dx.$$

Основным инструментом в теории Байеса является *условные вероятности* $\rho(x/h)\,dx$ и $(\rho(h/x)\,dh)$, симметрично определяемые для x и h:

$$dp(x,h) = \rho(x,h)\,dx\,dh = dp(x/h)\cdot dp(h) = dp(h/x)\cdot dp(x). \tag{2}$$

Отсюда

$$dp(x/h) = \frac{dp(x,h)}{dp(h)}; \quad dp(h/x) = \frac{dp(x,h)}{dp(x)}.$$

Теорема Байеса

Из (2) можно выразить одну условную вероятность через другую:

$$dp(h/x) = dp(x/h)\,\frac{dp(h)}{dp(x)}. \qquad (3)$$

Будем полагать x случайной величиной с параметром распределения h, также являющимся случайной величиной вследствие нашего незнания. Пусть до измерения наши знания о параметре h определялись распределением $dp_{apr}(h)$ – *априорным* распределением, – а после проведения опыта, давшего определенное значение $x = x_e$, наши знания о параметре h изменились: после получения результата мы имеем *апостериорное* распределение

$$dp_{post}(h/x_e) = dp(x_e/h)\,\frac{dp_{apr}(h)}{dp(x_e)} = N^{-1}\,dp(x_e/h)\,dp_{apr}(h). \qquad (4)$$

Параметр бинарного распределения

Применим методику Байеса к бинарной системе, где h есть вероятность выпадения “+” и соответственно $1-h$ – вероятность “-”.
Априорная вероятность $dp_0(h) = dh;\ 0 < h < 1$.
При выпадении плюса

$$dp(h\,/+) = 2\,h\,dh \equiv dp(h\,/0,1).$$

При выпадении минуса

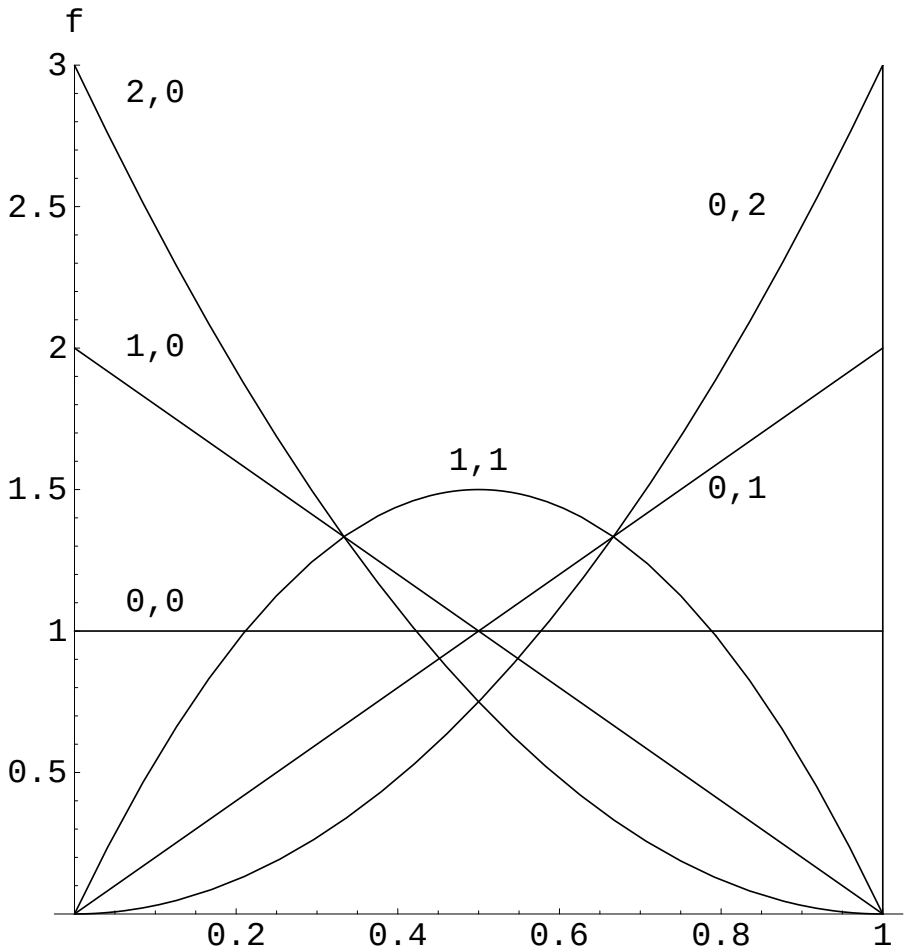
$$dp(h\,/-) = 2\,(1-h)\,dh \equiv dp(h\,/1,0).$$

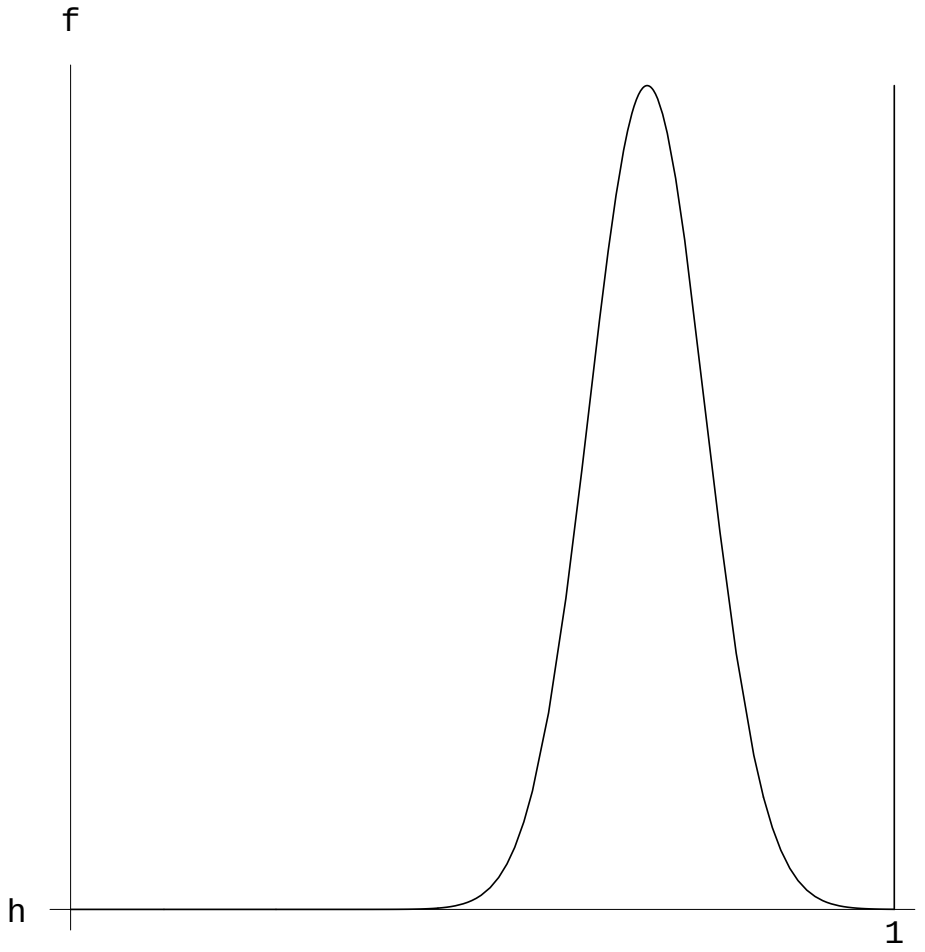
Результат эксперимента представляется парой целых чисел $(n_-,\ n_+)$.
Для последующих экспериментов эта вероятность становится априорной. Так, например,

$$dp(h\,/0,2) = N^{-1}\,h\,dp(h\,/0,1) = 3\,h^2\,dh; \quad dp(h\,/2,0) = 3\,(1-h)^2\,dh.$$

