МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ»

Ю.С. Радченко, В.Н. Верещагин

МЕТОДЫ ОБРАБОТКИ И ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Часть 3
Непараметрические методы обработки данных

Учебно-методическое пособие для вузов

Воронеж
Издательский дом ВГУ
2018

ђqdbyeY bȨ debȨbYзe d езаbсу Јqdbуbyè

ОГЛАВЛЕНИЕ

Введение

Данное учебно-методическое пособие является продолжением пособий «Методы обработки и планирования эксперимента. Часть 1. Оценка распределений и их параметров» и «Методы обработки и планирования эксперимента. Часть 2. ПРОВЕРКА ГИПОТЕЗ, АППРОКСИМАЦИЯ РАСПРЕДЕЛЕНИЙ». В данном учебном пособии рассматриваются задачи проверки непараметрических гипотез о свойствах выборки, а также подробно изложены ранговые алгоритмы обработки данных. Некоторые тесты не входят в классический курс математической статистики. Поэтому изло-

ђqdbyeY bȨ debȨbYзe d езаbсу Јqdbуbyè

$H_0: x_p = F^{-1}(p)$,

H_1: a) $F^{-1}(p) > x_p$,

b) $F^{-1}(p) < x_p$,

c) $F^{-1}(p) \neq x_p$.

3. Задача о масштабе.

Генеральная совокупность имеет закон распределения вида: $F(x,\theta) = F(x \cdot \theta)$. Здесь θ - параметр, характеризующий скорость роста функции распределения, или масштаба распределения.

Основная гипотеза: $H_0: \theta = 1$,

альтернативы H_1: a) $\theta > 1$,

b) $\theta < 1$,

c) $\theta \neq 1$.

4 Задача о независимости элементов выборки.

Пусть $(x_1, x_2 .. x_n)$- выборка из генеральной совокупности, с некоторой неизвестной функцией распределения.

Основная гипотеза H_0 : $F(x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} F^{(i)}(x_i)$,

альтернатива H_1: $F(x_1, x_2 \ldots x_n) \neq \prod_{i=1}^{n} F^{(i)}(x_i)$.

Общий подход при решении непараметрических задач заключается в преобразовании выборки с целью свести задачу к задаче, в которой фигурируют известные распределения, т.е. непараметрическую гипотезу свести к параметрической и простой для основной гипотезы.

Суть преобразований – выявление непараметрических, не зависящих от распределения фактов, позволяющих решать поставленную задачу. При этом используются:

1. Закон больших чисел при $n >> 1$.
2. Внутреннее свойство выборки $(x_1, x_2 .. x_n)$.

Преобразование выборки с целью выявления непараметрических фактов.

№	Тип преобразований	Используемые непараметрические свойства
1	Перестановка элементов выборки	Равновероятность перестановок при симметричности закона распределения $W(x_1,x_2\ldots x_n))$
2	Поэлементное приведение выборки в интервал [0,1] с помощью обращения F(x)	Равномерность приведений выборки в [0,1], в случае, когда $F(x)=F_0(x)$
3	Упорядочивание элементов выборки по величине	Получаем порядковые статистики, которые сходятся по вероятности к квантилю распределения $\frac{R}{n+1}$, где n – объем выборки, R – порядок статистики (ранг), т.е. положение ее в упорядоченном ряду
4	Отображение выборки на пространство ранговых векторов	Равновероятность ранговых векторов при симметричности закона распределения $W(x_1,x_2\ldots x_n))$

2. Порядковые статистики, ранги и их статистические свойства

Порядковые статистики

Пусть имеется выборка $X_1,..,X_n$ из некоторой генеральной совокупности. Конкретные значения, полученные в опыте $x_1,..,x_n$ - выборочные значения или реализация выборки. Упорядочим выборочные значения в порядке возрастания: $x^{(1)},..,x^{(n)}$, где $x^{(i)} \le x^{(i+1)}$.Такой ряд называется вариационным рядом. ($X^{(1)},..,X^{(n)}$) – последовательность случайных величин. $X^{(i)}$ - порядковая статистика, $X^{(1)}$ и $X^{(n)}$ - экстремальные статистики. $X^{(n)}$ - $X^{(1)}$ характеризует размах выборки.

