Определение объема выборки

Ранее мы рассмотрели методы построения доверительного интервала для математического ожидания генеральной совокупности. В каждом из рассмотренных случаев мы заранее фиксировали объем выборки, не учитывая ширину доверительного интервала. В реальных задачах определить объем выборки довольно сложно. Это зависит от наличия финансовых ресурсов, времени и легкости создания выборки.[1] Например, если нам необходимо оценить среднюю сумму накладных или долю ошибочных накладных в информационной системе компании, сначала следует выяснить, насколько точной должна быть оценка. Иначе говоря, следует задать ошибку выборочного исследования, допускаемую при оценке каждого из параметров. Кроме того, необходимо заранее определить доверительный уровень оценки истинного параметра генеральной совокупности.

Определение объема выборки для оценки математического ожидания

Чтобы определить объем выборки, необходимый для оценки математического ожидания генеральной совокупности, следует учесть величину ошибки выборочного исследования и доверительный уровень. Кроме того, необходима дополнительная информация о величине стандартного отклонения. Для того чтобы вывести формулу, позволяющую вычислить объем выборки, начнем с формулы (1) (о происхождении этой формулы см. Построение доверительного интервала для математического ожидания генеральной совокупности):

$$(1)\quad \bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

где $\bar{X}$ – среднее значение выборки, Z — значение стандартизованной нормально распределенной случайной величины, соответствующее интегральной вероятности, равной $1 - \alpha/2$, σ — стандартное отклонение генеральной совокупности, n – объем выборки

В этой формуле величина, добавляемая и вычитаемая из $\bar{X}$, равна половине длины интервала. Она определяет меру неточности оценки, возникающей вследствие ошибки выборочного исследования, которая обозначается символом e и вычисляется по формуле

$$(2)\quad e = Z\frac{\sigma}{\sqrt{n}}$$

Решив уравнение (2) относительно n, получим:

$$(3)\quad n = \left(\frac{Z\sigma}{e}\right)^2$$

Таким образом, для определения объема выборки необходимо знать три параметра:

1. Требуемый доверительный уровень, который влияет на величину Z, являющуюся критическим значением стандартизованного нормального распределения;[2]
2. Приемлемую ошибку выборочного исследования e;
3. Стандартное отклонение σ.

На практике вычислить эти величины непросто. Как определить доверительный уровень и ошибку выборочного исследования? Обычно ответить на этот вопрос могут лишь эксперты в предметной области (т.е. люди, понимающие смысл оцениваемых величин). Как правило, доверительный уровень равен 95% (в этом случае $Z = 1{,}96$).[3] Если требуется поднять доверительный уровень, обычно выбирают величину, равную 99%. Если можно ограничиться более низким доверительным уровнем, выбирают 90%. Определяя ошибку выборочного исследования, не стоит думать о ее величине (в

---

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 471–476

[2] Для определения размера выборки используется величина Z, а не t, поскольку для вычисления критического значения t размер выборки необходимо знать заранее. В большинстве случаев размеры выборки позволяют хорошо аппроксимировать t-распределение стандартизованным нормальным распределением.

[3] Интервал с доверительным уровнем 95% делится на две равные части. Первая часть лежит слева от математического ожидания генеральной совокупности, а вторая — справа. Значение величины Z, соответствующей вероятности 2,5% (площади 0,025), равно –1,96, а значение величины Z, соответствующей суммарной площади 0,975, равно +1,96. Для расчета удобно воспользоваться функцией Excel Z=НОРМ.СТ.ОБР(p), где p – вероятность, подставляя значения $p_1 = 2{,}5\%$ и $p_2 = 97{,}5\%$

принципе, любая ошибка нежелательна). Следует задать такую ошибку, чтобы полученные результаты допускали разумную интерпретацию.

Кроме доверительного уровня и ошибки выборочного исследования, необходимо знать стандартное отклонение генеральной совокупности. К сожалению, этот параметр почти никогда не известен. В некоторых случаях стандартное отклонение генеральной совокупности можно оценить на основе предшествующих исследований. В других ситуациях эксперт может учесть размах выборки и распределение случайной переменной. Например, если генеральная совокупность имеет нормальное распределение, ее размах приближенно равен 6σ (т.е. ±3σ в окрестности математического ожидания). Следовательно, стандартное отклонение приближенно равно одной шестой части диапазона. Если величину σ невозможно оценить таким способом, необходимо выполнить пилотный проект и вычислить стандартное отклонение по результатам.

Пример 1. Вернемся к задаче об аудиторской проверке. Предположим, что из информационной системы извлечена выборка, состоящая из 100 накладных, заполненных в течение последнего месяца. Компания желает построить интервал, содержащий математическое ожидание генеральной совокупности, доверительный уровень которого равен 95%. Как был определен объем выборки? Следует ли его уточнить?

Допустим, что после консультаций с экспертами, работающими в компании, статистики установили допустимую ошибку выборочного исследования равной ±5 долл., а доверительный уровень — 95%. Результаты предшествующих исследований свидетельствуют, что стандартное отклонение генеральной совокупности приближенно равно 25 долл. Таким образом, e = 5, σ = 25 и Z = 1,96 (что соответствует 95%-ному доверительному уровню). По формуле (3) получаем:

$$n = \left(\frac{1{,}96*25}{5}\right)^2 = 96$$

Следовательно, n = 96. Таким образом, объем выборки, равный 100, был выбран удачно и вполне соответствует требованиям, выдвинутым компанией.

