Стохастические модели

Колданов Александр Петрович

Задача о страховании жизни.

T - ~~остаток продолжительности~~ жизни с.в. A год смерти

$F_T(t) = P(T<t) \ \forall t \in \mathbb{R}$

$S(t) = G(t) = P(T \geq t) = 1 - F(t)$ - вероятность выживания

$L_0 = 100\,000$ - число наблюдений

L_x - число людей, которые прожили x лет

$S(t) = \frac{L_x}{L_0}$

Считаем, что T - непрерывная с.в.

Плотность $f(x)$

$F(t) = \int_{-\infty}^{x} f(x)dx \ \forall x \in \mathbb{R}$

Если дифференцируема, то $f(x) = F'(x)$

Кривая смертности

$f(x) = -S'(x) = \frac{d_x}{L_0}$

$P(x \leq T \leq x+1) = F(x+1) - F(x) \approx \frac{d_x}{L_0}$, d_x - число людей умерших в x лет

$= S(x) - S(x+1)$

$= \int_x^{x+1} f(u)du = f(x_{ср})(x+1-x) = f(x)$

$E\{x\}$ - среднее время жизни $\int_{-\infty}^{\infty} x f(x)dx = \int_0^{+\infty} S(x)dx$ если только пол. зн.

$D\{x\}$

Мода - самый популярный возраст смерти, $\max f(x)$

Медиана - $P(T \geq t_0) = P(T \leq t_0)$

μ_x - интенсивность смертности

$\mu_x = \frac{f(x)}{S(x)} = -\frac{S'(x)}{S(x)} = (\ln(S(x)))'$

$\ln S(x) = -\int_0^x \mu(u)du$

$S(x) = \exp\left(-\int_0^x \mu(u)du\right)$

80 лет 100 человек ×

30 ✓

$\mu_{80} = \frac{100}{30}$

Задача 2.3:

$P(50 \le T < 70) = 1 - S(70) - 1 + S(50) = S(50) - S(70) = 0{,}915 - 0{,}682$

$L_0 = 1000$

A - человек умер в промежутке $50 \le t < 70$

$P(X = k) = C_n^k p^k (1-p)^{n-k}$ - биномиальный закон

$E\{X\} = np$

$X_i = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}$ - конкретный человек

$X = \sum_{i=1}^{1000} X_i$

$E\{X\} = \sum_{i=1}^{n} E\{X\} = \sum_{i=1}^{n} (1 \cdot p + 0 \cdot (1-p)) = np$

$Var\{X\} = npq$

$\mu_{70} = \frac{f(x)}{S(x)} = \frac{S(75) - S(70)}{S(75)}$

Задача 2.2)

$f(x) = \begin{cases} \frac{x}{a^2} e^{-x/a}, & x > 0 \\ 0, & \text{иначе} \end{cases}$

$F(x) = \int_0^{+\infty} \frac{x}{a^2} e^{-x/a} = \frac{1}{a^2} \dots = 1$

0,682

Условная вероятность

2.5, 2.6, 2.7, 2.8

T - время жизни человека

$L_0 = 100\,000$

T_x - остаточное время жизни $= T - x$

$$F_x(t) = P(T_x < t \mid_{T \geq x}) = \frac{P(T - x \leq t,\ T \geq x)}{P(T \geq x)} = \frac{P(x \leq T \leq x + t)}{S(x)} =$$

$$= \frac{F_x(x+t) - F_x(x)}{S(x)} = \frac{S(x) - S(x+t)}{S(x)} = 1 - \frac{S(x+t)}{S(x)}$$

$$S_x(t) = 1 - F_x(t) = \frac{S(x+t)}{S(x)}$$

$$f_x(t) = \frac{f(x+t)}{f(x)}$$

$$\mu_x(t) = \frac{f_x(t)}{S_x(t)} = \frac{f(x+t)}{S(x)} \cdot \frac{S(x)}{S(x+t)} = \frac{f(x+t)}{S(x+t)} = \mu_{x+t}$$

Задача 2.24)

$x = 20$

1) $F_{20}(30) = \frac{L_{20+30}}{L_{20}} = \frac{89\,272}{97\,290}$

2) $P(20 < T_{20} < 50) = S_{20}(10) - S_{20}(50) = \frac{L_{40} - L_{70}}{L_{20}} = \frac{94086 - 64514}{97290}$

3) $P(T_{20} < 10) = 1 - S_{20}(10) = 1 - \frac{L_{30}}{L_{20}}$

Задача 2.2)

$$\int_0^{+\infty} \frac{x}{a^2} e^{-x/a} dx = 1$$

$$S(x) = \frac{x+a}{a} e^{-x/a}$$

Модели смертности

Модель Муавра:

$f(x) = 1/\omega$

$S(x) = 1 - x/\omega$

$\mu_x = \frac{1}{\omega - x}$

Мейкхам

$\mu_x = A + Hx + Be^{\alpha x}$

Гомперту

$\mu_x = Be^{\alpha x}$

Вейбулл

$\mu_x = kx^n$

Выводим время жизни без эффекта усталости:

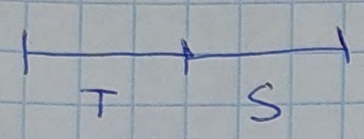
T – время жизни

A_T – человек проживет T лет

$G(T) = P(A_T)$ – жив в течение T лет

1) строго убывает

2) непрерывна и строго монотонна



$A_{T+S} = A_T \cap A_S$ (начало промежутков разное)

