УДК 311.17

А.В. Гласко, Л.Г. Садыхова

ОПРЕДЕЛЕНИЕ МИНИМАЛЬНОГО ОБЪЕМА ВЫБОРКИ РЕСПОНДЕНТОВ ДЛЯ ПРОВЕДЕНИЯ СОЦИОЛОГИЧЕСКОГО ИССЛЕДОВАНИЯ

Разработан метод определения минимального объема репрезентативной выборки респондентов для проведения социологического исследования.

E-mail: petronyi@mail.ru

Ключевые слова: *объем выборки, доверительная вероятность, социологическое исследование.*

Для успешного осуществления социального контроля нужна оперативная обратная связь между управляющими структурами и управляемыми субъектами [1, 2]. Эта обратная связь осуществляется, как правило, с помощью проведения социологических опросов.

В связи с этим целесообразно разработать метод определения минимального объема опрашиваемых при проведении социологического опроса.

Проводя социологическое исследование, принято считать, что совокупность участников опроса репрезентативна, т. е. адекватно и представительно отображает качественные характеристики той или иной группы населения. Другими словами, эта выборка респондентов позволяет считать результаты исследования достоверно отображающими свойства генеральной совокупности — объекта данного социологического исследования [1, 3, 4].

Пусть t — заданное время проведения социологического опроса. Естественно, что не все участники опроса способны должным образом сосредоточить свое внимание на поставленной перед ними задаче адекватно, т. е. обдуманно, самостоятельно и своевременно ответить на все вопросы анкеты [5]. Поэтому обозначим через τ время выполнения задания участником социологического опроса в течение заданного времени t.

Введем следующую случайную величину:

$$\eta(t) = \begin{cases} \tau, \text{ если } \tau < t; \\ t, \text{ если } \tau = t. \end{cases} \qquad (1)$$

Тогда величина $\eta(t)/t$ — доля от полного времени t, затрачиваемого на заполнение анкеты участником опроса.

Будем считать, что время заполнения анкеты респондентом прямо пропорционально количеству содержащихся в ней вопросов.

Определим среднюю долю от полного времени t, затрачиваемую на заполнение анкеты, по следующей формуле:

$$Y(t) = \left\langle \frac{\eta(t)}{t} \right\rangle, \tag{2}$$

где $\langle \cdot \rangle$ — символ математического ожидания величины, стоящей внутри угловых скобок.

Расчет показателя *Y(t).* Для дальнейшего изложения нам понадобится следующее утверждение.

Теорема 1. Пусть $F(x)$ — функция распределения величины τ. Тогда справедлива следующая формула:

$$Y(t) = 1 - \frac{1}{t}\int_0^t F(x)\,dx. \tag{3}$$

Доказательство. Пусть $f_\eta(x)$ — плотность распределения случайной величины (1). Тогда

$$f_\eta(x) = \begin{cases} f(x), \text{ если } x \in (0,t); \\ 1 - F(t), \text{ если } x = t, \end{cases}$$

где $f(x)$ — плотность распределения случайной величины τ..

Согласно формуле расчета математического ожидания смешанной случайной величины $\eta(t)$, имеем

$$\langle \eta(t) \rangle = \int_0^t x f(x) dx + t\left(1 - F(t)\right), \tag{4}$$

где первое слагаемое соответствует непрерывной, а второе — дискретной части величины (1).

Поскольку $f(x) = F'(x)$, то, интегрируя по частям, получаем

$$\int_0^t x f(x) dx = tF(t) - \int_0^t F(x) dx.$$

Подставим это выражение в (4):

$$\langle \eta(t) \rangle = t - \int_0^t F(x) dx.$$

Тогда с учетом формулы (2) получаем искомую формулу (3).

Формула (3) позволяет установить следующие свойства показателя $Y(t)$:

1) $0 \le Y(t) \le 1$;

2) $\lim\limits_{t \to +0} Y(t) = 1$;

3) $\lim\limits_{t \to \infty} Y(t) = 0$;

4) $Y(t) \ge 1 - F(t)$;

5) $Y'(t) = \dfrac{1}{t}\left[\dfrac{1}{t}\int\limits_0^t F(x)dx - F(t)\right]$;

6) $F(t) - \dfrac{1}{t}\int\limits_0^t F(x)dx = \dfrac{1}{t}\int\limits_0^t xf(x)dx.$

