ГЛАВА 2 СТАТИСТИКА, ВЕРОЯТНОСТЬ И ПОМЕХА

Статистика и вероятность используются в ЦОС для описания сигналов и процессов, которые их генерируют их. Например, основное применение ЦОС состоит в понижении уровня помех, шумов и других нежелательных составляющих в получаемых данных. Они могут быть частью измеряемого сигнала, возникать из-за недостатка в системе сбора данных или быть побочным продуктом некоторой операции ЦОС. Статистика и вероятность позволяют измерить и классифицировать эти вредные составляющие. Эта глава описывает основные концепции статистики и вероятности, а также их применение к записанным сигналам.

Терминология сигналов и диаграмм

Сигнал является описанием того, как один параметр относится к другому. Например, самый распространенный сигнал в аналоговой электронике это *напряжение*, которое изменяется *во времени*. Поскольку оба параметра могут быть представлены в непрерывном диапазоне значений, будем называть его *непрерывным сигналом*. В сравнении с ним, прохождение сигнала через аналого-цифровой преобразователь вызывает *квантование* этих двух параметров. Например, представьте, что преобразование было сделано с 12-разрядной точностью и скоростью 1000 выборок в секунду. Диапазон значений напряжения уменьшается до 4096 (2^{12}) возможных двоичных уровней, а время определяется с точностью одной миллисекунды. Сигнал, сформированный из квантованных параметров, называют *дискретным сигналом* или *цифровым сигналом*. В большинстве случаев, непрерывные сигналы существуют в природе, в то время как дискретные сигналы существуют в компьютерах (хотя можно найти исключения для обоих случаев). Существуют сигналы, в которых один параметр непрерывный, а другой дискретный. Поскольку эти смешанные сигналы встречаются довольно редко, они не имеют специального названия, и характер двух параметров должен быть подробно определен.

Рис.2-1. показывает два дискретных сигнала, которые могут быть получены цифровой системой сбора данных. *Вертикальная ось* может представлять напряжение, интенсивность освещенности, звуковое давление или бесконечное число других параметров. Поскольку неизвестно, что они представляют в частном случае, вертикальной оси присваивается общее название: *амплитуда*. Этот параметр имеет и другие названия: *ось Y*, *зависимая переменная*, *диапазон*, *ордината*.

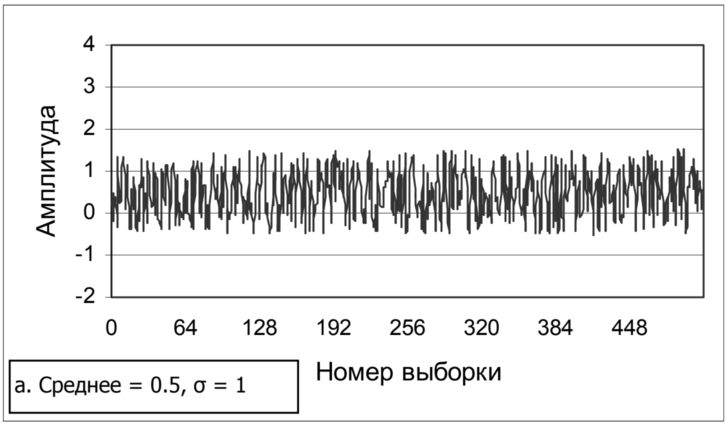
Горизонтальная ось представляет другой параметр сигнала, сопровождаемый такими названиями: *ось X*, *независимая переменная*, *область* и *абсцисса*. Наиболее распространенным параметром, представленным на горизонтальной оси, является *время*. Тем не менее, в специальных приложениях используются и другие параметры. Например, геофизики измеряют плотности пород на одинаково распределенных расстояниях вдоль поверхности земли. В общем случае, параметры на горизонтальной оси называют *номер выборки*. В случае непрерывного сигнала используются другие названия: *время*, *расстояние*, *x* и др.