$P(A_{T+S}) = P(A_T \cap A_S) = P(A_T) \cdot P(A_S | A_T)$

Если не учитываем эффект усталости

$P(A_S | A_T) = P(A_S)$

$P(A_T \cap A_S) = P(A_T) P(A_S)$ (1)

$G(T+S) = G(T) G(S) \quad \forall T, S$

$G(2T) = G(T)^2$ $\qquad nT = S \quad G(S) = G\left(\frac{S}{n}\right)^n$

$G(nT) = G(T)^n$ $\qquad G\left(\frac{T}{n}\right) = G(T)^{1/n}$

$G(rt) = G(t)^r$

$G(xt) = [G(t)]^x$

$t = 1 \Rightarrow G(x) = \underbrace{G(1)}_{=a}{}^x$

$G(x) = a^x = \exp(\underbrace{+\ln a}_{\lambda > 0} \cdot x)$

$$G(x) = \begin{cases} \exp(-\lambda x), & x \geq 0 \\ 0, & x < 0 \end{cases}$$ – Функция выживания

$$S(x) = \begin{cases} \exp\left(-\frac{kx^{n+1}}{n+1}\right), & x > 0 \\ 0 \end{cases}$$

$$f(x) = \begin{cases} kx^n \exp\left(-\frac{kx^{n+1}}{n+1}\right) \\ 0 \end{cases}$$

$M_x = kx^n$

$X - F(x)$

Основные понятия мат. статистики

Основной объект тер. вер-а $(\Omega, \mathcal{A}, P)$ ← одно распределение

мат. статистики $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ ← семейство распределений

1) $\mathcal{P} = \{f(x; \theta : \theta \in \mathbb{R}^n\}$
2) $\mathcal{P}$

Выборка – результаты наблюдений за случайной величиной

$X_1, \dots, X_n$ – выборка (состоит из случайных величин)

$x_1, \dots, x_n$ – конкретный набор данных

$X : F(x) \in \mathcal{P}$

X_i – случайная величина, описывающая i-е наблюдение

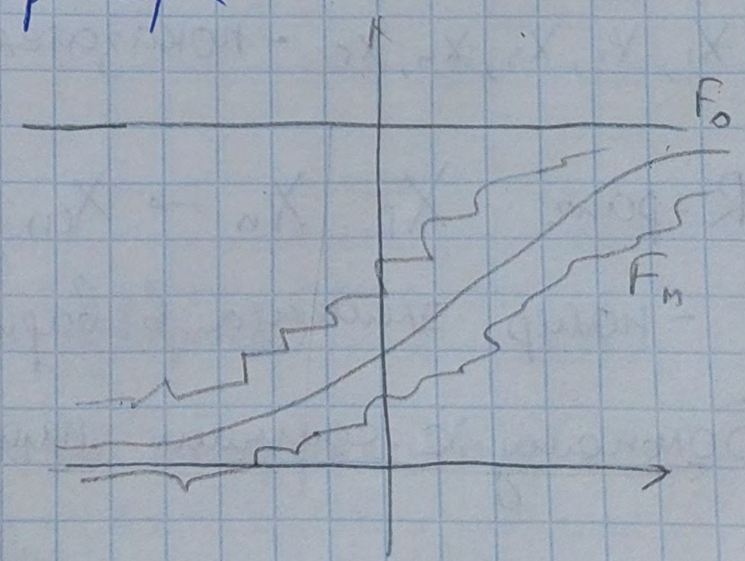
$T(X_1, \dots, X_n)$ – статистика

Типы задач

1) Оценка параметров
2) Проверка гипотез
3) Доверительное оценивание

$\hat{\alpha} = \hat{\alpha}(X_1, \dots, X_n)$ – случайная величина параметра

$\hat{\alpha} = \alpha(x_1, \dots, x_n)$ – конкретное число



$H = F_X(x) \in \mathcal{P} \subset \mathcal{P}$ – гипотеза согласия

$H_1 = F(x) = F_0(x)$ – гипотеза согласия

$|F_n(x) - F_0(x)| > C$

Виды гипотез:

1) Гипотиза согласия

2) Гипотиза однородности, отсутствия динамики → сравнения способа действия

H: лекарство не помогает

3. Гипотеза независимости

4. Параметрическая гипотеза

Тест суммы рангов Вилкоксона

1) Модель случайности. Набор объектов дан

2) Популяционная модель. Набор объектов выбирается случайно

Результаты наблюдений независимые с.в.

$X - F_x(x)$ $Y - F_y(y)$

H: $F_X(x) = F_Y(x)\ \forall x$

3)

Результа наблюдений зависимые с.в.

Используется тест знаковых рангов Вилкоксона

Модель случайности

Есть набор человек:

$\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$

Выбираем случайно 3х человек и даем лекарство, остальным плацебо.

плек	123	124
плац	45	35

$C_5^2 = C_5^3 = 10$

H: у лекарства нет эффекта K: у лекарства есть вариант

X_1, X_2, X_3, X_4, X_5 - показатели болезни людей (напр. давление)

R - ранг $X_1 \dots X_n \to X_{(1)} < X_{(2)} < \dots < X_{(n)}$

- номер элемента в вариационном ряду

Больному с самыми плохими результатами ставим ранг = 1.