Первые три свойства очевидны, четвертое — вытекает согласно следующей оценке:

$$\frac{1}{t}\int\limits_0^t F(x)dx \le F(t).$$

Пятое свойство следует из формулы (3), причем из него, согласно предыдущей оценке, заключаем, что $Y(t)$ как функция времени t монотонно убывает. Наконец, шестое свойство вытекает из формулы

$$\int\limits_0^t F(x)dx = tF(t) - \int\limits_0^t xf(x)dx.$$

Точечная оценка показателя *Y(t)*. Формула (2) позволяет найти точечную оценку показателя $Y(t)$. Пусть некоторое социологическое исследование заключается в изучении отношения людей к некой проблеме с помощью конкретных вопросов, содержащихся в анкете. При этом число участников опроса n, из которых k человек выполнили задание в течение времени t. Обозначим через τ_i время, затраченное на заполнение анкеты i-м участником в течение заданного времени t. Тогда точечной оценкой показателя $Y(t)$ служит следующая величина:

$$\hat{Y}_n(t) = \frac{1}{nt}\left[\sum_{i=1}^{k} \tau_i + (n-k)t\right]. \qquad (5)$$

Покажем это. Доля от полного времени t, затрачиваемая на выполнение задания i-м (справившимся с ним) участником, равна τ_i / t. Аналогичная доля затраченного времени для участника социологического опроса, который успел ответить на все вопросы анкеты, равна 1.

Поскольку количество участников первой группы составляет k человек, а второй — $n-k$, то оценкой показателя (2) служит величина

$$\hat{Y}_n(t)=\frac{\sum\limits_{i=1}^{k}\tau_i/t+(n-k)}{n}. \tag{6}$$

Отсюда следует искомая точечная оценка (5).

В оценке (5) величины k и τ_i $(i=1,2,...,k)$ случайные [6]. Поэтому возникает вопрос: смещена ли найденная оценка? В связи с этим докажем следующее утверждение.

Теорема 2. Точечная оценка $\hat{Y}_n(t)$, определенная формулой (5) для показателя $Y(t)$, несмещенная, т. е. справедлива следующая формула:

$$\left\langle \hat{Y}_n(t)\right\rangle = Y(t). \tag{7}$$

Доказательство. Из формулы (6) находим

$$\hat{Y}_n(t)-1+\hat{F}_n(t)=\frac{1}{nt}\sum_{i=1}^{k}\tau_i.$$

Здесь

$$\hat{F}_n(t)=k/n \tag{8}$$

— точечная оценка функции распределения $F(t)$ случайной величины τ в течение времени t, т. е.

$$F(t)=\Pr(\tau<t); \tag{9}$$

$\Pr(\cdot)$ — вероятность события, заключенного внутри скобок (от английского слова Probability — вероятность).

Отсюда получаем

$$\left\langle \hat{Y}_n(t)-1+\hat{F}_n(t)\right\rangle=\frac{1}{nt}\left\langle\sum_{i=1}^{k}\tau_i\right\rangle. \tag{10}$$

Будем считать, что длительности $\tau_1,\tau_2,...,\tau_k$ имеют одну и ту же функцию распределения

$$F_t(x)=\Pr\left((\tau<x)\middle|(\tau<t)\right),$$

где τ — одна из величин $\tau_1,\tau_2,...,\tau_k$; $(\tau<x)\big|(\tau<t)$ — обозначение события $\tau<x$ при условии, что $\tau<t$.

Единство функции распределения означает, что рассматриваемая совокупность респондентов как объектов социологического исследования однородна.

Найдем выражение для $F_t(x)$ при $x < t$.

Согласно теореме умножения вероятностей,

$$F_t(x) = \frac{\Pr\left((\tau < x) \bigcap (\tau < t)\right)}{\Pr(\tau < t)},$$

где $\bigcap$ — символ произведения событий.

Поскольку $x < t$, то

$$F_t(x) = \frac{F(x)}{F(t)},$$

где $F(\cdot)$ — функция распределения случайной величины τ, определенная формулой (9).

Тогда для соответствующей условной плотности распределения получаем

$$f_t(x) = \frac{f(x)}{F(t)},$$

где $f(x) = F'(x)$ — плотность распределения случайной величины τ $(x < t)$.

