ОПРЕДЕЛЕНИЕ МИНИМАЛЬНОГО ОБЪЕМА ВЫБОРКИ

О. А. Бакаева

В данной статье приведены способы нахождения оптимального объема выборки n для нормального закона распределения, распределения Стьюдента, а также биномиального закона в зависимости от известных параметров этих законов распределения.

В науке часто, чтобы определить какую-либо величину, приходится проделывать ряд испытаний. Но бывает так, что и в этом случае истинное значение показателя абсолютно точно измерить не удается, оно получается с определенной долей погрешности. Исходя из формул доверительного интервала для нормального, биномиального распределения и распределения Стьюдента находится минимальное количество экспериментов, необходимое для получения достоверной информации.

В современных условиях цена эксперимента бывает достаточно высокой как в переносном, так и в прямом смысле. Это может быть связано и с использованием дорогостоящего оборудования, и с оплатой труда специалиста, и непосредственно с затратами на сам опытный процесс. Поэтому задача определения минимального количества экспериментов для получения всей необходимой информации в целях ее последующей обработки является очень актуальной. На языке статистики эта задача сводится к определению минимального объема выборки.

Основная часть классической статистической теории предполагает нормальность распределения изучаемой случайной величины. Но на практике в большинстве случаев приходится сталкиваться с распределением, закон которого близок к одному из известных распределений, но далек от нормального. К наиболее употребительным распределениям можно отнести: непосредственно нормальное распределение и распределение Стьюдента, которые являются непрерывными, а также дискретное – биномиальное распределение. В зависимости от закона распределения и вычисляют необходимый объем выборки – n.

Нормальное распределение. Обычно в статистике решается задача определения доверительных интервалов, покрывающих параметр a, с надежностью γ и точностью δ, где a – математическое ожидание нормального распределения.

Пусть параметры распределения таковы: $M(\bar{X}) = a$, $\sigma(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$. Потребуем, чтобы выполнялось соотношение $P(|\bar{X}-a| < \delta) = \gamma$, где γ – заданная надежность, получим $P(|X-a| < \delta) = 2\Phi\Big(\dfrac{\delta}{\sigma}\Big)$, заменив X на $\bar{X}$ и σ на $\sigma(\bar{X}) = \dfrac{\delta}{\sigma}$. Тогда

$$P(|\bar{X}-a| < \delta) = 2\Phi\Big(\frac{\delta\sqrt{n}}{\sigma}\Big) = 2\Phi(t), \qquad (1)$$

где

$$t = \frac{\delta\sqrt{n}}{\sigma}. \qquad (2)$$

Найдя из последнего равенства $\delta = \dfrac{t\sigma}{\sqrt{n}}$, имеем право написать

$$P\Big(|\bar{X}-a| < \frac{t\sigma}{\sqrt{n}}\Big) = 2\Phi(t).$$

Приняв во внимание, что вероятность P задана и равна γ, окончательно имеем (чтобы получить рабочую формулу, выборочную среднюю обозначим за $\bar{x}$)

$$P\Big(\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}\Big) = 2\Phi(t) = \gamma. \qquad (3)$$

Смысл полученного отношения таков: с надежностью γ можно утверждать, что доверительный интервал $\Big(\bar{x} - \dfrac{t\sigma}{\sqrt{n}}, \bar{x} + \dfrac{t\sigma}{\sqrt{n}}\Big)$ покрывает неизвестный параметр a; точность оценки $\delta = t\sigma/\sqrt{n}$ Число t определяется из равенства $2\Phi(t) = \gamma$, или $\Phi(t) = \gamma/2$; по таблице функции Лапласа находят аргумент t, которому соответствует значение функции Лапласа, равное $\gamma/2$ [1].

Если известно математическое ожидание с наперед заданной точностью δ и надежностью γ, то минимальный объем выборки, который обеспечит эту точность, находят по формуле

$$n = \frac{t^2\sigma^2}{\delta^2} \qquad (*)$$

как следствие равенства $\delta = \dfrac{t\sigma}{\sqrt{n}}$.

