**Основы математической статистики**

Математическая статистика решает обратную задачу теории вероятностей — по известным результатам уже прошедших испытаний подбирается подходящая вероятностная модель и уточняются параметры этой модели. Впрочем, одно из главных прикладных значений статистики — возможность составления прогнозов на будущее, что возвращает нас к теории вероятности. Эти две научных дисциплины находятся в неразрывном единстве.

**1. Статистические ряды**

Пусть некоторый класс написал контрольную работу; учитель проверяет тетради и выставляет оценки в журнал:

3 4 2 5 3 5 4 4 4 3 2 4 4 3 5 4 3 5 4 3

Оценки приведены в том порядке, в каком учитель проверял тетради. Такого же вида результат получится, если выписать оценки по порядку их следования в колонке оценок в журнале (очевидно, это алфавитный список фамилий учеников). Сами же оценки при всех этих способах будут расположены совершенно безо всякого порядка.

**Неупорядоченный набор данных, непосредственно полученный в ходе исследования, называется *простым статистическим рядом*.**
**Множество обследованных объектов называется *выборкой*.**

В нашем случае выборка — это данный класс.

**Количество элементов $N$ в выборке называется *объемом выборки*.**
В нашем случае — это количество учеников в классе, найти его можно, пересчитав полученные оценки (не явившийся на контрольную ученик получает 2, так что кол-во оценок совпадает с количеством учеников).

**Уникальное значение $x_i$ измеряемого признака $X$ называется *вариантой*.**
В нашем случае — 4 варианты: 2, 3, 4, 5.

Для анализа данных удобно сперва расположить оценки в порядке возрастания.

**Статистический ряд, упорядоченный по возрастанию, называется *ранжированным*, иначе, *вариационным*.**

2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5

При больших объемах выборок необходимо сгруппировать значения, указав количество элементов выборки для каждой группы.

**Количество $n_i$ элементов в группе называется *частотой*.**

Группировка делается *простым* или *интервальным* способом.

Простая группировка возможна в случае небольшого количества различных вариант (как в нашем случае). В таком случае просто перечисляют все варианты и указывают частоты:

| | | | | |
|---|---|---|---|---|
| X | 2 | 3 | 4 | 5 |
| n | 2 | 6 | 8 | 4 |

Обратим внимание, что сумма частот равна объему выборки

$$\Sigma n_i = N \qquad (1)$$

Интервальная группировка делается путем разбиения всего диапазона значений на небольшое количество (желательно около десяти, оптимальное число дается формулой Стерджеса) равных интервалов, в качестве частоты выступает количество значений, попавших в данный интервал. Интервальная группировка теряет значительную часть информации, которая, однако, признается несущественной и мешающей эффективно анализировать данные.

Часто, особенно при обширных обследованиях, важны не частоты значений, а доли значений в общем объеме результатов.

**Частоты, соотнесенные к объему выборки, называются о*тносительными*:**

| | | | | |
|---|---|---|---|---|
| X | 2 | 3 | 4 | 5 |
| w | 0,1 | 0,3 | 0,4 | 0,2 |

$w_i = n_i/N$ (2)

Относительные частоты легко перевести в проценты и обратно:

| | | | | |
|---|---|---|---|---|
| X | 2 | 3 | 4 | 5 |
| w | 10% | 30% | 40% | 20% |

Очевидно, сумма относительных частот равна 1 (или 100%)

$\Sigma w_i = 1$ (3)

Относительная частота ведет себя во всем аналогично вероятности и является статистическим аналогом вероятности. Различие между ними состоит в том, что вероятность — это теоретическая величина, получаемая путем умозрительных рассуждений относительно опыта, который будет проводиться в некотором будущем, а относительная частота — измеренное на практике значение.

Относительная частота не всегда равна вероятности, более того, при проведении разных серий одинаковых испытаний обязательно будут получаться различные частоты, но они будут устойчиво колебаться вблизи вероятности, и эта близость будет тем более тесной, чем больше провести испытаний (чем большую выборку взять для исследования). Более точно об этом говорит одна из формулировок ***закона больших чисел (ЗБЧ):***

**Относительная частота стремится к вероятности при увеличении объема выборки**

**$\lim w_i = p_i$** **(4)**

**2. Средние значения**

Найти среднее значение измеряемого признака легко: нужно сложить все значения в простом либо ранжированном ряде и разделить на объем выборки.

