*Математические структуры и моделирование.*
2000, вып. 5. с.104–132.

УДК 519.24

# МАТЕМАТИЧЕСКОЕ ИССЛЕДОВАНИЕ МЕТОДИКИ ОРГАНИЗАЦИИ ОДНОГО СОЦИАЛЬНОГО МОНИТОРИНГА*

**А.К. Гуц, Р.Т. Файзуллин**

This article is publication of scientific report which was done for Omsk Center of humanitarian, social, economic and political investigations in 1996. The purpose of given work was seach of mathematical basis for «non-representative» selection of information, which could be used for social monitoring in Omsk city

## ВВЕДЕНИЕ

Для организации социального мониторинга в г.Омске необходимо получать значения наиболее важных для жизни города социально-экономических показателей. Для этого, как правило, проводятся социологические исследования, основанные на выборочных опросах населения. Опрашивается только часть жителей города. В результате опросов в распоряжение исследователя поступают так называемые *статистики*, являющиеся некоторыми производными числовыми характеристиками от выборочных данных. Например, среднее значение энергопотребления для тех жителей, которые были опрошены. Это выборочное среднее. Можно ли по нему судить о среднем значении энергопотребления всех жителей города? Такие вопросы являются стандартными для статистиков. Правильный ответ зависит от методики организации опроса, и, в частности, от того, какая часть населения была вовлечена в опрос и в какой мере эта часть населения *представляет* настроения основной массы жителей. Идея *представительности* (или «по-ученому» *репрезентативности*) – наиболее живучая в социологических исследованиях. О ней все время говорят, о ней никогда не

*Работа выполнена по заказу Центра гуманитарных, социально-экономических и политических исследований (г.Омск) в 1996 г.



E-mail: guts@univer.omsk.su, fz@univer.omsk.su
Омский государственный университет

забывают при проведении исследований и, тем не менее, она является наиболее неформализуемой на современном уровне социологической науки, и в силу этого именно с ней связаны главные провалы социологических прогнозов.

ГЭПИЦентр предложил следующую методику проведения мониторинга. Город разбит на 447 участков, которые городские власти определили из каких-то своих соображений с целью проведения выборов. Была проведена серия выборов, на основе которых определены 15 различных показателей $X^{(1)}, ..., X^{(15)}$ (активность избирателей, процент голосов за лидера и т.д.), каждый из которых имеет числовую характеристику для любого из 447 участков. Затем, опираясь на эти данные [1, 9], были отобраны пять специально *выделенных* участков (около 5000 избирателей) для проведения будущих опросов, в результате которых получаемые выборочные среднее и дисперсия будут использоваться как *оценки* соответствующих городских показателей (телефонизация, энергопотребление и др.), предлагаемые властям для реализации задач по управлению городом.

ГЭПИЦентр выделил пять участков на основе кластерного анализа данных $X^{(1)}, ...,X^{(15)}$ следующим образом. Берутся участки $i$, для которых самыми большими по модулю являются коэффициент корреляции и коэффициент близости по Cousine (?) [1]. Такими оказались участки $i = 52, 219, 281, 375, 411$.

Насколько удачен такой выбор? ГЭПИЦентр пытается показать, что статистически значимо с уровнем доверия в 95 % утверждение, что выборочное среднее для других тестов $X$ (телефонизация, энергопотребление и др.) по пяти участкам

$$\frac{1}{5}(x_{52} + x_{219} + x_{278} + x_{372} + x_{408}) = \frac{1}{447}(x_1 + ... + x_{447}), \qquad (0.1)$$

где $X^{(j)} = (x_1, ..., x_{447})$, то есть равно «городскому среднему» (см. [1]).

На основании такой оценки городского среднего ГЭПИЦентр предполагает возможным в будущем делать замеры только на указанных пяти участках и полученные статистики (оценки городского среднего) использовать в управлении городом.

По сути дела, проделанная ГЭПИЦентром работа – это: 1) формулировка идеи методики проведения мониторинга – «идея пяти участков», 2) ее апробация на некоторых социально-экономических показателях (тестах).

Возникают вопросы:

1) Насколько можно доверять тому способу разбиения на участки, который применяется городскими властями?

Хотя это дает возможность удешевлять обследования за счет того, что в распоряжении исследователей поступают регулярно обновляемые списки избирателей, составляемые властями под контролем Закона о выборах, тем не менее, сам способ разбиения может приводить к тому, что наравне с однородными по социальному составу участками имеются участки совершенно неоднородные. В какой мере это отразится на результатах мониторинга? Методика мониторинга должна гарантировать, что указанные моменты не отразятся на конечном результате.

2) Насколько корректен выбор 5 участков? Известно, что кластеризация сильно зависит от выбора метрики, лежащей в основе ее метода. Более того, сама кластеризация ГЭПИЦентра неадекватно отражает идею близости выборочного среднего к среднегородскому (см. Заключение).

3) Насколько корректно проверялась гипотеза (0.1)? ГЭПИЦентр использовал критерий Стьюдента, а он требует нормальности от $X$, которая в свою очередь подлежит проверке.

4) Каков доверительный интервал для «городского среднего», оцениваемого по выборочному среднему для пяти участков?

5) И самое главное. Будем исходить из того, что 5 участков выбраны и для 15 показателей гипотеза (0.1) верна. Делается стратифицированная выборка только в 5 указанных участках для нового «чисто экономического» 16-го показателя $X$. Будет ли гипотеза (0.1) верна для $X$? и почему?

ГЭПИЦентр предложил нам:

1) вывести и математически обосновать теоретические доверительные интервалы для статистик из отчета [1];

2) апробировать полученные доверительные интервалы на данных из отчета [1];

3) в случае неудачи в пункте 2) разобраться с методом выделения 5 участков, примененным в отчете [1].

Ниже мы решаем поставленные задачи. Но, по существу, делается большее: на основе выборочных данных ГЭПИЦентра проводится исследование жизнеспособности самой методики ГЭПИЦентра, в основе которой предположение о возможности оценивать «чисто экономические» показатели через «чисто политические», опираясь на «политизированные составляющие» этих «чисто экономических» показателей.

Данные ГЭПИЦентра, а это выборки, отражающие несколько выборных кампаний, проведенных в г.Омске, послужили для нас экспериментальной основой к проведению исследований вполне определенного среза реакций взрослого населения миллионного оборонно-промышленного города. Это не выборка, связанная с прогнозированием результатов голосования или возможных (но *не обязательно реализующихся!*) действий людей по тому или иному поводу. Это числовые данные, отражающие *действительные реакции людей в момент истины*, когда происходит само действие, а не регистрируется имитация действия, столь характерная для человека в том случае, когда он оказался в сфере внимания либо социологов, либо журналистов, либо представителей власти. Тем самым методика ГЭПИЦентра – это опора на единственно, что может быть достаточно надежным при проведении любых выборок, – на объективные данные (мы исключаем случай фальсификации результатов голосований), *представляющие* базовые экономико-политические показатели $X^{(1)}, ..., X^{(15)}$ (активность избирателей, процент голосов за лидера и т.д.). Это дает надежду, что анализ других показателей, спроецированных на «политическое пространство»,

порожденное показателями $X^{(1)}, ..., X^{(15)}$, унаследует какую-то часть «момента истины», заключенного в этих пятнадцати, и поэтому будет объективным и достоверным, но, естественно, на протяжении лишь вполне определенного времени, мера которого тем длиннее, чем стабильнее общество.

# Глава 1

# АНАЛИЗ СТАТИСТИК И ДЕТЕРМИНИСТСКОЕ ОБОСНОВАНИЕ МЕТОДИКИ

По представленным ГЭПИЦентром результатам голосований [10], точнее, по данным первичной социологической обработки результатов голосований, были проведены дополнительно к документу [1] исследования по оценке и апробации алгоритма выбора «репрезентативных» избирательных участков, то есть специально отобранных для «для организации социального мониторинга в социально-трудовой сфере».

Каждый показатель – это вектор размерности 447. Рассматривались 15 векторов $X^{(j)}$, каждый размерности $K = 447$, где 447 – это число избирательных участков города Омска. Каждый из векторов отражает определенную характеристику результатов голосования (например, активность на данных выборах, голосование за лидера и т.п.). Часто вектор $X^{(j)}$ будем называть $j$-м тестом. Совокупность векторов $X^{(j)}$, $j = 1, ..., 15$ представляют выборочные данные, которые и были подвергнуты математическому анализу.

## 1.1 Стандартная обработка данных и проверка гипотез о нормальности распределения

Для каждого вектора $X^{(j)}$, $j = 1, 2, ..., k = 15$, который в дальнейшем мы называем тестом,

$$X^{(j)} = (x_1^{(j)}, ..., x_{447}^{(j)}), \quad j = 1, 2, ..k,$$

было найдено городское среднее

$$\overline{X}^{(j)} = \frac{1}{447} \sum_{i=1}^{447} x_i^{(j)}$$

и (несмещенное) среднеквадратичное отклонение

$$\sigma_{(j)}^2 = \frac{1}{446} \sum_{i=1}^{447} (x_i^{(j)} - \overline{X}^{(j)})^2.$$

Отметим, что мы не включаем в рассмотрение численность зарегистрированных избирателей на участке, как равноправную рассматриваемым нами производным величинам от результатов голосования.

В формировании избирательных участков изначально присутствует произвол, или, иначе говоря, *формирование группы из пяти участков, проведенное ГЭПИЦентром, это один из удачных*, но *не единственный* способ агрегирования данных. Таким образом, рассмотрение численности избирателей было бы классическим примером артефакта. Конечно, плотность населения на участке или, более общо (и туманнее), степень урбанизации правомерно включать в рассмотрение, но получение подобных данных требует самостоятельных теоретических и полевых исследований.

