Расчёт объёма выборки эксперимента в условиях отсутствия нормальности данных

С.С. Акимов, магистрант, ФГБОУ ВПО Оренбургский ГУ

В настоящее время проблемам планирования эксперимента отводится большое значения во многих прикладных науках. Данное обстоятельство объясняется тем, что от правильной постановки эксперимента напрямую зависит конечный результат, и, в случае если эксперимент спланирован неверно, исследователь может получить в итоге

значительно искажённые данные, что сведёт на нет все его усилия.

Вопросом постановки эксперимента занимаются многие отечественные и зарубежные учёные [1–8]. Одним из ключевых вопросов при планировании эксперимента является объём выборки, необходимый для его проведения. Однако абсолютное большинство методов, на которых базируется расчёт объёмов экспериментальной выборки, опирается

на нормальный закон распределения [1]. При этом нет никаких априорных оснований утверждать, что результат эксперимента распределяется строго нормально, более того, зачастую можно найти обратные утверждения [9]. Провести же проверку на нормальность до расчёта объёма выборки чаще всего не представляется возможным, поскольку необходимые для проверки данные могут появиться только после проведения эксперимента [2].

В работах некоторых авторов можно обнаружить попытку обойти данный парадокс, используя более универсальные формулы, не ограниченные законом распределения [4]. Однако проверка адекватности данных методов не проводилась.

Цель исследования – определить наиболее адекватную методику планирования эксперимента в условиях отсутствия априорной информации о виде закона распределения, которому подчиняются результаты эксперимента.

Задачи исследования:

- рассмотреть несколько наиболее распространённых методов расчёта объёма выборки при планировании эксперимента;
- сформировать данные, подчиняющиеся различным законам распределения;
- провести проверку предлагаемых методик на сгенерированных данных и выбрать лучшую из них.

Материалы и методы исследования. Исследование проведено в лаборатории кафедры управления и информатики в технических системах Оренбургского государственного университета. Методы планирования эксперимента брали из трудов отечественных и зарубежных учёных. Для постановки задачи использовали генератор случайных чисел на базе Mathcad 15.

В качестве сравнения выбраны два известных метода вычисления объёма выборки.

Первый метод достаточно полно описан в работе «Определение размера выборки при планировании научного исследования» [10]. Согласно данному методу для количественных признаков необходимо использовать формулу

$$n = \frac{s^2 Z^2}{\Delta^2}, \quad (1)$$

где s^2 – дисперсии признака в группе;

Z – критическое значение нормального стандартного распределения для заданного уровня значимости;

Δ – минимальная (статистически значимая) величина различий, которую необходимо обнаружить.

Данную формулу можно обнаружить и в трудах других авторов [3, 5, 7, 8].

При этом в работах не встречается рекомендаций по определению выборочной дисперсии s^2 .

Второй способ, представленный в работе «Перспективные направления повышения надёжности

вычислительной техники и систем управления» [4], предлагает использование следующей формулы:

$$n = \frac{4}{P(1-P)\epsilon^2}, \quad (2)$$

где P – вероятность события;

ϵ – допустимая погрешность вычислений.

В качестве законов распределения, на которых произведена проверка, выбраны равномерный, нормальный и экспоненциальный. Данные законы являются очень распространёнными, а кроме того, обладают сильно различающимися кривыми распределения, поведение которых, безусловно, влияет на конечный результат.

Для проведения эксперимента на генераторе случайных чисел программы Mathcad 15 были сгенерированы три массива данных ($n = 1000$), подчиняющиеся равномерному (U), нормальному (N) и экспоненциальному (Exp) закону распределения, с такими параметрами, чтобы размах вариации для каждого был примерно одинаков.

Результаты исследования. Полученные параметры распределений для каждого из массивов представлены в таблице 1.

1. Основные выборочные параметры массивов данных, подчиняющиеся различным законам распределения

Параметры	U	N	Exp
Минимум	0,04	0,04	0,01
Максимум	19,99	18,49	19,83
Среднее	9,91	9,89	2,46
Дисперсия	32,81	8,56	5,64
Стандартное отклонение	5,73	2,93	2,38

Как показывают данные таблицы, при достаточных близких значениях минимумов и максимумов распределения данных имеются существенные различия в средних и дисперсии.

Произведём расчёт необходимого количества данных обоими способами.

Вероятность (P) задавалась на уровне 0,95, откуда $Z = 1,96$. Погрешность (Δ и ϵ) возьмём на уровне 0,5.

