Н.Ю. Энатская
Е.Р. Хакимуллин

Математическая статистика

Москва 2011

Министерство образования и науки
Российской Федерации
Государственное образовательное учреждение высшего
профессионального образования
«Московский государственный институт
электроники и математики
(технический университет)»

Н.Ю. Энатская
Е.Р. Хакимуллин

Математическая статистика

Утверждено Редакционно-издательским советом
института в качестве учебного пособия

Москва 2011

УДК 519
Э-61

Рецензенты: д. т. н. Е. Е. Тимонина (Российский государственный гуманитарный университет);
д. т. н., д. ф. н. В. А. Глазунов (Институт машиноведения им. А. А. Благонравова РАН)

Энатская Н. Ю., Хакимуллин Е. Р.
Э-61 Математическая статистика. Учебное пособие – Московский государственный институт электроники и математики. М., 2011. – 118с.
ISBN 978-5-94506-306-8

В первой части пособия рассмотрены дополнительные вопросы теории вероятностей, необходимые для изучения математической статистики, и начальные сведения по математической статистике.

Во второй части пособия подробно изложены вопросы, связанные с решением одной из основных задач математической статистики - параметрической задачи. Приведено много примеров.

Рекомендуется всем студентам МИЭМа, изучающим математическую статистику.

ISBN 978-5-94506-306-8 УДК 519

Оглавление.

Введение и обозначения

Настоящее пособие написано на основе многолетнего опыта проведения лекционных и практических занятий по математической статистике в МИЭМе, МАТИ, МГПУ. Как оказалось, не представляется возможным рекомендовать в качестве основного ни один из существующих учебников в силу разных причин: из-за большого их объема, из-за сложности изложения, из-за неполноты представленных тем, в то время как они же могут успешно служить в качестве дополнительной литературы. Учебное пособие должно отличаться от других источников информации по данной теме близостью к учебной программе предмета, компактностью, определяемой ограниченностью учебного времени и в то же время понятностью (с комментариями и большим числом примеров), логической законченностью и связанностью излагаемых вопросов. Здесь делается попытка создания такого учебного материала.

I часть.

В первой части рассматриваются некоторые вспомогательные темы теории вероятностей, не вошедшие в основной начальный курс, а также начальные темы математической статистики. Цель первой части – подготовка к восприятию математической статистики путем рассмотрения отдельных вопросов теории вероятностей и обсуждения основных понятий и структуры курса.

II часть.

Во второй части разобраны вопросы, связанные с решением параметрической задачи статистики.

Предполагается знание основного курса теории вероятностей.

Введем обозначения:

с.в. – случайная (ые) величина (ы);

L(x) – закон распределения с.в. X;

B (1, p) – бернуллиевское распределение;

B (n, p) – биномиальное распределение;

$\pi(\lambda)$ – пуассоновское распределение;

H (N, M, n) – гипергеометрическое распределение;

Ge (p) – геометрическое распределение;

сдGe(p) – сдвинутое геометрическое распределение;

πa (r,p) – сдвинутое геометрическое распределение;

OB (r, p) – отрицательное биномиальное распределение;

R [a, b] – равномерное распределение;

N [a, σ] – нормальное распределение;

$\Gamma_{\alpha,\lambda}$ и $\gamma(x;\alpha,\lambda)$ – гамма распределение с плотностью распределения $f(x;\alpha,\lambda)$;

$E(\lambda)$ – экспоненциальное распределение;

смысл параметров смотри в [1];

запись с.в. $X \sim A(\Theta)$ означает, что с.в. X имеет закон распределения A с параметром $\Theta=(\Theta_1,..,\Theta_k)$;
EX – математическое ожидание с.в. X;
DX – дисперсия с.в. X;
ф.п. – функция правдоподобия;
д.с. – достаточная статистика;
к.ф. – критерий факторизации;
ФПВ – формула полной вероятности;
МРК – неравенство Рао-Крамера;
КЭ – критерий эффективности (критерий НОМД)
НКБ – неравенство Коши-Буняковского;
ц.с. – центральная статистика;
д.и. – доверительный интервал;
ц.д.и. – центральный доверительный интервал;
$x_{(k)}$ – k-я порядковая статистика с.в. X;
х.ф. – характеристическая функция;
$\overline{x}$ – выборочное среднее;
э.с. – экспоненциальное семейство распределений;
=> – следует;
$\vec{x} = (x_1,\ldots,x_n)$ – выборка объема n из генеральной совокупности;
нз – независимость (ый, ые).

Часть I

В первой части пособия рассматриваются отдельные темы теории вероятностей, необходимые для изучения математической статистики, а также начальные сведения по математической статистике, включая решение непараметрической задачи и моделирование случайных последовательностей с заданными распределениями.

§1. Дополнительные сведения о распределениях

В общем курсе теории вероятностей рассматривался вопрос о законах распределения случайных величин наиболее распространённых простейших распределений. Здесь будут изучаться некоторые другие, более сложные распределения, широко используемые в математической статистике. Кроме непосредственных вычислений будем при этом пользоваться методом характеристических функций.

Изложение темы будет представлено в основном в виде задач с решениями с приведением сведений по изучаемым распределениям.

1. Гамма-распределение

Плотность гамма-распределения имеет вид

$$f(x;\, \alpha,\, \lambda) = \begin{cases} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha), & x \geq 0 \\ 0, & x<0 \end{cases},$$

где $\alpha, \lambda - \text{const}>0$ и $\Gamma(\alpha) = \int\limits_0^{\infty} x^{\alpha-1} e^{-x} dx$.

Свойства $\Gamma(\alpha)$.

$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$, при целом α: $\Gamma(\alpha+1) = \alpha!$; $\Gamma(0)=1$; $\Gamma(1)=1$.

Пусть запись $X \sim \Gamma_{\alpha,\lambda}$ означает, что случайная величина (с. в.) X имеет плотность распределения $f(x;\, \alpha,\, \lambda)$.

Задача 1.

Показать, что $f(x;\, \alpha,\, \lambda)$ является плотностью распределения.

Решение.

$f(x;\, \alpha,\, \lambda) \geq 0$. Остаётся проверить условие нормировки:

$$\int_0^\infty f\,(x;\alpha,\lambda)\,dx=\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx=\left\{\lambda x=y, dx=\frac{dy}{\lambda}\right\}=\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty (y/\lambda)^{\alpha-1}e^{-y}\frac{dy}{\lambda}=$$

$$=\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{1}{\lambda^\alpha}\int_0^\infty y^{\alpha-1}e^{-y}dy=\frac{\lambda^\alpha}{\Gamma(\alpha)}\frac{1}{\lambda^\alpha}\Gamma(\alpha)=1.$$

Задача 2.

Найти характеристическую функцию гамма-распределения.

Решение.

$$g_x(t)=g(t;a,\lambda)=\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-(\lambda-it)x}dx=\left\{\lambda-it=b, bx=y, dx=\frac{dy}{b}\right\}=$$

$$=\frac{\lambda^\alpha}{\Gamma(\alpha)b^\alpha}\int_0^\infty y^{\alpha-1}e^{-y}dy=\frac{\lambda^\alpha}{b^\alpha}=\frac{\lambda^\alpha}{(\lambda-it)^\alpha}=\left(1-\frac{it}{\lambda}\right)^{-\alpha}.$$

Задача 3.

Случайные величины X_1 и X_2 : независимы $X_1 \sim \Gamma_{\alpha,\lambda}$, $X_2 \sim \Gamma_{\alpha,\lambda}$. Найти закон распределения с.в. $Z= X_1 + X_2$.

Решение.

Решать задачу будем методом характеристических функций:

$g_z(t)= g_z(t;\alpha,\lambda)= g_x(t;\alpha_1,\lambda)\, g_x(t;\alpha_2,\lambda)=(1-it/\lambda)^{-(\alpha_1+\alpha_2)}$, откуда следует, что $X_1+ X_2 =Z$ есть гамма-распределение с плотностью $f\,(x;\lambda,\alpha_1+\alpha_2)$ (обозначим это распределение через $\Gamma_{\alpha,\lambda}$, где $\alpha=\alpha_1+\alpha_2$).

Задача 4.

Доказать, что, если с. в. X распределена по $\Gamma_{\alpha,\lambda}$, а с.в. $Y= \lambda X$, то с.в. Y распределена по $\Gamma_{1,\lambda}$.

Решение.

$f\,(x;\alpha=1,\lambda) = \lambda^\alpha x^{\alpha-1}e^{-\lambda x}/\Gamma(\alpha)$, $x\geq 0$; $\alpha,\lambda>0$;

$F_Y(x)=P(X<x/\lambda)=F_X(x/\lambda)$;

$$f_Y(x)=\frac{1}{\lambda}f_X\left(\frac{x}{\lambda}\right)=\frac{1}{\lambda}f\left(\frac{x}{\lambda};\alpha,\lambda\right)=\frac{1}{\lambda}\frac{\lambda^\alpha x^{\alpha-1}e^{\frac{-\lambda x}{\lambda}}}{\lambda^{\alpha-1}\Gamma(\alpha)}=\frac{e^{-x}x^{\alpha-1}}{\Gamma(\alpha)},\ x\geq 0,\ \lambda\geq 0,$$

т.е. $f_Y(x)=f(x;\alpha,1)$.

Задача 5.

Показать, что экспоненциальное распределение является частным случаем гамма распределения и выписать характеристическую функцию для экспоненциального распределения.

Решение.

$$f(x;\alpha{=}1,\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0, & x<0\Rightarrow g(t,1,\lambda)=\left(\dfrac{1-it}{\lambda}\right)^{-1}\end{cases}\quad —$$

характеристическая функция экспоненциального распределения.

Найдём моменты гамма-распределения. Воспользуемся известной связью характеристической функции с моментами с.в. X: $i^n MX = g^{(n)}(0)$. Тогда, если с.в. X~ $\Gamma_{\alpha,\lambda}$, то

$$EX=-g(0,\alpha,\lambda)=(-i)\left(\left(1-\frac{it}{\lambda}\right)^{-\alpha}\right)'\Bigg|_{t=0}=\alpha\left(1-\frac{it}{\lambda}\right)^{-\alpha-1}\left(-\frac{i}{\lambda}\right)\Bigg|_{t=0}=\frac{\alpha i}{\lambda}(-i)=\frac{\alpha}{\lambda}\Rightarrow$$

$$\Rightarrow EX=\frac{\alpha}{x};\ g''(0)=\frac{\alpha i}{\lambda}\left(1-\frac{it}{\lambda}\right)^{-\alpha-2}-(\alpha+1)\left(-\frac{i}{\lambda}\right)=-\frac{\alpha(\alpha+1)}{\lambda^2};\quad g''(0)=EX^2;$$

$$DX=EX^2-(EX)^2=-g''(0)\text{- }(EX)^2=\frac{\alpha(\alpha+1)}{\lambda^2}-\frac{\alpha^2}{\lambda^2}=\frac{\alpha}{\lambda^2}.$$

Итак, $EX=\dfrac{\alpha}{\lambda}$, $DX=\dfrac{\alpha}{\lambda^2}$.

Для экспоненциального распределения (с.в. $X\sim E(\lambda)$) получим отсюда моменты при $\alpha=1$: $EX=\dfrac{1}{\lambda}$, $EX^2=\dfrac{2}{\lambda^2}$, $DX=\dfrac{1}{\lambda^2}$.

2. Распределение χ^2.

Задача 1.

Найти распределение с.в. $\chi_n^2=\sum_{i=1}^{n}X_i^2$, где с.в. распределена по нормальному закону N(0,1) и все X_i независимы, $i=\overline{1,n}$.

Решение.

Найдём сначала распределение с.в. $\chi_1^2=X_1^2$, то есть (X_1^2<x)=P($|X_1|<\sqrt{x}$)= P (-$\sqrt{x}<X_1<\sqrt{x}$)=

$$=\frac{1}{2\sqrt{\pi}}\int_{-\sqrt{x}}^{\sqrt{x}}e^{\frac{-y^2}{2}}dy=\frac{2}{\sqrt{2\pi}}\int_0^{x}e^{\frac{-y^2}{2}}dy=\left\{y=u,dy=\frac{du}{2\sqrt{u}}\right\}=\frac{2}{\sqrt{2\pi}}\int_0^{x}e^{\frac{-u^2}{2}}\frac{du}{2\sqrt{u}}=$$

$$=\frac{1}{\sqrt{2\pi}}\int_0^{x}e^{\frac{-u^2}{2}}u^{\frac{1}{2}}du.$$

Отсюда следует, что плотность распределения с.в. χ_1^2 есть частный вид плотности гамма-распределения, т.к.:

$$f_{\chi_1^2}(x)=\frac{e^{\frac{-x}{2}}x^{\frac{1}{2}}}{\sqrt{2\pi}}=f\left(x;\frac{1}{2};\frac{1}{2}\right);\ \Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}.$$

Характеристическая функция с.в. χ_1^2 есть $g\left(t,\frac{1}{2},\frac{1}{2}\right)$ = $g_{\chi_1^2}(t)=$ =$(1–2it)^{-\frac{1}{2}}$.

Тогда по свойствам характеристической функции имеем

$$g_{\chi_n^2}(t)=\left(g_{\chi_1^2}(t)\right)^n=(1-2it)^{-\frac{n}{2}}\Rightarrow \text{с.в.}\chi_n^2\sim\Gamma_{\frac{n}{2},\frac{1}{2}}\Rightarrow f\left(x,\frac{n}{2},\frac{1}{2}\right)=$$

$$=\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}.$$

Найдём моменты распределения с.в. χ_n^2, т.е. найти $E\chi_n^2$ и $D\chi_n^2$.

$$\chi_n^2\sim\Gamma_{\alpha=\frac{n}{2},\lambda=\frac{1}{2}}\Rightarrow\quad E\chi_n^2=\frac{\alpha}{\lambda}=\frac{n\cdot 2}{2}=n;\ D\chi_n^2=\frac{\alpha}{\lambda^2}=\frac{n\cdot 4}{2}=2n.$$

Задача 2.

Найти предельное распределение для с.в. $\chi_{0n}^2=\frac{\chi_n^2-E\chi_n^2}{\sqrt{D\chi_n^2}}$ и с.в. χ_n^2, определённой в задаче 1, при $n\to\infty$.

Решение.

С.в. $\chi_n^2=\sum_{i=1}^{n}X_1^2$, где с.в. X_i распределена по N(0,1) и все с.в. X_i независимые, следовательно, с.в. X_i^2 тоже одинаково распределены, независимы и по центральной предельной теореме (теорема Леви) распределение с.в. $\chi_{0n}^2=\frac{\chi_n^2-E\chi_n^2}{\sqrt{D\chi_n^2}}$ сходится при $n\to\infty$ к нормальному закону N(0,1).

$E\chi_n^2=\frac{\alpha}{\lambda}=\frac{n\cdot 2}{2}=n;\ D\chi_n^2=\frac{\alpha}{\lambda^2}=\frac{n\cdot 4}{2}=2n$, т.е. распределение с.в. $\chi_{0n}^2=\frac{\chi_n^2-n}{\sqrt{2n}}$ при $n\to\infty$ сходится к N(0,1).

$$P\left(\chi_{0n}^2<x\right)_{n\to\infty}\to\left(\frac{1}{\sqrt{2n}}\right)^x e^{\frac{-u^2}{2}}du;$$

$$\chi_n^2 = \sqrt{2n}\,\chi_{0n}^2 + n;$$

$$P(\chi_{0n}^2 < x) = P(\sqrt{2n}\chi_{0n}^2 + n < x) = P\left(\chi_{0n}^2 < \frac{x-n}{\sqrt{2n}}\right)_{n\to\infty} \to \left(\frac{1}{2\pi}\right)^x \int_{-\infty}^{\frac{x-n}{\sqrt{2n}}} e^{-\frac{u^2}{2}}\,du.$$

Задача 3.

С.в. X распределена по закону $N(0,\sigma)$. Найти распределение с.в. $Z = \sum_{i=1}^{n} X_i^2$, где $\{X_i\}$ – независимые с.в. и её моменты.

Решение.

С.в. X_i/σ распределена по нормальному закону N(0,1), с.в. $U = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2$,

тогда $\sum_{i=1}^{n} X_i^2 = U\sigma^2 = Z.$

Найдём характеристическую функцию с.в. Z – $g_z(t)$ и её моменты EZ и DZ.

$g_z(t) = g_{\sigma^2 u}(t) = g_u(\sigma^2 t) = (1-2i\sigma^2 t)^{-\frac{n}{2}}$.

Из сравнения с.в. Z и гамма-распределения следует, что (так как $g_z(\sigma t) =$ $(1-2i\sigma^2 t)^{-\frac{n}{2}} = \left(1 - \frac{it}{1/2\sigma^2}\right)^{\alpha}$) с.в. Z распределена с плотностью распределения

$$f\left(x; \frac{1}{2\sigma^2}, \frac{n}{2}\right) = \begin{cases} \dfrac{\sigma^{-n} 2^{-\frac{n}{2}} e^{-\frac{x}{2\sigma^2}} x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, & x \geq 0 \\ 0, & x < 0. \end{cases};$$

$$EZ = -ig_z'(0) = -i\left(-\frac{n}{2}\right)(1-2i\sigma)^{-\frac{n+2}{2}}\left(-2i\sigma^2\right)\Big|_{t=0} = -\frac{in2i\sigma}{2} = n\sigma^2;$$

$$EZ^2 = g''(0) = i\frac{n}{2}\left(-\frac{n+2}{2}\right)\left(1-2i\sigma^2 t\right)^{-\frac{5n}{2}}\left(-2i\sigma^2\right)^2\Big|_{t=0} = \frac{n(n+2)}{4}4\sigma^4 =$$

$$= n(n+2)\sigma^4;$$

$$DZ = EZ^2 - (EZ)^2 = n^2\sigma^4 + 2n\sigma^4 - n^2\sigma^4 = 2n\sigma^4.$$

3. Распределение Стьюдента

С.в. X имеет распределение Стьюдента, если её плотность распределения имеет вид

$$S_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)\left(\frac{1+x^2}{n}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \ , \ -\infty < x < \infty.$$

Можно показать, с.в. $T = U\sqrt{\frac{n}{g}}$ распределена по закону Стьюдента, если с.в. U распределена по нормальному закону N(0,1), а с.в. g распределена по закону χ_n^2;

$-\infty < g < \infty$; $-\infty < U < \infty$; U, g – независимые с.в.

При n =1 закон Стьюдента называется законом Коши с плотностью распределения

$$f\ (x) = \frac{\Gamma(1)}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\right)\left(1+x^2\right)} = \frac{1}{\pi(1+x^2)}.$$

4. Распределение Фишера

Плотность распределения Фишера есть

$$f_{mn}(t) = \frac{\Gamma\left(\frac{n+m}{2}\right)t^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(t+1\right)^{\frac{m+n}{2}}}, \ t>0.$$

Можно показать, что если с.в. $X_1, X_2, \ldots, X_m$ независимы, $X = \sum_{i=1}^{m} X_i^2$,

$Y_1, Y_2, \ldots, Y_n$ - независимы, $Y = \sum_{j=1}^{n} Y_j^2$; X_i, Y_j и при всех $i=\overline{1,n}$; $j=\overline{1,m}$ распределены нормально по закону $N(0,\sigma^2)$, тогда с.в. $Z=\frac{X}{Y}$ распределена по Фишеру.

Найдём моменты с.в. Z:

$$EZ=\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^\infty t^{\frac{m}{2}+1-1}\frac{dt}{(1+t)^{\frac{m+n}{2}}}=\frac{\Gamma\left(\frac{m+n}{2}\right)\Gamma\left(\frac{m}{2}+1\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m+n}{2}\right)}=$$

$$=\frac{\Gamma\left(\frac{m}{2}+1\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$, так как бета функция $B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$;

$B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ и $B(x,y)=\int_0^\infty t^{x-1}\frac{dt}{(1+t)^{x+y}}$.

В частности, при $i=1$ и $i=2$ имеем

$$EZ=\frac{\Gamma\left(\frac{m}{2}+1\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}=\frac{\frac{m}{2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}-1\right)}=\frac{m}{n-2};$$

$$EZ^2=\frac{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}-2\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}=\frac{\left(\frac{m}{2}+1\right)\left(\frac{m}{2}\right)\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}-2\right)}{\Gamma\left(\frac{m}{2}\right)\left(\frac{n}{2}-2\right)\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}-2\right)}=\frac{\left(\frac{m}{2}+1\right)\left(\frac{m}{2}\right)}{\left(\frac{n}{2}-2\right)\left(\frac{n}{2}-1\right)}=$$

$$=\frac{(m+2)m}{(n-4)(n-2)};$$

$$DZ=EZ^2-(EZ)^2=\frac{(m+2)m}{(n-4)(n-2)}-\frac{m^2}{(n-2)^2}=\frac{2m(n-m)+4m(m-1)}{(n-4)(n-2)^2}.$$

5. Бета-распределение

С.в. Z имеет бета-распределение с параметрами p и q (p,q>0), если её плотность распределения имеет вид

$$f_\beta(x;p,q)=\frac{\Gamma(p+q)x^{p-1}(1-x)^{q-1}}{\Gamma(p)\Gamma(q)},\ x\in[0,1].$$

Задача 1.
Показать, что бета-распределение при p = q = 1 есть равномерное на [0,1] распределение.
Решение.
$f_\beta(x;p,q)\Big|_{p,q=1}=\dfrac{\Gamma(p+q)x^{p-1}(1-x)^{q-1}}{\Gamma(p)\Gamma(q)}\Big|_{p,q=1}=1$ – это плотность равномерного распределения на [0,1] при $x\in[0,1]$.

Задача 2.
Пусть $X_1,\ X_2,\ldots,\ X_m$ - независимы, равномерно распределены на [0,1], а $X_{(1)},\ X_{(2)},\ldots,\ X_{(n)}$ - упорядоченные в порядке возрастания исходные с.в. Показать, что плотность распределения с.в. $X_{(k)}$ имеет бета-распределение, где $X_{(k)}$ - k-ая порядковая статистика.
Решение.
Известно, что плотность распределения k-ой порядковой статистики $f_k(x)$ имеет вид: $f_k(x)=nC_{n-1}^{k-1}f(x)[1-F(x)]^{n-k}$, где f(x) и F(x) – соответственно плотность и функция распределения с.в. X.
В рассматриваемом случае $X\sim R[0,1]$, поэтому
$f_k(x)=nC_{n-1}^{k-1}x^{k-1}[1-x]^{n-k}=\dfrac{\Gamma(n+1)x^{k-1}(1-x)^{n-k}}{\Gamma(k)\Gamma(n-k+1)}$, то есть это – бета-
-распределение с параметрами $p=k$, $q=n-k+1$, $(p+q=k+n-k+1=n+1)$.

Задача 3.
Найти n-ый момент m_n, бета-распределения ($m_n=EZ^n$).
Решение.

$$m_n=\int_0^1 x^n f_\beta(x;p,q)dx=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^1 x^{n+p-1}(1-x)^{q-1}dx=\frac{\Gamma(p+q)\,B(n+p,q)}{\Gamma(p)\Gamma(q)}=$$
$$=\frac{\Gamma(p+q)\Gamma(n+p)\Gamma(q)}{\Gamma(p)\Gamma(q)\Gamma(n+p+q)}=\frac{\Gamma(p+q)\,\Gamma(n+p)}{\Gamma(p)\Gamma(n+p+q)},$$

так как $B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}dx$; (Re p>0, Re q>0); $B(p,q)=\dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, отсюда $m_1=\dfrac{\Gamma(p+q)\Gamma(p+1)}{\Gamma(p)\Gamma(p+q+1)}=\dfrac{p}{p+q}$; $m_2=\dfrac{\Gamma(p+q)\Gamma(p+2)}{\Gamma(p)\Gamma(p+q+2)}=\dfrac{p(p+1)}{(p+q)(p+q+1)}$;

$$DX=m_2-(m_1)^2=\frac{p(p+1)}{(p+q)(p+q+1)}-\frac{p^2}{(p+q)^2}=\frac{pq}{(p+q)^2(p+q+1)}.$$

§2. Условные распределения

Способы задания закона распределения двумерной случайной величины (с.в.) (X,Y).

1) Функция распределения F(x,y).
Определение.
$F(x,y)=P\{X<x, Y<y\}$. (1)
Определяется для дискретных и непрерывных составляющих.
Свойства F(x,y):
1) $0\le F(x,y) \le 1$;
2) F(x,y) – неубывающая функция по каждому аргументу;
3) $F(-\infty,\infty) = F(-\infty,y) = F(x,-\infty) = 0$;
4) $F(\infty,\infty) = 1$;
5) $F(x,\infty) = F_1(x)$, $F(\infty,y) = F_2(y)$, где $F_1(x)$ и $F_2(y)$ – маргинальные (по каждому аргументу) функции распределения.

2) Распределение дискретной с.в. (X,Y) можно задать таблицей распределения:

Y / X	x_1	x_2	…	x_n
y_1	p_{11}	p_{21}	…	p_{n1}
y_2	p_{11}	p_{22}	…	p_{n2}
…	…	…	…	…
y_m	p_{1m}	p_{2m}	…	p_{nm}

Свойства таблицы распределения:
а) для любых i и j $p_{ij} \ge 0$, где $p_{ij} = P\{X = x_i, Y = y_i\}$;
б) $\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1$.

Обозначим $p_i^{(1)} = P(X = x_i)$; $p_j^{(2)} = P(Y = y_j)$, которые при всех i и j задают маргинальные (одномерные) распределения с.в. X и Y соответственно.
По ФВП получаем

$$P\{X=x_i\} = \sum_{j=1}^{m} P(Y = y_j)\cdot P(X = x_i/Y = y_j) = \sum P(X = x_i, Y = y_j) = \sum_{j=1}^{m} p_{ij}; \qquad (2)$$

аналогично получаем

$$P\{Y=y_i\} = \sum_{i=1}^{n} p_{ij}. \qquad (3)$$

3) Распределение непрерывной с.в. (X,Y) можно задать плотностью распределения

$$f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y}; \qquad (4)$$

$f_1(x)=\int\limits_{-\infty}^{\infty} f(x,y)dy$ – маргинальная плотность распределения с.в. X;

$f_2(y)=\int\limits_{-\infty}^{\infty} f(x,y)dx$ – маргинальная плотность распределения с.в. Y;

$$F(x,y)=P(X<x,Y<y)\ ;\ F(x,y)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(u,v)dudv. \qquad (5)$$

2. Условные распределения и условные математические ожидания (у.м.о)

1) В дискретном случае

$$P\{X=x_i/Y=y_i\}=\frac{p_{ij}}{p_j^{(2)}};\ \ E(X/Y=y_i)=\sum_{i=1}^{n} x_i P\{X=x_i/Y=y_i\} \qquad (6)$$

– у.м.о. (при фиксированном условии – это число (const)).
Аналогично определяется $P\{Y=y_i/X=x_j\}$ и $E=(Y/X=x_i)$.

2) В непрерывном случае $\varphi(x/Y=y)=\frac{f(x,y)}{f_2(y)}$, $\psi(Y/X=x)=\frac{f(x,y)}{f_1(x)}$ – условные плотности при фиксированном значении другой координате.
У.м.о. для непрерывной с.в. при фиксированном условии есть

$$E(X/Y=y)=\int\limits_{-\infty}^{\infty} x\varphi(X/Y=y)dx$$

$$E(Y/X=x)=\int\limits_{-\infty}^{\infty} y\psi(Y/X=x)dy$$

– это числа (const).

Если рассматривать значение y с.в. Y как случайное, то E(X/Y) – с.в., зависящая от с.в. Y и называется у.м.о.

Простейшие свойства у.м.о $E(X/Y)$:

1) $E(C/Y) = C$;

2) $E(CX/Y) = CE(X/Y)$; (7)

3) $E(X+Y/Z) = E(X/Z)+E(Y/Z)$.

Если рассматривать значение x и y как возможные текущие значения с.в. X и Y, то $E(X/Y = y) = V_1(y)$, $E(Y/X = x) = V_2(x)$, где V_1 и V_2 - некоторые функции одного переменного. Тогда уравнения регрессии соответственно по x и y есть

$x = E(X/Y = y)$ – среднее значение с.в. X при фиксированном значении с.в. Y=y; (8)

$y = E(Y/X = x)$ – среднее значение с.в. Y при фиксированном значении с.в. X=x.

Если уравнение регрессии есть уравнение прямой (т.е. описывает линейную зависимость), то регрессия называется прямолинейной.

Тема условных распределений имеет важное теоретическое и практической значение, т.к. аппарат условных распределений и у.м.о. используется при решении различных статистических задач, а линии регрессии отражают зависимость между компонентами двумерной с.в. (X,Y).

Пусть одна из компонент двумерной с.в. (X,Y) недоступна для наблюдений, например, компонента Y. Тогда возникает задача прогноза возможных значений с.в. Y по наблюденным значениям с.в. X. Это означает поиск зависимости компонент X и Y: $\varphi(X)$, наилучшим образом приближающей возможные значения с.в. Y. Вид зависимости $\varphi(X)$ ищется, исходя из минимизации выбранной меры близости с.в. $\varphi(X)$ к Y. Наиболее распространенной является мера близости в среднем квадратическом:

$\mu = E(Y-\varphi(X))^2$. Тогда $\varphi(X)$ ищется из условий $\min\limits_{\varphi(X)} \mu = \min\limits_{\varphi(X)} E(Y-\varphi(X))^2$.

Ниже доказывается, что этот минимум достигается на $\varphi(X) = E(Y/X)$.

Свойства линий регрессии:

1) Если X и Y нз с.в., то линии регрессии параллельны координатным осям.

Доказательство.

Из нз с.в. X и Y следует, что

$x = E(X/Y = y) = EX$ прямая параллельна оси 0_y ($\| 0_y$)

или

$y = E(Y/X = x) = EY$ прямая параллельна оси $0x$ ($\| 0_x$)

линий регрессии есть $\begin{cases} x = EX \parallel 0_y \\ y = EY \parallel 0_x \end{cases}$;

2) Если $Y = aX + b$ (a и b – const), то линии регрессии совпадают и имеют уравнение $y = ax + b$

Доказательство.

Значения с.в. X и Y взаимно однозначно определяют друг друга и не являются случайными при задании одного из них, поэтому уравнения регрессии есть

$$\begin{cases} x = E(X/Y = y) = \dfrac{y-b}{a} \\ y = E(Y/X = x) = ax + b \end{cases},$$

а это уравнения одной и той же прямой: $y = ax + b$.

Приведем некоторые важные теоретические сведения по теме в виде следующих задач.

Задача 1.

Доказать, что когда X и Y – независимые с.в., то $f(x, y) = f_1(x)f_2(y)$; $\varphi(X/Y = y) = f_1(x)f_2(y)$; $\psi(Y/X = x) = f_2(y)$; $E(X/Y = y) = EX$; $E(Y/X = x) = EY$.

Решение.

По (6) имеем:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv = \int_{-\infty}^{x} f_1(u)du \int_{-\infty}^{y} f_2(v)dv,$$

откуда следует, что $f(x, y) = f_1(x)f_2(y)$.

$$\varphi(X/Y = y) = \frac{f(x, y)}{f_2(y)} = \frac{f_1(x)f_2(y)}{f_2(y)} = f_1(x).$$

Аналогично $\psi(Y/X = x) = f_2(y)$.

Для нз с.в. X и Y

$$E(X/Y = y) = \int_{-\infty}^{\infty} x\varphi(X/Y = y)dy = \int_{-\infty}^{\infty} xf_1(x)dx = EX.$$

Аналогично $E(Y/X = x) = EY$.

Задача 2.

Доказать, что $E(E(X/Y)) = EX$.

Решение.

а) X и Y – дискретные с.в., тогда, т.к. P(A/B)P(B) = P(AB) при A = ($X = x_i$), B = ($Y = y_j$), имеем

$$E(E(X/Y=y_j))=\sum_j(\sum_i x_i P(X=x_i/Y=y_j))P(Y=y_j)=$$

$$=\sum_j\sum_i x_i P(X=x_i, Y=y_i)=\sum_j\sum_i x_i p_{ij}=\sum_i x_i\sum_j(p_{ij})=\sum_i x_i p_i^{(1)}=EX,$$

где $p_i^{(1)}$ – маргинальное распределение с.в. X $(p_i^{(1)}=P(X=x_i))$.

б) Пусть теперь X и Y – непрерывные с.в., тогда

$$E(E(X/Y=y))=\int_{-\infty}^{\infty}(\int_{-\infty}^{\infty}x\varphi(X/Y=y)dx)f_2 y dy=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(xy)dxdy=$$

$$=\int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f(xy)dydx=\int_{-\infty}^{\infty}xf_1(x)dx=EX.$$

Задача 3.

Доказать, что $\min E(Y-\varphi(X))^2$ достигает на $\varphi(X)=E(Y/X)=m_Y(X)$, т.е. все возможные в указанном выше смысле прогнозируемые значения с.в. Y лежат на линии регрессии $y=E(Y/X=x)$.

Доказательство.

$$E(Y-\varphi(X))^2=E(Y-m_Y(X)+m_Y(X)-\varphi(X))^2=$$

$$=E[(Y-m_Y(X))+(m_Y(X)-\varphi(X))]^2= \qquad (*)$$

$$=E(Y-m_Y(X))^2+E(m_Y(X)-\varphi(X))^2+2E(Y-m_y(X))(m_y(X)-\varphi(X))$$

Если доказать, что (*) = 0, то получаем:

$$E(Y-\varphi(X))^2=E(Y-m_Y(X))^2+E(m_Y(X)-\varphi(X))^2\Rightarrow$$

$$\Rightarrow E(Y-\varphi(X))^2\ge E(Y-m_Y(X))^2$$

равенство достигается тогда, когда $\varphi(X)=m_Y(X)$, что и утверждалось.

Остается доказать, что $E(Y-m_Y(X))(m_Y(X)-\varphi(X))=0$.

Напомним свойства у.м.о.:

$$E(C/X)=C \qquad \text{(А)}$$

$$E(CX/Y)=CE(X/Y)=Cm_Y(X) \qquad \text{(Б)} \qquad (7)$$

$$E(X+Y/Z)=E(X/Z)+E(Y/Z) \qquad \text{(В)}$$

$$E(E(X/Y))=EX \qquad \text{(Г)}$$

Воспользуемся свойствами у.м.о. (7) и утверждением задачи 2.

