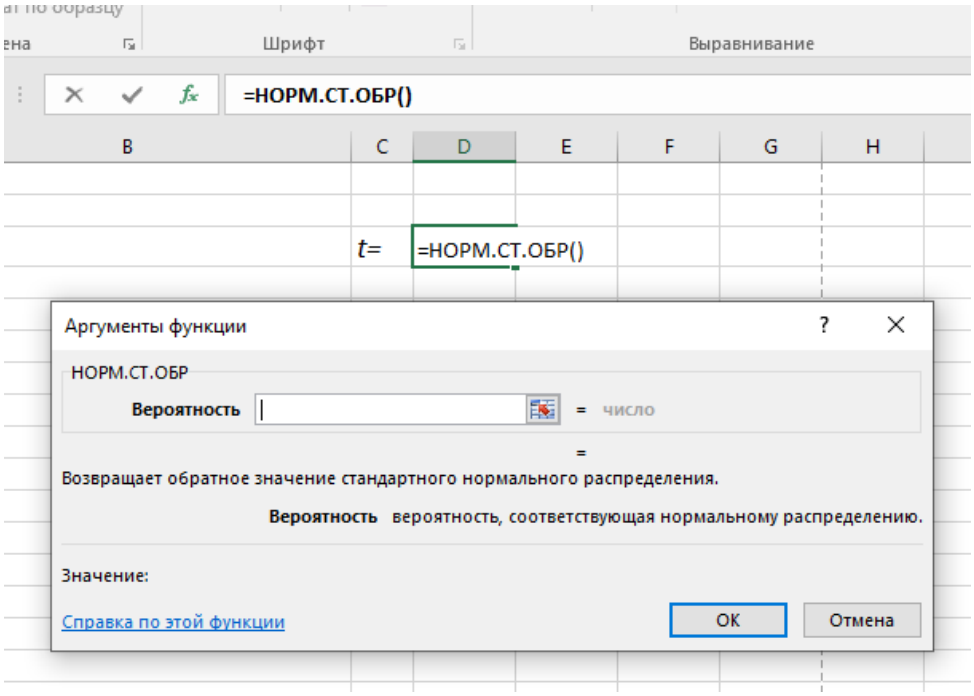
Лабораторная работа.

Определение оптимального объема выборочной совокупности

Цель. Научить определять минимальный объем выборочной совокупности.

Важная задача выборочного наблюдения заключается в определении объема выборки (n), при котором с заданной вероятностью средняя ошибка выборки не превосходит некоторой заранее заданной величины.

Для определения необходимой численности выборки исследователь должен задать уровень точности выборочной совокупности с определенной вероятностью: $\gamma = p/2 = \Phi(t)$ (p – доверительная вероятность, α – уровень значимости, то есть вероятность ошибки I рода, $\Phi(t)$ – интегральная функция Лапласа), которая позволяет найти соответствующее ей значение коэффициента доверия t, используемого в формулах. В Microsoft Excel для расчёта аргумента интегральной функции Лапласа t используются формулы: $p=1-\alpha$, $\Phi=p/2$, t=НОРМ.СТ.ОБР(Φ+0,5).



Величина, которую не должна превосходить средняя ошибка выборки, − это уже известная нам предельная ошибка выборки, обозначаемая Δ.

Необходимый объем случайной **повторной** выборки определяется по формуле:

$$n = \frac{t^2\sigma^2}{\Delta^2} \qquad (1)$$

Необходимый объем случайной **бесповторной** выборки определяется по формуле:

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2} \qquad (2)$$

Из формулы видно, что с увеличением предполагаемой ошибки выборки Δ значительно уменьшается необходимый объем выборки n. Так, увеличение допустимой ошибки выборки в 2 раза уменьшает необходимый ее объем в 4 раза. Необходимая численность выборки прямо пропорциональна дисперсии признака и коэффициенту доверия.

Одним из наиболее важных и в то же время сложных вопросов определения необходимого объема выборки в исследованиях является расчет показателя вариации изучаемого признака (σ). При подготовке выборочного наблюдения у его организаторов часто отсутствуют необходимые для этих вычислений данные. Основой оценки степени колеблемости изучаемого признака служат, как правило, материалы предыдущих обследований. Обращение к ним при отсутствии какой-либо другой информации вполне оправдано. Однако следует иметь в виду, что использование данных прошлых обследований имеет смысл только тогда, когда за прошедший до нового обследования период в генеральной совокупности не произошло значительных изменений.

Часто более точное представление об изучаемой совокупности может дать пробное обследование. По его данным можно рассчитать среднее квадратическое отклонение и дисперсию для последующего обоснования необходимого объема выборки.