Поскольку дисперсии для каждого из распределений отличаются, для первого случая рассчитано три результата:

- 1) при равномерном распределении $n = 257,19 \approx 257$;
- 2) при нормальном распределении $n = 67,15 \approx 67$;
- 3) при экспоненциальном распределении $n = 44,25 \approx 44$.

Для второго случая $n = 336,84 \approx 337$.

Сделаем подвыборки из выборок в соответствии с рассчитанным количеством, определим их выборочные параметры и рассчитаем абсолютные отклонения от параметров первоначальных выборок (табл. 2, 3).

2. Основные выборочные параметры массивов данных для подвыборок, рассчитанных при помощи первого метода

Параметр	U_1	$U-U_1$	N_1	$N-N_1$	Exp_1	$Exp-Exp_1$
Минимум	0,15	-0,12	3,74	-3,70	0,01	-0,21
Максимум	19,94	0,05	17,13	1,36	9,33	10,50
Среднее	9,15	0,77	10,17	-0,29	2,36	-3,17
Дисперсия	33,52	-0,71	9,50	-0,94	5,37	-3,64
Стандартное отклонение	5,79	-0,06	3,08	-0,16	2,32	-0,67

3. Основные выборочные параметры массивов данных для подвыборок, рассчитанных при помощи второго метода

Параметр	U_2	$U-U_2$	N_2	$N-N_2$	Exp_2	$Exp-Exp_2$
Минимум	0,04	-0,00	2,36	-2,32	0,01	0,00
Максимум	19,99	0,00	18,49	0,00	19,83	0,00
Среднее	9,43	0,48	9,79	0,10	2,61	-0,15
Дисперсия	33,20	-0,39	8,90	-0,34	5,82	-0,18
Стандартное отклонение	5,76	-0,03	2,98	-0,06	2,41	-0,04

По результатам таблицы видно, что параметры подвыборок несколько отличаются от параметров основных выборок. Причём если для равномерного и нормального закона отличия незначительны, то для экспоненциального распределения параметры имеют значительные расхождения.

При втором способе значения выборочных параметров определены значительно точнее. Из этого можно сделать вывод о том, что для распределений, имеющих отличие от нормального (в частности, значительно большие коэффициенты асимметрии и эксцесса, как экспоненциальное), первый метод определения количества данных для постановки эксперимента неприменим.

Выводы. В исследовании проведено сравнение двух распространённых методов расчёта объёма выборки при планировании эксперимента. Адекватность методов проверяли на равномерном, нормальном и экспоненциальном законах распределения. Определено, что метод, описанный в работе «Перспективные направления повышения надёжности вычислительной техники и систем управления» [4], более универсален и может применяться для данных, имеющих

значительное отличие от нормального закона распределения.

Литература

1. Акимов С.С. Оценка Хилла как ключевая оценка для распознавания тяжело- и легкохвостовых законов распределения вероятности // Научное обозрение. 2014. № 10. С. 349–352.
2. Акимов С.С. Расчёт вероятности дискретности для массива данных // Научное обозрение. 2013. № 6. С. 78–82.
3. Адлер Ю.П., Маркова Е.В., Грановский Ю.В. Планирование эксперимента при поиске оптимальных условий. 2-е изд., перераб. и доп. М.: Наука, 1976. 280 с.
4. Дианов В.Н. Перспективные направления повышения надёжности вычислительной техники и систем управления // Надёжность. 2004. № 3 (10). С. 33–47.
5. Зедгинидзе И.Г. Планирование эксперимента для исследования многокомпонентных систем. М.: Наука, 1976. 390 с.
6. Маркова Е.В., Лисенков А.Н. Планирование эксперимента в условиях неоднородности. М.: Наука, 1973. 220 с.
7. Налимов В.В., Чернова Н.Л. Статистические методы планирования экстремальных экспериментов. М.: Физматгиз, 1965. 340 с.
8. Хикс Ч.Р. Основные принципы планирования эксперимента / Пер. с англ. М.: Мир, 1967. 406 с.
9. Орлов А. И. Типовые ошибки при вхождении в прикладную статистику. [Электронный ресурс] URL: <http://forum.orlovs.pp.ru/viewtopic.php?t=97> (дата обращения 07.05.2015).
10. Койчубеков Б.К., Сорокина М.А., Мхитарян К.Э. Определение размера выборки при планировании научного исследования // Международный журнал прикладных и фундаментальных исследований. 2014. № 4. [Электронный ресурс] URL: <http://cyberleninka.ru/article/n/opredelenie-razmera-vyboroki-pri-planirovanii-nauchnogo-issledovaniya> (дата обращения: 07.05.2015).