Представим
$$E(X + m_Y(X))(m_Y(X) - \varphi(X)) \underset{(\Gamma)}{=} E\{E[(Y - m_Y(X))(m_Y(X) - \varphi(X))/X]\}$$
Рассмотрим
$$E[(Y - m_Y(X))(m_Y(X) - \varphi(X))/X] \underset{(Б)}{=} (m_Y(X) - \varphi(X))E(Y - m_Y(X)/X)$$
Покажем, что $E(Y - m_Y(X)/X) = 0$ по свойствам А, Б и В для у.м.о. имеем $E(Y - m_Y(X)/X) = m_Y(X) - m_Y(Y)$, что и завершает доказательство утверждения.

Формула полного математического ожидания (ФПМО)

Пусть X – дискретная с.в. с возможными значениями $x_1,\ldots,x_m$; $\{B_k\}\ (k = \overline{1,n})$ – полная группа несовместных событий. Тогда
$$E(X/B_k) = \sum_{i=1}^{m} x_i P(x_i/B_k)$$
$$EX = \sum_{k=1}^{n} P(B_k)E(X/B_k)E(X/B_k)$$
– это ФПМО, отсюда следует, что $EX = E(E(X/B_k))$.

Задача.

$Z = \sum_{i=1}^{y} X_i$, где Y – целочисленная с.в.; $\{X_i\}$ и Y – независимые с.в., с.в. $\{X_i\}$ – одинаково распределены. Найти EZ, зная $EX_i = m_x$ и $EY = m_y$.

Решение.

Положим $B_k = \{Y = k\}$, $k = \overline{1,n}$. Тогда по ФПМО
$$EZ = \sum_{k=1}^{m} P(Y = k)E(\sum_{i=1}^{y} x_i/Y = k) = \sum_{k=1}^{n} P(Y = k)km_x =$$
$$= m_x \sum_{k=1}^{n} kP(Y = k) = m_x m_y.$$

Замечание. Такой же результат в данной задаче был получен ранее методом производящих функций.

Задачи по теме

Задача 1.

Дано совместное распределение с.в. X и Y. Найти маргинальные распределения с.в. X и Y, EX, EY, условные распределения с.в. X при $Y = y_j$ (j=1,2); E(X/y) и F(x,y).

Y \ X	2	5	8
0.4	0.15	0.30	0.35
0.8	0.05	0.12	0.03

(Проверка: 0.15 + 0.30 + 0.35 + 0.05 + 0.12 + 0.03 = 1)

Решение.

По (2) и (3):

X	2	5	8
P	0.2	0.42	0.38

;

Y	0.4	0.8
P	0.8	0.2

EX = $2 \cdot 0.2 + 5 \cdot 0.42 + 8 \cdot 0.38 = 5.54$:

EY = $0{,}4 \cdot 0.8 + 5 \cdot 0.8 + 8 \cdot 0.2 = 0.48$.

По (4) $P(X = 2/Y = 0{,}4) = \frac{0.15}{0.8} = \frac{3}{16}$;

$P(X = 5/Y = 0{,}4) = \frac{0.3}{0.8} = \frac{3}{18}$;

$P(X = 8/Y = 0{,}4) = \frac{0.35}{0.8} = \frac{7}{16}$;

Аналогично получаем: $P(X = 2/Y = 0{,}8) = 0.25$;

$P(X = 5/Y = 0{,}8) = 0.6$;

$P(X = 8/Y = 0{,}8) = 0.15$.

$E(X/Y = 0{,}4) = 2 \cdot \frac{3}{16} + 5 \cdot \frac{3}{8} + 8 \cdot \frac{7}{16} = 5.75$;

$E(X/Y = 0{,}8) = 2 \cdot 0.25 + 5 \cdot 0.6 + 8 \cdot 0.15 = 4.7$.

Отсюда видно, что $E(X/Y)$ есть с.в. с рядом распределения:

E(X/Y)	4.7	5.75
P(Y)	0.2	0.8

Поэтому $E(E(X/Y)) = 4.7 \cdot 0.2 + 5.75 \cdot 0.8 = 5.54$.

Задача 2.

В счетчик направлен поток космических частиц. Число ξ_1 попавших в счетчик частиц в течение времени t распределено по закону Пуассона $\pi(\lambda t)$ ($\lambda > 0$). Каждая частица, независимо от других, может быть зарегистрирована с вероятностью p. ξ_2 – число зарегистрированных частиц за время t. Найти условное распределение с.в. ξ_1 при условии, что $\xi_2 = j$ и $E(\xi_1/\xi_2 = j)$.

Решение.

$$P(\xi_1 = k/\xi_2 = j) = \frac{P\{\xi_1 = k, \xi_2 = j\}}{P\{\xi_2 = j\}} \qquad ;$$

$$P(\xi_1 = k/\xi_2 = j) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}C_k^j p^j q^{k-j} = e^{-\lambda t}\frac{(\lambda tq)^k}{j!(k-j)!}\left(\frac{p}{q}\right)^j \qquad ;$$

$$P\{\xi_1 = j\} = \sum_{i=j}^{x} P\{\xi_1 = i, \xi_2 = j\} =$$

$$= e^{-\lambda t}\sum_{i=j}^{x}\frac{(\lambda tq)^i}{j!(i-j)!}\left(\frac{p}{q}\right)^j (\lambda tq)^i \sum_{i=j}^{x}\frac{(\lambda tq)^{i-j}}{(i-j)!} =$$

$$= \frac{e^{-\lambda t(1-q)}}{j!} = \frac{p^j(\lambda t)^j q^j}{q^j} = \frac{e^{-\lambda tp}(\lambda tp)^j}{j!} \Rightarrow \xi_2 \sim \pi(\lambda tp).$$

$$P(\xi_1 = k/\xi_2 = j) = \frac{e^{-\lambda t}(\lambda tq)^k p^j j!}{j!(k-j)!q^j e^{-\lambda tp}(\lambda t)^j p^j} =$$

$$= \frac{e^{-\lambda tq}(\lambda tq)^k}{(k-j)!(\lambda tq)^j} = \frac{e^{-\lambda tq}(\lambda tq)^{k-j}}{(k-j)!} \Rightarrow$$

условное распределение с.в. ξ_1 при $\xi_2 = j$ есть $\pi(\lambda tq)$.

$$E(\xi_1/\xi_2 = j) = \sum_{k=j}^{\infty} kP(\xi_1 = k/\xi_2 = j) = e^{-\lambda tq}\sum_{k=j}^{\infty}\frac{e^{-\lambda tq}(\lambda tq)^{k-j}}{(k-j)!}k =$$

$$= \left|k = (k-j) + j\right| = e^{-\lambda tq}\sum_{k=j}\frac{(\lambda tq)^{k-j}}{(k-j-1)!} + e^{-\lambda tq}\cdot j\sum_{k=j}\frac{(\lambda tq)^{k-j}}{(k-j)!} =$$

$$= e^{-\lambda tq}\lambda tq e^{\lambda tq} + e^{-\lambda tq}\cdot j e^{\lambda tq} = \lambda tq + j.$$

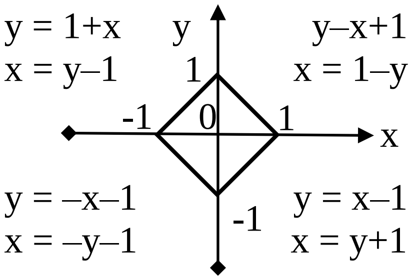
Задача 3.

Случайная точка (X,Y) распределена равномерно в области D, т.е. R[D]. Найти $f(x, y)$, $f_1(x)$, $f_2(e)$, $\varphi(x/y)$, $\psi(y/x)$ и уравнения регрессий.

Решение.

$$f(x,y)=\begin{cases}\frac{1}{2},(x,y)\in D\\ 0,(x,y)\notin D\end{cases};\quad f_1(x)=\begin{cases}0, x\le -1\\ 1+x,-1<x\le 0\\ 1-x, 0<x\le 1\\ 0, x>1\end{cases};$$

$$f_2(y)=\begin{cases}0, y\le 1\\ 1+y,-1<y\le 0\\ 1-y, 0<y\le 1\\ 0, y>1\end{cases}.$$



$$\varphi(x/y)=\frac{f(x,y)}{f_2(y)}=\begin{cases}0, y\le -1\\ \frac{1}{2(1+y)},-1<y\le 0\\ \frac{1}{2(1+y)},0<y\le 1\\ 0, y>1\end{cases};\quad \psi(y/x)=\begin{cases}0, x\le -1\\ \frac{1}{2(1+x)},-1<x\le 0\\ \frac{1}{2(1+x)},0<x\le 1\\ 0, x>1\end{cases}$$

$$x=E(X/y)=\int\limits_{-\infty}^{\infty}x\varphi(x/y)dx=\left\{\begin{matrix}\frac{1}{2(1+y)}\left.\frac{x^2}{2}\right|_{-y-1}^{y+1}=0, y\le 0\\ \frac{1}{2(1-y)}\left.\frac{x^2}{2}\right|_{y-1}^{1-y}=0, y\ge 0\end{matrix}\right\}=0;$$

т.е. $x=0$;

$$y=E(Y/x)=\int\limits_{-\infty}^{\infty}y\psi(y/x)dy=\left\{\begin{matrix}\frac{1}{2(1+x)}\left.\frac{y^2}{2}\right|_{-x-1}^{x+1}=0, x\le 0\\ \frac{1}{2(1-x)}\left.\frac{y^2}{2}\right|_{x-1}^{1-x}=0, x\ge 0\end{matrix}\right\}=0;$$

т.е. $y=0$.

Уравнения регрессии: x = 0, y = 0, т.е. линии регрессии – координатные оси.

Задача 4.

Случайная точка распределения на плоскости с плотностью (это двумерная нормальная плотность):

$$f(x,y)=\frac{1}{2\pi\pi_1\sigma_2\sqrt{1-r^2}}\cdot$$

$$\cdot\exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2}-\frac{2r(x-m_1)(x-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}.$$

Найти уравнения линий регрессии, где r – коэффициент корреляции, $m_1=EX$, $m_2=EY$, $\sigma_1^2=DX$, $\sigma_2^2=DY$.

Решение.

$$f_1(x)=\int_{-\infty}^{\infty}f(x,y)dy=\left\{\frac{x-m_1}{\sigma_1}=u,\frac{y-m_2}{\sigma_2}\right\}=$$

$$=\frac{\sigma_2}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\int_{-\infty}^{\infty}e^{-A/2}dv=$$

$$=\left\{A=\frac{1}{1-r^2}(u^2-2ruv+v^2)=u^2+\frac{(v-ru)^2}{1-r^2}\right\}=$$

$$=\frac{e^{\frac{-u^2}{2}}\sqrt{2(1-r^2)}}{2\pi\sigma_1\sqrt{1-r^2}}\int_{-\infty}^{\infty}e^{-t^2}dt=\frac{1}{\sqrt{2\pi\pi_1}}e^{\frac{-u^2}{2}}=\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left\{-\frac{(x-m_1)^2}{2\sigma_1^2}\right\}.$$

Аналогично получаем:

$$f_2(y)=\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left\{-\frac{(y-m_2)^2}{2\sigma_2^2}\right\};$$

$$\psi(y/x)=\frac{f(x,y)}{f_1(x)}=\frac{1}{\sigma_2\sqrt{2\pi(1-r^2)}}\exp\left\{-\frac{\left(y-m_2-r\frac{\sigma_2}{\sigma_1}(x-m_1)\right)^2}{2(1-r^2)\sigma_2^2}\right\},$$

что соответствует нормальному распределению:

$N\left(m_2+r\frac{\sigma_2}{\sigma_1}(x-m_1);\sigma_2^2(1-r^2)\right)$, откуда следует, что

$$E(y/x)=m_2+r\frac{\sigma_2}{\sigma_1}(x-m_1).$$

Аналогично получаем:

$$\varphi(x/y)=\frac{1}{\sigma_1\sqrt{2\pi(1-r^2)}}\exp\left\{-\frac{(x-m_1-r\frac{\sigma_1}{\sigma_2}(y-m_2))^2}{2(1-r^2)\sigma_1^2}\right\}, \text{ откуда}$$

$$E(x/y)=m_1+r\frac{\sigma_1}{\sigma_2}(y-m_2).$$

Таким образом, имеем уравнения линий регрессии:

$$x=m_1+r\frac{\sigma_1}{\sigma_2}(y-m_2);\ y=m_2+r\frac{\sigma_1}{\sigma_2}(y-m_1).$$

Задача 5.

Рабочий обслуживает n однотипных станков, расположенных в ряд на расстоянии a друг от друга. Рабочий приходит к ним в порядке появления неисправностей. X – длина перехода к следующему, требующему ремонта станку. Найти EX.

Решение.

Пусть событие B_k – рабочий находится у k-ого станка (k = 1,…,n); x_i – возможное значение с.в.; X – длина перехода к 1-ому станку.

$$P\{B_k\}=\frac{1}{n};\quad P\{X=x_i/B_k\}=\frac{1}{n};\ (k=\overline{1,n});$$

$$X=\begin{cases}(k-i)a,\ k\ge i;\\(i-k)a,\ k<i.\end{cases}$$

Тогда по (1) имеем:

$$E(X/B_k)=\sum_{i=1}^{k}(k-i)a\frac{1}{n}+\sum_{i=k+1}^{n}(i-k)a\frac{1}{n}=$$

$$=\frac{a}{n}\left[\frac{k(k-1)}{2}+\frac{(n-k)(n-k-1)}{2}\right]=\frac{a}{2n}(2k^2-2(n+1)k+n(n+1)).$$

$$EX=\frac{a}{2n^2}\sum_{k=1}^{n}(2k^2-2(n+1)k+n(n+1))=$$

$$=\frac{a}{2n^2}\left(\frac{2n(n+1)(2n+1)}{6}-\frac{2n(n+1)^2}{2}+n^2(n+1)\right)=\frac{a(n^2-1)}{3n}.$$

Задача 6.
С.в. $X \sim R[0,1]$; при $X = x$ с.в. $Y \sim R[x,1]$. Найти $f_1(x)$, $f_2(y)$, $\varphi(x/y)$, $\psi(y/x)$ и линии регрессии.

Решение.

$$f_1(x) = \begin{cases} 1, x \in [0,1] \\ 0, x \notin [0,1] \end{cases}; \quad \psi(y/x) = \begin{cases} \dfrac{1}{1-x}, x \le y \le 1 \\ 0, y \notin [x,1] \end{cases}$$

$$f(x,y) = f_1(x)\psi(y/x) = \begin{cases} \dfrac{1}{1-x}, x \le y \le 1 \\ 0, \text{в противном случае} \end{cases}$$

$$f_2(y) = \int_0^y f(x,y)dx = \int_0^y \frac{dx}{1-x} dx = -\ln(1-y), \quad y \in [0,1];$$

$$\varphi(x/y) = \frac{f(x,y)}{f_2(y)} = \begin{cases} -\dfrac{1}{(1-x)\ln(1-y)}, 0 \le x \le y \le 1 \\ 0, \text{в противном случае} \end{cases}$$

Регрессия с.в. X на Y: $x = E(x/y)$:

$$x = -\int_0^y \frac{xdx}{(1-x)\ln(1-y)} = \left| \begin{matrix} 1-x = z; dx = -dz \\ 0 \le x \le y, 1 \ge z \ge 1-y \end{matrix} \right| = \frac{1}{\ln(1-y)} \int_{1-y}^1 \frac{(1-z)dz}{z} =$$

$$= \frac{1}{\ln(1-y)} (\ln z - z) \Big|_{1-y}^{1} = -1 - \frac{y}{\ln(1-y)}.$$

Регрессия с.в. Y на X: $y=E(Y/x)$; $\quad y = \int_x^1 \dfrac{ydy}{1-x} = \dfrac{1+x}{2}$.

Задача 7.
С.в. $X_1 \sim R[0,1]$;при $X_1 = x_1$ с.в. $X_1 \sim R[x_1,1], \ldots$; при $X_{n-1} = x_{n-1}$ с.в. $X_n \sim R[x_{n-1},1]$.
Найти EX_n.

Решение.

$$E(X_n/X_{n-1} = x_{n-1}) = \int_{x_{n-1}}^1 x_n f(x_n/x_{n-1})dx_n = \int_{x_{n-1}}^1 \frac{x_n dx_n}{1-x_{n-1}} =$$

$$= \frac{1}{1-x_{n-1}} \frac{x_n^2}{2}\Bigg|_{x_{n-1}}^{1} = \frac{1+x_{n-1}}{2}.$$

Это ясно и сразу, т.к. с.в. $X_n \sim R[x_{n-1},1]$ при $X_{n-1} = x_{n-1}$.

$EX = E(E(X/B_k))$;

X – дискретная с.в. с возможными значениями $x_1,\ldots,x_m$;

$\{B_k\}(k=1,\ldots,n)$ – полная группа несовместимых событий, тогда

$E(X/B_k) = \sum_{i=1}^{m}(x_i/B_k)$, $EX = \sum_{k=1}^{n} P(B_k)\cdot E(X/B_k)$. Отсюда следует:

$EX_n = E(E(X_n/X_{n-1} = x_{n-1}))$, т.е. $EX_n = \frac{1}{2}EX_{n-1} + \frac{1}{2}$ – рекуррентное соотношение $EX_1 = \frac{1}{2}$, откуда

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

$$= \frac{1}{2^3}EX_{n-3} + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} = \ldots = \frac{1}{2^{n-1}}EX_1 + \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{n-1}} =$$

$$= \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^n} = \frac{1}{2}\,\frac{1-\left(\frac{1}{2}\right)^n}{1-\frac{1}{2}} = 1 - \frac{1}{2^n}.$$

§3. Основные понятия математической статистики

Математическая статистика занимается разработкой, научным обоснованием и применением методов обработки данных наблюдений с целью получения информации о распределении изучаемой случайной величины.

Пусть X – наблюдаемая случайная величина (с.в.)

Все значения, которые может принимать с.в. X, называют **выборочным пространством** или **генеральной совокупностью;** n возможных значений с.в. X называются **выборкой** (n – объем выборки):

$\vec{x} = (x_1, x_2, x_3, ..., x_n)$.

Конкретные n наблюденных значений с.в. X называются **реализацией выборки**.

Для того чтобы изучение с.в. по выборке имело смысл, необходимо, чтобы выборка не искажала, а отражала свойства генеральной совокупности. Такая выборка называется **представительной** (репрезентативной). Математически это означает, что элементы выборки должны быть **независимыми, одинаково распределенными с.в. с тем же законом распределения**, что и у изучаемой с.в. X.

Если упорядочить значения выборки по величине $X_{(1)} \le X_{(2)} \le .... \le X_{(n)}$, то ряд $X_{(1)}, X_{(2)}, ..., X_{(n)}$ называется **вариационным рядом** или **рядом порядковых статистик**, а элементы ряда называются **вариантами** или **порядковыми статистиками**.

Элементы вариационного ряда не обладают теми же свойствами, которыми обладала выборка, т.е. они не являются независимыми, одинаково распределенными с.в. и каждый имеет свой закон распределения, не совпадающий с законом распределения исходной с.в. X (доказательство см. ниже).

Проиллюстрируем это на примере крайних порядковых статистик с.в. X с функцией распределения $F(x)$:

$X_{(1)} = \min(x_1, ..., x_n) = U$ и $x_{(n)} = \max(x_1, ..., x_n) = V$; Зависимость их очевидна.

Найдем закон распределения для с.в. U и V. Обозначим: $G_U(x)$ – функция распределения с.в. U и $G_V(x)$ – функция распределения с.в. V;

$G_U(x) = P\{U<x\} = P\{\min(x_1,....,x_n)<x\}$ = (сформулируем эквивалентное событие в терминах элементов выборки, переходя к дополнительному событию, и учитывая свойства элементов выборки) $= 1 - P\{x_i \ge x\} =$

$$= 1 - \prod_{i=1}^{n} P\{x_i \ge x\} = 1 - (1 - F(x))^n .$$

Найдем плотность распределения с.в. U есть $\mathbf{g_U(x) = G_u'(x) =}$
$\mathbf{= n(1-F(x))^{n-1} f_0(x)}$.

Аналогичными рассуждениями найдем закон распределения с.в. V:

$G_V(x) = P\{x_{(n)} < x\} = P\{x_1 < x, ..., x_n < x\} = \prod_{i=1}^{n} F(x)$. Тогда плотность распределения с.в. V есть $\mathbf{g_v(x) = G_V(x) = nF_0(x)^{n-1} f(x)}$.

Статистика – это любая функция от наблюдений, не зависящая от неизвестных параметров распределения.

Семейство распределений $F_\theta(x)$ – множество распределений одного аналитического вида со всеми возможными значениями параметра
$\theta = (\theta_1, \ldots, \theta_k)$.

Основные задачи математической статистики

Непараметрическая задача статистики

Это первоначальная (грубая) обработка данных вида: наблюдаемая с.в. X и реализация выборки $\vec{x} = (x_1,....,x_n)$ с целью получения информации о виде распределения с.в. X, т.е. нахождения семейства распределений, к которому оно относится. В результате решения задачи получается, что $L(X) \in F_\theta(x)$, где $L(X)$ – распределение с.в. X, а $\theta = (\theta_1,....,\theta_k)$ – неизвестный параметр распределения с.в. X.

Параметрическая задача статистики

Эта задача решается после решения непараметрической задачи (когда семейство распределений $F_\theta(x)$ для с.в. X уже определено) и состоит в уточнении значения параметра θ или функции $\tau(\theta)$ от этого параметра в общем случае, где $\theta = (\theta_1,....,\theta_k)$, т.е. в построении оценки $t(\vec{x}) = \hat{\tau}$ функции $\tau(\theta) = \tau$.

При решении этой задачи различают два подхода: точечное и доверительное оценивание.

1) Точечное оценивание

Суть точечного оценивания в том, что для $\tau(\theta)$ строится одна статистика $t(x) = \hat{\tau}$, которая принимается за оценку $\tau(\theta)$, т.е. $t(x) = \hat{\tau}$.

«Хорошей» оценкой является такая оценка, которая наиболее близка к истинному значению $\tau(\theta)$, т.е. когда ее значения в каком-то смысле сконцентрированы вокруг истинного значения $\tau(\theta)$.

Математически это означает желательность следующих свойств оценки:

а) несмещенность

$Et(x) = \tau(\theta)$, где $Et(x)$ – среднее значение с.в. X (при малых выборках это требование очень важно, при больших n достаточно выполнения свойства асимптотической несмещенности: $Et(x) \neq \theta$, но $Et(x) \to \theta$, при $n \to \infty$;

б) состоятельность

$t(x) \xrightarrow[n\to\infty]{} \tau(\theta)$, что означает $P\{|t(x) - \tau(\theta)| > \varepsilon\} \xrightarrow[n\to\infty]{} 0$;

в) эффективность, которая определяется только для несмещенных оценок и определяется **дисперсией** $Dt(x)$. Тогда значение дисперсии несмещенной оценки называется ее **эффективностью**, которая используется при сравнении качества несмещенных оценок: та оценка лучше, у которой дисперсия меньше, т.е. эффективность больше.

Оценка называется **оптимальной**, если она несмещенная и имеет минимально возможную дисперсию.

2) Доверительное оценивание

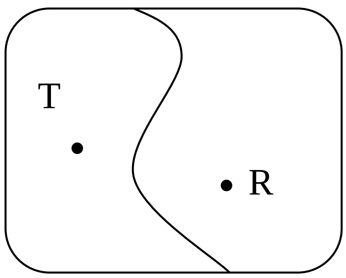
Ограничимся сначала рассмотрением этого подхода в случае одномерного параметра. Тогда решение состоит в построении двух статистик: $t_1 = t_1(\vec{X})$ и $t_2 = t_2(\vec{X})$ таким образом, чтобы $P\{t_1(\vec{X}) < \tau(\theta) < t_2(\vec{X})\} = \gamma$, где γ – некоторое значение, называемое доверительной вероятностью, а интервал $I = (t_1(x),\ t_2(x))$ – доверительным интервалом (этот интервал желательно строить наиболее коротким). Тогда γ – надежность оценивания, а длина интервала $(t_1(x),\ t_2(x))$ определяет точность оценивания, поэтому при заданном γ этот интервал желательно строить более коротким.

В случае бо́льшей размерности неизвестного параметра аналогично определяется доверительная область, содержащая истинное значение $\tau(\theta)$.

Проверка статистических гипотез

Это проверка предположений (гипотез) о распределении изучаемой с.в. Если гипотеза H_0 описывает одно распределение, то она называется **простой**, а если несколько, то **сложной**. Та гипотеза H_0, которую нужно проверить, называется **нулевой** или **основной**; H_1 – **альтернативная гипотеза.**

Чтобы проверить гипотезу, нужно построить критерий, т.е. правило, по которому нулевая гипотеза принимается или отвергается. Это значит, что выборочное пространство нужно разделить границей на две зоны: T – область принятия гипотезы (основной гипотезы H_0) и R – критическая область, где она отвергается.



Пусть X – изучаемая с.в., $x_1, \ldots, x_n$ – выборка ее наблюдаемых значений. Если точка попадает в T, то гипотеза принимается, если точка попадает в R, то гипотеза отвергается.

Как выбрать границу наилучшим образом?

Это делают на основе анализа возможных ошибок. При решении могут возникнуть ошибки двух родов.

Пусть α – вероятность ошибки 1$^{\text{го}}$ рода (это вероятность того, что отвергли верную гипотезу (H_0): $\alpha = P\{x \in R/H_0\}$), β – вероятность ошибки

$2^{го}$ рода (это вероятность того, что приняли неверную гипотезу (H_0): $\beta = P\{\vec{X} \in R/H_1\}$).

α называется ещё уровнем значимости или размером критерия.

Оказывается, что одновременно вероятности ошибок $1^{го}$ и $2^{го}$ рода (α и β) невозможно сделать сколь угодно маленькими. Поэтому два критерия, если они сравнимы, нужно сравнивать по качеству, определяемому вероятностями ошибок α и β следующим образом: фиксируется вероятность ошибки $1^{го}$ рода α для первого и второго критериев, по этому значению α строится граница и вычисляется вероятность ошибки $2^{го}$ рода β. Тогда тот критерий лучше, у которого меньше вероятность ошибки $2^{го}$ рода β.

Говорят, что такой критерий более мощный: $W = 1 - \beta$ – мощность критерия.

Задача.

Пусть проверяют партии деталей, упакованных по 1000 штук. Партию считают хорошей, если при контроле 20 наугад выбранных из нее деталей бракуют $X \le 1$ штуки.

Пусть гипотеза H_0 состоит в том, что партия хорошая, а гипотеза H_1 – партия плохая. Тогда $\alpha = P(X>1/\ H_0)$ – риск заказчика; $\beta = P(X \le 1/\ H_1)$ – риск изготовителя.

§4. Порядковые статистики

Пусть X – изучаемая случайная величина (с.в.), $\vec{x} = x_1, \ldots, x_n$ – выборка наблюдений над с.в. X объёма n. Тогда если упорядочить по возрастанию элементы выборки : $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, то последовательность

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ называется вариационным рядом, а его элементы – порядковыми статистиками (или вариантами) с соответствующими номерами ($X_{(i)}$ – $i^{ая}$ порядковая статистика). $X_{(1)}, X_{(n)}$ – крайние порядковые статистики.

Порядковые статистики широко используются при решении многих задач статистики, поэтому должны быть заранее изучены в первую очередь по следующим направлениям: установление связи распределения изучаемой с.в. X с законами распределений порядковых статистик разных размерностей; нахождение выражений моментов порядковых статистик через закон распределения с.в. X; изучение закона распределения размаха выборки ($W = X_{(n)} - X_{(1)}$). Все полученные здесь формулы по указанным направлениям ниже проиллюстрированы на примере равномерного распределения.

1. Основные формулы для порядковых статистик

Через $F(x)$ и $f(x)$ будем обозначать соответственно функцию распределения и плотность распределения изучаемой с.в. X, а через $F_{(k)}(x)$ и $f_{(k)}(x)$ – соответственно функцию распределения и плотность распределения $k^{ой}$ порядковой статистики.

1.1) Законы распределения крайних порядковых статистик $X_{(1)}, X_{(n)}$.

а)$F_{(1)}(x) = P(X_{(1)}<x) = 1 - P(x_{(1)} \ge x) = 1 - P(x_1 \ge x, \ldots, x_n \ge x) = 1 - \prod\limits_{i=1}^{n} P(x_i \ge x) =$

$= 1 - (P(x_1 \ge x))^n = 1 - (1 - F(x))^n$;

$f_{(1)}(x) = F'_{(1)}(x)$, т.е. $f_{(1)}(x) = n(1 - F(x))^{n-1} f(x)$.

б) $F_{(n)}(x) = P(X_{(n)} < x) = P(x_1 < x, \ldots, x_n < x) = \prod\limits_{i=1}^{n} P(x_i < x) = F^n(x)$;

$f_{(n)}(x) = nF^{n-1}(x)f(x)$.

1.2) Закон распределения k-ой порядковой статистики $X_{(k)}$.

$f_{(k)}(x)$ – плотность распределения $k^{ой}$ порядковой статистики.

Тогда $f_{(k)}(x)dx$ – элемент вероятности для $k^{ой}$ порядковой статистики, т.е. $f_{(k)}(x)dx = P(X_{(k)} \in (x ; x + dx)) = P(A)$, где $A = \{X_{(k)} \in (x, x + dx)\}$.

Проинтерпретируем событие A в терминах элементов выборки в соответствии с рисунком 1.

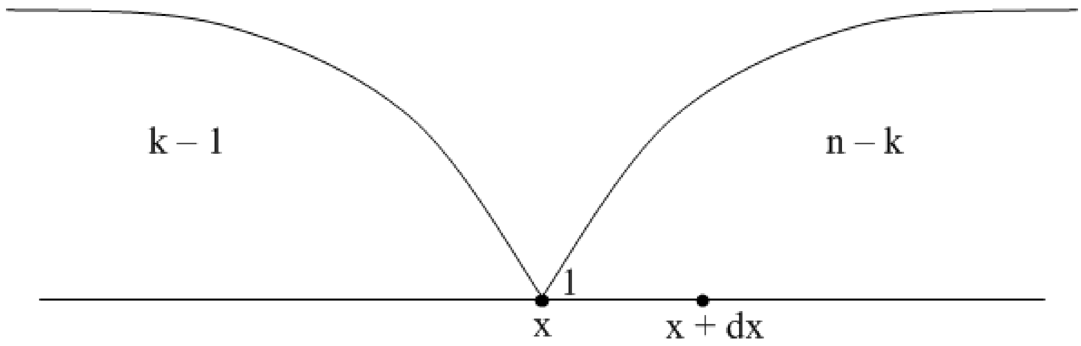


Рис.1

Событие А состоит в одновременном появлении следующих трёх событий:

	с вероятностью	числом вариантов
1. какое-то наблюдение $\in(x;x+dx)$;	$f(x)dx$	n
2. какие-то (k–1) остальных наблюдений левее x;	$F^{k-1}(x)$	C_{n-1}^{k-1}
3. остальные (n–k) правее x.	$(1-F(x))^{n-k}$	1

Тогда $f_{(k)}(x)dx = P(x_{(k)}\in(x;\, x + dx)) = nC_{n-1}^{k-1}f(x)dxF^{k-1}(x)(1-F(x))^{n-k}$.
Отсюда плотность распределения $k^{ой}$ порядковой статистики:
$f_{(k)}(x) = nC_{n-1}^{k-1}f(x)F^{k-1}(x)(1-F(x))^{n-k}$.

Замечание 1. Если в эту формулу подставить $k = 1$, $k = n$, то получим законы распределения крайних порядковых статистик, найденных в пункте 1.

1.3) Найдём закон совместного распределения k-ой и l-ой порядковой статистик (k < l).

$f_{(kl)}(u,v)dudv = P(X_{(k)}\in(u,u + du),\ X_{(l)}\in(v,v + dv))$ – элемент вероятности с.в. $(X_{(k)}, X_{(l)})$.

Пусть событие $A = \{X_{(k)}\in(u,u + du),\ X_{(l)}\in(v,v + dv)\}$. Переформулируем событие A в терминах элементов выборки в соответствии с рисунком 2.

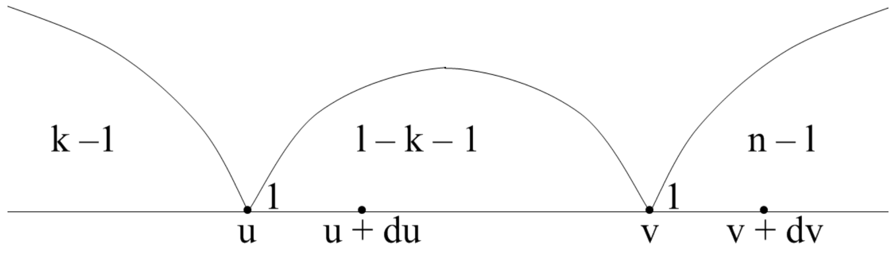


Рис. 2

Событие А состоит в одновременном появлении следующих событий:

	с вероятностью	числом вариантов
1. какое-то одно наблюдение $\in(u;u+du)$;	$f(u)du$	n
2. какое-то одно наблюдение $\in(v;v+dv)$;	$f(v)dv$	$n–1$
3. какие-то $(k–1)$ из оставшихся левее u;	$F^{k-1}(u)$	C_{n-1}^{k-1}
4. какие-то $(l–k–1)$ из оставшихся $\in(u;v)$;	$(F(v) – F(u))^{l-k-1}$	C_{n-k-1}^{l-k-1}
5. оставшиеся $(n–1)$ правее v.	$(1–F(v))^{n-1}$	1

Тогда

$f_{(ki)}(u,v) = n(n-1)C_{n-2}^{k-1}C_{n-k-1}^{l-k-1}f(u)f(v)F^{k-1}(u)(F(v)-F(u))^{l-k-1}(1-F(v))^{n-1}$

Замечание. Подставляя $k = 1$ и $l = n$, получим закон распределения крайних порядковых статистик:

$f_{(1n)}(u,v) = n(n-1)f(u)f(v)(F(v)-F(u))^{n-2}$.