**$\overline{X} = \Sigma x_i / N$** **(5)**

При использовании группировки вариант в числителе этой дроби возникает многократное сложение одинаковых значений:

2+2+3+3+3+3+3+3+4+4+...,

где количество одинаковых слагаемых есть не что иное, как частота данного значения, поэтому среднее вычисляют следующим образом:

$$\overline{X} = \Sigma\, x_i\, n_i\, /N \qquad (6)$$

Легко заметить, что последнюю формулу можно переписать иначе:

$$\bar{X} = \sum x_i \frac{n_i}{N}$$

или, окончательно,

$$\overline{X} = \Sigma\, x_i\, w_i \qquad (7)$$

где использованы относительные частоты. В нашем примере (используем последнюю формулу)

$\overline{X}$ = 2·0,1 + 3·0,3 + 4·0,4 + 5·0,2 = 3,7

Поскольку вероятность является ожидаемой величиной для относительной частоты, замена в последней формуле относительных частот на вероятности приведет к ожидаемой величине для среднего значения. Эта величина так и называется — математическое ожидание
Итак, математическое ожидание — это теоретически вычисленное до проведения испытаний ожидание среднего значения. Вычисляется полностью аналогично среднему значению, но с заменой относительных частот на вероятности.

**3. Дисперсия**

Дисперсия — это усредненная характеристика разброса значений вокруг среднего значения. Каждое значение отклоняется от среднего на некоторую величину, называемую отклонением:

$$\Delta\, x_i = x_i - \overline{X} \qquad (8)$$

***Дисперсией*** **называется средний квадрат отклонения:**

$$\mathbf{D(X) = \overline{(\Delta x)^2}} \qquad \mathbf{(9)}$$

В последней формуле полезно написать вычисление среднего явно:

$$\mathbf{D(X) = \sum (x_i - \overline{X})^2 w_i} \qquad \mathbf{(10)}$$

Ожидаемое теоретическое значение дисперсии получим, как обычно, заменой частот на вероятности, а среднего — на мат. ожидание::

$$\mathbf{D = \sum (x_i - M)^2 p_i}$$

***Стандартным (среднеквадратичным) отклонением*** **(СКО) называется квадратный корень из дисперсии:**

$$\sigma = \sqrt{D} \qquad \mathbf{(11)}$$

В вычислении теоретического ожидаемого и практически полученного значений СКО, очевидно, нет отличий.

В нашем примере
$D = (3 - 3{,}7)^2 \cdot 0{,}1 + (3 - 3{,}7)^2 \cdot 0{,}3 + (4 - 3{,}7)^2 \cdot 0{,}4 + (5 - 3{,}7)^2 \cdot 0{,}2 = 0{,}81$
$\sigma = 0{,}9$

**4. Мода**

***Модой*** $M_o(X)$ **называется наиболее часто встречающееся значение.**

В нашем примере

$$M_o = 4$$

Очевидно, полученное на практике значение моды подвержено колебаниям вблизи истинного значения. Истинное значение моды можно узнать, только зная вероятности.

Существуют случайные величины, имеющие несколько мод одновременно — ***полимодальные*** и не имеющие моды вообще — ***безмодальные.***

**5. Медиана**

***Медиана*** $M_e(X)$ **— это значение, стоящее в середине вариационного ряда.**

Если число вариант четно, то в середине ряда оказываются два числа, в качестве медианы берут среднее значение этих чисел.

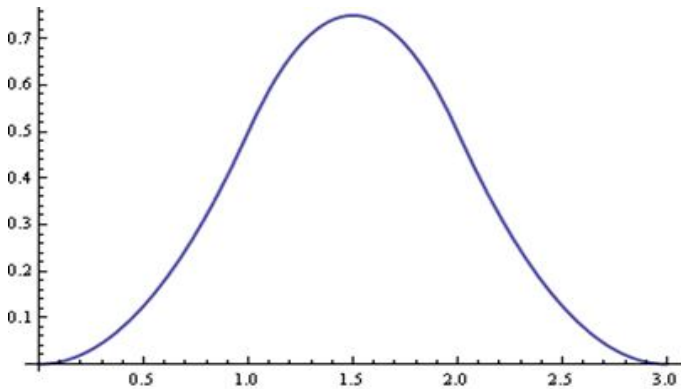
В нашем примере число вариант — 20, в середине стоят числа 4 и 4, значит,

$$M_e = 4$$

Близость медианы и среднего значения говорит о симметричности распределения: в основном значения лежат вблизи среднего, сильных отклонений и в большую, и в меньшую стороны, мало. Большая разница между ними говорит о резких перекосах распределения, например, если случайная величина X — это доходы населения, и выявлена большая разница между медианой и средним, это говорит об очень большой концентрации доходов в руках маленькой группы людей.

**6. Нормальное распределение**

В предыдущем пункте мы затронули вопрос о больших отклонениях от среднего. В большинстве случаев, отклонения редки, и чем отклонения больше, тем реже они встречаются (например, люди намного ниже и намного выше среднего роста встречаются редко, истинные гиганты и карлики — это вообще единичные случаи на Земле). Идеализированной схемой такого распределения является **нормальное распределение (распределение Гаусса).**

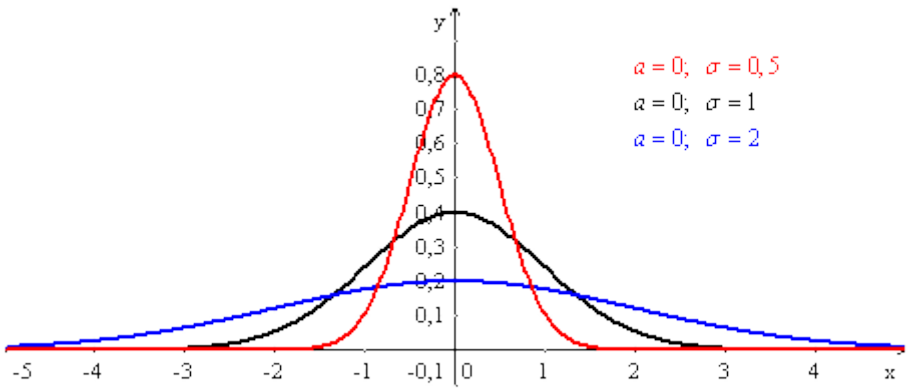


(задание: определите на глаз, чему равны мода, медиана и мат. ожидание распределения, представленного на рисунке).

Математическая формула, задающая нормальное распределение, имеет следующий вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \textbf{(12)}$$

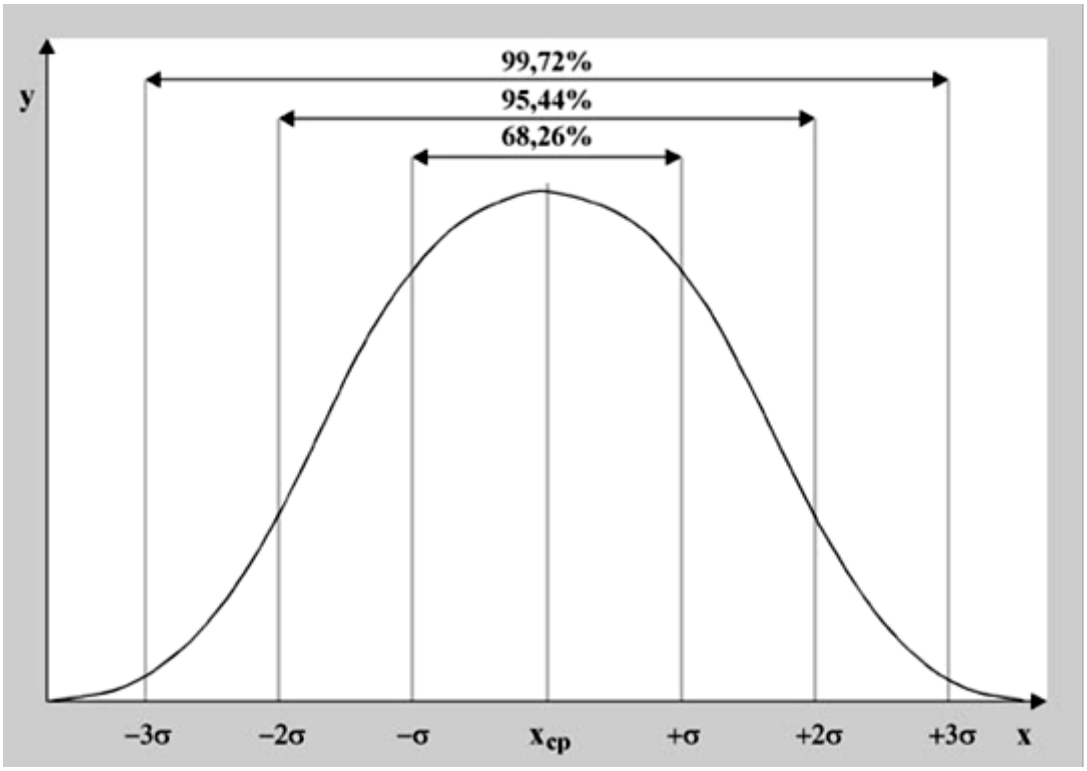
Здесь буквой μ обозначено математическое ожидание, оно же часто в литературе обозначается буквой a, изредка буквой m. Это общепринятое соглашение обозначения математического ожидания при нормальном распределении. σ - среднеквадратичное отклонение. Следующий рисунок демонстрирует, как изменяется распределение в зависимости от величины СКО при одинаковом μ = 0:



