Сюжеты с кружка по эконометрике

Последнее обновление:

22 июля 2019 г.

Содержание

1 Пределы по вероятности

1.1 Разминка

Найдите:

1. $\text{plim}\bar{X}_n$
2. $\text{plim}\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$
3. Считая, что пары $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ независимы и одинаково распеределены, найдите: $\text{plim}\frac{\sum_{i=1}^{n}(x_i-\bar{x_i})(y_i-\bar{y})}{n-1}$

1.2 Задача про русскую народную скромность

Отчаянный исследователь Эйзенхорн пытается предсказать результаты студентов по эконометрике и строит следующую регрессию:

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

y_i - результат по эконометрике
x_i - количество съеденных бургеров

Для получения данных Эйзенхорн заставил студентов заполнить табличку. Получилось, что регрессор подчиняется следующему утверждению:

$$x_i^* = x_i + \alpha + \nu_i$$

α - Параметр русской народной скромности
ν_i - случайная величина, отражающая несовершенство памяти
$E(x_i) = \mu_x, \mathbb{E}(u_i) = 0, \mathbb{E}(\nu_i) = 0, Var(x_i) = \sigma_x^2, Var(u_i) = \sigma_u^2, Var(\nu_i) = \sigma_\nu^2$

Найдите:
$\text{plim}\hat{\beta}_2$
$\text{plim}\hat{\beta}_1$

1.3 Смещённость оценок

Дети Империума Аида и Саша каждый день покупают айфоны.

x_i - количество купленных айфонов

y_i - потраченная сумма денег

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Они оба записывают в списки количество купленных айфонов.

Аида записывает: $x_i^a = x_i + \nu_i$

Саша записывает: $x_i^s = x_i + \phi_i$

Четвёрки $(x_i, u_i, \nu_i, \phi_i)$ независимы. Величины внутри четвёрок также независимы.

$E(x_i) = \mu_x, \mathbb{E}(u_i) = 0, \mathbb{E}(\nu_i) = 0, E(\phi_i) = 0, Var(x_i) = \sigma_x^2, Var(u_i) = \sigma_u^2, Var(\nu_i) = \sigma_\nu^2, Var(\phi_i) = \sigma_\phi^2$

1. $\hat{y}_i = \hat{\beta}_1^a + \hat{\beta}_2^a x_i^a$
   Найдите $plim\hat{\beta}_2^a$
2. $\hat{y}_i = \hat{\beta}_1^s + \hat{\beta}_2^s x_i^s$
   Найдите $plim\hat{\beta}_2^s$
3. Придумайте оценку $\hat{\beta}_2$, у которой $plim\hat{\beta}_2 = \beta_2$

1.4

$y_i = \beta_1 + \beta_2 x_i + u_i$ $\mathbb{E}(u_i) = 0$, $Var(u_i) = \sigma^2$, $x_i = i$

Найдите $plim\hat{\beta}_2$

1.5

$y_i = \beta_1 + \beta_2 x_i + u_i$ $\mathbb{E}(u_i) = 0$, $Var(u_i) = \sigma^2$, $x_i = \begin{cases} 0, & \text{если } i = 1; \\ 1, & \text{если } i > 1; \end{cases}$

Найдите $plim\hat{\beta}_2$

1.6

$y_i = \beta_1 + \beta_2 x_i + u_i$ $\mathbb{E}(u_i) = 0$, $Var(u_i) = 2^i\sigma^2$, Чётные x_i равны нулю, нечётные x_i равны единице.

Найдите $plim\hat{\beta}_2$

1.7 Гетерогенный эффект воздействия и ошибка специффикации

Рассмотрим модель $y_i = \beta_i \cdot x_i + \alpha_i \cdot w_i + \epsilon_i$.

Здесь y_i – производительнось труда i-го работника; x_i – стаж труда i-го работника, $E(x_i) > 0$, $0 < Var(x_i) < \infty$; w_i - бинарная переменная, равная единице, если i-ый работник посетил курсы повышения кваллификации, и равная нулю в противном случае; α_i - изменение производительности труда i-ого работника в результате посещения курсов повышения кваллификации. Обратите внимание, что эффект различен для разных работников, то есть является гетерогенным. $E(\epsilon_i | x_i, \alpha_i, w_i) = 0$, а векторы $v_1, \ldots, v_n$ - независимы и одинаково распределены, где $v_i' = (x_i, w_i, \alpha_i, e_i)$.

Исследователя интересует *средний* эффект воздействия курсов повышения кваллификации α : $\alpha = E(\alpha_i)$. В качестве его оценки он использует МНК-оценку коэффициента при переменной w в регрессии: $\hat{y} = \hat{\beta}_i \cdot x_i + \hat{\alpha}_i \cdot w_i$.

Обратите внимание, что исследователь игнорирует гетерогенность, мотивируя тем, что его интересует только *средний* эффект воздействия.

Дополнительно известно, что в действительности эффект от посещения курсов повышения кваллификации зависит от опыта работника: $\alpha_i = \gamma \cdot x_i$, где $\gamma > 0$.

1. Будет ли оценка, полученная исследователем, состоятельной? Если нет, то можете ли вы определить направление ее асимптотического смещения?
2. В предыдущем пункте на кружке получилось доказать только что при $w_i = 1$: $\text{plim}_{n\to\infty} \hat{\alpha} = 0 < \gamma \cdot E(x_1)$. Выясните, можно ли подобрать w_i так, что $\text{plim}_{n\to\infty} \hat{\alpha} > \gamma \cdot E(x_1)$.

3. Пусть теперь известно, что x_i и w_i независимы. Ответьте на вопросы предыдущего пункта.

1.8

Привести пример, когда одновременно выполняется:

$$\lim_{n\to\infty} E(\hat{\theta}_n - \theta) = 0$$

$$\text{plim}_{n\to\infty}(\hat{\theta}_n - \theta) \neq 0$$

Также приведите пример, когда выполняется обратное соотношение.

1.9

Дано:

$$y_i = \beta_1 + \beta_2 \cdot x_i + u_i$$

$$v_i = \begin{pmatrix} u_i \\ x_i \end{pmatrix};\ E(v_i) = \begin{pmatrix} 0 \\ \mu_x \end{pmatrix};$$

$$Var(v_i) = \begin{pmatrix} \sigma_u^2 & c \\ c & \sigma_x^2 \end{pmatrix}$$

Найдите $\text{plim}_{n\to\infty} \hat{\beta}_2^{\text{МНК}}$

1.10 Решения

Разминка

1. $\text{plim}\bar{X}_n = E(X_i)$
2. $\text{plim}\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \text{plim}\frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1} \cdot \frac{n}{n} = \left(\text{plim}\frac{\sum_{i=1}^n X_i^2}{n} - \text{plim}\bar{X}^2\right) \cdot \text{plim}\frac{n}{n-1} = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = Var(X_1)$
3. $\text{plim}\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \text{plim}\frac{\sum_{i=1}^n (x_i - \bar{x}y_i)}{n-1} = \text{plim}\frac{\sum_{i=1}^n (x_i y_i - \bar{x}y_i)}{n-1} = \left(\text{plim}\frac{\sum_{i=1}^n x_i y_i}{n} - \mathbb{E}(x_i)\,\mathbb{E}(y_i)\right) \cdot \text{plim}\left(\frac{n}{n-1}\right) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) = (X, Y)$

Задача про русскую народную скромность

$\text{plim}\hat{\beta}_2 = \frac{\text{plim}\left(\frac{\sum_{i=1}^n (x_i^* - \bar{x}^*)(y_i - \bar{y})}{n-1}\right)}{\text{plim}\left(\frac{\sum_{i=1}^n (x_i^* - \bar{x}^*)^2}{n-1}\right)} = \frac{Cov(x_1^*, y_1)}{\text{Var}(x_1^*)} = \frac{Cov(x_1 + \alpha + \nu_1, \beta_1 + \beta_2 x_1 + \epsilon_1)}{\text{Var}(x_1^*)} = \frac{\beta_2 \text{Var}(x_1)}{\text{Var}(x_1^*)} = \frac{\beta_2 \text{Var}(x_1)}{\text{Var}(x_1 + \alpha + \nu_1)} = \frac{\beta_2 \text{Var}(x_1)}{\text{Var}(x_1) + \text{Var}(\nu_1)} = \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}$

Выводы:

1. Оценки занижаются, $\text{plim}\hat{\beta}_2 \neq \beta_2$
2. α не влияет на смещение
3. При $\sigma_\nu^2 = 0$ (все студенты идеально помнят свои x_i, и просто занижают из-за русской национальной скромности) $\Rightarrow \text{plim}\hat{\beta}_2 = \beta_2$