1.4) Закон совместного распределения k-ой, l-ой и m-ой порядковых статистик (k<l<m).

$f_{(klm)}(u,v,c)dudvdc = P(X_{(k)}\in(u;u+du),\ X_{(l)}\in(v;v+dv),\ X_{(m)}\in(c;c+dc))$.

Пусть событие $A = (X_{(k)}\in(u;u+du),\ X_{(l)}\in(v;v+dv),\ X_{(m)}\in(c;c+dc))$. Переформулируем событие A в терминах элемента выборки в соответствии с рисунком 3.

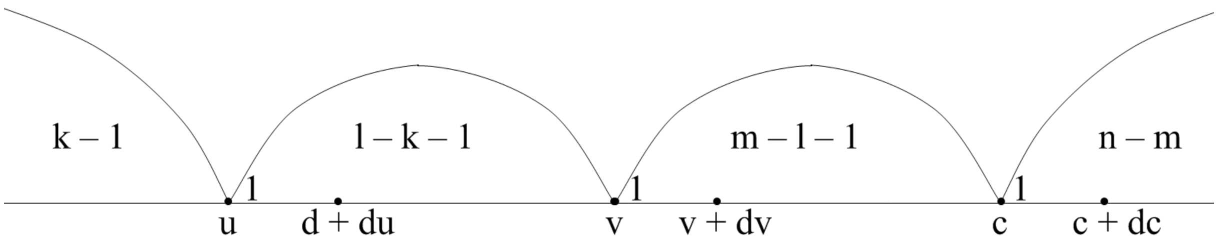


Рис.3

Событие A состоит в одновременном появлении следующих событий:

	с вероятностью	числом вариантов
1. какое-то одно наблюдение $\in(u;u+du)$;	$f(u)du$	n
2. какое-то одно наблюдение $\in(v;v+dv)$;	$f(v)dv$	$n–1$

3. какое-то одно наблюдение $\in(c;c+dc)$;	$f(c)dc$	$n-2$
4. какие-то $(k-1)$ из оставшихся левее u;	$F^{k-1}(u)$	C_{n-3}^{k-1}
5. какие-то $(l-k-1)$ из оставшихся $\in(u;v)$;	$(F(v)-F(u))^{l-k-1}$	C_{n-k-2}^{l-k-1}
6. какие-то $(m-l-1)$ из оставшихся $\in(v;c)$;	$(F(c)-F(v))^{m-l-1}$	C_{n-l-1}^{m-l-1}
7. оставшиеся $(n-m)$ правее c.	$(1-F(c))^{n-m}$	1

Тогда: $f_{(klm)}(u,v,c) = n(n-1)(n-2)C_{n-3}^{k-1}C_{n-k-2}^{l-k-1}C_{n-l-1}^{m-l-1}f(u)f(v)f(w)F^{k-1}(u)(F(v) -$
$- F(u))^{l-k-1}(F(c) - F(v))^{n-l-1}(1 - F(c))^{n-m}$.

1.5) Совместное распределение первых r порядковых статистик ($r \le n$).

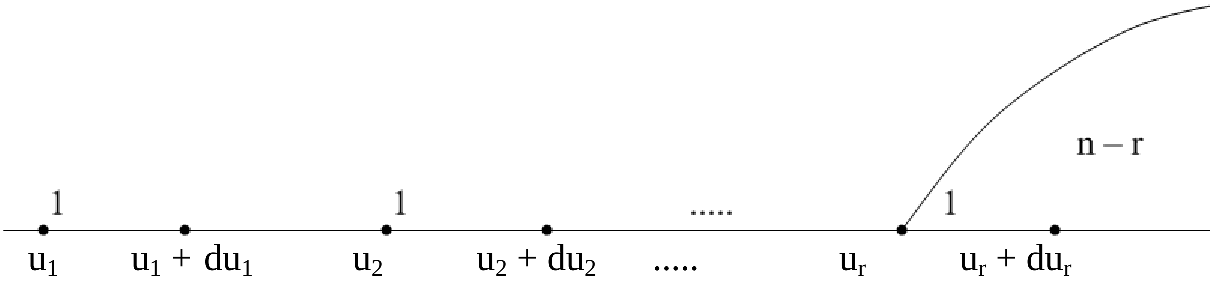


Рис.4

$f_{(12..r)}(u_1, u_2, \dots, u_r) = P(X_{(1)}\in(u_1,u_1+du_1),\ X_{(2)}\in(u_2,u_2+du_2),\dots, X_{(r)}\in(u_r,u_r+du_r))=$
$= P(A)$, где событие A, в соответствии с рисунком 4 , состоит в одновременном появлении следующих событий:

	с вероятностью	числом вариантов
1. какое-то одно наблюдение $\in(u_1;u_1+du_1)$;	$f(u_1)du_1$	n
2. какое-то одно из остальных $\in(u_2;u_2+du_2)$;	$f(u_2)du_2$	$n-1$
…….	…..	…..
r. какое-то одно из остальных $\in(u_r;u_r+du_r)$;	$f(u_r)du_r$	$n-r+1$
r+1. остальные $(n-r)$ наблюдений левее u_r.	$(1-F(u_r))^{n-r}$	1

$f_{(1\ldots r)}(u_1,\ldots,u_r) = n(n-1)(n-2)\ldots(n-r+1)f(u_1)\ldots f(u_r)(1-F(u_r))^{n-r}$ – плотность распределения первых r порядковых статистик.

1.6) Размах выборки.

Назовём $W = X_{(n)} - X_{(1)}$ размахом выборки. Найдём закон распределения размаха выборки H(t):

$$H(t) = P(W<t) = \iint\limits_{0\le v-u\le t} f_{(1n)}(uv)dudv =$$

$$= n(n-1)\int_{-\infty}^{\infty} f(u)du \int_{u}^{u+t} (F(v)-F(u))^{n-2}f(v)dv =$$

$$= n(n-1)\int_{-\infty}^{\infty} f(u)\left(\frac{(F(v)-F(u))^{n-1}}{n-1}\Big|_{u}^{u+t}\right)du = n\int_{-\infty}^{\infty} f(u)(F(u+t)-F(u))^{n-1}du .$$

$H(t) = n\int_{-\infty}^{\infty} f(u)(F(u+t)-F(u))^{n-1}du$ – закон распределения размаха выборки.

1.7) Моменты порядковых статистик.

$$EX_{(k)} = \int_{-\infty}^{\infty} xf_{(k)}(x)dx \quad ; \quad EX_{(k)}^2 = \int_{-\infty}^{\infty} x^2 f_{(k)}(x)dx \; ; \; DX_{(k)} = EX_{(k)}^2 - (EX_{(k)})^2$$

$$K_{X_{(k)}X_{(l)}} = EX_{(k)}X_{(l)} - EX_{(k)}EX_{(l)} \; ; \qquad EX_{(k)}X_{(l)} = \iint_{-\infty}^{\infty} uvf_{kl}(u,v)dudv$$

2.Порядковые статистики равномерного распределения

Пусть с.в. $X\sim R[0,1]$; с.в. $Y\sim R[a,b]$, тогда $Y=(b-a)X + a$

$$f_x(x) = \begin{cases} 0, & x\notin[0,1] \\ 1, & x\in[0,1] \end{cases} ; \qquad f_y(x) = \begin{cases} 0, & x\notin[a,b] \\ 1/(b-a), & x\in[a,b] \end{cases} ;$$

$$F_x(x) = \begin{cases} 0, & x<0 \\ x, & x\in[0,1] \\ 1, & x>1 \end{cases} ; \qquad F_y(x) = \begin{cases} 0, & x<a \\ \dfrac{x-a}{b-a}, & x\in[a,b]; \; 1 \le k \le n. \\ 1, & x > b \end{cases}$$

2.1) Связи распределений и моментов случайных величин $X_{(k)}$ и $Y_{(k)}$:
Очевидно, что $Y_{(k)} = (b - a)X_{(k)} + a$. Тогда $F_{Y_{(k)}}(x) = P\{Y_{(k)} < x\} =$

$$=P\left\{X_{(k)} < \frac{x-a}{x-b}\right\} = F_{X_{(k)}}\left(\frac{x-a}{b-a}\right) \quad \Rightarrow \quad f_{Y_{(k)}}(x) = \frac{1}{b-a} f_{X_{(k)}}\left(\frac{x-a}{b-a}\right)$$

$EY_{(k)} = (b - a)EX_{(k)} + a$; $DY_{(k)} = (b - a)^2 DX_{(k)}$; $K_{Y_{(k)}Y_{(l)}} = (b - a)^2 K_{X_{(k)}X_{(l)}}$,

$1 \le k < l \le n$;

$W_X = X_{(n)} - X_{(1)}$; $W_Y = Y_{(n)} - Y_{(1)} = (b - a)(X_{(n)} - X_{(1)}) = (b - a)W_X$;

$EW_Y = (b - a)EW_X$; $DW_Y = (b - a)^2 DW_X$.

Замечание. На основании п.1 достаточно изучать порядковые статистики случайной величины $X_{(k)}$: $f_{X_{(k)}}(x) = nC_{n-1}^{k-1} f(x)F^{k-1}(x)[1 - F(x)]^{n-1}$.

2.2) Вычисление моментов порядковых статистик равномерного распределения с.в. X~R[0,1].

$$f_k(x) = f_{X_{(k)}}(x) = \begin{cases} 0, \ x \notin [0,1] \\ nC_{n-1}^{k-1} x^{k-1}(1 - x)^{n-k}, \ x \in [0,1] \end{cases} ;$$

$$F_k(x) = F_{X_{(k)}}(x) = \begin{cases} 0, \ x < 0 \\ n\, C_{n-1}^{k-1} \int\limits_0^x t^{k-1}(1 - t)^{n-k}dt, \ x \in [0,1]. \\ 1, \ x > 1 \end{cases}$$

Далее используются функции:

$\Gamma(p) = \int\limits_0^\infty x^{p-1}e^{-x}dx$ и $B(p,q) = \int\limits_0^1 x^{p-1}(1 - x)^{q-1}dx$; ($\Gamma(p + 1) = p\Gamma(p)$, при целом p

$\Gamma(p + 1) = p!$; $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$)

$$EX_{(k)} = \int\limits_0^1 xf_k(x)dx = n\, C_{n-1}^{k-1} \int\limits_0^1 x^k(1 - x)^{n-k}dx = n\, C_{n-1}^{k-1} B(k + 1, n - k + 1) =$$

$$= n\, C_{n-1}^{k-1} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{n!}{(k-1)!(n-k)!} \frac{k!(n-k)!}{(n+1)!} = \frac{k}{n+1} \quad \Rightarrow \ EX_{(1)} = \frac{1}{n+1} \ ;$$

$$EX_{(n)}=\frac{n}{n+1}\quad\Rightarrow\quad EY_{(1)}=a+\frac{b-a}{n+1}\ ;\ EY_{(n)}=\frac{n(b-a)}{n+1}+a=b-\frac{b-a}{n+1};$$

$$DX_{(k)}=EX^2_{(k)}-(EX_{(k)})^2\ ;$$

$$EX^2_{(k)}=\int_0^1 x^2 f_k(x)dx=n\,C_{n-1}^{k-1}\int_0^1 x^{k+1}(1-x)^{n-k}dx=n\,C_{n-1}^{k-1}B(k+2,n-k+1)=$$

$$=n\,C_{n-1}^{k-1}\frac{\Gamma(k+2)\,\Gamma(n-k+1)}{\Gamma(n+3)}=\frac{n!}{(k-1)!(n-k)!}\frac{(k+1)!(n-k)!}{(n+2)!}=\frac{k(k+1)}{(n+1)(n+2)};$$

$$DX_{(k)}=\frac{k(k+1)}{(n+1)(n+2)}-\frac{k^2}{(n+1)^2}=\frac{k(n-k+1)}{(n+1)^2(n+2)}\Rightarrow DX_{(1)}=\frac{n}{(n+1)^2(n+2)}=DX_{(n)};$$

$$K_{X_{(k)}X_{(l)}}=EX_{(1)}X_{(n)}-EX_{(1)}EX_{(n)}\ ;$$

$$EX_{(1)}X_{(n)}=\iint uvf_{1n}(u,v)dudv\ ,\text{ где}$$

$$f_{(ln)}(x,y)=\begin{cases}n(n-1)f(x)f(y)(F(y)-F(x))^{n-2},\ x,y\in[0,1]\\ 0,\text{ в противном случае}\end{cases}=\begin{cases}n(n-1)(y-x)^{n-2},\ x,y\in[0,1]\\ 0,\text{ в противном случае}\end{cases}$$

$$EX_{(1)}X_{(n)}=n(n-1)\int_0^1\int_u^1 uv(v-u)^{n-2}dudv=n(n-1)\int_0^1 u\Big(\int_u^1 v(v-u)^{n-2}dv\Big)du\ ;$$

$$I=\int_u^1 v(v-u)^{n-2}dv=\int_u^1 (v-u)^{n-1}dv+u\int_u^1 (v-u)^{n-2}dv=$$

$$=\frac{(v-u)^n}{n}\Bigg|_u^1+u\frac{(v-u)^{n-1}}{n-1}\Bigg|_u^1=\frac{(1-u)^n}{n}+\frac{u(1-u)^{n-1}}{n-1}\quad\Rightarrow$$

$$\Rightarrow EX_{(1)}EX_{(n)}=\frac{1}{n-1}\int_0^1 u(u-1)^n du+\frac{1}{n}\int_0^1 u^2(1-u)^{n-1}du=$$

$$=\frac{1}{n-1}B(2,n+1)+\frac{1}{n}B(3,n)=\frac{\Gamma(2)\Gamma(n+1)}{(n-1)\Gamma(3+n)}+\frac{\Gamma(3)\Gamma(n)}{n\Gamma(n+3)}=\frac{1}{n+2}\quad\Rightarrow$$

$$K_{X_{(k)}X_{(n)}} = \frac{1}{n+2} - \left(\frac{n}{n+1}\right)^2 = \frac{1}{(n+1)^2(n+2)}$$

$$\left(\text{Без доказательства}: K_{X_{(k)}X_{(n)}} = \frac{k(n-l+1)}{(n+1)^2(n+2)}\right).$$

По п.1). найти моменты для $Y_{(k)}$ самостоятельно.

2.3) Моменты размаха выборки из равномерного распределения.

$W_X = X_{(n)} - X_{(1)}$; $EW_X = EX_{(n)} - EX_{(1)} = \dfrac{n}{n+1} - \dfrac{1}{n+1} = \dfrac{n-1}{n+1}$;

$DW_X = EW_X^2 - (EW_X)^2$, где $EW_X^2 = E(X_{(n)} - X_{(1)})^2 =$

$= EX_{(n)}^2 + EX_{(1)}^2 - 2EX_{(1)}X_{(n)}$;

$$EX_{(1)}^2 = \frac{2}{(n+1)(n+2)};\ EX_{(n)}^2 = \frac{n(n+1)}{(n+1)(n+2)};\ EX_{(1)}X_{(n)} = \frac{1}{n+2} \quad \Rightarrow$$

$$\Rightarrow\ DW_X = \frac{2+n(n+1)}{(n+1)(n+2)} - \frac{2}{n+2} - \left(\frac{n-1}{n+1}\right)^2 = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

Можно вычислить DW_X по-другому:

$$DW_X = D(X_{(n)} - X_{(1)}) = DX_{(1)} + DX_{(n)} - 2K_{X_{(k)}X_{(l)}} =$$

$$= \frac{2n}{(n+1)^2(n+2)} - \frac{2}{(n+1)^2(n+2)} = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

2.4) Распределение размаха выборки из равномерного распределения.

$$H_X(t) = P(W_X<t) = n\int_{-\infty}^{\infty} (F(u + t) - F(u))^{n-1}f(u)du$$

Найдём $H_X(t)$ сначала для $t \in [0,1]$. ($f(u) = f_x(t)$ – обозначение);

$$F(u) = \begin{cases} 0, u < 0 \\ u, u \in [0,1]; \\ 1, u > 1 \end{cases}$$

$$F(u+t) = \begin{cases} 0, u + t < 0 \Rightarrow u < -t \\ u + t, 0 \le u + t \le 1 \Rightarrow -t \le u \le 1-t , \\ 1, u + t >1 \Rightarrow u > 1-t \end{cases}$$

или удобнее для дальнейшего представления $F(u)$ и $F(u + t)$ в следующем виде:

$$F(u) = \begin{cases} 0, & u < -t \\ 0, & -t \le u \le 0 \\ u, & 0 < u \le 1-t \\ u, & 1-t < u \le 1 \\ 1, & u > 1 \end{cases}; \quad F(u+t) = \begin{cases} 0, & u < -t \\ u+t, & -t \le u \le 0 \\ u+t, & 0 < u \le 1-t \\ 1, & 1-t < u \le 1 \\ 1, & u > 1 \end{cases} \Rightarrow$$

$$\Rightarrow \quad F(u+t) - F(u) = t, \begin{cases} 0, & u < -t \\ u+t, & -t \le u \le 0 \\ & 0 < u < 1-t \\ 1-u, & 1-t < u \le 1 \\ 0, & u > 1 \end{cases} \Rightarrow$$

$$H_X(t) = n\int_0^{1-t} t^{n-1}du + n\int_{1-t}^{1} (1-u)^{n-1}du = n(1-t)t^{n-1} + n\frac{t^n}{n} = n(1-t)t^{n-1} + t^n$$

для $t \in [0,1] \Rightarrow$

$$H_X(t) = \begin{cases} 0, & t < 0 \\ t^n + n(1-t)t^{n-1}, & t \in [0,1] \\ 1, & t > 1 \end{cases} \Rightarrow$$

$$h_X(t) = \begin{cases} 0, & t \notin [0,1] \\ nt^{n-1} - n(n-1)(1-t)t^{n-2} - nt^{n-1}, & t \in [0,1] \end{cases}$$

Для нахождения $H_Y(t)$ используем связь случайных величин X и Y :
$Y = (b-a)X + a$

$$H_Y(t) = P(W_Y < t) = P\left(W_X = X_{(n)} - X_{(1)} < \frac{t}{b-a}\right) = \left(H_X \frac{t}{b-a}\right) \Rightarrow$$

$$h_Y(t) = \frac{1}{b-a}\, h_X\left(\frac{t}{b-a}\right).$$

Самостоятельно выписать $H_Y(t)$ и $h_Y(t)$ в явном виде.

§5. Моделирование распределений случайных величин

Часто для исследования вероятностных характеристик с.в. приходится создавать соответствующий статистический материал. Основные идеи и приемы такого моделирования подробно изложены, например, в [1] и [2], приведем здесь краткие сведения по этому вопросу.

С.в. с любым распределением могут быть получены из равномерного распределения на отрезке [0.1] чисел $\{r_i\}$, вырабатываемых на ЭВМ датчиком случайных чисел.

Моделирование дискретной с.в. и метод маркировки (м.м.)

Для моделирования n значений дискретной с.в. X с заданным рядом распределения

X	x_1	x_2	…	x_m
P	p_1	p_2	…	p_m

Нужно отрезок [0,1] разбить на m отрезков: $[0;p_1]$, $[p_1;p_1+p_2]$, …, $[p_1+p_2+\ldots+p_{m-1};1]$ и промаркировать реализацию n случайных чисел $\{r_i\}$ по построенным интервалам. Тогда попадание случайного числа r в i–ый интервал будет означать, что i – е значение разыгрываемой с.в. $x_i = x_j$ $(i=\overline{1,n})$;

Задача 1.

Разыграть (смоделировать) 6 значений с.в. X, имеющей ряд распределения

X	1	3	8
p	0,34	0,25	0,41

Решение.

С помощью датчика с.в. получим 6 случайных чисел $\{r_i\}$:

0,43; 0,61; 0,21; 0,84; 0,70; 0,5.

Промаркируем эти случайные числа по отрезкам:

[0;0.34], [0,34;0,59],[0,59;1]; получим, что случайные числа $\{r_i\}$ соответственно попали в интервалы с номерами: 2, 3, 1, 3, 3, 1, что отвечает следующей смоделированной последовательности: $\{x_i\}$:3, 8, 1, 8, 8, 1.

Моделирование непрерывной с.в. (метод обратных функций (м.о.ф.))

В основе моделирования непрерывной с.в. лежит следующий, легко доказываемый факт: если F(x) – непрерывная строго монотонная функция распределения и $F^{-1}(x)$ – обратная к ней функция, то с.в. $X = F^{-1}(R)$, где с.в. R равномерно распределена на отрезке [0,1], имеет функцию распределения F(x).

Доказательство.

Действительно, $P\{F^{-1}(R) < x\} = P\{R < F(x)\} = F(x)$, что и утверждалось.

Таким образом, если функцию $F^{-1}(x)$ можно явно вычислить, то моделирование n значений $\{x_i\}$ с.в. X производится из n значений случайных чисел $\{r_i\}$ по формуле

$$x_i = F^{-1}(r_i). \qquad (1)$$

Приведем примеры применения м.о.ф., используя обозначение L(x) – закон распределения с.в. X.

Задача 2.

Смоделировать получение значений с.в. X в следующих случаях:

а) $L(X) = R[a,b]$.

Решение.

$F(x) = (x-a)/(x-b)$, при $a < x < b$; по (1) имеем $r_i = (x_i - a)/(b-a)$ =>

$x_i = r_i(b-a) + a$;

б) $L(x) = E(L)$.

Решение.

$F(x) = 1 - \exp\{-Lx\}, x \geq 0$; по (1) имеем $r_i = 1 - \exp\{-Lx_i\}$ =>

$x_i = -(1/L)\ln(1-r_i)$, а так как с.в. $(1-r_i)$ распределена так же, как и с.в. r_i, то возможное значение с.в. x_i можно получить по более простой формуле $x_i = -(1/L)\ln r_i$.

Замечание 1. Если $F(x)$ не строго монотонная функция, то обратную к ней функцию определяют при моделировании значений с.в. X следующим выражением: $F^{-1}(r_i) = \sup\{x_i : F(x_i) < r_i\}$.

Ограниченность применения м.о.ф. связана с невозможностью явного вычисления обратной функции $F^{-1}(x)$. В таких ситуациях используются различные частные приемы моделирования.

Метод Неймана

Этот метод является достаточно общим для разыгрывания непрерывной с.в. X, когда она определена на конечном отрезке [a,b] и плотность ее ограничена: $f(x) \leq M$, где M – постоянная.

Разыгрывание возможного значения с.в. X производят по следующему алгоритму:

а) получают две равномерно распределенные на отрезке [0,1] с.в. R: r_1 и r_2;

б) вычисляют координаты случайной точки $Z=(z_1,z_2)$, где $Z_1=a+r_i(b-a)$, $z_2=r_2M$;

в) если точка Z лежит под кривой $y=f(x)$ (т.е. $z_2\le f(z_1)$), то полагают, что возможное значение с.в. X есть $x_i=z_1$; в противном случае точку отбрасывают и переходят к следующей точке, повторяя алгоритм с пункта а) до получения возможного с.в. X.

В качестве примеров остановимся на практическом моделировании значений наиболее распространенных распределений.

Замечание 2. Если смоделированы возможные значения с.в. X_1, то для моделирования реализаций с.в. $X_2=aX_1+b$ нужно применить это линейное преобразование к полученным возможным значениям с.в. $X_1: x_{2i}=ax_{1i}+b.$

Задача 3. Смоделировать получение значений с.в. X в следующих случаях:

а) L(X)=R[a,b]. Эта задача решена выше м.о.ф. Та же формула для моделирования получается на основании замечания 2:

$x_i=r_i(b-a)+a.$

б) L(X)=N(a, b). Если L(X)=N(1,0) и моделирование возможных значений с.в. X_0 x_{0i} проведено, то на основании замечания 2 получение возможных значений с.в. X x_i можно производить по формуле $x_i=bx_{0i}+a.$ Таким образом, задача сводится к моделированию с.в. x_0.

Приведем два способа моделирования возможных значений стандартной нормальной с.в. x_0.

1) По теореме Леви (ЦПТ) возможные значения с.в. X могут быть получены при достаточно большом n по формуле

$$x_{0i}=\left(\sum_{i=1}^{n}r_i-n/2\right)/\sqrt{n/12}.$$

Часто на практике достаточно брать n=12. Тогда имеем следующую простую формулу моделирования таких с.в.: $x_{0i}=\sum_{i=1}^{12}r_i-6.$

2) Две возможные реализации x_{01} и x_{02} вычисляются из двух значений чисел r_1 и r_2 по формулам

$x_{01}=\sqrt{-2\ln r_2}\cos(2\pi r_1)$;

$x_{02}=\sqrt{-2\ln r_2}\sin(2\pi r_1)$.

Первый из этих способов легче реализуется, но приводит к приближенно нормальным числам.

Задача 4.

Биномиальное распределение B(n,p). Возможное значение биномиально распределенной с.в. X~B(n,p) может быть получено по общей схеме с $p_i = P(x=i) = C_n^i p_i (1-p)^{n-1}$, где $i = \overline{0,n}$, или проще, методом браковки: берется n независимых равномерно распределенных случайных чисел $r_1,\ldots,r_n$, тогда количество чисел из них, меньших p, есть возможное значение x_i с.в. X~B(n,p).

В частности, при n=1 получаем возможное значение бернулевской с.в.

Задача 5.

Геометрическое распределение G(p), $(P(X=i) = p(1-p)^{i-1}; i \geq 1,2,\ldots)$

$x_i = [\ln r_i / \ln(1-p)]$, где $[\ldots]$ – означает целую часть числа.

С помощью описанного приема для моделирования возможного значения геометрического распределения могут быть получены возможные значения с.в., распределенной по:

а) сдвинутому геометрическому закону cgG(p),

$(P(Y=i) = p(1-p)_i, i = 1,2,\ldots)$: $y_i = x_i - 1$;

б) по закону Паскаля πa(r,p), ($P(U=i) = C_{i-1}^{r-1} p^r (1-p)^{i-r}, i = r, r+1$) по формуле $u_i = \sum_{i=1}^{r} x_i$;

в) отрицательному биномиальному закону OB(r, p),

$(P(V=i) = C_{r+i-1}^{i} p^r (1-p)^i, i = 0,1,2,\ldots)$ по формуле:

$v_i = \sum_{i=1}^{r} y_i$.

Задача 6.

Пуассоновское распределение $\pi(\lambda)$, $(P(X=k)) = (\lambda e^{-\lambda} / k!)$, k=0,1,2,…)

Возможное значение с.в. X может быть получено по одному из соотношений:

$x_i = \min\{n : \prod_{i=1}^{n+1} r_i < e^{-\lambda}$;

$x_i = \max\{n : \prod_{i=1}^{n} r_i \geq e^{-\lambda}$.

Задача 7.

Полиномиальное распределение $P(n, p_1, \ldots, p_m, m)$, где n – число опытов, m – число возможных исходов в каждом опыте, а $p_i(i=\overline{1,m})$ – вероятность i–ого исхода $P(x_1 = v_1, \ldots, x_m = v_m) = n! \prod_{i=1}^{m} \frac{p_i}{v_i}$; $\left(\sum_{i=1}^{m} v_i = n, \sum_{i=1}^{m} p_i = 1 \right)$.

Для моделирования:

а) разбивают отрезок на m частей: $\Delta_\mu = \left[\sum_{i=1}^{m-1} p_i, 1 \right]$;

б) получают n чисел r_1, r_2, r_n;

в) для каждого $r_j = (j = \overline{1,n})$ находят отрезок Δ_ι, которому r_j принадлежит;

г) для каждого отрезка $\Delta_i(i = \overline{1,m})$ определяют число v_1 попавших чисел из $r_1, r_2, \ldots, r_n$, тогда считают, что x_i приняла значение v_i, а с.в. $x = (x_1, x_2, \ldots, x_m)$ – приняла значение $((v_1, \ldots, v_m)$.

§6. Непараметрическая задача статистики.

Исходной информацией является случайная величина X и выборка наблюдений над ней $\vec{x} = (x_1, x_2, x_3, ..., x_n)$. Требуется найти вид распределения (семейство распределений), к которому относится распределение с.в. X. Для решения этой (непараметрической) задачи существует целый ряд обработок или представлений данной выборки.

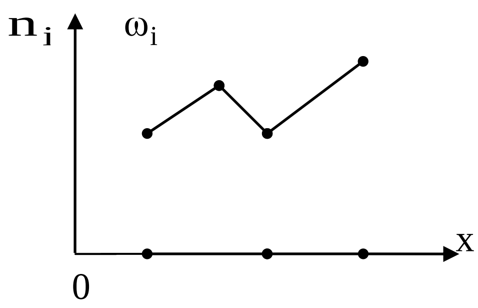
1. Статистический (или частотный) ряд – это таблица вида:

X	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k
$\omega_i = \frac{n_i}{n}$	ω_1	ω_2	...	ω_k

Здесь $x_1,...,x_k$ – возрастающая последовательность разных наблюденных значений с.в. X; n_i – число наблюденных значений с.в.

$X = x_i$, $i = \overline{1,n}$; $\sum_{i=1}^{k} n_i = n$ –объем выборки; $\sum_{i=1}^{k} \omega_i = 1$.

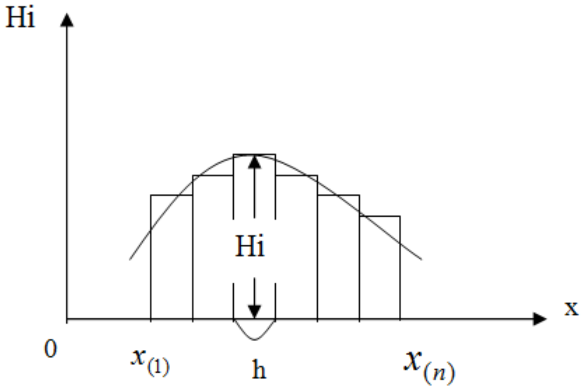
2. Полигон частот (относительных частот) – это ломаная линия (строится для дискретной с.в.), соединяющая точки $(x_i, n_i)((x_i, \omega_i))$.



Полигоны частот и относительных частот отличаются друг от друга только масштабом по оси ординат. Полигон относительных частот является графическим статическим аналогом соответствующего представления ряда распределения для точек (x_i, p_i).

3. Гистограмма частот (относительных частот) (строится для непрерывной с.в.)

– это ступенчатая фигура, сглаживающая кривая для которой является статистическим аналогом кривой распределения с.в. X.



Гистограмма строится следующим образом: на оси абсцисс отмечают отрезок $(x_{(1)}, x_{(n)})$, делят его на равные отрезки длиной h таким образом, чтобы ни один отрезок не был пуст, а далее над $i-м$ отрезком строят прямоугольники высотой H_i, где

$$H_i = \begin{cases} \dfrac{n_i}{h} & \text{для гистограммы частот;} \\ \dfrac{\omega_i}{h} = \dfrac{n_i}{nh} & \text{для гистограммы относительных частот;} \end{cases}$$

Площадь ступенчатой фигуры – гистограммы

$$S = \begin{cases} \sum\limits_{i=1}^{k} \dfrac{n_i}{h} h = \sum\limits_{i=1}^{k} n_i = n & \text{для гистограммы частот} \\ \sum\limits_{i=1}^{k} \dfrac{n_i}{nh} h = \sum\limits_{i=1}^{k} \omega_i = 1 & \text{для гистограммы относительных частот;} \end{cases}$$

Гистограммы частот и относительных частот отличались только масштабом по оси ординат. Кривую распределения с.в. X приближает сглаживающая кривая гистограммы относительных частот, т.к. в этом случае

$S = \sum\limits_{i=1}^{k} \omega_i = 1$ как и площадь под графиком плотности распределения.

4. Эмпирическая функция распределения $F_n(x)$ (э.ф.р.)

$F_n(x) = \dfrac{k}{n}$, где n – объем выборки, k – число наблюдений < x.

Задача.

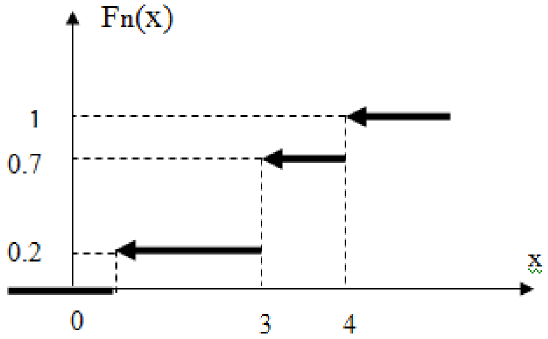
Пусть наблюденная выборка значений с.в. X есть 3, 1, 4, 4, 3, 3, 3, 1, 3, 4. Построить э.ф.р. $F_n(x)$.

Решение.

Построим статический ряд

X	1	3	4
n_i	2	5	3

$$\sum_{i=1}^{3} n_i = n = 10;$$



$$F_n(x) = \begin{cases} \dfrac{0}{10} = 0\,,\ x \le 1 \\ \dfrac{2}{10} = 0.2\,,\ 1 < x \le 3 \\ \dfrac{2+5}{10} = 0.7\ ,\ 3 < x \le 4 \\ \dfrac{2+5+3}{10} = 1\ ,\ \ x > 4 \end{cases}$$

Свойства $F_n(x)$ (совпадают со свойствами F(x)):

1. $0 \le F_n(x) \le 1;$
2. $F_n(x)(-\infty) = 0; F_n(+\infty) = 1;$
3. $F_n(x)$ – неубывающая по x;
4. $F_n(x)$ – непрерывна слева.

Распределение и моменты э.ф.р. $F_n(x)$.

