СЕРГЕЙ ЛИТВИНОВ,
кандидат социологических наук, ведущий специалист DP маркетингового агентства IRS

Использование понятия относительной погрешности оценивания для расчета выборки из генеральной совокупности с малой долей признака

Abstract

The paper elucidates how to determine a sample size from the general population if the analyzed variable's value is small. The Ukrainian sociologist Mykola Churylov was the first who highlighted this problem in the national scientific literature. The author argues that it is necessary to calculate both absolute and relative error. He also points to the sample size from the general population (in case of small variable's value) that should be larger than while using traditional formulas. A new enhanced approach presented in the paper includes using within a representative random sample only a priori characteristics.

Важную роль в эмпирической социологии играет выборочный метод, основанный на теоретико-вероятностном подходе к оценке характеристик генеральной совокупности (ГС) по характеристикам выборочной совокупности (ВС). Специфика предметной сферы и социальная ситуация "обычного" социологического исследования (ограниченность средств, технических и человеческих ресурсов, требование оперативности, труднодоступность элементов эмпирического объекта исследования) стимулируют поиск метода, позволяющего строить обоснованные заключения относительно всей совокупности на основе данных обследования ее части. Именно это имеет целью выборочный метод. Он применяется везде, где следует исходить из указанных выше условий — при изучении бюджетов домохозяйств, структуры свободного времени граждан, при контроле качества продукции и услуг, при статистических обследованиях населения, при массовых социологических опросах.

Выборочный метод охватывает всю проблематику, связанную с отбором единиц, вычислением характеристик выборки, а также с формулированием суждений о количественных характеристиках совокупности, на основании которой построена эта выборка [1]. Каждая интегральная статистическая характеристика выборки (точнее, ее математическое ожидание), например доля признака, среднее значение некоторой величины, ее дисперсия, является оценкой соответствующей характеристики ГС. Характеристики ГС принято называть параметрами. Характеристики ВС, рассчитанные по тому же правилу, суть оценки параметров ГС. Поскольку ВС, по определению, не тождественна ГС, выборочные оценки отличаются от действительных значений параметров. Количественная мера их расхождения называется погрешностью выборки. В зависимости от источника возникновения и свойств погрешностей репрезентативности их разделяют на систематические и случайные. Систематические погрешности возникают из-за недостатков планирования выборки, неслучайности отбора единиц, невалидности процедуры обследования и инструментария измерения первичных характеристик и т.п. Случайные погрешности связаны со стохастичностью отклонений выборочных оценок от генеральных параметров вследствие недетерминированности отбора одних элементов ГС и исключения других (под случайным мы понимаем недетерминированное, стохастическое, а под вероятностным — особый вид причинной детерминации — вероятностную детерминацию). Говоря об ошибках репрезентативности, мы будем иметь в виду именно предельные случайные погрешности, которые, в отличие от систематических, могут быть статистически оценены до осуществления процедуры отбора единиц ВС. Если расчетная случайная погрешность репрезентативности не превышает приемлемых для исследователя границ, то выборка считается репрезентативной. В противном случае она должна считаться нерепрезентативной.

Погрешность репрезентативности всегда имеет конкретное содержание и характеризует не качество выборки в целом, а точность выборочной оценки какого-то параметра генеральной совокупности. Выборка, в достаточной мере репрезентативная относительно одного параметра, например доли некоторого признака, может быть недостаточно репрезентативной относительно доли другого признака [2]. На практике достижима репрезентативность по двум-трем, в лучшем случае — еще нескольким признакам одновременно. Парадокс выборочного метода заключается в том, что репрезентативность ВС по исследуемому признаку требует исчерпывающего знания о его распределении в генеральной совокупности, реконструкция которого как раз и составляет задачу выборочного исследования. В современной теории выборочного метода существует несколько возможностей выхода из этого замкнутого круга.