Если исходная выборка $(X_1,..X_n)$ является независимой, то в вариационном ряде $\left(X^{(1)},..X^{(n)}\right)$ элементы являются зависимыми. Кроме того, распределения значений вариационного ряда имеют различный вид. Если $F(x)$, $W(x)$ - законы распределения элементов исходной выборки, то плотность вероятности j – й порядковой статистики $X^{(j)}$, $1\le j\le n$ имеет вид

$$W_j(x)=\frac{(n-1)!}{(j-1)!(n-j)!}[F(x)]^{j-1}[1-F(x)]^{n-j}W(x)$$

На рис.1 приведены графики распределений порядковых статистик их нормально распределенной выборки $W(x)=exp\left(-x^2/2\right)/\sqrt{2\pi}$ при n=5.

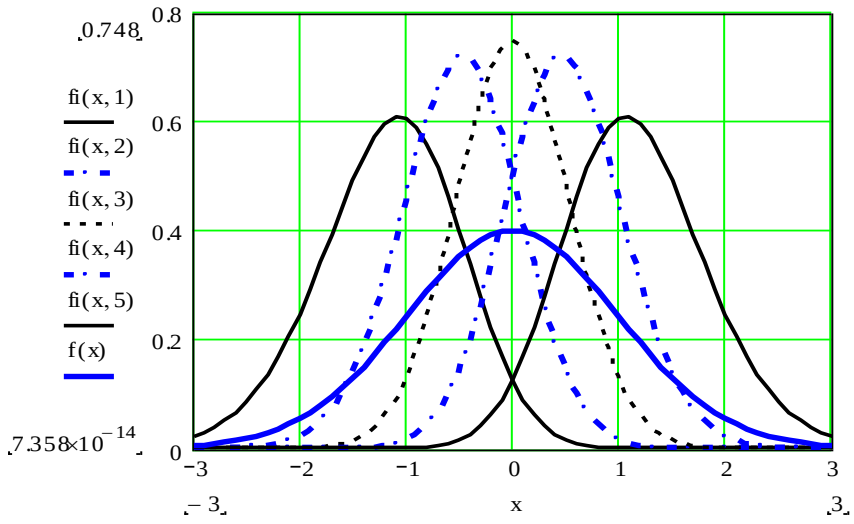

Рис. 1

Очевидно, что порядковые статистики имеют негауссовские распределения, а также различные математические ожидания и дисперсии

$$< X^{(j)} >\equiv m_j = \frac{n!}{(j-1)!(n-j)!}\int_{-\infty}^{\infty} u\left[F(u)^{j-1}\right]\left[1-F(u)\right]^{n-j} W(u)du$$

$$D_j = < \left(X^{(j)}\right)^2 > - m_j^2, \qquad \text{где}$$

$$< \left(X^{(j)}\right)^2 >= \frac{n!}{(j-1)!(n-j)!}\int_{-\infty}^{\infty} u^2\left[F(u)^{j-1}\right]\left[1-F(u)\right]^{n-j} W(u)du$$

2.2. Ранги выборки. Пусть имеется выборка

$X^T =$

	0	1	2	3	4	5	6	7	8	9
0	1.561	1.321	1.527	1.049	0.314	2.044	1.879	2.556	4.192	2.809

Упорядоченная выборка имеет вид

$PX^T =$

	0	1	2	3	4	5	6	7	8	9
0	0.314	1.049	1.321	1.527	1.561	1.879	2.044	2.556	2.809	4.192

Номер элемента X_i в упорядоченной выборке PX^T и есть ранг R_i данного элемента в исходной выборке. Ранги - целые числа.

$RX^T =$

	0	1	2	3	4	5	6	7	8	9
0	5	3	4	2	1	7	6	8	10	9

.

Ранг является функцией выборки и представляет собой случайную величину дискретного типа с возможными значениями 1,2,.., *n*. Совокупность рангов выборки называется ранговым вектором. Ранговый вектор - случайный целочисленный вектор. Реализациями (возможными значениями) этого вектора являются всевозможные перестановки чисел 1,2,..,*n*. Число возможных перестановок ***n!***

Необходимость исследования рангов определяется следующими обстоятельствами.

