Разведочный анализ

В годы Второй мировой войны разведка союзников пыталась определить объем производства тяжелых немецких танков. Сведения из разных источников были противоречивыми, и статистики разработали метод оценки на основе серийных номеров захваченных танков. Чем больше отличаются эти серийные номера, тем больше танков было произведено. Таблица демонстрирует превосходство статистических методов оценки над разведданными:

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

Производство тяжелых немецких танков; * – согласно захваченным после войны документам

Предыдущая глава Оглавление Следующая глава

В статистике термин «разведочный анализ» ввел Джон Тьюки в 1970-х. Суть метода – преобразование данных наблюдений и их наглядное представление, позволяющие выявить внутренние закономерности в данных.

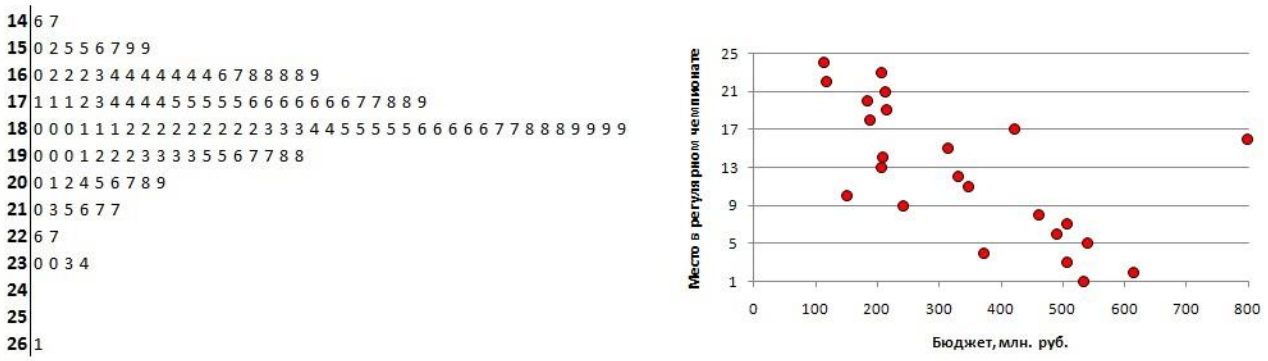
Генеральная совокупность и выборка

Генеральная совокупность – множество всех рассматриваемых объектов, выборка – часть генеральной совокупности, извлекаемая для анализа. Пример, генеральной совокупности – сплошная перепись населения, пример выборки – экзитпол. Цель разведочного анализа – сделать выводы о генеральной совокупности на основе одной или нескольких выборок. Числа, которые характеризуют генеральную совокупность, называют *параметрами* и обозначают греческими буквами. Числа, описывающие выборку, называют *статистиками* и обозначают латинскими буквами.

Основной принцип подготовки выборки – обеспечить чтобы каждая единица генеральной совокупности имела равный шанс попасть в выборку. Выборы президента США 1936 г. продемонстрировали, к чему могут привести прогнозы, основанные на неаккуратно сделанной выборке. Авторитетное издание Литерари дайджест, много лет удивлявшее американцев своими точными прогнозами, предсказывало поражение Рузвельта. Журнал разослал десять миллионов опросных листов владельцам телефонов и автомобилей, и обработал два миллиона ответов. К сожалению, выборка журнала была не беспристрастной. В нее попали преимущественно более обеспеченные слои общества, которые были недовольны «Новым курсом», и жаждали перемен.

Визуализация данных

Для начала данные полезно представить в наглядном виде. Это может быть диаграмма ствол и листья, диаграмма рассеяния или что-то еще. Важно, чтобы визуальный образ содержал все данные.



Визуализация данных выборки: (а) диаграмма «ствол и листья», (б) диаграмма рассеяния

Диаграмма «ствол и листья» была предложена Тьюки еще до эры ПК (Excel не строит эту диаграмму автоматически). На рисунке представлены средние температуры июля в Москве, начиная с 1879 г. Например, число 18,9°C состоит из ствола 18 и листа 9. На диаграмме легко увидеть минимальное (14,6°C) и максимальное (26,1°C) значения. Большинство данных попадают в диапазон 16…20°C, а сами значения образуют распределение близкое к нормальному со средним около 18°C, и довольно

широким хвостом в области больших значений. Также можно увидеть, что значение 26,1°C не просто максимально, а экстремально. Помните жару 2010 г.?