Во-первых, можно рассуждать следующим образом. Пусть нам нужно количественно оценить параметры неизвестного генерального распределения признака *A*, вместе с тем нам известно распределение признака *B*, связанного с *A*. Тогда можно осуществить выборку, репрезентативную относительно признака *B*, предположив, что чем более репрезентативной будет выборка по *B*, тем выше будет ее репрезентативность относительно *A*. Связь между переменными может быть обоснована теоретически или установлена эмпирически, скажем, на основе наличия корреляции (ковариации) при-

знаков. Рассмотрим, например, такие признаки работников большого предприятия, как "стаж работы", "опытность" (уровень усвоенных профессиональных умений и навыков) и "возраст". Наиболее очевидна связь между первым и вторым признаками — работники с бóльшим стажем работы являются более опытными. Менее очевидна зависимость между признаками "возраст" и "стаж работы". Как правило, большинство работников старшего возраста имеют больший стаж работы, поскольку для них характерна сходная периодичность жизненного пути: в нашем обществе приблизительно в одном возрасте начинают и заканчивают обучение, в одном возрасте начинается и трудовая деятельность. А вот стереотипная очевидность связи между возрастом работника и его опытностью обманчива: на предприятии при определенных обстоятельствах, скажем в условиях перепрофилирования производства, могут преобладать старшие работники с меньшим стажем работы в данной области, а значит менее опытные, нежели младшие, уже освоившие новый технологический процесс. В этом случае наличие связи может быть установлено лишь эмпирически, исходя из предыдущих социологических исследований или статистических обследований.

Довольно часто приходится иметь дело с ситуацией, отличной от описанной выше, когда всякая надежная основа выборки отсутствует. Если это так, то можно: 1) исходя из теоретической модели объекта исследования сделать предположение относительно вида генерального распределения, но в социологии подобный подход скорее исключение, чем правило; 2) воспользоваться следствиями центральной предельной теоремы теории вероятностей (ЦПТ). Данный подход стал каноническим при построении выборки [3].

Вообще говоря, для идеального соблюдения условий применимости ЦПТ необходимо осуществить серию подобных выборок, но при известных допущениях обходятся одной. Для оценки параметров ГС в этом случае применяют точечные и интервальные приближения. Конечно, в силу самой сущности выборочного метода точечная оценка будет отличаться от действительного значения параметра на некоторую стохастическую величину. Поскольку действительное значение неизвестно (его надо оценить), то можно предположить, что погрешность оценивания (отклонение) не превысит определенной заранее заданной нами величины. Отклонение является статистикой, т.е. каждому значению отклонения соответствует вероятность, с которой отклонения реализуются в длинной серии испытаний (при больших объемах выборки). Это означает, что отклонение $\left|\hat{X} - X\right|$ точечной выборочной оценки $\hat{X}$ от "истинного" значения X параметра не превышает заданной погрешности Δ с вероятностью P. Чтобы осуществить интервальное оценивание X, необходимо и достаточно построить доверительный интервал для X. Доверительным интервалом называется рассчитанный по выборочной оценке $\hat{X}$ интервал значений параметра X, в котором находится его действительное значение с доверительной вероятностью P. Доверительный интервал имеет вид

$$\hat{X} - \Delta \le X \le \hat{X} + \Delta. \qquad (1)$$

В отдельных случаях перед исследователем стоит задача сравнения точности выборочных оценок. Количественной мерой точности в статистике принято считать относительную погрешность выборки [2], которая показывает, на какую часть своей величины точечная оценка отличается от дейст-

вительного значения параметра в пределах единичного доверительного интервала (при $t = 1$ и $P(|\hat{X} - X| < \sigma) \approx 0{,}683$):

$$E_{\mu} = \mu_{\hat{X}} / \hat{X}, \qquad (2)$$

$\mu_{\hat{X}}$ — среднее квадратическое отклонение выборочной оценки генерального параметра.

Точность оценки следует рассматривать в качестве важнейшего критерия репрезентативности выборки. В то же время социологи необоснованно пользуются понятием абсолютной погрешности как единственной характеристики репрезентативности. Для выборок из генеральной совокупности с малой долей признака это приводит к коллизии, на которую обратил внимание Н.Чурилов [4]. Использование традиционных формул приводит здесь к ошибочным выводам о небольшом объеме репрезентативной выборки. Так, для признака, доля которого в квазибесконечной ГС составляет 10%, объем репрезентативной выборки при погрешности в 5% (примем доверительную вероятность равной 0,95) равен 138 единицам, а значит, в выборке окажется всего 14 (±7 с доверительной вероятностью 95%) единиц, обладающих искомым признаком, что явно недостаточно для анализа. Данное обстоятельство можно объяснить большой относительной погрешностью выборки, которая в приведенном примере приближается к 50%.