Каждый вектор-тест $X^{(j)}$ центрировался по формуле

$$x_i^{(j)} \to x_i^{(j)} - \overline{X}^{(j)},\ i = 1, 2, ...447,$$

а затем нормировался –

$$x_i^{(j)} \to \frac{x_i^{(j)}}{((x_1^{(j)})^2 + ... + (x_{447}^{(j)})^2)^{1/2}}.$$

Далее в этой главе имеем дело с такими векторами, то есть вначале центрированными, а затем нормированными.

Возникает вопрос о характере функции распределения выборочных данных ГЭПИЦентра. В отчете [1] неявно предположено, что тесты имеют нормальное распределение или, что допустимо считать их нормальными.

Было осуществлено несколько проверок гипотезы о нормальности случайных величин $X^{(j)}$, представленных 15 выборками-тестами, которые показывают, что *все тесты не отвечают случайным величинам с нормальным распределением*. Поэтому *использование критерия Стьюдента в отчете [1] не является обоснованным*.

Данное заключение сделано на основе следующих расчетов.

Для вычисления коэффициента вариации вектора-теста $j$ или, иначе говоря, для 447 реализаций соответствующей случайной величины применялась формула [2, с.107]:

$$v_{(j)} = \frac{\sigma_{(j)}}{\overline{X}^{(j)}}.$$

Для реализации нормального закона данное отношение должно лежать в пределах интервала $0.08-0.4$, среднее значение – 0.25 [2, с.107]. Ниже приведены значения для 15 тестов:

| | | |
|---|---|---|
| 1) .2566, | 2) .1055, | 3) .4867, |
| 4) .2570, | 5) .1516, | 6) .5146, |
| 7) .2574, | 8) .2997, | 9) .3185, |
| 10) .2563, | 11) .0826, | 12) .1375, |
| 13) .5922, | 14) .3387, | 15) .5931. |

Как можно видеть, данный тест на нормальность проходят выборки с номерами: 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 14.

Вычисленные выборочные коэффициенты асимметрии и эксцесса для 15 векторов-тестов:

Коэффициент асимметрии –

$$A^{(j)} = \frac{1}{447\sigma^3_{(j)}} \sum_{i=1}^{447} (x_i^{(j)} - \overline{X}^{(j)})^3$$

1) 1.9204, 2) .9352, 3) 3.1324,
4) 1.9290, 5) 1.5321, 6) 3.9161,
7) 1.8921, 8) .7601, 9) 3.0463,
10) 1.9189, 11) .4007, 12) .4007,
13) 2.7776, 14) 1.2547, 15) 2.5607,

Эксцесс –

$$E^{(j)} = \frac{1}{447\sigma^4_{(j)}} \sum_{i=1}^{447} (x_i^{(j)} - \overline{X}^{(j)})^4$$

1) 4.4936, 2) 4.0021, 3) 16.2068,
4) 4.5195, 5) 5.9499, 6) 21.4883,
7) 4.8916, 8) .6750, 9) 16.4988,
10) 4.4902, 11) 2.7244, 12) 2.7244,
13) 7.3284, 14) 2.4295, 15) 10.2354.

Средние значения ассоциированных нормально распределенных случайных величин (в случае исходного нормального закона) равны нулю, и отклонения от среднего подчиняются так называемому закону трех сигм [2, с.108]. Среднеквадратичные отклонения для теоретического распределения асимметрии и эксцесса вычисляются по явным формулам:

$$S_A^2 = \frac{6(K-1)}{(K+1)(K+3)},$$

$$S_E^2 = \frac{24K(K-2)(K-3)}{(K-1)^2(K+3)(K+5)}.$$

Для $K = 447$ пороговые значения $3\sigma$ соответственно равны: $3S_A = 0.34563$ и $3S_E = 0.6866$. Выборочные значения (из таблиц выше) для $A^{(j)}$ и $E^{(j)}$ соответственно превышают этот порог, что *приводит нас к обоснованному выводу о неприемлемости гипотезы нормальности для всех векторов-тестов* [2].

Проверка гипотезы нормальности методом Колмогорова-Смирнова [2, с.100], также показывает, что гипотеза нормальности неприемлема. Следует отметить, что применение критерия Колмогорова-Смирнова является в данном случае даже излишним, так как более грубые методы – сравнение вариаций, асимметрии, эксцесса – позволяют гарантированно отвергнуть гипотезу

нормальности. На рисунках 1, 2, 3 приведены гистограммы, отвечающие нескольким векторам-тестам.

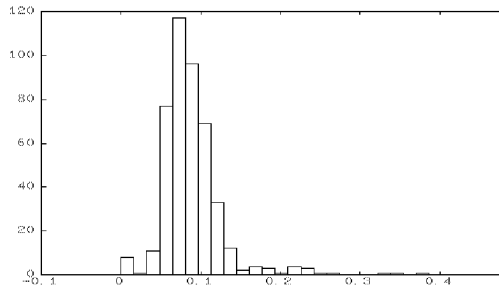


Рис 1. Гистограмма для теста 3

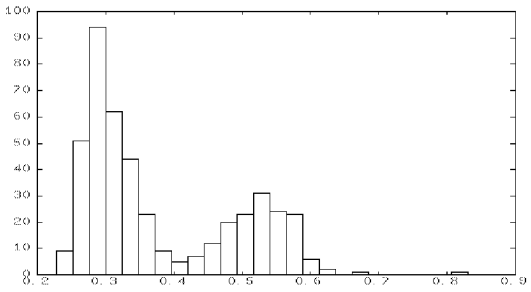


Рис 2. Гистограмма для теста 8

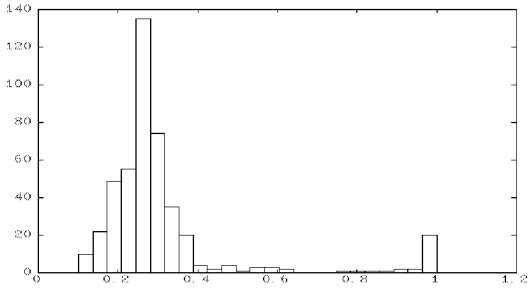


Рис 3. Гистограмма для теста 13

Высота каждого столбика на приведенных выше гистограммах равна количеству участков, показавших соответствующий процент активности в данном тесте. Координата $x$ изменяется от нуля до единицы, что отвечает диапазону от нуля до 100 процентов. Тест 3 (активность по избирательным участкам при голосовании за Конституцию) представлен гистограммой на рисунке 1, тест 8 (активность по избирательным участкам при голосовании за лидера (за сильную личность)) – на рисунке 2, тест 13 (активность на выборах 20.03.94) – на рисунке 3.

Явственно прослеживается асимметрия для тестов 3 и 13, что сразу же исключает гипотезу о нормальности законов распределения. Рассмотрение рисунка 2 (теста 8) позволяет даже сделать предположение о двух независимых законах, которым подчинены две группы участков, причем в силу наблюдаемой асимметрии эти законы также далеки от нормального.

## 1.2 Вывод об избыточности данных

По таблице полученных векторов была определена корреляционная матрица $\|k_{(st)}\|$ для системы из 15 векторов, как матрица взаимных скалярных произведений нормированных векторов-тестов.

Ниже выписаны абсолютные значения элементов данной матрицы $\|k_{(st)}\|$ по строкам:

1) 1.0 .12 .36 1.0 .20 .49 .94 .00 .37 1.0 .00 .00 .83 .04 .38
2) .12 1.0 .56 .12 .13 .50 .06 .17 .51 .12 .09 .09 .21 .05 .26
3) .36 .56 1.0 .36 .24 .75 .25 .08 .81 .36 .24 .24 .45 .04 .47
4) 1.0 .12 .36 1.0 .20 .49 .94 .00 .37 1.0 .01 .01 .83 .04 .38
5) .20 .13 .24 .20 1.0 .24 .15 .21 .21 .20 .21 .21 .18 .09 .13
6) .49 .50 .75 .49 .24 1.0 .37 .10 .76 .49 .25 .25 .61 .04 .57
7) .94 .06 .25 .94 .15 .37 1.0 .01 .26 .94 .03 .03 .75 .03 .30
8) .00 .17 .08 .00 .21 .10 .01 1.0 .06 .00 .02 .02 .06 .03 .10
9) .37 .51 .81 .37 .21 .76 .26 .06 1.0 .37 .20 .20 .47 .07 .58
10) 1.0 .12 .36 1.0 .20 .49 .94 .00 .37 1.0 .00 .00 .83 .04 .38
11) .00 .09 .24 .01 .21 .25 .03 .02 .20 .00 1.0 1.0 .08 .03 .13
12) .00 .09 .24 .01 .21 .25 .03 .02 .20 .00 1.0 1.0 .08 .03 .13
13) .83 .21 .45 .83 .18 .61 .75 .06 .47 .83 .08 .08 1.0 .11 .46
14) .04 .05 .04 .04 .09 .04 .03 .03 .07 .04 .03 .03 .11 1.0 .31
15) .38 .26 .47 .38 .13 .57 .30 .10 .58 .38 .13 .13 .46 .31 1.0

Можно сделать вывод об избыточности рассматриваемой системы векторов-тестов, так как оказалось, что группы векторов с номерами $1, 4, 7, 10$ и $11, 12$ попарно имеют коэффициенты корреляции Пирсона, сравнимые с единицей.