Распределение при малых n

На левом графике представлены априорная вероятность с меткой (0, 0) – горизонтальная прямая $h = 1$ и апостериорные вероятности при одном испытании (исходы (1,0) и (0,1)) и двух испытаниях ((n_-, n_+)=(2,0), (1,1), (0,2)).





После реализации $(n-m,\ m)$ распределение вероятности параметра h определяется β - распределением:

$$dp(h\,/(n-m,m)) = \frac{(n+1)!}{m!\,(n-m)!}\,h^m\,(1-h)^{n-m}. \qquad (5)$$

При некоторых достаточно больших значениях n и m график апостериорного распределения представлен на правом графике. Таким образом параметр h, определяющий результат статистического испытания, сам стал случайной величиной с распределением (5). Наиболее вероятное значение

$$\frac{d}{dh}\,\ln(\rho(h)) = \frac{m}{h} - \frac{n-m}{1-h} = 0; \quad h_0 = \frac{m}{n}.$$

Математическое ожидание h

$$< h >= \int_0^1 h\,dp(h) = \frac{m+1}{n+2}; \quad < 1-h >= \frac{n-m+1}{n+2}.$$

Дисперсия:

$$< h^2 >= \frac{(m+1)(m+2)}{(n+2)(n+3)}; \quad D_h = \frac{(m+1)(n-m+1)}{(n+2)^2(n+3)}.$$

Прогноз

Теперь мы хотим расчитать вероятности появления l плюсов в последующих, еще не проведенных k испытаниях. Если бы h была величина заданная, то

$$p(l\,/k,h) = \frac{k!}{l!\,(k-l)!}\,h^l\,(1-h)^{k-l}.$$

Теперь по формуле Байеса (2) можно вычислить *совместное распределение* случайных величин l и h, а затем, проинтегрировав по h, получим *частную* вероятность числа l:

$$dp(l,h/k) = \frac{k!}{l!\,(k-l)!}\,\frac{(n+1)!}{m!\,(n-m)!}\,h^{m+l}\,(1-h)^{n+k-l-m}\,dh;$$

$$p(l/k,n,m) = \int_h dp(l,h\,/k)\,dp(h\,/n,m) =$$

$$\frac{k!\,(n+1)!\,(m+l)!\,(n+k-m-l)!}{l!\,(k-l)!\,m!\,(n-m)!\,(n+k+1)!}.$$

Это – распределение случайной величины l.

Теоремы о параметрах прогноза

- Математическое ожидание равно математическому ожиданию по параметру от условного математического ожидания $< x/h >= \int_x x\, dp(x/h)$, усредненному по распределению случайного параметра.

$$< x >=< (< x/h >_x) >_h \equiv << x >> .$$

- Дисперсия имеет более сложную конструкцию:

$$D_x = \int_x (x - << x >>)^2\, dp(x) =$$

$$\int_h \int_x ((x - < x/h >) + (< x/h > - << x >>)^2\, dp(x/h)\, dp(h) =$$

$$\int_h \left(\int_x ((x - < x/h >)\, dp(x/h) \right)^2\, dp(h) +$$

$$\int_h (< x/h > - << x >>)^2\, dp(h) =< D_{x/h} > + D_{<x/h>}. \quad (6)$$

Она равна условной дисперсии, усредненной по параметру,

Прогноз параметров биномиального распределения

Например, в рассмотренном выше случае прогноза на k испытаний после n испытаний, в которых выпало m плюсов для случайной величины l, математическое ожидание

$$<< l >>=< k\,h >= \frac{k\,(m+1)}{(n+2)}.$$

Дисперсия в соответствии с (6) ведет себя сложнее:

$$D_l =< k\,h\,(1-h) > + \left(< (k\,h)^2 > - < k\,h >^2\right) =$$

$$\frac{k\,(m+1)\,(n-m+1)}{(n+2)\,(n+3)} + k^2\,\frac{(m+1)\,(n-m+1)}{(n+2)^2\,(n+3)} =$$

$$(k + \frac{k^2}{n+2})\,\frac{(m+1)\,(n-m+1)}{(n+2)\,(n+3)}.$$

Пока объем прогноза много меньше объема проделанного эксперимента $k \ll n$, дисперсия прогноза растет линейно по k. При объеме прогноза, превышающем объем эксперимента, зависимость от k оказывается квадратичной.

Относительная погрешность

Проследим за относительной погрешностью $\delta = \sigma_I / << I >>$, где $\sigma_I = \sqrt{D_I}$ – норма:

$$\delta = \sqrt{\left(\frac{1}{k} + \frac{1}{n+2}\right)} \sqrt{\frac{(m+1)(n-m+1)}{(n+2)(n+3)}} \frac{(n+2)}{(m+1)}.$$

Здесь n и m – объем и результат проведенного эксперимента, а k – объем прогноза. При $k \to \infty$ относительная ошибка имеет конечный предел в отличие от похожей ситуации в законе больших чисел теории вероятностей.

Параметры нормального распределения

Вследствие центральной предельной теоремы многие непрерывные случайные величины распределены нормально.

$$dp(x\,/\mu, D) = \frac{1}{\sqrt{2\,\pi\,D}}\,e^{-\frac{(x-\mu)^2}{2\,D}}.$$

Вероятность появления набора случайных величин $(x_1,\, x_2,\, \ldots,\, x_n)$ пропорциональна

$$\rho(x_1, x_2, \ldots, x_n\,/\mu, D) = \frac{1}{(2\pi\,D)^{n/2}}\,e^{-\frac{(\mu-\bar{x})^2}{2\,D}-\frac{\Delta^2}{2\,D}},$$

где показатель экспоненты есть результат раскрытия суммы

$$\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\,D} = \frac{1}{2\,D}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{n}{2\,D}(\mu-\bar{x})^2 =$$

$$\frac{n(\mu-\bar{x})^2}{2\,D}+\frac{\Delta^2}{2\,D};\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i;\quad \sum_{i=1}^{n}(x_i-\bar{x}) = 0;\quad \Delta^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Применение формулы Байеса