Как же оптимизировать расчет объема выборки на основе имеющейся априорной информации про долю биномиального признака? Н.Чурилов для этой цели вместо традиционных формул предлагает оценивать объем ВС на основе коэффициента вариации выборочной оценки (фактически, стандартной относительной погрешности выборки) [4]:

$$n = \frac{N}{(NpE_p^2 / (1-p)) + 1} = \frac{1}{(pE_p^2 / (1-p)) + 1/N}.$$

Если p мала и $p E_p^2 \sim 1/N$ (объем выборки значительно меньше объема генеральной совокупности), $n \approx 1 / p E_p^2$. (3)

E_p — стандартная относительная погрешность оценки генеральной доли по выборочной ($E_p = E_{\mu}$ для параметра "доля признака" p)[1].

Но доля признака в генеральной совокупности нам, как правило, заранее не известна. Еще один недостаток формулы (3) состоит в том, что в нее входят как генеральный параметр (априорная информация), так и величина, производная от выборочной статистики (апостериорная информация).

Чтобы преодолеть этот недостаток, нужно перейти от использования абсолютных погрешностей к относительным величинам как показателям репрезентативности выборочной совокупности. Один из возможных подходов состоит в следующем.

Неоднозначность использования предельной абсолютной погрешности при планировании выборочной совокупности состоит в том, что при отсутствии информации о доле признака p приравнивается 0,5 и стандартная относительная погрешность оценивания доли E_p получается **минимальной** (см. формулу (13) ниже). Это приводит к занижению объема репрезентативной для доли p выборки, особенно существенному в случае малой p. Для

[1] Все условные обозначения по статье расшифрованы в Приложении.

любого другого значения доли биномиального признака, такой, что $p < 0{,}5$, погрешность E_p будет больше минимальной. Поэтому для адекватной репрезентации доли исследуемого признака или другого, связанного с исследуемым признаком, распределение которого известно, желательно использовать приблизительную априорную информацию о доле, а выборку рассчитывать исходя из оценки относительной погрешности.

Выразим относительную погрешность выборки через абсолютную погрешность Δ и в результате получим оценку относительной погрешности оценивания:

$$V_\Delta = \frac{\Delta}{\hat{X}} = \frac{t\mu_{\hat{X}}}{\hat{X}} = tV_\mu. \qquad (4)$$

Но $n \approx \dfrac{\hat{\sigma}^2 t^2}{\Delta^2}$, тогда $n \approx \dfrac{\hat{\sigma}^2 t^2}{\Delta^2} \approx \dfrac{\hat{V}_X^2 \hat{X}^2 t^2}{E_\Delta^2 \hat{X}^2} = \dfrac{\hat{V}_X^2 t^2}{t^2 E_\mu^2} = \dfrac{\hat{V}_X^2}{E_\mu^2}$,

$$n \approx (\hat{V}_X / E_\mu)^2, \qquad (5)$$

где $E_\Delta = \Delta / \hat{X} = t\mu_{\hat{X}} / \hat{X} = tE_\mu$.

Соотношение (5) фиксирует тот факт, что объем выборки прямо пропорционален квадрату величины, которая показывает, во сколько раз относительная погрешность оценивания меньше коэффициента вариации соответствующего параметра (или величины, равной отношению стандартного отклонения параметра в ГС к абсолютной погрешности выборочной оценки параметра). Для того, чтобы точность повысилась в r раз (в r раз уменьшилась относительная погрешность оценивания), нужно увеличить объем ВС в r^2 раз.

Найдем априорное выражение для стандартной относительной погрешности. Из того, что выборочный коэффициент вариации равен $\hat{V}_X = \hat{\sigma}_X / \hat{X}$, а $\mu_X = \sqrt{\dfrac{N-n}{N(n-1)}\hat{\sigma}_X^2} \underset{\frac{n}{N} \le 0{,}05}{\approx} \sqrt{\dfrac{1}{n-1} - \dfrac{1}{N}}\hat{\sigma}_X$ или $\mu_X \approx \dfrac{\hat{\sigma}_X}{\sqrt{n}}$, можно получить, учитывая (2), что стандартная относительная погрешность оценки равна:

$$E_\mu = \frac{\hat{V}_X \hat{X} / \sqrt{n}}{\hat{X}} = \frac{\hat{V}_X}{\sqrt{n}}; \qquad (6)$$