Следовательно, математическое ожидание затраченного времени на выполнение задания в течение заданной длительности времени t

$$\langle \tau | t \rangle = \frac{1}{F(t)} \int_0^t x f(x) dx. \qquad (11)$$

Согласно тождеству Вальда [7],

$$\left\langle \sum_{i=1}^{k} \tau_i \right\rangle = \langle k \rangle \langle \tau | t \rangle.$$

Учитывая формулу Бернулли

$$\langle k \rangle = nF(t) \qquad (12)$$

и формулу (11), находим

$$\frac{1}{nt} \left\langle \sum_{i=1}^{k} \tau_i \right\rangle = \frac{1}{t} \int_0^t x f(x) dx.$$

Следовательно, согласно соотношению (10),

$$\left\langle \hat{Y}_n(t) - 1 + \hat{F}_n(t) \right\rangle = \frac{1}{t}\int_0^t xf(x)dx. \qquad (13)$$

Для слагаемого в левой части выражения (13) с учетом (8) и (12) получаем

$$\left\langle \hat{Y}_n(t) \right\rangle - 1 + F(t) = \frac{1}{t}\int_0^t xf(x)dx. \qquad (14)$$

Правые части формул (14) и свойства 6 показателя $Y(t)$ равны, а значит, равны и левые части, т. е.

$$\left\langle \hat{Y}_n(t) \right\rangle - 1 + F(t) = F(t) - \frac{1}{t}\int_0^t F(x)dx.$$

Отсюда находим

$$\left\langle \hat{Y}_n(t) \right\rangle = 1 - \frac{1}{t}\int_0^t F(x)dx.$$

Правая часть найденного выражения, согласно формуле (3), равна $Y(t)$, следовательно, искомое соотношение (7) получено, что и требовалось доказать.

Нижняя доверительная граница показателя *Y(t)*. При малых объемах выборки n степень доверия к точечной оценке $\hat{Y}_n(t)$ крайне низка. Поэтому докажем следующее утверждение.

Теорема 3. Пусть p — заданная доверительная вероятность. Тогда нижней доверительной границей показателя $Y(t)$ служит следующая величина:

$$\underline{Y}_n(t) = \hat{Y}_n(t) - \sqrt{\frac{-\ln(1-p)}{2n}}. \qquad (15)$$

Доказательство. Для установления формулы (15) воспользуемся неравенством Хевдинга [8]

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon \right) \leq \exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)} \right), \qquad (16)$$

где X_i — случайная величина, удовлетворяющая условию $a_i \leq X_i \leq b_i$ $(i = 1, 2, ..., n)$; $\mu = \left\langle \frac{1}{n}\sum_{i=1}^{n} X_i \right\rangle$; $\varepsilon > 0$ — произвольное число.

Примем следующие обозначения:

$$X_i = \begin{cases} \tau_i / t, \text{если} \quad \tau_i \in (0,t); \\ 1, \ \text{если} \quad \tau_i = t; \end{cases}$$

$$a_i = 0\,, \ b_i = 1 \ \text{при} \ i = 1,2,...,n\,.$$

Тогда, согласно (6), имеем

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{Y}_n(t). \tag{17}$$

Для определения величины μ воспользуемся теоремой 2, согласно которой

$$\mu = Y(t). \tag{18}$$

Поскольку $\sum\limits_{i=1}^{n}(b_i - a_i) = n,$ подставляя соотношения (17) и (18) в неравенство (16), получаем

$$\Pr\left(\hat{Y}_n(t) - Y(t) \ge \varepsilon\right) \le \exp\left(-2n\varepsilon^2\right).$$

Следовательно,

$$\Pr\left(\hat{Y}_n(t) - Y(t) < \varepsilon\right) > 1 - \exp\left(-2n\varepsilon^2\right),$$

откуда находим

$$\Pr\left(Y(t) > \hat{Y}_n(t) - \varepsilon\right) > 1 - \exp\left(-2n\varepsilon^2\right). \tag{19}$$

Поскольку число $\varepsilon > 0$ произвольное, выберем его из следующего условия:

$$1 - \exp\left(-2n\varepsilon^2\right) = p, \tag{20}$$

где p — заданная доверительная вероятность, $0 < p < 1$.

Решая уравнение (20), находим

$$\varepsilon = \sqrt{\frac{-\ln(1-p)}{2n}}.$$

Следовательно, согласно соотношениям (19) и (20), имеем

$$\Pr\left(Y(t) > \underline{Y}_n(t)\right) > p,$$

где значение $\underline{Y}_n(t)$ определено формулой (15), что и доказывает теорему 3.