$\text{plim}\,\hat{\beta}_1 = \text{plim}(\bar{y} - \hat{\beta}_2\bar{x}) = \text{plim}\,\bar{y} - \text{plim}\,\hat{\beta}_2 \cdot \text{plim}\,\bar{x} = \text{plim}\left(\frac{\sum_{i=1}^n \beta_1 + \beta_2 x_i + \epsilon_i}{n}\right) - \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\mu_x = \text{plim}(\beta_1 + \beta_2\bar{x} + \bar{\epsilon}) - \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\mu_x = \beta_1 + \beta_2\mu_x - \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\mu_x = \beta_1 + \frac{\beta_2\mu_x(\sigma_x^2 + \sigma_v^2) - \beta\sigma_x^2\mu_x}{\sigma_x^2 + \sigma_\nu^2} = \beta_1 + \frac{\beta_2\mu_x\sigma_\nu^2}{\sigma_x^2 + \sigma_\nu^2}$

Смещённость оценок

1. $\operatorname{plim} \hat{\beta}_2^a = \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}$ где $\sigma_x^2 + \sigma_\nu^2 = \operatorname{Var}(x_1^a)$
2. $\operatorname{plim} \hat{\beta}_2^s = \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\phi^2}$ где $\sigma_x^2 + \sigma_\phi^2 = \operatorname{Var}(x_1^s)$
3. $\operatorname{plim} \frac{\sum_{i=1}^n (x_i^s - \bar{x}^s)(x_i^a - \bar{x}^a)}{n-1} = \sigma_x^2$

   $\operatorname{plim} \frac{\sum_{i=1}^n (x_i^a - \bar{x}^a)^2}{n-1} = \operatorname{Var}(x_1^a) = \sigma_x^2 + \sigma_\nu^2$

   $\operatorname{plim} \hat{\beta}_2^a = \operatorname{plim} \frac{\sum_{i=1}^n (x_i^a - \bar{x}^a)(y_i - \bar{x}^a)}{\sum_{i=1}^n (x_i^a - \bar{x}^a)^2} = \beta_2 \frac{\beta_2 \sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}$

   $\hat{\beta}_2 = \hat{\beta}_2^a \frac{\sum_{i=1}^n (x_i^a - \bar{x}^a)^2/(n-1)}{\sum_{i=1}^n (x_i^a - \bar{x}^a)(x_i^s - \bar{x}^s)/(n-1)} = \frac{\sum_{i=1}^n (x_i^a - \bar{x}^a)(y_i - \bar{y})}{\sum_{i=1}^n (x_i^a - \bar{x}^a)(x_I^s - \bar{x}^s)} = \beta_2^*$

   $\operatorname{plim} \beta_2^* = \beta_2$

   Другая оценка:

   $\beta_2^{**} = \frac{\sum_{i=1}^n (x_i^s - \bar{x}^s)(y_i - \bar{y})}{\sum_{i=1}^n (x_i^a - \bar{x}^a)(x_I^s - \bar{x}^s)}$

   $\operatorname{plim} \beta_2^{**} = \beta_2$

   Ещё одна оценка:

   $\beta_2^{***} = \beta_2^{**} \frac{n}{n-1}$

1.4

$\operatorname{plim} \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Подставим в это выражение $y_i = \beta_1 + \beta_2 x_i + u_i$. Получим сумму в из трёх слагаемых. Рассмотрим каждое в отдельности.

1. $\frac{\sum_{i=1}^n (x_i - \bar{x})\beta_1}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$
2. $\frac{\sum_{i=1}^n (x_i - \bar{x})\beta_2 x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_2 \frac{\sum_{i=1}^n x_i^2 - x_i \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_2$
3. Подсчёт третьей части затруднён так как есть случайная составляющая. Законом Больших чисел мы также не можему воспользоваться. Однако есть другой способ.

   Пусть $\mathbb{E}(R_n) \to \mu, \operatorname{Var}(R_n) \to 0$

   Тогда согласно неравенству Чебышёва $P(|R_n - \mu| > \epsilon) \to 0$. То есть, если мы докажем что дисперсия этого слагаемого стремится к нулю, то предел по вероятности будет равен математическому ожиданию.

   Обозначим третье слагаемое как C_n

   $C_n = \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

   $\mathbb{E}(C_n) = 0$

   $\operatorname{Var}(C_n) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_u^2}{\sum_{i=1}^n ((x_i - \bar{x})^2)^2} = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma_u^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sigma_u^2}{\sum_{i=1}^n i^2 - n\left(\frac{\sum_{i=1}^n i}{n}\right)^2}$

   Далее было объяснение почему эта штука в знаменателе равна нулю. Нужен график, позже отрисую.

   Следовательно, $\operatorname{plim} \hat{\beta}_2 = \beta_2$

1.5

Аналогично считаем, что $\operatorname{plim} \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \operatorname{plim} A_n + \operatorname{plim} B_n + \operatorname{plim} C_n$

Помня, что $x_1 - \bar{x} = -\frac{n-1}{n}$, и что $x_2 - \bar{x} = \frac{1}{n}$ вычислим $\sum_{i=1}^n (x_i - \bar{x})^2$:

$\left(-\frac{n-1}{n}\right)^2 + \sum_{i=1}^n \frac{1}{n^2} = \left(-\frac{n-1}{n}\right)^2 + \frac{n-1}{n} = \frac{n-1}{n^2}(n - 1 + 1) = \frac{n-1}{n}$

Теперь можно вычислить $\operatorname{plim} C_n$

$\operatorname{plim} C_n = plim \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \operatorname{plim} \frac{-\frac{n-1}{n}u_1}{\frac{n-1}{n}} + \operatorname{plim} \frac{\frac{1}{n}\sum_{i=1}^{n}u_i}{\frac{n-1}{n}} = -\operatorname{plim} u_1 + 0 = -u_1$

Таким образом, $\operatorname{plim} \hat{\beta}_2 = \beta_2 - u_1$

2 Двухшаговый МНК

2.1 Бытовуха

Шаг 1: Оценим модель

$$\hat{x}_i = \hat{\alpha}_1 + \hat{\alpha}_1 z_i$$

Шаг 2: Оценим модель

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{x}_i$$

$$v_i = \begin{pmatrix} u_i \\ x_i \\ z_i \end{pmatrix};\ E(v_i) = \begin{pmatrix} 0 \\ \mu_x \\ \mu_z \end{pmatrix};$$

$$Var(v_i) = \begin{pmatrix} \sigma_u^2 & c & 0 \\ c & \sigma_x^2 & d \\ 0 & d & \sigma_z^2 \end{pmatrix}$$

Найдите $\operatorname{plim}_{n\to\infty} \hat{\beta}_2, \operatorname{plim}_{n\to\infty} \hat{\beta}_1$

Хинт: Полученная оценка $\hat{\beta}_2$ будет в итоге оценкой метода инструментальных переменных.

2.2 Бытовуха, версия хардкор

Рассматривается модель:

$$\theta_1 + \theta_2 x_i + u_i$$

где x_i стохастический эндогенный регрессор:

$$x_i = \alpha_0 + \alpha_1 p_i + \alpha_2 q_i + \epsilon_i$$

$$v_i = \begin{pmatrix} u_i \\ x_i \\ p_i \\ q_i \\ \epsilon_i \end{pmatrix};\ E(v_i) = \begin{pmatrix} 0 \\ \mu_x \\ \mu_p \\ \mu_q \\ 0 \end{pmatrix};$$

$$Var(v_i) = \begin{pmatrix} \sigma_u^2 & C_{xu} & 0 & 0 & 0 \\ C_{xu} & \sigma_x^2 & C_{px} & C_{qx} & 0 \\ 0 & C_{px} & \sigma_p^2 & C_{pq} & 0 \\ 0 & C_{qx} & C_{pq} & \sigma_q^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\epsilon^2 \end{pmatrix}$$

Шаг 1: Оценим модель

$$\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 p_i + \hat{\alpha}_2 q_i$$

Шаг 2: Оценим модель

$$\hat{y}_i = \hat{\theta}_1 + \hat{\theta}_2 \hat{x}_i$$

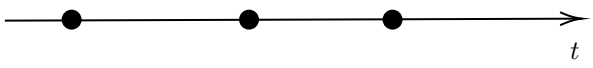
1. Найдите $\operatorname{plim}_{n\to\infty} \hat{\theta}_2, \operatorname{plim}_{n\to\infty} \hat{\theta}_1$ и проверьте состоятельность этих оценок.
2. Пусть ваша выборка состоит из 1000 наблюдений, причем вы располагаете данными о средних выборочных значениях переменных:
   $\bar{X} = \bar{Y} = \bar{P} = 0,\ \bar{Q} = \bar{PQ} = \bar{XQ} = \bar{P^2} = \bar{YQ} = 1,\ \bar{Q^2} = 1.5,\ \bar{XP} = \bar{YP} = 2$
   Вычислите состоятельную оценку параметра θ_2 из предыдущего пункта.