Введем индикаторную функцию $J(z) = \begin{cases} 0, z \le 0 \\ 1, z > 0 \end{cases}$. Тогда

$F_n(x) = \dfrac{k}{n} = \dfrac{1}{n}\sum_{i=1}^{n} J(x - x_i)$, где при фиксированном значении x с.в. J(x–X) ~B (1,p=F(x)), так как

J(x–X)	0	1

P	$P(x - X \le 0) = P(X \ge x) =$ $1 - P(X < x) = 1 - F(x)$	$P(x - X < 0) = P(X < x)$ $= F(x)$

Известно, что тогда $Y = \sum_{i=1}^{n} J(x - x_i) \sim B(n, p)$, где p=F(x)

$EY = np = nF(x), DY = np(1-p) = nF(x)(1-F(x)),$

$P\left\{F_n(x) = \frac{k}{n}\right\} = P\{Y = k\} = C_n^k (F(x))^k (1-F(x))^{n-k};$

$EF_n(x) = E\frac{Y}{n} = \frac{1}{n}EY = \frac{1}{n}nF(x) = F(x);$

$DF_n(x) = D\frac{Y}{n} = \frac{1}{n^2}DY = \frac{1}{n^2}nF(x)(1-F(x)) = \frac{F(x)(1-F(x))}{n} \to 0,$ при $n \to \infty$

Таким образом, при фиксированном значении x $EF_n(x) = F(x)$,
а это значит, что $F_n(x)$ – несмещенная оценка для EX; $DF_n(x) \to 0$ *при* $n \to \infty$,
а это значит, что $F_n(x)$ – состоятельная оценка для F(x).

Сходимость $F_n(x)$ к $F(x)$.

а) Обозначим $F_n(x) = F_n$; $F(x) = F$. Воспользуемся неравенством Чебышева:

$P(|\xi - E\xi| \ge \varepsilon) \le \frac{D\xi}{\varepsilon^2}$, положим $\xi = F_n(x)$, тогда $E\xi = F_n(x)$ и

$D\xi = \frac{F(x)(1-F(x))}{n}$.

Получим $P\{|F_n - F| \ge \varepsilon\} \le \frac{F(x)(1-F(x))}{n\varepsilon^2} \to 0,$ при $n \to \infty \Rightarrow F_n \xrightarrow{\text{п.в.}} F$ при $n \to \infty$.

б) Воспользуемся теоремой Колмогорова. Если $\xi_1, \ldots, \xi_n$ – последовательность независимых с.в. и выполнено условие $\sum_{n=1}^{\infty} \frac{D\xi_n}{n^2} < \infty$, то $\xi_n \xrightarrow{\text{п.н.}} \xi$ при $n \to \infty$, т.е. $P\{\xi_n > \xi\} = 0$.

Положим $\xi_n = F_n$ и проверим условие теоремы Колмогорова:

$\sum_{n=1}^{\infty} \frac{DF_n(x)}{n^2} = \sum_{n=1}^{\infty} \frac{F(x)(1-F(x))}{n^3} < \infty \Rightarrow F_n \xrightarrow{\text{п.н.}} F$ при $n \to \infty$.

Докажем, что из сходимости п.н. $\Rightarrow$ сходимость п.в.

Доказательство.

Пусть $\xi_n \xrightarrow{\text{п.н.}} \xi$, т.е. $P\{\xi_n \not\to \xi\} = 0$ или $P\{\xi_i > \xi\} = \varepsilon$.

Хотя бы при одном $m \ge n\} = P(A)$, а сходимость п.в. означает, что

$P\{|\xi_n - \xi| > \varepsilon\} = P(B)$, но $B \subset A \Rightarrow P(B) \le P(A) \Rightarrow$ что и требовалось доказать, т.е. $\xi_n \xrightarrow{\text{п.в.}} \xi, n \to \infty$.

Замечание. Из доказанного утверждения следует, что т.к. при $n \to \infty$ $F_n \xrightarrow{\text{п.н.}} F$ по теореме Колмогорова, то $F_n \xrightarrow{\text{п.в.}} F$, что было установлено раньше непосредственно по неравенству Чебышева.

Точность и надежность $F_n(x)$ для $F(x)$ при фиксированном значении x.

а) Пусть n велико. Тогда по следствию из интегральной теоремы Муавра – Лапласа

$$P = P\{|F_n - F| \le \varepsilon\} = P\left\{\left|\frac{k}{n} - P\right| \le \varepsilon\right\} \approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{F(1-F)}}\right) \ge 2\Phi(2\varepsilon\sqrt{n}) \to 1$$

при $n \to \infty$ т т. $F(1-F) \le 4^{-1}$.

Здесь P – надежность оценки F_n для F с точностью ε, а $2\Phi(2\varepsilon\sqrt{n})$ – нижняя граница надежности.

б) При любом n по неравенству Чебышева имеем

$P = P\{|F_n - F| \le \varepsilon\} \ge 1 - \frac{DF_n}{\varepsilon^2} = 1 - \frac{F(1-F)}{n\varepsilon^2} \ge 1 - \frac{1}{4n\varepsilon^2}$ – нижняя граница надежности оценки F_n для F с точностью ε.

Часть II

II часть пособия по математической статистике предполагает знание 1–й части и посвящена решению одной из основных задач математической статистики – параметрической задачи. Исходными данными здесь являются изучаемая случайная величина (с.в) X, выборка $\vec{x} = (x_1,...,x_n)$ значений с.в. X с точностью до параметра $\theta = (\theta_1,...,\theta_n)$. Требуется оценить заданную функцию от $\theta : \tau(\theta)$.

Рассматриваются два подхода к решению задачи: точечное и доверительное оценивание. Обсуждаются методы построения таких оценок.

Большое внимание отводится вопросу анализа качества полученных оценок с различных точек зрения и при различной природе неизвестного параметра распределения с.в. X.

Изложение теоретического материала сопровождается многочисленными примерами с подробными комментариями, а также предложены задачи для самостоятельного решения.

Настоящее пособия имеет целью оказание помощи студентам подбором соответствующего материала и пояснениями при решении поставленных задач.

§1. Выборочные моменты. Их свойства

1. Начальные сведения

В качестве оцениваемых функций в параметрической задаче статистики часто встречаются выборочные моменты. Пусть X – наблюдаемая с.в., $\vec{x}$ =(x_1, x_2,...., x_n)– выборка объема n.

$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} x_i^r$ – r – ый начальный выборочный момент; при r = 1;

$\alpha_i = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное соединение;

$\hat{\mu}_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$ – r – ый центральный выборочный момент; при r = 2;

$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ – выборочная дисперсия;

$\hat{\sigma} = \hat{\sigma}_x = \sqrt{S_1^2}$ – выборочное среднеквадратическое отклонение;

При EX=m – известном

$\hat{m}_r = \frac{1}{n}\sum (x_i - m)^r$ – r – ый центральный выборочный момент;

если выборки значений с.в. X и У есть соответственно $x_1,\dots,x_n; y_1,\dots,y_n$, то

$\hat{K}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ – выборочная корреляция;

$\hat{r}_{xy} = \frac{\hat{K}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$ – выборочный коэффициент корреляции;

$\hat{S}_k = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$ – выборочный коэффициент ассиметрии;

$\hat{\varepsilon} = \frac{\hat{\mu}_4}{\hat{\sigma}^4} - 3$ – выборочный коэффициент эксцесса

2. Свойства выборочных моментов

Сформулируем их в форме задач.

1) Исследовать $\hat{\alpha}_r$ на несмещенность и состоятельность для α_r r – го начального момента $\left(\alpha_r = EX^r\right)$.

Решение.

$E\widehat{\alpha_r} = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right) = \frac{1}{n}\cdot n \cdot EX^r = EX^r = \alpha_r$, т.е. $\hat{\alpha}_r$ – несмещенная оценка для α_r; по теореме Хинчина при $\xi_i = x_i^r$ имеем: если $Ex_i^r < \alpha_r < \infty$, то для с.в. $\xi_1,\dots,\xi_n$ выполняется ЗБЧ, т.е. $\hat{\alpha}_r$ – состоятельная оценка для α_r. В частности, при r = 1 $\overline{x}$ – несмещенная состоятельная оценка для EX.

2) Исследовать на несмещенность и состоятельность S_1^2 для DX.

Решение.

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\sum_{i=1}^{n} x_i + \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2;$$

$$E\overline{x} = EX;\ D\overline{x} = \frac{DX}{n} \Rightarrow E\overline{x^2} = D\overline{x} + (E\overline{x})^2 = \frac{DX}{n} + (EX)^2;$$

$$ES_1^2 = \frac{1}{n}nEX^2 - E\overline{x}^2 = EX^2 - \frac{DX}{n} - (EX)^2 = \left(1 - \frac{1}{n}\right)DX = \frac{n-1}{n}DX \underset{n\to\infty}{\to} DX \Rightarrow$$

$\Rightarrow S_1^2$ – смещенная, но асимптотически несмещенная оценка для DX;

смещение $b = ES_1^2 - DX = -\frac{DX}{n} < 0 \Rightarrow S_1^2$ в среднем занижает значение DX.

Т.к. смещение линейно, можно его подправить:

$$S_2^2=\frac{n}{n-1}S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2;$$

$ES_2^2=\frac{n}{n-1}ES_1^2=DX\Rightarrow S_2^2$ – несмещенная оценка для DX.

Состоятельность S_1^2 для DX следует из теоремы Хинчина для $\{\xi_i\}=\{x_1^2\}$ и состоятельности $\bar{x}$ для DX, так как

$$S_1^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-x^{-2}\to EX^2-(EX)^2=DX.$$

3) Исследовать $\hat{m}_r$ для m_r следует из теоремы Хинчина, если в качестве ξ_i взять x_i-m:

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-E\xi_i\right|>\varepsilon\right\}=P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2-m_r\right|>\varepsilon\right\}\underset{n\to\infty}{\to}0$$

Отсюда, в частности, следует, что при известном EX = m $\hat{m}_2$ – несмещенная и состоятельная оценка для DX.

4) Исследовать точность и надежность $\bar{x}$ для EX = m.

Решение.

Обозначим L(z) – закон распределения с.в. Z.

Если $\gamma=P(|\bar{x}-m|>\varepsilon)$, то ε – точность, а γ – надежность оценки $\bar{x}$ для m; найдем распределение статистики $\bar{x}$. Обозначим

$$S_n=\bar{x}=\sum_{i=1}^{n}\frac{x_i}{n};\ S_n^*=\frac{S_n-ES_n}{\sqrt{DS_n}};ES_n=E\bar{x}=m;DS_n=D\bar{x}=\frac{Dx}{n}=\frac{\sigma^2}{n};$$

$$S_n^*=\frac{\bar{x}-m}{\sigma/\sqrt{n}}$$

тогда по теореме Леви $L(S_n^x)=L\left(\frac{\bar{x}-m}{\sigma/\sqrt{n}}\right)\xrightarrow[n\to\infty]{}N(0,1)$;

$$\bar{x}=S_n^*\cdot\frac{\sigma}{\sqrt{n}}+m\Rightarrow L(\bar{x})\xrightarrow[n\to\infty]{}N\left(m,\frac{\sigma}{\sqrt{n}}\right)$$

$$\gamma=P(|\bar{x}-m|>\varepsilon)=P(m-\varepsilon<\bar{x}<m+\varepsilon)=$$

$$\Phi\left(\frac{m+\varepsilon-m}{\sigma/\sqrt{n}}\right)-\Phi\left(\frac{m-\varepsilon-m}{\sigma/\sqrt{n}}\right)=2\Phi(\varepsilon\sqrt{n}/\sigma).$$

Таким образом, надежность $\bar{x}$ для EX с точностью ε есть $2\Phi(\varepsilon\sqrt{n}/\sigma)$ и может быть увеличена за счет выбора большего n.

Приведем числовые примеры.

С какой вероятностью (надежностью) совершается ошибка < ε =0.3 при замене EX на $\overline{x}$ при:

а) n = 20; $\overline{x} = 4.52$; $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 2.35$.

Решение.

По 4). При $D\overline{x} = \frac{DX}{n} \approx \frac{2.35}{20} = 0.1175; \sigma \approx \sqrt{\frac{2.35}{20}} = 0.343;$

$\gamma = P(|\overline{x} - EX| < \varepsilon) \approx 2\Phi(0.3/0.343) \approx 0.618$.

б) $n = 100$; $\overline{x} = 4.52$; $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 2.35$.

Решение.

$\sigma \approx \sqrt{\frac{2.35}{100}} = 0.1533$; $\gamma = P(|\overline{x} - m| < 0.3) \approx 2\Phi(0.3/0.1533) \approx 0.9$.

При сравнении результатов а) и б) наглядно видно повышение надежности γ с ростом n.

5) Исследовать на несмещенность $\hat{K}_{xy}$ для K_{xy}.

Решение.

$\hat{K}_{xy} = \frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})$; $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $MX = m_x, MY = m_y$.

$$E\hat{K}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = Ex_iy_i - E\overline{x}y_i - E\overline{y}x_i + E\overline{x}\overline{y} \qquad (*)$$

$Ex_iy_i = K_{xy} + m_xm_y$ т.к. x_i и y_i независимы;

$$Ey_i\overline{x} = E\left(y_i \cdot \frac{1}{n}\sum_{j=1}^{n} x_j\right) = \frac{1}{n}Ex_iy_i + \frac{1}{n}\sum_{i\neq j}Mx_iy_i = \frac{K_{xy} + m_xm_y}{n} + \frac{n-1}{n}m_xm =$$

$$= \frac{K_{xy}}{n} + m_xm_y;$$

$$E\overline{x}\overline{y} = \frac{1}{n^2}E\left(\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i\neq j} Ex_iy_i + \frac{1}{n^2}\cdot nEx_iy_i = \frac{n(n-1)}{n^2}m_xm_y +$$

$$+\frac{1}{n}\left(K_{xy} + m_xm_y\right) = m_xm_y - \frac{1}{n}m_xm_y + \frac{K_{xy}}{n} + \frac{1}{n}m_xm_y = m_xm_y + \frac{K_{xy}}{n}$$

Подставим все это в выражение (*):

$$E\hat{K}_{xy} = K_{xy} + m_xm_y - \frac{K_{xy}}{n} - m_xm_y - \frac{K_{xy}}{n} - m_xm_y + \frac{K_{xy}}{n} + m_xm_y$$

$$= \left(1 - \frac{1}{n}\right)K_{xy} = \frac{n-1}{n}\frac{K_{xy}}{n}$$

$\Rightarrow \hat{K}_{xy}$ – смещенная, но асимптотически несмещенная оценка для K_{xy}.

Смещение оценки $\hat{K}_{xy}$ есть $b = \left(\frac{n-1}{n} - 1\right)K_{xy} = -\frac{K_{xy}}{n}$ линейное. Подправим смещение оценки $\hat{K}_{xy}$: $\tilde{K}_{xy} = \frac{n}{n-1}\hat{K}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)(y_i - \bar{y})$ – несмещенная оценка для K_{xy}.

3. Законы распределения и моменты статистик $\bar{x}$ и S_1^2 для X~N(m, σ).

а) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $g_{x_i}(t) = \exp\{itm - \sigma^2 t^2/2\}$;

$$g_x = \sum_{i=1}^{n}\frac{x_i}{n}(t) = \left(g_{x_i}\left(\frac{t}{n}\right)\right)^n = e^{i\frac{t}{n}mn - \frac{n^2\sigma t^2}{2n^2}} = e^{itm - \frac{\left(\frac{\sigma}{\sqrt{n}}\right)^2 t^2}{2}} \Rightarrow$$

$$\Rightarrow \bar{x} \sim N(m, \frac{\sigma}{\sqrt{n}}) \Rightarrow \begin{cases} E\bar{x} = m \\ D\bar{x} = \frac{\sigma^2}{n} = \frac{DX}{n} \end{cases}$$

б) $S_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - x^2$

Подвергнем $\{x_i\}$ линейному преобразованию с матрицей преобразования следующего вида:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n\text{-}11} & a_{n\text{-}12} & \cdots & a_{n\text{-}1n} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix}, \text{т.е.} \begin{cases} y_i = \sum_{j=1}^{n} a_{ij}x_j,\ i = \overline{1, n-1} \\ y_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} x_j = \sqrt{n}\,\bar{x} \end{cases} \qquad (1)$$

при выполнении условий
$$\begin{cases} \sum_{j=1}^{n} a_{ij}^2 = 1,\ i = \overline{1, n} \\ \sum_{k=1}^{n} a_{ik}a_{jk} = 0,\ i = \overline{1, n-1};\ j = \overline{1, n-1} \\ \sum_{j=1}^{n} a_{ij} = 0,\ i = \overline{1, n-1} \end{cases} \qquad (2)$$

Покажем, что (2) – ортогональные преобразования, т.е. (по определению) такие $y_i = \sum_{j=1}^{n} a_{ij} x_j$, для которых выполняются условия:

$$\begin{cases} \sum_{j=1}^{n} a_{ij}^2 = 1 \\ \sum_{k=1}^{n} a_{ik} a_{jk} = 0, i, j = \overline{1,n} \end{cases}, \qquad (3)$$

т.е. нужно показать, что для данного преобразования (1) условия (2) следуют из условий (3). Действительно, $\sum_{j=1}^{n} a_{ij}^2 = 1$ по (2), (и при

i=n: $\sum_{j=1}^{n} a_{ij}^2 = n \cdot \left(\frac{1}{\sqrt{n}}\right)^2 = 1$, а из условия $\sum_{k=1}^{n} a_{ik} a_{jk} = 0$ при i=n имеем

$\sum_{k=1}^{n} a_{nk} a_{jk} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} a_{jk} = 0 \Rightarrow \sum_{k=1}^{n} a_{jk} = 0$. Таким образом, доказано, что для преобразований (1) выполнены условия (3), а значит (1) – ортогональное

преобразование, сохраняющее сумму квадратов: $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2$.

$x_1,\ldots,x_n$ – независимые с.в. как элементы выборки, каждая из которых ~N(m,σ), значит, их совместное распределение нормально (с плотностью, равной произведению их одинаковых плотностей). Известно, что ортогональное преобразование переводит нормальный вектор $(x_1,\ldots,x_n)$ в другой нормальный вектор $(y_1,\ldots,y_n)$.

Найдем моменты с.в. $\{y_i\}$ (i=1,n) при условиях (2), обозначив $Ex_i = EX = m, Dx_i = DX = \sigma^2, i = \overline{1,n}$;

$$Ey_i = E\left(\sum_{j=1}^{n} a_{ij} x_j\right) = \sum_{j=1}^{n} a_{ij} \cdot Ex_j = m \sum_{j=1}^{n} a_{ij} = 0, i = \overline{1, n-1}$$

$$Ey_n = \sqrt{n} E\bar{x} = m\sqrt{n};\ Dy_i = D\sum_{j=1}^{n} a_{ij} x_j = \sum_{j=1}^{n} a_{ij}.\ Dx_j = \sigma^2 \sum_{j=1}^{n} a_{ij}^2 = \sigma^2;$$

при i≠j

$$K_{y_i y_j} = E_{y_i y_j} - Ey_i Ey_j = E\left(\sum_{k=1}^{n} a_{ik} x_k \sum_{l=1}^{n} a_{jl} x_l\right) = \sum_{k,l}^{n} a_{ik} a_{jl} E(x_k x_1) =$$

$$= EX^2 \sum_{k=1}^{n} a_{ik} a_{jk} + (EX)^2 \sum_{k=1}^{n} a_{ik} \sum_{l=1}^{n} a_{jl} x_k = EX^2 \sum_{k=1}^{n} a_{ik} \sum_{k=1}^{n} a_{jk} - (EX)^2 \sum_{k \neq 1} a_{ik} a_{jl} = 0,$$

А это значит для нормального вектора $(y_1,...,y_n)$ независимость его компонент с учетом следующего замечания (без доказательств).

Замечание 1. Для гауссовского вектора некоррелированность его компонент эквивалентна их независимости.

Т.к. ортогональное преобразование сохраняет сумму квадратов и $y_n^2 = n\overline{x}^2$, то верно равенство:

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n} y_n^2 = \frac{1}{n}\sum_{i=1}^{n-1} y_i^2 = \sum_{i=1}^{n-1}\left(\frac{y_i}{\sqrt{n}}\right)^2 = \sum_{i=1}^{n-1}(y_i^*)^2,$$

где $y_i^* = \dfrac{y_i}{\sqrt{n}}$; $g_{y_i}(t) = e^{-\frac{\sigma^2 t^2}{2}}$; $g_{y_i^*}(t) = g_{y_i}\left(\dfrac{t}{\sqrt{n}}\right) = e^{-\frac{\sigma^2 t^2}{2n}} \Rightarrow \alpha\lambda$

$$\Rightarrow \quad y_i^* \sim N\left(0, \sigma^* = \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \sum_{i=1}^{n-1}(y_i^*)^2 = S_1^2 \sim \Gamma_{\alpha,\lambda},$$

где $\alpha = (n-1)/2$, а $\lambda = \dfrac{1}{2\sigma^{*2}}$, что соответствует плотности гамма-распределения $f(x,\alpha,\lambda) = \begin{cases} 0, x < 0 \\ \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0 \end{cases}$, где $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$;

$$ES_1^2 = \frac{\alpha}{\lambda} = \frac{(n-1)2\sigma^2}{2n} = \frac{n-1}{n}\sigma^2;\ DS_1^2 = \frac{\alpha}{\lambda^2} = \frac{(n-1)4\sigma^4}{2n^2} = \frac{2(n-1)}{n^2}\sigma^4;$$

$$\frac{y_i}{\sigma^*} = \frac{y_i\sqrt{n}}{\sigma} \sim N(0,1) \Rightarrow \sum_{i=1}^{n^*=n-1}\left(\frac{y_i}{\sigma^*}\right) \sim x_{n-1}^2;$$

$$\sum_{i=1}^{n-1}\left(\frac{\sqrt{n}y_i}{\sigma}\right) = \frac{nS_1^2}{\sigma^2} \sim \chi_{n-1}^2 \Rightarrow \begin{cases} E\dfrac{nS_1^2}{\sigma^2} = n-1 \\ D\dfrac{nS_1^2}{\sigma^2} = 2(n-1) \end{cases} \Rightarrow \begin{cases} ES_1^2 = \dfrac{n-1}{n}\sigma^2 \\ DS_1^2 = \dfrac{2(n-1)}{n^2}\sigma^4 \end{cases}$$

Замечание 2. Не нарушая общности, можно считать, что исходные с.в. $x_i \sim N(0,\sigma)$, $i = \overline{1,n}$, т.к. S_1^{*2} для $x^* \sim N(m,\sigma), i = \overline{1,m}$ и S_1^2 для $x_i \sim N(0,\sigma), i = \overline{1,n}$ совпадают, покажем это: $x_i = x_i^* - m \Rightarrow x_i^* = x_i + m;$;

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ;$$

$$S_1^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i^* - \overline{x^*})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + m - \bar{x} - m)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = S_1^2,$$ т.к.

$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - m) = \bar{x} + m$. Т.е. S_1^2 и S_1^{*2} совпадают.

Замечание 3. Для матрицы ортогонального преобразования $A = (a_{ij}), i,\ j = \overline{1,n}$ $A^T = A^{-1}$ (где A^T – транспонированная матрица, а A^{-1} – обратная матрица).

Тогда из условий ортогональности (3) $\Longleftrightarrow$ условия:

$$\begin{cases} \sum\limits_{k=1}^{n} a_{ki}a_{kj} = 0, i \neq j \\ \sum\limits_{k=1}^{n} a_{ki}^2 = 1 \end{cases}, \qquad (4)$$

т.е. условия (3) и (4) эквивалентны.

Покажем, что с учетом замечания 2 и эквивалентности условий (2) и (3) для рассматриваемых преобразований (1) сохраняются суммы квадратов, т.е. $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2$. Действительно:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n-1}\left(\sum_{j=1}^{n} a_{ij}x_j\right)^2 + (\sqrt{n}\bar{x})^2 = \sum_{i=1}^{n-1}\sum_{j=1}^{n} a_{ij}^2 x_j^2 + \sum_{i=1}^{n-1}\sum_{j_1}\sum_{j_2} a_{ij_1}a_{ij_2}x_{j_1}x_{j_2}$$

$$+\frac{1}{n}(x_1 + \ldots + x_n)^2 = \sum_{j=1}^{n} x_j^2 \sum_{i=1}^{n-1} a_{ij}^2 + \sum_{j_1} x_{j_1} \sum_{j_2} x_{j_2} \sum_{i=1}^{n-1} a_{ij_1}a_{ij_2} + \frac{1}{n}\sum_{i=1}^{n-1} x_j^2 +$$

$$\frac{1}{n}\sum_{j_1}\sum_{j_2} x_{j_1}x_{j_2}, j_1 \neq j_2$$

Отдельно вычисляем следующие суммы:

$$\sum_{i=1}^{n-1} a_{ij_1}a_{ij_2} = 0 - a_{nij_1}a_{nij_1} = -\frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{n}} = -\frac{1}{n};$$

$$\sum_{i=1}^{n-1} a_{ij}^2 = \sum_{i=1}^{n} a_{ij}^2 - a_{nj}^2 = 1 - \frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{n}} = 1 - \frac{1}{n}$$, поэтому

$$\sum_{i=1}^{n} y_i^2 = \sum_{j=1}^{n} x_j^2\left(1 - \frac{1}{n}\right) + 2\left(-\frac{1}{n}\right)\sum_{j_1<j_2} x_{j_1}x_{j_2} + \frac{1}{n}\sum_{j=1}^{n} x_j^2 + \frac{2}{n}\sum_{j_1<j_2} x_{j_1}x_{j_2} =$$

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

что и утверждалось.

Замечание 4. Если $X \sim N(m,\sigma)$, то $\overline{x}$ и S_1^2 независимы.

Действительно, по (1) $\overline{x} = \dfrac{y_n}{\sqrt{n}}$, а $S_1^2 = \dfrac{1}{n}\sum_{i=1}^{n} y_i^2$, т.е. $\overline{x}$ и S_1^2 зависят от разных, независимых между собой с.в. соответственно y_n и $(y_1,..,...,y_{n\text{-}1})$, а значит, $\overline{x}$ и S_1^2 независимы между собой.

§2. Свойства точечных оценок

1. Задача точечного оценивания

Исследуется случайная величина X, распределение которой относится к параметрическому множеству $F_\theta(x)$, где $\theta = (\theta_1, \ldots, \theta_к)$ – неизвестный k – мерный параметр. В дальнейшем будем обозначать это короче: $X \sim F_\theta(x)$.

Имеется выборка наблюденных значений случайной величины $X: \tau = \tau(\theta)$ объема n. Требуется построить точечную оценку (статистику) для данной функции $\tau = \tau(\theta)$ и исследовать качество данной оценки.

Замечание. При такой постановке задачи нет проблемы построения оценки для τ, так как требования к ее качеству (близости к истинному значению) не высказаны. Математическая задача возникает тогда, когда эти требования математически формализованы. А в связи с тем, что они не все и не всегда выполнимы, будем называть их желательными свойствами оценок.

2. Простейшие свойства точечных оценок $\mathbf{t = t(\vec{x})}$

а) **Несмещенность**: $Et(\vec{x}) = \tau(\theta)$(EX– математическое ожидание с.в. X) или асимптотическая несмещенность $Et(\vec{x}) \xrightarrow[n\to\infty]{} \tau(\theta)$.

б) **Состоятельность**: $t(\vec{x}) \xrightarrow{п.в.} \xrightarrow[n\to\infty]{} \tau(\theta)$, т.е.

$P\{|t(x) - \tau(\theta)| > \varepsilon\} \xrightarrow[n\to\infty]{} 0$ для любого $\varepsilon - const > 0$.

в) **Эффективность** несмещенной оценки характеризуется дисперсией $Dt(\vec{x})$ и используется для сравнения качества несмещенных оценок.

Эти свойства или пожелания к качеству точечной оценки объединены стремлением достичь определенной степени концентрации возможных значений оценки $\hat{\tau} = t(\vec{x})$ вокруг истинного значения оцениваемой функции $\tau(\theta)$.

Одновременное выполнение этих желательных свойств не всегда возможно, поэтому представление о «хорошей» оценке зависит от цели и возможностей исследования, определяющих приоритетные свойства оценки. Так, для малых выборок часто важна несмещенность оценки, а для больших – асимптотическая несмещенность и состоятельность. А иногда, сознательно отказываясь от одних свойств оценок, добиваются выполнения других, более важных с точки зрения исследования свойств.

3. Оптимальные оценки

Если же оценка является несмещенной с минимальной дисперсией, то она называется оптимальной.

Теорема.

Если оптимальная оценка существует, то она единственна.

Доказательство (от противного).

Пусть не так, т.е. существуют две оптимальные оценки для $\tau = \tau(\theta)$: $t_1 = t_1(\vec{x})$ и $t_2 = t_2(\vec{x})$. Тогда $Et_1 = Et_2 = \tau$; $Dt_1 = Dt_2 = \sigma^2$- минимальная возможная дисперсия несмещенных оценок для τ.

Построим оценку $t_3 = t_3(\vec{x}) = \dfrac{t_1 + t_2}{2}$ и изучим ее свойства. $Et_3 = \tau \Rightarrow t_3$ – снова несмещенная оценка для τ.

$Dt_3 = \dfrac{1}{4}\left(Dt_1 + Dt_2 + 2K_{t_1t_2}\right) = \dfrac{1}{4}\left(2\sigma^2 + 2K_{t_1t_2}\right) = \dfrac{\sigma^2 + K_{t_1t_2}}{2} \le \sigma^2$, т.к. по неравенству Коши-Буняковского $\left|K_{t_1t_2}\right| \le \sigma^2$. Но Dt_3 не может быть $< \sigma^2$, по условию. Отсюда следует, что $Dt_3 = \sigma^2$ (поэтому t_3 – оптимальная оценка для τ), а это значит, что $\left|K_{t_1t_2}\right| = \sigma^2$ или $\left|r_{t_1t_2}\right| = 1$ и t_1 и t_2 линейно зависимы $t_2 = at_1 + b$, где a и b – const.

Тогда $Et_1 = aEt_2 + b$ или $\tau = a\tau + b \Rightarrow a = 1, b = 0 \Rightarrow t_1 = t_2$, что и утверждалось.

4. Общий подход к сравнению оценок

Функция потерь -$G(t(\vec{x}), \tau(\theta)) = G(t, \tau)$ – это любая неотрицательная функция, дающая потерю (ущерб) в результате того, что за $\tau = \tau(\theta)$ принята ее оценка $\hat{\tau} = t(\vec{x}) = t$;

Однако важным являются не единичные потери, а средние при многократном использовании оценки вместо истинного значения оцениваемой функции. Поэтому введем функцию риска $R(t, \tau) = E(G(t, \tau))$ - это средние потери относительно выбранной функции потерь.

Часто функция потерь выбирается в виде $G(t, \tau) = (t - \tau)^2$ и называется квадратичной функцией потерь, а соответствующий риск $R(t, \tau) = E(t - \tau)^2$ – квадратичным риском.

Для несмещенных оценок квадратичный риск

$$R(t, \tau) = E(t - \tau)^2 = E(t - Et)^2 = Dt,$$

поэтому сравнение качества несмещенных оценок по квадратичному риску лежит в русле этого общего подхода и совпадает с их сравнением по эффективности.

5. Смещение оценки

Пусть $b = Et - \tau$. Тогда b называется смещением оценки t для $Et = \tau + b$. Если b=0, оценка называется несмещенной, если b>0 (b<0), то оценка в среднем завышает (занижает) истинное значение оцениваемой функции. В случае линейного смещения его легко устранить, т.е. подправить оценку по смещению. Пусть $Et = a\tau + b$, где a и b – const, тогда получаем, что $E\left(\frac{t-b}{a}\right) = \tau \Rightarrow \hat{\tau} = \frac{t-b}{a}$ – несмещенная оценка для τ.

6. Связь смещения, квадратичного риска и дисперсии оценки

Пусть $Et = a\tau + b \Rightarrow \tau = Et - b;$

Тогда $R(t,\tau) = E(t-\tau)^2 = E(t - Et + b)^2 = E(t - Et)^2 + 2bE(t - Et) + b^2$

$\Rightarrow R(t,\tau)Dt + b^2$ – эта формула часто упрощает вычисление квадратичного риска.

7. Достаточное условие состоятельности несмещенных и асимптотически несмещенных оценок

Теорема 1.

Для состоятельности несмещенной оценки t_n достаточно, чтобы $Dt_n \xrightarrow[n\to\infty]{} 0.$

Доказательство.

Пусть $Et_n = \tau$ и $Dt_n \xrightarrow[n\to\infty]{} 0$ воспользуемся неравенством Чебышева $P\{|\xi - E\xi| \ge \varepsilon\} \le \frac{D\xi}{\varepsilon^2}$, получив $\xi = t_n$ или, с учетом несмещенности t для $\tau : P\{|\xi - E\xi| \ge \varepsilon\} \le \frac{D\xi}{\varepsilon^2} \xrightarrow[n\to\infty]{} 0$, откуда и следует утверждение.

Теорема 2.

Для состоятельности асимптотически несмещенной оценки t для τ достаточно, чтобы $Dt_n \xrightarrow[n\to\infty]{} 0.$

Доказательство.

Пусть $Et_n = \tau + \varepsilon_n; \varepsilon_n, Dt_n \to 0; \tau = Et_n - \varepsilon_n$. Рассмотрим событие

$A = \{|t_n - \tau| \ge \varepsilon\} = \{|t_n - Et_n + \varepsilon_n| \ge \varepsilon\} \subset \{|t_n - Et_n| + |\varepsilon_n| \ge \varepsilon\} =$

$= \{|t_n - Et_n| \ge \varepsilon - |\varepsilon_n|\} = B$

тогда из того, что $A \subset B \Rightarrow P(A) \le P(B)$, тогда по неравенству Чебышева имеем $P\{|t_n - \tau| \ge \varepsilon\} \le P\{|t_n - Et_n| \ge \varepsilon - |\varepsilon_n|\} \le \frac{Dt_n}{(\varepsilon - |\varepsilon_n|)^2} \xrightarrow[n\to\infty]{} 0$, что и доказывает утверждение.

Замечание. Результаты этих теорем в указанных условиях часто упрощают установление состоятельности.

8. Выборочные моменты

Выборочные моменты – распространенный вид оцениваемых функций от неизвестного параметра θ распределения с.в. X. Приведем сначала наиболее общие формы выборочных моментов ($\vec{x} = x_1,...,x_n$) – выборка)

$\hat{L}_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$ – начальный r-ый выборочный момент; при r=1

$\hat{L}_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – выборочное среднее с.в. X;

$\bar{\mu}_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$ центральный r-ый выборочный момент; при r=2

$\bar{\mu}_2 = S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — выборочная дисперсия. При известном EX=m

$\hat{m}_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^r$ – выборочный r-ый центральный момент.