Теперь по формуле Байеса

$$dp(\mu, D\,/x_1, x_2, \ldots, x_n) = dp(\mu, D\,/\bar{x}, \Delta^2) =$$

$$N^{-1}\,\frac{e^{-\frac{\Delta^2}{2\,D}}}{D^{n/2}}\,e^{-\frac{n\,(\mu-\bar{x})^2}{2\,D}}\,dp_0(\mu, D). \qquad (7)$$

Здесь случайными величинами являются μ и D, а экспериментальные данные входят только в виде двух комбинаций: среднего значения $\bar{x}$ и суммарной квадратичной ошибки Δ^2. Множитель $dp_0(\mu, D)$ есть априорная вероятность распределения μ и D. Максимальная неопределенность для μ – это равновероятность всех μ: $dp_0(\mu) \sim d\mu$. Для D это не так, так как D – положительно определенная величина: $dp_0(D) \sim dD/D$. После такой подстановки распределение (30) принимает вид:

$$dp(\mu, D\,/\bar{x}, \Delta^2) = N^{-1}\,\frac{e^{-\frac{\Delta^2}{2\,D}}}{D^{n/2+1}}\,e^{-\frac{n\,(\mu-\bar{x})^2}{2\,D}}\,d\mu\,dD. \qquad (8)$$

Распределение дисперсии

Распределение (8) несложно проинтегрировать по μ, так как это гауссов интеграл:

$$\int\limits_{-\infty}^{\infty} e^{-\frac{n\,(\mu-\bar{x})^2}{2\,D}}\,d\mu = \sqrt{\frac{2\,\pi\,D}{n}},$$

после чего останется распределение дисперсии D

$$dp(D\,/\Delta^2) = N^{-1}\sqrt{\frac{2\,\pi}{n}}\,\frac{e^{-\frac{\Delta^2}{2\,D}}}{D^{(n-1)/2+1}}\,dD.$$

Это есть Г-распределение для величины $u = \frac{\Delta^2}{2\,D}$.

$$D = \frac{\Delta^2}{2\,u};\quad dD = -\frac{\Delta^2}{2\,u^2}\,du;$$

$$dp(u) = N^{-1}\sqrt{\frac{2\,\pi}{n}}\left(\frac{2}{\Delta^2}\right)^{\frac{n-1}{2}}\,u^{(n-1)/2-1}e^{-u}\,du =$$

$$\frac{1}{\Gamma(\frac{n-1}{2})}\,u^{(n-1)/2-1}e^{-u}\,du.$$

Отсюда определяется нормировочный множитель в (8):

$$N = \sqrt{\frac{2\,\pi}{n}} \left(\frac{2}{\Delta^2}\right)^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2}). \qquad (9)$$

Математическое ожидание дисперсии

$$< D >= \frac{\Delta^2}{2} \left\langle \frac{1}{u} \right\rangle = \frac{\Delta^2}{2} \frac{1}{\Gamma(\frac{n-1}{2})} \int\limits_0^\infty u^{(n-3)/2-1} e^{-u}\, du = \qquad (10)$$

$$\frac{\Delta^2}{2} \frac{\Gamma(\frac{n-3}{2})}{\Gamma(\frac{n-1}{2})} = \frac{\Delta^2}{n-3}.$$

Для оценки математического ожидания дисперсии нужно более трех замеров.

Распределение математического ожидания

Подставив в выражение (8) найденное значение нормировочного множителя (9)

$$dp(\mu, D\,/\bar{x}, \Delta^2) = \sqrt{\frac{n}{2\,\pi}}\left(\frac{\Delta^2}{2}\right)^{\frac{n-1}{2}} \frac{1}{\Gamma(\frac{n-1}{2})}\,\frac{e^{-\frac{\Delta^2}{2\,D}}}{D^{(n/2+1)}}\,e^{-\frac{n\,(\mu-\bar{x})^2}{2\,D}}\,d\mu\,dD, \tag{11}$$

собрав показатель экспоненты в одну функцию w, получаем статистически независимое распределение для случайных величин μ и w:

$$dp(\mu, w) = \left(\sqrt{\frac{n}{\pi\,\Delta^2}}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\,\frac{d\mu}{\left(1+\frac{n(\mu-\bar{x})^2}{\Delta^2}\right)^{\frac{n}{2}}}\right)\cdot\left(\frac{e^{-w}}{\Gamma(\frac{n}{2})}\,w^{\frac{n}{2}-1}\,dw\right) =$$

$$dp(\mu)\cdot dp(w); \quad w = \frac{1}{2\,D}\,(\Delta^2 + n(\mu-\bar{x})^2).$$

Первый сомножитель в этом распределении определяет *распределение Стьюдента* с $m = n-1$ *степенями свободы*.

Математическое ожидание и дисперсия параметра μ

Из соотношения

$$\mu = \bar{x} + y_{n-1}\sqrt{\frac{\Delta^2}{n(n-1)}}; \quad m = n-1$$

Можно получить математическое ожидание и дисперсию оцениваемой величины μ:

$$< \mu >= \bar{x}; \quad D_\mu = \frac{\Delta^2}{n(n-1)}\,\frac{n+1}{n-1} = \left(1+\frac{1}{n}\right)\,\frac{\Delta^2}{(n-1)^2},$$

а также определить *доверительный интервал*

$$\mu = \bar{x} \pm\, y_{n-1}(\eta)\sqrt{\frac{\Delta^2}{n\,(n-1)}}\,,$$

где $y_{n-1}(\eta)$ – *квантиль* надежности η распределения Стьюдента с $n-1$ степенями свободы: $\pm y_{n-1}(\eta)$ отсекают под кривой площадь η (например, 0.9, 0.95), определяющую *надежность вывода*. Квантили заданной надежности распределения Стьюдента с m степенями свободы приводятся в таблицах.

Прогноз

При заданных μ и D распределение случайной величины x определяется нормальным распределением

$$dp(x/\mu, D) = \frac{1}{\sqrt{2\pi D}} \cdot e^{-\frac{(x-\mu)^2}{2D}}\, dx.$$

Но сами эти параметры оказались случайными величинами, определяемыми распределением (11), так что с учетом этой вероятности прогнозируемое значение x имеет распределение

$$dp(x) = \int_{\mu,D} dp(x/\mu, D)\, dp(\mu, D) = N^{-1}\, dx \int d\mu\, dD \left(e^{-\frac{f}{2D}}\right) \frac{1}{D^{\frac{n+1}{2}+1}} \tag{12}$$

где

$$f = \Delta^2 + n(\mu - \bar{x})^2 + (x-\mu)^2 = \Delta^2 + (\mu - \bar{\mu})^2 + \frac{n(x-\bar{x})^2}{n+1};$$

$$\bar{\mu} = \frac{n\bar{x} + x}{n+1}.$$

После интегрирования по μ и D, уменьшающего степень D в знаменателе на 3/2, распределение прогнозируемой величины x оказывается распределением Стьюдента с $m = n - 1$ степенью свободы:

$$dp(x) = N^{-1} \frac{dx}{\left(1 + \frac{n\,(x-\bar{x})^2}{(n+1)\,\Delta^2}\right)^{\frac{n}{2}}} \sim \frac{dS}{\left(1 + \frac{S^2}{m}\right)^{\frac{m+1}{2}}}.$$

$$x = \bar{x} + S_{n-1}\sqrt{\frac{(n+1)\,\Delta^2}{n\,(n-1)}}; \quad < x >= \bar{x}; \quad D_x = \frac{(n+1)\,\Delta^2}{n(n-3)}.$$

В отличие от дисперсии параметра μ, уменьшающейся при увеличении числа замеров, так как Δ^2 растет пропорционально n, дисперсия x с ростом числа замеров стабилизируется. Наши знания о параметрах растут, но объективная неопределенность в системе не уменьшается.