3 Идеи распределений

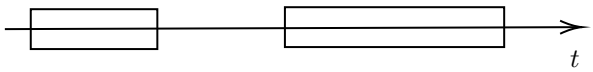
3.1 Экспоненциальное распределение и распределение Пуассона через АПП

Аксиомы Пуассоновского потока (АПП):

1. Время непрерывно.
2. На оси времени происходят точечные происшествия.

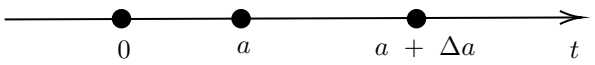


   «Точечные» означает, что происшествие не длится, скажем, 2 секунды, а случается «в точке» на оси времени.
3. Количества происшествий на непересекающихся интервалах независимы.



4. Закон распределения количества происшествий стабилен во времени. Это означает, что на двух временных интервалах одинаковой длины количества происшествий распределены одинаково.
5. На малом интервале времени Δt:
   - вероятность двух и более происшествий мала по сравнению с Δt:
     $$\mathbb{P}(\text{2 и более происшетсвий}) = o(\Delta t).$$
   - вероятность ровно одного происшествия пропорциональна Δt с точностью до o-малых:
     $$\mathbb{P}(\text{ровно 1 происшествие}) = \lambda \Delta t + o(\Delta t).$$

Посмотрим, как из АПП можно вывести экспоненциальное распределение и распределение Пуассона. Рассмотрим следующий участок на оси времени:



Попробуем найти вероятность того, что за промежуток времени $[0; a]$ произойдёт ровно 0 происшествий:

$$\mathbb{P}(0 \text{ происшествий за } [0; a]).$$

Для этого рассмотрим вероятность того, что ровно 0 происшествий произойдёт за промежуток времени $[0; a + \Delta a]$:

$$\mathbb{P}(0 \text{ происшествий за } [0; a + \Delta a]).$$

Так как $[0;a]$ и $[a;a+\Delta a]$ – непересекающиеся интервалы, то по аксиоме 3 количество происшествий на них независимы, а значит совместная вероятность раскладывается в произведение:

$$\mathbb{P}(0 \text{ происшествий за } [0;a+\Delta a]) = \mathbb{P}(0 \text{ происшествий за } [0;a]) \times \mathbb{P}(0 \text{ происшествий за } [a;a+\Delta a]).$$

А $\mathbb{P}(0$ происшествий за $[a;a+\Delta a])$ можно рассчитать по аксиоме 5. На этом интервале может произойти ноль, одно или два и боле событий. Так как нас интересует первый вариант, вычтем из единицы вероятности второго и третьего вариантов:

$$\mathbb{P}(0 \text{ происшествий за } [a;a+\Delta a]) = 1 - \mathbb{P}(1 \text{ происшествие за } [a;a+\Delta a]) - \\ - \mathbb{P}(2 \text{ и более происшествий за } [a;a+\Delta a]).$$

А эти вероятности – ровно то, что стоит в аксиоме 5, только в нашем случае интервал Δt – это Δa. Таким образом, получаем:

$$\begin{aligned}\mathbb{P}(0 \text{ происшествий за } [a;a+\Delta a]) &= 1 - \mathbb{P}(1 \text{ происшествие за } [a;a+\Delta a]) - \\ &- \mathbb{P}(2 \text{ и более происшествий за } [a;a+\Delta a]) = \\ &= 1 - \lambda\Delta a - o(\Delta a) - o(\Delta a) = 1 - \lambda\Delta a - o(\Delta a) \text{ (по свойствам } o\text{-малых)}.\end{aligned}$$

Вернёмся к начальному выражению:

$$\mathbb{P}(0 \text{ происшествий за } [0;a+\Delta a]) = \mathbb{P}(0 \text{ происшествий за } [0;a]) \times (1 - \lambda\Delta a - o(\Delta a)).$$

Раскроем скобки:

$$\mathbb{P}(0 \text{ происшествий за } [0;a+\Delta a]) - \mathbb{P}(0 \text{ происшествий за } [0;a]) = \mathbb{P}(0 \text{ происшествий за } [0;a])(-\lambda\Delta a) + o(\Delta a),$$

так как $\mathbb{P}(0$ происшествий за $[0;a]) \times o(\Delta a) = o(\Delta a)$ (вероятность лежит в пределах от нуля до единицы, то есть мала).

Поделим обе части на Δa:

$$\frac{\mathbb{P}(0 \text{ происшествий за } [0;a+\Delta a]) - \mathbb{P}(0 \text{ происшествий за } [0;a])}{\Delta a} = \mathbb{P}(0 \text{ происшествий за } [0;a])(-\lambda) + \frac{o(\Delta a)}{\Delta a}.$$

А теперь возьмём предел правой и левой части при $\Delta a \to 0$. Заметим, что слева стоит ни что иное как производная $\mathbb{P}(0$ происшествий за $[0;a])$ (по определению производной), если мы рассматриваем эту вероятность как функцию. Также по определению:

$$\lim_{\Delta a \to 0} \frac{o(\Delta a)}{\Delta a} = 0.$$

Обозначим вероятность $\mathbb{P}(0$ происшествий за $[0;a])$ как $z(a)$, то есть явно скажем, что это некоторая функция:

$$\mathbb{P}(0 \text{ происшествий за } [0;a]) \equiv z(a).$$

Тогда выражение выше можно записать как:

$$z'_a = -\lambda z(a) + 0.$$

Общее решение этого дифференциального уравнения:

$$z(a) = Ce^{-\lambda a}.$$

Заметим, что $z(a)$ уже похожа на функции распределения экспоненциального распределения и распределения Пуассона. Восстановим константу C. Логично предположить, что $z(0) = 1$, то есть вероятность того, что за промежуток времени $[0;0]$ произойдёт 0 происшествий, равна единице. Отсюда:

$$1 = Ce^0 \Rightarrow C = 1.$$

Таким образом:

$$z(a) = e^{-\lambda a}.$$

Что такое $z(a)$? По нашей записи это вероятность того, что в промежуток времени $[0; a]$ произойдёт 0 происшествий.

Распределение Пуассона как раз моделирует случайную величину, которая показывает число событий, произошедших за фиксированный промежуток времени. Предположим, что $a = 1$, то есть $z(a) = e^{-\lambda}$ (промежуток времени фиксирован от 0 до 1). Таким образом, $z(a)$ – это вероятность того, что случайная величина, имеющая Пуассоновское распределение, примет значение 0 на промежутке времени $[0; 1]$. Покажем это:

$$X \sim pois(\lambda),$$

$$\mathbb{P}(X = 0) = \frac{e^{-\lambda} \times \lambda^0}{0!} = e^{-\lambda}.$$

Заметим, что $z(a)$ можно интерпретировать по-другому: если это вероятность того, что за промежуток времени $[0; a]$ произойдёт 0 происшествий, то это же вероятность того, что первое происшествие произойдёт после точки a на временной оси. Пусть случайная величина Y показывает время до первого происшествия. Тогда:

$$z(a) = \mathbb{P}(Y \geq a).$$

Перепишем данное выражение в терминах функции распределения:

$$P(Y \leq a) = 1 - z(a) = 1 - e^{-\lambda a}.$$

Получили функцию распределения экспоненциального распределения! Экспоненциальное распределение моделирует время между двумя происшествиями. У нас получилось, что $Y \sim exp(\lambda)$ и Y – время до первого происшествия, то есть экспоненциальное распределение также показывает время от начала отсчёта до первого происшествия. Ввиду свойства отсутствия памяти, два этих определения эквивалентны (так как после наступления нового происшествия, точка отсчёта смещается в точку этого происшествия).