$K_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ – выборочная корреляция (здесь $x_1,...,x_n$; $y_1,...,y_n$ – выборки значений соответственно с.в. X и Y);

$\hat{\sigma} = \hat{\sigma}_X = \sqrt{S_1^2}$ – выборочное среднее квадратичное отклонение с.в. X,

$\hat{r}_{XY} = \frac{\hat{K}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$ – выборочный коэффициент корреляции;

$\hat{S}_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\hat{\sigma}} = \frac{\mu_3}{\hat{\sigma}}$ – выборочный коэффициент асимметрии;

$\hat{\varepsilon} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\hat{\sigma}} - 3 = \frac{\mu_4}{\hat{\sigma}} - 3$ – выборочный эксцесс.

9. Примеры

9.1. Примеры на определение свойств оценок

1) Проверить на состоятельность и несмещенность выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ для математического ожидания EX.

Решение.

$E\bar{x} = \frac{1}{n} \cdot n \cdot Ex = Ex$, то есть статистика $\bar{x}$ – несмещенная оценка для EX.

$D\bar{x} = \frac{1}{n^2} \cdot n \cdot D\bar{x} = \frac{D\bar{x}}{n}$ при $n \to \infty$, откуда следует и состоятельность статистики $\bar{x}$ для EX. Этот факт сразу следует и из теоремы Хинчина (ЗБЧ).

Замечание 1.Отсюда получаем, например, что статистика $\bar{x}$ есть несмещенная состоятельная оценка для параметра L распределения Пуассона П(L), параметра а нормального распределения N(а,б).

Замечание 2.Если при исследовании смещенности оценки T(x) получается линейная функция L от параметра θ, то для построения несмещенной оценки для θ нужно применить к оценке T(x) преобразование L^{-1}.

2) Проверить на состоятельность и несмещенность выборочную дисперсию $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ для дисперсии DX.

Решение.

Преобразуем выражение для S_1^2 :

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_1^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2. \qquad (1)$$

Тогда

$$ES_1^2 = \frac{1}{n}\sum_{i=1}^{n}Ex_i^2 - E\bar{x}^2 = \frac{1}{n}Ex^2 - (Dx + (Ex)^2) = Ex^2 - \frac{1}{n}(Dx) + (Ex)^2) =$$

$$= \frac{n-1}{n}Dx \xrightarrow[n\to\infty]{} Dx$$

то есть, S_1^2 – смещенная, но асимптотически несмещенная оценка для DX. Подправим S_1^2 оценку для DX. По замечанию 2 оценка $S_2^2 = \frac{n}{n-1}S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – несмещенная оценка для DX.

Состоятельность оценки S_1^2 для DX следует (по определению) из теоремы Хинчина, примененной к каждому слагаемому выражению (1).

3) Самостоятельно показать, что статистика $S_3^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ при известном значении EX = m является несмещенной для дисперсии.

В дальнейшем будем использовать обозначение L(X) – закон распределения с.в. X.

4) L(X) = $R[0,\theta]$. Проверить свойства оценки $\hat{Q} = 2x_1$ для θ.

Решение.

$E\hat{Q} = 2Ex_1 = \frac{2Q}{2} = Q;\ \hat{Q} = 2x_1$ от объема выборки n не зависит, поэтому при $n \to \infty$ не сходится к θ, то есть является несостоятельной оценкой для θ. Здесь мы имеем пример несмещенной и несостоятельной оценки для θ.

5) L(X) = $R[0,\theta]$. Проверить свойства оценки $\hat{Q} = 2x_{(n)}$ для θ. В случае смещенности подправить ее.

Решение.

$Ex_{(n)} = \int_0^Q x^n \frac{n}{Q^n} dx = \frac{n}{n+1} Q \to Q$, то есть данная оценка $\hat{Q}$ является смещенной, но асимптотически несмещенной. Подправим ее. По замечанию 2 получаем, что оценка $\tilde{Q} = \frac{n+1}{n}\hat{Q} = \frac{n+1}{n} x_{(n)}$ – несмещенная оценка для параметра θ.

Для исследования состоятельности оценки $\hat{Q}$ вычислим дисперсию оценки $\hat{Q}$:

$Ex_{(n)}^2 = \int_0^Q x^{n+1} \frac{n}{Q^n} dx = \frac{nQ^2}{n+2}$;

$Dx_{(n)} = \frac{nQ^2}{(n+2)(n+1)} - (Ex_{(n)})^2 = \frac{Q^2 n}{(n+1)^2} \xrightarrow[n\to\infty]{} 0$, откуда следует, что оценки $\hat{Q}$ и $\tilde{Q}$ являются состоятельными.

6) С.в. X распределена по закону Коши $f(x;Q) = \frac{1}{n}(1+(x-Q)^2)$. Состоятельна ли оценка $\hat{Q} = \sum_{i=1}^{n} x_i = \bar{x}$ для θ?

Решение.

Функция распределения закона Коши есть $F(x) = \frac{1}{2} + \frac{1}{\pi}\mathrm{arctg}(x-Q)$.

Характеристическая функция (х.ф.) с.в. X $g(t) = \exp\{\mathrm{int}\,Q - n|t|\}$

Х.ф. с.в. $Y = \sum_{i=1}^{n} x$; $g_Y(t) = \exp\{\mathrm{int}\,Q - n|t|\}$;

Х.ф. с.в. $Q = \frac{\sum_{i=1}^{n} x}{n}$; $g_Q(t) = g_Y\left(\frac{t}{n}\right) = \exp\{\mathrm{int}\,Q - |t|\}$, то есть с.в. X и θ имеют одинаковое распределение (данное распределение Коши).

$P\left\{\left|\hat{Q} - Q\right| > h\right\} = P\{|x_i - Q| > h\} = 1 - P\{-h + Q < x_i < h + Q\} =$

$1 - (F(h+Q) - F(-h+Q)) = 1 - \left(\frac{1}{2} + \frac{1}{\pi} arctgh\right) = 1 - \frac{2}{\pi} arctgh \quad \not\to 0$

При $n \to \infty$ это означает несостоятельность приведенной оценки $\hat{Q}$ для θ.

9.2. Примеры на простейшие свойства точечных оценок

1) Найти распределение и моменты статистики $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ если с.в. X ~ N (m,σ) , $X_1,\ldots,X_n$ – выборка значений λ. $g_{X_i}(t) = \exp\left(itm - \frac{\delta^2 t^2}{2}\right)$;

$g_{\overline{X}=\sum_{i=1}^{n}\frac{x_i}{n}}(t) =$

$= \left(g_{X_i}\left(\frac{t}{n}\right)\right)^n = \exp\left(i\frac{t}{n}mn - \frac{n\sigma^2 t^2}{2n^2}\right) = \exp\left(itm - \left(\frac{\sigma}{\sqrt{n}}\right)^2 \cdot \frac{t^2}{2}\right)$ => $\overline{X}$ ~ $N(m, \frac{\sigma}{\sqrt{n}})$ => $M\overline{X}$=m; $D\overline{X} = \frac{\sigma^2}{n}$.

Примеры, когда несмещенной оценки нет.

2) С.в. X ~ R[0,θ]; $\tau(\theta) = \frac{1}{\theta}$. Пусть t(x) – несмещенная оценка для

$\tau(\theta) = \frac{1}{\theta}$ тогда $\frac{1}{\theta} = \int_0^\theta \frac{t(x)}{\theta}dx \Rightarrow \int_0^\theta t(x)dx = 1 \Rightarrow t(x) = \frac{1}{\theta}$, но это не статистика

$\Rightarrow$ несмещенной оценки для $\tau(\theta) = \frac{1}{\theta}$ нет.

3) С.в. X ~ П(θ); $\tau(\theta) = \frac{1}{\theta}$. Пусть t(x)- несмещенная оценка для $\tau(\theta) = \frac{1}{\theta}$,

тогда $\frac{1}{\theta} = \sum_{x=0}^{\infty} t(x)\frac{\theta^x e^{-\theta}}{x!}$ или $\sum_{x=0}^{\infty} t(x)\frac{\theta^{x+1}}{x!} = e^\theta = \sum_{x=0}^{\theta}\frac{\theta^x}{x!} \Rightarrow \tau(x) = \frac{1}{\theta}$, т. е. это не статистика $\Rightarrow$ несмещенной оценки для $\tau(\theta) = \frac{1}{\theta}$ нет.

Примеры бесполезных (осциллирующих) несмещенных оценок.

4) С.в. X ~ Г(θ); $\tau(\theta) = \theta$; $P(X=x) = \theta(1-\theta)^{x-1}$, $\theta \in [0;1]$, x=1,2,… Пусть t(x) – несмещенная оценка для $\tau(\theta) = \theta$, тогда $\sum_{x=1}^{\infty} t(x)\,\theta(1-\theta)^{x-1} = \theta$ =>

$\sum_{x=1}^{\infty} t(x)\,(1-\theta)^{x-1}=1 \Rightarrow t(x)=\begin{cases}1, x=1\\ 0, x\geq 2\end{cases}$ - это «плохая» оценка для θ, т.к. θ – вероятность успеха в одном опыте и не равна нулю по смыслу, кроме того, оценка $t(x)$ дает нулевую вероятность успеха в одном опыте, если успех не происходит в первом опыте, что не соответствует действительности.

5) С.в. $X \sim П(\theta)$; $\theta \in [0;\infty)$; $\tau(\theta)=e^{-2\theta}$. Пусть $t(x)$ – несмещенная оценка для $\tau(\theta)$, тогда $\sum_{x=0}^{\infty} t(x)\frac{\theta^x e^{-\theta}}{x!}=e^{-2\theta} \Rightarrow \sum_{x=0}^{\infty} t(x)\frac{\theta^x}{x!}=e^{-\theta} \Rightarrow$

$\sum_{x=0}^{\infty} t(x)\frac{\theta^x}{x!}=\sum_{x=0}^{\infty}\frac{(-1)^x\theta^x}{x!} \Rightarrow t(x)=\begin{cases}1, x=2k\\ -1, x=2k+1\end{cases}$, k=0,1,2,…- это «плохая» оценка для $\tau(\theta)=e^{-2\theta}$, т.к. $e^{-2\theta}>0$, и $t(x)$ реагирует только на четность значения с.в. X.

6) С.в. $X \sim B(m=\theta, p)$; $\tau(\theta)=\theta$; $\hat\theta=\frac{x_1}{p}$ – предлагаемая оценка для θ.

$E\hat\theta=E\frac{x_1}{p}=\frac{1}{p}\theta p=\theta \Rightarrow \hat\theta$ – несмещенная оценка для θ; $D\hat\theta=E\hat\theta^2-(E\hat\theta)^2=$

$=\sum_{x=0}^{n}\frac{x^2}{p^2}C_\theta^x p^x(1-p)^{\theta-x}=\frac{1}{p^2}(\theta p(1-p)+\theta^2p^2)-\theta^2=$

$\frac{\theta(1-p)}{p}+\theta^2-\theta^2=\frac{\theta(1-p)}{p}$.

Если p близко к 1, то $D\hat\theta$ мала и $\hat\theta$ – «хорошая» оценка для θ,а при p малом $D\hat\theta$ велика и $\hat\theta$ – «плохая» оценка для θ. Значения $\hat\theta$ не целые, поэтому в качестве оценок для θ следует выбрать натуральные числа, ближайшие к $\frac{x_1}{p}$.

Примеры несостоятельных оценок.

7) С.в. $X \sim R[0,\theta]$; $\tau(\theta)=\theta$; $\hat\theta=2x_1$; $E\hat\theta=2Ex_1=2\frac{\theta}{2}=\theta \Rightarrow \hat\theta$ – несмещенная оценка для $\tau(\theta)$, но $\hat\theta$ - несостоятельная оценка, т.к. не зависит от n.

8) С.в. $X \sim П(\theta)$; $\hat\theta=x_1$; $E\hat\theta=\theta$, но $\hat\theta$ несостоятельная оценка, т.к. не зависит от n.

9) Пример несостоятельной оценки, зависящей от n. С.в. X ~ Коши с плотностью распределения $f(x,\theta)=\frac{1}{\pi(1+(x-\theta)^2)}$; $\hat\theta=\bar{x}$;

$F(x)=\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}(x-\theta)$; характеристическая функция $g_x(t)=e^{it\theta-|t|}$, тогда

$g_{\sum\limits_{i=1}^{n} x_i}(t)=(g_{x_i}(t))^n=\exp\{-int\theta-n|t|\}$; $g_{\bar{x}}(t)=g_{\sum\limits_{i=1}^{n} x_i}(\frac{t}{n})=e^{it\theta-|t|}=g_x(t)\Rightarrow$

$\{|\hat{\theta}-\theta|\geq\varepsilon\}$=P $\{|x_i-\theta|\geq\varepsilon\}$=1-P$\{-\varepsilon+\theta\leq x<\varepsilon+\theta\}$= 1-(F($\varepsilon+\theta$)-F($-\varepsilon+\theta$))=

$=1-\frac{2}{\pi}\operatorname{arctg}\varepsilon\not\to 0$ при $n\to\infty$, а это означает несостоятельность оценки $\hat{\theta}$.

Определение.
Несмещенная оценка с минимальной дисперсией называется **оптимальной**.

Теорема единственности оптимальной оценки.
Пусть $t_1=t_1(x)$ и $t_2=t_2(x)$ – две оптимальные оценки для $\tau=\tau(\theta)$ с дисперсией $Dt_1=Dt_2=\sigma^2$. Тогда $t_3=\frac{t_1+t_2}{2}$ тоже несмещенная оценка для $\tau(\theta)$, т.к. $Et_3=\frac{Et_1+Et_2}{2}=\frac{2\tau}{2}=\tau$.

$Dt_3=\frac{1}{4}(Dt_1+Dt_2+2K_{t_1t_2})=\frac{\sigma^2+K_{t_1t_2}}{2}\leq\sigma^2$, т.к. по неравенству Коши-Буняковского $|K_{t_1t_2}|\leq\sqrt{Dt_1Dt_2}$. Но σ^2 - минимальная возможная дисперсия несмещенных оценок для $\tau\Rightarrow Dt_3=\sigma^2\Rightarrow K_{t_1t_2}=\sigma^2\Rightarrow t_1$ и t_2 линейно зависимы $t_1=at_2+b$, где a и b – const. Из несмещенности оценок t_1 и $t_2\Rightarrow Et_2=aEt_1+b$ или $\tau=a\tau+b\Rightarrow a=1, b=0\Rightarrow t_1=t_2$, что и утверждалось.

9.3. Примеры на сравнение качества оценок по среднему квадратичному риску

10) X~R[0, θ]; $\tau=\tau(\theta)=\frac{\theta}{2}$; $\hat{\tau}_1=\frac{x_1}{2}$; $\hat{\tau}_2=\bar{x}$; $\hat{\tau}_3=\frac{x_{(n)}}{2}$. Проверим их на несмещенность:

$EQ_1=Ex=\frac{\theta}{2}; E\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}EX_i=n\cdot\frac{1}{n}\cdot\frac{\theta}{2}=\frac{\theta}{2}; E\frac{x_{(n)}}{2}=\int\limits_0^Q\frac{x}{2}f_n(x)dx=$

$=\{F_n(x)=F^n(x)\to\ f_n(x)=F_n{}'(x)=nF^{n-1}(x)f(x);\ f(x)=\frac{\theta}{2};\ F(x)=\frac{x}{\theta}$

при $x \in (0;\theta]$ $\}= \dfrac{n}{2\theta^n}\int_0^Q x_n dx = \dfrac{n\theta^{n+1}}{2\theta^4(n+1)} = \dfrac{n}{2(n+1)}\theta;$

$b_{\tau 3} = (\dfrac{n}{n+1} - 1)\theta = \dfrac{\theta}{n+1} < 0;$ $E(\dfrac{x_{(n)}}{2})^2 = \dfrac{n}{2\theta^n}\int_0^Q x^n dx = \dfrac{n\theta^{n+2}}{2\theta^n(n+2)} = \dfrac{n\theta^2}{4(n+2)};$

$D\hat{\tau}_3 = \dfrac{n\theta^2}{4(n+2)} - \dfrac{n^2\theta^2}{4(n+1)^2} = \dfrac{n\theta^2(n+1)^2 - n^2\theta^2(n+2)}{4(n+1)^2(n+2)} =$

$= \dfrac{n^3\theta^2 + 2n^2\theta^2 + n\theta^2 - n^3\theta^2 - 2n^2\theta^2}{4(n+1)^2(n+2)} = \dfrac{n\theta^2}{4(n+1)^2(n+2)};$ $b_{\tau_3} = E\hat{\tau}_3 - \tau = \dfrac{n\theta}{2(n+1)} - \dfrac{\theta}{2} =$

$-\dfrac{\theta}{2(n+1)} < 0;$ $R(\hat{\tau}_1,\tau) = R_1 = D\tau_1 = \dfrac{\theta^2}{48};$ $R(\hat{\tau}_2,\tau) = R_2 = D\tau_2 = \dfrac{\theta^2}{48n};$

$R(\hat{\tau}_3,\tau) = R_3 = D\hat{\tau}_3 + b^2_{\tau_3} = = \dfrac{n\theta^2}{4(n+1)^2(n+2)} + \dfrac{\theta^2}{4(n+1)^2} = \dfrac{\theta^2}{2(n+1)^2(n+2)}$. Теперь при n=1,2,… можно сравнивать риски данных оценок.

11) $X \sim N(a,\sigma^2)$. Сравнить по риску оценки для σ^2: $S_1^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ и

$S_2^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Решение.

Раньше получено, что S_1^2 – смещенная оценка для σ^2 со смещением $b = -\dfrac{\sigma^2}{n}$; S_2^2 – несмещенная оценка ($ES_1^2 = \dfrac{n-1}{n}\sigma^2; ES_2^2 = \sigma^2$).

Т.к. элементы выборки $x_1,...,x_n$ связаны равенством $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, то

$\dfrac{nS_1^2}{\sigma^2} \sim X^2_{n-1} \Rightarrow EX^2_{n-1} = n-1,\ DX^2_{n-1} = 2(n-1) \Rightarrow ES_1^2 = \dfrac{n-1}{n}\sigma^2 = \sigma^2 - \dfrac{\sigma^2}{n};$

$DS_1^2 = \dfrac{2\sigma^4(n-1)}{n^2};$ $S_2^2 = \dfrac{n}{n-1}S_1^2 \Rightarrow ES_2^2 = \sigma^2;$

$DS_2^2 = R_2 = R(S_2^2,\sigma^2) = \dfrac{2\sigma^4(n-1)\cdot n^2}{n^2\cdot(n-1)^2} = \dfrac{2\sigma^4}{n-1} \xrightarrow[n\to\infty]{} 0 \Rightarrow$ S_2^2 – состоятельная оценка для σ^2 (состоятельность S_1^2 для σ^2 определена раньше)

$R_n = R_1(S_2^2,\sigma^2) = DS_1^2 + b^2 = = \dfrac{2\sigma^4(n-1)}{n^2} + \dfrac{\sigma^4}{n^2} = \dfrac{2(n-1)\sigma^4}{n^2}$. Сравним риски R_1 и R_2: при n>1 $R_1 < R_2$, т.е. S_1^2 по риску лучше, чем S_2^2.

9.4. Примеры на построение «лучшей» оценки из данного класса

При выборе оценок часто ограничиваются рассмотрением определенного класса оценок. Тогда в пределах этого класса при установлении приоритетных требований ищется «лучшая» в указанном смысле оценка.

1) С.в. $X \sim N(a, \sigma^2 = \theta)$, θ – неизвестный параметр. В классе оценок $V(k) = \sigma^2(k) = \frac{k}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ построить оценку S_3^2 для σ^2 с наименьшим квадратичным риском. Найти $R(S_3^2, \sigma^2)$.

Решение.

Ранее рассматривались следующие оценки для σ^2: $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ и $S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Получено, что S_1^2 – смещенная, но асимптотически несмещенная оценка для σ^2 ($ES_1^2 = \frac{n-1}{n}\sigma^2, b_{S_1^2} = -\frac{\sigma^2}{n}$ – смещение оценки для S_1^2); S_2^2 – несмещенная оценка для σ^2. $\frac{nS_1^2}{\sigma^2} \sim X_{n-1}^2$ (выборка связана равенством $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$) $\Rightarrow EX_{n-1}^2 = n-1$; $DX_{n-1}^2 = 2(n-1) \Rightarrow ES_1^2 = \frac{n-1}{n}\sigma^2$;

$DS_1^2 = D(\frac{\sigma^2}{n}X_{n-1}^2) = \frac{2\sigma^4(n-1)}{n^2}$; $S_2^2 = \frac{n}{n-1}S_1^2 \Rightarrow ES_2^2 = \sigma^2$;

$DS_2^2 = \frac{\cdot n^2}{(n-1)^2}DS_1^2 = \frac{2\sigma^4}{n-1}$; $S_1^2 = \sigma^2\ (k = \frac{n-1}{n})$; $S_2^2 = \sigma^2\ (k = 1) \Rightarrow S_1^2$ и $S_2^2 \in V(k)$.

Найдем квадратичный риск оценки $\sigma^2(k)$:

$R(\sigma^2(k), \sigma^2) = D\sigma^2(k) + b^2_{\sigma^2(k)}$; $D\sigma^2(k) = D(kS_2^2) = \frac{2k^2\sigma^4}{n-1}$;

$b^2_{\sigma^2(k)} = E\sigma^2(k) - \sigma^2 = E(kS_2^2) - \sigma^2 = (k-1)\sigma^2 \Rightarrow R(\sigma^2(k), \sigma^2) =$

$$\frac{2k^2\sigma^4}{n-1}+(k-1)^2\sigma^4 = \frac{\sigma^4}{n-1}(2k^2+k^2(n-1)-2k(n-1)+n-1)=$$
$$=\frac{\sigma^4}{n-1}(k^2(n+1)-2k(n-1)+n-1).$$

2) Найдем $k=k_0(n)=k_0$, при котором f(k)=$(k^2(n+1)-2k(n-1)+n-1)$ достигает минимального значения и при этом значении $k=k_0$ риск оценки $\sigma^2(k)$ $R(\sigma^2(k),\sigma^2)$ будет минимален.

Искомое значение $k=k_0-$ абсцисса вершины параболы $f(k) \Rightarrow$

$\Rightarrow k_0=\dfrac{n-1}{n+1}$, тогда искомая оценка из V(k) с минимальным риском

$$S_3^2=\sigma^2\,(k_0=\frac{n-1}{n+1})=\frac{1}{n+1}\sum_{i=1}^{n}(x_i-\overline{x})^2 .$$

$$R(S_3^2,\sigma^2)=R(\sigma^2(k),\sigma^2)=\frac{\sigma^4}{n-1}((\frac{n-1}{n+1})^2(n+1)-\frac{2(n-1)}{n+1}+n-1)=\frac{2\sigma^4}{n-1}.$$

T_1 и T_2 – независимые оценки для $\tau(\theta)=\theta$: $ET_1=\theta+d_1$, $ET_2=\theta+d_2$, $d_1,d_2\neq 0$ (d_1 и d_2- const). В классе V: $\{\ T=aT_1+bT_2, |a|,|b|<\infty\ \}$ построить несмещенную оценку для θ и найти ее дисперсию, если $DT_1=\sigma_1^2$, $DT_2=\sigma_2^2$.

Решение.

$ET_1=(a+b)\theta+ad_1+bd_2$; $DT_1=a^2\sigma_1^2+b^2\sigma_2^2$; T – несмещенная оценка для

θ, если $\begin{cases} a+b=1 \\ ad_1+bd_2=0 \\ d_1\neq d_2 \end{cases} \Rightarrow \begin{cases} a^*=\dfrac{d_2}{d_1-d_2} \\ b^*=\dfrac{d_1}{d_1-d_2} \end{cases}$ Тогда $T=a^*T_1+b^*T_2=$

$$=\frac{d_2}{d_1-d_2}T_1+\frac{d_1}{d_1-d_2}T_2;\ R(T(\theta)=DT=(\frac{d_2}{d_1-d_2})^2\sigma_1^2+(\frac{d_1}{d_1-d_2})^2\sigma_2^2, d_1\neq d_2.$$

При $d_1=d_2$ несмещенной оценки для θ в классе V не существует.

Выводы.

а) пример 2 дает алгоритм построения несмещенной оценки по двум любым смещенным с разным смещением;

б) по примеру 2 при $d_1=d_2$ можно строить примеры, когда не существует несмещенной оценки для θ в классе V.

3) T_1 и T_2 - независимые оценки для $\tau(\theta)=\theta$; $DT_1=\sigma_1^2, DT_2=\sigma_2^2$. В классе V: $\{\ T=aT_2+(1-a)T_1, 0\le a\le 1\ \}$ найти несмещенную оценку T^* для θ наименьшей возможной дисперсией DT^*.

Решение.

$ET = aT_2 + (1-a)T_1 = \theta \Rightarrow T$ – несмещенная оценка для θ.

$DT = f(a) = a^2\sigma_2^2 + (1-a)^2\sigma_1^2$; достигает минимума при a, для которого $(DT)' = f(a)' = 0$ (т.к. коэффициент при a^2 положителен и равен $u = \sigma_1^2 + \sigma_2^2$). $f'(a) = 2a\sigma_2^2 + 2(1-a)\sigma_1^2 = 0 \Rightarrow a^* = \dfrac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$.

Таким образом, искомая оценка $T^* = \dfrac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} T_2 + \dfrac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} T_1$;

$$DT^* = \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_2^2 + \left(\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^2 \sigma_1^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

4) $x_1,\dots,x_n; y_1,\dots,y_m$ – две независимые выборки $\sim N(\mu(\sigma^2)$, (μ и σ^2 – неизвестные параметры). В классе V: $\{\ a\bar{x} + (1-a)\bar{y}, 0 \le a \le 1\ \}$ ($\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$) найти оценку T^* для μ с минимальной дисперсией DT^*.

Решение.

$ET = aE\bar{x} + (1-a)E\bar{y} = a\mu + (1-a)\mu = \mu \Rightarrow V$ – класс несмещенных оценок.

Тогда мы находимся в условиях задачи 3, из которой следует, что единственной несмещенной оценкой в классе V с наименьшей дисперсией является

$$T^* = \frac{D\bar{x}}{D\bar{x} + D\bar{y}}\bar{y} + \frac{D\bar{y}}{D\bar{x} + D\bar{y}}\bar{x} = \left\{\begin{matrix} D\bar{x} = \dfrac{\sigma^2}{n} \\ D\bar{y} = \dfrac{\sigma^2}{m} \end{matrix}\right\} = \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}\bar{y} + \frac{\frac{\sigma^2}{m}}{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}\bar{x} = \frac{m\bar{y} - n\bar{x}}{n+m}.$$

Из задачи 3 $DT^* = \dfrac{D\bar{x}D\bar{y}}{D\bar{x} + D\bar{y}}\bar{y} = \dfrac{\sigma^2}{n+m} \Rightarrow T^* \sim N(\mu(\dfrac{\sigma^2}{n+m})$.

Вывод.

Пример 4 дает алгоритм построения лучшей по риску несмещенной оценки по данным для μ при $X \sim N(\mu(\sigma^2)$: $\bar{x}$ и $\bar{y}$.

5) С.в. $X \sim B(1, p = \theta)$, θ – неизвестный параметр, $\tau(\theta) = \theta$. Сравнить квадратичные риски оценок $T_1 = \bar{x}$ и $T_2 = \bar{x} + \dfrac{0{,}5 - \bar{x}}{\sqrt{n} + 1}$ (оценка Ходжеста-Лемана) для θ.

Решение.

$ET_1=\theta; ET_2=\theta-\frac{\theta}{\sqrt{n}+1}+\frac{1}{2(\sqrt{n}+1)}\Rightarrow T_1-$ несмещенная оценка, а T_2- смещенная, но асимптотически несмещенная оценка для θ. Причем T_2 несмещенно оценивает лишь значение $\theta=0{,}5$. Сравним T_1 и T_2 по квадратичным рискам.

$R_1=R_1(T_1,\theta)=DT_1=\frac{\theta(1-\theta)}{n}; R_2=R_2(T_2,\theta)=DT_2+b_{T_2}^2$, где $b_{T_2}^2=ET_2-\theta$ $=-\frac{\theta}{\sqrt{n}+1}+\frac{1}{2(\sqrt{n}+1)}=\frac{1-2\theta}{2(\sqrt{n}+1)}$; $DT_1\xrightarrow[n\to\infty]{}0\Rightarrow T_1-$ состоятельная оценка для θ.

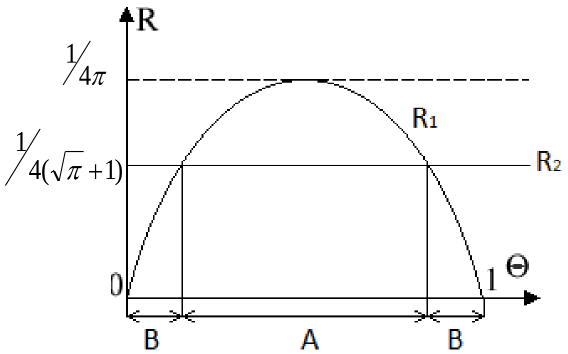
$$T_2=\bar{x}+\frac{0{,}5}{\sqrt{n}+1}+\frac{\bar{x}}{\sqrt{n}+1}=\frac{\sqrt{n}}{\sqrt{n}+1}\bar{x}+\frac{0{,}5}{\sqrt{n}+1};$$

$DT_2=\frac{n}{(\sqrt{n}+1)^2}\frac{Dx}{n}=\frac{\theta(1-\theta)}{(\sqrt{n}+1)^2}{}_1\xrightarrow[n\to\infty]{}0\Rightarrow T_2-$ состоятельная оценка для θ.

$R_2=\frac{1-2\theta}{2(\sqrt{n}+1)}+\frac{(1-2\theta)^2}{4(\sqrt{n}+1)^2}=\frac{1}{4(\sqrt{n}+1)^2}$, т.е. R_2 не зависит от θ.

Изобразим R_1 и R_2 графически как функции от θ.

Из графика следует, что, если есть основание считать, что значение θ близко к 0,5 (средняя зона А), то по риску лучше оценка T_2 ,если же значение близко к 0 или к 1 (крайние зоны В), то лучше оценка T_1 для θ.



§3. Достаточные статистики (д.с.)

Определение. Статистика T=T(x) называется достаточной для семейства распределений $F_\theta(x)$ (θ – неизвестный параметр), если вероятность любой выборки $\vec{x} = (x_1,...,x_n)$ не зависит от значения неизвестного параметра θ. (смысл д.с.: содержит в себе всю информацию о неизвестном параметре θ).

Определение. Функцией правдоподобия (ф.п.) $L_\theta(\vec{x}) = L$ называется вероятность данной выборки $\vec{x} = (x_1,...,x_n)$, т.е.

$$L = \begin{cases} \prod_{i=1}^{n} P(X = x_i), \text{ если с.в.X дискретна;} \\ \prod_{i=1}^{n} f(x_i), \text{ если с.в.X непрерывна, } f(x) - \text{плотность ее распределения;} \end{cases}$$

Для опознавания и построения д.с. приведём **критерий факторизации**:

Теорема 1. (Критерий факторизации (к.ф.)).

Для того чтобы T(x) была д.с. для семейства распределения $F_\theta(x)$ (θ – неизвестный параметр), необходимо и достаточно чтобы функция правдоподобия $L_\theta(x_1,...,x_n)$ имела вид $L = \varphi(t(x) = T, \theta)h(x)$. (1)

Тогда из критерия следует, что, приводя к виду (1), получим вид д.с. {из вида (1) следует, что д.с. определяется неоднозначно => достаточных статистик бесконечно много.}

Д.с. существует, так как, например, выборка является д.с..

Доказательство.

1) **Достаточность**.

Пусть (1) выполнено, тогда покажем, что T(x) является д.с., то есть распределение любой выборки при её фиксированном значении не зависит от параметра θ.

$$P\{X = \vec{x}/T(x) = T\} = \{\text{по теореме умножения}\} =$$

$$= \frac{P(X = \vec{x}, T(x) = T)}{P(T(x) = T)} = \frac{L}{\sum_{\vec{y}:T(y)=T} L(\vec{y})} =$$

$$= \frac{\varphi(T(x) = T, \theta)h(x)}{\sum_{\vec{y}:T(y)=T} \varphi(T, \theta)h(y)} = \frac{h(x)}{\sum_{\vec{y}:T(y)=T} h(y)} \text{ не зависит от } \theta.$$

2) **Необходимость**.

Пусть T(x) – д.с. Тогда покажем, что (1) имеет место:

$$P\{X=\vec{x}/T(x)=T\}=\frac{P(X=\vec{x},T(x)=T)}{P(T(x)=T)}=\frac{P(X=\vec{x})}{P(T(x)=T)}=h(x)$$ тогда по теореме умножения имеем: $P(X=\vec{x})=L=P(X=\vec{x}/T(x)=T)P(T(x)=T)\equiv(1)$ (т.к. $P(T(x)=T)=\varphi(T,\theta)$), а $P(X=\vec{x}/T(x)=T)=h(x)$ не зависит от θ из определения д.с. $T(x)$.

Пример применения критерия факторизации:

С.в. $X\sim\pi$ ($\lambda=\theta$) Найти д.с. для этого семейства.

Решение.

$L=\prod_{i=1}^{n}P(X=x_i)$ – функция правдоподобия в дискретном случае.

$$L=\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!}=\frac{\theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}}{\prod_{i=1}^{n}x_i!}\quad\Rightarrow\quad T(x)=\sum_{i=1}^{n}x_i \text{ - д.с. (т.к. } \theta^{\sum_{i=1}^{n}x_i}e^{-n\theta}=\varphi(T(x),\theta))$$

(проверка по определению проведена ниже).

Теорема 2 (Рао – Блэкуэлла - Колмогорова).

Оптимальная оценка, если она существует, является функцией от достаточной статистики.

(Оптимальная оценка – несмещённая оценка (н.о.) с минимально возможной дисперсией).

Доказательство.

