

ОПРЕДЕЛЕНИЕ ОБЪЕМА ВЫБОРОЧНОЙ СОВОКУПНОСТИ ПРИ ЗАДАННОМ ЗНАЧЕНИИ ОШИБКИ

Эд. А. АБГАРЯН

Теория выборки представляет собой важную часть социологического исследования. Это обстоятельство вовсе не исключает возможность сплошного опроса: напротив, сплошной опрос был и остается наиболее точным методом получения информации. Однако выборочные обследования позволяют сократить затраты на сбор и обработку социологической информации и добиться большей оперативности. Но суждение в целом служит источником статистических ошибок. Поэтому центральной проблемой теории выборки является проблема соотношения генеральной и выборочной совокупностей (определение объема выборочной совокупности при заданном значении ошибки), т. е. проблема количественно-качественной репрезентации.

Выборочным называется такое обследование, при котором наблюдению подвергается часть единиц совокупности, отобранных на основе научно разработанных принципов, обеспечивающих получение достаточно достоверных данных для характеристики совокупности в целом.

Например, если проводится выборочное обследование по стране, то население больших и малых городов, а также население деревни должно встречаться в выборке примерно в тех же соотношениях, которые характерны для всего населения страны. Тогда частоты, вычисленные по выборке, будут являться приближенными значениями соответствующих частот для всего населения. Что же происходит в тех случаях, когда частоты выборочной совокупности не соответствуют частотам изучаемой совокупности?

Рассмотрим частный пример. Допустим, обследуется завод, где численность рабочих первого разряда составляет 116 человек, второго разряда—165 человек, третьего разряда—208 человек, четвертого разряда—326 человек, пятого разряда—243 человека и шестого разряда—97 человек. Таблица распределения примет следующий вид (табл. 1):

Т а б л и ц а 1

Квалификация рабочих	1 разряд	2 разряд	3 разряд	4 разряд	5 разряд	6 разряд	Всего
Численность . . .	116	165	208	326	243	97	1155
Частота	0,100	0,142	0,180	0,277	0,210	0,084	≈ 1

Далее допускаем, что, в зависимости от квалификации, численность рабочих, постоянно выписывающих газету, распределяется следующим образом (таблица 2):

Т а б л и ц а 2

Квалификация рабочих	1 разряд	2 разряд	3 разряд	4 разряд	5 разряд	6 разряд	Всего
Численность рабочих, выписывающих газеты	29	25	94	150	160	76	554
% выписывающих	25	27,2	45,2	46,2	65,8	78,3	47,96

Таким образом, мы выяснили, что в генеральной совокупности в действительности рабочие, которые постоянно выписывают газету, составляют 47,96%. Это же число, или, по крайней мере, близкое к нему число, мы должны получить и при выборочном обследовании, если выборка определена правильно. Но если в выборке при прочих равных условиях представительство из первой группы (I разряд) будет больше, а представительство из шестой группы будет меньше (6 разряд), то общий процент (47,96%) будет занижен. Итак, необходимым условием для правильной выборки является пропорциональное представительство из каждой группы.

Однако это условие не является достаточным, так как в ряде случаев не обеспечивается необходимое количество представителей внутри группы. Каким же образом можно определить число единиц, которые будут компетентны в отражении той совокупности (группы), которую они представляют?

На этот вопрос нам может дать ответ следующее рассуждение.

Рассмотрим генеральную совокупность из «N» членов. Допустим, что из «N» членов «K» человек выписывают газету, а «L» человек не выписывают. Тогда:

$$N = K + L.$$

Далее допускаем, что «п» членам генеральной совокупности мы задали вопрос: выписываете ли вы газету? с возможными ответами «да» или «нет». Оказалось, что «к» человек сказали «да» и «l» человек — «нет». Опросив «п» человек, мы можем рассуждать следующим образом.

Если бы процентное соотношение между сказавшими «да» и «нет» из выборочной совокупности «п» соответствовало бы процентному соотношению «да» и «нет» генеральной совокупности «N», то из общего числа «п» опрошенных сказали бы «да» $\frac{nK}{N}$ членов. В действительности «да» сказали «к» членов. Поэтому имеет место отклонение:

$$\left| k - n \frac{K}{N} \right|. \quad (1)$$

Данное отклонение, как уже доказано¹, является величиной порядка:

$$\sigma^2 = \frac{KLn(N-n)}{N^2(N-1)}. \quad (2)$$

¹ В. Л. ван дер Варден, Математическая статистика, М., стр. 52.