Покажем, что распределение Пуассона можно вывести для любого числа происшествий. Например, проверим, что наши рассуждения верны и для $\mathbb{P}$(1 происшествие за $[0; a]$). Будем рассматривать фиксированный промежуток времени $[0; 1]$ (то есть при $a = 1$). Ожидаемый результат:

$$X \sim pois(\lambda).$$

$$\mathbb{P}(X = 1) = e^{-\lambda}\lambda.$$

Для краткости обозначим искомую вероятность за $u(a)$:

$$\mathbb{P}(1 \text{ происшествие за } [0; a]) \equiv u(a).$$

Применим ту же схему. Вероятность того, что произошло ровно 1 происшествие за промежуток времени $[0; a + \Delta a]$ раскладывается в сумму вероятностей, что ровно 1 происшествие произошло либо за $[0; a]$, либо за $[a; a + \Delta a]$. Первая вероятность – это произведение вероятностей, что за $[0; a]$ произошло ровно одно происшествие, а за $[a + \Delta a]$ произошло ноль происшествий. Вторая вероятность – это произведение вероятностей, что за $[0; a]$ произошло ноль происшествий, а за $[a + \Delta a]$ произошло ровно одно происшествие. Запишем эти рассуждения в наших обозначениях:

$$u(a + \Delta a) = u(a)(1 - \lambda\Delta a - o(\Delta a)) + z(a)(\lambda\Delta a + o(\Delta a)).$$

Снова раскроем скобки и поделим обе части на Δa:

$$\frac{u(a + \Delta a) - u(a)}{\Delta a} = \lambda z(a) - \lambda u(a) + \frac{o(\Delta a)}{\Delta a}.$$

В пределе при $\Delta a \to 0$ получаем:

$$u'_a = \lambda e^{-\lambda a} - \lambda u(a).$$

Условие $u(0) = 0$ опять же выполняется. Если решить данное дифференциальное уравнение, получим:

$$u(a) = \frac{e^{-\lambda a}\lambda a}{1!}.$$

При $a = 1$ получаем:

$$u(a) = e^{-\lambda}\lambda.$$

Получили, что ожидали. Можно проверить результаты и для большего числа происшествий.

Важный факт, который получаем из двоякой интерпретации $z(a)$: если предполагаем, что время между двумя происшествиями распределено экспоненциально, то их количество распределено по Пуассону. И наоборот, если считаем, что количество происшествий распределено по Пуассону, то время между двумя происшествиями распределено экспоненциально.

3.2 Гамма-распределение: $\gamma(k, \lambda)$

Гамма распределение будем рассматривать на примере ловли червячков птичкой. По смыслу, случайная величина $s \sim \gamma(k, \lambda)$ показывает суммарное время на ловлю k червячков, если матожидание количества червячков за единицу времени равно λ. Раз это суммарное время, то гамма-распределение – это ещё и сумма случайных величин, распределённых экспоненциально с параметром λ.

Выведем общую формулу гамма-распределения по индукции, начав со случая $\gamma(3, \lambda)$. Пусть $Y_1, Y_2, Y_3 \sim exp(\lambda)$. Тогда:

$$f(y_1, y_2, y_3)dy_1 \wedge dy_2 \wedge dy_3 = \lambda e^{-\lambda y_1}\lambda e^{-\lambda y_2}\lambda e^{-\lambda y_3}dy_1 \wedge dy_2 \wedge dy_3.$$

Обозначим: $S_3 = Y_1 + Y_2 + Y_3$, $S_2 = Y_1 + Y_2$, $S_1 = Y_1$; $R_2 = S_1/S_2$, $R_3 = S_2/S_3$, $R_4 = S_3/S_4$. По смыслу, R_i – какая доля времени потрачена на поимку $(i-1)$ червячка, если известно время на поимку i червячков. Перейдём от Y_1, Y_2, Y_3 к R_2, R_3, S_3.

$$Y_1 = S_1 = R_2 \times R_3 \times S_3.$$
$$Y_2 = S_2 - S_1 = (1 - R_2) \times R_3 \times S_3.$$
$$Y_3 = S_3 - S_2 = (1 - R_3) \times S_3.$$

Далее нам нужно подставить эти выражения в дифференциальную форму выше. Прежде чем сделать это, выведем часть «с птичками».

$$\begin{aligned} dy_1 \wedge dy_2 = d(r_2r_3s_3) \wedge d((1-r_2)r_3s_3) = (d[r_2]r_3s_3 + r_2d[r_3s_3]) \wedge (-d[r_2]r_3s_3 + (1-r_2)d[r_3s_3]) = \\ = (1-r_2)r_3s_3d[r_2] \wedge d[r_3s_3] + r_2r_3s_3d[r_2] \wedge d[r_3s_3] = r_3s_3d[r_2] \wedge d[r_3s_3]. \end{aligned}$$

$$\begin{aligned} dy_1 \wedge dy_2 \wedge dy_3 = \\ = r_3s_3d[r_2] \wedge d[r_3s_3] \wedge (-d[r_3]s_3 + (1-r_3)ds_3) = \\ = r_3s_3d[r_2] \wedge (d[r_3]s_3 + r_3d[s_3]) \wedge (-d[r_3]s_3 + (1-r_3)d[s_3]) = \\ = r_3s_3d[r_2] \wedge (s_3(1-r_3)d[r_3]d[s_3] + s_3r_3d[r_3]d[s_3]) = \\ = r_3s_3^2dr_2dr_3ds_3. \end{aligned}$$

Теперь подставим всё, что нашли, в дифференциальную форму:

$$f(y_1, y_2, y_3)dy_1 \wedge dy_2 \wedge dy_3 = \lambda^3 e^{-\lambda s_3}r_3s_3^2dr_2dr_3ds_3.$$

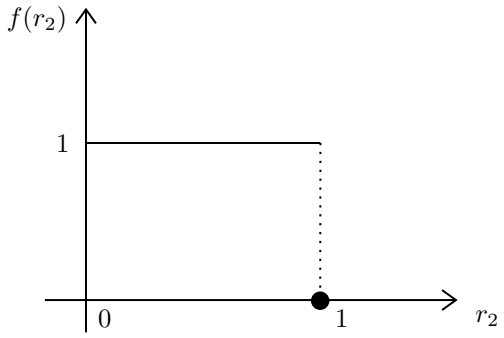
Отсюда:

$$f(r_2, r_3, s_3) = \lambda^3 e^{-\lambda s_3}r_3s_3^2.$$

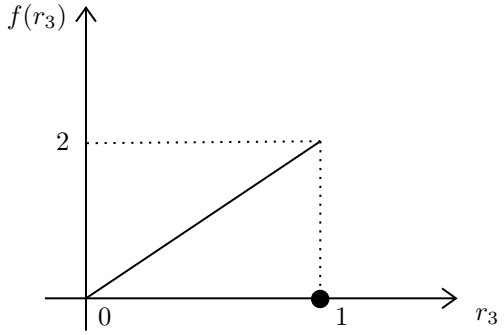
Замечаем, что совместная функция плотности раскладывается на произведение индивидуальных функций плотности:

$$f(r_2, r_3, s_3) = f(r_2)f(r_3)f(s_3) = [C_1][C_2r_3][C_3e^{-\lambda s_3}s_3^2].$$

Восстановим константы. Понятно, что $r_2 \sim U[0, 1]$, (от 0 до 1, так как это доля времени) – а значит, $C_1 = 1$:



Далее понимаем, что r_3 тоже распределено от 0 до 1, так как это доля времени. Посмотрим на функцию плотности r_3:



Чтобы интеграл под функцией плотности равнялся 1, C_2 должна быть равна 2. Выходит, что $C_3 = \frac{\lambda^3}{3}$.

На самом деле, константы нам не так важны. Обобщим, что получили про гамма-распределение. Если S_k – суммарное время на ловлю k червячков, если ожидаемое количество пойманных червячков за единицу времени равно λ, то $S_k \sim \gamma(k, \lambda)$ и $f(s) = Ce^{-\lambda s}s^{k-1}$.

Покажем, что это верно и для следующего шага индукции. Перейдём от r_2, r_3, s_3 к r_2, r_3, r_4, s_4. Тогда $s_3 = s_4 \times r_4$. Вернёмся к дифференциальной форме:

$$1 \times 2r_3 \times \frac{\lambda^3}{2}e^{-\lambda s_3}s_3^2 dr_2 dr_3 ds_3.$$

При добавлении y_4 форма домножится на $\wedge\lambda e^{-\lambda y_4}dy_4$. Понятно, что $y_4 = (1 - r_4)s_4$.