Рассмотрение собственных чисел корреляционной матрицы, при исключенных векторах, например $4, 7, 10$ и $11$, показало, что оставшаяся система векторов в совокупности линейно независима, минимальное по модулю собственное число $\lambda_m$ симметричной матрицы корреляции

$$\left(\|k_{(st)}\|_{s,t=1}^{m} - \lambda_m \|\delta_{(st)}\|_{s,t=1}^{m}\right) U = 0,\ m = 1, ..., k = 11$$

монотонно убывает при увеличении размерности $m$ миноров $\|k_{(st)}\|_{s,t=1}^{m}$ корреляционной матрицы, и в итоге имеет порядок $10^{-2}$ (и равно 0.05).

Для определения собственных чисел матриц миноров применялся численный метод Якоби-Эберляйн, численно устойчивый в случае возможных совпадений собственных чисел [3, с.355].

Следует отчетливо понимать, что имеющаяся у нас матрица и ее миноры – всего лишь один из возможных исходов для экспериментов, подобных проведенному ГЭПИЩентром. Что подразумевается под этими словами? Обратим внимание на хорошо коррелированные векторы 1, 4, 7, 10. Малые различия, которые присутствуют в компонентах данных векторов, обусловлены различным способом агрегирования данных (рассмотрением активности на выборах, активности за лидера и.т.д.). Очень хорошая корреляция данных векторов-тестов

означает не что иное, как то, что в действительности рассматривается лишь одна значимая случайная величина – активность избирателей на выборах. Таким образом, нам принципиально безразлично, какой именно вектор из набора 1, 4, 7, 10 векторов-тестов следует подставлять в нашу матрицу. Очевидно, что варьироваться подобным образом могут и другие векторы, причем следует отметить, что вариации возможны не только в результате артефакта эксперимента, как это произошло с векторами 1,4,7,10, но и в зависимости от реальных внешних причин, происходящих во времени и влияющих на политическую реакцию населения, – не следует забывать, что мы рассматриваем политический процесс по времени, а не какую-то абстрактную реализацию некой случайной величины.

Поэтому рассматриваемая матрица является лишь представителем целого класса матриц, которые отличаются друг от друга на величины, зависящие от вариаций векторов-тестов, которые в свою очередь определяются способом агрегирования данных или другими случайными факторами. Если варьируя векторы-тесты мы можем получить вырожденную матрицу, то с некоторой вероятностью случайные величины, реализацией которых являются векторы, будут в совокупности линейно зависимы. Если же произвольно выбранный способ агрегирования дает нам матрицу, для которой минимальное по модулю (вообще-то они все положительны) собственное число существенно меньше возможных вариаций векторов-тестов, то это означает, что соответствующие случайные величины зависимы с большой степенью вероятности.

В нашем конкретном случае это означает, что число векторов-тестов, основанных на обработке результатов голосования, может быть выбранным не более, чем $8-9$, причем предпочтительнее выбор 8. Вариации норм векторов 1, 4, 7 достигают величины порядка сотых, и мы имеем минимальное собственное число (по модулю) такого же порядка для корреляционной матрицы порядка 8 – 9.

Поэтому следует отметить, что полученные в результате полевых исследований данные, конечно, не являются бесполезными, так как они позволяют экспериментально определить степень устойчивости данных от способа агрегирования и, что очень важно, число независимых параметров, которые определяют срез реакции населения.

## 1.3 Кластеризация и пространственная близость

Было проведено упорядочение номеров $(i)$ избирательных участков по степени близости соответствующих выборочных значений $x_i^{(j)}$ тестов $X^{(j)}$ на $i$-ом участке к значениям городских средних $\overline{X}^{(j)}$ (относительно среднеквадратичного отклонения $\sigma_{(j)}$) одновременно для всех 15 тестов. Для этого рассматри-

вался вектор

$$\Delta_i = (\frac{\mid x_i^{(1)} - \overline{X}^{(1)} \mid}{\sigma_{(1)}}, ..., \frac{\mid x_i^{(15)} - \overline{X}^{(15)} \mid}{\sigma_{(15)}}) \tag{1.1}$$

и вычислялась его норма $\|\Delta_i\|$.

Оказалось, что вне зависимости от выбора конкретной меры близости, как то малости эвклидовой нормы вектора $\Delta_i$, или нормы-максимум, или нормы по подмножествам индексов $j$, упорядочение участков устойчиво. Иначе говоря, первые десять номеров остаются среди первых пятнадцати, первые тридцать номеров остаются среди первых 50 и аналогично первая сотня номеров остается среди первых 150, вне зависимости от выбора нормы близости. Ниже приведены списки первых тридцати номеров участков при упорядочении по эвклидовой (1) и норме-максимум (2), дающих наиболее близкие значения тестов к своим городским средним:

(1)

1845, 1701, 1672, 1837, 1330, 1595, 1662, 1668,

1770, 1604, 1850, 1347, 1847, 1862, 1846, 1608,

1808, 1666, 1350, 1677, 1859, 1802, 1265, 1811,

1589, 1775, 1844, 1720, 1778, 1644

(2)

1845, 1701, 1662, 1770, 1604, 1595, 1787, 1608,

1808, 1668, 1837, 1835, 1847, 1350, 1330, 1872,

1666, 1347, 1704, 1775, 1613, 1802, 1619, 1589,

1778, 1850, 1720, 1651, 1865, 1846

Конечно, хорошо известен факт эквивалентности норм в конечномерном случае, но следует отметить, что данный математический факт еще не гарантирует выявленной устойчивости номеров, точнее говоря, не предсказывает степени устойчивости.

Более того, было выявлено, что наблюдается пространственная консолидация избирательных участков с соответствующими номерами, которые на карте города образуют вытянутые конфигурации, а не хаотические пятна. Данный факт содержит в себе большее, чем просто очевидную близость номеров.

Это служит экспериментальным доказательством существования того, что на языке статистики называется совместной функцией распределения, причем данная функция, по всей видимости, является гладкой по пространственным переменным.

Отметим, что факт пространственной консолидации отчетливо наблюдается и в случае независимого исследования, проведенного в г. Надыме, несмотря на то, что в Надыме имеется всего 40 избирательных участков.

На рисунке 4 показано расположение первых 30 избирательных участков, наилучших по степени близости к средним значениям по городу Омску [11].

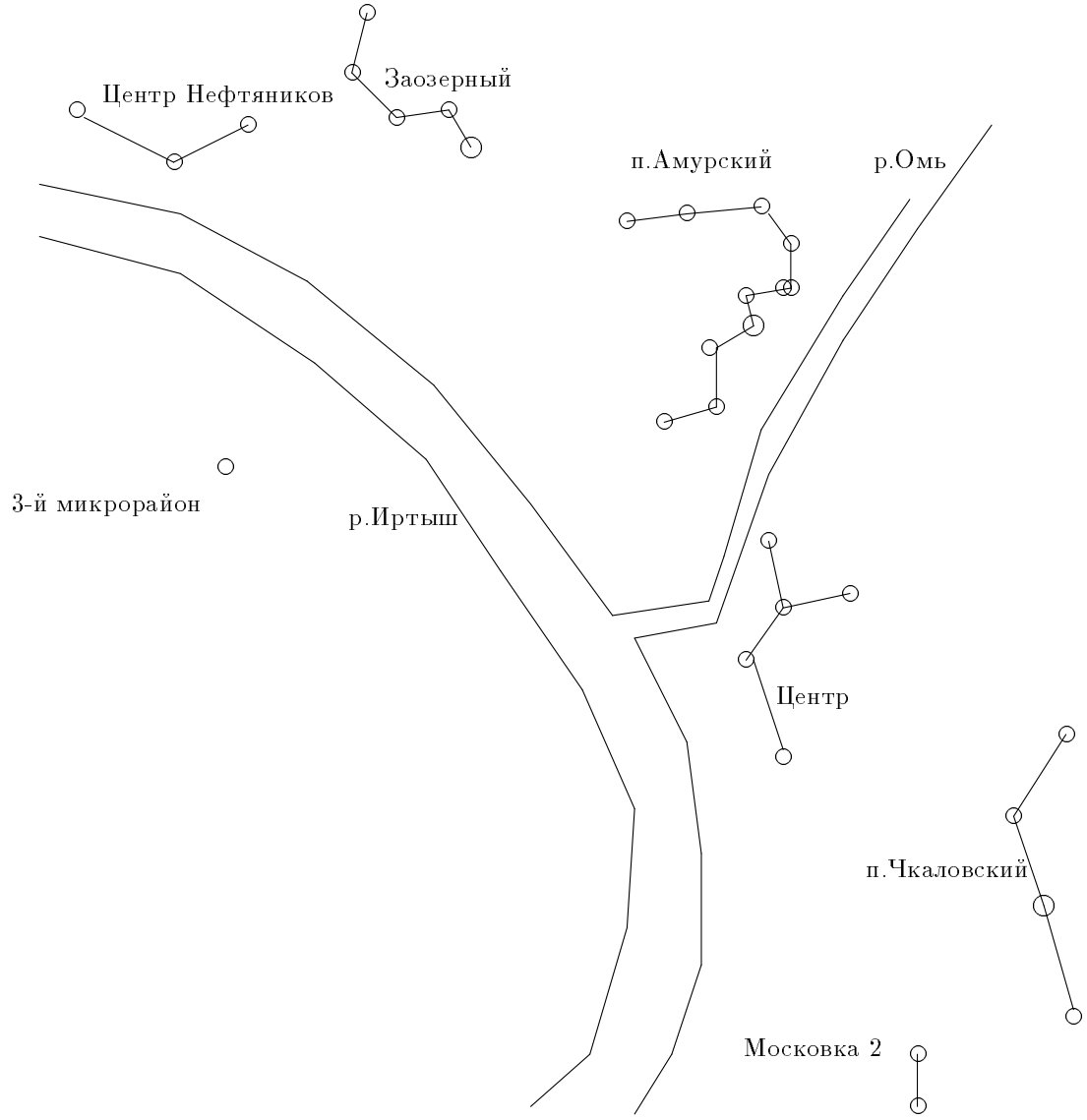


Рис 4. Схематичное территориальное распределение участков, показавших наиболее близкий к среднему по городу социологический срез (по результатам голосований)

Следует отметить, что ранее *выбранные ГЭПИЦентром с помощью алгоритма кластеризации [1] участки 52(1259), 219(1691), 281(1714), 375(1808),*

*411(1844), являются приемлемыми и с точки зрения предложенного здесь упорядочения, а также представительными для выявленных конфигураций. Они, за исключением 219 избирательного участка, принадлежат первой сотне номеров (но и 219 имеет номер 111).*

Кроме того, данные участки являются приемлемыми представителями зон со средним политическим откликом, равномерно представляющим город по районам. Конечно, следует отметить, что у ГЭПИЦентра в его отчете [1] выпадает важная зона, непосредственно лежащая в центре города. Эта зона учтена в предлагаемой нами кластеризации. Поэтому участки, выбранные ГЭПИЦентром, можно дополнить участками из нашей кластеризации, например, в центре города.