с поправкой на конечность ГС и при условии, что мы планируем выборку больше нескольких десятков единиц:

$$E_\mu = \sqrt{\frac{N-n}{N(n-1)}\hat{V}_X^2 \hat{X}^2} \cdot \frac{1}{\hat{X}} = \frac{\hat{V}_X}{\sqrt{n-1}}\sqrt{1 - \frac{n}{N}} \underset{\frac{n}{N} \le 0{,}05}{\approx} \sqrt{\frac{1}{n} - \frac{1}{N}} \cdot \hat{V}_X. \quad (7)$$

Тогда из (7) получим уточнение формулы (5) с поправкой на конечность ГС:

$$n = \frac{\hat{V}_X^2 + E_\mu^2}{E_\mu^2 + \hat{V}_X^2 / N}. \qquad (8)$$

По заданной величине относительной погрешности можно вычислить $\hat{\sigma}_X$, а на ее основании — доверительные интервалы, доверительные вероятности и объемы репрезентативных относительно доли признака выборок.

$$\hat{V}_X = \hat{\sigma}_X / \hat{X} = E_\mu \sqrt{n-1}; \quad \hat{\sigma}_X = \hat{X}\hat{V}_X = E_\mu \hat{X}\sqrt{n-1}, \quad (9)$$

или

$$\hat{\sigma}_X = E_\mu \hat{X}\sqrt{n-1} / \sqrt{1-\frac{n}{N}} = E_\mu \hat{X}\sqrt{\frac{N(n-1)}{N-n}}. \quad (10)$$

В случае доли признака $E_\mu = \dfrac{\mu_p}{\hat{p}} = \dfrac{\sqrt{\dfrac{p(1-p)}{n-1}}\sqrt{1-\dfrac{n}{N}}}{\hat{p}}$. (11)

Для больших ***n*** исходя из теоремы Бернулли можно считать, что в формуле (11) $\hat{p} / p \approx 1$, а

$$E_\mu = \sqrt{\frac{p(1-p)}{n-1}}\sqrt{1-\frac{n}{N}}. \quad (12)$$

Стандартная относительная погрешность оценки доли биномиального признака приближенно равна

$$E_p \approx V_p = \sqrt{\frac{1-p}{np}}. \quad (13)$$

Для выборочной оценки доли признака из (12), учитывая, что $E_\Delta = t\,E_p$, получаем:

$$n \approx \frac{1-p}{pE_p^2} = t^2\,\frac{1-p}{pE_\Delta^2}. \quad (14)$$

Для малых известных p относительная погрешность оценивания должна быть минимальной при абсолютной погрешности, равной Δ, иначе апостериорная относительная погрешность оценивания с высокой вероятностью будет больше, чем принятая нами априорная относительная погрешность E_Δ в формуле (14):

$$E_\Delta(\min) = \frac{\Delta}{\hat{p}(\max)} = \frac{\Delta}{p+\Delta}. \quad (15)$$

Преобразуем (14) в соответствии с (15) и получим:

$$n_1 = t^2\left(\frac{p}{\Delta}+1\right)^2\left(\frac{1}{p}-1\right) = t^2\left(\frac{1}{\varepsilon}+1\right)^2\left(\frac{1}{p}-1\right), \quad (16)$$

$\varepsilon = \Delta / p$ — критерий точности выборки, по смыслу аналогичный относительной погрешности, но отображающий не выборочную оценку доли, а ее действительную величину в ГС.

Формула (16) представляется более удовлетворительной, чем (3), так как в нее входит лишь априорная информация о генеральной совокупности, а не гипотезы относительно выборочной статистики, что было бы нелогично.

Табулируем n_1 для различных p и Δ, учитывая, что $\Delta \le p$ и $\boldsymbol{t} = 2$ ($P = 0{,}95$):

Таблица 1

Объем n_1 простой вероятностной выборки, пригодной для репрезентации доли признака p в ГС по выборочной оценке со статистической погрешностью Δ[1]

p\Δ	0,01	0,05
0,01	1584	–
0,05	2736	304
0,10	4356	324
0,15	5803	363
0,20	7056	400
0,25	8112	432
0,30	8969	457
0,35	9627	475
0,40	10086	486
0,45	10345	489
0,50	10404	484