Рассчитаем новую часть «с птичками».

$$ds_3 \wedge dy_4 = d(s_4 r_4) \wedge d((1 - r_4)s_4) = (d[s_4]r_4 + s_4 d[r_4]) \wedge (-d[r_4]s_4 + (1 - r_4)d[s_4]) = (r_4 s_4 + s_4(1 - r_4))d[r_4] \wedge d[s_4] = \\ = s_4 dr_4 ds_4.$$

Понятно, что $s_3^2 = (s_4 r_4)^2$. Тогда новая дифференциальная форма имеет вид (подставляем всё, что получили):

$$1 \times 2r_3 \times 3r_4^2 \times \frac{\lambda^4}{3!}s_4^3 e^{-\lambda s_4}dr_2 dr_3 dr_4 ds_4.$$

Совпадает с нашей формулой выше с точностью до константы. Аналогично можно сделать следующий шаг индукции для $ds_4 \wedge dy_5$.

Так как гамма-распределение – это сумма экспоненциальных, легко вывести матожидание и дисперсию гамма-распределения:

$$\mathbb{E}(\gamma) = k \times \frac{1}{\lambda}.$$

$$Var(\gamma) = Var(Y_1 + \cdots + Y_k) = kVar(Y_1) = k \times \frac{1}{\lambda^2}.$$

Понятно, что при $k = 1$ гамма-распределение – это экспоненциальное распределение.

Полное общее определение: $S_k \sim \gamma(k, \lambda)$, где S_k – суммарное время на ловлю k червячков, λ – ожидаемое количество пойманных червячков за единицу времени. Тогда:

$$f(s_k) = \frac{\lambda^k}{(k-1)!} e^{-\lambda s_k} s_k^{k-1}.$$
$$\mathbb{E}(S_k) = \frac{k}{\lambda}.$$
$$Var(S_k) = \frac{k}{\lambda^2}.$$

3.3 Бета-распределение: $\beta(a, b)$

По смыслу: ловим $(a+b)$ червячков, a отдаём, b оставляем себе. Тогда, если R – доля времени на поимку a червячков, то то $R \sim \beta(a, b)$.

Пусть Z_1 – время на поимку a червячков, а Z_2 – доля на поимку b червячков. Тогда $Z_1 \sim \gamma(a, \lambda)$, $Z_2 \sim \gamma(b, \lambda)$. Дифференциальная форма:

$$f(z_1, z_2) dz_1 \wedge dz_2 = \frac{\lambda^a}{(a-1)!} e^{-\lambda z_1} z_1^{a-1} \times \frac{\lambda^b}{(b-1)!} e^{-\lambda z_2} z_2^{b-1} dz_1 \wedge dz_2.$$

Перейдём к нашим обозначениям r и s. Понятно, что $Z_1 = Y_1 + \cdots + Y_a$, $Z_2 = Y_{a+1} + \ldots$ Тогда обозначим:

$$s = Z_1 + Z_2.$$
$$r = \frac{Z_1}{Z_1 + Z_2}.$$

Тогда $Z_1 = rs$, $Z_2 = (1-r)s$. Часть «с птичками»:

$$d(rs) \wedge d((1-r)s) = s dr \wedge ds \text{ (считали выше).}$$

Подставляем всё в дифференциальную форму:

$$\frac{\lambda^{a+b}}{(a-1)!(b-1)!} e^{-\lambda s} r^{a-1} (1-r)^{b-1} s^{(a-1)+(b-1)+1} dr \wedge ds.$$

Перегруппируем:

$$\frac{\lambda^{a+b}}{(a+b-1)!} s^{a+b-1} e^{-\lambda s} \times \frac{(a+b-1)!}{(a-1)!(b-1)!} r^{a-1} (1-r)^{b-1} dr \wedge ds.$$

Замечаем, что до знака $\times$ стоит функция плотности гамма-распределения $f(s)$ с параметрами $k = a+b$ и λ. А после этого знака – $f(r)$, и по нашему обозначению r, это и есть функция плотности бета-распределения.

Выпишем отдельно: если R – доля времени на поимку a червячков, а всего ловим $(a+b)$ червячков, то $R \sim \beta(a, b)$ и:

$$f(r) = \frac{(a+b-1)!}{(a-1)!(b-1)!} r^{a-1} (1-r)^{b-1}.$$
$$\mathbb{E}(R) = \frac{a}{a+b}.$$

3.4 Нормальное распределение (через Гаусса)

Предпосылки:

1. Есть истинная величина μ.
2. $y_i = \mu + u_i$, u_i – независимы и симметричны около 0 (то есть свидетели равновероятно завышают и занижают показания).

3. $\bar{y} = \hat{\mu}_{ML}\ \forall y_1 \dots y_{n+1}$.

4? $f(u)$ – дифференцируемая.

Тогда $u_i \sim \mathbb{N}(0, \sigma^2)$.

Выпишем правдоподобие:

$$L = f(y_1 - \mu)f(y_2 - \mu)\dots f(y_{n+1} - \mu).$$

$$\ell = \sum_{i=1}^{n+1} \ln(f(y_i - \mu)).$$

$$\ell'_\mu = -\sum_{i=1}^{n+1} \frac{f'(y_i - \mu)}{f(y_i - \mu)}.$$

По предпосылке 3:

$$\sum_{i=1}^{n+1} \frac{f'(y_i - \bar{y})}{f(y_i - \bar{y})} = 0\ \forall y_1 \dots y_{n+1}.$$

Так как выполняется для любых y_i, возьмём конкретные показания. Для них должно выполняться:

$$\begin{aligned} y_1 - \bar{y} &= a, \\ y_2 - \bar{y} &= a, \\ &\dots \\ y_n - \bar{y} &= a, \\ y_{n+1} - \bar{y} &= -na. \end{aligned}$$

Получаем:

$$n\frac{f'(a)}{f(a)} + \frac{f'(-na)}{f(-na)} = 0\ \forall n, a.$$

Так как f – симметричная дифференцируемая:

$$\frac{f'(na)}{f(na)} = n\frac{f'(a)}{f(a)}.$$

Обозначим: $h(x) = \dfrac{f'(x)}{f(x)}$. Получаем, что $h(na) = nh(a)\ \forall n, a$. Это означает, что $h(x) = kx$. Получаем дифференциальное уравнение:

$$\frac{f'(x)}{f(x)} = kx.$$

Решение:

$$f(x) = C_1 e^{\frac{kx^2}{2}}.$$

3.5 Распределения через максимальную энтропию

Отступление: про энтропию.

1. $\mathbb{E}(y) = 10$, $\sigma = 1$, $H(y) \to \max \Rightarrow y \sim N(0{,}1)$.
2. $y \geq 0$, $\mathbb{E}(y) = 5$, $H(y) \to \max \Rightarrow y \sim Exp(0)$.
3. $y \in [a, b]$, $H(y) \to \max \Rightarrow y \sim U[a, b]$.

Интересный вопрос: что больше: $D_{KL}(N(0{,}1)||U[0, 1])$ или $D_{KL}(U[0{,}1]||N(0, 1)$? Посмотрим по смыслу: в перовом случае мы выбираем число из нормального распределения, а задаём вопросы про числа из равномерного распределения. Значит, существует положительная вероятность, что мы выберем такое число из нормального распределения, которое не накрывается равномерным, то есть, мы можем никогда не угадать это число. Это означает, что $D_{KL}(N||U) = +\infty$. Во втором же случае наоборот: загадали из равномерного, а спрашиваем про нормальное, то есть рано или поздно число будет угадано. Это означает, что $D_{KL}(U||N) =$ большое, но конечное число. То есть $D_{KL}(U||N) < D_{KL}(N||U)$.

3.5.1 Нормальное распределение (через максимальную энтропию)

Пусть $X \sim N(\mu, \sigma^2)$, $\mathbb{E}(y) = \mu$, $Var(y) = \sigma^2$. Тогда $H(x) \geq H(y)$. Покажем это.

$$H(X) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \log_{\frac{1}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx.$$

Для краткости:

$$H(X) = \int f(x) \log_{\frac{1}{2}} f(x) dx.$$

Для удобства перейдём в наты. Напоминание:

$$\log_{\frac{1}{2}} x \times \log_e \frac{1}{2} = \ln x.$$

$$\ln \frac{1}{2} < 0.$$

Тогда:

$$H(X) = -\int f(x) \ln f(x) dx.$$

Раскроем вторые скобки:

$$H(X) = -\int f(x) \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right) dx.$$

Знаем, что:

$$Var(X) = \mathbb{E}(X-\mu)^2 = \int f(x)(x-\mu)^2 dx.$$

А это как раз второй член в скобках. Получается, что числитель при взятии интеграла обратится в диспресию X и сократится со знаменателем. Получаем:

$$H(X) = -\left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \right) = \frac{1}{2} \left(1 + \ln\left(2\pi\sigma^2\right) \right).$$

Пусть $Y \sim q$. Тогда:

$$CE(Y||X) = -\int q(x) \left(\ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right) dx = \frac{1}{2} \left(1 + \ln\left(2\pi\sigma^2\right) \right).$$

Получили, что:

$$H(X) = CE(Y||X) \geq H(Y) \text{ (доказано ниже).}$$

3.5.2 Экспоненциальное распределение (через максимальную энтропию)

Докажем, что если $\mathbb{E}(X) = 1/\lambda$ и $X \geq 0$, то $H(X)$ максимальна при $X \sim exp(\frac{1}{\lambda})$.