Учитывая, что характеристиками стандартного нормального распределения являются $a = 0$ и $\sigma = 1$, то формула (1) примет вид:

$$P(|\bar{X}| < \delta) = 2\Phi(\delta\sqrt{n}) = 2\Phi(t), \qquad (4)$$

где

$$t = \delta\sqrt{n}. \qquad (5)$$

Из последнего равенства следует, что минимальный объем выборки будет равен:

$$n = \frac{t^2}{\delta^2}. \qquad (**)$$

Также можно использовать аппроксимацию $t \approx 4,91[\alpha^{0,14} - (1-\alpha)^{0,14}]$. Тогда получается [2]

$$n = 24{,}1081\left\{\frac{\sigma}{\delta}\left[\alpha^{0,14} - (1-\alpha)^{0,14}\right]\right\}^2$$

Как показывает полученная формула, минимальное число опытов прямо пропорционально квадрату значения t, которое находится по табличным значениям функции Лапласа, $\Phi(t) = \gamma/2$, где γ – это надежность. То есть с увеличением надежности минимальное число элементов увеличивается в параболической зависимости. С другой стороны, минимальное число опытов обратно пропорционально точности, с которой измеряется среднее значение признака. С увеличением δ, т. е. с уменьшением точности, число элементов уменьшается, а с уменьшением δ, т. е. с увеличением точности, число элементов, наоборот, увеличивается.

О применимости формул (*) и (**) относительно общего количества экспериментов речь пойдет ниже.

Известно, что при неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Поэтому практически при $n > 30$ можно вместо распределения Стьюдента пользоваться нормальным распределением. Однако важно, что для малых объемов выборок ($n < 30$), в особенности для малых значений n, замена распределения нормальным приводит к грубым ошибкам, а именно к неоправданному сужению доверительного интервала, т. е. к повышению точности оценки. Например, если $n = 5$ и $\gamma = 0{,}99$, то пользуясь распределением Стьюдента, имеем $t_\gamma = 4{,}6$, а используя функцию Лапласа, найдем $t_\gamma = 2{,}58$, т. е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стьюдента. То обстоятельство, что распределение Стьюдента при малой выборке дает широкий доверительный интервал вовсе не свидетельствует о непригодности метода Стьюдента, а объясняется тем, что малая выборка содержит малую информацию об интересующем нас признаке.

Распределение Стьюдента определяется параметром n – объемом выборки (или числом степеней свободы $k = n - 1$) и не зависит от неизвестных параметров a и σ; эта особенность является его большим достоинством.

При достаточно больших значениях n объема выборки выборочная и исправленная дисперсии различаются мало. На практике пользуются исправленной дисперсией, если примерно $n < 30$ (напомним, что именно при небольших размерах выборок и используется распределение Стьюдента, тогда как при $n > 30$ практически любая случайная величина аппроксимируется нормальным распределением).

При неизвестной дисперсии необходимый объем выборки определяется из соотношения

$$\delta = \frac{\epsilon}{\bar{x}} = \frac{t_\alpha s}{\sqrt{n}\bar{x}}, \qquad (6)$$

где t_α – α-квантиль распределения Стьюдента при $f = n$ степенях свободы; s и $\bar{x}$ – выборочные оценки соответственно стандартного отклонения и среднего значения [2].

Необходимые значения $\frac{t_\alpha(n)}{\sqrt{n}}$ рассчитаны и могут быть найдены по таблицам [2, табл. 49].

Определение объема выборки происходит в следующей последовательности. Сначала по заданным величинам $\delta = \dfrac{\epsilon}{\bar{x}}$ и α и предполагаемому значению коэффициента вариации $\nu = \dfrac{s}{\bar{x}}$ находят по таблице значение $\dfrac{t_\alpha(n)}{\sqrt{n}}$ и по нему определяют искомое значение n. Если для найденного объема выборки n выборочное значение окажется больше предполагавшегося, то эксперимент должен быть продолжен.

Замечание. Если $\alpha = 0{,}975$, то, как частный случай, из выражения

$$t_{0,975}(n) = 2\sqrt{\frac{n}{n-2}} \quad (7)$$

следует, что объем выборки

$$n = \left(\frac{2s}{\epsilon}\right)^2 + 2. \quad (8)$$

В этом случае по заданной абсолютной ошибке ϵ и предполагаемому стандартному отклонению s может быть непосредственно определен объем необходимой выборки n.

Биномиальное распределение. Пусть производятся независимые испытания с неизвестной вероятностью p появления события A в каждом испытании. Ставится задача найти доверительный интервал для оценки вероятности, в случае биномиального распределения это можно будет сделать с помощью относительной частоты $p = \dfrac{m}{n}$. Учитывая, что

$$P(|X-a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right), \quad (9)$$

и заменив случайную величину X и ее математическое ожидание a соответственно случайной величиной W и ее математическим ожиданием p, получим приближенное (так как относительная частота распределена приближенно нормально) равенство

$$P(|W-p| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right) = \gamma. \quad (10)$$