Два параметры, формирующие сигнал, обычно не взаимозаменяемы. Параметр по оси Y (зависимая переменная) называют *функцией* от параметра по оси X (независимой переменной). Другими словами, независимая переменная описывает как и когда принята каждая выборка, в то время как зависимая переменная представляет собой измеренную величину. Задав определенную величину по оси X, можно всегда найти соответствующую величину по оси Y, обычно обратное не верно.

Широко используемым термином в ЦОС является слово *область*. Например, если сигнал использует время как независимую переменную (параметр по горизонтальной оси), то говорят, что он находится *во временной области*. Другой распространенный сигнал в ЦОС использует в качестве независимой переменной *частоту*. В этом случае говорят, что сигнал находится *в частотной области*. Подобным образом, сигналы, использующие независимой переменной *расстояние*, находятся в *пространственной области* (мерой пространства выступает расстояние). Тип параметра по горизонтальной оси *является* областью сигнала. Что делать в том случае, если ось X определяется чем-то общим, например, номером выборки? Авторы чаще обращаются к таким сигналам, как к сигналам во временной области. Это поясняется тем, что в большинстве случаев выборки берутся через равные интервалы времени, и не существует другой характеристики для спецификации этого параметра.

Хотя сигналы на Рис.2-1 являются дискретными, на диаграмме они отображаются как сплошные линии. Это поясняется тем, что число выборок велико, чтобы быть различимым в виде отдельных точек. На графиках, которые отображают сигнал с менее 100 выборками, обычно показывают отдельные точки. Непрерывные линии могут прорисовываться или нет, в зависимости от того, как автор хочет представить отдельные точки. Например, непрерывная линия может показать, что происходит *между* выборками, или помочь читателю проследить тренд в зашумленных данных. Всегда проверяйте разметку горизонтальной оси, чтобы проверить с какими сигналами идет работа: дискретными или непрерывными. Не полагайтесь на представление графика иллюстратором.

Переменная N широко применяется в ЦОС для представления общего количества выборок сигнала. Например, для сигналов на Рис.2-1, N=512.



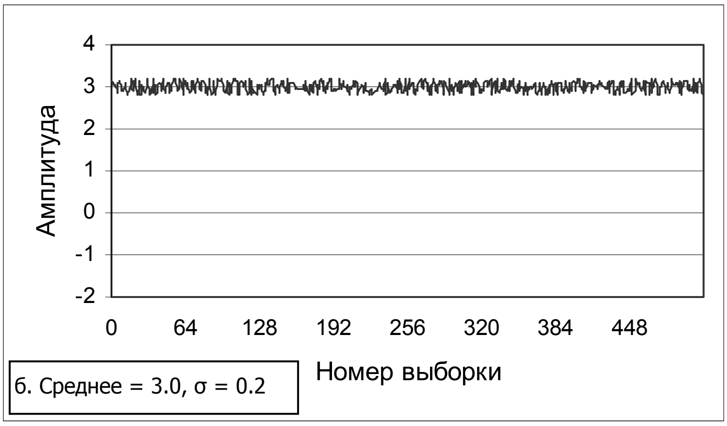


Рис.2-1. Пример двух цифровых сигналов с различными средними значениями и среднеквадратическими отклонениями.

Чтобы сохранить организацию данных, каждой выборке присвоен *номер выборки* или *индекс*. Для обозначения номера выборки используют два способа. В первом случае индекс нумеруется от 1 до N (т.е. от 1 до 512), во втором индекс нумеруется от 0 до N-1 (т.е. от 0 до 511). Математики чаще используют первый принцип обозначения, в то время как в ЦОС больше применяется второй. В этой книге будет использоваться второй способ обозначения (от 0 до N-1).

Среднее значение и стандартное отклонение

Среднее значение (математическое ожидание или ожидаемое значение), обозначаемое через **μ** (строчная греческая буква *мю*), определяет в статистике среднее значение сигнала. Чтобы найти среднее нужно сложить все значения выборок и разделить сумму на количество выборок. В математической форме это выглядит следующим образом:

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} X_i$$

Уравнение 2-1. Расчет среднего значения сигнала. Сигнал содержится в x_0 до x_{N-1}, i - индекс, проходящий через эти значения, μ – среднее значение.