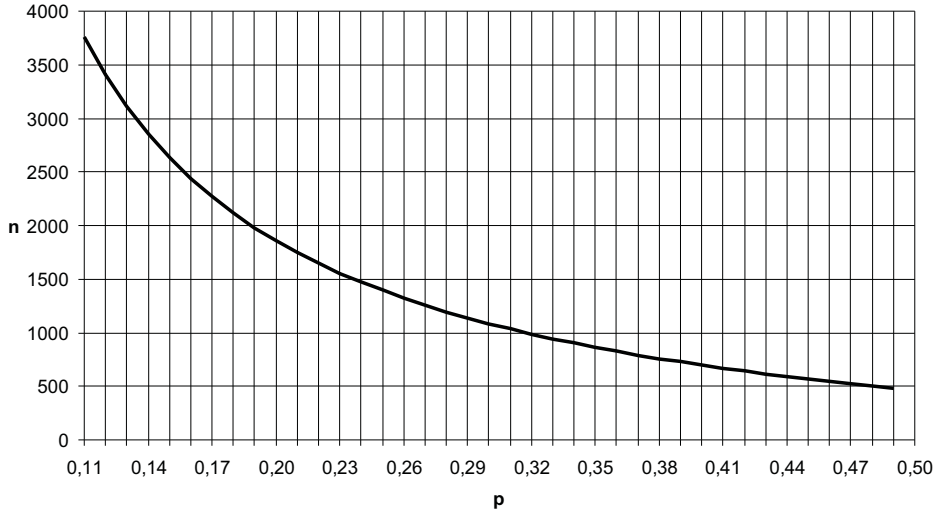

Рисунок. Зависимость оптимального объема $n = t^2\,(1/\varepsilon + 1)^2\,(1/p - 1)$ простой вероятностной выборки, пригодной для репрезентации доли признака p (интервал от 0,11 до 0,5) в ГС по выборочной оценке со статистической погрешностью не больше $\Delta = 0{,}1\,p$ ($P = 0{,}95$)

Полезно также иметь таблицу объема ВС $n_2 \approx (V_X / E_\mu)^2 = t^2\,(V_X / E_\Delta)^2, \boldsymbol{t} = 2$ ($P = 0{,}95$):

1 См. рисунок.

Таблица 2

Объем n, простой вероятностной выборки, пригодной для репрезентации параметра ГС с коэффициентом вариации V_X по выборочной оценке с относительной статистической погрешностью E_Δ

$E_\Delta \backslash V_X$	0,01	0,05	0,1	0,2	0,3
0,05	100	–	–	–	–
0,1	400	–	–	–	–
0,2	1600	–	–	–	–
0,3	3600	144	–	–	–
0,4	6400	264	–	–	–
0,5	10000	400	100	–	–
0,6	–	584	144	–	–
0,7	–	784	196	–	–
0,8	–	1024	256	–	–
0,9	–	1296	324	–	–
1,0	–	1600	400	100	–

* *Примечание.* Пометкой "–" обозначены значения, не соответствующие условиям аппроксимации генеральной доли нормальным распределением выборочных оценок.

Рассмотрим различия между методами расчета выборочной совокупности, о которых речь шла выше, на примере. Пусть необходимо провести масштабное электоральное исследование, которое нацелено, в частности, на изучение характеристик электората политических партий, которые преодолевают 3%-й проходной барьер на выборах в парламент[1].

(а) В соответствии с традиционным подходом, число единиц ВС (генеральную совокупность можно считать квазибесконечной) равно: $n \approx \sigma^2 t^2 / \Delta^2$, $\sigma^2 = p\,(1-p)$, $p = 0{,}03$ – доля признака в ГС. Примем в качестве критерия удовлетворительной точности выборки $\varepsilon = \Delta / p = 0{,}2$, $t = 1{,}96$.

$$\text{Тогда } n = \lim_{p \to 0,03} \frac{p(1-p)\cdot t^2}{\Delta^2} \approx \frac{t^2}{\varepsilon\Delta} = \frac{t^2}{\varepsilon^2 p},$$

$$n \approx 1{,}96^2 / (0{,}2^2 \cdot 0{,}03) \approx 3201.$$

(б) Согласно подходу Н.Чурилова, $n = 1/(p\,E_p^2)$.