Отметим, что не вполне оптимальный выбор ГЭПИЦентром участков был обусловлен абстрактным алгоритмом кластеризации, не предполагавшим физической подоплеки кластеризации, как то пространственная консолидация участков со средним откликом.

Возвращаясь к пространственной консолидации, обратим внимание на то, что на самом деле можно построить линии уровня, как отклонения от среднего отклика, по всем рассматриваемым величинам. Более того, по каждой величине (активность, голосование за лидера и т.п.) можно построить аналогичные картины линий уровня.

Данная информация может оказаться полезной с точки зрения организации адекватной избирательной кампании и, как следствие, позволит учесть желания людей, проживающих на данной территории.

## 1.4 Приближения среднего по 5 выделенным участкам

Из набора рекордных по степени близости к средним значениям $\overline{X}^{(j)}$ номеров избирательных участков нами выбирались случайным образом 5 номеров, и по ним вычислялось среднее по городу для каждого теста и по всем тестам сразу. Оказалось, что для любых 5 номеров из первых тридцати, отклонение $w$ (или $ww$) от среднего по тестам (относительно дисперсии по тесту) лежит в пределах $0.05 - 0.2$, где

$$w^{(j)} = \frac{1}{\sigma^{(j)}} \mid \overline{X}^{(j)} - \frac{1}{5}(x_{i_1}^{(j)} + ... + x_{i_5}^{(j)}) \mid,$$

$$w = \frac{1}{15}(w^{(1)} + .... + w^{(15)}),$$

и

$$ww = \sqrt{\frac{1}{15}(w^{(1)})^2 + .... + (w^{(15)})^2)}.$$

Для наихудшего возможного случая – равномерного закона, реализацией которого может быть соответствующий тест, это дает отклонение порядка 0.05 размаха выборки.

Также отметим, что участки $52, 219, 281, 375, 411$, исследуемые ГЭПИЦентром [1], являются приемлемым набором ($w$ и $ww$ порядка 0.18). Но следует отметить, что первые участки сохраняют свою приемлемость и при формировании наборов по $2, 3, 4$ участка даже при взвешенном представительстве, тогда как выделенные ГЭПИЦентром участки теряют свои привлекательные качества.

Несмотря на это, выделенные ГЭПИЦентром участки представляют среднее, что все-таки лучше, чем произвольные пять из 447 для произвольных $w$ и $ww$ порядка $0.35 - 0.5$.

## 1.5 Принципиальная возможность экстраполяции

Основываясь на полученных результатах, правомерно попытаться экстраполировать выводы о представительности территорий выбранных избирательных участков для оценки средних значений по городу и других определяющих параметров. В сущности, рассмотренные ранее векторы-тесты – это производные величины от чисто физических параметров, описывающих условия проживания, и от параметров, характеризующих степень осознания условий проживания.

Рассмотрим функцию $H(i)$, определенную на плоском графе, вершинами которого являются избирательные участки, а ребра определяют преимущественные пути перераспределения некоторых ресурсов между ареалами, коррелированными с вершинами графа. Это вполне допустимая аппроксимация для представления функции двух переменных, определенной на некоторой плоской области, которая в свою очередь представляет населенный пункт.

Под значениями функции $H$ мы будем понимать «функцию уровня жизни» прилегающей к участку территории. Следует отметить, что под такой функцией мы понимаем не общеупотребительное и достаточно смутное социологическое содержимое, а строго конкретное функциональное определение, о значимости которого можно дискутировать, но которое, по крайней мере, можно применять в вычислениях, и значимость которого можно проверять.

Выбор функции $H$ неоднозначен, одним из возможных вариантов может быть задание значения в вершине $i$ величины:

$$H(i) = \alpha_{i1}^2 + \alpha_{i2}^+ .... + \alpha_{in}^2 + ...,$$

где двойной индекс означает соответственно номер участка и номер ресурса, определяющего, в определенной мере, уровень жизни.

Предполагается, что проводятся локальные усреднения по подобластям и значения средних считаются значениями $H$ на вершинах графа. Более естественным будет рассматривать не саму функцию, а ее отклонения от среднего значения $V = H - \overline{H}$, причем считать отклонения нормированными. Статистическим аналогом введенной в рассмотрении функции $V$ будет двумерно распределенная случайная величина с нулевым математическим ожиданием и еди-

ничным среднеквадратичным отклонением. Отметим, что и далее мы будем придерживаться детерминистского подхода, не забывая при этом проводить статистические аналогии.

Естественно предполагать нормированность ресурсов, то есть, что сумма квадратов ресурса $k$ по всем вершинам равна единице.

Приведенные выше слова «в какой то мере» определяют неоднозначность функции $H$. Вернее была бы запись вида:

$$H(i) = Z_1\alpha_{i1}^2 + Z_2\alpha_{i2}^+.... + Z_n alpha_{in}^2 + ...,$$

где коэффициенты $Z_k$ определяют значимость ресурса для конкретной группы населения. Например, степень обеспеченности углем очень значима для населения, проживающего в частном секторе, и наоборот, домашние телефоны представляются чем-то нереальным и не относящимся к непосредственным нуждам этой группы населения.

Конечно, данные коэффициенты не являются предметом дискуссии, они должны быть определены экспериментально. Кроме того, $H$ не является наблюдаемой физической величиной; на самом деле мы имеем дело с реакциями населения, в частности, с рассмотренными ранее результатами выборов, которые и служат тестами на значения функции $H$.

Более строго рассматриваются $l$ функций $f_m(i)$, $m = 1, 2, ..l$, которые, кроме того, что они являются функциями от вершины, являются нелинейными функциями от $H(i)$. На языке статистики можно сказать, что случайная величина $f_m$ коррелирована со случайной величиной $H$. Как видим, в статистическом подходе опускается индекс $i$, что, по существу, сводит пространственную функциональную зависимость к реализациям случайной величины, что также не вполне достаточно.

Рассмотрим зависимость:

$$f_m(i) = a_m + b_m V(i) + c_m V^3(i) + ... + g_m(i),$$

где $a_m, b_m, c_m, ...$ – коэффициенты ряда функции $f_m$ по функции $H$, а $g(i)$ некоторая функция, не зависящая от $V$.

Рассмотрим среднее от функции $f_m$, иначе говоря, возьмем приближение интеграла по всем вершинам графа. Первое слагаемое даст нам $a_m$, второе не будет присутствовать, третье и далее также не привнесут вклада. Среднее от функции $g_m(i)$ по всем вершинам равно нулю, в противном случае среднее относится на счет константы $a$ и переопределенная функция $g$ будет иметь уже нулевое среднее.

Очевидно, что если $f_m$ существенно зависит от $V$, то среднеквадратичное функции $g_m$ должно быть существенно меньше, чем среднеквадратичное слагаемых, содержащих степени $V$. В противном случае предыдущая запись была бы просто формальным представлением функции.

В случае существенной зависимости, при пространственном представлении функций $f_m(i)$, узлы, в которых $V$ достигает нуля, должны представлять связную совокупность и характеризоваться лишь осцилляциями функции $g_m$, малыми по сравнению с среднеквадратичным отклонением $V$.

Как видно из предыдущего, консолидация избирательных участков по признаку среднего отклика дает нам именно такие связные конфигурации, причем конфигурации совпадают для попарно независимых векторов-тестов.

Отсюда следует, что дополнительная функция, столь же зависимая от $V$, как и те, которые мы рассматривали ранее, должна давать средний отклик на тех же консолидированных участках, как и рассмотренные ранее функции.

Ответ на вопрос, является ли рассматриваемая функция $f_{l+1}$ столь же зависимой от $V$, как и предыдущие векторы-тесты, представляет в общем случае нетривиальную задачу. Но для некоторых функций эту зависимость можно сделать обоснованней и прийти к выводу о ее (по крайней мере, не худшей) зависимости от $V$. Например, степень телефонизации населения не зависит от политических реакций того же населения на данный момент времени, но, по всей видимости, обратная зависимость есть. Можно сказать, что одна из этих величин – определяющая, а вторая – производная. Амплитуда колебаний функции $g$ может лишь увеличиться при переходе от определяющих к производным величинам.

Таким образом, отклонения от среднего по городу для политических реакций населения мажорируют отклонения от среднего для определяющих величин.

## 1.6 Ошибки оценки среднего для произвольного вектора-теста

Основываясь на материалах предыдущих параграфов, определим ошибку оценки городского среднего в некотором смысле произвольного 16-го теста $Y$ выборочными средними для конечного числа участков.