1. M - белых, N-M - черных

1) белый шар на iтм месте

2) белый шар раньше черного

3) на i, j шары разных цветов (одинаковых)

1) $\frac{(n-1)!\,m}{n!} = \frac{m}{n}$

2) $\frac{m}{n}$

3) $\frac{(n-m)m(n-2)!}{n!} = \frac{(n-m)m}{n\cdot n-1}\cdot 2$

4) $\frac{m(m-1)(n-2)!}{n!} = \frac{m(m-1)}{n\cdot(n-1)}$

$+$

$\frac{(n-m)(n-m-1)(n-2)!}{n!} = \frac{(n-m)(n-m-1)}{n(n-1)}$

Всего N объектов. n случайных обработаем, m оставлен. $n+m=N$

Вероятность возникновения конкретного разбиения:

$\frac{1}{C_N^n} = \frac{n!\,m!}{N!}$

$S_1 S_2 \dots S_n\ R_1 \dots R_m$ - ранги $\quad T = \sum_{i=1}^{n+m}$

$P_H(S_1 = x_1\ S_2 = x_2 \dots S_n = x_n) = \frac{1}{C_N^n}$ (при предположении истинности гипотезы)

Это верно только если мы считаем, что лекарство не действует.

$W_S = \sum_{i=1}^{n} S_i$ - статистика

$\delta = \begin{cases} 1, & W_S \geq C \\ 0, & W_S < C \end{cases}$ - решающее правило

гипотеза отклоняется

гипотеза подтверждается

$P_H(W_S > C) \leq \alpha \quad \alpha$ - выбираем

Используем k потому что:(

1) консервативная точка зрения (если сильно уверены)

2) Мы знаем распределение, если лекарство неэффективно.

$P(\sum_{i=1}^{n} S_i = x) = \frac{r(x,n,m)}{C_N^n}$ - число перестановок.

Статистика Мана-Уитни:

$W_{XY} = S - \frac{n(n+1)}{2} \quad \left(S = \sum_{i=1}^{n} S_i \right)$

$W_{YX} = R - \frac{m(m+1)}{2}$

Всего N элементов, n-обрабатывали, m-контроль

Расставим ранги в общем ряду:

S_1 - перед ним $(S_1 - 1)$ элементов, которые не обрабатывали

S_1 - перед ним $(S_2 - 2)$ -//-//

$\vdots$

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

$W_{XY} = \sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_i > X_j)$

Y_i - хар-тики объектов, которых обрабатывали

X_i - хар-тики объектов, которые не обрабатывали

$P_H(W_{XY} > C) \leq \alpha$

$E\{W_{XY}\} = \frac{nm}{2}$

$D(I(Y_i > X_i)) = E(I^2(Y_i > X_j)) - [E(I(Y_i > X_j))]^2 = 1/2 - (1/2)^2 = 1/4$

$D(\sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_i > X_j)) = E(\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{n} \sum_{L=1}^{m} I(Y_i > X_j) I(Y_k > X_L)) - (E(I(Y_i > X_j)))^2 =$

$=$

1. $i = k \quad j = L$

2. $i = k \quad j \neq L$

$E(I(Y_i > X_j)(Y_i > X_L)) = P(Y_i > X_j, Y_i > X_j) = 1/3$

$Z_1, \ldots, Z_N$ - случайные величины, характеристика случайно выбранного объекта.

Задача:

Лагутин Мат. мат. стат

Наблюдение $a = \{1.3, 1.7, 1.7, 2.5\}$, ранги $r = \{1, 2.5, 2.5, 4\}$

2 выбираем, 2 оставляем на контроль

Лидерман Введение в мат. стат.

$P(S_1^* = x_1, S_2^* = x_2) =$

$P(S_1^* = 1, S_2^* = 2,5) = 2/6$

$P(S_1^* = 2,5, S_2^* = 4) = 2/6$

$P(S_1^* = 2,5, S_2^* = 2,5) = 1/6$

$P(S_1^* = 1, S_2^* = 4) = 1/6$

T	3.5	5	6.5
P	2/6	2/6	2/6

Пусть $n = m = 3$ $\quad (2, 2, 9)$ $\quad$ ~~2 2~~

$d_k = \{2$ $\quad (4, 9, 9)$ $\quad$ 2 2 4 9 9 9

1.5 1.5 3 5 5 5

$W_{XY} =$

$P(S_1 = 1.5, S_2 = 1.5, S_3 = 1,5)$

Если есть совпадающие характеристики, то

$$W_{XY} = \sum_i \sum_j I^*(Y_i > X_i)$$

$$I^* = \begin{cases} 1 & Y_i > X_j \\ 1/2 & Y_i = X_j \\ 0, & Y_i < X_j \end{cases}$$

$$D(W_S^*) = \frac{mn(N+1)}{12} - \frac{mn \sum_{i=1}^{L} (d_i^3 - d_i)}{12N(N-1)}$$

Популяционные модели

$\exists$ бесконечная генеральная выборка

N - объектов, n - обрабатываем, m - контроль

$Y_1, \dots, Y_n$ - случайные величины обработанных элементов $= F$

$X_1, \dots, X_n$ - —//— контроля

Случайные имеют распределение $F_Y(y)$, контроль $F_X(x)$

$H: F_X(x) = F_Y(x)$ - гипотеза однородности

Предполагаем, что X, Y - непрерывные с.в.