1. Ранговый вектор содержит часть информации, содержащейся в исходной выборке, т.к. с помощью упорядочения исходной выборке $\{X_i\}$ ставится в однозначное соответствие пара векторов - вектор порядковых статистик $\{X_{(R)}\}$ и ранговый вектор $\{R_i\}$. Располагая значениями $\{X_{(R)}\}$ и $\{R_i\}$ можно восстановить исходную выборку.

2. Используя информацию, содержащуюся в рангах, можно строить статистические процедуры, которые являются достаточно простыми ввиду ***целочисленности*** рангов.

3. Ранговые процедуры обладают свойствами ***непараметричности*** (не требуют знания закона распределения), при определенных условиях являются весьма эффективными .

4. Ранговые процедуры особенно важны, когда наблюдения носят не количественный, а качественный (нечисловой) характер и результаты наблюдений можно упорядочить.

Рангом R_i элемента выборки X_i называется число значений выборки, не превышающих X_i, $X_k \le X_i$.Чтобы посчитать ранг i-го элемента введем функцию $C(t) = \begin{cases} 1, t \ge 0 \\ 0, t < 0 \end{cases}$.

Тогда $R_i = \sum_{k=1}^{n} C(X_i - X_k)$ *, i=1,…,n, определяет случайный ранговый вектор* $\{R_i\}$. $\{R_i\}$ *и* $X^{(i)}$ *полностью в совокупности содержат всю информацию, которая находится в исходной выборке.*

2.3.Статистические свойства рангов при инвариантности к перестановкам

Рассмотрим свойства ранговых векторов, когда распределение выборки инвариантно по отношению к перестановкам аргументов (элементы выборки независимы и одинаково распределены).

Пусть упорядочивание производится непосредственно по алгебраическим значениям измеряемых величин. Т.к. $\{X_{(R)}\}$ и $\{R_i\}$ находятся во взаимно однозначном соответствии с исходной выборкой $\{X_i\}$, то

$$f_x(x_1, x_2, \ldots x_n) = f\left(\{x_{(R)}\}, \{r_i\}\right).$$

Здесь $x_{(R)}$ и r_i - соответственно значения элементов векторов порядковых статистик и рангового вектора. Ввиду инвариантности плотности вероятности к перестановке аргументов:

$$f\left(\{x_{(R)}\}, \{r_i\}\right) = f\left(x_{(1)}, x_{(2)}, \ldots x_{(n)}\right).$$

Безусловное распределение рангового вектора является равномерным:

$$P\left(\{R_i\}\right) = \frac{1}{N!}.$$

Безусловное распределение порядковой статистики определяется соотношением:

$$f\left(\{x_{(R)}\}\right) = N! f\left(x_{(1)}, \ldots x_{(n)}\right).$$

Поэтому можно переписать

$$f\left(\{x_{(R)}\}, \{r_i\}\right) = N! \frac{1}{N!} f\left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right) = f\left(\{x_{(R)}\}\right) P\left(\{R_i\}\right).$$

Откуда следует формулировка следующей теоремы.

Теорема Гаека. При инвариантности выборки по отношению к перестановкам упорядоченная статистика и ранговый вектор статистически независимы.

Следствие 1. Для независимых выборочных значений с одинаковыми и симметричными относительно нуля распределениями случайный вектор $\{\mathrm{sign}(X_i)\}$ и $\left\{\left|X_{(R_i^+)}\right|\right\}$, $\{R_i^+\}$ тоже независимы и их распределения имеют вид:

$$P\left(\{sign\ X_i\}\right) = \left(\frac{1}{2}\right)^n,\ \mathrm{P}\left(\left|\mathrm{R}^+\right|\right) = \frac{1}{\mathrm{n}!},\ \mathrm{f}\left(\left\{\left|\mathrm{x}_{(\mathrm{R}^+)}\right|\right\}\right) = 2^{\mathrm{n}}\,\mathrm{n}! \prod_{\mathrm{R}=1}^{\mathrm{n}} \mathrm{f}\left(|\mathrm{x}|_{(\mathrm{R}^+)}\right).$$

Следствие 2. При статистической независимости двух случайных выборок **X** и **Y** их упорядоченные статистики и ранговые векторы, независимые для каждой из выборок, независимы между собой.

Информативность рангов и способ ранжировки

Важным для обеспечения информативности рангов, а также эффективности статистических процедур на их основе, является выбор способа ранжировки. Он зависит от задачи.