Рассмотрим вектор-тест $Y^l$, полученный в результате линейной комбинации известных нам 15 векторов-тестов:

$$Y^l = \gamma_1 X^{(1)} + ..... + \gamma_1 5 X^{(15)},$$

где коэффициенты $\gamma_j$ представляют собой компоненты нормированного вектора, выбранные с помощью генератора случайных чисел (равномерный закон), число $l$ определяет количество генерируемых векторов $Y$. Мы сформировали таким образом 500 векторов; каждый из них – это выборка-тест, который мог бы быть измерен по всем 447 участкам.

Другими словами, мы как бы провели замеры 500 новых показателей (тестов), «порожденных» исходными 15 тестами ГЭПИЦентра. Отметим, что, рассматривая произвольную линейную комбинацию заданных векторов-тестов, мы просто следуем предположению о том, что *любая реакция населения или базовая характеристика обеспечения населения взаимосвязана с факторами, определяющими политические и некоторые другие реакции населения, учтенные ГЭПИЦентром в тестах* $X^{(1)}, ..., X^{(15)}$.

Насколько обоснованно такое предположение? Допустим, что мы упустили один из важных факторов при обработке тестов $X^{(1)}, ..., X^{(15)}$ и он проявится

в некотором другом важном тесте, не связанном с рассмотренными ГЭПИЦентром, но зависящим от базовых характеристик проживания.

Какова вероятность такого события? Вспомним, что обработка результатов голосования показала избыточность 15 тестов, вполне хватает и 8-9. То же показывает обработка результатов голосования в г.Надыме, где рассматривались уже 22 теста. Значит, стараясь отразить независимые характеристики результатов голосования, мы выделили 9 определяющих факторов и не смогли, как оказалось, в дополнительных 6 (г.Омск) и 13 (г.Надым) тестах выделить 10-й важный фактор. Рассматривая определение важного фактора как независимое испытание (по получаемому в полевых работах вектору-тесту можно определить определяется новый или старый фактор), мы получим вероятность, что в дополнительных тестах 10-й фактор не выявился с вероятностью $P_1 = (9/10)^6$ для г. Омска и с вероятностью $P_2 = (9/10)^1 3$ для Надыма. Кроме того, количество факторов (9) одно и то же для Омска и Надыма, а это означает, что необходимо рассматривать вероятность неучета 10-го фактора порядка произведения вероятностей $P_1$ и $P_2$, что в итоге дает число порядка 0.1.

С точки зрения линейной алгебры, предположение об учете всех определяющих параметров означает, что результат любого аналогичного измерения (после операций центрирования и нормировки) будет точно представляться как линейная комбинация наших 15 векторов $X^{(1)}, ..., X^{(15)}$.

Мы формировали выборки, давшие 500 векторов $Y^l$. Для каждого $Y^l$ вычислялась ошибка $\epsilon_k$:

$$\epsilon_l = \frac{\sigma_l}{\overline{Y}^l - p_m},$$

где $\sigma_l$ – среднеквадратичное отклонение для вектора-теста $Y^l$; $\overline{Y}^l$ – среднее по тесту, а $p_m$ – выборочное среднее по $m$ выделенным участкам.

Оказалось, что для рассматриваемого ГЭПИЦентром числа участков $m = 5$, ошибка $\mid \epsilon_l \mid$ не превышает 0.5. Средняя ошибка:

$$\sum_l \epsilon_l = 0.$$

Выбор наилучших участков принципиально картину не меняет.

Можно утверждать, что с вероятностью, большей, чем 0.95, ошибка оценки городского среднего для, в общем-то призвольно взятого, нового теста $Y^l$ выборочным средним по 5 выделенным ГЭПИЦентром участкам не превысит 40% среднеквадратичного отклонения $\sigma_l$ по городу. С вероятностью 0.9 можно гарантировать точность 0.35.

Рассмотрим другие возможные меры ошибки, эквивалентные оценке через среднеквадратичное. В наихудшем случае – случае равномерного распределения, максимальная ошибка определения среднего будет равна примерно 0.2 размаха выборки. В случае закона распределения близкого к нормальному максимальная ошибка будет ограничена интервалом 0.0032-0.16 значения среднего. Это следует из оценки вариации, использованной нами ранее.

Но мы можем усреднять и по меньшему числу участков, чем 5, например, из имеющихся у нас участков организовать выборки по 2, 3, 4 участка и оценить вариацию среднего при различных усреднениях. Отметим, что оценки среднего по выборкам из 2 и 3 участков не коррелируют с оценками среднего по 5 участкам.

Если данная вариация велика, то можно сделать обоснованный вывод, что усреднение по данным 5 участкам не совсем удовлетворительно. Данная процедура хорошо известна под названием – метод складного ножа [5, с.10] Таким образом, отсеиваются векторы $Y$ с ошибкой оценки среднего выше, чем $0.1 - 0.15\sigma$.

Также, если известно, что корреляция вектора-теста $Y$ с некоторым из данных 15 векторов-тестов значима (т.е больше 0.7), то можно гарантировать с вероятностью 0.95, что ошибка оценивания среднего лежит в интервале от нуля до 0.15 среднеквадратичного по тесту $Y$. На рисунках 5, 6, 7 приведены гистограммы распределения ошибки $\epsilon$, для усреднения по 5 выделенным участкам, для усреднения по 12 участкам, для случая значимой корреляции вектора $Y$ (0.7) с одним из заданных векторов.

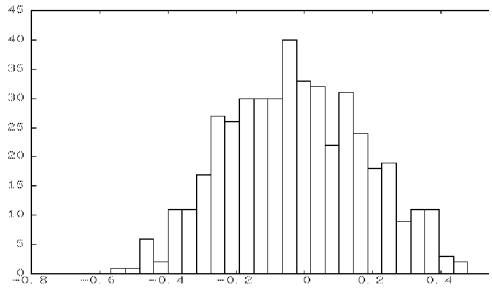


Рис 5. Гистограмма для ошибки $\epsilon_l$ для участка 5

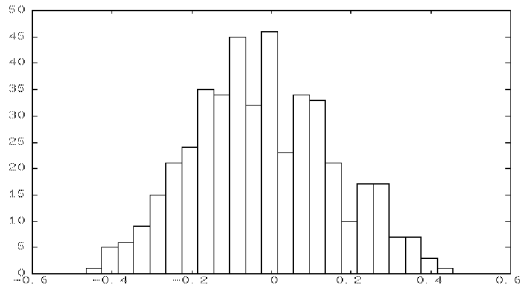


Рис 6. Гистограмма для ошибки $\epsilon_l$ для участка 12

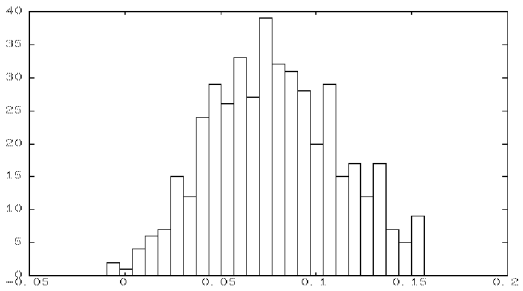


Рис 7. Гистограмма для ошибки $\epsilon_l$ в случае значимой корреляции с 1 тестом

Вышеприведенные оценки вероятности сделаны на основе численного эксперимента, диапазон изменения ошибки определяется подобными гистограммами.

# Глава 2

# СТАТИСТИЧЕСКОЕ ОБОСНОВАНИЕ МЕТОДИКИ И ВЫВОД ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

## 2.1 Постановка задачи

### 2.1.1. Задача оценки средних городских показателей

Жители города разбиты по $K$ участкам. Проводятся измерения некоторых жизненно-важных величин для этих участков. Например, число проголосовавших на участке за лидера, деленное на число голосовавших избирателей участка, число телефонов на участке, деленное на число избирателей, и т.д. В принципе, измеряемое число может принимать любое (вещественное) значение. Естественно считать, что интересующая нас величина измеряется с ошибкой, так как на процесс измерения влияют самые разнообразные *случайные* факторы. Более того, сама измеряемая величина как явление общественной жизни является случайной, в том смысле, что нельзя ее считать досконально и точно известной даже для соответствующей административной службы: избиратели уезжают или умирают без предупреждения, телефоны ставятся и снимаются более или менее регулярно, и это не всегда своевременно отражается в документах телефонных служб.

Как оценить среднее значение интересующей нас величины (теста) $X$, называемой средним по городу, и ее возможное отклонение, если использовать знание значений этой величины только на некоторых $n$ участках, с какой-то долей уверенности, что представленные нам цифры отражают реальную ситуацию? Отметим специально, что для ГЭПИЦентра особый интерес представляет либо $n = 5$, либо $n$ порядка 10.

Эта задача является традиционной для статистики.

Для ее решения используем методы *математической* статистики, которая в отличие от статистики предъявляет, во-первых, строго доказанные оценки, а, во-вторых, предупреждает в каких рамках их следует применять. Вольности в интерпретации чреваты ошибками при принятии решений.

Будем ниже считать, что участки с номерами $i = 1, ..., 5$ – это специально выделенные ГЭПИЦентром.

### 2.1.2. Математическая формулировка задачи

Итак, для каждого участка задано значение (реализация) $x_i$ случайной величины $X_i$, $i = 1, ..., K$. Все $X_i$ независимы и имеют одно и то же распределение

$F(x)$. Пусть существуют среднее $\mu = \mathbf{M}X_i$ и дисперсия $\sigma^2(F) = \mathbf{M}(X - \mu(F))^2$. По сути дела, мы имеем одну случайную величину $X$ с распределением $F(x)$ и ее реализации $x_1, ..., x_n$ (выборка).