$E_p = E_\Delta / t \approx \varepsilon / t$.

$$\text{Отсюда: } n = \frac{1}{p}\cdot\frac{1}{(\varepsilon/t)^2} = \frac{t^2}{\varepsilon^2 p}, \quad n \approx 1{,}96^2 / (0{,}2^2 \cdot 0{,}03) \approx 3201.$$

В данном случае подход (б) является лишь корректной формулировкой традиционного подхода (а).

(в) Применим для расчета ВС формулу (16), основанную на предположении о максимальной возможной относительной погрешности:

$n = t^2\,(1/\varepsilon + 1)^2\,(1/p - 1)$, $n = 1{,}96^2\,(5+1)^2\,(1/0{,}03 - 1) \approx 4472$.

[1] Пример вымышленный, так как в реальности для указанной цели случайный отбор респондентов малоэффективен.

Как видим, выборка из генеральной совокупности с малой долей признака должна быть большего объема, чем можно судить, исходя из привычных формул (см. рис.). Больший объем выборки обусловлен тем, что использование принципа минимизации относительной ошибки позволяет избежать занижения дисперсии биномиального признака с малой долей на этапе планирования выборки. При использовании традиционного подхода априорное занижение дисперсии очень вероятно, и поэтому выборочные оценки доли оказываются неточными, с неоправданно широкими доверительными границами.

Заметим, что почти все приведенные способы оценки объема выборочной совокупности применимы лишь: (1) для больших выборок ($n > 100$); (2) для несдвинутых выборок или таких, для которых величиной сдвига B можно пренебречь, когда $B / \hat{\sigma} < 0{,}1$ [5]; (3) для генеральной совокупности с коэффициентом вариации исследуемого признака $\sigma / \mathrm{X} < \sqrt{n} / 3$, где n — объем выборочной совокупности. Если коэффициент вариации больший, то при большом объеме ВС (ГС квазибесконечна и $\hat{V}_X \approx V_X$) нижняя граница доверительного интервала параметра X с вероятностью $P(t = 3) \approx 0{,}997$ стремится к ($X_{\min} \approx \hat{X} - 3\hat{\sigma} / \sqrt{n}) < 0$, что лишено смысла, поэтому следует использовать скорректированные формулы вычисления предельной погрешности выборки. Можно предположить, что в последнем случае к распределению вероятностей погрешности оценивания ЦПТ (вследствие несоблюдения условий применимости теоремы Ляпунова) неприменима, поэтому его нельзя считать нормальным. Соответственно, теряют смысл традиционные формулы построения интервальных оценок, а сами интервалы не будут симметричными [1; 6]. Распределение погрешностей случайных выборок из генеральной совокупности с малой долей признака аппроксимируется не нормальным распределением, а распределением Пуассона или подобным ему скошенным распределением.

Таким образом, выборки из генеральных совокупностей с малой долей признака должны рассчитываться на иных статистических основаниях, нежели обычные, или же их следует заменять одним из методов направленного отбора, монографическим исследованием, использовать процедуры бустинга.

ПРИЛОЖЕНИЕ

Аббревиатуры и основные условные обозначения

X– параметр распределения статистической величины в ГС

$\hat{X}$– выборочная оценка параметра распределения статистической величины в ГС

σ – среднеквадратическое отклонение признака в ГС

t — квантиль нормального распределения оценок в интервальном оценивании характеристик ГС

$\hat{\sigma}$ – среднеквадратическое отклонение признака в ВС

μ_X — среднее квадратическое отклонение выборочной оценки генерального параметра от “истинного” значения

V_X – генеральный коэффициент вариации признака

$\hat{V}_X$ – коэффициент вариации признака в ВС

$E_\mu = \mu_X / \hat{X}$ — стандартная относительная погрешность выборки (выборочной оценки параметра $\hat{X}$)

N — объем ГС

n — объем ВС

Литература

1. *Шварц Г.* Выборочный метод. Руководство по применению статистических методов оценивания. — М., 1978.

2. Общая теория статистики / Под ред. А.Я.Боярского, Г.Л.Громыко. — М., 1985.

3. *Гнеденко Б.В.* Курс теории вероятностей. — М., 1988.

4. Оперативные социологические исследования : Учебное пособие. — Минск, 1997.

5. *Кокрен У.* Методы выборочного исследования. — М., 1976.

6. *Орлов А.И.* Социология: методология, методы, математические модели. — М., 1992. — С. 28–50.