В электронике *среднее значение* часто называется величина **DC** (direct current – постоянный ток). В отличие от него, **AC** (alternating current - переменный ток) привязывается к тому, как сигнал изменяется около среднего значения. Если сигнал повторяет свою форму во времени, как синусоида или меандр, его отклонение может быть описано в виде амплитуды от пика до пика. К сожалению, большинство сигналов не обладают определенным значением от пика до пика, а имеют случайный характер, как показано на Рис.2-1. В этом случае должен использоваться обобщенный метод, называемый *стандартным отклонением*, обозначаемый через **σ** (строчная греческая буква *сигма*).

Выражение $|x_i - \mu|$ показывает как i-тое значение *отклоняется* (отличается) от среднего значения. *Среднее отклонение* сигнала вычисляется через суммирование отклонений всех выборок и делением на их количество, *N*. Заметьте, что используется абсолютное значение отклонения, иначе отрицательные и положительные величины дадут среднее значение около нуля. Хотя среднее отклонение является удобным и простым, оно не используется в статистике, поскольку не соответствует физике работы сигналов. В большинстве случаев, важным параметром является не *отклонение* от среднего значения, а *мощность* отклонения. К примеру, когда случайные сигналы объединяются в электрической схеме, результирующий шум равен суммарной *мощности* индивидуальных сигналов, а не их суммарной *амплитуде*.

Стандартное (среднеквадратичное) отклонение похоже на *среднее отклонение*, за исключением того, что усредняется мощность, а не амплитуда. Это достигается возведением в квадрат каждого отклонения перед суммированием. В завершение операции проводится извлечение квадратного корня. В математической форме, стандартное отклонение рассчитывается как:

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{N-1} (X_i - \mu)^2$$

Уравнение 2-2. Расчет стандартного отклонения сигнала. Сигнал содержится в x_i, μ – среднее значение, σ – стандартное отклонение.

Заметьте, что усреднение проводится делением на N-1, вместо N. Это является особенностью этого выражения, которое будет обсуждаться позже. Термин σ^2 часто встречается в статистике и называется *дисперсия*. Стандартное отклонение является мерой того, как сигнал отличается от среднего значения. Дисперсия представляет мощность этого отклонения. Другим термином, часто используемым в электронике, является среднеквадратичное значение (rms value – root-mean-square value). По определению стандартное отклонение измеряет только AC порцию сигнала, в то время как среднеквадратичное значение измеряет как AC, так DC компоненты. Если сигнал не содержит DC компоненты, его среднеквадратическое значение идентично стандартному отклонению. Рис.2-2. показывает связь между стандартным отклонением и значением от пика до пика некоторых распространенных сигналов.