Распределение Пуассона

При заданном μ

$$p_m = \frac{\mu^m}{m!}\, e^{-\mu}.$$

Если проведено k замеров m_1, $m_2, \ldots,$ m_k, то по формуле Байеса ($n = m_1 + m_2 + \cdots + m_k$)

$$dp(\mu) = \frac{k^{n+1}}{n!}\, \mu^n\, e^{-k\,\mu}\, d\mu. \tag{13}$$

$$< \mu >= \frac{n+1}{k}; \quad < \mu^2 >= \frac{(n+1)\,(n+2)}{k^2}. \tag{14}$$

Распределение Паскаля

При заданном γ

$$p_m = (1-\gamma)\,\gamma^m.$$

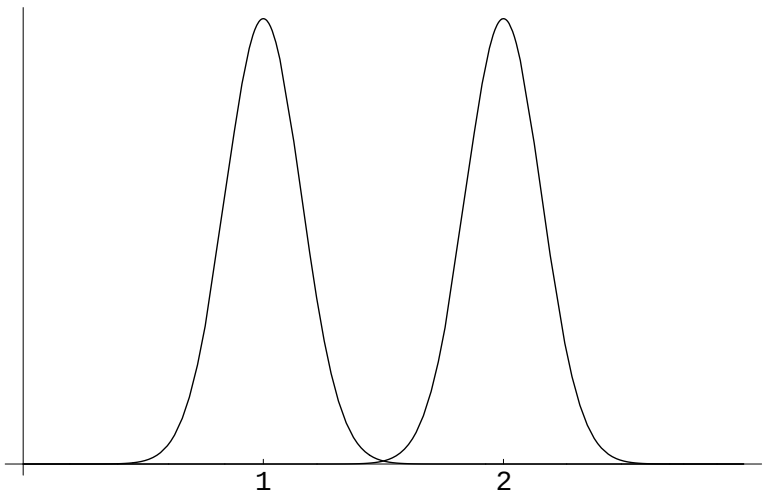
Если проведено k замеров m_1, $m_2, \ldots,$ m_k, то по формуле Байеса ($n = m_1 + m_2 + \cdots + m_k$)

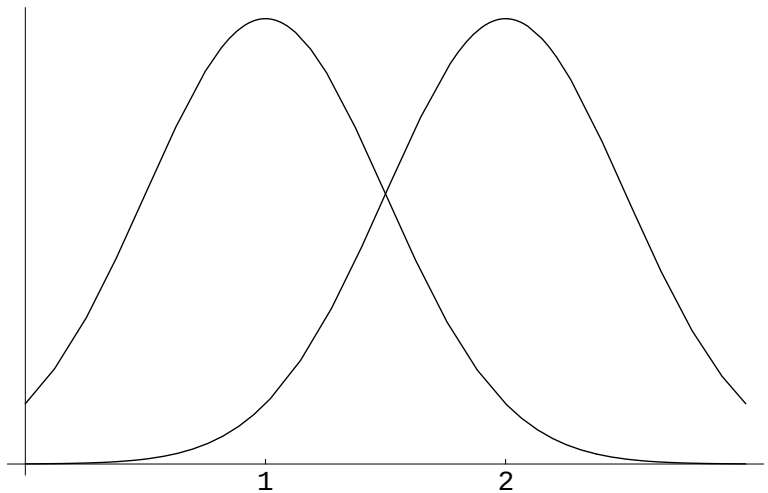
$$dp(\gamma) = \frac{(n+k+1)!}{k!\,n!}\,(1-\gamma)^k\,\gamma^n\,d\gamma. \tag{15}$$

$$<\gamma>=\frac{n+1}{n+k+2}; \quad <\gamma^2>=\frac{(n+1)\,(n+2)}{(n+k+2)\,(n+k+3)}; \tag{16}$$

Различение математических ожиданий

Если в двух различных сериях измерений получены слегка различные значения $\bar{x}$, являются ли эти различия результатом замера одной и той же величины, или же в каждой серии имелась специфика, приведшая к измерению разных величин? На рисунках видно, что вывод определяется не разностью, а отношением этой разности к погрешности измерений:





Для установления математического критерия различности средних нужно найти методом Байеса распределение параметров μ_1 и μ_2 в этих сериях с длинами выборок n_1 и n_2:

$$dp(\mu_1, \mu_2, \sigma) = N^{-1} e^{-\frac{\Delta_1^2}{2\sigma^2} - \frac{\Delta_2^2}{2\sigma^2}} \left(e^{-\frac{n_1 (\mu_1 - \bar{x}_1)^2 + n_2 (\mu_2 - \bar{x}_2)^2}{2\sigma^2}} \right) \frac{d\mu_1 \, d\mu_2 \, d\sigma}{\sigma^{n_1+n_2+1}}.$$

Основную роль играет выражение в показателе экспоненты

$$s = n_1 (\mu_1 - \bar{x}_1)^2 + n_2 (\mu_2 - \bar{x}_2)^2.$$

Выделим из переменных μ_1 и μ_2 их разность

$$\mu_1 = \mu + \frac{n_2}{n_1 + n_2} \lambda; \quad \mu_2 = \mu - \frac{n_1}{n_1 + n_2} \lambda; \quad \lambda = \mu_1 - \mu_2.$$

Тогда

$$s = (n_1 + n_2)(\mu - \bar{\bar{x}})^2 + \frac{n_1\, n_2}{n_1 + n_2}\,(\lambda - \delta)^2,$$

где

$$\bar{\bar{x}} = \frac{n_1\, \bar{x}_1 + n_2\, \bar{x}_2}{n_1 + n_2}; \quad \delta = \bar{x}_1 - \bar{x}_2.$$

После интегрирования по μ зависимость от λ определится стьюдентовской функцией с $m = n_1 + n_2 - 2$ степенями свободы

$$\left(1 + \frac{n_1\, n_2\,(\lambda - \delta)^2}{(n_1 + n_2)\,(\Delta_1^2 + \Delta_2^2)}\right)^{-(n_1+n_2-1)/2} = \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2},$$

откуда можно определить

$$< \mu_1 - \mu_2 > = < \lambda > = \bar{x}_1 - \bar{x}_2; \quad D_\lambda = \frac{(n_1 + n_2 - 2)(n_1 + n_2)}{n_1\, n_2}\,\frac{\Delta_1^2 + \Delta_2^2}{(n_1 + n_2 -}$$

Значимость разности $\bar{x}_1 - \bar{x}_2$ определяется ее отношением к корню из дисперсии.