Пусть $T=T(x)$ – д.с. для $F_\theta(x)$ и $T_1=T(x)$ – н.о. для $\tau=\tau(\theta)$.

Рассмотрим статистику $H(t)=E(T_1/T)$ (при $T=t$ имеем $H(t)=E(T_1/T=t)$ (1))

Тогда $EH(t)=E(E(T_1/T))=ET_1=\tau$ – н.о. для τ.

Теперь сравним $DH(t)$ с DT_1: $DT_1=E(T_1-ET_1)^2=E(T_1-\tau)^2=E(T_1-H(T)+$
$+H(T)-\tau)^2=E(T_1-H(T))^2+2E(T_1-H(T))(H(T)-\tau)+E(H(T)-\tau)^2$.

Если показать, что $E(T_1-H(T))(H(T)-\tau)=0$ (♦), то, так как $E(H(T)-\tau)^2\geq 0$, а $E(\tau-H(T))^2=DH(T)$, получим $DT_1\geq DH(T)$. Остаётся доказать (♦):

$E(T_1-H(T))(H(T)-\tau)=E((T_1-H(T))(H(T))-\tau E(T_1-H(T))$;

$E(T_1-H(T))=ET_1-EH(T)=\tau-\tau=0$;

По формуле полной вероятности при гипотезах $\{T=t\}$ получаем:

$E((T_1-H(T))(H(T))=\int_{-\infty}^{\infty}(E(T_1/T=t)-H(t))H(t)g(t)dt=0$, где – $g(t)$ плотность распределения статистики $H(t)$, так как $E(T/T=t)-H(t)=0$ по (1).

Теорема 2 доказана.

Определение полноты достаточной статистики.
Достаточная статистика T(x) называется **полной**, если для любой функции f(T(x))=f(T) из того, что Ef(T)=0 для всех θ следует, что f(T)=0 почти всюду (то есть в отдельных областях она не ноль, но эти области имеют меру ноль).

Теорема 3.
Если существует полная д.с., то вся функция от неё является оптимальной оценкой для своего математического ожидания.
Доказательство.
Пусть T=T(x) – полная д.с.; H(t) – произвольная функция от T(x); $E_\theta H(T)=\tau(\theta)$.
Тогда из определения полноты д.с. T(x) следует единственность для н.о. H(T) для τ(θ), так как в противном случае существует другая $H_1(T)$ н.о. $E(H(T) - H_1(T))=0 \Rightarrow H(T) = H_1(T)$. п.в. (из определения полноты д.с.)
Из предыдущей теоремы 2 следует, что оптимальную оценку следует искать в классе функций, зависящих от д.с. – T(x). Но так как H(T) – единственная н.о. для τ(θ), зависимая от T, то она является оптимальной оценкой для τ(θ).

Следствия из теорем.
Пусть T=T(x) – полная д.с.; τ(θ) – оцениваемая функция. Тогда:
1) когда оптимальная оценка существует, она является функцией д.с. и однозначно определяется из уравнения несмещённости: EH(T) = τ(θ) (где T – полная д.с.);
2) оптимальная оценка, если она существует, ищется по формуле
$\tau^* = H(T) = E(T_1/T=t)$; где T_1 – н.о. для τ(θ).

Задачи.
Доказать, что:
1) любая взаимно однозначная функция W(x) от д.с. – тоже является д.с.
Доказательство.
Пусть – T(x) д.с., тогда P{x / T(x)=t}=P{x/W(T(x)) = =W(t)}, что и требовалось доказать.

2) любая выборка – д.с.
Доказательство.
$P\left\{\frac{x_1,\ldots,x_n}{x_1',\ldots,x_n'}\right\}=1$, то есть не зависит от θ.

3) вариационный ряд – д.с.
Доказательство следует из задач 1), 2).

4) эмпирическая функция распределения – д.с.
Доказательство следует из задач 1), 3).

5) для семейства распределений $\Pi(\theta)$найти д.с. и проверить её по определению.
Решение.
Функция правдоподобия имеет вид

$$L(x)=\frac{e^{-\theta}\theta^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i}, \text{ по к.ф. получим } T(x)=\sum_{i=1}^{n}x_i \text{ - д.с.}$$

Проверка.

$$P\left\{\vec{x}/T(x)=\sum_{i=1}^{n}x_i=t\right\}=\frac{P\left\{X=x_1,\ldots,X=x_{n-1},X=t-\sum_{i=1}^{n-1}x_i\right\}}{P\left\{\sum_{i=1}^{n}x_i=t\right\}}=$$

$$=\frac{e^{-\theta}\theta^{x_1}}{x_1!}\ldots\frac{e^{-\theta}\theta^{t-\sum_{i=1}^{n-1}x_i}}{\left(t-\sum_{i=1}^{n-1}x_i\right)!}\frac{t!}{\theta^t n^t e^{-n\theta}}=\frac{t!}{n^t\left(t-\sum_{i=1}^{n-1}x_i\right)!\prod_{i=1}^{n-1}(x_i)!}, \text{ то есть не зависит от}$$

θ.

6) найти д.с. для семейства распределений $R[0,\theta]$ и проверить её по определению.
Решение.
Функция правдоподобия имеет вид

$$L(x)=\underbrace{(1/\theta^n)W(\theta-x_{(n)})}_{\varphi(T(x)=x_{(n)},\theta)}\underbrace{1}_{h(x)}, \text{ где } W(x)=\begin{cases}1, x\geq 0\\ 0, x<0\end{cases} \Rightarrow\ x_{(n)} \text{ – д.с.}$$

Проверка.

$$f(x_1,\ldots,x_n/x_{(n)}=t)=\frac{1}{t}, \text{ то есть не зависит от } \theta.$$

7) найти д.с. для семейства γ – распределений с плотностью распределения: $f(x,\alpha,\lambda)=\lambda^{\alpha}e^{-\lambda x}x^{\alpha-1}\dfrac{1}{\Gamma(\alpha)}$, $0<x<\infty$.

$$L(x)=\lambda^{n\alpha}\exp\left\{-\lambda\sum_{i=1}^{n}x_i\right\}(x_1,\ldots,x_n)^{\alpha-1}\left(\frac{1}{\Gamma(\alpha)}\right)^n W(x_{(n)})$$, по к.ф. получим

$$T(x)=\sum_{i=1}^{n}x_i$$ - д.с..

8) **экспоненциальным семейством** (э.с.) называется семейство распределений с плотностью f(x) или вероятностью P(X=x) вида exp{A(Q)B(x)+C(Q)+D(X)}. Найти д.с. э.с.

Решение.

$$L(x)=\exp\left\{A(Q)\sum_{i=1}^{n}B(x_i)+nC(Q)\right\}\exp\left\{\sum_{i=1}^{n}D(x_i)\right\}$$, по к.ф. получим

$$T(x)=\sum_{i=1}^{n}B(x_i)$$ - д.с., так как $\exp\left\{A(Q)\sum_{i=1}^{n}B(x_i)+nC(Q)\right\}=\varphi(T(x=x,\theta))$;

$$\exp\left\{\sum_{i=1}^{n}B(x_i)\right\}=h(x).$$

К э.с. относится γ – распределение, биномиальное, пуассоновское, геометрическое, Паскаля и другие. Методом приведения к виду э.с. с учётом результата задачи 8 найти д.с. для перечисленных выше законов распределения.

Нахождение д.с. путём приведения к экспоненциальному виду.

9) с.в. $X\sim\gamma(x;\alpha;\theta)$;

$$f(x;\alpha;\theta)=\begin{cases}\dfrac{\theta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-x\theta}, x\ge 0\\ 0, x<0\end{cases}\qquad \theta \text{ – неизвестно;}$$

$$L=\frac{\theta^{n\alpha}}{[\Gamma(\alpha)]^n}(x_1,\ldots,x_n)^{\alpha-1}e^{-\theta\sum_{i=1}^{n}x_i}X(x_{(1)})=(\bullet);\quad X(z)=\begin{cases}1, z>0\\ 0, z\le 0\end{cases};\quad \min_{\underline{i=1,n}}x_i=x_{(1)}.$$

$$(\bullet)=\exp\left\{\ln\frac{\theta^{n\alpha}}{[\Gamma(\alpha)]^n}(x_1,\ldots,x_n)^{\alpha-1}e^{-\theta\sum_{i=1}^{n}x_i}X(x_{(1)})\right\}=$$

$$=\exp\left\{n\alpha\ln\theta-n\ln\Gamma(\alpha)+(\alpha-1)\sum_{i=1}^{n}\ln x_i-\theta\sum_{i=1}^{n}x_i+\ln X(x_{(1)})\right\}=$$

$$= \exp\left\{ \underbrace{-\theta\sum_{i=1}^{n} x_i}_{A(Q)\sum B(x_i)} + \underbrace{n\alpha\ln\theta}_{C(Q)} + \underbrace{n\ln\Gamma(\alpha) + \ln X(x_{(1)})}_{D(X)} \right\} => \sum_{i=1}^{n} B(x_i) = \sum_{i=1}^{n} x_i - \text{ д.с.}$$

10) с.в. $X \sim B(n,\theta)$;

$$f_0 = P(X = x_i) = C_n^{x_i}\theta^{x_i}(1-\theta)^{n-x_i} = \exp\left\{x\ln\theta + (n-x)\ln(1-\theta) + \ln C_n^{x_i}\right\} =$$

$$= \exp\left\{ \underbrace{x[\ln\theta - \ln(1-\theta)]}_{B(x)A(\theta)} + \underbrace{n\ln(1-\theta)}_{C(\theta)} + \underbrace{\ln C_n^{x_i}}_{D(x)} \right\} => \sum_{i=1}^{n} B(x_i) = \sum_{i=1}^{n} x_i - \text{ д.с.}$$

11) с.в. $X \sim \Gamma_{\alpha,\lambda}$, при $x \ge 0$;

$$f(x;\alpha;\theta) = \frac{\theta^\alpha x^{\alpha-1} e^{-x\theta}}{\Gamma(\alpha)} = \{x \ge 0\} = \exp\{-\theta - (\alpha-1)\ln\theta + \alpha\ln\theta - \ln\Gamma(\alpha)\} =$$

$$= \exp\left\{ \underbrace{-\theta x}_{A(\theta)B(x)} + \underbrace{\alpha\ln\theta}_{C(\theta)} - \underbrace{(\alpha-1)\ln x - \ln\Gamma(x)}_{D(x)} \right\} => \sum_{i=1}^{n} x_i - \text{ д.с..}$$

12) с.в. $X \sim N(\theta,\sigma^2)$; $f_0(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\} =$

$$= \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{ \underbrace{\frac{x^2}{2\sigma^2}}_{D(x)} + \underbrace{\frac{x\theta}{\sigma^2}}_{A(\theta)B(x)} - \underbrace{\frac{\theta^2}{2\sigma^2}}_{C(\theta)} \right\} => \sum_{i=1}^{n} x_i - \text{ д.с..}$$

13) $X \sim \pi(\theta)$;

$$P(X = x) = \frac{e^{-\theta}\theta^x}{x!} = \{x = 0,1,2,\ldots\} = \exp\left\{ \underbrace{-\theta}_{C(\theta)} + \underbrace{x\ln\theta}_{B(x)A(\theta)} - \underbrace{\ln(x!)}_{D(x)} \right\} => \sum_{i=1}^{n} x_i - \text{ д.с.,}$$

где x=0,1,2,..

Задачи на полную д.с.

14) с.в. $X \sim B(1,\theta)$, где θ – неизвестно, $\theta \in (0,1)$. Найти д.с. и проверить её на полноту.

Решение.

Функция правдоподобия $L(x)$ имеет вид

$$L(x)=\underbrace{\theta^{\sum\limits_{i=1}^{n}x_i}(1-\theta)^{n-\sum\limits_{i=1}^{n}x_i}}_{\varphi(T(x)=\sum\limits_{i=1}^{n}x_i,\theta)}\underbrace{1}_{h(x)} \Rightarrow r=\sum_{i=1}^{n}x_i - \text{ д.с.}$$

С.в. $r\sim B(n,\theta)$. Пусть $f(r)=f\left(\sum\limits_{i=1}^{n}x_i\right)$ такова, что $E_\theta(f(r))=0$ при всех $\theta\in(0,1)$, то есть $E_\theta f(r)=\sum\limits_{r=0}^{n}f(r)C_n^r\theta^r(1-\theta)^{n-r}=0 \Rightarrow f(r)=0$ почти всюду как коэффициенты многочлена, то есть r – полная д.с.

Теорема 4.
Необходимым и достаточным условием полной д.с., относящейся к э.с., является совпадение размерности статистики и неизвестного параметра.

Примеры на теорему 3.

15) с.в. $X\sim B(1,\theta), r=\sum\limits_{i=1}^{n}x_i$. По теореме 3 установить, для функции от θ будут ли эффективными следующие оценки:

а) $\frac{r}{n}$; б) $\frac{r(r-1)}{n(n-1)}$; в) $r\frac{r-n}{r}(n-1)$.

Решение.

а) $E\left(\frac{r}{n}\right)=E\left(\frac{\sum\limits_{i=1}^{n}x_i}{n}\right)=\theta$, то есть для θ;

б) $E\left(\frac{r(r-1)}{n(n-1)}\right)=\frac{1}{n(n-1)}E\left[\left(\sum\limits_{i=1}^{n}x_i\right)^2-\left(\sum\limits_{i=1}^{n}x_i\right)\right]=\frac{n(n-1)\theta^2}{n(n-1)}=\theta^2$, так как

$$\left(\sum_{i=1}^{n}x_i\right)^2=D\left(\sum_{i=1}^{n}x_i\right)+\left[E\left(\sum_{i=1}^{n}x_i\right)\right]^2=n(1-\theta)\theta+n^2\theta^2=n\theta-n\theta^2+n^2\theta^2.$$

в) решить самостоятельно.
Решить самостоятельно задачи.

16) с.в. $X\sim R[0,\theta]$, (θ - неизвестно). Показать, что $T(x)=x_{(n)}$ – полная д.с.

17) с.в. $X\sim R[\theta_1,\theta_2]$, ($\theta_1$, θ_2 - неизвестны). Показать, что $T(x)=\left(x_{(1)},x_{(n)}\right)$ – полная д.с.

Примеры полных и неполных достаточных статистик.

1) С.в. $X \sim R[\theta_1,\theta_2]$; $x_1,\dots,x_n$ – выборка значений с.в. X. Доказать, что $T(x)=\left(x_{(1)},x_{(n)}\right)$ – полная д.с.

Решение.

$$L = W(\theta_2 - x_{(n)})\frac{W(x_{(1)}-\theta_1)}{(\theta_2-\theta_1)^2}, \text{ где } W(z)=\begin{cases}0, z<0\\ 1, z\ge 0\end{cases}.$$

По к.ф. получаем, что $T(x)=\left(x_{(1)},x_{(n)}\right)$ – д.с..

$f_{1,n}(u,v)=n(n-1)f(u)f(v)[F(v)-F(u)]^{n-2}$ – плотность совместного распределения с.в. $x_{(1)}$ и $x_{(n)}$; F(x) – функция распределения с.в. X:

$$F(x)=\frac{(x-\theta_1)}{(\theta_2-\theta_1)^2} \text{ при } \theta_1 \le x \le \theta_2;$$

$$f_{1,n}(u,v)=\frac{n(n-1)}{(\theta_2-\theta_1)^n}(v-u)^{n-2}, \theta_1 \le u < v \le \theta_2.$$

пусть $g(t_1,t_2)$ такова, что для всех $\theta_1<\theta_2$ выполняется соотношение:

$$Eg(t_1,t_2)=\iint\limits_{D:\theta_1\le t_1<t_2\le\theta_2} g(t_1,t_2)f_{1,n}(t_1,t_2)dt_1dt_2=0=$$

$$\iint\limits_D g(t_1,t_2)(t_2-t_1)^{n-2}dt_1dt_2=0 \text{ для всех } \theta_1<\theta_2.$$

Продифференцировав последнее выражение сначала по θ_1, потом по θ_2, получим: $g(t_1,t_2)$=0, откуда следует полнота д.с. T (x).

2) с.в. $X \sim \Pi(\lambda)$; $x_1,\dots,x_n$ – выборка значений с.в. X. Доказать, что статистика $T(x)=S=\left(S_{1l}=\sum_{i=1}^{l}x_i, S_{2l}=\sum_{i=l+1}^{n}x_i\right)$ – д.с., но неполная.

Решение.

Функция правдоподобия L(x) имеет вид

$$L=\prod_{i=1}^{n}\frac{\lambda^{x_i}\exp\{-\lambda\}}{x_i!}=\frac{\lambda^{S_{1l}}\lambda^{S_{2l}}\exp\{-n\lambda\lambda}{\prod_{i=1}^{n}x_i!} \Rightarrow T(x) \text{ – д.с. по к.ф.}$$

Докажем неполноту д.с. T(x). Рассмотрим функцию

$U(S)=\dfrac{S_{1l}(n-1)}{1-S}\ne 0$, в то время как MS(U)=0 => T(x)=S – неполная д.с.

3) с.в. $X \sim R[a,b]=R[a_1Q+b_1,a_2Q+b_2]$. Доказать, что статистика $T(x)=\left(x_{(1)},x_{(n)}\right)$ – достаточная, но неполная.

Решение.

Ранее получено, что $Ex_{(1)}=a+\frac{b-a}{n-1}, Ex_{(n)}=b-\frac{b-a}{n-1}.$

$Ex_{(1)}=Q\frac{a_1n+a_2}{n+1}+\frac{b_1n+b_2}{n+1}, Ex_{(n)}=Q\frac{a_2n+a_1}{n+1}+\frac{b_2n+b_1}{n+1};$

Рассмотрим функцию

$W(T(x))=\frac{(n+1)x_{(1)}-(b_1n+b_2)}{a_1n+a_2}+\frac{(n+1)x_{(n)}-(b_2n+b_1)}{a_2n+a_1}=0,$ в то время как

E(T(x))=0 => T(x) – неполная д.с.

Решить самостоятельно аналогичные задачи 4 и 5 для случаев:

4) с.в. $X\sim B(1,Q)$; $T(x)=\left(S_{1l}=\sum_{i=1}^{l}x_i, S_{2l}=\sum_{i=l+1}^{n}x_i\right)$;

5) с.в. $X\sim R\left[Q-\frac{1}{2},Q+\frac{1}{2}\right]$; $T(x)=\left(x_{(1)},x_{(n)}\right)$ (решить с использованием задачи 3 и непосредственно);

6) с.в. $X\sim R[0,Q]$; Доказать полноту д.с. $T(x)=x_{(n)}$;

7) с.в. $X\sim R[Q_1,Q_2]$; Используя теорему 3 (о полных д.с.), определить, для каких функций от неизвестных параметров Q_1,Q_2 будут эффективными оценки:

а) $\frac{(x_{(1)}+x_{(n)})}{2}$; б) $(n+1)\frac{(x_{(1)}-x_{(n)})}{n-1}$.

Д.с. могут быть использованы для улучшения имеющихся н.о. неизвестного параметра θ в соответствии с теоремой 2.

Примеры.

1) С.в. $X\sim \Pi(\lambda)$; $Q=x_1,x_1,\ldots,x_n$ - выборка значений с.в. X. Улучшить оценку для Q.

Решение.

$ET_1=Ex=Q \Rightarrow Q=T_1-$ н.о. $T(x)=\sum_{i=1}^{n}x_i-$ д.с.

$$H(t) = \left(E\frac{x_1}{\sum_{i=1}^{n} x_i} = t \right) = \sum_{k=0}^{t} kP\left\{ x_1 = \frac{k}{\sum_{i=1}^{n} x_i} = t \right\} = \sum_{k=0}^{t} \frac{kP\left\{ x_1 = k, \sum_{i=2}^{n} x_i = t - k \right\}}{P\left\{ \sum_{i=1}^{n} x_i = t \right\}} =$$

$$\sum_{k=0}^{t} k \frac{Q^k \exp\{-Q\}}{k!} \frac{Q^{t-k}(n-1)^{t-k}\exp\{-(n-1Q)\}}{(t-1)!} \frac{t!}{Q^t n^t \exp\{-nQ\}} =$$

$= \sum_{k=0}^{t} kC_t^k \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{t-k} = \frac{t}{n}$, так как последняя сумма = EZ, где

$Z \sim B\left(t, \frac{1}{n}\right)$, то

$$DH(t) = D\left(\frac{t}{n}\right) = \frac{D\left(\sum_{i-1}^{n} x_i\right)}{n} = \frac{nQ}{n^2} = \frac{Q}{n}; \quad DT_1 = Dx_1 = Q => \begin{matrix} \{DT_1 = DH(t), n = 1 \\ \{DT_1 > DH(t), n > 1 \end{matrix}.$$

Решить самостоятельно аналогичные задачи в случаях:

1) с.в. $X \sim B(1, Q); T_1 = x_1;$
2) с.в. $X \sim R[0, Q]; T_1 = 2x_1.$

§4. Неравенство Рао-Крамера (НРК)

Напомним определение функции правдоподобия $L = L(x) = L(\vec{x}) = L_\theta(\vec{x})$:

$$L(\vec{x}) = \begin{cases} \prod_{i=1}^{n} P(x = x_i), & \text{если с.в. } X - \text{дискретная} \\ \prod_{i=1}^{n} f(x_i), & \text{если с.в. } X - \text{непрерывная} \end{cases}.$$

Эта функция обозначает вероятность данной выборки и называется **функцией правдоподобия.**

Теперь сформулируем **условия регулярности** семейства распределений $x \sim F_\theta(x)$:

1. Дифференцируемость L.
2. $\{x:\ L \neq 0\}$ не зависит от неизвестного параметра θ .

Теорема 1. Неравенство Рао-Крамера дает нижнюю грань дисперсий несмещенных оценок в регулярном случае.

Рассмотрим задачу точечного оценивания при $x \sim F_\theta(x)$, $\tau = \tau(\theta)$, тогда, если выполняются условия регулярности и $t(x)$ – несмещенная оценка для $\tau = \tau(\theta)$, то дисперсия не может быть как угодно малой при построении разных оценок, то есть $Dt(x) \geq \dfrac{(\tau')^2}{E\left(\dfrac{d\ln L}{d\theta}\right)^2}$. (1)

Если в неравенстве (1) достигается равенство, то оценка $t(x)$ называется несмещенной оценкой с минимальной дисперсией (НОМД) или **эффективной оценкой**.

Доказательство неравенства Рао-Крамера опирается на преобразованное неравенство Коши-Буняковского (НКБ).

Преобразуем неравенство Коши-Буняковского (НКБ)

$|r_{xy}| \leq 1$, причем $|r_{xy}|=1$ тогда и только тогда, когда с.в. X и Y линейно зависимы. Рассмотрим неравенство $|r_{xy}| \leq 1$ при условии MX=0 и MY=0. Получим $\left|r_{xy}\right| = \dfrac{\left|K_{xy}\right|}{\sqrt{DX \cdot DY}} \leq 1$ (по определению r_{xy}), тогда $\left|K_{xy}\right| \leq \sqrt{DX \cdot DY}$. (2)

$K_{xy} = EXY - EX \cdot EY = EXY$; $DX = EX^2 - (EX)^2 = EX^2$ и $DY = EY^2$.

Подставляя значения дисперсии в формулу (2), имеем:

$\left|EXY\right| \leq \sqrt{EX^2 \cdot EY^2}$ или $(EXY)^2 \leq EX^2 \cdot EY^2$. (*)

В этой форме и будем использовать неравенство Коши-Буняковского.

План доказательства:

1. Выписать условие нормировки и продифференцировать его по θ.

2. Для несмещенной оценки $t(x)$ для $\tau=\tau(\theta)$ выписать уравнение несмещенности и продифференцировать обе части этого уравнения по θ.
3. Из результата пункта 2 вычесть результат пункта 1, умноженный на τ.
4. Применить НКБ в форме (*).

Проводим доказательство НРК по плану.

1. $\int\limits_{R^n} Ld\vec{x}=1$, т.е. $0=(1)'=\int\limits_{R^n}\left(\frac{dL}{d\theta}\cdot\frac{1}{L}\right)\cdot Ld\vec{x}=\int\limits_{R^n}\frac{d\ln L}{d\theta}\cdot Ld\vec{x}=E\left(\frac{d\ln L}{d\theta}\right)$,

т.е. $E\left(\frac{d\ln L}{d\theta}\right)=0$.

(3)

2. $t=t(x)$ – несмещенная оценка для $\tau=\tau(\theta)$.

Выпишем уравнение несмещенности: $\int\limits_{R^n} tLd\vec{x}=\tau$ и продифференцируем его по θ:

$$\tau'=\int\limits_{R^n} t\cdot\frac{dL}{d\theta}\cdot\frac{1}{L}\cdot Ldx=\int\limits_{R^n} t\cdot\frac{d\ln L}{d\theta}\cdot Ldx=E\left(t\cdot\frac{d\ln L}{d\theta}\right),\text{т.е. } E\left(t\cdot\frac{d\ln L}{d\theta}\right)=\tau'. \quad (4)$$

3. Обозначим $u=\frac{d\ln L}{d\theta}$ и вычтем из уравнения (4) уравнение (3), умноженное на τ, получим:

$E(u\cdot t)-\tau\cdot Eu=\tau'$, воспользуемся свойствами мат. ожидания EX

$E(ut-\tau u)=E(u(t-\tau))=\tau'$, возведем обе части равенства в квадрат

$(E(u(t-\tau)))^2=(\tau')^2$.

4. Воспользуемся НКБ в форме (*); это возможно, т.к. $Eu=0$ по формуле (3), а $E(t-\tau)=0$, т.к. t – несмещенная оценка для τ.

$(\tau')^2\le Eu^2\cdot E(t-\tau)^2$, но т.к. $E(t-\tau)^2=Et^2-(Et)^2=Dt$, то $Dt\le\frac{(\tau')^2}{Eu^2}$, но

$u=\frac{d\ln L}{d\theta}$, поэтому $Dt\le\frac{(\tau')^2}{E\left(\frac{d\ln L}{d\theta}\right)^2}$, что и требовалось доказать.

Критерий НОМД(Критерий эффективности (КЭ))

Дисперсия будет наименьшей, если знак неравенства заменить знаком равенства, а это по неравенству Коши-Буняковского возможно только в том случае, если $u=\frac{d\ln L}{d\theta}$ и $t-\tau$ линейно зависимы, т.е. искомый критерий формулируется следующим образом: $\frac{d\ln L}{d\theta}=a(\theta)(t-\tau)+b$. Покажем, что $b=0$. Действительно, возьмем математическое ожидание от обеих частей последнего равенства. Тогда $E\frac{d\ln L}{d\theta}=0$ по (3), а $E(t-\tau)=0$, т.к. $t-$ несмещенная оценка для τ, откуда получаем, что $в=0$, тогда критерий НОРД окончательно имеет вид $\frac{d\ln L}{d\theta}=a(\theta)(t-\tau)$. (5)

В этом случае получим более простой вид дисперсии: $Dt=\frac{(\tau')^2}{E\left(\frac{d\ln L}{d\theta}\right)^2}$. (6)

Возведем обе части (5) в квадрат и возьмем мат. ожидание от обеих частей:

$E\left(\frac{d\ln L}{d\theta}\right)^2=a^2(\theta)\cdot E(t-\tau)^2=a^2(\theta)Dt$. Тогда из формулы (6), имеем:

$$Dt=\frac{(\tau')^2}{a^2(\theta)Dt} \text{ или } (Dt)^2=\frac{(\tau')^2}{a^2(\theta)}, \text{ т.е. } Dt=\left|\frac{\tau'}{a(\theta)}\right|. \quad (7)$$

Теорема 2.

Если существует НОМД для θ, то НОМД существует для любой линейной функции от θ и не существует ни для какой другой функции от θ.

Доказательство. Пусть $t(x)$– НОМД для θ, а $\tau = A\theta + B$. Используем критерий НОМД:

$$\frac{d\ln L}{d\theta}=a(\theta)(t-\theta)=\frac{a(\theta)}{A}(At-A\theta)=\frac{a(\theta)}{A}(At+B-A\theta-B)=$$
$$a_1(\theta)\cdot(At+B-(A\theta+B))$$

По критерию получим, что $At + B$ – НОМД для $\tau = A\theta + B$.

$$D(At+B)=A^2Dt=A^2\left|\frac{\tau'}{a(\theta)}\right|$$

Замечание 1.

Ни для какой функции от $\tau = \tau(\theta)$, для которой есть НОМД, кроме линейной, не существует НОМД.

Доказательство.

Пусть $t=t(x)$ –НОМД для $\tau=\tau(\theta)$, $\tau_1=\tau_1(\theta)=\varphi(\tau)$ – нелинейная функция от τ. Предположим противное, т.е., t_1-НОМД для τ_1. Тогда по КЭ (5) имеем:

$$\frac{d\ln L}{d\theta}=a(\theta)(t-\tau); \quad (а)$$

$$\frac{d\ln L}{d\theta}=a_1(\theta)(t_1-\tau_1)=a_1(\theta)(t_1-\varphi(\tau)), \quad (б)$$

где $a(\theta) = a_1(\theta)\psi(\theta)$, тогда из (б) получаем:

$$\frac{d\ln L}{d\theta} = \left(\frac{a(\theta)}{\psi(\theta)}\right)(t_1 - \varphi(\tau)) = a(\theta)\left(\frac{t_1}{\psi(\theta)} - \frac{\varphi(\tau)}{\psi(\theta)}\right). \qquad \text{(b)}$$

Из (а) и (в) => $t - \tau = \dfrac{t_1}{\psi(\theta)} - \dfrac{\varphi(\tau)}{\psi(\theta)}$, поэтому, т.к. t и t_1 зависят только от $\vec{x}$, а τ и φ(τ) только от θ, то ψ(θ)=c_1– const и $t = \dfrac{t_1}{C_1} + C_2$, где c_2– const. Тогда

$$\frac{d\ln L}{d\theta} = a(\theta)\left(\frac{t - C_2}{C_1} - \frac{\varphi(\tau)}{C_1}\right) = a(\theta)(t - \tau) \Rightarrow C_1 t - C_1\tau = t - C_2 - \varphi(\tau) \Rightarrow \varphi(\tau) -$$

линейная функция от τ, а это противоречит предположению противного и значит, что утверждение замечания 1 доказано.

Практически это утверждение применяется в следующем случае: если получен вид (5) $\dfrac{d\ln L}{d\theta} = a(\theta)(t - \tau)$ и требуется ответить на вопрос о существовании НОМД для какой-либо функции от θ, отличной от линейной; ответ, очевидно, отрицательный.

Замечание 2.

Если построена оценка t(x) – НОМД для τ(θ), то для $\tau_1(\theta) = A\tau(\theta) + B$ НОМД $t_1(x) = At(x) + B$.

В более общем случае при выводе критерия можно говорить о τ(θ).

Замечание 3.

Если для θ НОМД существует, то ее дисперсию можно получить из неравенства Рао-Крамера, заменяя знак неравенства равенством, или по критерию (формула (6)).

Если НОМД не существует, то смысл неравенства Рао-Крамера состоит в том, что дает нижнюю грань дисперсии, которая не достигается.

Примеры.

Пример 1. Проверим, является ли $\hat{\theta} = \dfrac{\overline{x}}{m} = \dfrac{\sum_{i=1}^{n} x_i}{n \cdot m}$ НОМД для с.в. X~B(m,θ), где θ – вероятность успеха в 1-м опыте и τ(θ) = θ.

Известно, что EX = mθ, DX = mθ(1 – θ), тогда найдем $D\hat{\theta}$:

$$D\hat{\theta} = \frac{D\left(\sum x_i\right)}{m^2 \cdot n^2} = \frac{nm\,\theta(1-\theta)}{m^2 \cdot n^2} = \frac{\theta(1-\theta)}{mn}$$

Вычислим нижнюю грань дисперсии несмещенных оценок по неравенству Рао-Крамера. Покажем несмещенность $\hat{\theta}$: $E\hat{\theta} = \dfrac{1}{mn} m \cdot n\theta = \theta$.

$$L=\prod_{i=1}^{n}P(x=x_i)=\prod_{i=1}^{n}C_m^{x_i}\theta^{x_i}(1-\theta)^{m-x_i}=\left(\prod_{i=1}^{n}C_m^{x_i}\right)\cdot\theta^{\sum x_i}\cdot(1-\theta)^{nm-\sum x_i}.$$

$\ln L=\sum\ln C_m^{x_i}+\sum x_i\cdot\ln\theta+(mn-\sum x_i)\ln(1-\theta)$, тогда

$$\frac{d\ln L}{d\theta}=\frac{\sum x_i}{\theta}-\frac{mn-\sum x_i}{1-\theta}=\frac{\sum x_i-\theta\sum x_i-mn\theta+\theta\sum x_i}{\theta(1-\theta)}=\frac{\sum x_i-mn\theta}{\theta(1-\theta)}$$

получим: $E\left(\frac{d\ln L}{d\theta}\right)^2=\frac{1}{\theta^2(1-\theta)^2}\cdot E(\sum x_i-nm\theta)^2=$

$$E\left(\frac{d\ln L}{d\theta}\right)^2=\frac{1}{\theta^2(1-\theta)^2}\cdot\left[E(\sum x_i)^2-2mn\theta\cdot E\sum x_i+n^2m^2\theta^2\right].$$

Пользуясь тем, что $E(\sum x_i)^2=D(\sum x_i)+(E\sum x_i)^2$, $D\sum x_i=nm\theta(1-\theta)$, $(E\sum x_i)^2=n^2m^2\theta^2$, получим:

$$E\frac{d\ln L}{d\theta}=\frac{1}{\theta^2(1-\theta)^2}\cdot\left(mn\theta-mn\theta^2+m^2n^2\theta^2-2m^2n^2\theta^2+m^2n^2\theta^2\right)=\frac{mn}{\theta(1-\theta)}.$$

Теперь, применяя неравенство Рао-Крамера и учитывая, что $\tau'=1$, получим: $Dt\ge\frac{\theta(1-\theta)}{mn}$. Видим, что $\hat{\theta}$ – НОМД, т.к. результат вычисления $D\hat{\theta}$ дисперсии совпадает с нижней гранью из неравенства Рао-Крамера.

Пример 2.

