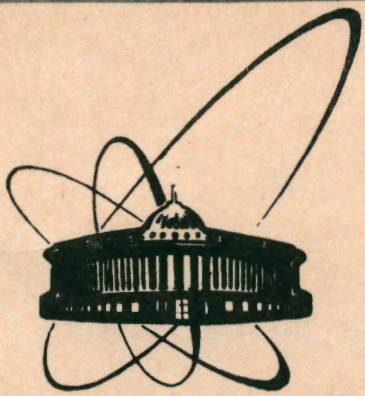

объединенный
институт
ядерных
исследований
дубна

Д11-92-139

П.В.Зрелов, В.В.Иванов

# ФУНКЦИИ РАСПРЕДЕЛЕНИЯ СТАТИСТИКИ $\omega_n^3 = n^{3/2} \int_{-\infty}^{\infty} [S_n(x) - P(x)]^3 dP(x)$ ДЛЯ МАЛЫХ $n$

Направлено в журнал "Математическое моделирование"

1992

# 1 Введение

Среди статистических методов, используемых в различных областях науки и техники, широкое распространение получили методы, основанные на непараметрических статистиках. В частности, в задачах физики высоких энергий, таких как планирование эксперимента, обработка экспериментальных данных и их сопоставление с теоретическими моделями, активно используются непараметрические критерии на основе статистики $\omega_n^2$ [1,2].

В работе [3] предложена новая статистика такого типа

$$\omega_n^3 = n^{\frac{3}{2}} \int_{-\infty}^{\infty} [S_n(y) - P(y)]^3 dP(y), \qquad (1)$$

где $P(y)$ - теоретическая, а $S_n(y)$ - эмпирическая функции распределения случайной величины $y$, $n$ - объем выборки. Для практических применений удобен алгебраический вид статистики (1)

$$\omega_n^3 = -\frac{\sqrt{n}}{8} \sum_{i=1}^{n} \left[2P(y_i) - \frac{2i-1}{n}\right] \left\{ \left[2P(y_i) - \frac{2i-1}{n}\right]^2 + \frac{1}{n^2} \right\}, \qquad (2)$$

где $y_1 < y_2 < ... < y_n$ – вариационный ряд по выборке объёма $n$. В [3] вычислены среднее значение и дисперсия распределения $\omega_n^3$, а также таблицы процентных точек для малых $n$.

Процентные точки были получены с помощью геометрического метода Монте-Карло в соответствии с алгоритмом, описанным в работе [4]. Однако точность значений процентных точек невысока и может оказаться недостаточной для практических применений.

Ввиду того, что ошибка в оценке функции распределения с помощью указанного метода убывает как $\frac{1}{\sqrt{N}}$ ($N$ - число разыгранных случайным образом выборок объёма $n$ значений статистики $\omega_n^3$), для достижения более высокой точности, необходимо увеличить $N$ на $2 \div 3$ порядка. Это ведёт к пропорциональному росту времени вычислений, превращая рассматриваемую задачу в трудноразрешимую на ЭВМ проблему.

В настоящей работе предложен метод вычисления процентных точек функции распределения $\omega_n^3$ для малых объемов выборки $n$, представляющий собой модификацию подхода М.Нотта [5], используемого для определения процентных точек распределения статистики $\omega_n^2$. Метод основан на численном определении характеристической функции статистики $\omega_n^3$ и её последующей инверсии.

# 2 Метод вычислений

Введем переменную $x = P(y)$. Когда $y$ пробегает всю действительную ось, $x$ изменяется в интервале (0,1); при этом равенство (2) принимает вид

$$\omega_n^3 = -\frac{\sqrt{n}}{8}\sum_{i=1}^{n}\left(2x_i - \frac{2i-1}{n}\right)\left[\left(2x_i - \frac{2i-1}{n}\right)^2 + \frac{1}{n^2}\right]. \qquad (3)$$

Раскроем скобки в выражении (3) и приведем подобные члены:

$$\omega_n^3 = -\frac{\sqrt{n}}{2}\sum_{i=1}^{n}\left(2x_i^3 - 3x_i^2\frac{2i-1}{n} + 2x_i\frac{3i^2-3i+1}{n^2}\right) + C_n. \qquad (4)$$

Константа $C_n$ в (4) равна

$$C_n = \frac{1}{4n^{\frac{5}{2}}}\sum_{i=1}^{n}\left(4i^3 - 6i^2 + 4i - 1\right) = \frac{1}{4}n^{\frac{3}{2}}, \qquad (5)$$

что совпадает с максимально возможным значением $\omega_n^3$ [3].

