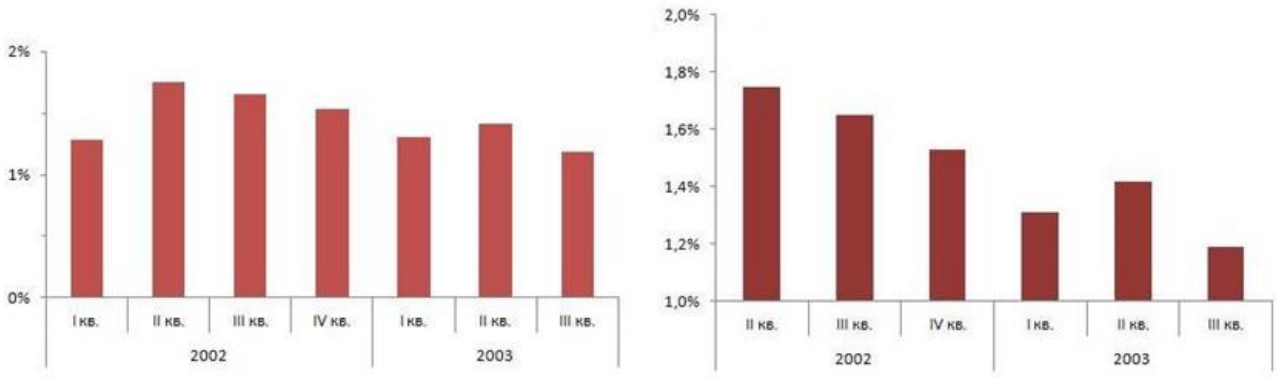
## Ложь, наглая ложь и статистика

Выражение «Существуют три вида лжи: ложь, наглая ложь и статистика» получило известность благодаря Марку Твену, который приписал его премьер-министру Великобритании Бенджамину Дизраэли. Я помню, как однажды, работая в издательстве, хотел продемонстрировать сокращение числа ошибок в рекламных объявлениях. Исходные данные выглядели неплохо (левый рисунок), но самое первое значение явно не вписывалось в обнаруженную тенденцию, и я его… просто отбросил. Теперь график вместо колебаний вокруг значения 1,5%, однозначно демонстрировал успехи))




Доля объявлений, вышедших с ошибками: (а) полные данные; (б) после коррекции

Предыдущая глава Оглавление Следующая глава

## Фактор лжи

Эдвард Тафти предложил использовать *фактор лжи* – показатель правдивость отображения информации. Честная диаграмма должна отражать исходные данные без искажения, то есть иметь фактор лжи равный единице. Чем больше визуальное искажение, тем выше фактор лжи:

$$\text{Фактор лжи} = \frac{\text{размера эффекта, показанного на графике}}{\text{размер эффекта, присущего данным}}$$

$$\text{Размер эффекта} = \frac{|\text{второе значение} - \text{первое значение}|}{\text{первое значение}}$$

К наиболее наглым приемам искажение визуальной информации относятся: отсчет не от нуля и растягивание вертикальной шкалы (см. выше правый рисунок).

## Неполные данные

В 1986 г. газета The Guardian запустила рекламный ролик. Бритоголовый молодой человек убегает от въезжающего в кадр автомобиля. Звучит закадровый голос: «Событие, увиденное с одной стороны, дает одно впечатление». Тут же бегущего показывают в другом ракурсе: он догоняет бизнесмена с портфелем, очевидно собираясь напасть на него или вырвать портфель. «Увиденное с другой стороны, оно создает совсем иное впечатление». Следующая смена плана — и мы видим всю сцену сверху. Прямо на голову бизнесмена летит груз, сорвавшийся с крана. Бритоголовый оттаскивает бизнесмена в сторону, спасая тому жизнь, и груз падает рядом. «Но только увидев картину целиком, вы можете понять, что происходит», — заключает закадровый голос. Ролик «Точка зрения» и поныне приводят в пример как одну из лучших телереклам.

Статистика – не ложь, но статистические данные, хотя они и правдивы, бывают куда более пластичны, чем можно ожидать от сухих цифр.

## Мусор на входе – мусор на выходе

Модели управления рисками, использовавшиеся до финансового кризиса 2008 года, были довольно точными. Концепция стоимости под риском (Value at risk, VaR) позволяла компаниям вычислить величину своего капитала, которая может быть потеряна в случае реализации тех или иных сценариев. Проблема состояла в том, что математические модели были основаны на неверном допущении – нормальном распределении колебаний цен. Какие бы качественные алгоритмы расчета не использовались, результат был заведомо обречен.

Таким образом, для получения корректных статистических выводов необходимо грамотно сделать выборку, и применить подходящую модель.

## Мухи отдельно, котлеты отдельно

Голливудские киностудии для продвижения недавно отснятых лент игнорируют этот принцип, и сравнивают несопоставимое. Большинство рейтингов кассовых сборов опираются на номинальные доллары, игнорируя инфляцию. Как, например, выглядит пятерка самых кассовых (на внутреннем рынке США) фильмов всех времен по состоянию на 2011 год?