$P(X = Y) = 0$

$$W_{XY} = \sum_j \sum_i I(Y_j > X_i)$$

$$E_H\{I(Y > X)\} = 1/2$$

$$E_H\{I(Y > X)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{xy}(x, y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_x(x) f_y(y)\,dx\,dy =$$

$$= \int_{-\infty}^{\infty} F_y(x) \cdot f_x(x)\,dx$$

Пример:

$A, B \qquad A < B$

$P(x = A) = p$

$P(x = B) = q$

Возьмем 3 объекта из 2х групп:

$AAA \to (2, 2, 2) \qquad p^3$

$AAB \to (1{,}5, 1.5, 3) \qquad p^2 q$

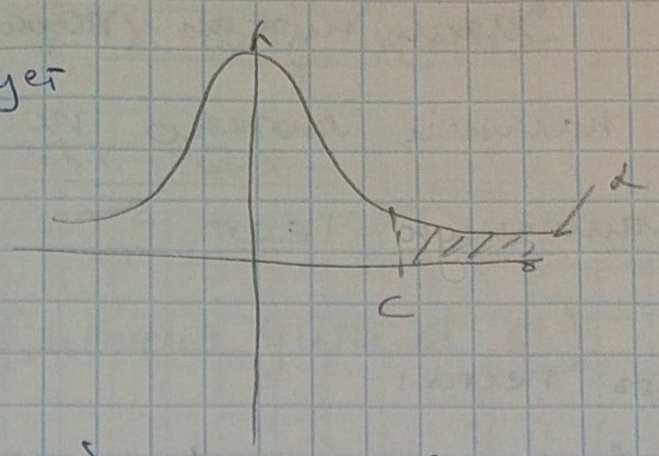
$ABB \to (1, 2.5, 2.5) \qquad p q^2$

$BBB \to (2, 2, 2) \qquad q^3$

1	1.5	2	2.5	3
pq^2	$2p^2q$	p^3+q^3	$2pq^2$	p^3q

1) H: эффект от нового способа отсутствует

$$\varphi(x) = \begin{cases} 1, & T(x) > C \\ 0, & T(x) \le C \end{cases}$$



$P_H(T(x) > C) = \alpha$ (0,05)

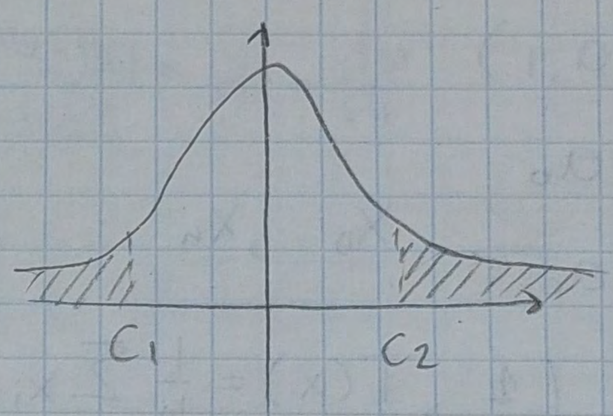
Односторонняя альтернатива. – новый способ лучше (отдаем предпочтение гипотезе)

2) H: новый способ = стандартному

K: новый способ отличается

Двухсторонняя альтернатива – новый способ отличается

$$\varphi_{двух}(x) = \begin{cases} 1, & T(x) < C_1 \text{ или } T(x) > C_2 \\ 0, & C_1 < T(x) < C_2 \end{cases}$$



$P_H(T(x) < C_1 \text{ или } T(x) > C_2) = \alpha$

2 решения: отличается или нет

3) H 1 новый способ = 2 новый способ

3 решения (не отличаются, 1й лучше, 2й лучше)

Двухсторонняя гипотеза отвергается реже. (= получаем меньше значимых результатов)

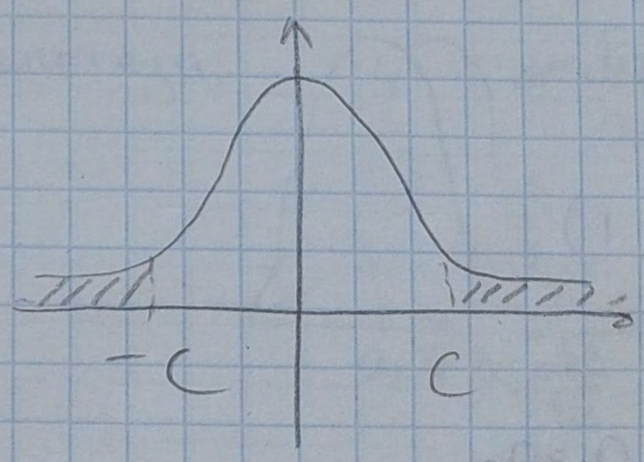
Значим является отвержение теории.

4) H_1: 1 > 2

H_2: 2 < 1

$P_{2<1}(2>1) = \alpha$

$P_{2>1}(2<1) = \alpha$



$$\varphi(x) = \begin{cases} 2>1: T(x) > C \\ 2<1: T(x) < -C \end{cases}$$

$-C < T(x) < C$ – продолжаем наблюдение, поступаем как хотим (зона неопределенности) (по данным мы не можем достоверно сделать вывод)

Зачем нужны параметрические модели:

При их наличии можно построить оптимальный или ассимптотически оптимальный тест.

Мощность теста:

H: здоров

K: болен

$$\varphi(x) = \begin{cases} d_k \\ d_H \end{cases}$$

$P(d_k|H) \le \alpha$ – ошибка I рода

$P(d_H|k)$ – ошибка II рода

Обычно ограничивается ошибка I рода.