Обозначим через $z_n^3$ величину

$$z_n^3 = C_n - \omega_n^3. \qquad (6)$$

Случайная величина $z_n^3$ принимает значения на отрезке $[0, \frac{1}{2}n^{\frac{3}{2}}]$, а соответствующая ей функция распределения $F(z)$ равна нулю на $(-\infty, 0)$. Используя характеристическую функцию $\Phi(t)$ величины $z_n^3$, можно записать двойное неравенство [5]

$$F(h) \le \frac{h}{\lambda} + \frac{2}{\pi}\sum_{k=1}^{\infty}\frac{sin\frac{hk\pi}{\lambda}}{k} Re\left[\Phi\left(\frac{k\pi}{\lambda}\right)\right] \le F(h) + [1 - F(\lambda)], \quad 0 < h < \lambda, \qquad (7)$$

обеспечивающее достаточно простую процедуру вычисления функции распределения $F(z)$.



## 2.1 Вычисление характеристической функции

Характеристическая функция $\Phi_n(t)$ случайной величины $z_n^3$ по определению равна:

$$\Phi_n(t) = n!\int_0^1\int_0^{x_n}\dots\int_0^{x_2} exp\left[it\frac{\sqrt{n}}{2}\sum_{j=1}^{n}\left(2x_j^3 - 3x_j^2\frac{2j-1}{n} + 2x_j\frac{3i^2-3j+1}{n^2}\right)\right]dx_1\dots dx_n. \qquad (8)$$

Введем обозначение

$$I(x,t,k) = k!\int_0^x\int_0^{x_k}\dots\int_0^{x_2} exp\left[it\frac{\sqrt{n}}{2}\sum_{j=1}^{k}\left(2x_j^3 - 3x_j^2\frac{2j-1}{n} + 2x_j\frac{3j^2-3j+1}{n^2}\right)\right]dx_1\dots dx_k. \qquad (9)$$

Заметим, что $I(1,t,n) = \Phi_n(t)$.

Запишем рекуррентное соотношение

$$I(x,t,k+1) = (k+1)\int_0^x exp\left\{it\frac{\sqrt{n}}{2}\left[2y^3 - 3y^2\frac{2k+1}{n} + 2y\frac{3k^2+3k+1}{n^2}\right]\right\}I(y,t,k)dy, \qquad (10)$$

справедливое при $k = 2,3,\dots,n-1$. Для того чтобы оно имело место и при $k = 0,1$, определим $I(x,t,1)$ в виде

$$I(x,t,1) = \int_0^x exp\left[it\frac{\sqrt{n}}{2}\left(2x_1^3 - \frac{3x_1^2}{n} + \frac{2x_1}{n^2}\right)\right]dx_1, \qquad (11)$$

а $I(x,t,0)$ положим равным 1 для любых $x$ и $t$.

Функцию $\Phi_n(t)$ не удается вычислить из (8) аналитически, однако, используя (9), ёе величину для любого заданного $t$ можно определить путём численного интегрирования. Для этого воспользуемся применённым в [5] методом вычисления $I(x,t,k)$ на решетке $x_j = \frac{j}{m}$ $(j = 1,2,\dots,m)$, где $m$ достаточно велико; величины $I(x,t,k+1)$ определяются на основании значений $I(x,t,k)$ в узлах выбранной решетки.