1. «Аватар» (2009)
2. «Титаник» (1997)
3. «Темный рыцарь» (2008)
4. «Звездные войны. Эпизод IV» (1977)
5. «Шрек 2» (2004)

Голливуд хотел бы создать у нас впечатление, что каждый его очередной блокбастер грандиознее и прибыльнее предыдущего. Как выглядела бы пятерка самых успешных с коммерческой точки зрения американских фильмов за всю историю существования кино в США с поправкой на инфляцию?

1. «Унесенные ветром» (1939)
2. «Звездные войны. Эпизод IV» (1977)
3. «Звуки музыки» (1965)
4. «Инопланетянин» (1982)
5. «Десять заповедей» (1956)

В реальных величинах «Аватар» оказывается на 14-м месте, а «Шрек 2» опускается на 31-е.

## Проценты

Выраженный в процентах показатель, если он рассчитан на основе небольшого числа случаев, скорее всего, искажает реальную картину. Если же проценты приводятся с одним или двумя знаками после запятой, определенно вас хотят ввести в заблуждение, преподнося результаты как весьма точные.

Еще один прием: суммирование процентов там, где это неуместно. Посмотрите, как убедительно выглядит пример на страницах The New York Times Book Review: «Разрыв между растущими ценами на книги и авторскими заработками, обусловлен ростом производственных издержек. Редакционные затраты за последние десять лет выросли на 10-12%; расходы на материалы – на 6-9%, торговые издержки и расходы на рекламу – на 10%. В общей сложности рост затрат достиг 26–31%». Но если каждая из статей увеличится примерно на 10%, общие затраты должны возрасти на те же 10%!

## Большие данные и … переподгонка

Обработка данных в Excel и специализированных пакетах настолько упростилась, что аналитики, нажав несколько кнопок, могут провести множество сложных статистических тестов. А если использовать нелинейную регрессию, можно описать практически любой набор данных. Серьезные статистики называют это произвольной подгонкой под кривые, или переподгонкой.

Это означает, что ваша модель хорошо аппроксимирует данные и объясняет не только возможные *зависимости*, но и *случайные* отклонения. Вы можете получить модель, которая замечательно описывает ваши исторические данные, но она не подойдет для каких-то других данных, и уж тем более для прогноза. Лучшая защита от переподгонки – построение моделей на основе теории.

Нассим Николас Талеб пишет:

> … исследователь … может обнаружить статистические взаимосвязи – и создать иллюзию результата. В огромных массивах данных большие отклонения – это куда чаще шум или вариации, а не информация или сигнал. Если я работаю с набором из 200 случайных переменных, совершенно не зависящих друг от друга, почти невозможно не обнаружить высокую корреляцию на уровне, скажем, 30%, однако эта корреляция будет абсолютно ложной.

## Доверительный интервал

Часто объектом манипуляций становятся результаты опросов населения. Яркие заголовки могут говорить о преобладании той или иной точки зрения, а вот ошибки и методология выборочного исследования приводятся мелким шрифтом или вовсе опускаются. Чтобы обосновать полученные точечные оценки, необходимо указывать объем выборки и границы доверительного интервала. Если

за первого кандидата отдали голоса 23%, а за второго – 19%, это может выглядеть успехом. Но при небольшой выборке и доверительном интервале в 5% полные итоги выглядят так…

- первый кандидат 23±5%
- второй кандидат 19±5%

… что интервалы в значительной мере перекрываются.

## Управление на основе данных… может заводить в тупик

Управление на основе статистики способно изменить к лучшему поведение людей и организаций. Если вы можете определить долю бракованных изделий, сходящих с производственного конвейера, и эти дефекты обусловлены ситуацией на заводе, то выплата работникам премии за сокращение количества бракованных изделий изменит их поведение. Каждый из нас реагирует на стимулы. Статистика измеряет важные для нас результаты; стимулы подталкивают нас к их улучшению… или к приукрашиванию статистики.

## Литература

Чарльз Уилан. Голая статистика. Самая интересная книга о самой скучной науке. – М.: Манн, Иванов и Фербер, 2016. — 352 с. Конспект: http://baguzin.ru/wp/?p=14151

Дарелл Хафф. Как лгать при помощи статистики — М.: Альпина Паблишер, 2015. – 168 с. Конспект: http://baguzin.ru/wp/?p=12682

Как с помощью диаграммы приукрасить действительность? или о факторе лжи Эдварда Тафти: http://baguzin.ru/wp/?p=2086. К сожалению Эдвард Тафти не дает разрешение на издание его книг за пределами США.

Гектор Макдональд. Правда. Как политики, корпорации и медиа формируют нашу реальность, выставляя факты в выгодном свете. – М.: Альбина Паблишер, 2019. – 368 с. Конспект: http://baguzin.ru/wp/?p=20135

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

Нассим Николас Талеб. Антихрупкость. Как извлечь выгоду из хаоса. – М.: КоЛибри, 2014. – 768 с. Конспект: http://baguzin.ru/wp/?p=7903