Дискриминатор

Дискриминатор распределения величины $\delta\gamma$ строится из первого (m_1) и второго (m_2) моментов распределения которые могут быть вычислены через математические ожидания величин γ_1 и γ_2. Он определяется как отношение

$$d = \frac{m_1^2}{m_2} \leq 1. \tag{17}$$

Первый момент (угловыми скобками мы обозначаем математическое ожидание)

$$m_1 = < \delta\gamma > = < \gamma_1 > - < \gamma_2 > . \tag{18}$$

Дисперсия величины $\delta\gamma$ является суммой дисперсий γ_1 и γ_2, а с другой стороны вычисляется через m_1 и m_2:

$$D_{\delta\gamma} = m_2 - m_1^2 = D_{\gamma_1} + D_{\gamma_2}. \tag{19}$$

Отсюда второй момент определяется через параметры первой и второй серий

$$m_2 = D_{\gamma_1} + D_{\gamma_2} + (<\gamma_1> - <\gamma_2>)^2 \tag{20}$$

и, подставляя в (17), окончательно определяем выражение для дискриминатора:

$$d = \frac{(<\gamma_1> - <\gamma_2>)^2}{D_{\gamma_1} + D_{\gamma_2} + (<\gamma_1> - <\gamma_2>)^2} \tag{21}$$

Для вычисления дискриминатора как функции экспериментально замеренных параметров нужно методом Байеса найти распределение математического ожидания параметра γ, но не нужно находить явный вид распределения разности двух параметров.

Биномиальное распределение

$$dp(\gamma) = \frac{(n+1)!}{m!\,(n-m)!}\,\gamma^m\,(1-\gamma)^{n-m}\,d\gamma \tag{22}$$

$$<\gamma>=\frac{m+1}{n+2};\quad <\gamma^2>=\frac{(m+1)\,(m+2)}{(n+2)\,(n+3)}. \tag{23}$$

$$D_\gamma =<\gamma^2> - <\gamma>^2= \frac{(m+1)\,(n-m+1)}{(n+2)^2\,(n+3)}. \tag{24}$$

$$d(m_1,n_1,m_2,n_2) = \frac{(\frac{m_1+1}{n_1+2}-\frac{m_2+1}{n_2+2})^2}{\frac{(m_1+1)(n_1-m_1+1)}{(n_1+2)^2(n_1+3)}+\frac{(m_2+1)(n_2-m_2+1)}{(n_2+2)^2(n_2+3)}+(\frac{m_1+1}{n_1+2}-\frac{m_2+1}{n_2+2})^2}.$$

Распределение Пуассона

$$dp(\mu) = \frac{k^{n+1}}{n!}\,\mu^n\,e^{-k\,\mu}\,d\mu. \tag{25}$$

$$<\mu>=\frac{n+1}{k};\quad <\mu^2>=\frac{(n+1)\,(n+2)}{k^2};\quad D_\mu=\frac{n+1}{k^2}. \tag{26}$$

Распределение Паскаля

$$dp(\gamma) = \frac{(n+k+1)!}{k!\,n!}\,(1-\gamma)^k\,\gamma^n\,d\gamma. \tag{27}$$

$$<\gamma>=\frac{n+1}{n+k+2};\quad <\gamma^2>=\frac{(n+1)\,(n+2)}{(n+k+2)\,(n+k+3)}; \tag{28}$$

$$D_\gamma = \frac{(n+1)\,(k+1)}{(n+k+2)^2\,(n+k+3)}. \tag{29}$$

Нормальное распределение

$$dp(\mu, D\,/x_1, x_2, \ldots, x_n) = dp(\mu, D\,/\bar{x}, \Delta^2) =$$

$$N^{-1}\,\frac{e^{-\frac{\Delta^2}{2\,D}}}{D^{n/2}}\,e^{-\frac{n\,(\mu-\bar{x})^2}{2\,D}}\,dp_0(\mu, D). \qquad (30)$$

$$dp(\mu, w) = \left(\sqrt{\frac{n}{\pi\,\Delta^2}}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\,\frac{d\mu}{\left(1+\frac{n(\mu-\bar{x})^2}{\Delta^2}\right)^{\frac{n}{2}}}\right)\cdot\left(\frac{e^{-w}}{\Gamma(\frac{n}{2})}\,w^{\frac{n}{2}-1}\,dw\right) =$$

$$dp(\mu)\cdot dp(w); \quad w = \frac{1}{2\,D}\,(\Delta^2 + n(\mu-\bar{x})^2).$$

Первый сомножитель в этом распределении определяет *распределение Стьюдента* с $m = n-1$ *степенями свободы* с математическим ожиданием и дисперсией