Положим

$$\Delta I(x,t,k) = I\left(x + \frac{1}{m}, t, k\right) - I(x,t,k) \qquad (12)$$

и

$$f_r = (k+1)exp\left\{it\frac{\sqrt{n}}{2}\left[2\left(\frac{r}{m}\right)^3 - 3\left(\frac{r}{m}\right)^2\frac{2k+1}{n} + 2\left(\frac{r}{m}\right)\frac{3k^2+3k+1}{n^2}\right]\right\}I\left(\frac{r}{m}, t, k\right). \qquad (13)$$

Используя (10) и (12), получим

$$\Delta I(x,t,k+1) = (k+1)\int_{x}^{x+\frac{1}{m}} exp\Big\{it\frac{\sqrt{n}}{2}\Big[2y^3-3y^2\frac{2k+1}{n}+ \\ +2y\frac{3k^2+3k+1}{n^2}\Big]\Big\}I(y,t,k)dy. \qquad (14)$$

Для вычисления $\Delta I(x,t,k+1)$ использовались квадратурные интерполяционные формулы, взятые из раздела $E$ таблиц [6]:

$$\Delta I(0,t,k+1) = (251f_0+646f_1-264f_2+106f_3-94f_4)/(720m),$$

$$\Delta I(\tfrac{1}{m},t,k+1) = (-19f_0+346f_1+456f_2-74f_3+11f_4)/(720m),$$

$$\Delta I(\tfrac{r}{m},t,k+1) = [802(f_r+f_{r+1})-93(f_{r-1}+f_{r+2})+ \\ +11(f_{r-2}+f_{r+3})]/(1440m) \quad \text{для} \quad 1<r<m-2,$$

$$\Delta I(\tfrac{m-2}{m},t,k+1) = (-19f_m+346f_{m-1}+456f_{m-2}-74f_{m-3}+ \\ +11f_{m-4})/(720m),$$

$$\Delta I(\tfrac{m-1}{m},t,k+1) = (251f_m+646f_{m-1}-264f_{m-2}+106f_{m-3}- \\ -19f_{m-4})/(720m).$$

Положив $I(0,t,k+1)=0$ для любого $t$, запишем тождество:

$$I(\frac{r}{m},t,k+1) = \sum_{\rho=1}^{r}\Delta I\Big(\frac{\rho-1}{m},t,k+1\Big), \quad r=1,2,...,m, \qquad (15)$$

которое определяет процедуру вычисления характеристической функции $\Phi_n(t) = I(1,t,k)$. Так как для всех рассмотренных $n$ и ряда значений $t$ увеличение числа разбиений $m$ от 500 до 800 не приводит к изменениям $F(h)$, превышающим по абсолютной величине $10^{-10}$, то для всех $n$ параметр $m$ задавался равным 500.

## 2.2 Инверсия характеристической функции

Процедура инверсии характеристической функции основывается на неравенстве (7), из которого видно, что точность определения $F(h)$ зависит от величины $\lambda$ и параметра $K$, ограничивающего число членов бесконечной суммы. В случае выбора $\lambda \geq \{z_n^3\}_{max} = \frac{1}{2}n^{\frac{3}{2}}$ разность $1-F(\lambda)$ обращается в нуль, и точность вычисления $F(h)$ определяется только значением $K$. Выбор величины $\lambda$ несколько большей максимального значения $z_n^3$ позволяет дополнительно контролировать точность определения функции $F(h)$ по её близости к единице при значениях $h$, превосходящих $\frac{1}{2}n^{\frac{3}{2}}$. Такой выбор $\lambda$ требует дополнительного увеличения числа



членов $K$, что приводит к пропорциональному росту времени счёта на ЭВМ из-за увеличения времени вычислений каждого последующего члена суммы. Поэтому для обеспечения заданной точности вычисления процентных точек значение параметра $K$ изменялось от 400 для выборки $n=1$ до 2100 в случае $n=10$.

На заключительном этапе значения функции распределения $F(h)$ величины $z_n^3$, вычисленные с мелким равномерным шагом $\Delta h$, преобразовывались в значения функции распределения $F_n(x)$ статистики $\omega_n^3$ согласно соотношению $F(h) = 1-F_n(x)$, где $x = C_n - h$. Полученные величины использовались для определения значений процентных точек $Z_p$, вычислявшихся с помощью подпрограммы $E105$ из библиотеки CERN [7], которая реализует интерполяцию функции одного аргумента методом конечных разностей.