Отсюда квадратическое отклонение будет равно:

$$\sigma = \frac{1}{N} \sqrt{\frac{KL\pi(N-\pi)}{N-1}} \quad (2')$$

Доля выписывающих и не выписывающих газету в генеральной совокупности будет выражаться соответственно как:

$$\frac{K}{N} = p \text{ и } \frac{L}{N} = q, \quad (3)$$

где: p —доля выписывающих газету в генеральной совокупности;

q —доля не выписывающих газету в генеральной совокупности.

Отсюда ясно, что $p+q=1$ или $q=1-p$.

Наша задача — определить « p » с наперед заданной ошибкой, то есть указать такие числа p_1 и p_2 , где имело бы место выражение: $p_1 \leq p \leq p_2$ с определенной, наперед заданной вероятностью (доверительной вероятностью). Эту вероятность обозначим через $1-2\beta$. Обратно: вероятность того, что найденная « p » не лежит в указанных пределах, равна « 2β ».

Пусть « g »—произвольное положительное число. Известно, что неравенство

$$\left| k - n \frac{K}{N} \right| \leq g \delta, \quad (4)$$

выполняется с вероятностью

$$1 - 2\beta = \frac{2}{\sqrt{2\pi}} \int_0^{\delta} e^{-\frac{1}{2}t^2} dt. \quad (5)$$

Из выражения (5) видно, что, задавшись значениями « 2β », можно определить « g ». Далее преобразуем формулу (4) следующим образом: в левой части вместо $\frac{K}{N}$ запишем его значение —долю, а в правой части подставим значение « δ » и разделим обе части на « n ». Тогда формула (4) примет вид:

$$|h - p| \leq g \sqrt{\frac{p(1-p)}{n} \cdot \frac{N-\pi}{N-1}}, \quad (6)$$

где

$$h = \frac{K}{n}.$$

Возведя обе части выражения (6) в квадрат и приведя его к виду, удобному для решения, мы получим:

$$p^2 - 2hp + h^2 = c(p - p^2), \quad (7)$$

где

$$c = \frac{g^2}{n} \cdot \frac{N - n}{N - 1}. \quad (8)$$

Решая уравнение (7) относительно «р», мы получим:

$$p_{12} = \frac{(2h + c) \pm \sqrt{4hc(1 - h) + c^2}}{2(1 + c)}. \quad (9)$$

Таким образом, доверительные границы, между которыми предположительно заключено истинное значение «р», определяются следующими формулами:

$$p_1 = \frac{(2h + c) - \sqrt{4hc(1 - h) + c^2}}{2(1 + c)}, \quad (10)$$

$$p_2 = \frac{(2h + c) + \sqrt{4hc(1 - h) + c^2}}{2(1 + c)}. \quad (11)$$

Определим «п», при котором вероятность «р» лежит (с доверительной вероятностью $1 - 2\beta$) в наперед заданных пределах.

Возьмем $p_2 - p_1 = \mu$;

где μ — заданная величина ошибки.

Используя выражения (10), (11), мы получаем следующее выражение:

$$p_2 - p_1 = \frac{\sqrt{4hc(1 - h) + c^2}}{1 + c} = \mu. \quad (12)$$

Определяя отсюда «с», мы получаем квадратное уравнение, решением которого является:

$$c = \frac{2h(1 - h) - \mu^2}{1 - \mu^2} \cdot \left(\sqrt{1 + \frac{\mu^2(1 - \mu^2)}{[2h(1 - h) - \mu^2]}} - 1 \right). \quad (13)$$

Из выражения (8) видно, что $c > 0$, поэтому решение со знаком «минус» перед корнем неприемлемо. Определив «с», мы можем из выражения (8) определить «п»:

$$n = \frac{g^2 N}{c(N - 1) + g^2}. \quad (14)$$

Таким образом, если нам нужно узнать, при каком «п» истинное значение отношения находится внутри трубки, т. е. какая выборка определяет ее с доверительной вероятностью $1 - 2\beta$, то мы должны:

1. Определить $p = \frac{k'}{n'}$, для чего необходимо проделать некоторое число «п» проб.
2. По формуле (13) определить «с».
3. По заданной величине доверительной вероятности « $1 - 2\beta$ » определить «g» (с помощью таблиц).
4. Пользуясь формулой (14), определить величину «п».