Определение минимального объема выборки респондентов для проведения социологического исследования. Формула (15) позволяет установить минимальное количество участников, необходимое для проведения выборочного социологического исследования. С этой целью докажем следующее утверждение.

Теорема 4. Пусть $\underline{Y}_n(t)$ — заданная нижняя доверительная граница показателя $Y(t)$ при заданной доверительной вероятности p $(0 < p < 1)$. Тогда минимальный объем n_0 выборки респондентов для проведения выборочного социологического исследования в течение заданного периода времени t определяется из следующего выражения:

$$n_0 = \min\left(n \middle| n \geq \frac{-\ln(1-p)}{2\left[1-\underline{Y}_n(t)\right]^2} \right). \tag{21}$$

Доказательство. Из формулы (15) находим

$$n = \frac{-\ln(1-p)}{2\left[\hat{Y}_n(t) - \underline{Y}_n(t)\right]^2}.$$

Поскольку, согласно формуле (5), $\hat{Y}_n(t) \leq 1$, то

$$n \geq \frac{-\ln(1-p)}{2\left(1-\underline{Y}_n(t)\right)^2}. \tag{22}$$

Следовательно, минимальный объем n_0 выборки респондентов для проведения социологического исследования определяется из неравенства (22) как наименьшее целое число, что и доказывает теорему 4.

Пример. Определить минимальный n_0 объем выборки респондентов для проведения выборочного социологического исследования в течение заданного времени t при условии, что с доверительной вероятностью $p = 0{,}865$ нижняя доверительная граница показателя $\underline{Y}_n(t) \geq 2/3$.

Решение. Условие $p = 0{,}865$ отражает меру соответствия выборочной совокупности респондентов генеральной, а $\underline{Y}_n(t) \geq 2/3$ — нижнюю доверительную границу ожидаемого значения средней доли от полного времени t, затрачиваемой на выполнение задания представителем генеральной совокупности. Поскольку

$$p \simeq 1 - e^{-2},$$

то, согласно (22), получаем

$$n \geq \frac{-\ln\left(e^{-2}\right)}{2 \cdot 1/9}.$$

Правая часть здесь равна 9, а значит, согласно (21), минимальный объем выборки респондентов для проведения социологического исследования $n_0 = 9$.

Из выражения (21) следует, что если доверительная вероятность $p \to 1$, то минимальный объем выборки респондентов увеличивается, а если $p \to +0$, то уменьшается.

Оба вывода хорошо согласуются с логикой проведения любого выборочного эксперимента [9].

Таким образом, разработан метод определения минимального объема выборки респондентов для проведения социологического исследования в течение времени t при заданном значении нижней доверительной границы средней доли от полного времени t, затрачиваемой на выполнение задания представителем генеральной совокупности при заданной доверительной вероятности p.

Работа выполнена при поддержке РФФИ (проект №10-08-00607-а).

СПИСОК ЛИТЕРАТУРЫ

1. Кравченко А.И., Тюрина И.О. Социология управления. М.: Академический проект, 2005. – 1136 с.
2. Толстова Ю.Н. Может ли социология «разговаривать» на языке математики? // Социолог. исслед. – 2000. – № 5. – С. 107–116.
3. Гуц А.К., Фролова Ю.В. Математические методы в социологии. – М.: ЛКИ, 2007. – 280 с.
4. Чуриков А.Б. Случайные и неслучайные выборки в социологических исследованиях // Социальная реальность. – 2007. – № 4. – С. 89–95.
5. Гласко А.В. К проблеме управления сознанием // Фундаментальные проблемы системной безопасности / Вычислительный центр РАН. – 2012. – Вып. 3. – С. 582–585.
6. Fararo T., Butts C. Advances in generative structuralism: Structured agence and multilevel dynamics // The Journal Of Mathematical Sociology. – 1999. – V. 24. – No 1. – P. 1–65.
7. Skvorets J. Looking Backwards into Future: Mathematical Sociology Then and Now // Sociological Theory. – 2000. – V. 18. – P. 510–517.
8. Hoeffding W. Probability inequalities for sums of bounded random variables // J. Amer. Statist. Ass. – 1963. – V. 58. – No 301. – P. 13–30.
9. Давыдов А.А. Системный подход в социологии: новые направления, теории и методы анализа социальных систем. – М.: Эдиториал УРСС, 2005. – 396 с.

Статья поступила в редакцию 28.09.2012