Получим эту же оценку по критерию. $X\sim B(m,\theta)$; $\tau(\theta)=\theta$.

$\frac{d\ln L}{d\theta}=\frac{\sum x_i-mn\theta}{\theta(1-\theta)}$ (см. вычисления в Примере 1).

Приведем это выражение к виду (5): $\frac{d\ln L}{d\theta}=a(\theta)(t-\tau)$:

$\frac{d\ln L}{d\theta}=\frac{mn}{\theta(1-\theta)}\cdot\left[\frac{\sum x_i}{mn}-\theta\right]$, тогда $a(\theta)=\frac{mn}{\theta(1-\theta)}$ и $t(x)=\frac{\sum x_i}{mn}$ – НОМД.

Учитывая что $\tau'=1$, применим формулу (7) и получим, что

$$Dt=\left|\frac{\tau'}{a(\theta)}\right|=\frac{1}{a}=\frac{\theta(1-\theta)}{mn}.$$

Критерий НОМД для экспоненциального семейства (э.с.).

Пусть $\left.\begin{matrix}f(x)\\P(X)=x\end{matrix}\right\}=\exp\{A(\theta)+B(x)+C(\theta)+D(x)\}.$ (8)

Тогда $L=L_\theta(\vec{x})=\exp\left\{A(\theta(\sum_{i=1}^{n}B(x_i)+nC(\theta)+\sum_{i=1}^{n}D(x_i)\right\}$;

$$\frac{d\ln L}{d\theta}=\sum_{i=1}^{n}B(x_i)\frac{dA(\theta)}{d\theta}+n\frac{dC(\theta)}{d\theta}=\sum_{i=1}^{n}B(x_i)A'(\theta)+nC'(\theta)=$$

$$\frac{A'(\theta)}{n}\left(\frac{\sum B(x_i)}{n}-\frac{C'(\theta)}{A'(\theta)}\right),$$

следовательно по критерию $t^*(\vec{x})=\frac{1}{n}\sum B(x_i)$ – НОМД для

$$\tau^*=\tau(\theta)=\frac{C'(\theta)}{A'(\theta)}, \tag{9}$$

тогда по (7) $D\tau^*=\left|\frac{\tau^{*\prime}n}{A'(\theta)}\right|$,а значит НОМД существует и для любой линейной функции от τ.

Теорема 3.
Если НОМД для $\tau = \tau(\theta)$ существует, то распределение с.в. X относится к экспоненциальному семейству.
Доказательство. Пусть $t=t(\vec{x})$ – НОМД для $\tau = \tau(\theta)$. Тогда $\frac{d\ln L}{d\theta}=a(\theta)(t-\tau)$. Проинтегрировав по θ обе части последнего равенства, получим:
$\ln L=A(\theta)t(\vec{x})+C(\theta)+D(x)$, что и доказывает утверждение.

Примеры использования формулы (9).

Примеры функций $\tau = \tau(\theta)$, для которых существует НОМД при следующих распределениях:
а) с.в. $X \sim B(1,0)$;

$$P(X=x_i)=\theta^{x_i}(1-\theta)^{1-x_i}=\exp\{x_i\ln\theta+(1-x_i)\ln(1-\theta)\}=$$

$$=\exp\left\{\underbrace{x_i}_{B(x_i)}\underbrace{(\ln\theta-\ln(1-\theta))}_{A(\theta)}+\underbrace{\ln(1-\theta)}_{C(\theta)}\right\}\Rightarrow\tau^*=\frac{C'(\theta)}{A'(\theta)}=-\frac{\frac{1}{1-\theta}}{\frac{1-\theta}{\theta}\cdot\frac{1}{(1-\theta)^2}}=-\theta;$$

$\left(D\tau=\frac{1}{\theta(1-\theta)}\right)\Rightarrow$ для $\tau = \tau(\theta) = A\theta + B$, где A и B – любые const, существует НОМД;
б) с.в. $X \sim B(m,\theta)$;

$$P(X=x_i)=C_m^{x_i}\theta^{x_i}(1-\theta)^{m-x_i}=\exp\left\{\ln C_m^{x_i}+x_i\ln\theta+(m-x_i)\ln(1-\theta)\right\}=$$

$$=\exp\left\{\underbrace{x_i}_{B(x_i)}\underbrace{(\ln\theta-\ln(1-\theta))}_{A(\theta)}+\underbrace{m\ln(1-\theta)}_{C(\theta)}+\underbrace{C_m^{x_i}}_{D(x_i)}\right\}\Rightarrow\tau^*=\frac{C'(\theta)}{A'(\theta)}=-\frac{\frac{m}{1-\theta}}{\frac{1-\theta}{\theta}\cdot\frac{1}{(1-\theta)^2}}=-m\theta;$$

$\left(D\tau^*=\frac{m}{\theta(1-\theta)}\right)\Rightarrow$ для $\tau=\tau(\theta)=A\theta+B$, где A и B – любые const, существует НОМД;

в) с.в. $X\sim\pi(\theta)$;

$$P(X=x_i)=\frac{\theta^{x_i}e^{-\theta}}{x_i!}=\exp\left\{\underbrace{x_i}_{B(x_i)}\underbrace{\ln\theta}_{A(\theta)}-\underbrace{\theta}_{C(\theta)}-\underbrace{\ln(x_i!)}_{D(x_i)}\right\}\Rightarrow\tau^*=\frac{C'(\theta)}{A'(\theta)}=\theta;$$

$\left(D\tau\tau=\frac{1}{1/\theta}=\theta\right)\Rightarrow$ для $\tau=A\theta+B$, где A и B – любые const, существует НОМД;

г) с.в. $X\sim\Gamma(\theta)$;

$$P(X=x_i)=\theta(1-\theta)^{x_i-1}=\exp\{\ln\theta+(x_i-1)\ln(1-\theta)\}=$$

$$=\exp\left\{\underbrace{x_i}_{B(x_i)}\underbrace{\ln(1-\theta)}_{A(\theta)}+\underbrace{\ln\theta-\ln(1-\theta)}_{C(\theta)}\right\}\Rightarrow\tau^*=\frac{C'(\theta)}{A'(\theta)}=\frac{1-\theta}{\theta}\cdot\frac{(1-\theta)}{(1-\theta)^2}=\frac{1}{\theta};$$

$\left(D\tau^*=\frac{1}{\theta^2}\cdot\frac{1}{1-\theta}=\frac{1}{\theta^2(1-\theta)}\right)\Rightarrow$ для $\tau=\frac{A}{\theta}+B$, где A и B – любые const, существует НОМД.

§5. Методы получения точечных оценок

Постановка задачи. Пусть имеется некоторая выборка $\vec{x}: x_1, x_2, \ldots, x_n$, $X \sim F_\theta(x)$ $\theta \in \Theta$, где Θ – параметрическое пространство. Требуется построить оценку для θ или $\tau(\theta)$.

Различаются два подхода к решению этой задачи в зависимости от понимания природы неизвестного параметра.

1 – ый подход реализуется в случае, когда θ является неизвестной постоянной, т.е. $\theta = \text{const}$. В данной ситуации используется **метод подстановки**.

Суть метода. Выбирается некоторая мера расхождения теоретического и эмпирического распределения и строится некоторый функционал от этой меры. Оценка для неизвестного параметра ищется таким образом, чтобы этот функционал принимал некоторое значение, соответствующее минимуму расхождения теоретических и практических результатов.

Метод подстановки объединяет ряд конкретных методов, которые различаются по мере различия теории и практики:

1. Метод моментов (ММ).
2. Метод максимального правдоподобия (ММП).
3. Метод минимального χ^2.
4. Метод минимального расстояния.

Существуют и другие методы, но мы рассмотрим только первые, т.к. они используются наиболее часто.

2 − ой подход реализуется в случае, когда θ является случайной величиной, т.е $\theta \sim H_\theta(t)$. В данной ситуации используются **Байесовские оценки.**

Методы подстановки

1. Метод моментов (ММ)

Совокупность неизвестных параметров будем рассматривать как k−мерный вектор $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Тогда оценки метода моментов (ОММ) являются решениями системы k уравнений, составленных путем приравнивания k теоретических моментов соответственно k эмпирическим.

Достоинством ММ является простота его применения.

Недостатком ММ является то, что он не гарантирует качества оценок, хотя часто оценки, полученные этим методом, являются «удачными».

Пример 1.

С.в. $X \sim \pi(\theta)$. Найти ОММ – $\tilde{\theta}$ для неизвестного параметра θ.

Решение.

$\tilde{\theta} = EX = \bar{x}$.

Пример 2.

С.в. $X \sim N\left(a, \sigma^2\right)$, тогда $\theta = (\theta_1, \theta_2)$, где $\theta_1 = a$, $\theta_2 = \sigma^2$. Найти ОММ для θ.

Решение.

$$\begin{cases} \tilde{a} = \bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i \\ \tilde{\sigma}^2 = S_2^1 = \dfrac{1}{n} \sum\limits_{i=1}^{n} (x_i - \bar{x})^2, \text{ отсюда } \tilde{\theta} = \left(\bar{x}, S_2^1\right). \end{cases}$$

Пример 3.

С.в. $X \sim B\,(m,\ p = \theta)$. Найти ОММ для θ (m–известно).

Решение.

$$EX = m\theta = \bar{x}, \text{ следовательно, } \tilde{\theta} = \frac{\bar{x}}{m} = \frac{\sum\limits_{i=1}^{n} x_i}{mn}.$$

Пример 4.

С.в. $X \sim R[0;\theta]$. Найти ОММ для θ.

Решение.

$$EX = \frac{\tilde{\theta}}{2} = \bar{x}, \text{ тогда ОММ } \tilde{\theta} = 2\bar{x} = \frac{2\sum\limits_{i=1}^{n} x_i}{n}.$$

2. Метод максимального правдоподобия (ММП)

Оценка максимального правдоподобия (ОМП) $\hat{\theta}$ выбирается таким образом, чтобы функция правдоподобия $L = L_\theta(x)$ принимала наибольшее значение.

Различается два случая нахождения ОМП: регулярный и нерегулярный.

Регулярный случай.

Имеем следующие условия регулярности:

1) дифференцируемость L по θ;

2) множество $\{x : L \neq 0\}$ не зависит от неизвестного параметра θ.

Условие регулярности обеспечивает достижение максимума функции L(x), поэтому в регулярном случае ОМП ищется из следующей системы уравнений (УМП):

$$\begin{cases} \dfrac{\partial L}{\partial \theta_j} = 0 \\ j = \overline{1,k} \end{cases} \quad \text{или из системы уравнений} \quad \begin{cases} \dfrac{\partial \ln L}{\partial \theta_j} = 0 \\ j = \overline{1,k} \end{cases},$$

т.к. функции L и ln L достигают максимума в тех же точках.

Пример 5.

С.в. $X \sim \pi(\theta)$. Найти ОМП для θ.

Решение.

$$P(X = x_i) = \frac{\theta^{x_i} \cdot e^{-\theta}}{x_i!}, \text{ тогда } \quad L = \frac{\theta^{\sum\limits_{i=1}^{n} x_i} \cdot e^{-n\theta}}{\prod\limits_i x_i!}.$$

$$\ln L = \sum_{i=1}^{n} x_i \cdot \ln\theta - n\theta - \sum_{i=1}^{n} \ln x_i!\ ; \quad \frac{d\ln L}{d\theta} = \frac{\sum\limits_{i=1}^{n} x_i}{\theta} - n.$$

Решим УМП и получим, что $\hat{\theta} = \sum\limits_{i=1}^{n} x_i = \overline{x}$.

В данном случае получили, что ОМП совпадает с ОММ.

Очевидно, что вычисления ММП значительно сложнее, чем ММ, но ММП теоретически обосновывает гарантии качества оценок.

Теорема 1.

Если существует НОМД для $\tau(\theta)$, то в регулярном случае она является ОМП для $\tau(\theta)$.

Доказательство.

Используем критерий НОМД: $\dfrac{\partial \ln L}{\partial \theta} = a(\theta)\,(t(x) - \tau(\theta))$, где t(x) – НОМД для $\tau(\theta)$. Тогда, если $\dfrac{\partial \ln L}{\partial \theta} = 0$, то $t(x)$ – ОМП для $\tau(\theta)$.

Пример 6 для функции от неизвестного параметра θ, для которой не существует НОМД, а ОМП существует.

Пусть с.в. $X \sim Ge\,(\theta),\ \tau(\theta) = \theta$. Найдем ОМП для $\tau(\theta) = \theta$.

$P\left(X=x_i\right)=\theta\left(1-\theta\right)^{x_i-1}$, тогда $L=\theta^n\left(1-\theta\right)^{\sum\limits_{i=1}^{n} x_i-n}$;

$$\ln L-n\ln\theta+\left(\sum_{i=1}^{n}x_i-n\right)\cdot\ln\left(1-\theta\right);$$

$$\frac{d\ln L}{d\theta}=\frac{n}{\theta}-\frac{\sum\limits_{i=1}^{n}x_i-n}{1-\theta}=\frac{n-n\theta-\theta\sum\limits_{i=1}^{n}x_i+n\theta}{\theta\left(1-\theta\right)}=\frac{n-\theta\sum\limits_{i=1}^{n}x_i}{\theta\left(1-\theta\right)}=\frac{n}{1-\theta}\left(\frac{1}{\theta}-\frac{\sum\limits_{i=1}^{n}x_i}{n}\right)=$$

$$=\frac{n}{1-\theta}\left(\frac{1}{\theta}-\overline{x}\right).$$

Найдем ОМП из УМП

$\dfrac{n-\theta\sum\limits_{i=1}^{n}x_i}{\theta(1-\theta)}=0$, откуда следует, что $n=\theta\sum\limits_{i=1}^{n}x_i$. Тогда $\hat{\theta}=\dfrac{1}{\overline{x}}$ – ОМП.

Воспользуемся критерием НОМД:

$$\frac{d\ln L}{d\theta}=\frac{n-\theta\sum\limits_{i=1}^{n}x_i}{\theta\left(1-\theta\right)}=\frac{n\theta}{\theta\left(1-\theta\right)}=\left(\frac{1}{\theta}-\frac{\sum\limits_{i=1}^{n}x_i}{n}\right)=-\frac{n\theta}{\theta\left(1-\theta\right)}\left(\overline{x}-\frac{1}{\theta}\right),$$ следовательно,

по критерию НОМД имеем $\bar{x}$ – НОМД для $\dfrac{1}{\theta}$, тогда для $\tau(\theta)=\theta$ НОМД не существует.

По данной теореме, если НОМД для $\tau(\theta)$ существует, то ОМП совпадает с ней. Это и является гарантией качества ОМП.

Теорема 2.

Все решения УМП (в регулярном случае), т.е. ОМП являются функциями достаточной статистики.

Доказательство.

Пусть $\hat{\theta}$ – достаточная статистика, $g_\theta(t)$ – плотность её распределения $\hat{\theta}$, а $f(x_k)$ – плотность распределения исследуемой с.в. в точке x_k.

$L=\prod_{k=1}^{n} f(x_k)=\prod_{k=1}^{n} f\left(x_k/\hat{\theta}=t\right)\cdot g_\theta(t)$, тогда $\ln L=\sum_{k=1}^{n}\ln f\left(x_k/\ \hat{\theta}=t\right)+n\cdot\ln g_\theta(t)$,

следовательно, $\frac{d\ln L}{d\theta}=\frac{\partial\sum_{k=1}^{n}\ln f\left(x_k/\hat{\theta}=t\right)}{\partial\theta}+n\cdot\frac{\partial\ln g_\theta(t)}{\partial\theta}=0$. Полученное УМП принимает вид $n\cdot\frac{d\ln g_\theta(t)}{d\theta}=0$, т.к. $\sum_{k=1}^{n}\ln f\left(x_k/\hat{\theta}=t\right)$ не зависит от θ в силу того, что $\hat{\theta}$ – достаточная статистика; следовательно,
$\frac{d\sum_{k=1}^{n}\ln f\left(x_k/\hat{\theta}=t\right)}{d\theta}=0$.

Решение уравнения $n\cdot\frac{d\ln g_\theta(t)}{d\theta}=0$ будет находиться в терминах достаточной статистики, что и означает выполнение утверждения теоремы.

Теорема 3. Инвариантность ОМП.
Если параметры θ и ν связаны непрерывной взаимно-однозначной зависимостью $\nu=\varphi(\theta)$ и $\hat{\theta}$ – ОМП для θ, то $\hat{\nu}=\varphi(\hat{\theta})$ – ОМП для ν.

Доказательство.
Пусть $f(x)$ – плотность распределения изучаемой с.в. X. Тогда $f_\theta(x)\le f_{\hat{\theta}}(x)$ в некоторой окрестности точки $\hat{\theta}$; следовательно, т.к . φ непрерывна, $f_{\varphi^{-1}(\nu)}(x)\le f_{\varphi^{-1}(\hat{\nu})}(x)$. Это неравенство означает утверждение теоремы в силу характера зависимости φ, т.е. $f_\nu(x)\le f_{\hat{\nu}}(x)$ в некоторой окрестности и $\hat{\nu}=\varphi\left(\hat{\theta}\right)$ является ОМП для ν.

Замечание 1. Если задача состоит в построении ОМП для некоторой $\tau(\theta)$ и $\tau(\theta)$ удовлетворяет условиям теоремы 3, то она сводится к более простой – нахождению $\hat{\theta}$ – ОМП для θ, тогда $\hat{\tau}(\theta)=\tau(\hat{\theta})$ есть ОМП для $\tau(\theta)$.

Замечание 2. (без доказательства) При достаточно общих условиях ОМП состоятельны, асимптотически нормальны и асимптотически эффективны при $n\to\infty$.

Особый интерес представляет нахождение ОМП в **нерегулярном случае:** ОМП находится из смысла метода (ММП), т.е. как значение, при котором функция правдоподобия принимает наибольшие значения.

Пример 7.

С.в. $X \sim R[0,\theta]$ Найти ОМП для θ, т.е. $\hat{\theta}$.

Решение.

Это нерегулярный случай.

$$L = \begin{cases} 0, & x \notin [0,\theta] \\ \dfrac{1}{\theta^n}, & x \in [0,\theta]. \end{cases}$$

Наибольшее значение L, равное $\dfrac{1}{\theta^n}$, получается при наименьшем возможном значении θ, если наблюдались значения $x: x_1,\ldots,x_n$, т.е. при

$$\theta = x_{(n)}.$$

Пример 8 **того, что ОМП бесконечно много.**

С. в. $X \sim R[0,1+\theta]$

$$L = \begin{cases} 1, & \theta \le x_{(1)} \le x_{(n)} \le \theta + 1 \\ 0, & \text{в противном случае.} \end{cases}$$

Очевидно, что max L = 1, но

$$\begin{cases} \theta \le x_{(1)} \\ \theta \le x_{(n)} - 1 \end{cases} \Rightarrow \hat{\theta} \in [x_{(n)} - 1; x_{(1)}].$$

Точек на отрезке для $\hat{\theta}$ бесконечно много, следовательно, значений $\hat{\theta}$ бесконечно много. Остается только один вопрос: существует ли такой отрезок?

Из выражения для L имеем: $\theta \le x_{(1)} \le x_{(n)} \le \theta + 1$, значит, $x_{(1)} \ge x_{(n)} - 1$, т.е. отрезок существует.

Далее предлагаются к рассмотрению примеры на построение ОМП и ОММ как с объяснениями, так и для самостоятельного решения.

Примеры нахождения ОММ и ОМП.

Пример 9.

С.в. $X \sim \gamma(x;\, \alpha, \lambda)$, α, λ – const > 0 , плотность распределения с.в. X есть

$$f(x)=\begin{cases} 0, -x<0 \\ \dfrac{\lambda^{\alpha}e^{-\lambda x}x^{\alpha-1}}{\Gamma(\alpha)}, x\ge 0 \end{cases}, \qquad \text{где } \Gamma(\alpha)=\int_0^{\infty} x^{\alpha-1}e^{-x}dx.$$

Найти ОММ для α и λ.

Решение.

$$\begin{cases} \dfrac{\alpha}{\lambda}=\overline{x} & (1)\\ \dfrac{\alpha(\alpha+1)}{\lambda^2}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^2 . & (2)\end{cases}$$

Поделим (2) на (1), возведенное в квадрат. Получаем

$$\frac{\alpha(\alpha+1)\lambda^2}{\lambda^2\alpha^2}=\frac{\sum\limits_{i=1}^{n}x_i^{\ 2}}{n\overline{x}^2}\Rightarrow \alpha\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)=nx\Rightarrow \hat{\alpha}=\frac{n\overline{x}^2}{\sum\limits_{i=1}^{n}x_i-n\overline{x}^2};$$

$$\hat{\lambda}=\frac{\hat{\alpha}}{\hat{x}}=\frac{n\overline{x}}{\sum\limits_{i=1}^{n}x_i-n\overline{x}} .$$

Пример 10.

С.в. $X \sim R[m-\alpha, m+\alpha]$, α – известно, m – неизвестно. Найти ОМП для параметра m.

Решение.

Вычислим функцию правдоподобия:

$$L=\begin{cases} 0, x\in[m-\alpha, m+\alpha] \\ \dfrac{1}{(2\alpha)^2}, x\in[m-\alpha, m+\alpha]. \end{cases}\Rightarrow$$

максимального значения $L=\dfrac{1}{(2\alpha)^n}$ достигает при всех

$x_i=[m-\alpha, m+\alpha], i=\overline{1,n}$. Отсюда следует, что множество ОМП для m есть пересечение всех отрезков $[x_i-\alpha, x_i+\alpha], i=\overline{1,n}$, т.е. это любое значение отрезка $[x_{(n)}-\alpha, x_{(1)}+\alpha]=\left[\max\limits_{i=\overline{1,n}} x_i-\alpha, \min\limits_{i=\overline{1,n}} x_i+\alpha\right]$, при этом $\overline{x}=\frac{1}{n}\sum\limits_{i=1}^{n} x_i$ может не принадлежать этому отрезку.

Пример 11.

С.в. $X \sim R[m-\alpha, m+\alpha]$, где параметры m и α оба неизвестны. Найти ОМП для параметров m и α.

Решение.

По решению примера 10 $\max L=\frac{1}{(2\alpha)^n}$ **достигается при возможном минимальном значении α, совместном с выборкой, когда** $m \in [x_{(n)}-\alpha, x_{(1)}+\alpha]$**, т.е. когда** $x_{(n)}-\alpha \le x_{(1)}+\alpha \Rightarrow \alpha \ge \frac{x_{(n)}-x_{(1)}}{2} \Rightarrow \hat{\alpha}=\frac{x_{(n)}-x_{(1)}}{2}$ **– ОМП для α. Тогда получаем:**

$$\hat{m}=\left[x_{(n)}-\alpha, x_{(1)}+\alpha\right]=\left[x_{(n)}-\frac{x_{(n)}-x_{(1)}}{2}, x_{(1)}+\frac{x_{(n)}-x_{(1)}}{2}\right]=$$

$$=\left[\frac{x_{(n)}-x_{(1)}}{2}, \frac{x_{(n)}-x_{(1)}}{2}\right] \Rightarrow$$

ОМП для параметра m: $\hat{m}=\frac{x_{(n)}+x_{(n)}}{2}$**.**

Пример 12.

С.в. $X \sim R[a\theta_1+b, c\theta_2+d]$, где a, b, c, d – известные константы. $\theta=(\theta_1,\theta_2)$ – неизвестный параметр распределения. Найти ОМП и ОММ для $\theta=(\theta_1,\theta_2)$.

Решение.

$$L=\begin{cases}\frac{1}{c\theta_2-a\theta_1+d-b}, & a\theta_1+b \le x_{(1)} \le x_{(n)} \le c\theta_2+d \\ 0, & \text{в противном случае}\end{cases}$$

ф.п. L принимает наибольшее значение при минимальном значении θ $_2$ и максимальном значении θ_1, а т.к. $\theta_1 \le \frac{x_{(1)}-b}{a}$, $\theta_2 \ge \frac{x_{(n)}-d}{c}$, то

$$\text{ОМП} = \hat{\theta} = \begin{cases} \hat{\theta}_1 = \dfrac{x_{(1)}-b}{a} \\ \hat{\theta}_2 = \dfrac{x_{(2)}-d}{c} \end{cases}$$

$$\text{ОММ:} \quad \begin{cases} \overline{x} = \dfrac{\sum\limits_{i=1}^{n} x_i}{n} = \dfrac{a\theta_1 + b + c\theta_2 + d}{2} \\ \dfrac{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} = \dfrac{\left(c\theta_2 + d - z\theta_1 - b\right)^2}{12}. \end{cases}$$

Байесовские оценки (решения)

Под решением здесь понимается выбор оценки неизвестного параметра распределения $\theta = (\theta_1, \dots, \theta_K)$.

Решающей функцией $d = \delta(x)$ называется правило (функция), ставящее в соответствие каждому результату наблюдения некоторое решение d. Областью определения величины $\delta(x)$ является множество значений наблюдаемой случайной величины (с.в.) X , областью значений – множество решений D .

Чтобы функцию $\delta(x)$ выбрать наилучшим образом, нужно сравнить последствия использования различных функций $\delta(x)$. Для этого задается функция потерь $W(\theta,d)$, значение которой определяется выбранным решением, если с.в. X имеет функцию распределения $F_\theta(x)$ с неизвестным параметром θ. Тогда при многократном применении $\delta(x)$ (при повторении опыта) определяются средние потери $EW(\theta,d)=R(\theta,d)$, называемые функцией риска.

Цель состоит в выборе решающей функции $\delta(x)$, минимизирующей функцию риска $L(\theta,d)$.

Байесовский подход отличается от небайесовского тем, что неизвестный параметр θ считается не фиксированной постоянной, а с.в. с известным распределением $\pi(\theta)$ – априорным распределением.

После завершения наблюдений над с.в. X из априорного распределения неизвестного параметра можно получить его апостериорные распределение $\pi(\theta/X)$, и выбор байесовского решения естественно связывать с ожидаемыми потерями при этом апостериорном распределении $\pi(\theta/X)$.

Доказано, что при байесовском подходе минимальные средние потери (риск) по всем возможным значениям параметра θ и всем реализациям с.в. X достигается при той же решающей функции $\delta(x)$, при которой получаются минимальные средние потери при апостериорном распределении параметра θ.

Апостериорный риск вычисляется по формуле

$$R(d,x)=\begin{cases}\sum\limits_i W(\theta_i,d)\,\pi(\theta_i/x_j) & \text{в дискретном случае}\\ \int\limits_{\Theta} W(\theta,d)\,\pi(\theta/x_j)d\theta & \text{в непрерывном случае}\end{cases} \qquad (1)$$

$$\pi(\theta_i/x_j)=\frac{\pi(\theta_i)P(x_i/\theta_j)}{P(x_j)};\quad \pi(\theta/x_j)=\frac{\pi(\theta)P(x_j/\theta)}{P(x_j)}. \qquad (2)$$

Здесь $P(x_j/\theta_i)=P(X=x_j/\theta_i)$; $P(x_j/\theta)=f(x_j/\theta)$ – плотность распределения с.в. X в непрерывном случае при $X = x_j$.

$$P(x_j)=\begin{cases}\sum\limits_i \pi(\theta_i)P(x_j/\theta_i) & \text{в дискретном случае}\\ \int\limits_{\Theta} \pi(\theta)P(x_j/\theta)\,d\theta & \text{в непрерывном случае}\end{cases} \qquad (3)$$

Формула (2) – формула Байеса, а (3) – формула полной вероятности.

Формулы (1) – (3) даны для нахождения байесовского решения минимизацией апостериорного риска по одному наблюдению. Если же ставится задача сделать это по выборке, то в формулах (2) и (3) следует заменить вероятности $P(x_j/\theta_i)$ и $P(x_j/\theta)$ на соответствующие функции правдоподобия $L_\theta(x_1,\ldots,x_k)$.

Разберем подробно две задачи оценивания неизвестного параметра θ

путем нахождения байесовского решения в дискретном и непрерывном случаях с использованием формул (1) , (2) , (3) .

Задача 1.

Пусть с.в. X имеет Бернуллиевское распределение B (1,θ) , априорное распределение $\pi(\theta)$ **задано рядом распределения:**

θ	1/4	1/2
3	1/3	1/3

,

а функция потерь $W(\theta_i, d_j)$ задается таблицей:

$W(\theta_i, d_j)$	**d_1**	**d_2**
θ_1=1/4	**1**	**4**
θ_2=1/2	**3**	**2**

Здесь D = (d_1,d_2) – множество решений состоит из двух точек: d_1: (θ =1/4) и d_2 = (θ =1/2)

Найти байесовскую оценку для неизвестного параметра θ.

Решение.

По (1) $R(d, X_j) = \sum_{i=1}^{2} W(\theta_i, d)\ \pi(\theta_i/X_j)$**, где j = 1,2 (т.к. с.в. X принимает два значения: X = 0, X = 1).**

По (2) $\pi(\theta_i/x_j) = \dfrac{\pi(\theta)P(x_j/\theta_i)}{P(x_j)}$ **.**

По (3) $P(x_j) = \sum_{i=1}^{2} \pi(\theta_i)P(x_j/\theta_i)$, а $P\left(x_j/\theta_i\right) = \theta_i^{x_j}\left(1-\theta_i\right)^{1-x_j}$

по условию задачи, откуда получаем

$P(X_1 = 0) = \pi(\theta_1 = 1/4)\, P(x_1 = 0/\theta_1 = 1/4) +$

$+ \pi(\theta_2 = 1/2)\, P(x_1 = 0/\theta_2 = 1/2) = (1/3)(3/4) + (2/3)(1/2) = 7/12.$

Аналогично

$P(X_2 = 1) = 3/12;$

$$\pi(\theta_1/X_1) = 3/7; \qquad \pi(\theta_2/X_1) = 4/7;$$

$$\pi(\theta_1/X_2) = 1/5; \qquad \pi(\theta_2/X_2) = 4/5.$$

Таким образом, найдено апостериорное распределение параметра θ.

Вычислим и сравним теперь апостериорные риски при каждом наблюдаемом значении с.в. X ($X_1 = 0$ и $X_2 = 1$) для всех решений, и в качестве байесовского выберем то из них, при котором получается меньший риск.

а) Пусть $X = x_1 = 0$, тогда по (1)

$$R(d_1, x_1) = W(\theta_1, d_1)\pi(\theta_1, x_1) + W(\theta_2, d_1)\pi(\theta_2, x_1) = 1 \cdot (3/7) + 2 \cdot (4/7) =$$
$$= 15/7.$$

Аналогично получаем, что $R(d_2, x_1) = 4 \cdot (3/7) + 2 \cdot (4/7) = 12/5$.

При $X = x_1 = 0$ меньший апостериорный риск при решении d_1, а при $X = x_2 = 1$ – при d_2, таким образом получаем байесовское правило:

$$\delta(x) = (\delta(0) = d_1, \delta((1) = d_2).$$

б) Пусть с.в. X имеет экспоненциальное распределение с неизвестным параметром θ, априорное распределение которого есть гамма-распределение с плотностью

$$g(\theta(k;r)=\frac{r^{r}\theta^{k-1}\exp\{-r\theta\}}{\Gamma(k)};\theta\ge 0;$$

$$\left(\Gamma(k)=\int_{0}^{\infty}x^{k-1}\exp\{-x\}dx;\ k\Gamma(k)=\Gamma(k+1)\right).$$

А функция потерь $W(d,\theta)=(\theta-t)^2$**, где решение** $d:t=\hat{\theta}$ **– оценка для θ.**

Найти байесовскую оценку для параметра θ по

а) одному наблюдению X,

б) по выборке $x_1,\dots,x_n$.

Решение.

а) Найдем апостериорное распределение

$$\pi(\theta/x)=\frac{\pi(\theta)P(x/\theta)}{\int\limits_{\Theta}\pi(\theta)P(x/\theta)d\theta}=\frac{r^{k}\theta^{k-1}\exp\{-rx\}\,\theta\exp\{-\theta x\}\Gamma(k)}{\Gamma(k)r^{k}\int_{0}^{\infty}\theta^{k-1}\exp\{-r\theta\theta\}Q\exp-\theta x\}d\theta}=$$

$$=\frac{\theta^{k}\exp\{-\theta(x+r)\}}{\int_{0}^{\infty}\theta^{k}\exp\{-\theta(r+x)\}dt}=\frac{\theta^{k}\exp\{-\theta(x+r)\}}{\left(1/(x+r)^{k+1}\right)\int_{0}^{\infty}(\theta(x+r))^{k}\exp\{-\theta(x+r)\}d\theta(x+r)}=$$

$$=\frac{\theta^{k}\exp\{-\theta(x+r)\}(x+r)^{k+1}}{\Gamma(k+1)}.$$

$\pi(\theta/x)=\dfrac{\theta^{k}\exp\{-\theta(x+r)\}(x+r)^{k+1}}{\Gamma(k+1)}$ есть гамма–распределение с плотностью $g(\theta;k+1,x+r)$.

Байесовскую оценку для θ находим, минимизируя апостериорный риск $R(d,x)$ по t.

$$R(d,x)=\int_{0}^{\infty}(\theta-t)^2g(\theta;k+1,x+r)d\theta.$$

Берем производную по t и приравниваем ее к нулю. Получаем

$$\int_0^{\infty}(Q-t)\,g\,(Q;k+1,x+r)\,dQ=0,\ \text{откуда}\ t=\int_0^{\infty}\theta g(\theta;k+1,x+r)d\theta=$$

$$=\left((x+r)^{k+1}/\Gamma(k+1)\right)\int_0^{\infty}\theta^{k+1}(x+r)^{k+1}\exp\{-\theta\,(x+r)\}\left(1/(x+r)^{k+2}\right)d\theta(x+r)=$$

$$=\frac{(k+1)\Gamma(k+1)}{\Gamma(k+1)(x+r)}=\frac{(k+1)}{(x+r)}\ ,$$

то есть $\hat{\theta}=t=(k+1)/(x+r)$ **байесовская оценка для неизвестного параметра θ по наблюденному значению x.**

Апостериорный байесовский риск тогда вычисляется по формуле

$$R\left(\theta,x_j\right)=\int_0^{\infty}(\theta-((k+1)/(x+r))^2 g(\theta;k+1,x+r)d\theta=$$

$$=\int_0^{\infty}\theta^2 g(\theta;k+1,x+r)d\theta-x(k+1)/(x+r)\int_0^{\infty}\theta g(\theta;k+1,x+r)d\theta+$$

$+((k+1)/(x+r))^2$, где интегралы $\int_0^{\infty}\theta^2 g(\theta;k+1,x+r)d\theta$ и

$\int_0^{\infty}\theta t(\theta;k+1,x+r)d\theta$ аналогично преобразуются, как было показано выше, выражаются через гамма−функцию, и могут быть досчитаны самостоятельно.