Упрощенные формулы

Заметим, что иногда для определения выборки, когда генеральная совокупность достаточно велика, могут быть использованы более простые выражения.

Допустим, нас интересует мнение одного миллиона рабочих, а опрашиваем только тысячу человек. Тогда:

$$\frac{N - n}{N - 1} \approx 1,$$

т. е. тысяча по сравнению с миллионом является величиной малого порядка, а следовательно, формула (8) может быть представлена как:

$$n = \frac{g^2}{c}. \quad (15)$$

Отсюда, если выражение, определяющее «с», меньше единицы, то:

$$\sqrt{1 + \frac{\mu^2(1 - \mu^2)}{[2h(1 - h) - \mu^2]}} \approx 1 + \frac{1}{2} \frac{\mu^2(1 - \mu^2)}{[2h(1 - h) - \mu^2]^2}.$$

Поэтому формулу (13) приближенно можно записать так:

$$c \approx \frac{\mu^2}{4h(1 - h) - 2\mu^2}. \quad (16)$$

Формула для оценок

Из формулы (14) видно, что чем меньше «с», тем больше «п». С другой стороны, если обозначить:

$$x = 4h(1 - h) - 2\mu^2,$$

то будем иметь:

$$c = \frac{1}{2(1 - \mu^2)} \left(-x + \sqrt{x^2 - 4\mu^2(1 - \mu^2)} \right).$$

И отсюда видно, что при $x=0$ «с» принимает наибольшее значение, равное:

$$c = \frac{\mu}{\sqrt{1 - \mu^2}}. \quad (17)$$

А затем, по мере роста «х», «с» монотонно убывает, а «х» тем больше, чем больше $4h(1 - h)$.

Максимальное значение $4h(1 - h)$ достигает при $h = \frac{1}{2}$.

При этом $4h(1 - h) = 1$, а значит $x = 1 - 2\mu^2$.

$$c = \frac{\mu^2}{1 - \mu^2}. \quad (18)$$

или, если пренебречь « μ^2 » по сравнению с единицей, то

$$c \cong \mu^2. \quad (19)$$

В соответствии с этим

$$n = \frac{g^2 N}{\mu^2(N-1) + g^2}. \quad (20)$$

А при $\frac{N-n}{N-1} \approx 1$ будем иметь:

$$n = \frac{g^2}{\mu^2}. \quad (21)$$

Если принять, что « g » = 2,58, $\mu = 0,05$ (5% ошибки), то по формуле (21) мы получим, что:

$$n = \frac{(2,58)^2}{(0,05)^2} = 2662.$$

Другими словами, при сколь угодно увеличении генеральной совокупности, если определена социальная группа, то для получения альтернативного ответа максимальное число опросов может быть 2662 (при допущении 5%-ой ошибки). При более точных определениях числа опроса достаточно в формуле (21) взять соответствующее значение « μ ».

ՆԱԽԱԳԵՍ ՏՐՎԱԾ ՍԽԱԼՈՎ ԸՆՏՐՈՂԱԿԱՆ ԱՄՐՈՂՋՈՒԹՅԱՆ ԾԱՎԱԼԻ ՈՐՈՇՈՒՄԸ

ԷԴ. Ա. ԱՐԳԱՐՅԱՆ

Ա մ փ ո փ ո լ մ

Ընտրողության տեխնիկան օգտագործվում է սոցիոլոգիական ուսումնասիրության առաջնային՝ էմպիրիկ փաստերի հաշվարման և մշակման կարևոր էտապում: Այն հնարավորություն է ընձեռում արագացնել փաստերի հոսքը, կրճատել ուսումնասիրության ժամկետը և ծախսերը:

Հոգվածում հիմնական ուղադրությունը գարծված է ընտրողության ծավալի որոշման մաթեմատիկական ապարատին, որը հնարավորություն է տալիս, մեթոդի էությունից ըխող սխալների և ընտրության ծավալի ցանկալի հարերակցություն գտնել: