**Малые выборки в конкурентной разведке**

Не так давно я прочитал замечательную книгу Дугласа Хаббарда [«Как измерить всё, что угодно. Оценка стоимости нематериального в бизнесе»](http://baguzin.ru/wp/?p=2511). Несколько тем меня заинтересовали, и я решил вернуться к их более подробному изучению позже. Сегодня – одна из таких тем: определение количества элементов генеральной совокупности на основании наблюдения нескольких номеров выборки. Например, имея на руках несколько счетов (накладных) конкурента за определенный период, можно определить количество всех счетов, выписанных им за этот период. По номерам билетов метро (автобуса, электрички) можно оценить число перевозимых пассажиров.

Для начала фрагмент упомянутой книги Хаббарда.

\* \* \*

**Серийная выборка**

В учебниках по статистике метод серийной выборки[[1]](#footnote-1) обычно не освещается. Не обсуждали бы его и мы, называйся наша книга «Как измерить большинство вещей». Но в период Второй мировой войны данный подход широко использовался для сбора разведывательной информации[[2]](#footnote-2). Он может пригодиться и сегодня для решения некоторых бизнес-проблем. В годы Второй мировой разведчики союзников собирали сведения о производстве вражеской военной техники, в том числе немецких танков «Магк V». Сведения о производстве этих тяжелых танков были очень противоречивыми, и руководство разведки союзников не знало, чему верить. В 1943 г. работавшие на союзников статистики разработали метод определения объема производства на основе серийных номеров захваченных танков. Серийные номера следовали друг за другом и включали какую-нибудь дату. Однако по одному серийному номеру еще нельзя было точно сказать, когда был произведен первый танк серии (поскольку серия могла начинаться не с номера 001). Здравый смысл подсказывает, что минимальное месячное производство должно равняться, по крайней мере, разности между последним и первым серийными номерами машин, захваченных в данном месяце. Но нельзя ли узнать из этих номеров что-нибудь еще?

Подойдя к захваченным танкам как к случайной выборке из всей генеральной совокупности, статистики поняли, что могут рассчитать вероятность производства в разных объемах. Например, если в каком-то месяце было выпушено 1000 танков, то вероятность случайно захватить 10 произведенных в этот период машин с серийными номерами, отличающимися друг от друга не более чем на 50 единиц, крайне мала. Разброс номеров 10 танков, случайным образом выбранных из 1000, должен быть больше. Если же в этом месяце было произведено только 80 танков, то захват 10 из них с такими близкими номерами, по крайней мере, возможен.

Таблица демонстрирует данные о выпуске танков «Магк V» по сведениям разведки и расчетам статистиков в сравнении с фактическим производством (сведения из захваченных после войны документов). Сравнение подтверждает эффективность статистического метода, основанного на анализе серийных номеров захваченных машин.

**Таблица. Производство немецких танков «Магк V» в период Второй мировой войны**

|  |  |  |  |
| --- | --- | --- | --- |
| Месяц, год | Оценка разведки | Расчеты статистиков | Фактическое производство\* |
| Июнь, 1940 | 1000 | 169 | 122 |
| Июнь, 1941 | 1550 | 244 | 271 |
| Август, 1942 | 1550 | 327 | 342 |

\* Согласно захваченным после войны документам.

Более того, дать оценку, значительно более точную, чем исходные данные разведки, наверное, можно было по номерам нескольких танков. Рисунок 1 показывает, как по случайной выборке предметов с серийными номерами определяют размер всей генеральной совокупности. Следуя указаниям рисунка, рассмотрим пример, когда число трофеев составляет восемь объектов (которыми могут быть товары конкурирующей фирмы, найденные в мусоре страницы ее отчета и т.д.). Самый большой серийный номер — 100 220, а самый маленький — 100 070, так что в результате этапа 1 мы получаем 150. Результат этапа 2 — около 1,0 (в этой точке кривая верхней границы пересекает вертикальную линию для нашего размера выборки — 8). На этапе 3 мы производим простые вычисления (1 + 1,0) х 150 = 300 и получим значение верхней границы CI. Повторив эти шаги для нижней границы и среднего значения, получаем 90%-ный доверительный интервал 156–300 со средним значением 195 (обратите внимание, что среднее — это не середина диапазона, поскольку распределение асимметрично). Так что статистики могли сделать свои выводы, располагая всего восемью номерами захваченных танков.



Рис. 1. Выборочное обследование серийных номеров

Два предостережения: если захвачены машины одной бронетанковой части, ни одна из них не может считаться отобранной случайно, поскольку у них могут быть близкие номера. Однако обычно это сразу видно по самим номерам. Кроме того, когда на самом деле нумерация серии не совсем последовательная (то есть каждый следующий номер присваивался не следующему танку) и какие-то номера пропущены, данный метод требует определенной модификации. Отметим: необходимо, чтобы распределение используемых номеров легко обнаруживалось. Например, если используются только четные номера или номера, отличающиеся друг от друга на пять, то это должно быть очевидным из выборки.

Как это можно применить в бизнесе? «Серийные номера» (то есть последовательные серии) можно найти в современном мире где угодно. Так, компании бесплатно предоставляют конкурентам информацию о своем объеме производства, просто указывая на товарах серийные номера, которые может увидеть любой покупатель. (Однако, чтобы быть случайной, такая выборочная совокупность должна состоять из товаров, купленных в разных магазинах.) Аналогичным образом несколько страниц из выброшенного отчета конкурента или цифр из квитанции могут многое рассказать об остальных страницах отчета или обо всех квитанциях за данный день. Я вовсе не призываю вас копаться в отбросах, но исследование содержимого мусорных контейнеров нередко позволяет решить интересные задачи по измерению.

\* \* \*

Эта же статистическая проблема подробно описана в Википедии: [German tank problem](http://en.wikipedia.org/wiki/German_tank_problem). Страница доступна только на английском языке. В общем виде задача формулируется следующим образом. Имеется несколько наблюдений номеров, входящих в генеральную совокупность номеров, созданных за какой-то период. Чему равно количество номеров генеральной совокупностиза этот период?



где ***N*** – оценка общего количества элементов генеральной совокупности, ***m*** – самый большой из наблюдаемых номеров (выборочный максимум); ***k*** - количество наблюдаемых номеров (объем выборки).

Дисперсия рассчитывается по формуле:



а для ***k*** << ***N***



\* \* \*

А теперь небольшой эксперимент. Мне нравятся книги, которые издает «Альпина Паблишер» (ранее «Альпина Бизнес Букс»), и у меня довольно много книг этого издательства. Каждой книге присваивается уникальный номер – [ISBN](http://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B6%D0%B4%D1%83%D0%BD%D0%B0%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D1%81%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D0%BD%D1%8B%D0%B9_%D0%BA%D0%BD%D0%B8%D0%B6%D0%BD%D1%8B%D0%B9_%D0%BD%D0%BE%D0%BC%D0%B5%D1%80) (рис. 2).



Рис. 2. Расшифровка уникального кода книги – ISBN: группа *а* (978) – признак товара: книга; *b* (5) – язык книги: русский; *c* (9614) – признак издательства: Альпина Паблишер; *d* (0942) – номер, присвоенный издательством конкретной книге; *e* (0) – контрольная цифра.

Кроме того, в течение последнего года я приобрел несколько книг издательства «Манн, Иванов и Фербер». Так вот, оценкой годового выпуска издательств мы сейчас и займемся. В таблице на рис. 3 представлены данные по году издания и части ISBN, относящейся к группе d – номер, присвоенный издательством конкретной книге. А на рис. 4 представлены вычисления, позволяющие применить методику Хаббарда (см. рис. 1).



Рис. 3. Годы выпуска и части номеров ISBN, относящиеся к конкретному изданию



Рис. 4. Вычисления, позволяющие применить метод Хаббарда, изложенный на рис. 1. Дельта – разность между максимальным и минимальным номером в течение года.

Для первой строки (Альпина Паблишер за 2005 год) получаем (см. инструкции на рис. 1):

1. Вычтите из самого большого самый маленький серийный номер в выборке – 119
2. Найдите размер выборки на горизонтальной оси (5) и соответствующую ему точку на кривой верхней границы
3. Найдите на вертикальной оси значение А, самое близкое к этой точке на кривой (2,0), и добавьте к нему 1; умножьте полученный итог на результат шага 1. Это — верхняя граница 90%-ного интервала для всех предметов, имеющих серийные номера: (2,0 + 1) \* 119 = 360
4. Повторите шаги 2 и 3 для среднего значения (0,5 + 1) \* 119 = 180 и нижней границы (0,08 + 1) \* 119 = 130. Подробности вычисления представлены в Excel-файле



Рис. 5. Оценка числа книг, выпущенных издательствами «Альпина Паблишер» и «Манн, Иванов и Фербер»: точки – средние значения, вертикальные линии простираются от минимума до максимума.

Насколько адекватной является оценка? Если для издательства «Альпина Паблишер» суммировать средние значения за 2005–2011, то получим 2067, в то же время (см. рис. 3) за 2012-й год уже есть один ISBN за номером 1793. То есть, наша оценка средних значений приблизительно на 15% завышена… Много это или мало, при таком простом методе оценки, решать вам!

1. На мой взгляд, либо автор выбрал неудачный термин, либо переводчики что-то начудили. Но… на мой взгляд, термин использован неудачно: имелось ввиду исследование серийных номеров на основе *случайной выборки*, а получилась *серийная выборка*. Вот как определяется последняя, например, [здесь](http://www.kukiani.ru/index.php?page=content&subpage=s&r=9&p=26&s=97): «В *серийной выборке* вместо случайного отбора единиц генеральной совокупности осуществляется отбор групп (серий, гнезд). Внутри отобранных серий производится *сплошное наблюдение*». [↑](#footnote-ref-1)
2. Leo A. Goodman. [SERIAL NUMBER ANALYSIS](http://www.jstor.org/discover/10.2307/2280780?uid=3738936&uid=2129&uid=2&uid=70&uid=4&sid=21100760163031) // Journal of the American Statistical Association, Vol. 47, No. 260 (Dec., 1952), pp. 622-634 [↑](#footnote-ref-2)