Аналогично нормальному распределению:

$$H(X) = -\int_0^{+\infty} f(x)(\ln\lambda - \lambda x) dx = -(\ln\lambda - 1) = 1 - \ln\lambda.$$

$$CE(Y||X) = -\int_0^{+\infty} q(x)(\ln\lambda - \lambda x) dx = -(\ln\lambda - 1) = 1 - \ln\lambda.$$

$$H(X) = CE(Y||X) \geq H(Y).$$

4 Энтропия

Рассмотрим игру. Пусть мы загадываем один из четырёх вариантов со следующими вероятностями:

Y	Пуассон	Фишер	Коши	Арбуз
Prob.	1/2	1/4	1/8	1/8

Нужно найти, какое среднее число бинарных вопросов нужно задать, чтобы угадать Y, используя *оптимальную стратегию.*

Логично, что при данном распределении вероятностей сначала уместно задать вопрос: «Это Пуассон?», потому что Пуассон загадывается наиболее вероятно. Если это Пуассон, ты игра оканчивается, если нет – то следующим вопросом узнаём, Фишер ли это. Если же это и не Фишер, то можем спросить либо про Коши, либо про арбуз, и после этого точно завершить игру. Посмотрим, какое минимальное количество вопросов нужно задать в лучшем случае, чтобы отгадать каждый вариант, используя оптимальную стратегию:

Y	Prob.	Q
Пуассон	1/2	1
Фишер	1/4	2
Коши	1/8	3
Арбуз	1/8	3

Заметим, что $Q = \log_{\frac{1}{2}} \text{Prob}$. Найдём матожидание Q:

$$\mathbb{E}(Q) = \sum \mathbb{P}(Y = y_i) \times \log_{\frac{1}{2}} \mathbb{P}(Y = y_i).$$

Это матожидание и есть энтропия. По смыслу, энтропия показывает среднее число бинарных вопросов, которые нужно задать при использовании оптимальной стратегии, чтобы завершить игру. Обозначение: $H(y)$, $H(p)$.

Теперь рассмотрим кросс-энтропию (из одного распределения в другое). Для этого вернёмся к нашей игре и будем считать, что вероятности, которые мы рассматривали до этого, – это истинное распределение вероятностей A. Также считаем, что это истинное распределение нам неизвестно, и свою стратегию отгадывания мы будем строить по другому распределению, B:

Y	Пуассон	Фишер	Коши	Арбуз
A	1/2	1/4	1/8	1/8
B	1/4	1/2	1/8	1/8

Так как мы строим стратегию по распределению B, то сначала уместно задать вопрос по Фишера и только затем про Пуассона. Соответственно, минимальное количество задаваемых вопросов так же изменится:

Y	Prob. (по A)	Q (по B)
Фишер	1/4	1
Пуассон	1/2	2
Коши	1/8	3
Арбуз	1/8	3

Таким образом, изменится и матожидание Q, то есть энтропия. Теперь это матожидание называется кросс-энтропией из A в B. По смыслу: кросс-энтропия показывает среднее число бинарных вопросов, которое, как мы считаем в соответствии с распределением B, нужно задать при использовании оптимальной стратегии, чтобы завершить игру, истинное распределение вероятностей которой – A. Обозначение: $CE(a||b)$, $Ha(b)$. Понятно, что $H(a) = CE(a||a)$. То есть:

$$CE(A||B) = \sum \mathbb{P}_A(Y = y_i) \times \log_{\frac{1}{2}} \mathbb{P}_B(Y = y_i).$$

Рассчитаем энтропию и кросс-энтропию для нашей игры.

$$H(A) = \frac{1}{2} \times 1 + \frac{1}{3} \times 2 + 2 \times 3 \times \frac{1}{8} = 1\frac{3}{4}.$$

$$CE(A||B) = \frac{1}{4} \times 1 + \frac{1}{2} \times 2 + 2 \times 3 \times \frac{1}{8} = 2.$$

Получили, что $CE(a||b) > H(a)$. Совпадение ли это? Прежде чем выяснить это, введём следующую величину.

Дивергенция Куйбака-Лейблера показывает, насколько в среднем больше вопросов потребуется задать, чем могли бы при использовании оптимальной стратегии. По смыслу, это некоторая мера потери. Обозначение:

$$D_{KL}(a||b) = CE(a||b) - H(a).$$

Теорема:

$$CE(a||b) \geq CE(a||a),$$
$$D_{KL}(a||b) \geq 0.$$

Доказательство: Построив графики $y = x - 1$ и $y = \ln x$, можно убедиться, что $x - 1 \geq \ln x$. Подставим $x = \frac{a_i}{b_i}$. Получаем:

$$\frac{a_i}{b_i} - 1e \geq \ln \frac{a_i}{b_i}$$
$$\sum (a_i - b_i) \geq \sum (b_i \ln a_i - b_i \ln b_i).$$

Так как A и B – распределения вероятностей, слева стоит разность $1 - 1 = 0$. Получаем:

$$\sum (b_i \ln a_i) - \sum (b_i \ln b_i) \leq 0.$$

Перейдём от натуральных логарифмов к логарифмам с основанем $\frac{1}{2}$. Для этого домножим обе части неравенства на $\log_{\frac{1}{2}} e$, потому что:

$$\log_e a \times \log_{\frac{1}{2}} e = \log_{\frac{1}{2}} e^{log_e a} = \log_{\frac{1}{2}} a.$$

Тогда получим:

$$\sum (b_i \log_{\frac{1}{2}} a_i) - \sum (b_i \log_{\frac{1}{2}} b_i) \geq 0,$$

так как $\log_{\frac{1}{2}} e < 0$. Таким образом,

$$\sum (b_i \log_{\frac{1}{2}} a_i) \geq \sum (b_i \log_{\frac{1}{2}} b_i)$$
$$CE(b||a) \geq CE(b||b).$$

В непрерывном случае всё сохраняется, но все суммы заменяются на интегралы. Энтропия в непрерывном случае плохо интерпретируется. Например:

$$D_{KL}(a||b) = \int_{-\infty}^{+\infty} a \log(b) dt - \int_{-\infty}^{+\infty} a \log(a) dt = \int_{-\infty}^{+\infty} a \log\left(\frac{b}{a}\right) dt.$$

В дискретном случае, скорее, хотим минимизировать энтропию, а в непрерывном – D_{KL}.

Пример из Variational Bayes. Имеется апостериорное распределение $P(\theta|y)$, хитрое и странное. Возьмём другое распределение, $q(\theta)$, так, чтобы оно было простым и было похоже на апостериорное. Предположим, что:

$$q(\theta) = q_1(\theta_1) \times q_2(\theta_2).$$
$$\theta_1 \sim N(\mu_1, \sigma_1^2).$$
$$\theta_2 \sim N(\mu_2, \sigma_2^2).$$

Что проще минимизировать по q (ответ в разделе про VB):

1. $D_{KL}(q||p) = \int q \ln \frac{p}{q} dt.$
2. $D_{KL}(p||q) = \int p \ln \frac{q}{p} dt.$

5 Тета-метод

Тета-метод – частный случай ETS(AAN), поэтому он не подходит для анализа сезонных рядов. Напоминание: ETS(AAN):

$$\varepsilon \sim N(0, \sigma^2)$$
$$\begin{cases} y_t = l_{t-1} + b_{t-1} + \varepsilon_t, \\ l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t, \\ b_t = b_{t-1} + \beta\varepsilon_t. \end{cases}$$

ML: α, β, σ^2, l_0, b_0.