# 3 Таблицы процентных точек $\omega_n^3$-распределения

В таблице 1 приведены процентные точки ($Z_p$) распределения статистики $\omega_n^3$ для $n = 1(1)5$, а в таблице 2 – для $n = 6(1)10$. В таблицах представлены точки $Z_p > 0$, поскольку из-за симметрии $\omega_n^3$ - распределения относительно нуля имеет место равенство

$$F_n(Z_p) = 1 - F_n(-Z_p). \qquad (16)$$

Точность вычисленных значений процентных точек оценивалась путём варьирования числа $K$, а также попарным сравнением абсолютных величин процентных точек, симметричных относительно $Z_p = 0$. Анализ показал, что для выбранных $m$ и $K$ процентные точки вычислены с точностью,не худшей 1 – 2 единиц пятого знака после запятой.

# 4 Заключение

Предложен метод вычисления функции распределения новой непараметрической статистики $\omega_n^3$, представляющий собой модификацию метода М.Нотта [5] (используемого для вычисления процентных точек распределения статистики $\omega_n^2$), и основанный на численном определении характеристической функции и её последующей инверсии. С высокой точностью, не худшей 1 – 2 единиц пятого знака после запятой, получены таблицы процентных точек $\omega_n^3$-распределения для выборок малых объемов: $n = 1(1)10$.

Полученные в настоящей работе таблицы позволили выполнить анализ экспериментальных данных (включающий выделение маловероятных событий), накопленных в опытах на синхрофазотроне ОИЯИ с помощью спектрометра **МАСПИК** [9], а также провести детальное сравнение некоторых традиционных статистических методов идентификации заряженных частиц [10] с методом на основе статистики $\omega_n^3$ [8].

Таблица 1. Процентные точки $Z_p$ случайной величины $\omega_n^3$:
$F_n(Z_p) = Pr\{\omega_n^3 < Z_p\}$, $n = 1, 2, ..., 5$

| Процентные точки $Z_p$ | | | | | |
|---|---|---|---|---|---|
| $F_n(Z_p)$ | n=1 | n=2 | n=3 | n=4 | n=5 |
| .50 | .00000 | .00000 | .00000 | .00000 | .00000 |
| .51 | .00250 | .00146 | .00131 | .00126 | .00124 |
| .52 | .00501 | .00294 | .00262 | .00253 | .00249 |
| .53 | .00753 | .00442 | .00394 | .00382 | .00376 |
| .54 | .01006 | .00592 | .00529 | .00514 | .00507 |
| .55 | .01262 | .00744 | .00667 | .00649 | .00641 |
| .56 | .01522 | .00898 | .00808 | .00790 | .00781 |
| .57 | .01784 | .01056 | .00954 | .00936 | .00926 |
| .58 | .02051 | .01217 | .01104 | .01088 | .01077 |
| .59 | .02323 | .01381 | .01261 | .01247 | .01234 |
| .60 | .02600 | .01550 | .01424 | .01413 | .01398 |
| .61 | .02883 | .01724 | .01595 | .01588 | .01569 |
| .62 | .03173 | .01902 | .01775 | .01770 | .01749 |
| .63 | .03470 | .02086 | .01965 | .01962 | .01937 |
| .64 | .03774 | .02276 | .02166 | .02163 | .02134 |
| .65 | .04088 | .02472 | .02378 | .02375 | .02340 |
| .66 | .04410 | .02676 | .02604 | .02598 | .02557 |
| .67 | .04741 | .02887 | .02842 | .02832 | .02785 |
| .68 | .05083 | .03107 | .03095 | .03078 | .03025 |
| .69 | .05436 | .03337 | .03362 | .03337 | .03278 |
| .70 | .05800 | .03578 | .03646 | .03611 | .03545 |
| .71 | .06176 | .03833 | .03947 | .03900 | .03827 |
| .72 | .06565 | .04104 | .04266 | .04206 | .04125 |
| .73 | .06967 | .04396 | .04604 | .04528 | .04442 |
| .74 | .07382 | .04713 | .04964 | .04871 | .04778 |
| .75 | .07813 | .05061 | .05346 | .05234 | .05136 |
| .76 | .08258 | .05443 | .05753 | .05619 | .05519 |
| .77 | .08718 | .05862 | .06186 | .06030 | .05629 |
| .78 | .09195 | .06320 | .06648 | .06468 | .06369 |
| .79 | .09689 | .06822 | .07142 | .06937 | .06844 |
| .80 | .10200 | .07369 | .07670 | .07440 | .07358 |
| .81 | .10729 | .07966 | .08236 | .07981 | .07913 |
| .82 | .11277 | .08618 | .08844 | .08566 | .08517 |
| .83 | .11844 | .09329 | .09498 | .09201 | .09173 |
| .84 | .12430 | .10107 | .10205 | .09894 | .09889 |
| .85 | .13038 | .10959 | .10972 | .10656 | .10673 |
| .86 | .13666 | .11895 | .11805 | .11499 | .11534 |
| .87 | .14315 | .12924 | .12716 | .12442 | .12485 |
| .88 | .14987 | .14061 | .13718 | .13500 | .13542 |
| .89 | .15682 | .15321 | .14827 | .14691 | .14724 |
| .90 | .16400 | .16726 | .16066 | .16040 | .16057 |
| .91 | .17142 | .18300 | .17465 | .17578 | .17577 |
| .92 | .17909 | .20078 | .19071 | .19351 | .19330 |
| .93 | .18701 | .22105 | .20953 | .21423 | .21387 |
| .94 | .19518 | .24443 | .23236 | .23886 | .23851 |
| .95 | .20363 | .27183 | .26159 | .26888 | .26890 |
| .96 | .21234 | .30464 | .30006 | .30676 | .30793 |
| .97 | .22132 | .34524 | .35176 | .35725 | .36108 |
| .98 | .23059 | .39829 | .42636 | .43142 | .44022 |
| .99 | .24015 | .47592 | .55265 | .56721 | .58339 |