б) Дана выборка $x_1,...,x_n$. Найдем сначала апостериорное распределение $\pi(\theta/x_1,...,x_n)$.

Аналогично случаю а) получаем байесовскую оценку в виде

$$t = \int_0^{\infty} \theta\pi\left(\theta / x_1, \ldots, x_n\right) d\theta =$$

$$= \left(\left(\sum_{i=1}^{n} x_i + r \right)^{n+k} \Big/ \Gamma(n+k) \right) \int_0^{\infty} \theta^{n+k} \exp\left\{ -\theta \left(\sum_{i=1}^{n} x_i + r \right) \right\} d\theta =$$

$$= \left(\sum_{i=1}^{n} x_i + r \right)^{n+k} \Gamma(n+k+1) / \Gamma(n+k) \left(\sum_{i=1}^{n} x_i + r \right)^{n+k+1} = \left(\sum_{i=1}^{n} + r \right),$$ то есть

$\hat{\theta} = t = (n+k) \Big/ \sum_{i=1}^{n} x_i + r$ — байесовская оценка для параметра в случае б).

Апостериорный риск тогда в этом случае б) есть

$$R(d, x_1, \ldots, x_n) = \int_0^{\infty} \left(\theta - (n-k) \Big/ \left(\sum_{i=1}^{n} x_i + r \right) \right)^2 g\left(\theta; n+k, \sum_{i=1}^{n} x_i + r \right) d\theta$$ и вычисляется аналогично случаю а) и может быть получен самостоятельно.

§6. Доверительное оценивание

Постановка задачи.

$\bar{x} = x_1, \ldots, x_n$ − выборка объёма n наблюдений над случайной величиной X, распределение которой относится к параметрическому семейству $F\theta(x)$, где $\theta = (\theta_1, \ldots, \theta_к)$ и $\theta \in \Theta$ (Θ − параметрическое множество). Требуется оценить некоторую функцию $\tau = \tau(\theta)$. Доверительное оценивание τ означает нахождение k −мерной области, заключающей неизвестное значение функции τ с заданной доверительной вероятностью γ. Подробнее остановимся на рассмотрении случая $k = 1$ и $\tau(\theta) = \theta$. Тогда искомое доверительное множество становится доверительным интервалом, и задача состоит в построении двух статистик $t_1 = t_1(\vec{x})$ и $t_2 = t_2(\vec{x})$ (концов доверительного интервала $J = (t_1, t_2)$, заключающего в себе неизвестное значение параметра θ с заданной доверительной вероятностью
γ: $\gamma = P(t_1 < \theta < t_2))$.

При доверительном оценивании заданное значение γ (обычно близкое к единице) означает надёжность оценивания $\tau(\theta)$ с точностью, определяемой размером доверительной области. При построении доверительного интервала для параметра θ его длина – точность оценивания, а γ – заданная надежность. Поэтому желательно строить кратчайший доверительный интервал, соответствующий наибольшей точности при данном γ .

Общий приём при нахождении доверительного интервала состоит в построении центральной статистики (ц.с.) $Z = Z(\theta)$, т.е. такой статистики, распределение которой не зависит от неизвестного параметра θ. Если $Z(\theta)$ непрерывна и монотонна по θ , то это обеспечивает однозначную эквивалентность событий $\{t_1^* < Z < t_2^*\}$ и $\{t_1 < \theta < t_2\}$. Тогда, если удалось найти $t_1^* = t_1^*(\theta)$ и $t_2^* = t_2^*(\theta)$ − нижнюю и верхнюю доверительные границы, то решая неравенство $t_1^* < Z < t_2^*$ относительно θ, находим значения t_1 и t_2 − искомые границы доверительного интервала для неизвестного параметра θ.

Остаётся обсудить две проблемы: построение центральной статистики $Z = Z(\theta)$ и нахождение значений t_1^* и t_2^* из уравнения:

$$\gamma = P(t_1^* < Z < t_2^*) . \tag{1}$$

Начнём с **первой проблемы**. Идеями построения ц.с. могут быть следующие:

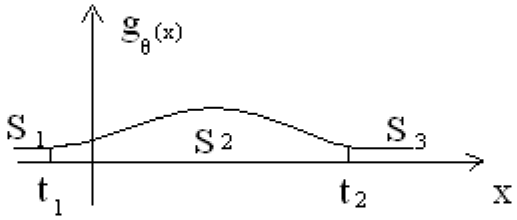
1) замена исходной с.в. на новую, зависящую от неизвестного параметра θ , распределение которой не зависит от θ;
2) стандартизация имеющейся точечной оценки;
3) использование результатов ЦПТ или асимптотической нормальности ОМП (для построения асимптотических доверительных интервалов).

Вторая проблема состоит в нахождении значений t_1^* и t_2^* из уравнения (1). Требуется сформулировать дополнительное ограничение на t_1^*

и t_2* , чтобы это было возможно, т.е. чтобы уменьшить число неизвестных в уравнении (1) с двух до одной.

При решении этой проблемы различают, обычно, два случая: **регулярный** и **нерегулярный** (раньше определён регулярный случай требованиями дифференцируемости функции правдоподобия L и независимости области, в которой $L \neq 0$, от неизвестного параметра θ).

Регулярный случай. Строят центральный доверительный интервал (ц.д.и.). Определим ц.д.и. Пусть $g_\theta(x)$ − кривая распределения неизвестного параметра θ :



Тогда $J = (t_1,t_2)$ − ц.д.и., если площади S_1 и S_3 одинаковы.

Определим x_p – p − квантиль распределения $F(x)$, если x_p − корень уравнения $F(x) = p$.

Теперь выразим значения t_1 и t_2 в терминах квантили распределения параметра θ с функцией распределения $G(\theta)$: $S_1+S_2+S_3=1$.
Пусть $S_2 = \gamma$, тогда
$S_1 = S_3 = \frac{1-\gamma}{2}$, откуда следует, что $t_1 - \frac{1-\gamma}{2}$, − квантиль, а $t_2 = \left(\frac{1-\gamma}{2} + \gamma\right) = \frac{1+\gamma}{2}$ − квантиль распределения. Значит ц.д.и. – интервал между $\frac{1-\gamma}{2}$ и $\frac{1+\gamma}{2}$ – квантилями распределения $G(\theta)$. Таким образом, требование построения ц.д.и. и есть необходимое дополнительное требование в уравнении (1).

Нерегулярный случай. В качестве искомого доверительного интервала в уравнении (1) выбирают крайнюю зону значений неизвестного параметра. Тогда число неизвестных в уравнении (1) уменьшается до одного.

Примеры построения точных доверительных интервалов (д.и.).
1. С.в. $X \sim R[2\theta-1, 3\theta+4]$. Построить д.и. с уровнем доверия γ для неизвестного параметра θ.
Решение.
Введем новую случайную величину (с.в.) $Y = X-2\theta+1$, тогда с.в. $Y \sim R[0,\theta+5] = R[0, \theta\ *]$, где $\theta = \theta +5$. Построим д.и. с доверительным уровнем γ для параметра θ.

Обозначим с.в. Z = Y(n)/ θ *, тогда Fz (Y) = P{Y(n) < θ *y}=y ^{n} – это функция распределения максимума выборки n равномерно распределенных на [0,1] значений $y_1,\ldots,y_n$. Распределение с.в. Z не зависит от θ, поэтому Z – **ц.с.:**

$$p\left\{ t_\gamma < Y_{(n)}/\theta^* < 1\right\} = \gamma = P\{t_\gamma < Z < 1\} = F_Z(1) - F_Z(t_\gamma) = 1 - t_\gamma^n, \text{ откуда}$$

$$\gamma = P\left\{Y_{(n)} < \theta^* = \theta + 5 < Y_{(n)} \Big/ \sqrt[n]{1-\gamma}\ \right\} =$$

$$= P\{x(n) - 2\theta - 4 < \theta < \frac{x_{(n)} - 2\theta + 1}{\sqrt[n]{1-\nu}} - 5\},$$

или с вероятностью γ выполнены неравенства:

$$\begin{cases} \dfrac{x_{(n)} + 1 - 5\sqrt[n]{1-\nu}}{\sqrt[n]{1-\nu} - 2} > \theta \\ (x_{(n)} - 4)/3 < \theta \end{cases},$$

то есть $P\left\{\dfrac{x_{(n)} - 4}{3} < \theta < \dfrac{x_{(n)} + 1 - 5\sqrt[n]{1-\gamma}}{\sqrt[n]{1-\gamma} - 2}\right\} = \gamma$.

Требуемый интервал (t_1, t_2) для θ построен:

$$t_1 = \frac{x_{(n)} - 4}{3};\ t_2 = \frac{x_{(n)} + 1 - 5\sqrt[n]{1-\gamma}}{\sqrt[n]{1-\gamma-2}}.$$

2. С.в. X ~ E(a θ +b). Построить д.и. для неизвестного параметра θ с уровнем доверия γ, где a и b – const, a > 0.

Решение.

Обозначим θ * = a θ +b и построим сначала д.и. для параметра θ с уровнем доверия γ.

X~E(θ*), $F(x) = 1 - \exp\{-\theta^* x\}$, $x \ge 0$,

$$F_{(1)}(x) = 1 - (1 - F(x))^n = 1 - \exp\{-\theta^* nx\}.$$

Введем новую с.в. Y= $\theta^* x_{(1)}$.

$F_Y(x)$ = P{x(1)<x/ θ*} = F(1)(x/ θ*) = 1– exp{−nx}– это функция распределения минимума выборки экспоненциально распределенных значений $y_1,\ldots,y_n$ с.в. Y~E(1), поэтому Y– **ц.с.** и

$$\gamma = P\{0 < Y < \theta^* x_{(1)} < t_\gamma\} = P\{0 < \theta^* < t_\gamma / x_{(1)}\} =$$

$$= F_Y(t_\gamma) - F_Y(0) = 1 - \exp\{-nt_\gamma\}, \text{ откуда } t_\gamma = -(\ln(1-\gamma))/n.$$

Тогда

$$P\left\{0<\theta^*<-\frac{\ln(1-\gamma)}{nx_{(1)}}\right\}=P\left\{0<a\theta+b<-\frac{\ln(1-\gamma)}{nx_{(1)}}\right\}=$$

$$=P\left\{-b/a<\theta^*<-\frac{\ln(1-\gamma)}{anx_{(1)}}-b/a\right\},$$

то есть требуемый интервал (t_1, t_2) для θ построен:

$$t_1=-\frac{b}{a},\quad t_2=\frac{\ln(1-\gamma)}{ax_{(1)}}-\frac{b}{a}.$$

Рассмотрим приведенные выше идеи построения центральной статистики на примерах, и с их помощью - нахождения доверительных интервалов для параметра θ.

3. Построить доверительный интервал для m: с.в. X~N (m; σ), где σ – известный параметр, m – неизвестно.

Построим статистику $\bar{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i$ и найдем её распределение методом характеристических функций. С.в. X~N(m;σ) ⇒

$$\Rightarrow g_{x_i}(t)=e^{itm-\frac{\sigma^2\cdot t^2}{2}}\Rightarrow g_{\bar{x}}(t)=\left(g_{x_i}\left(\frac{t}{n}\right)\right)^n=\left(e^{i\frac{t}{n}m-\frac{\sigma^2t^2}{2n^2}}\right)^n=$$

$$=e^{itm-\frac{\left(\frac{\sigma}{\sqrt{n}}\right)^2t^2}{2}}\Rightarrow \bar{x}\sim N\left(m;\frac{\sigma}{\sqrt{n}}\right)\Rightarrow M\bar{x}=m;\ D\bar{x}=\frac{\sigma^2}{n};$$

$$\sigma\bar{x}=\sqrt{D\bar{x}}=\frac{\sigma}{\sqrt{n}}\cdot I\left(\frac{\bar{x}-M\bar{x}}{\sqrt{D\bar{x}}}\right)\sim N(0;1)\Rightarrow T=\frac{\bar{x}-M\bar{x}}{\sqrt{D\bar{x}}}=\frac{(\bar{x}-m)\sqrt{n}}{\sigma}-$$ центральная статистика ⇒

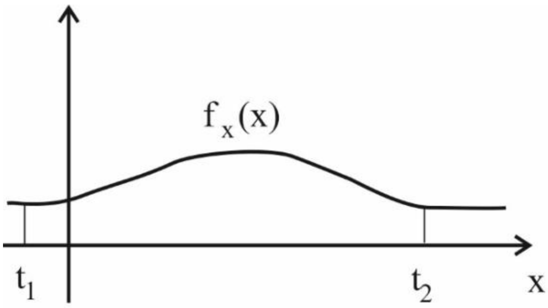
$$\gamma=P(|\bar{x}-m|<\delta)=P(|\bar{x}-M\bar{x}|<\delta)=P\left(|T|<\frac{\delta\sqrt{n}}{\sigma}=t_\gamma\right)=2\Phi(t_\gamma)\Rightarrow$$

⇒ t_γ находим по таблице функции Лапласа. Тогда с заданной вероятностью γ должно выполняться неравенство:

$$\frac{|\bar{x}-m|\sqrt{n}}{\sigma}=|T|<t_\gamma,\quad \text{или}\quad |\bar{x}-m|<\frac{t_\gamma\delta}{\sqrt{n}},\quad \text{или}\quad \bar{x}-\frac{t_\gamma\sigma}{\sqrt{n}}<m<\bar{x}+\frac{t_\gamma\sigma}{\sqrt{n}}\Rightarrow$$

искомый доверительный интервал $Y = (t_1, t_2) = \left(\bar{x} \mp \frac{t_\gamma \delta}{\sqrt{n}}\right)$

- центральный доверительный интервал. Если $f_X(x)$ плотность распределения случайной величины X ~N (m,σ), то (смотри рисунок).



$$t_1 = x_{\frac{1-\gamma}{2}} = \bar{x} - \frac{t_\gamma \sigma}{\sqrt{n}}; \quad t_2 = x_{\frac{1+\gamma}{2}} = \bar{x} + \frac{t_\gamma \sigma}{\sqrt{n}}.$$

Числовые примеры.
а) С.в. X ~ N (m;σ=3) . Найти доверительный интервал J для неизвестного параметра m с доверительной вероятностью $\gamma = 0{,}95$.
Решение.

t_γ находим по таблице функции Лапласа из условия $\Phi(t_\gamma) = \frac{\gamma}{2} = 0{,}475 \Rightarrow$

$$t_\gamma = 1{,}96; \; Y = \left(\bar{x} \mp \frac{1{,}96 \cdot 3}{\sqrt{36}}\right) = (\bar{x} \mp 0{,}98).$$

Например, при $\bar{x} = 4$ имеем J = (3,02; 4,98).
б) В предыдущей задаче найти минимальный объем выборки n , при котором с заданной точностью $\delta = 0{,}98$ и надежностью $\gamma = 0{,}95$ можно оценить параметр m.
Решение.

$$\delta = \frac{t_\gamma \sigma}{\sqrt{n}} \Rightarrow n = \frac{t_\gamma^2 \sigma^2}{\delta^2} = \frac{(1{,}96)^2 \cdot 3^2}{(0{,}98)^2} \approx 36.$$

Примеры построения асимптотических доверительных интервалов J для неизвестного параметра θ с уровнем доверия γ по выборке: $x_1, x_2, \ldots, x_n$, где n − объём выборки, для следующих распределений:

1. С.в. X~B(N,θ).

Решение.
Рассмотрим статистику

$$\hat\theta = \frac{1}{nN}\sum_{i=1}^{n} x_i;\ M\hat\theta = \hat\theta;\ D\hat\theta = \frac{1}{n^2N^2}\cdot nN\theta(1-\theta) = \frac{\theta(1-\theta)}{nN};\ \text{тогда}$$

$$L\left(\frac{\hat\theta - M\hat\theta}{\sqrt{D\hat\theta}}\right)\xrightarrow[n\to\infty]{} N(0,1) \Rightarrow T = \frac{\hat\theta - M\hat\theta}{\sqrt{D\hat\theta}} = \frac{(\hat\theta - \hat\theta)\sqrt{nN}}{\sqrt{\theta(1-\theta)}} \quad -$$

– **центральная статистика** $\Rightarrow$

$$\gamma = P\left(\left|\hat\theta - \theta\right| < \delta\right) = P\left(\left|\hat\theta - M\hat\theta\right| < \delta\right) = P\left(|T| < \frac{\delta\sqrt{nN}}{\sqrt{\theta(1-\theta)}} = t_\gamma\right) = 2\Phi\left(t_\gamma\right) \Rightarrow$$

$\Rightarrow$значение t_γ находим по таблице функции Лапласа. Тогда с заданной вероятностью γ должно выполняться неравенство:

$$\frac{\left|\hat\theta - \theta\right|\sqrt{nN}}{\sqrt{\theta(1-\theta)}} < t_y \Rightarrow \frac{(\hat\theta - \theta)^2 nN}{(1-\theta)\theta} < t_y^2 \Rightarrow$$

$$\Rightarrow f(\theta) = \hat(\hat\theta^2 - 2\theta\hat\theta + \theta^2)nN - t_y{}^2\theta(1-\theta) < 0 \Rightarrow$$

$$\Rightarrow \quad \theta^2(nN + t_y{}^2) - 2\theta\,(\hat\theta nN + t_y{}^2) + \hat\theta^2 nN < 0.$$

Решая неравенство относительно θ, получим искомый доверительный интервал $J = (\theta_1,\theta_2)$, где θ_1 и θ_2 - корни уравнения $f(\theta)=0$.

2. С.в. $X\sim\pi(\theta)$

Решение. Рассмотрим статистику

$$\hat\theta = \bar x = \frac{1}{n}\sum_{i=1}^{n} x_i;\ M\hat\theta = \theta;\ D\hat\theta = \frac{DX}{n} = \frac{\theta}{n};\ \text{тогда}$$

$$L\left(\frac{\bar x - M\bar x}{\sqrt{Dx}}\right)\xrightarrow[n\to\infty]{} N(0,1) \Rightarrow T = \frac{\bar x - M\bar x}{\sqrt{Dx}} = \frac{(\bar x - \theta)\sqrt{n}}{\sqrt{\theta}}$$ – центральная статистика $\Rightarrow \gamma = \left(P\left|\bar x - \theta\right| < \delta\right) = P\left(\left|\bar x - M\bar x\right| < \delta\right) = P\left(|T| < \frac{\delta\sqrt{n}}{\sqrt{\theta}} = t_\gamma\right) = 2\Phi(t_\gamma)$, где значение t_γ находят по таблице функции Лапласа*.

$$\bar x \sim N\left(\theta, \frac{\sigma}{\sqrt{n}}\right);\quad \frac{nS_1^2}{\sigma^2} \sim \chi^2_{n\text{-}1}$$

(было показано раньше)

Тогда с заданной вероятностью γ должно выполняться неравенство:

$$\frac{(\bar x - \theta)\sqrt{n}}{\sqrt{\theta}} < t_\gamma \Rightarrow \frac{(\bar x - \theta)^2 n}{\theta} < t_\gamma^2 \Rightarrow \left(\bar x^2 - 2\theta\bar x + \theta^2\right) n - t_\gamma^2\theta = f(0) < 0.$$

Отсюда находят искомый доверительный интервал $J = (\theta_1, \theta_2)$, где θ_1 и θ_2 – корни уравнения $f(\theta) = 0$.

*Табличная функция Лапласа $\Phi(x) = \frac{1}{2\pi}\int_0^x e^{-\frac{t^2}{2}}dt$;

3. С.в. X ~ N (θ = m,σ), σ− неизвестно

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\hat{\sigma} = \sqrt{S_1^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$;

Покажем, что случайная величина $T = \frac{\bar{x}-\theta}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{(\bar{x}-\theta)\sqrt{n}}{\hat{\sigma}}$ – центральная статистика. Известно, что случайная величина $\frac{Z\sqrt{n}}{\sqrt{V}}$ имеет распределение Стьюдента с плотностью

$$S_{n-1}(t) = \frac{\Gamma\frac{n}{2}}{\sqrt{\pi(n-1)\cdot\Gamma\frac{n-1}{2}}}\cdot\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$$ – четная табличная функция, где

с.в. Z ~ N(0,1), с.в. V ~ $\chi^2_{n\text{-}1}$. Покажем, что случайная величина T распределена по Стьюденту, а значит, является центральной статистикой.

Действительно, $T = \frac{(\bar{x}-\theta)\sqrt{n}}{\hat{\sigma}}\cdot\frac{\sigma}{\sqrt{n}\cdot\hat{\sigma}}\cdot\sqrt{n}$, где $Z = \frac{(\bar{x}-\theta)\sqrt{n}}{\hat{\sigma}} \sim N(0,1)$, а

$\frac{\sigma}{\sqrt{n}\cdot\hat{\sigma}} = \frac{1}{\sqrt{V}} \Rightarrow V = \frac{n\cdot\hat{\sigma}^2}{\sigma^2} = \frac{n\cdot S_1^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow L(T)$– распределение Стьюдента.

Тогда $\gamma = P(|T| < t_\gamma) = P\left(\frac{|\bar{x}-\theta|\sqrt{n}}{\hat{\sigma}} < t_\gamma\right) = P\left(|\bar{x}-\theta| < \frac{t_\gamma\cdot\hat{\sigma}}{\sqrt{n}} = \sigma\right)$, где t_γ находится по таблицам $S_{n-1}(t)$ $S_{n-1}(t)$ − четная функция ⇒

$P|T| < t_\gamma = 2\cdot\int_0^{t_\gamma} S_{n-1}(t)dt = \gamma$. Поэтому $J = \left(\bar{x} \mp \frac{t_\gamma\hat{\sigma}}{\sqrt{n}}\right)$ – искомый доверительный интервал.

Числовой пример.

$\bar{x} = (2{,}5;\ 2;\ -2{,}3;\ 1{,}9;\ -2{,}1;\ 2{,}4;\ 2{,}3;\ -2{,}5;\ 1{,}5;\ 1{,}7)$; n=10; $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 0{,}4$;

$$S_1^2=\frac{1}{10}\sum_{i=1}^{10}\left(x_i-\bar{x}\right)^2=5;\ \hat{\sigma}=\sqrt{5};\quad n-1=9;\quad \gamma=0{,}95;\quad t_\gamma=2{,}26$$

$$\sigma=\frac{t_\gamma}{\sqrt{n}}=\frac{2{,}26\cdot\sqrt{5}}{\sqrt{10}}\approx 1{,}58\Rightarrow J=\left(0{,}4\mp 1{,}58\right)=\left(1{,}58;1{,}98\right).$$

4. С.в. $X\sim N(m,\sigma)$, m−неизвестен. Найти д.и. J_δ для δ.

Известно, что статистика $\chi_{n-1}^2=\dfrac{S_1^2}{\sigma^2}$, где $S_1^2=\dfrac{1}{10}\sum\limits_{i=1}^{n}\left(x_i-\bar{x}\right)^2$, имеет

χ_{n-1}^2 распределение с плотностью $k_n(x)=f(x)=\dfrac{x^{\frac{n-3}{2}}\cdot e^{-\frac{x}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}$

(это табличная функция).

Построим новую случайную величину $Y=\sqrt{X^2}=\dfrac{S_1\sqrt{n}}{\sigma}$, где $S_1=\sqrt{S_1^2}$.

Найдём плотность распределения случайной величины Y: g(y).

Определение. $Y=\varphi(x)$; $f(x)$ − плотность распределения случайной величины X; $g(y)$ − плотность распределения случайной величины Y. $\varphi(x)$ − монотонная функция; $x=\psi(y)\Rightarrow g(y)=f\left(\psi(y)\right)\left|\psi'(y)\right|$ $\qquad(*)$

Воспользуемся формулой $(*)$: $y=\sqrt{x}=\varphi(x);\ x=y^2=\psi(y);\ \psi'(y)=2y$;

$$g(y)=\frac{\left(y^2\right)^{\frac{n-3}{2}}\cdot e^{-\frac{y^2}{2}}}{2^{\frac{n-1}{2}}\cdot\Gamma\left(\frac{n-1}{2}\right)}\cdot 2y=\frac{y^{n-2}e^{-\frac{y^2}{2}}}{2^{\frac{n-3}{2}}\cdot\Gamma\left(\frac{n-1}{2}\right)}$$ – это табличная функция «хи»

– χ.

По данному γ ищем доверительный интервал J для σ из условия:

$\gamma=P(|\sigma-S_1|<\delta=P(S_1-\delta)=P\{S_1(1-q)<\sigma<S_1(1+q)\}$, где $q=\dfrac{\delta}{S_1}$.

а) При $q<1$ имеем

$$\gamma=P\left\{\frac{1}{S_1(1+q)}\right\}<\frac{1}{q}<\frac{1}{S_1(1-q)}=P\left\{\frac{\sqrt{n}}{1+q}<\frac{S_1\sqrt{n}}{\sigma}<\frac{\sqrt{n}}{1-q}\right\}=P\left\{\chi_1<\chi<\chi_2\right\},$$

где $\chi_1 = \dfrac{\sqrt{n}}{1+q}$; $\chi_2 = \dfrac{\sqrt{n}}{1-q}$, так что $\gamma = \int\limits_{x_1}^{x_2} g(y)dy$.

По таблице распределения «хи» по γ и n находим q, а S_1 вычисляем по выборке $\Rightarrow J_\sigma = \{S_1(1-q) < \sigma < S_1(1+q)\}$.

б) При q >1

$\gamma=P\{S_1\ (1-q) < \sigma < S_1\ (1+q)\} = P\{0< \sigma< S_1\ (1+q)\}$, так как

$\sigma \ge 0 \Rightarrow J = \{0, S_1(1=q)\}$, так как $\gamma=P\{\ \sigma< S_1(1+q)\}=$

$= P\left\{\dfrac{1}{\sigma} > \dfrac{1}{S_1(1+q)}\right\} = P\left\{\dfrac{S_1\sqrt{n}}{\sigma} > \dfrac{\sqrt{n}}{1+q}\right\} = P\{\chi > \chi_1\}$, где $\chi_1 = \dfrac{\sqrt{n}}{1+q}$, так что

$$\gamma = \int\limits_{\gamma_1}^{\infty} g(y)dy.$$

В итоге получаем:

$$J_\sigma = \begin{cases} \{S_1(1-q)\}, S_1(1+q), & \text{при } q<1 \\ \{0, S_1(1+q)\}, & \text{при } q>1. \end{cases}$$

Числовые примеры.

X ~ N(m,σ), m − неизвестно. Найти J для σ.

1. n=25; $\sqrt{S_1^2 = S_1 =}0{,}8$; γ=0,95.

Решение. По таблице распределения χ находим значение q=0,32 $\Rightarrow$ искомый д.и. J для σ есть

J={$0{,}8\cdot(1\mp 0{,}32)$}={0,544;1,056}

2. n=10; $S_1 = 0{,}16$; γ=0,99

Решение. По таблице распределения χ находим значение q=1,8 (q>1) $\Rightarrow$ искомый д.и. для σ есть J={0;0,16(1+1,8)}={0;0,448}.

5. С.в. X~N(m, σ^2) **,** γ − заданный уровень доверия. Найти J_δ и $J_\delta{}^2$ д.и. для δ и δ^2=DX в случаях:

а) EX=m – известно

б) EX=m – неизвестно

Решение.

а) Построим статистику $S_0^{\hat{2}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m) = \sum_{i=1}^{n}\left(\frac{x_i - m}{\sqrt{n}}\right)$. Найдём распределение S_0^2 методом характеристических функций.

$$g_{\frac{x_i - m}{\sqrt{n}}}(t) = g_{\frac{x_i}{\sqrt{n}} - \frac{m}{\sqrt{n}}}(t) = \left\langle g_{ax+b(x)} = e^{itb} g_x(at)\right\rangle =$$

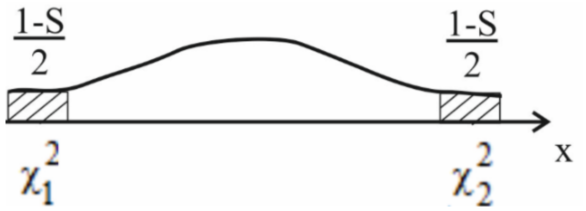
$$= e^{it\frac{m}{\sqrt{n}}} e^{im\frac{t}{\sqrt{n}}} e^{-\frac{\sigma^2 t^2}{2n}} = e^{\frac{\left(\frac{\sigma}{\sqrt{n}}\right)^2 \cdot t}{2}} \Rightarrow \frac{x_i - m}{\sqrt{n}} \sim$$

$$\sim N\left(0, \sigma^* = \frac{\sigma}{\sqrt{n}}\right) \Rightarrow S_0^2 = (\sigma^*)^2 \cdot u, \text{ где } \quad u \sim \chi_n^2 \Rightarrow \frac{S_0^2}{(\sigma^*)^2} \sim \chi_n^2 \text{ или } \frac{S_0^n n}{\sigma^2} \sim \chi_n^2$$

$$\Rightarrow \int_{\chi_1^2}^{\chi_2^2} k_n(x) \cdot dx = \gamma, \quad \text{где} \quad k_n(x) - \text{плотность распределения } \chi_n^2,$$

а значения χ_1^2 и χ_2^2 находятся по таблице из условий:

$$\begin{cases} \int_0^{\chi_1^2} k_n(x) \cdot dx = \frac{1-\gamma}{2} \\ \int_{\chi_2^2}^{\infty} k_n(x) \cdot dx = \frac{1-\gamma}{2} \end{cases}$$



Тогда

$$\gamma = P\left\{\chi_1^2 < \frac{nS_0^2}{\sigma^2} < \chi_2^2\right\} = P\left\{\frac{nS_0^2}{\chi_2^2} < \sigma^2 < \frac{nS_0^2}{\chi_1^2}\right\} \Rightarrow$$

$$J_{DX} = \left\{\frac{nS_0^2}{\chi_2^2}; \frac{nS_0^2}{\chi_1^2}\right\} \Rightarrow J_\sigma\left\{\frac{\sqrt{n} \cdot S_0}{\chi_1}; \frac{\sqrt{n} \cdot S_0}{\chi_2}\right\}.$$

б) Построим статистику $S_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Известно, что $\frac{nS_1^2}{\sigma^2} \sim \chi_{n-1}^2$ с плотностью распределения $k_{n-1}(x)$, γ – задано. Тогда из условий

$$\int_{\chi_1^2}^{\chi_2^2} k_{n-1}(x) \cdot dx = \gamma$$

и

$$\begin{cases} \int\limits_0^{\chi_1^2} k_{n-1}(x)\cdot dx = \dfrac{1-\gamma}{2} \\ \int\limits_{\chi_2^2}^{\infty} k_{n-1}(x)\cdot dx = \dfrac{1-\gamma}{2} \end{cases}$$

находим по таблицам χ^2 значения χ_1^2 и χ_2^2, откуда получаем

$$P\left\{\chi_1^2 < \frac{nS_1^2}{\sigma^2} < \chi_2^2\right\} = P\left\{\frac{nS_1^2}{\chi_2^2} < \sigma^2 < \frac{nS_1^2}{\chi_1^2}\right\} \Rightarrow$$

$$\Rightarrow J_{DX} = \left\{\frac{nS_1^2}{\chi_2^2}; \frac{nS_1^2}{\chi_1^2}\right\} \Rightarrow J_{\sigma} = \left\{\frac{\sqrt{n}\cdot S_1}{\chi_1}; \frac{\sqrt{n}\cdot S_1}{\chi_2}\right\},$$

где $S_1 = \sqrt{S_1^2}$, а χ_1 и χ_2 имеют распределение «хи»$-\ \chi$ (табличное).

Числовой пример.

N=22; $\bar{x} = 3{,}031$; $\hat{\sigma}^2 = S_1^2 = 0{,}03$; $\gamma = 0{,}98$; $\chi_1^2 = 8{,}897$; $\chi_2^2 = 38{,}932$

(по таблицам χ^2)

$$\frac{nS_1^2}{\chi_1^2} = 0{,}0715; \quad \frac{nS_1^2}{\chi_1^2} = 0{,}0163; \quad J_{DX} = \{0{,}0163; 0{,}0715\}.$$

Библиографический список.

1. Энатская Н.Ю., Е.Р. Хакимуллин. Теория вероятностей. Уч. пос., МИЭМ, М., 2010.
2. Боровков А.А. Теория вероятностей. М.: Наука, 1986.
3. Ротарь В.И. Теория вероятностей. М.: Высшая школа, 1992.
4. Дейвид Г. Порядковые статистики. М.: Наука, 1979.
5. Ивченко Г.И., Медведев Ю.И. Математическая статистика. М.: Высшая школа, 1984.
6. Ивченко Г.И., Чистяков В.П. Методические указания по использованию ЭВМ в преподавании математической статистики. МИЭМ. М., 1987.
7. Морозова Н.М., Энатская Н.Ю. Моделирование и обработка статистических данных (методические указания). МГИЭМ. М., 1990.
8. Кендалл М. Дж., Стюарт А. Теория распределений. М.: Наука, 1966.
9. Ивченко Г.И., Энатская Н.Ю. Основные сведения по распределениям. МИЭМ. М., 2008.

ЭНАТСКАЯ Наталия Юрьевна
ХАКИМУЛЛИН Евгений Робертович

Математическая статистика

Редактор Е. С. Резникова
Технический редактор О. Г. Завьялова

Подписано в печать . Формат 60х84/16.
Бумага офсетная №2. Ризография. Усл. печ. л.
Уч. – изд. л. . Изд. № . Тираж . Заказ .
Московский государственный институт электроники и математики.
109028, Москва, Б. Трехсвятительский пер., 3.
Отдел оперативной полиграфии Московского государственного института электроники и математики.
115054, Москва, ул. М. Пионерская, 12.