$X \sim N(a,1)$

$H: a \le a_0$

$K: a > a_0$ $\quad X_1, \dots, X_n$

$$\varphi(x) = \begin{cases} 1, & T(x) = \frac{1}{n}\sum x_i > C \\ 0, & T(x) < C \end{cases}$$

$P_{a_0}(T(x) > C) = \alpha$ – ошибка I рода.

$P_a(T(x) < C) = \beta$ – ошибка II рода. $a > a_0$

Мощность теста:

$\beta = P_a(T(x) > C) \ \forall a \in \mathbb{R}$

Пример:

$X \sim N(a, 1)$

$H: a = a$

$K: a = a_1 \quad a_1 > a_0$

Построить оптимальный тест:

$$\varphi(x) = \begin{cases} 1, & \frac{f_k(x_1, \dots, x_n)}{f_n(x_1, \dots, x_n)} > C \\ 0 \end{cases}$$

$$f_k(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{(x_i - a_1)^2}{2}\right)}$$

$$f_n(x_1, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{(x_i - a_0)^2}{2}\right)}$$

Лемма Неймана-Пирсона говорит, что такой тест контролирует α и минимизирует вероятность ошибку II рода.

$$\frac{f_k(x_1,\dots,x_n)}{f_n(x_1,\dots,x_n)} = \prod_{i=1}^{n} e^{-\frac{(x_i-a_1)^2}{2}+\frac{(x_i-a_0)^2}{2}} = e^{\sum_{i=1}^{n}\frac{(x_i-a_0)^2-(x_i-a_1)^2}{2}} =$$

$$= e^{\sum_{i=1}^{n}\frac{\cancel{x_i^2}-2a_0x_i+a_0^2-\cancel{x_i^2}+2a_1x_i-a_1^2}{2}} = e^{\sum_{i=1}^{n}\frac{2a_1x_i-2a_0x_i+a_0^2-a_1^2}{2}}$$

$$\varphi(x) = \begin{cases} 1, & e^{\sum_{i=1}^{n}\left[x_i(a_1-a_0)+\frac{a_1^2}{2}+\frac{a_0^2}{2}\right]} > C \\ 0 \end{cases}$$ (можно уйти от e)

$$P_{a_0}(\varphi(X)=1)=\alpha$$

$$\varphi(x) = \begin{cases} 1, & \sum_{i=1}^{n} x_i > C \quad (a_1>a_0) \\ 0 \end{cases}$$

$P_{a_0}(\sum x_i > C) = \alpha$ т.к. $x_i \sim N(0,1)$, то $\sum x_i \sim N(na_0, n)$

Это правило работает $\forall a_1 > a_0$

$$P_{a_0}(\sum x_i > C) = \int_C^{+\infty} \frac{1}{\sqrt{2\pi n}} \exp\left(-\frac{(u-na_0)^2}{2n}\right) du \overset{t=\frac{u-na_0}{\sqrt{n}}}{=\!=\!=} dt = \frac{du}{\sqrt{n}}$$

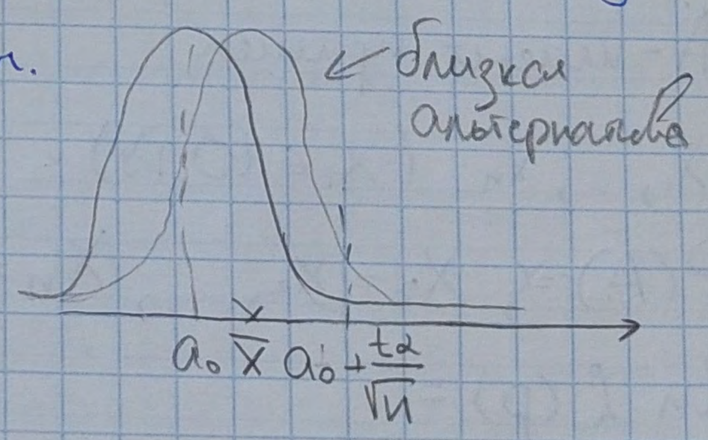
$$\frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{t^2}{2}\right) dt = \alpha$$

$$\frac{u-na_0}{\sqrt{n}} = t_\alpha$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x \exp\left(-\frac{t^2}{2}\right) dt = \alpha$$ t_α - решение уравнение

$$\varphi(x) = \begin{cases} 1, & \sum x_i > \sqrt{n_0}\, t_\alpha + na_0 \quad \left(\frac{1}{n}\sum x_i = \overline{X} > a_0 + \frac{t_\alpha}{\sqrt{n_0}}\right) \\ 0 \end{cases}$$

Данные не противоречат гипотезе a не гипотеза верна, т.к. могут быть верны близкие альтернативные гипотезы.



Осн. проблематика мат. статистики:

Есть наблюдения $x_1, \dots, x_n$.

Есть сл. $X \sim F(x; \theta)$ – параметрическая постановка задачи

$X \sim F_x \in \mathcal{F}$ – непараметрическая –//–//–

Задача уточнить распределение по наблюдения.

Типы задач:

1) Оценивание параметров (выборочные аналоги, метод макс. правдоподобия)

2) Проверка гипотез

Выборочные аналоги:

$x_1, \dots, x_n$

$F_x(x) = P(X < x) \quad x \in R$

$F_{\hat{x}}(x) = \sum_{i=1}^{n} I(x_i < x)/n$

$E\{x\} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$

Метод максимального правдоподобия:

В качестве параметра θ возьмем такой, при котором вероятность получить данную выборку максимальна.