Таблица 2. Процентные точки $Z_p$ случайной величины $\omega_n^3$:
$F_n(Z_p) = Pr\{\omega_n^3 < Z_p\}$, $n = 6, 7, ..., 10$

| Процентные точки $Z_p$ | | | | | |
|---|---|---|---|---|---|
| $F_n(Z_p)$ | n=6 | n=7 | n=8 | n=9 | n=10 |
| .50 | .00000 | .00000 | .00000 | .00000 | .00000 |
| .51 | .00123 | .00122 | .00121 | .00120 | .00120 |
| .52 | .00247 | .00245 | .00244 | .00241 | .00240 |
| .53 | .00373 | .00370 | .00368 | .00364 | .00363 |
| .54 | .00502 | .00498 | .00494 | .00490 | .00488 |
| .55 | .00634 | .00629 | .00625 | .00620 | .00617 |
| .56 | .00772 | .00765 | .00759 | .00754 | .00751 |
| .57 | .00915 | .00906 | .00899 | .00894 | .00890 |
| .58 | .01064 | .01053 | .01045 | .01040 | .01035 |
| .59 | .01220 | .01207 | .01197 | .01192 | .01187 |
| .60 | .01382 | .01367 | .01357 | .01351 | .01346 |
| .61 | .01551 | .01536 | .01525 | .01518 | .01513 |
| .62 | .01727 | .01711 | .01700 | .01692 | .01686 |
| .63 | .01912 | .01895 | .01884 | .01875 | .01869 |
| .64 | .02105 | .02087 | .02076 | .02066 | .02060 |
| .65 | .02309 | .02289 | .02278 | .02268 | .02261 |
| .66 | .02523 | .02502 | .02490 | .02481 | .02473 |
| .67 | .02749 | .02727 | .02714 | .02706 | .02697 |
| .68 | .02987 | .02966 | .02952 | .02943 | .02934 |
| .69 | .03237 | .03218 | .03204 | .03194 | .03184 |
| .70 | .03503 | .03485 | .03471 | .03458 | .03447 |
| .71 | .03785 | .03767 | .03753 | .03738 | .03727 |
| .72 | .04084 | .04065 | .04051 | .04036 | .04024 |
| .73 | .04403 | .04384 | .04367 | .04353 | .04340 |
| .74 | .04741 | .04724 | .04704 | .04689 | .04678 |
| .75 | .05103 | .05086 | .05065 | .05048 | .05035 |
| .76 | .05491 | .05470 | .05450 | .05431 | .05418 |
| .77 | .05905 | .05882 | .05860 | .05842 | .05829 |
| .78 | .06349 | .06324 | .06299 | .06283 | .06271 |
| .79 | .06827 | .06800 | .06774 | .06756 | .06744 |
| .80 | .07340 | .07309 | .07285 | .07268 | .07254 |
| .81 | .07895 | .07861 | .07835 | .07819 | .07807 |
| .82 | .08495 | .08460 | .08433 | .08418 | .08405 |
| .83 | .09146 | .09108 | .09086 | .09070 | .09057 |
| .84 | .09857 | .09818 | .09794 | .09782 | .09768 |
| .85 | .10634 | .10595 | .10577 | .10563 | .10550 |
| .86 | .11489 | .11452 | .11435 | .11423 | .11410 |
| .87 | .12435 | .12401 | .12391 | .12376 | .12365 |
| .88 | .13486 | .13462 | .13451 | .13440 | .13430 |
| .89 | .14667 | .14652 | .14647 | .14635 | .14626 |
| .90 | .16004 | .16001 | .15999 | .15989 | .15983 |
| .91 | .17536 | .17547 | .17545 | .17541 | .17539 |
| .92 | .19314 | .19337 | .19340 | .19342 | .19347 |
| .93 | .21409 | .21446 | .21457 | .21467 | .21477 |
| .94 | .23931 | .23977 | .23999 | .24027 | .24047 |
| .95 | .27039 | .27102 | .27151 | .27195 | .27226 |
| .96 | .31015 | .31107 | .31196 | .31266 | .31319 |
| .97 | .36395 | .36563 | .36710 | .36820 | .36905 |
| .98 | .44420 | .44752 | .45002 | .45185 | .45336 |
| .99 | .59220 | .59901 | .60385 | .60775 | .61077 |