$$< y_m >= 0; \quad D_{y_m} = \frac{m}{m-2} = \frac{n-1}{n-3}. \qquad (31)$$

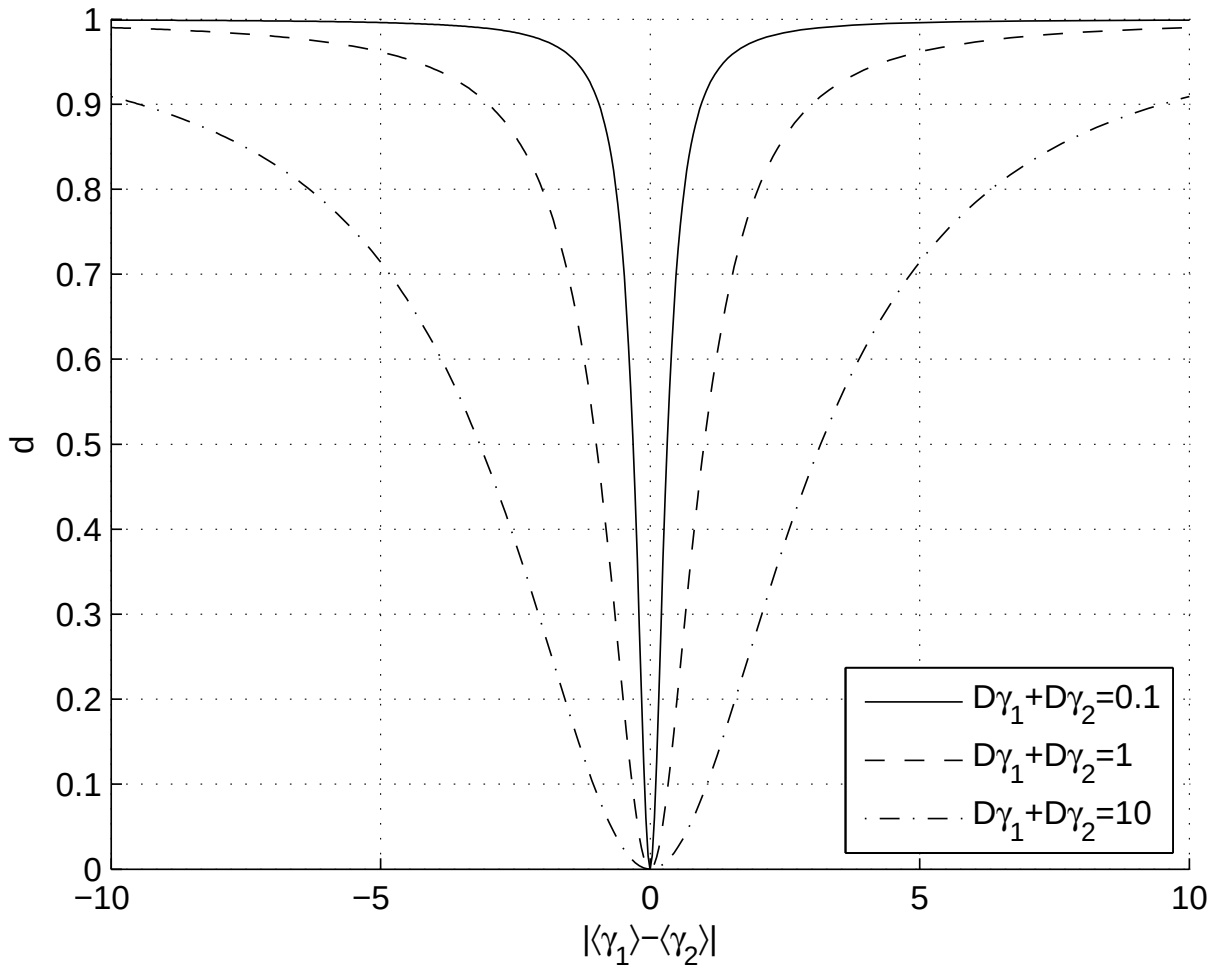
Из соотношения

$$\mu = \bar{x} + y_{n-1}\sqrt{\frac{\Delta^2}{n\,(n-1)}}$$

можно получить математическое ожидание и дисперсию оцениваемой величины μ:

$$< \mu >= \bar{x}; \quad D_\mu = \frac{\Delta^2}{n(n-1)}\,\frac{n-1}{n-3} = \frac{\triangle^2}{n\,(n-3)}. \qquad (32)$$

Испытания метода



Полиномиальное распределение

k возможных событий: 1, 2, ..., k. (Неизвестные) вероятности $p_1, \dots p_k$, причем $\sum_{i=1}^{k} p_i = 1$. n-кратный эксперимент:

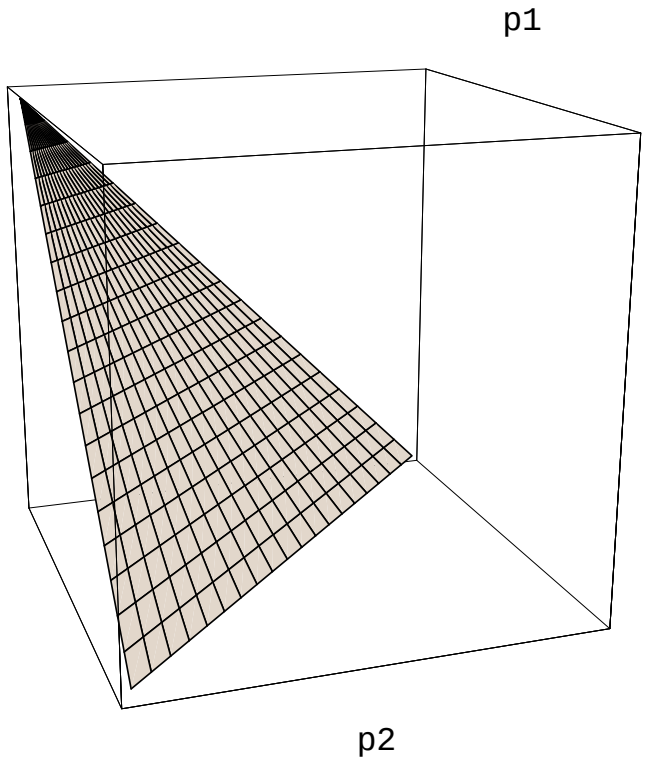
$$m_1,\ m_2, \dots,\ m_k; \quad \sum_{i=1}^{k} m_i = n.$$

Полиномиальное распределение при известных p_i:

$$P(m_1, m_2, \dots) = \frac{n!}{m_1! m_2! \dots m_k!}\, p_1^{m_1}\, p_2^{m_2} \dots p_k^{m_k}.$$

Априорная вероятность

$$dP_0(p_1, \dots, p_k) = \frac{dp_1\, dp_2\, ..dp_{k-1}}{(k-1)!}.$$



Объем $k-1$-мерного симплекса с единичными сторонами $1/(k-1)!$.

$$dP(p_1, p_2, .., dp_{k-1}) =$$

$$\frac{(n+k-1)!}{m_1!\, m_2!, ..m_k!}\, p_1^{m_1}\, p_2^{m_2} .. p_{k-1}^{m_{k-1}} (1 - p_1 .. - p_{k-1})^{m_k}$$

$$\times dp_1\, dp_2\, ..dp_{k-1}. \qquad < p_i >= \frac{m_i + 1}{n+k}.$$

Статистические гипотезы. Критерий согласия χ^2

Применяется для проверки гипотезы о том, что эксперимент объема n с k состояниями, в результате которого в каждом i-м состоянии выпало r_i событий, определяется вероятностями p_i, оцененными какими-либо статистическими методами. При этом

$$\sum_{i=1}^{k} r_i = n. \tag{33}$$

Математические ожидания чисел r_i равны $< r_i >= m_i = n\,p_i$, а при достаточно больших m_i (больших 10-и) отклонения $r_i - m_i$ вследствие центральной предельной теоремы подчиняются нормальному распределению с дисперсией

$$D_i = n\,p_i\,(1 - p_i) \approx n\,p_i,$$

если состояний достаточно много и каждое p_i мало.