Зная примерную величину средней, дисперсию можно найти из соотношения $\sigma \approx \frac{1}{3}\bar{x}$. Если известны x_{max} и x_{min}, то можно определить среднее квадратическое отклонение в соответствии с правилом «трех сигм»: $\sigma = \frac{1}{6}(x_{max} - x_{min})$, так как в нормальном распределении в размахе вариации «укладывается» 6σ ($\bar{x} \pm 3\sigma$).

При расчете n не следует гнаться за большими значениями t и малыми значениями Δ, так как это приведет к увеличению объема выборки, а следовательно, к увеличению затрат средств, труда и времени, вовсе не являющемуся необходимым.

Рассмотрим несколько примеров расчета объема выборки при различных способах отбора.

Пример 1. Для определения средней длины детали следует провести выборочное обследование методом **случайного повторного** отбора. Какое количество деталей надо отобрать, чтобы ошибка выборки не превышала 3 мм с вероятностью 0,997 при среднем квадратическом отклонении 6 мм (ошибка и среднее квадратическое отклонение заданы исходя из технических нормативов).

Решение: при γ=0,997 по функции Лапласа t=3, объем выборки рассчитывается по формуле (1):

$$n = \frac{t^2\sigma^2}{\Delta^2}$$

$$n = \frac{3^2 \cdot 6^2}{3^2} = 36 \text{ деталей.}$$

Пример 2. В микрорайоне города проживает 5000 семей. В порядке **случайной бесповторной** выборки предполагается определить средний размер семьи. Какое количество семей нужно обследовать, если ошибка выборочной средней не должна превышать 0,8 человек с вероятностью 0,954; а среднее квадратическое отклонение, полученное из материалов предыдущих обследований, равно 3 человека.

Решение: при γ=0,954 по функции Лапласа t=2, тогда необходимая численность выборки n равна по формуле (2):

$$n = \frac{t^2\sigma^2 N}{\Delta^2 N + t^2\sigma^2}$$

$$n = \frac{2^2 \cdot 3^2 \cdot 5000}{0{,}8^2 \cdot 5000 + 2^2 \cdot 3^2} = \frac{4 \cdot 9 \cdot 5000}{0{,}64 \cdot 5000 + 4 \cdot 9} = \frac{180000}{3236} = 56 \text{ семей.}$$

Отметим, что существует и более простая формула:

$$n = \frac{t^2 \cdot s^2}{\varepsilon^2}, \qquad (3)$$

где ε – необходимая точность оценки, t – нормированное отклонение, с которым связана та или иная доверительная вероятность p: $t = \frac{x_i - \mu}{\sigma}$. t определяется как аргумент интегральной функции Лапласа по заданной доверительной вероятности при условии: $\Phi(t) = \frac{p}{2}$.

Пример 3. Изучается влияние лечебного препарата на массу мышей (для этого обычно берут две группы лабораторных мышей – опытную и контрольную). После месяца испытаний в опытной группе масса животных варьировалась следующим образом: 80 г, 75 г, 62 г, 70 г, 68 г и 71 г.

Определить с точностью оценки: ε=1 г и надёжностью в 90%: является ли число мышей в опытной группе достаточным для того, чтобы проводимое исследование было достоверным.

Решение. Найдём выборочное среднее и несмещённую выборочную дисперсию.

$$\bar{x} = \frac{80+75+62+70+68+71}{6} = 71 \text{ (г)}.$$

$$S^2 = \left[\frac{80^2+75^2+62^2+70^2+68^2+71^2}{6} - (71)^2\right] \cdot \frac{6}{6-1} = 37{,}6 \text{ (г}^2\text{)}.$$

Тогда значение аргумента функции Лапласа при заданной доверительной вероятности p = 0,9 равно $t \approx 1{,}64$. По формуле (3) имеем:

$$n = \frac{t^2 \cdot s^2}{\varepsilon^2} = \frac{1{,}64^2 \cdot 37{,}6}{1^2} \approx 101{,}13 \approx 102.$$

Таким образом, выборка из 102 мышей обеспечит заданную точность и выбранную надёжность.

Замечание. При определении минимального объёма выборки округление чисел осуществляется не в математическом, а в статистическом смысле.

Задачи для самостоятельного решения

Задача 1. На склад завода поступило 100 ящиков готовых изделий по 80 шт. в каждом. Для установления среднего веса деталей следует провести серийную выборку деталей методом механического отбора так, чтобы с вероятностью 0,954 ошибка выборки не превышала 2 г. На основе предыдущих обследований известно, что дисперсия серийной выборки равна 4. Определить необходимый объем выборки.

Задача 2. Определите, сколько семей необходимо охватить собственно-случайной выборкой для определения доли семей, не имеющих детей с вероятностью 0,954 и предельной ошибкой 2%. Известно, что в регионе проживают 600 семей, и по результатам ранее проведенных обследований доля семей, не имеющих детей, составляет 25%.