Требуется найти интервальные оценки для $\mu(F)$ и $\sigma^2(F)$.

Пусть

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X}_n)^2$$

– выборочные среднее и (смещенная) дисперсия.

Предположим, что величина $X$ положительна и ограничена $X \leq \tau$ (тогда $\sigma(F) \leq \tau/2$). Такое предположение, если вспомнить о практической подоплеке нашей задачи, вполне разумно и естественно. Но тогда справедливо неравенство [6, с.52]

$$\mathbf{P}_{F_n}(\mid \overline{X}_n - \mu(F) \mid \leq \delta) \geq \gamma, \tag{2.1}$$

где

$$\gamma = 1 - nexp\left(\frac{\delta^2 n}{4\tau^2}\right)$$

и $F_n$ – функция распределения выборочного среднего $\overline{X}_n$, которое рассматривается как случайная величина. То есть с уровнем доверия (значимости) $\gamma$ или, иначе говоря, в $\gamma 100$ % из 100 неизвестное нам среднее $\mu(F)$ будет лежать в интервале

$$(\overline{X}_n - \delta,\ \overline{X}_n + \delta).$$

Для использования формулы (2.1) необходимо знать функцию $F_n(t)$, и это главное препятствие на пути к успеху в решении задачи.

По центральной предельной теореме при $\mathbf{M}X^2 < \infty$

$$\lim_{n\to\infty} \mathbf{P}_{\overline{X}_n}\left(\sqrt{n}\frac{\overline{X}_n - \mu(F)}{\sigma(F)} \leq t\right) = \Phi(t).$$

Значит, при достаточно большом $n$ можно считать, что выборочное среднее $\overline{X}_n$ имеет нормальное $F_n = \mathcal{N}(\mu, \sigma^2(F)/n)$ распределение. Но как раз $\sigma^2(F)$ неизвестно, поэтому неизвестна и $F_n$. Более того, замена $F_n$ на $\mathcal{N}(\mu, \sigma^2(F)/n)$ возможна с ошибкой порядка $1/\sqrt{n}$; в лучшем случае – порядка $1/\sqrt{n^3}$ [7, с.276-277]. В условиях, предлагаемых ГЭПИЦентром $n = 5$ ошибка лежит в пределах $0.09 - 0.45$, то есть составляет в худшем, и наиболее вероятном случае, 45 %. Это много! Напомним, что все 15 тестов не имеют нормального распределения.

Поэтому нужно как-то обосновать гипотезу ГЭПИЦентра (0.1) без предположения о нормальности тестов.

## 2.2 Доверительный интервал для городского среднего в случае произвольного распределения $F$ с $\mathrm{M}X^2 < \infty$

### 2.2.1. Непараметрический метод Чоу-Роббинса

Найти доверительный интервал для $\mu(F)$, не зная ее функции распределения, безнадежно. Гарантированная оценка среднего $\mu(F)$ без какой-либо априорной информации о $F$ невозможна. Например, такая ситуация может случиться при «больших выбросах» величины $X$ на отдельных участках [6, с.49-50]. Изложим метод Чоу-Роббинса получения доверительного интервала для среднего «фиксированной ширины» [8, §10.10].

Предположим, что $F$ имеет конечный второй момент. Зададим числа $\delta$ и $\gamma$ и последовательность $a_n$, сходящуюся к $a$, где $\Phi(a) = (1+\gamma)/2$.

Найдем число

$$n(\delta) = \min_n \left\{ n \le \frac{a_n^2}{\delta^2} \left( \frac{1}{n} + s_n^2 \right) \right\}$$

– число остановки.

Тогда

$$\lim_{\delta \to 0} \mathbf{P}_{F_n} ( \mid \overline{X}_{n(\delta)} - \mu(F) \mid \le \delta ) \ge \gamma.$$

### 2.2.2. Апробация метода Чоу-Роббинса на статистиках ГЭПИЦентра

Имеются данные $x_1, ..., x_K$ по $(K = 447)$ участкам. Посмотрим чему окажется равным число остановки $n(\delta)$ для различных $\delta$ при уровне доверия $\gamma = 0.95$.

Применим следующий алгоритм. Берем $\delta_k = 1/k$, $(k = 1, 2, ...)$ и $a_n = a = 1.96 = \Phi^{-1}(0.975)$. Для данного $k$ находим все такие $n$, которые не меньше, чем $4k^2(1/n + s_n^2)$. Из них берем самое маленькое. Это и есть $n(\delta_k)$.

В таком случае с уровнем доверия 95% городское среднее лежит в интервале

$$(\overline{X}_{n(\delta_k)} - 1/k,\ \overline{X}_{n(\delta_k)} + 1/k)$$

при одном, но очень суровом условии, – число $k$ достаточно большое.

Поскольку [8, теорема 10.10.2]

$$\lim_{k \to \infty} \frac{n_(\delta_k)}{a^2 k^2 \sigma^2(F)} = 1,$$

то следует, задавая последовательно $k = 1, 2, ...$, находить число остановки $n_(\delta_k)$. При малой дисперсии $\sigma^2(F)$ есть шанс, что оно при очень больших $k$ будет лежать еще в пределах $K$. Более того, если обнаруженное ГЭПИЦентром для пятнадцати показателей (тестов) $X$ экспериментальное наблюдение

$$\overline{X}_5 \approx \mu(F) \tag{2.2}$$

отражает объективную социальную закономерность, то следует ожидать, что число остановки будет еще лежать в пределах от 5 до 10 при очень больших $k$. Рост $k$ – это сужение доверительных границ, то есть все более точная оценка городского среднего $\mu(F)$. Отметим, что формула дает оценку среднеквадратичного отклонения

$$\sigma(F) \sim \sqrt{\frac{n_(\delta_k)}{a^2 k^2}} \quad npuk \to \infty.$$

Сказанное выше означает следующее:

1) если для всех статистик ГЭПИЦентра при нарастании $k$ число остановки будет достаточно долго лежать в пределах от 5 до 10, то выводы ГЭПИЦентра относительно наблюдения (2.2) небезосновательны;

2) если 1) не верно, то есть число остановок быстро превысит разумные пределы (более 20 (?) участков), то усилия, предпринятые ГЭПИЦентром по поиску облегчения и удешевления полевых работ, малоосновательны;

Большего трудно достичь, поскольку нам почти ничего не известно о распределении $F$.

Указанный алгоритм легко реализуется на компьютере.

Ниже даются кривые, характеризующие поведение доверительного интервала в зависимости от числа используемых участков (ось $X$). На рис. 8-12 горизонтальная прямая – это городское среднее; кривая, «зажатая» между двумя другими похожими кривыми, – кривая выборочных средних.

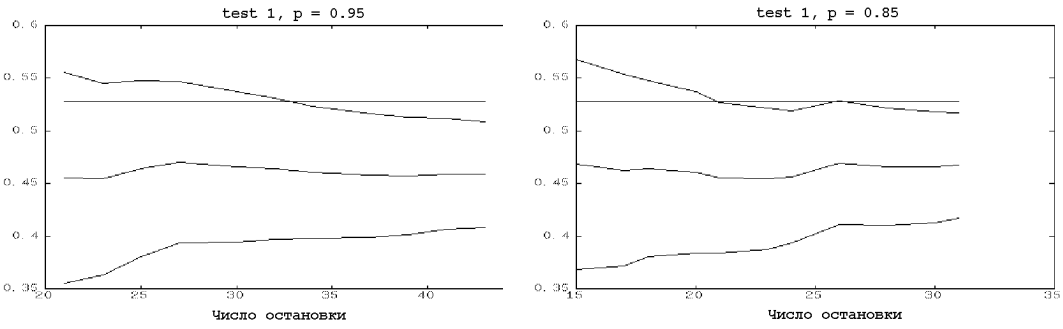


Рис 8. Доверительный интервал для теста 1

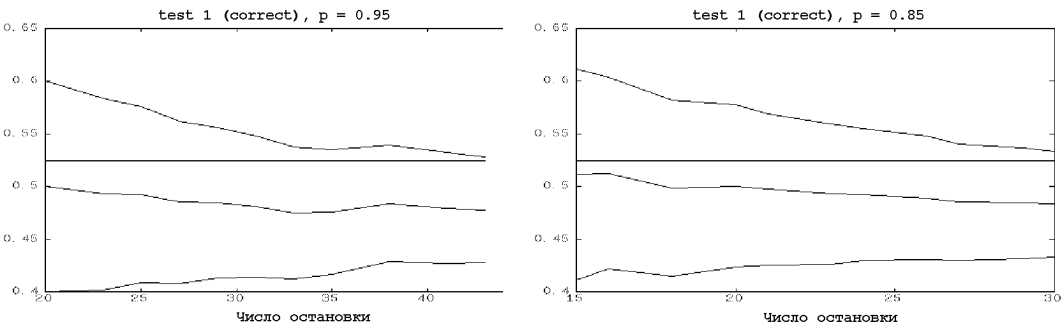


Рис 9. Доверительный интервал для теста 1 с учетом выделенных участков

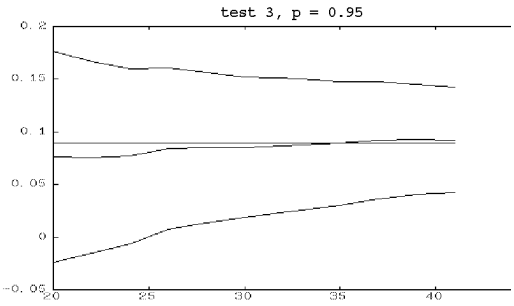


Рис 10. Доверительный интервал для теста 3

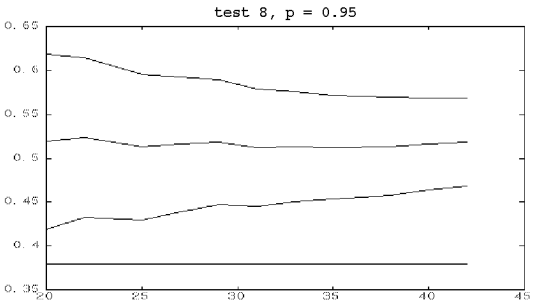


Рис 11. Доверительный интервал для теста 8

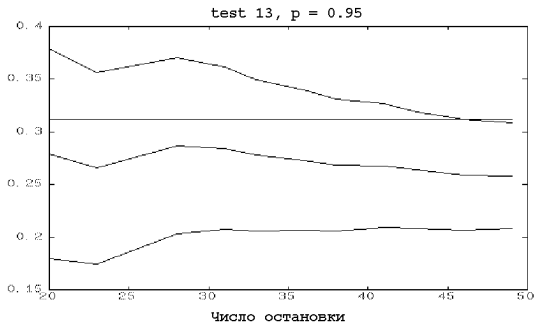


Рис 12. Доверительный интервал для теста 13

На рис.8, 9 видно, как влияют специально выбираемые участки на характеризацию городского среднего посредством выборочных средних. В целом, для описания городского среднего хватает участков в количестве, на порядок меньшем, чем общее число участков 447 (тем не менее, для теста 8 не всегда хватает и 45 участков. – см. рис. 2, 11).

## 2.3 Доверительный интервал для городского среднего значения при асимптотически нормальном распределении $F$ с $\mathrm{M}X^2 < \infty$

В §1.2 говорилось о сложностях, связанных с применением центральной предельной теоремы для выборочного распределения $F_n$ . Однако, если сама величина $X$ имеет нормальное распределение $\mathcal{N}(\mu(F), \sigma^2(F))$, то и $F_n$ нормальное $\mathcal{N}(\mu(F), \sigma^2(F)/n)$. Делать предположение о нормальности $X$ не хочется. Более того, как показал анализ статистик ГЭПИЦентра, величины, отвечающие всем 15 тестам $X^{(1)}, ..., X^{(15)}$, не имеют нормального распределения (см. §1.1).

Таким образом, остается надеяться на метод Чоу-Роббинса. Но он слишком груб. Полезно для подстраховки иметь альтернативный способ получения доверительного интервала для $\mu(F)$. Можно, в случае малого отклонения от нормального закона для $X$, надеяться на то, что доверительный интервал,

полученный всеже при допущении предположения о нормальности, окажется вполне согласующимся с тем интервалом, что получен на основе метода Чоу-Роббинса. В результате есть шанс иметь более убедительную оценку городского среднего. На каком пути должно появляться допущение о нормальности? Этот путь должен контролировать ошибку, получаемую нами в том случае, когда интересующая нас величина считается нормально распределенной.

Предположим, что каждое $X_i$ $(i = 1, ..., K)$ имеет измеряемое значение (реализацию) вида

$$x_i = \frac{1}{K_i} \sum_{l=1}^{K_i} z_{i,l}, \tag{2.3}$$

где $z_{i,1}, ..., z_{i,K_i}$ – случайная выборка внутри участка, отвечающая показателю $Z_i$, рассматриваемому как случайная величина (см. §2.1), тогда при объеме выборки порядка 400 (число, например, избирателей на участках порядка 1000), вполне обоснованно считать $X_i$ нормально распределенной величиной $\mathcal{N}(\mu_i, \sigma_i^2)$. Но тогда среднее $\overline{X}_n$ является нормальной величиной

$$\mathcal{N}(\frac{1}{n} \sum_{i=1} n\mu_i, \ \frac{1}{n^2} \sum_{1} n\sigma_i^2).$$

Таким образом, можно находить доверительные интервалы традиционным способом на основе распределения Стьюдента [8, с.644] с уровнем доверия $\gamma$:

$$\overline{X}_n - t_{(1+\gamma)/2}[n-1]\frac{s_n}{\sqrt{n-1}} < \mu(F) < \overline{X}_n + t_{(1+\gamma)/2}[n-1]\frac{s_n}{\sqrt{n-1}}, \tag{2.4}$$

где $t_\alpha[\nu]$ – $\alpha$-квантиль распределения величины $t$ с $\nu$ степенями свободы;

$$S_\nu(t_\alpha[\nu]) = \mathbf{P}_t(t < t_\alpha[\nu]) = \alpha.$$

Вычисление доверительного интервала (2.4) для статистик ГЭПИЦентра при $n = 5$ позволит проверить отчасти наблюдение ГЭПИЦентра (2.2). Дело в том, что, во-первых, неизвестна степень случайности в отборе данных $z_{i,l}$, а, во-вторых, не все показатели $x_i$ представимы в виде (2.3).

Полезен ли такой путь? Судя по Приложению к Отчету [1], ГЭПИЦентр *предполагает в будущем проводить работы по схеме, подпадающей под формулы* (2.3) *и* (2.4). Точнее, планируется ГЭПИЦентром делать выборки $z_{i,l}, l = 1, ..., K_i$ объема $K_i = 100$, $i = 1, ..., 5$ для каждого из пяти заранее отобранных участков (принцип отбора обосновывается в Гл.1) для снятия показателей по тесту $Z$, из которых формируется величина $X$ по формуле (2.3). Объем выборки $z_{i,l}$ не столь хорош (будем надеяться на опыт группы, производящей выборку объема 100), как хотелось бы, но позволяет применить центральную предельную теорему, и, следовательно, считать, что $X_i$ являются нормально распределенными величинами $\mathcal{N}(\mu_i, \sigma_i^2)$.

## 2.4 Доверительный эллипсоид для $k$ тестов

К сказанному в предыдущем параграфе полезно добавить следующее. Как показывает вычисление коэффициентов корреляции для тестов $X^{(1)}, ..., X^{(15)}$, а они, как правило, отличны от 0 и 1, мы не можем рассматривать тесты как независимые, особенно в случае их нормального распределения. Поэтому для получения более корректных результатов необходимо использовать многомерный статистический анализ.

Пусть даны «чисто политические» тесты $X^{(1)}, ..., X^{(15)}$ и новый «чисто экономический» $Y$. Все эти 16 показателей коррелируют друг с другом. Как оценить среднее для $Y$ через $\overline{Y}_5$, используя информацию о верности гипотезы (0.1) для $X^{(1)}, ..., X^{(15)}$ и метод доверительных интервалов?

Переобозначим $X, X^{(1)}, ...$ через $Y^{(1)}, ..., Y^{(k)}$, где $1 \leq k \leq 16$, то есть не все из тестов $X^{(j)}$ (и даже $X$) рассматриваются.

Будем предполагать, что величины $Y^{(1)}, ..., Y^{(k)}$ имеют совместное нормальное распределение $\mathcal{N}(\{\mu^{(j)}\}, \{\|\sigma_{st)}\|\})$, где

$$\mu^{(j)} = \mathbf{M}Y^{(j)}$$

– городское среднее величины $Y^{(j)}$; а $\|\sigma_{st)}\|$ – ковариационная матрица:

$$\sigma_{(st)} = \mathbf{M}[(Y^{(s)} - \mathbf{M}Y^{(s)})(Y^{(t)} - \mathbf{M}Y^{(t)})] \quad s, t = 1, ..., k$$

(см. [7, с.177]).

Основание для этого предположения дают не статистики ГЭПИЦентра, а его планы на будущее [1], хотя несомненно нужно проводить исследования на совместную нормальность отдельно в каждом конкретном социологическом обследовании по $k$ тестам.

Пусть для каждого теста $Y^{(j)}$ дана выборка $y_1^{(j)}, ..., y_n^{(j)}$ – замеры этого теста на участках с номерами $1, ..., n$

$$u_{(st)} = \sum_{i=1}^{n} (y_i^{(s)} - \overline{Y}_n^{(s)})(y_i^{(t)} - \overline{Y}_n^{(t)}), \quad s, t = 1, ..., k,$$

где

$$\overline{Y}_n^{(j)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(j)}$$

– элементы *матрицы внутреннего рассеяния*. Если $n > k$, то есть в случае, когда число отобранных для мониторинга участков $n$ больше числа тестов, участвующих в мониторинге $k$, с вероятностью 1 матрица $\|u_{(st)}\|$ положительно определена [7, с.555]. *Таким образом, при пяти участках и более пяти тестах вероятность иметь положительно определенную матрицу рассеяния может отличаться от 1.*

Как показал Хотеллинг [7, с.592-593], с уровнем доверия $\gamma$ доверительный эллипсоид для совокупности городских средних $(\mu^{(1)}, ..., \mu^{(k)})$ имеет вид

$$n \sum_{s,t=1}^{k} u^{(st)} (\overline{Y}^{(s)_n} - \mu^{(s)})(\overline{Y}^{(s)_n} - \mu^{(s)}) < \frac{1 - z_\gamma}{z_\gamma}, \tag{2.5}$$

где

$z_\gamma$ – $100\gamma$-процентный квантиль бета-распределения $Be(\frac{1}{2}(n-k), \frac{1}{2}k)$,

$$\|u^{(st)}\| = \|u_{(st)}\|^{-1}.$$

Можно сделать грубую оценку, увеличивающую уровень доверия $\gamma$, заменив эллипсоид на шар

$$\sum_{j=1}^{k}(\overline{Y}^{(j)_n} - \mu^{(j)})^2 < \frac{(1-z_\gamma)\max_j \lambda_{(j)}}{nz_\gamma},$$

где $\lambda_{(j)}$ – собственное число матрицы рассеяния $\|u_{(st)}\|$. Другими словами, с уровнем доверия $\gamma$ вектор городских средних $(\mu^{(1)}, ..., \mu^{(k)})$ лежит в шаре с центром в векторе выборочных средних по $n$ и с радиусом

$$r_\gamma = \sqrt{\frac{(1-z_\gamma)\max_j \lambda_{(j)}}{nz_\gamma}}.$$

Это дает $\gamma$-доверительный интервал для городского среднего $\mu(X)$ 16-ого теста $X$

$$(\overline{X}_5 - r_\gamma,\ \overline{X}_5 + r_\gamma). \tag{2.6}$$

## ЗАКЛЮЧЕНИЕ

Проведенный анализ статистик ГЭПИЦентра показал:

1) Все пятнадцать тестов $X^{(j)},\ j = 1, ..., 15$ ГЭПИЦентра не отвечают случайным величинам с нормальным распределением. Поэтому использование критерия Стьюдента в отчете [1] для проверки гипотезы (0.1) не является обоснованным. Другими словами, нельзя считать доказанным, что для 95 из 100 реализаций (выборок) $x_1, ..., x_5$ на особо и заранее отобранных 5 участках теста $X^{(j)}$ среднее

$$\overline{X}_5^{(j)} \approx \overline{X}^{(j)}, \tag{3.1}$$

как это имело место при одной выборке, оказавшейся в распоряжении ГЭПИЦентра.