Тогда величину

$$z_i = \frac{r_i - n\,p_i}{\sqrt{n\,p_i}} = \frac{r_i - m_i}{\sqrt{m_i}}$$

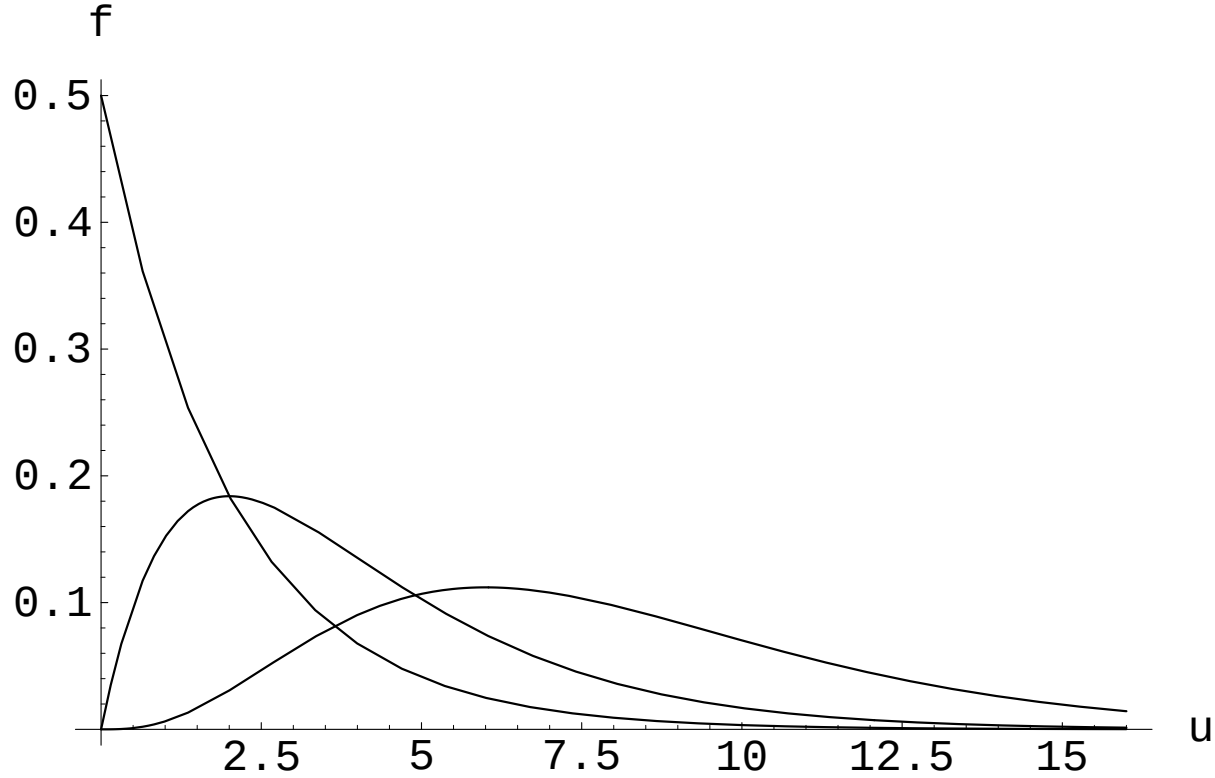
можно полагать нормально распределенной с единичной дисперсией, а величина

$$u = \sum_{i=1}^{k} z_i^2 = \sum_{i=1}^{k} \frac{(r_i - m_i)^2}{m_i} \tag{34}$$

распределена примерно по закону χ^2.
Число статистических степеней свободы в χ^2 - распределении:

$$\nu = k - 1 - t. \tag{35}$$

Площадь под кривой χ^2-распределения с ν степенями свободы от нуля до вычисленного значения u определяет вероятность *неверности гипотезы*. Если u велико, вероятность отклонения гипотезы оказывается высокой.



Регрессия

Определению подлежит зависимость

$$y(x) = a_1\, u_1(x) + a_2\, u_2(x) + \cdots + a_m\, u_m(x) + \delta, \qquad (36)$$

где $u_1(x),\ u_2(x), \ldots, u_m(x)$ – заданные *регрессионные функции*, коэффициенты $a_1,\ a_2, \ldots, a_m$ – подлежащие определению *регрессионные коэффициенты*.
Схема данных: имеется N пар данных $(x_i,\ y_i);\ i = 1, \ldots N$.

$$\delta_i = y_i - \sum_{j=1}^{m} a_j\, u_j(x_i); \quad j = 1, \ldots, N.$$

При заданных регрессионных коэффициентах a_j и дисперсии D вероятность появления набора ошибок δ_i определяется произведением гауссовых распределений

$$dp(\delta_1, \ldots, \delta_N / D, a_1, \ldots a_m) \sim \prod_{i=1}^{N} \left(\frac{e^{-\frac{\delta_i^2}{2D}}}{\sqrt{D}} \right) \sim \frac{1}{D^{\frac{N}{2}}} e^{-\frac{f}{2D}}, \quad (37)$$

где

$$f = f(a_1, .., a_m) = \sum_{i=1}^{N} \delta_i^2 = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{m} a_j \, u_j(x_i) \right)^2 . \quad (38)$$

Распределение коэффициентов a_i и дисперсии

Коэффициенты a_i, D – случайные величины с априорной мерой вероятности

$$dp_{apr}(a_1, a_2, \ldots, a_n, D) \sim da_1\, da_2 \ldots da_n\, \frac{dD}{D},$$

что приводит к распределению параметров a_j и D:

$$dp(a_1, a_2, \ldots, a_n, D) = N^{-1}\, \frac{1}{D^{\frac{N}{2}}}\, \left(e^{-\frac{f}{2D}}\right) da_1\, da_2 \ldots da_n\, \frac{dD}{D}. \tag{39}$$

Функция f квадратична по коэффициентам a_j:

$$f = \left(\sum_{i=1}^{N} y_i^2\right) - 2\sum_{j=1}^{m} a_j \left(\sum_{i=1}^{N} y_i\, u_j(x_i)\right) + \sum_{j=1}^{m}\sum_{k=1}^{m} a_j\, a_k \left(\sum_{i=1}^{N} u_j(x_i)\, u_k(x_i)\right)$$

Обозначим выражения в круглых скобках как $< y^2 >$, $< y\, u_j > \equiv b_j$ и $< u_j\, u_k > \equiv K_{jk}$. Минимум f при

$$b_j = \sum_{l=1}^{m} K_{jl}\, \bar{a}_l; \quad \bar{a}_l = \sum_{l=1}^{m} K_{lj}^{-1}\, b_j. \tag{40}$$

Регрессия прямой линией

$y = a + b\,x$ с двумя регрессионными функциями $u_1 = 1$, $u_2 = x$.

$$K_{11} =< 1 \cdot 1 >= N;\ K_{12} = K_{21} = \sum_{i=1}^{N} x_i = N\,\bar{x};\ K_{22} = \sum_{i=1}^{N} x_i^2.$$

Чтобы матрица K_{ij} была диагональной, удобнее взять за исходные функции $u_1 = 1$, $u_2 = x - \bar{x}$, тогда $< u_1 \cdot u_2 >= \sum(x_i - \bar{x}) = 0$ и матрица K_{jl} имеет лишь диагональные элементы:

$$K_{11} = N; \quad K_{22} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \equiv N\,L^2.$$

Наилучшие коэффициенты $\bar{a}$ и $\bar{b}$ определяются из (40):

$$\bar{a} = \frac{< y >}{N}; \quad \bar{b} = \frac{< y\,(x - \bar{x}) >}{N\,L^2}.$$

Квадратичная форма при наилучших коэффициентах определяет суммарную квадратичную погрешность:

$$\Delta^2 = \sum_{i=1}^{N} (y_i - \bar{a} - \bar{b}\,(x_i - \bar{x}))^2,$$

а при произвольных a и b:

$$f = \Delta^2 + N\,(a - \bar{a})^2 + N\,L^2\,(b - \bar{b})^2,$$

и формула (39) приводит к распределению параметров a, b, D:

$$dp(a, b, D) = N^{-1}\,\frac{1}{D^{\frac{N}{2}}}\,\left(e^{-\frac{\Delta^2 + N\,(a-\bar{a})^2 + N\,L^2\,(b-\bar{b})^2}{2\,D}}\right) da\,db\,\frac{dD}{D}. \quad (41)$$