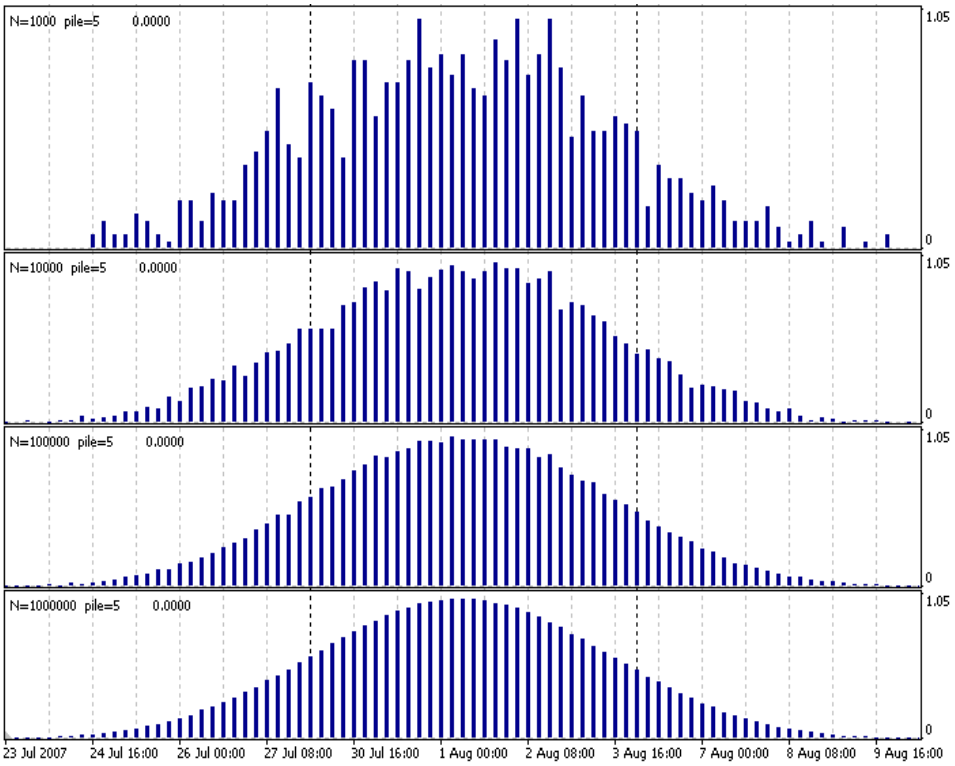

Для нормального распределения верно «правило трех сигм»:

**Вероятность отклонения от среднего, больше чем на $3\sigma$ , равна 0,08%, то есть, практически, такое событие может считаться невозможным.**

Следующий рисунок уточняет распределение вероятности для интервалов по СКО:

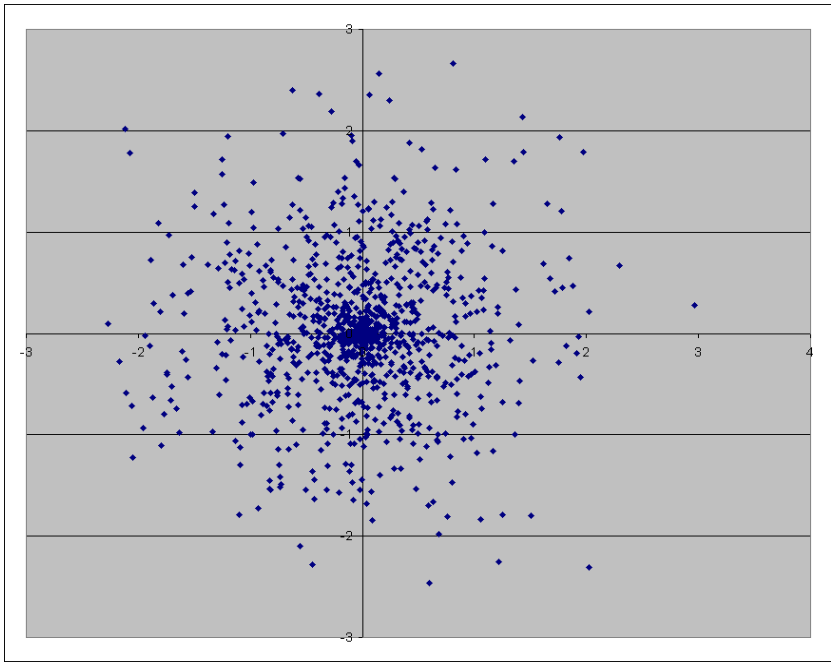



Очевидно, при практическом измерении нормально распределенной величины на идеальную кривую будут накладываться искажения, называемые ***шумами***. На следующем рисунке представлены результаты измерений при разном относительном уровне шумов. Снижение относительного уровня шумов возможно в абсолютном большинстве задач, условием является независимость абсолютного уровня шумов от количества испытаний. Это позволяет снизить относительный уровень шумов за счет увеличения числа испытаний, что, очевидно, является альтернативной формулировкой закона больших чисел.




Нормальное распределение возникает всякий раз, когда на измеряемую величину оказывают влияние множество случайных факторов, каждый из которых вносит небольшой вклад в отклонение от среднего.

Классический пример — стрельба по мишени (из ружья, пушки и т. д.). На отклонение от центра мишени влияют: колебания воздуха во время полета снаряда, коления ствола орудия при выстреле, отклонения массы снаряда от номинала, неидеальности формы снаряда и т. д. - сотни малых отклонений от идеальной картины. Результат их совместного влияния — нормальное (двумерное) распределение отклонений снаряда от центра мишени:

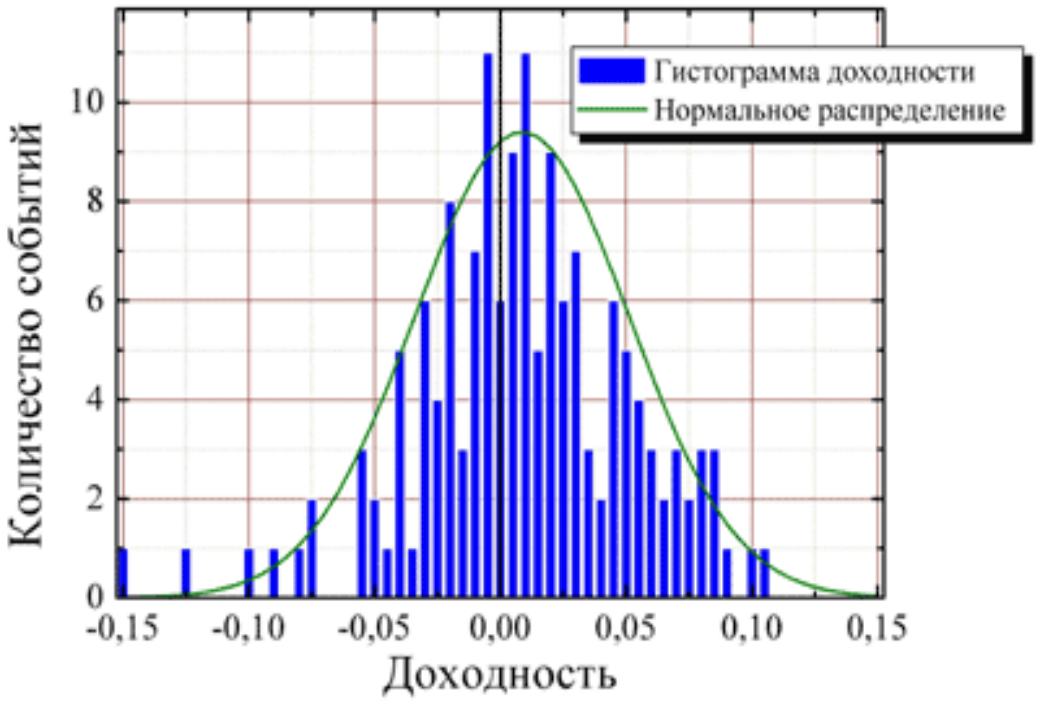


Точная формулировка закономерности возникновения нормального распределения в системах со множеством суммарно действующих малых факторов дается одной из формулировок центральной предельной теоремы (ЦПТ):