Тета-метод: $\beta = 0$, $l_1 = y_1$ (положим линию долгосрочного уровня так, чтобы было попадание на первое наблюдение). Следовательно:

$$y_t = l_{t-1} + b + \varepsilon_t.$$
$$l_t = l_{t-1} + b + \alpha\varepsilon_t.$$
$$b_t = b_0 = b.$$
$$y_1 = b_1 = b.$$

Тогда:

$$\Delta y_t = \Delta l_{t-1} + \Delta b + \Delta\varepsilon_t.$$
$$\Delta y_t = b + \alpha\varepsilon_{t-1} + \varepsilon_t - \varepsilon_{t-1}.$$
$$\Delta y_t = b + \varepsilon_t + (\alpha - 1)\varepsilon_{t-1}.$$

А это ARIMA(0, 1, 1).

Тета-метод получается лучше, чем ETS-AAN.

6 Байесовский подход

Задача:

1. Данные: $y_1 = 3$, $y_2 = 5$.
2. Модель: $y_i \sim N(\mu, \frac{1}{\tau})$ и независимы.
3. Априорное мнение о μ и τ: $\tau \sim \gamma(2, 7)$, $\mu|\tau \sim N(1, \frac{2}{\tau})$.

Нужно найти апостериорное распределение $f(\mu, \tau|y)$:

a) Явно: N-IG (нормальное обратное гамма, нормальное обратное Уишарта): $\sigma^2 \sim IG$, $\mu|\sigma^2 \sim N$.

b) Описать явно алгоритм MH-RW.

c) Описать VB (вариационный Байес).

6.1 Явно (N-IG)

Априорное распределение:

$$f(\tau) = Ce^{-7\tau}\tau^{2-1} \propto e^{-7\tau}\tau.$$
$$f(\mu|\tau) = \frac{1}{\sqrt{2\pi\frac{2}{\tau}}}e^{-\frac{1}{2}(\mu-1)^2\frac{\tau}{2}}.$$

Модель:

$$f(y_i|\mu,\tau) = \frac{1}{\sqrt{2\pi\frac{1}{\tau}}}e^{-\frac{1}{2}(y_i-\mu)^2\tau}.$$

Тогда выводим апостериорное распределение:

$$f(\mu,\tau|y_1,y_2) = \frac{f(\mu,\tau,y_1,y_2)}{f(y_1,y_2)} \propto f(\mu,\tau,y_1,y_2) = f(\mu,\tau)f(y_1,y_2|\mu,\tau) =$$
$$= [\text{априорное распределение} \times \text{функция правдоподобия}] = f(\tau)f(\mu|\tau)f(y_1|\mu,\tau)f(y_2|\mu,\tau) \propto$$
$$\propto \tau e^{-7\tau}\sqrt{\tau}e^{-\frac{1}{2}(\mu-1)^2\frac{\tau}{2}}\sqrt{\tau}e^{-\frac{1}{2}(3-\mu)^2\tau}\sqrt{\tau}e^{-\frac{1}{2}(5-\mu)^2\tau} \propto \tau^{2.5}e^{-7\tau}e^{-\frac{1}{2}\tau(\frac{(\mu-1)^2}{2})+(\mu-3)^2+(\mu-5)^2} =$$
$$= \tau^{2.5}e^{-7\tau}e^{-\frac{1}{2}\tau(2.5\mu^2-17\mu+34.5)} = \tau^{2.5}e^{-7\tau}e^{-\frac{1}{2}2.5\tau((\mu-\frac{17}{5})^2+(\frac{69}{5}-(\frac{17}{5})^2))} \propto \underbrace{\tau^{2.5}e^{-7\tau-\frac{\tau}{4}(69-\frac{17^2}{5})}}_{\gamma(2.5-0.5+1;\ 7+\frac{69-\frac{17^2}{5}}{4})} \times \underbrace{\underbrace{C}_{\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\frac{1}{2.5\tau}}}} e^{-\frac{1}{2}2.5\tau(\mu-\frac{17}{5})^2}}_{N(\frac{17}{5};\frac{1}{2.5\tau})}.$$

В итоге, снова получили NIG. Было:

$$\tau \sim \gamma(a,b).$$
$$\mu|\tau \sim N(\mu_0, \frac{k}{\tau}),$$
$$y_1 \dots y_n.$$

Стало:

$$\tau|y_1,y_2 \sim \gamma(a^*,b^*),$$
$$\mu|\tau,y_1,y_2 \sim N(\mu_0^*, \frac{k^*}{\tau}),$$

Параметры априорного распределения можно восстановить в общем виде, например:

$$a^* = a + \frac{n}{2},$$

потому что к изначальному $a-1$ в степени τ придёт $n \times 0.5$ из функций плотности y_i и 0.5 из функции плотности μ, а в конце нужно вычесть одну 0.5 и прибавить 1 для восстановления константы нормального распределения. Получаем, что $a^* = a - 1 + 0.5n + 0.5 - 0.5 + 1$.

6.2 MH-RW

Цель – получить набор $\left(\mu^{[1]},\tau^{[1]}\right)$, $\left(\mu^{[2]},\tau^{[2]}\right)$, $\left(\mu^{[3]},\tau^{[3]}\right)$, ... Вопрос – как это сделать.

Шаг 1: Генерируем случайно $\tau^{[1]} \sim f(\tau)$. Получаем $\mu^{[1]}$. В нашем случае:

$$\tau^{[1]} \sim \gamma(2,7),$$
$$\mu^{[1]} \sim N(1, \frac{2}{\tau^{[1]}}).$$

Шаг 2: Сочиняем предложение:

$$\tau^{[\text{prop}]} = \tau^{[1]} + N(0;4),$$
$$\mu^{[\text{prop}]} = \mu^{[1]} + N(0;1).$$

Прибавляемые распределения – вопрос настройки алгоритма.

Шаг 3: С некоторой вероятностью предложение одобряется, и тогда $\tau^{[2]} = \tau^{[\text{prop}]}$, $\mu^{[2]} = \mu^{[\text{prop}]}$, а затем переходим к шагу 2. Если предложение не одобряется, то просто переходим к шагу 2.

Как получить вероятность перехода?

Основная идея: если $\theta^{[n]}$ $(\theta = (\mu, \tau))$ уже генерируется из $f(\theta|y_1, y_2)$, то **не портить!**

Достаточное условие:

$$s(\theta_A \to \theta_B) = s(\theta_B \to \theta_A),$$

то есть количество переходов из A в B за единицу времени равно количеству переходов из B в A за единицу времени.

Распишем это условие в плотностях:

$$f(\theta_A|y) \times f_{N(0,1)}(\mu_B - \mu_A) \times f_{N(0,4)}(\tau_B - \tau_A) \times \alpha(\theta_A \to \theta_B) =$$
$$= f(\theta_B|y) \times f_{N(0,1)}(\mu_A - \mu_B) \times f_{N(0,4)}(\tau_A - \tau_B) \times \alpha(\theta_B \to \theta_A),$$

где $\alpha(\cdot)$ – вероятность одобрения перехода, указанного в скобках. Функции плотности нормального распределения сокращаются, и получаем следующее условие:

$$\frac{\alpha(\theta_B \to \theta_A)}{\alpha(\theta_A \to \theta_B)} = \frac{f(\theta_A|y)}{f(\theta_B|y)}.$$

А $f(\theta_A|y)$ и $f(\theta_B|y)$ знаем из пункта а):

$$f(\theta_A|y) = f(\mu_A, \theta_A|y_1, y_2),$$
$$f(\theta_B|y) = f(\mu_B, \theta_B|y_1, y_2).$$

Пример: $f(\theta_A|y) = 0.0012$, $f(\theta_B|y) = 0.0036$. Тогда, например, $\alpha(\theta_A \to \theta_B) = 1$, $\alpha(\theta_B \to \theta_A) = \frac{1}{3}$. Для вероятностей одобрения можно взять и другие числа (например, $\frac{1}{5}$ и $\frac{1}{15}$), но не берут, так как сходимость дольше (3 предложения на один шаг против 15 предложений на один шаг).

В компьютере хранятся не сами функции плотности, а их логарифмы, чтобы не было выхода за размеры хранения чисел. Тогда:

$$\ln \alpha(\theta_B \to \theta_A) - \ln \alpha(\theta_A \to \theta_B) = \ln f(\theta_A|y) - \ln f(\theta_B|y).$$

6.3 Вариационный Байес

Явный алгоритм – точный, на конечной выборке – примерный. MH-RW – асимптотически точный. VB – примерный. То есть через VB получаем распределение, похожее на апостериорное, но сколько ни жди, точным он не будет.

Идея:

$$f(\mu, \tau|y),$$
$$\tau|y \sim \gamma(\ldots),$$
$$\mu|\tau, y \sim N(\ldots),$$

μ и τ зависимы.