Пример: страховая компания

A – клиент пришел

$X_1, \dots, X_n \quad (X_i \in \{0,1\})$

$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = p^{\sum x_i}(1-p)^{n - \sum x_i} = \mathcal{L}(p)$

$\ln \mathcal{L}(p) =$

$\hat{p} = \frac{\sum x_i}{n}$ (параметр, при котором функция правдобия максимальна)

$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

Свойства оценок:

1) Несмещенность:

$E_\theta \hat{\theta}(X_1, \dots, X_n) = \theta$

$\int_{-\infty}^{+\infty} \dots \int \hat{\theta}(x_1, \dots, x_n) \prod_{i=1}^{n} f(x_i, \theta) = 0$

2) Эффективность:

$E\{(\hat{\theta}-\theta)^2\} = \min$

$D\{\hat{\theta}\} \geq \frac{1}{nI(\theta)}$, где $I(\theta)$ - информационное кол-во Фишера

3) Состоятельность

$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$

Пример несмещенности:

X

$E\{X\} = a = ?$

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$E\{\overline{X}\} = E\{\frac{1}{n}\sum x_i\} = \frac{1}{n}\sum E\{x_i\} = a$

Проверка гипотез:

Задача: построить тест:

$$\varphi(x_1,\dots,x_n) = \begin{cases} 1, & T(x) \geq C \\ 0, & T(x) < C \end{cases}$$

$P_H(T(x) > C) = \alpha$

$H: p_0(x)$

$K: p_1(x)$ — разные распределения

Лемма Неймана-Пирсона:

$X_1,\dots,X_n$

$$\varphi(X_1,\dots,X_n) = \begin{cases} 1, & T = \frac{p_1(X_1,\dots,X_n)}{p_0(X_1,\dots,X_n)} > C \\ 0, & T < C \end{cases}$$

$P_H = P_0(T > C) = \alpha \quad \forall \alpha \in [0,1]$

Такой тест существует и вероятность ошибки второго рода минимальна

(наиболее мощный тест)

Пример:

$X \sim N(a, \sigma^2)$

$H: a \leq a_0$

$K: a > a_0$

Разобрать

Доверительный интервал:

Значения параметра, при котором (отвергаем или) принимаем гипотезу:

Модель:

$X \sim F_x(x)$

$Y = F_x(x)$

$F_y(y) = P(Y < y) = P(F_x(x) < y) = P(X < F_x^{-1}(y)) = F_x(F_x^{-1}(y)) = y$

Многомерное распределение:

$X = (X_1, \dots, X_n)^T$

$F_{x_1, x_2, \dots, x_n}(x_1, x_2, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n) \quad \forall (x_1, x_n) \in \mathbb{R}^n$

Пример:

N - об

Гипергеометрическое распределение:

$$P(X_1 = m_1, X_2 = m_2, \dots, X_n = m_n) = \frac{C_{M_1}^{m_1} C_{M_2}^{m_2} \dots C_{M_n}^{m_n}}{C_N^n}$$

Полиномиальное распределение:

$$P(X_1 = m_1, \dots, X_n = m_n) = \frac{n!}{m_1! \, m_2! \dots m_n!} p_1^{m_1} p_2^{m_2} \dots p_n^{m_n}$$

Многомерное нормальное распределение

$$P(X_1 = m_1, \dots, X_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} (\bar{x} - \bar{a})^T \Sigma^{-1} (x - a)$$

Есть монетка, надо получить распределение с $p = 1/3$

$$\left.\begin{matrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{matrix}\right\}$$
$$1 \quad 1$$

$$P(0,0) = \lim_{n\to\infty} \sum_{i=0}^{n} \left(\frac{1}{4}\right)^i = \frac{1/4}{1-1/4} = \frac{1}{3}$$

Задача: есть кубик, нужно получить равн. распр. с $p = 1/7$

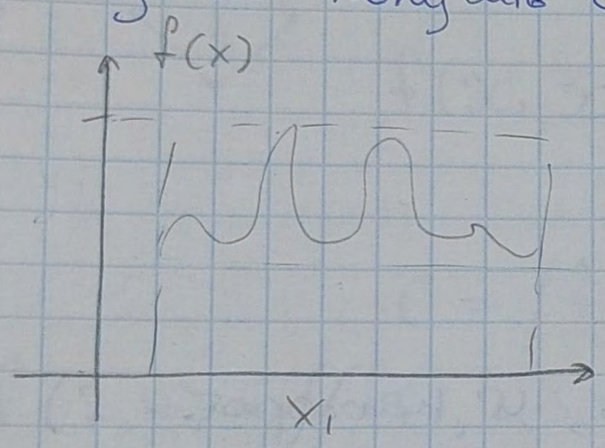
(i_1, i_2)

$i_1 \neq i_2 \to i_1$

$i_1 = i_2$, но не 6 $\to 7$

$i_1 = i_2 = 6 \to$ перебрасываем

Задача: получить с.в. из заданного распределения:



Генерируем пары $(x_1, y_1), (x_2, y_2), \ldots$

Если $y_i < f(x_i)$, то берем точку x_i, иначе отбрасываем

Частное распределение: (маргинальная)

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 < x_1, X_2 < +\infty, X_3 < +\infty, \ldots)$$

$$f_{X_1}(x_1) = \iint_{-\infty\,-\infty}^{+\infty\,+\infty} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,dx_2\,dx_n$$

Независимость:

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \sum_{i=1}^{n} f_{X_i}(x_i)$$

X, Y, Z

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

$$f_{XYZ}(x,y,z) = f_{Z|XY}(z|xy) \cdot f_X(x,y)$$

$X_1, \ldots, X_N$

$Y_1, \ldots, Y_N$

$f_{X_1,\ldots,X_N,Y_1,\ldots,Y_n} = f_{X_1,\ldots,X_n} \cdot f_{Y_1,\ldots,Y_n}$ если вектора независимы (т.е. $\forall$ с.в. из перво

вектора независима со всеми с.в. из второго вектора и наоборот).