**Сумма одинаково распределенных случайных величин имеет распределение, близкое к нормальному.**

**7. Использование теоретических распределений**

Следующий рисунок демонстрирует результаты измерений доходности населения в некоторый период и нормальное распределение, наиболее подходящее к данным результатам:

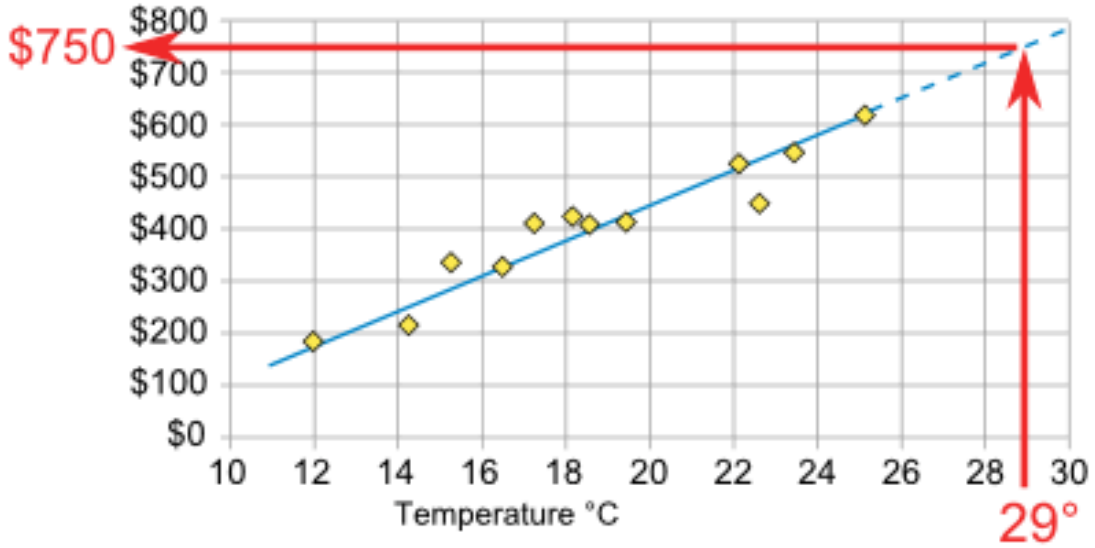



Подбор функции, наиболее подходящей к результатам измерений, называется ***аппроксимацией***. Найденная функция называется ***функцией регрессии*** или ***трендом***.