Как известно, для биномиального распределения дисперсия находится по формуле $D(W) = \dfrac{pq}{n}$, а среднее квадратическое отклонение как квадратный корень из дисперсии $\sigma = \sqrt{D(W)} = \sqrt{\dfrac{pq}{n}}$, где $q = 1-p$ – вероятность не появления события A, тогда подставив данные выражения в формулу (10), получают:

$$P(|W-p| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sqrt{pq}}\right) = 2\Phi(t) = \gamma, \quad (11)$$

где

$$t = \frac{\delta\sqrt{n}}{\sqrt{pq}}. \quad (12)$$

Следовательно,

$$P\left(|W-p| < t\sqrt{\frac{pq}{n}}\right) = 2\Phi(t) = \gamma. \quad (13)$$

Можно выразить точность $\delta = t\sqrt{\frac{pq}{n}}$, откуда минимальный объем выборки, если вероятность p появления события известна, находится по формуле:

$$n = \frac{\sqrt{t^2pq}}{\delta^2}. \quad (***)$$

где t – значение функции Лапласа. Если вероятность появления события явно не задана, то находим ее из соотношения $p = \frac{m}{n}$, где m – число появления события, а n – число испытаний. Тогда минимальный объем выборки будет

$$n = t^2\left(\frac{m}{n\delta^2} - \frac{m^2}{n^2\delta^2}\right). \quad (****)$$

Если n достаточно велико и вероятность p не очень близка к нулю и к единице, то можно считать, что относительная частота распределена приближенно нормально.

Также можно аппроксимировать практически любое распределение нормальным при достаточном объеме выборки. Об этом свидетельствует и Центральная предельная теорема А. М. Ляпунова. Отсюда следует, что практически все статистические распределения должны приближаться к нормальному распределению как к идеальной предельной форме, если только можно располагать достаточно большим числом наблюдений. То есть, если объем выборки > 30 и случайная величина близки к нормальному распределению, то минимальный размер выборки определяется соотношением $n = \dfrac{t^2\sigma^2}{\delta^2}$. А если объем выборки < 30 и дисперсия неизвестна, то исходя из распределения Стьюдента и табличных значений $\dfrac{t_\alpha(n)}{\sqrt{n}}$, так как при новых условиях формула $(*)$ не гарантирует того, что полученное число экспериментов будет достаточным.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Гмурман В. Е.** Теория вероятностей и математическая статистика : учеб. пособие для студентов вузов / В. Е. Гмурман. – 8-е изд., стер. – М. : Высш. шк., 2002. – 479 с.

2. **Кобзарь А. И.** Прикладная математическая статистика / А. И. Кобзарь. – М. Физматлит, 2006. – 816 с.

Поступила 03.11.10.

О СТРУКТУРЕ ПАКЕТА ПРОБЛЕМНО-ОРИЕНТИРОВАННЫХ ПРОГРАММ, ИСПОЛЬЗУЕМЫХ ПРИ МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ ДИНАМИЧЕСКИХ СИСТЕМ ТРАНСПОРТА*

Н. А. Базеева, Ю. И. Голечков, Е. В. Щенникова

Рассмотрены вопросы математического моделирования транспортных динамических систем. Описаны структура и функциональные возможности соответствующего пакета проблемно-ориентированных программ.

Применение программного обеспечения ПЭВМ для исследования динамических характеристик железнодорожных транспортных средств рассматривалось в работах [1–2; 5] и др. В данной работе представлена структура пакета проблемно-ориентированных программ, предназначенного для математического моделирования транспортных динамических систем более широких классов.

Пусть транспортная динамическая система описывается многомерным матричным дифференциальным уравнением второго порядка

$$A\ddot{x} + B\dot{x} + Cx = Q(t, x, \dot{x}), \quad x \in R^n, \qquad (1)$$

где A, B, C – квадратные матрицы (соответственно матрицы масс, демпфирования и жесткости); $Q(t, x, \dot{x})$ – заданная нелинейная вектор-функция времени, перемещения и скорости (обобщенная возмущающая сила); x – вектор обобщенных координат; R^n – евклидово пространство. Такая динамическая система возникает при описании и изучении колебательных процессов летательных аппаратов в воздушном потоке, колебаний корпусов кораблей и подводных лодок при волнении в открытом море, колебаний элементов и узлов подвижного состава железнодорожного и автомобильного транспорта при движении по неровному пути.

Предложенный пакет содержит набор проблемно-ориентированных программ по математическому моделированию движения и оптимизации динамических параметров железнодорожных и автомобильных транспортных средств, а также программу графической иллюстрации полученных результатов, написанные в математической интегрированной среде *Maple* [3–4]. Здесь же приведены описания, тексты программ и даны указания по их активизации.

* *Работа частично поддержана РФФИ (проект № 10-08-00826-а).*