Условное распределение:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Условная независимость:

$$P(A \cap B | C) = P(A|C) \cdot P(B|C)$$

$$f_{XY|Z}(x, y | z) = f_{X|Z}(x|z) \cdot f_{Y|Z}(y|z)$$

$X \perp\!\!\!\perp Y | Z$ – условная независимость

X, Y, Z $\|\sigma_{ij}\|$ – матрица cov (норм. распр., $E = 0$)

обр. матрица

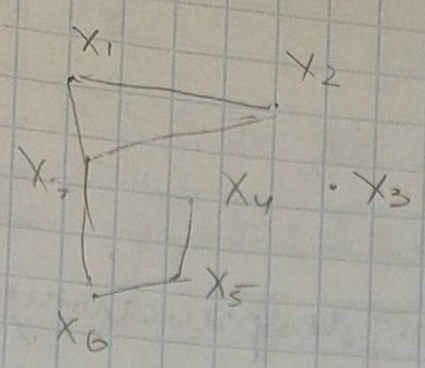
$$f(x, y, z) = C \exp\left\{ \sum_{i=1}^{3} \sum_{j=1}^{3} \sigma^{ij} x_i x_j \right\}$$

Частный коэф. коррел:

X, Y, Z

$$\rho^{x,y} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{1 - \rho_{xz}^2}\sqrt{1 - \rho_{yz}}}$$

Из условной независимости не следует марginальная независимость, и наоборот.



$$f(x_1,\dots,x_7)=\prod_{C\in G} h_{C_i}(X_{i_1},\dots,X_{i_k}),$$

тогда и только есть условно независимые величины (удовлет. одному из марковских свойств).

Марковские свойства:

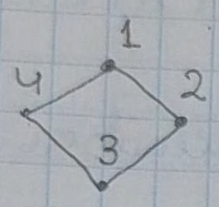
1) Если нет ребра, то $X_i \perp\!\!\!\perp X_j \mid$ все остальные (попарное)

2) Граница вершины: множество всех смежных

Любая другая вершина, не входящая в границу, условно независима.

3) Есть 3 множества вершин.

Если любой путь из X в Y проходит через Z, то $X \perp\!\!\!\perp Y \mid Z$



$$G=\{(1,2),(2,3),(3,4),(1,4)\}$$

$$f(x_1,x_2,x_3,x_4)=h(x_1,x_2)\,h(x_2,x_3)\,h(x_3,x_4)\,h(x_1,x_4)\cdot 1/C$$

$$C=\int f(x_1,x_2,x_3,x_4)\,dx_1\,dx_2\,dx_3\,dx_4$$

$$f(2,4)=\iint f(1,2,3,4)\,d_1 d_3=\frac{1}{C}\iint h(1,2)\,h(2,3)\,h(3,4)\,h(1,4)\,d_1 d_3=$$

$$=\frac{1}{\sqrt{C}}\int h(1,2)\,h(1,4)\,d_1\int h(2,3)\,h(3,4)\,d_3$$

$$f(1,3)=1/C\int h(1,2)\,h(2,3)\,d_2\int h(3,4)\,h(1,4)\,d_4$$

$$f(2,4\mid 1,3)=\frac{1/C\;h(1,2)\,h(2,3)\,h(3,4)\,h(1,4)}{1/C\int h(1,2)\,h(2,3)\,d_2\int h(3,4)\,h(1,4)\,d_4}$$

Если расписать, то получится $f(2,4\mid 1,3)=f(1\mid 2,4)\cdot f(3\mid 2,4)$

Сетевые модели и марк их связи

$(\underset{\text{вершин}}{V}, \overset{\text{ребра}}{E}, \underset{\text{веса}}{\gamma})$

В отличии от графовой модели, веса показывают маргинальную зависимость.

Зависимость мат. ожиданий не говорит о зависимости с.в.

Корелляция Пирсона:

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{E\{(X-E(X))(Y-E(Y))\}}{\sqrt{D(X)D(Y)}}$$

Независимость $\Rightarrow$ кор. $= 0$

Исключение: нормальный закон независимость $\Leftrightarrow$ кор. $= 0$

Коремируемость $\Rightarrow$ зависимость

$X \quad E\{X\} = 0 \quad E\{X^{2x+1}\} = 0$

$Y = X^2$

$$\rho = \frac{E(X(X^2 - EX)}{} = EX^3 - EX\,E(EX^2) = 0$$

С.в. связаны функционально, но они не коремярованы.

Корреляция Пирсона измеряет степень линейности корелляционной связи.