Диаграмма рассеяния на рисунке показывает для чемпионата КХЛ обратную зависимость места, занятого в сезоне 2008–2009, от бюджета клуба.

Сводки данных

Вместо демонстрации всех данных выборку удобно представить несколькими числами – сводками данных, среди которых наиболее известны:

- меры центральной тенденции: среднее, медиана и мода;
- меры рассеяния (разброса): размах, дисперсия, стандартное отклонение.

Несмотря на то, что сводки очень полезны, не забывайте, что они могут скрывать важные подробности. Наверное, самой популярной сводкой является среднее значение. Однако, использование среднего негласно подразумевает, что распределение генеральной совокупности близко к нормальному.

Допустим вы измерили средний рост 100 случайно взятых людей, и получили значение 1,650 м. Добавление к этой выборке самого высокого человека в мире с ростом 2,51 м, увеличит среднее 101 человека менее чем на сантиметр до 1,659 м.

По данным Credit Suisse богатство в мире на одного взрослого человека составляет в среднем $51 600. Если к 100 случайным людям добавить Джеффа Безоса, самого богатого на планете, то среднее для этих 101 человека взлетит до $1,49 млрд.

Всё дело в том, что рост людей на планете подчиняется нормальному распределению, а богатство – нет.

Среднее значение чувствительно к выбросам. Более универсальным показателем меры центральной тенденции является медиана – такое число, что половина из элементов выборки больше него, а другая половина меньше. Длинный хвост практически не влияет на медиану.

Блочная диаграмма или диаграмма ящик с усами

Идеальным графическим представлением сводки данных является блочная диаграмма, введенная также Джоном Тьюки (а эту диаграмму Excel строит автоматически). Продолжая пример со средними температурами июля в Москве, ниже приведена блочная диаграмма за период 1987–2016 гг.

	A	B	C	D	E
1		Июл			
2	1987	16,8			
3	1988	21,6			
4	1989	19,2			
5	1990	17,5			
6	1991	18,1			
7	1992	18,6			
8	1993	17,5			
9	1994	17,6			
10	1995	17,5			
11	1996	18,9			
12	1997	18,7			
13	1998	18,9			
14	1999	21,7			
15	2000	19,3			
16	2001	23			
17	2002	22,6			
18	2003	20,6		Минимум	16,80
19	2004	19		1-й квартиль	18,38
20	2005	19,3		Медиана	18,95
21	2006	18		3-й квартиль	20,90
22	2007	18,8		Максимум	26,10
23	2008	19,1			
24	2009	18,8		2-е наибольшее значение	23,40
25	2010	26,1			
26	2011	23,4			
27	2012	20,9			
28	2013	18,9			
29	2014	21,1			
30	2015	18,3			
31	2016	20,9			

Визуализация статистических данных в Excel с помощью диаграммы ящик с усами

Джон Тьюки. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981. – 696 с. Конспект: http://baguzin.ru/wp/?p=15897

Дуглас Хаббард. Как измерить всё, что угодно. Оценка стоимости нематериального в бизнесе. – М.: Олимп-Бизнес, 2009. – 320 с. Конспект: http://baguzin.ru/wp/?p=2511

Малые выборки в конкурентной разведке: http://baguzin.ru/wp/?p=2880

Левин, Дэвид М., Стефан, Дэвид, Кребиль, Тимоти С., Беренсон, Марк Л. Статистика для менеджеров с использованием Microsoft Excel, 4-е изд. — М.: Издательский дом «Вильямс», 2004. — 1312 с. Конспект: http://baguzin.ru/wp/?p=5285

Сара Бослаф. Статистика для всех. – М.: ДМК Пресс, 2017. – 586. Конспект: http://baguzin.ru/wp/?p=19047

Визуализация статистических данных с помощью диаграммы ящик с усами: http://baguzin.ru/wp/?p=17422