## Литература


[1] Ludlam T. et al.: *Phys.Rev.D, 1973, v.8, No.5, p.408; Phys.Rev.D, 1977, v.16, No.1, p.100.*

[2] Зрелов П.В., Иванов В.В.: *ОИЯИ, Р10-86-812*, Дубна, 1986.

[3] Зрелов П.В., Иванов В.В.: *ОИЯИ, Р10-88-321*, Дубна, 1988.

[4] Зрелов П.В., Иванов В.В.: *ОИЯИ, Р10-86-547*, Дубна, 1986.

[5] Knott M. *The Distribution of the Gramer – Von Mises Statistic for Small Sample Sizes.* J. Roy. Statist. Soc. B36, 3, 430 – 436, 1973.

[6] Interpolation and Allied Tables (1956). London: H.M.S.O.

[7] James F.: *CERN Computer Program Library E105.*

[8] Зрелов П.В., Иванов В.В.: *ОИЯИ, Р10-89-739*, Дубна, 1989.

[9] Ажгирей Л.С. и др.: В кн.: *Труды совещания по исследованиям в области релятивистской ядерной физики. ОИЯИ, Д2-82-568*, Дубна, 1982, с.83.

[10] Ramaha Murty P.V., Demeester G.D.: *Nucl. Instr. and Meth., 1967, 56, p.93.*



Рукопись поступила в издательский отдел
26 марта 1992 года.





Зрелов П.В., Иванов В.В.


Д11-92-139


Функция распределения статистики

$$\omega_n^3 = n^{3/2} \int_{-\infty}^{\infty} [S_n(x) - P(x)]^3 dP(x)$$ для малых $n$


Рассмотрен численный метод определения функции распределения новой непараметрической статистики $\omega_n^3$ для малых объемов выборки, позволивший с высокой точностью вычислить процентные точки для $n = 1, 2, ..., 10$.



Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1992



Zrelov P.V., Ivanov V.V.


D11-92-139


The Distribution Fuctions

of the $\omega_n^3 = n^{3/2} \int_{-\infty}^{\infty} [S_n(x) - P(x)]^3 dP(x)$ Statistics

for Small $n$


A numerical method of the distribution function calculation of a new nonparametric $\omega_n^3$ statistic for small sample sizes is considered. It allowed us to calculate with a high accuracy the percentage points for $n = 1, 2, ..., 10$.



The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna 1992