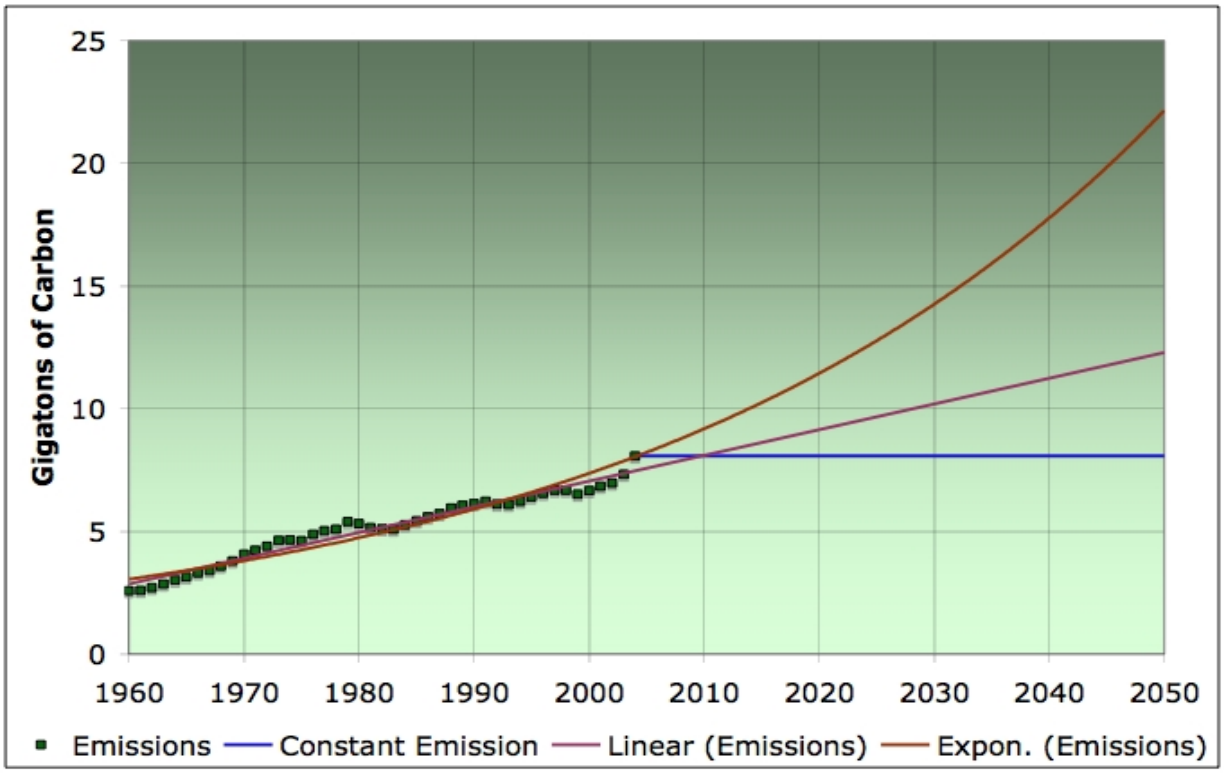
Когда найден тренд, можно осуществить два вида прогнозов:

1) предсказать неизвестное промежуточное значение между двумя известными, такое предсказание называется ***интерполяцией***;
2) предсказать неизвестное значение за границами измеренного интервала (до или после

реальных результатов), такое предсказание называется ***экстраполяцией***.




Следующий рисунок демонстрирует проблему, возникающую из-за использования различных моделей аппроксимации (красная лини — линейная модель, коричневая линия — экспоненциальная модель):




Очевидно, краткосрочные прогнозы достаточно хороши при использовании любой модели, для долгосрочных прогнозов одних только экспериментальных данных недостаточно,

необходимо глубокое теоретическое понимание сущности явления и теоретически обоснованный выбор правильной модели.

**Литература**


1. Гмурман, В.Е. Теория вероятностей и математическая статистика: Учебное пособие для бакалавров / В.Е. Гмурман. - М.: Юрайт, 2013. - 479 с.
2. Горлач, Б.А. Теория вероятностей и математическая статистика: Учебное пособие / Б.А. Горлач. - СПб.: Лань, 2013. - 320 с.
3. Калинина, В.Н. Теория вероятностей и математическая статистика: Учебник для бакалавров / В.Н. Калинина. - М.: Юрайт, 2013. - 472 с.
4. Климов, Г.П. Теория вероятностей и математическая статистика / Г.П. Климов. - М.: МГУ, 2011. - 368 с.
5. Кобзарь, А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. - М.: ФИЗМАТЛИТ, 2012. - 816 с.
6. Колемаев, В.А. Теория вероятностей и математическая статистика: Учебник / В.А. Колемаев, В.Н. Калинина. - М.: КноРус, 2013. - 376 с.
7. Кочетков, Е.С. Теория вероятностей и математическая статистика: Учебник / Е.С. Кочетков, С.О. Смерчинская, В.В. Соколов. - М.: Форум, НИЦ ИНФРА-М, 2013. - 240 с.
8. Кремер, Н.Ш. Теория вероятностей и математическая статистика: Учебник для студентов вузов / Н.Ш. Кремер. - М.: ЮНИТИ-ДАНА, 2012. - 551 с.
9. Кричевец, А.Н. Математическая статистика для психологов: Учебник для студ. учреждений высш. проф. образования / А.Н. Кричевец, А.А. Корнеев, Е.И. Рассказова. - М.: ИЦ Академия, 2012. - 400 с.
10. Лебедев, А.В. Теория вероятностей и математическая статистика: Учебное пособие / Л.Н. Фадеева, А.В. Лебедев; Под ред. проф. Л.Н. Фадеева. - М.: Рид Групп, 2011. - 496 с.