$|\rho(X,Y)| = 1 \Leftrightarrow Y = \alpha X + \beta$

$E\{Y|X\} = f(x)$ - регрессия Y на X

Если это $\neq const$, то между X и Y имеется корел. связь

$$f(x) = \sum_{i=0}^{\infty} a_i x^i$$

В случае лин. зависимости: $a_0 = E\{y\} a_1 = a_1(\rho)$, $a_j = a_j(1-\rho)$, $j \geq 2$

т.е. $a_i \to 0$ $i \geq 2$, $\rho \to 1$

МНК

$x_1, \dots, x_n$

$y_1, \dots, y_n$

$$\sum (y_i - a_0 - a_1 x_i - \dots)^2 \to \min_{a_i}$$

$X \quad Y$

$E\{(Y-\alpha-\beta X)^2\} \to \min\limits_{\alpha,\beta}$

$f'_\alpha = \begin{cases} 2E(Y-\alpha-\beta X)=0 \\ E\{X(Y-\alpha-\beta X)\}=0 \end{cases} \qquad \begin{cases} E(Y)-\alpha-\beta E(X)=0 \\ E(XY)-\alpha E(X)-\beta E(X^2)=0 \end{cases}$

$\alpha = E(Y) - \beta E(X)$

$E(XY) - (E(Y)-\beta E(X))E(X) - \beta E(X^2) = 0$

$-\beta(-E^2(X)+E(X^2)) + E(XY) - E(X)E(Y) = 0$

$\beta = \dfrac{E(XY)-E(X)E(Y)}{E(X^2)-E^2(X)} = cov(X,Y)/D(X)$

$\alpha = E(Y)$

$E\{Y|X\} = E\{Y\} + \rho\dfrac{\sigma_y}{\sigma_x}(X-E\{X\})$

$\alpha = E\{Y\} - \rho\dfrac{\sigma_y}{\sigma_x}E\{X\}$

В классе эллиптических распределений коэффициент корреляции не зависит от функции g.

$X = U_1+U_2+\ldots+U_m+V_1+\ldots+V_n$

$Y = U_1+U_2+\ldots+U_m+W_1+\ldots+W_n$

$\rho = \dfrac{cov(X,Y)}{\sqrt{D(X)D(Y)}} = \dfrac{m}{m+n}$

Знаки коэф. корреляции можно оценивать как число общих независимых факторов.

2 подхода к оценке с. в. через $P(x)$ или $E(x)$.

X, Y

$P(X\cdot Y>0)$

$P(X>0, Y>0) = 2/\pi \arcsin\rho$ (для N и всего класса эллиптических

Недостатки корреляции Пирсона:

1) Интерпретируемость

2) Не сохраняется при функциональной зависимости

Частный коэффициент корреляции:

$X_1, X_2, \dots, X_N$

$X_1^* = X_1 - \sum_{i=3}^{N} \alpha_i X_i$

$X_2^* = X_2 - \sum_{i=3}^{N} \beta_i X_i$

α_i и β_i находим методом наименьших квадратов.

Част. коэф. корреляции это $\rho_{X_1^*, X_2^*}$

Коэффициент корреляции в условном распределении:

$$f_{x_1, x_2 | x_3 \dots x_N} = \frac{f_{x_1, \dots, x_N}}{f_{x_3} \dots f_{x_n}}$$

Эти 2 коэффициента совпадают в нормальном законе.

$$\rho^{xy} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{(1-\rho^2_{xz})(1-\rho^2_{yz})}}$$

Вероятностные меры связи

X, Y – с.в.

$P((X-x_0)(Y-y_0)>0)$ – это вероят

Мера того, что с.в. ведут себя согласовано отн. (x_0, y_0)

Варианты таких точек:

1) Медиана – Мера связи Блюквиста – Краскала

$x_0 = med(X)$ (оба ↑ и ↓)

$y_0 = med(Y)$ Зн. Кор. сохраняется при монотонных преобразован

$P((X-med X)(Y-med Y)>0) = P((f(X)-med f(X))(f(Y)-med f(Y))>0)$ (1)

Коэффициент корреляции:

$$2P((X-x_0)(Y-y_0)>0)-1 = P((X-x_0)(Y-y_0)) - P((X-x_0)(Y-y_0)<0)$$

2) мат. ожидание – мера связи Фехнера

$x_0 = E(X)$

$y_0 = E(Y)$

Пример: доходность акций. Если ~~ра~~ вер. мера $= 0{,}7$. Значит обе акции идут вверх от своего среднего в 70%.

Коэффициент корреляции Кэндэла

$$2P((X_2-X_1)(Y_2-Y_1)>0)-1$$

(X_1, Y_1), (X_2, Y_2) – независимы от (X_1, Y_1), но одинаково распределены.

$$// P((X-x)(Y-y) f_{X,Y}(x,y)\,dx\,dy =$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\iint_{x\ y}^{+\infty\ +\infty} f_{XY}(u,v)\,du\,dv\, f_{XY}(x,y)\,dx\,dy = P(u>X, v>Y)$$

Коэффициент корреляции Спирмена

(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)

$P((X_1-X_2)(Y_1-Y_3)>0)$

Если (X,Y) – эллиптическое распр, то все меры связи, кроме Спирмена функ. связаны:

Фехнера = Кендалл = Бломквист – Краскалл = $\frac{2}{\pi}\arcsin\rho$

Ранговая корреляция

ученики	A	B	C	D	E	F	G	H	I	J
математ.	7	4	3	10	6	2	9	8	1	5
музыка	5	7	3	10	1	9	6	2	8	4

I	F	C	B	J	E	A	H	G	D
1	2	3	4	5	6	7	8	9	10
8	9	3	7	4	1	5	2	6	10

$P = 21 \quad Q = -24$

$S = 21 - 24 = -3$

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$