Пример 2. Некая промышленная компания на Среднем Западе производит электрические изоляторы. Если во время работы изолятор выходит из строя, происходит короткое замыкание. Чтобы проверить прочность изолятора, компания проводит испытания, в ходе которых определяется максимальная сила, необходимая для разрушения изолятора. Сила измеряется в фунтах нагрузки, приводящей к разрушению изолятора (рис. 1, столбец А). Предположим, что нам необходимо оценить среднюю силу разрушения изолятора с точностью +25 фунтов при 95%-ном доверительном интервале для этой величины. Данные, полученные в предыдущем исследовании, свидетельствуют, что стандартное отклонение равно 100 фунтов. Определите требуемый объем выборки.

Решение. Итак, e = 25, σ =100, доверительный уровень 95% (т.е. Z = 1,96) (рис. 1).

	A	B	C	D	E	F
1	Сила разрушения изолятора		Параметры			
2	1870		Приемлемая ошибка выборочного исследования	*e*	25	фунтов
3	1728		Стандартное отклонение	σ	100	фунтов
4	1656		Доверительный уровень	(1 - α)	95%	
5	1610		Альфа	α	5%	
6	1634		Критическое значение стандартизованной нормально распределенной случайной величины	Z	1,96	=НОРМ.СТ.ОБР((E14+1)/2)
7	1784					
8	1522		Объем выборки	*n*	61,5	=(E16*E13/E12)^2
9	1696					

Рис. 1. Определение объема выборки

Таким образом, n = 62 (дробные результаты, как правило, округляют с избытком до ближайшего целого).

Определение объема выборки для оценки доли признака в генеральной совокупности

Выше мы рассмотрели способ определения объема выборки для оценки математического ожидания генеральной совокупности. Предположим теперь, что нам необходимо определить долю накладных, не соответствующих правилам, принятым компанией (начальные условия см. пример 1 выше).

Сколько накладных следует извлечь из информационной системы, чтобы построенный интервал имел заданный доверительный уровень? Для ответа на этот вопрос применим тот же подход, что и при определении объема выборки для оценки математического ожидания.

Ошибка выборочного исследования определяется по формуле (2). При оценке доли признака величину *σ* следует заменить на величину $\sqrt{p(1-p)}$. Таким образом, формула для ошибки выборочного исследования принимает следующий вид:

$$(4) \quad e = Z\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Выражая *n* через остальные величины, получаем следующую формулу:

$$(5) \quad n = \left(\frac{Z\sqrt{p(1-p)}}{e}\right)^2$$

Таким образом, для определения объема выборки необходимо знать три параметра:

1. Требуемый доверительный уровень, по которому определяется величина *Z*.
2. Допустимую ошибку выборочного исследования *e*.
3. Истинную долю успехов *p*.

На практике вычислить эти величины нелегко. Если известен доверительный уровень, можно вычислить критическое значение стандартизованного нормального распределения *Z*. Ошибка выборочного исследования *e* определяет точность, с которой оценивается доля успехов в генеральной совокупности. Третий параметр — доля успехов в генеральной совокупности *p* — это именно тот параметр, который нам необходимо оценить. Итак, как оценить диапазон изменения величины *p* по его выборочным значениям?

Существуют два способа. Во-первых, во многих ситуациях для оценки величины *p* можно использовать результаты предыдущих исследований. Во-вторых, если данные о предыдущих исследованиях недоступны, можно попытаться оценить параметр *p* так, чтобы исключить недооценку объема выборки. Обратите внимание на то, что в формуле (5) величина *p(1 – p)* стоит в числителе. Следовательно, необходимо найти максимальное значение этой величины. Очевидно, что оно достигается при *p* = 0,5.

Таким образом, если доля признака в генеральной совокупности *p* заранее неизвестна, для определения объема выборки следует задать *p* = 0,5. В этом случае объем выборки будет переоценен, что приведет к дополнительным затратам на ее создание. Если истинная доля успехов в генеральной совокупности сильно отличается от 0,5, доверительный интервал окажется значительно уже, чем требовалось. Оценка параметра *p* в этом случае будет весьма точной, однако за это придется заплатить дополнительными временны*м*и и финансовыми ресурсами.

Вернемся к задаче об аудиторской проверке. Предположим, аудитор желает построить интервал, содержащий долю ошибочных накладных, доверительный уровень которого равен 95%. Допустимая точность равна ±0,07. Результаты предыдущих проверок свидетельствуют, что доля ошибочных накладных не превышает 0,15. Таким образом, *e* = 0,07, *p* = 0,15 и *Z* = 1,96 (что соответствует 95%-ному доверительному уровню). По формуле (5) получаем:

$$n = \left(\frac{1{,}96\sqrt{0{,}15(1-0{,}15)}}{0{,}07}\right)^2 = 99{,}96$$

Таким образом, объем выборки, равный 100, был выбран совершенно правильно и вполне соответствует требованиям, выдвинутым компанией.

Определение объема выборки, извлекаемой из конечной генеральной совокупности

Для определения объема выборки, извлеченной из конечной генеральной совокупности без возвращения, необходимо использовать поправочный коэффициент. Например, при оценке математического ожидания выборочная ошибка вычисляется по следующей формуле:

$$e = Z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

При оценке доли признака ошибка выборочного исследования равна:

$$e = Z\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Чтобы вычислить объем выборки для оценки математического ожидания или доли признака, применяются формулы:

$$n_0 = \frac{Z^2\sigma^2}{e^2} \text{ и } n_0 = \frac{Z^2 p(1-p)}{e^2}$$

где n_0 — объем выборки без учета поправочного коэффициента для конечной генеральной совокупности. Применение поправочного коэффициента приводит к следующей формуле:

$$n = \frac{n_0 N}{n_0 + (N-1)}$$

Предыдущая заметка Построение доверительного интервала для математического ожидания генеральной совокупности

Следующая заметка

К оглавлению Статистика для менеджеров с использованием Microsoft Excel