2) Недоказанное – не значит неверное! Компьютерные эксперименты (§1.6) с генерацией реализаций теста вида

$$X = \gamma_1 X^{(1)} + ..... + \gamma_{15} X^{(15)}$$

показали, что с вероятностью 0.95 ошибка

$$\epsilon_l = \frac{\overline{X} - \overline{X}_5}{\sigma} \tag{3.2}$$

для выделенных ГЭПИЦентром участков не превышает 0.4 (где $\sigma$ – среднеквадратичное отклонение для вектора-теста $X$).

Это весомый довод в пользу гипотезы ГЭПИЦентра (0.1) или (3.1).

3) В работе предложен метод построения доверительных интервалов на основе теоремы Чоу-Роббинса, требующий лишь конечности вторых моментов. Апробация метода для тестов ГЭПИЦентра $X^{(j)}$, $j = 1, ..., 15$ показало, что с уровнем доверия 85 - 95 % ошибка от 10 до 20 % достигается для $n = 15 - 20$ и для $n = 20 - 40$ участков соответственно. При $n < 5$ ситуация ухудшается.

Оценка улучшается, если брать специально отобранные в §1.3 участки. Отбор участков должен производится до взятия выборок. Тогда не будет нарушаться «парадигма случайного отбора», поскольку речь идет уже о стратифицированных выборках. Таким образом, следует, скорее, говорить о подтверждении методики ГЭПИЦентра, а не о противоречии принципам организации статистических выборок.

Сила метода Чоу-Роббинса – в отсутствии каких-либо обременительных ограничений на распределение величины $X$, а слабость – в грубости оценки необходимого объема $n$ или в понижении уровня значимости.

4) В §2.3, 2.4 показано, как можно уточнить доверительные интервалы Чоу-Роббинса для исследуемых «экономических» тестов. Для этого требуется от ГЭПИЦентра отработать саму методику взятия выборки $x_1, ..., x_5$ «экономических тестов» $X$.

«Экономические тесты» $X$ ГЭПИЦентра разноплановы: каждая выборка $x_1, ..., x_5$ извлекалась по индивидуальному плану, и, следовательно, величины $X$ могут иметь различные распределения (так и есть в действительности). Или каждый раз надо «возиться» с отдельно взятым тестом, подбирая для него персональный метод нахождения доверительного интервала, или совершенствовать путь, намеченный в §2.3, 2.4.

5) В §1.3 предложен другой способ выделения «репрезентативных участков». При этом было выявлено, что наблюдается пространственная консолидация «репрезентативных» избирательных участков с соответствующими номерами, которые на карте города образуют вытянутые конфигурации, а не хаотические пятна. При предложенной кластеризации реализована основная идея ГЭПИЦентра: участки выделяются по степени близости выборочного среднего по этим участкам к среднему по городу сразу для всех 15 тестов! Это видно из формулы (1.1) для $\Delta_i$, если просуммировать по $i = 1, ..., 5$ каждую $j$-ю компоненту векторов $\Delta_i$ –

$$\forall j = 1, ..., 15 \left( \frac{\frac{1}{5}\sum_{i=1}^{5} x_i^{(j)} - \overline{X}^{(j)}}{\sigma_{(j)}} \to 0 \right).$$

Тем самым ГЭПИЦентр получает процедуру кластеризации, свободную от необходимости интуитивного выбора, присущего отчету [1]. Важно отметить, что наш метод кластеризации участков показывает, что выделенные ГЭПИЦентром участки приемлемы и, в какой-то мере, неслучайны.

6) Имеются основания предположить (см. §1.5, 1.6), что новые тесты являются линейной комбинацией 15 тестов ГЭПИЦентра. Другими словами, для

произвольного теста $Y$ можно подвергнуть анализу в рамках идей ГЭПИЦентра только «политизированную составляющую» $Y_P$ – проекцию вектора-теста $Y$ на «политическое пространство» с базовыми векторами-тестами $X^{(1)}, ..., X^{(15)}$. В таком случае гипотеза ГЭПИЦентра с уровнем доверия 0.95 позволяет считать, что ошибка оценки (3.2) не превысит 0.4.

7) В силу сказанного в 6) каждый новый тест $Y$ должен рассматриваться в комплексе с 15 тестами $X^{(1)}, ..., X^{(15)}$. Однако вместо линейного разложения $Y$ по $X^{(1)}, ..., X^{(15)}$ можно использовать методы многомерного статистического анализа, предполагая, например, существование совместного нормального распределения. Это допущение нормальности вполне приемлемо, если корректно делать выборки с учетом пути, указанного в §2.3. Тогда для оценки городского среднего можно использовать доверительный эллипсоид (2.5) или интервал (2.6) из §2.4.

8) Следует строго ограничивать число тестов при измерении показателей (взятии выборки) в зависимости от числа участков, использованных в исследовании (см. §2.4). При пяти участках, в случае числа тестов более пяти, вероятность иметь положительно определенную матрицу рассеяния может отличаться от 1. Даже если матрица рассеяния вдруг окажется положительно определенной, а число тестов превышало разумный предел, это может означать, что получаемые на таком материале выводы являются скорее исключением, чем правилом.

В заключение следует добавить, что практика несомненно внесет коррективы в методику ГЭПИЦентра. Без практики нет полной ясности и с математическим обоснованием этой методики, поскольку математика базировалась на некоторых очевидных для авторов допущениях. Но вот проблема: очевидны ли эти допущения для окружающей нас реальности?

# ПРИЛОЖЕНИЕ

## Список тестов ГЭПИЦентра

тест 1 — активность избирателей при выборах в Совет Федерации;

тест 2 — процентное соотношение проголосовавших за лидера в Совет Федерации (СФ);

тест 3 — процентное соотношение проголосовавших против всех в СФ;

тест 4 — активность избирателей по партийным спискам в СФ;

тест 5 — за лидера партии в СФ;

тест 6 — против всех партий в СФ;

тест 7 — активность при голосовании в Государственную Думу (ГД);

тест 8 — за лидера в ГД;

тест 9 — против всех партий в ГД;

тест 10 — активность избирателей при голосовании за Конституцию;
тест 11 — процент, проголосовавших «за» Конституцию;
тест 12 — процент, проголосовавших «против» Конституции;
тест 13 — активность избирателей;
тест 14 — за лидера;
тест 15 — против всех.

Тесты 1 – 12 относятся к голосованию 12.12.93.
Тесты 13 – 15 относятся к голосованию 20.03.94.

## Литература

1. *Построение территориальной выборки для организации социального мониторинга в социально-трудовой сфере. Формирование сети респондентов для проведения опросов* // Отчет по теме: х/д N 18-94 от 12.09.94. Омск: ГЭПИЦентр, 1995.

2. Герасимович А.И., Матвеева Я.И. *Математическая статистика.* Минск: Высшая школа, 1978.

3. Уилкинсон Дж., Райнш К. *Справочник алгоритмов на Алголе. Линейная алгебра.* М.: Машиностроение, 1976.

4. Беллман Р. *Введение в теорию матриц.* М.: Наука,1976.

5. Эфрон Б. *Нестандартные методы непараметрической статистики.* М.: Наука,1984.

6. Вапник В.Н. *Восстановление зависимостей по эмпирическим данным.* М.:Наука, 1979.

7. Уилкс С. *Математическая статистика.* М.: Наука, 1967.

8. Закс Ш. *Теория статистических выводов.* М.: Мир, 1975.

### АРХИВНЫЕ ИСТОЧНИКИ

**(Научный архив Центра гуманитарных, социально-экономических и политических исследований – 1)**

9. Отдел 03 фонд 02 сектор 01 дело 152. Построение территориальной выборки для организации социального мониторинга в социально-трудовой сфере. Формирование сети респондентов для проведения опросов // Отчет по теме: х/д N 18-94 от 12.09.94. Омск: ГЭПИЦентр, 1995.

10. Отдел 04 фонд 03 сектор 05 дело 005. Результаты Всероссийского референдума о доверии президенту РФ 25.04.1993 г. Отдел 04 фонд 03 сектор 05 дело 018. Результаты голосования по выборам в Государственную Думу РФ 12.12.1993 г.

11. Отдел 04 фонд 03 сектор 05 дело 033. Описание границ избирательных участков г.Омска по выборам в Государственную Думу РФ 12.12.1993 г.