После интегрирования по D эта формула дает *двумерное распределение Стьюдента* для коэффициентов a и b, пропорциональное

$$dp(a,b) \sim \frac{da\,db}{\left(1+\frac{N\left((a-\bar{a})^2+L^2\,(b-\bar{b})^2\right)}{\Delta^2}\right)^{\frac{N}{2}}}.$$

Если же сначала проинтегрировать по b, что просто уменьшит на $1/2$ степень по D в знаменателе, а затем по D, то результатом будет распределение коэффициента a – распределение Стьюдента с $N-2$ степенями свободы:

$$dp(a) \sim \frac{da}{\left(1+\frac{(N-1)\,(a-\bar{a})^2}{\Delta^2}\right)^{\frac{N-1}{2}}} \sim \frac{dy}{\left(1+\frac{y^2}{m}\right)^{\frac{m+1}{2}}}. \qquad (42)$$

Аналогично, интегрирование по a и D приводит к распределению Стьюдента для коэффициента b:

$$dp(b) \sim \frac{db}{\left(1 + \frac{(N-1)\,L^2\,(b-\bar{b})^2}{\Delta^2}\right)^{\frac{N-1}{2}}} \sim \frac{dy}{\left(1 + \frac{y^2}{m}\right)^{\frac{m+1}{2}}}. \qquad (43)$$

Из этих выражений находим математические ожидания и дисперсии коэффициентов a и b:

$$< a >= \bar{a} = \frac{1}{N}\sum_{i=1}^{N} y_i; \quad < b >= \bar{b} = \frac{1}{N\,L^2}\sum_{i=1}^{N}(x_i - \bar{x})\,y_i;$$

$$D_a = \frac{(N-2)}{(N-4)}\frac{\Delta^2}{(N-1)}; \quad D_b = \frac{(N-2)}{(N-4)}\frac{\Delta^2}{(N-1)\,L^2}$$

Каждый из них распределен по Стьюденту с $m = N - 2$ степенями свободы.

Регрессионный прогноз

Статистическая неопределенность коэффициентов регрессии a и b приводит к необходимости при прогнозировании проводить усреднение по этим параметрам с распределением (41):

$$dp(x) = \int_{a,b,D} dp(x\,/a,b,D)\,dp(a,b,D). \qquad (44)$$

Этот интеграл все того же гауссова типа, который после интегрирования по D приводит к распределению Стьюдента. Нужно расписать квадратичную форму в показателе экспоненты:

$$f = \Delta^2 + N(a-\bar{a})^2 + N\,L^2\,(b-\bar{b})^2 + (y-a-b\,(x-\bar{x}))^2.$$

Теперь нужно найти минимум этой формы по a и b:

$$\tilde{a} = \frac{\bar{a}\,(N\,L^2 + x(x-\bar{x})^2) + L^2(y-\bar{b}\,(x-\bar{x}))}{(N+1)\,L^2 + (x-\bar{x})^2}.$$

$$\tilde{b} = \frac{\bar{b}\,L^2\,(N+1) + (x-\bar{x})\,(y-\bar{a})}{(N+1)\,L^2 + (x-\bar{x})^2}.$$

При этом сама квадратичная форма оказывается равной

$$f = \frac{N\,L^2(y - \bar{a} - \bar{b}\,(x - \bar{x}))^2}{(N+1)\,L^2 + (x - \bar{x})^2} + \Delta^2.$$

Добавление новой точки увеличит степень D в знаменателе на $1/2$, однако интегрирование по a и b две половинки съедает, так что окончательно

$$dp(y(x)) \sim \frac{dy}{\left(1 + \frac{N\,L^2\,(y - \bar{a} - \bar{b}\,(x - \bar{x}))^2}{((N+1)L^2 + (x - \bar{x})^2)\Delta^2}\right)^{\frac{N-1}{2}}} \sim \frac{dS}{\left(1 + \frac{S^2}{m}\right)^{\frac{m+1}{2}}}, \quad (45)$$

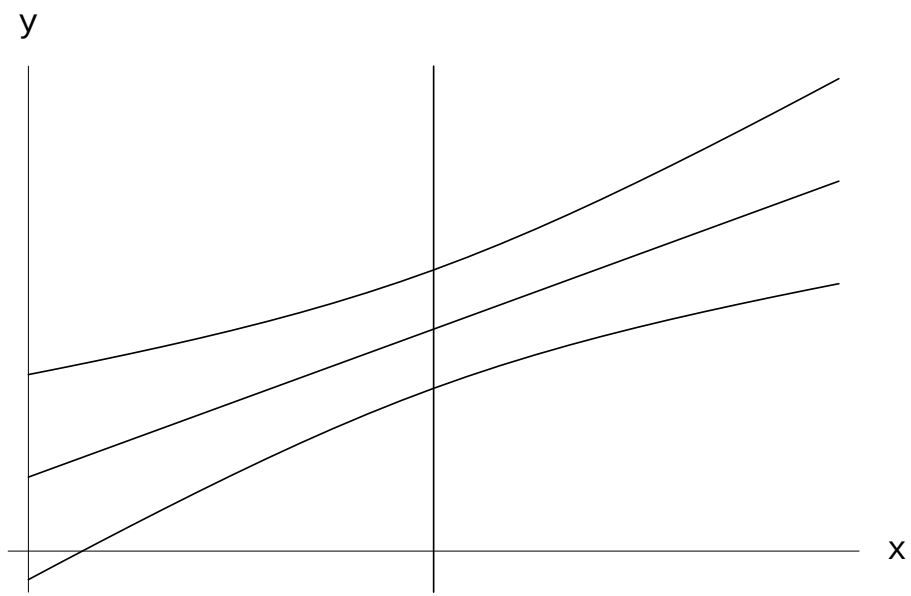
Это распределение Стьюдента с $m = N - 2$ степенями свободы. Сама прогнозируемая величина является суммой регулярной части – наиболее вероятной регрессионной функции – и случайной величины S, распределенной по Стьюденту:

$$y(x) = \bar{a} + \bar{b}\,(x - \bar{x}) + S_{N-2}\sqrt{\Delta^2\left(1 + \frac{1}{N} + \frac{(x - \bar{x})^2}{N\,L^2}\right)} \quad (46)$$

Математическое ожидание y лежит на регрессионной кривой $y = \bar{a} + \bar{b}\,(x - \bar{x})$, а дисперсия ее различна при различных x:

$$D_x = \frac{(N-2)}{(N-4)}\,\Delta^2\,\left(1 + \frac{1}{N} + \frac{(x-\bar{x})^2}{N\,L^2}\right) =$$

$$= \frac{(N-2)\,(N+1)}{(N-4)\,N}\,\Delta^2\,\left(1 + \frac{(x-\bar{x})^2}{(N+1)\,\Delta^2}\right) \qquad (47)$$



Дисперсия, а вместе с ней и доверительный интервал заданной надежности, расширяются от середины графика к концам. Минимум дисперсии в точке $x = \bar{x}$.