Чаще всего: $q(\mu, \tau) = q(\mu)q(\tau)$. Приближаем распределениями: $\mu \sim N(?; ?)$, $\tau \sim N(?; ?)$, параметры которых надо подобрать.

$$D_{KL}(q||f(\theta|y)) \geq 0,$$
$$D_{KL}(f(\theta|y)||q) \geq 0.$$

Что проще минимизировать по q:

1. $D_{KL}(q||f(\theta|y)) = \int_\theta \underbrace{q(\theta)}_{\text{простая}} \ln \frac{f(\theta|y)}{q(\theta)} d\theta.$
2. $D_{KL}(f(\theta|y)||q) = \int_\theta \underbrace{f(\theta|y)}_{\text{адская}} \ln \frac{q(\theta)}{f(\theta|y)} d\theta.$

Вывод: по q легче минимизировать первую функцию. Упростим далее:

$$\int_\theta q(\theta) \ln \frac{f(\theta|y)}{q(\theta)} d\theta = \int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta - \int_\theta q(\theta) \ln f(y) d\theta = \int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta - \ln f(y).$$

Получается, что минимизировать нужно только уменьшаемое:

$$\int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta = \int_\theta q(\mu)q(\tau) \ln \frac{f(\theta,y)}{q(\theta)} d\theta \to \min_q.$$

Процедура VB с высоты птичьего полёта:

Шаг 1: Подбираем $q_1(\mu)$, чтобы $\min \int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta$.

Шаг 2: Подбираем $q_2(\tau)$, чтобы $\min \int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta$. Далее переходим к шагу 1.

Как подбирать? Сделаем вспомогательное упражнение. В $\int_\theta q(\theta) \ln \frac{f(\theta,y)}{q(\theta)} d\theta$ выделить часть, которая зависит от q_1.

$$\iint q_1(\mu)q_2(\tau) \ln \frac{f(\mu,\tau,y)}{q_1(\mu)q_2(\tau)} d\mu d\tau =$$
$$= \iint q_1(\mu)q_2(\tau) \ln f(\mu,\tau,y) d\mu d\tau - \iint q_1(\mu)q_2(\tau) \ln q_1(\mu) d\mu d\tau - \iint q_1(\mu)q_2(\tau) \ln q_2(\tau) d\mu d\tau.$$

Помня, что минимизируем по q_1, видим, что последний член – это константа, так как q_1 интегрируется в 1, а остальное не зависит от q_1. В двух первых членах вынесем $q_2(\tau)$ за знак внутреннего интеграла и скомбинируем. Получаем:

$$\int q_2(\tau) \int q_1(\mu)(\ln f(\mu,\tau,y) - \ln q_1(\mu)) d\mu d\tau.$$

Как это минимизировать по q? Отступление: уравнение Эйлера.

<Продолжение следует>

7 Уравнение Эйлера

$$\int_0^1 g(y, \dot{y}, t)dt \to \max_y .$$

Пример:

$$\int_0^1 (\dot{y} - 7)^2 dt \to \min_y .$$

Понятно, что $y(t) = 7t + k$.

Рассмотрим возмущения нашей функции:

$$y(t) + \delta v(t) = \tilde{y}(t),$$

где δ – число, а $v(t)$ – возмущение. Тогда:

$$\dot{\tilde{y}}(t) = \dot{y}(t) + \delta \dot{v}(t).$$

Идея: если нашли экстремум, модифицировать функцию каким угодно возмущением не выгодно, а значит, можно минимизировать по δ, то есть:

$$\frac{d\left[\int_0^1 f(y + \delta v; \dot{y} + \delta \dot{v}; t)dt\right]}{d\delta} = 0, \ \forall v.$$

Пусть все функции «хорошие», тогда:

$$\int_0^1 \frac{df(\dots)}{d\delta} dt = 0.$$
$$\int_0^1 \left(f'_y(y + \delta v; \dot{y} + \delta \dot{v}; t)v + f'_{\dot{y}}(y + \delta v; \dot{y} + \delta \dot{v}; t)\dot{v}\right) dt = 0.$$

Подставим $\delta = 0$:

$$\int_0^1 \left(f'_y(y; \dot{y}; t)v + f'_{\dot{y}}(y; \dot{y}; t)\dot{v}\right) dt = 0.$$

Проинтегрируем второе слагаемое по частям:

$$\int_0^1 f'_{\dot{y}}\dot{v}dt = f'_{\dot{y}}v \Big|_0^1 - \int_0^1 \frac{df'_{\dot{y}}}{dt} v dt.$$

Рассмотрим простую задачу с фиксированными краями (функция должна начинаться и заканчиваться в каких-то точках). Возмущение на краях равно 0, поэтому первое слагаемое равно 0. Тогда получаем:

$$\int_0^1 \left(f'_y(y; \dot{y}; t)v - \frac{df'_{\dot{y}}}{dt} v\right) dt = 0.$$
$$\int_0^1 \left(f'_y(y; \dot{y}; t) - \frac{df'_{\dot{y}}}{dt}\right) v dt = 0.$$

Так как всё выражение должно быть равно 0 для $\forall \delta$, то выражение в скобках должно быть равно 0. То есть:

$$f'_y = \frac{df'_{\dot{y}}}{dt}.$$

Это и есть уравнение Эйлера-Лагранжа.

Пример:

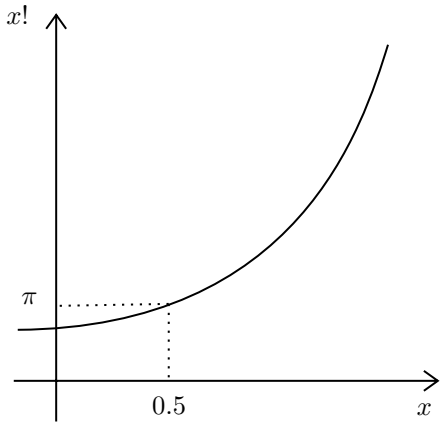
$$\int_0^\pi (y^2 - (\dot{y})^2)dt \to \min_y.$$
$$y(0) = 1, y(\pi) = \frac{1}{2}.$$

Применяем уравнение Эйлера:

$$2y = \frac{d(-2\dot{y})}{dt} \Rightarrow y + \ddot{y} = 0.$$
$$\lambda^2 + 1 = 0 \Rightarrow \lambda = \pm i.$$
$$y(t) = a\cos t + b\sin t.$$

8 Разное

1. Гамма-функция:



2. Задачка на пробит.

$$y_i = \begin{cases} 1, \text{ if } y_i^* > 0, \\ 0, \text{ if } y_i^* \le 0. \end{cases}$$
$$y_i^* = \beta_1 + \beta_2 x_i + u_i$$

Вопрос: как изменятся коэффициенты, если u_i распределено не $N(0,1)$, а $u_i \sim N(0,4)$.

Рассмотрим эту модель:

$$y_i^* = \beta_1 + \beta_2 x_i + u_i,$$
$$u_i \sim N(0,4).$$

Поделим левую и правую часть на 2:

$$\frac{y_i^*}{2} = \frac{\beta_1}{2} + \frac{\beta_2}{2} x_i + \frac{u_i}{2}.$$

Переобозначим:

$$y_i^* = \tilde{\beta_1} + \tilde{\beta_2} x_i + \tilde{u_i}.$$

А это классическая модель с $\tilde{u_i} \sim N(0,1)$. Поэтому ответ: увеличатся в два раза.

3. Как считать сложные интегралы с параметром.

Пусть есть интеграл:

$$\int_{-\infty}^{+\infty} e^{-ax^2} dx = f(a).$$

Введём единицы измерения. Пусть x измеряется в слонах. Тогда dx тоже измеряется в слонах. Сделаем так, чтобы $f(a)$ тоже измерялось в слонах. Понятно, что тогда e^{-ax^2} должно измеряться «ни в чём». Чтобы это произошло, a должно измеряться в [слонах$^{-2}$]: тогда в степени экспоненты единицы измерения сократятся. А так как $f(a)$ тоже измеряется в слонах, то $f(a)$ должна быть такой, чтобы [слоны$^{-2}$] переходили в [слоны]. Это означает, что $f(a) = c\dfrac{1}{\sqrt{a}}$. В итоге, восстановили интеграл с точностью до константы.

Другой пример:

$$\int_{-\infty}^{+\infty} e^{-\frac{x^3}{a}}\,dx = f(a).$$

Рассуждения аналогичные. a должна измеряться в [слонах^3]. Тогда $f(a)$ должна переводить [слонов^3] в [слоны]. Тогда $f(a) = c\sqrt[3]{a}$.