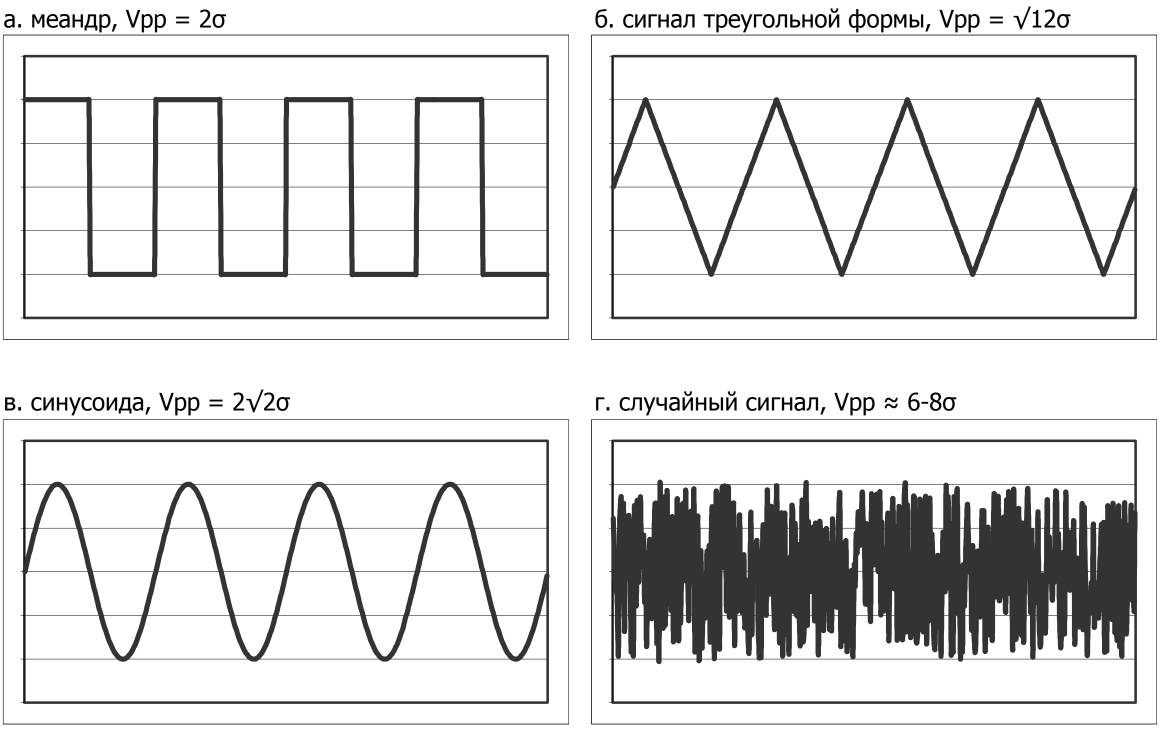


Рис.2-2. Отношение удвоенной амплитуды к стандартному отклонению для некоторых распространенных сигналов.

Таблица 2-1 содержит компьютерную программу для расчета среднего значения и стандартного отклонения с использованием уравнений 2-1 и 2-2. Программы в этой книге предназначены для демонстрации алгоритмов и в большинстве случаев являются прямолинейными. Если программная логика становится более ясной, хорошим стилем программирования пренебрегают. Например, используется упрощенная версия языка BASIC, включены номера строк, из управляющих структур применяется только цикл FOR…NEXT, отсутствуют операторы ввода/вывода и т.п. Поэтому воспринимайте эти программы как альтернативный способ понимания приведенных уравнений ЦОС.

```
100 'ВЫЧИСЛЕНИЕ СРЕДНЕГО ЗНАЧЕНИЯ И СТАНДАРТНОГО ОТКЛОНЕНИЯ
110 '
120 DIM X[511]              'сигнал содержится в X[0] ... X[511]
130 N% = 512                'N% - число точек в сигнале
140 '
150 GOSUB XXXX              'некоторая подпрограмма загрузки сигнала в X[]
160 '
170 MEAN = 0                'найти среднее значения из уравнения 2-1
180 FOR I% = 0 TO N%-1
190     MEAN = MEAN + X[I%]
200 NEXT I%
210 MEAN = MEAN/N%
220 '
230 VARIANCE = 0            'найти стандартное отклонение из уравнения 2-2
240 FOR I% = 0 TO N%-1
250     VARIANCE = VARIANCE + ( X[I%] - MEAN )^2
260 NEXT I%
270 VARIANCE = VARIANCE/(N%-1)
280 SD = SQR(VARIANCE)
290 '
300 PRINT MEAN SD           'вывести результаты
310 '
320 END
```

Таблица 2-1.

В BASICе символ «%» после переменной обозначается целое число. Все другие переменные являются числами с плавающей запятой. Глава 4 детально обсуждает эти типы переменных.

Приведенный метод вычисления среднего значения и стандартного отклонения удовлетворяет большинству приложений, однако, он имеет два ограничения. Во-первых, среднее значение гораздо больше, чем значение стандартного отклонения. В уравнении 2-2 производится вычитание двух чисел, которые очень близки по значению, что может привести к ошибкам округления в расчетах. Во-вторых, часто предпочтительно пересчитать среднее значение и стандартное отклонение для новой выборки сигнала. Такой тип вычисления мы будем называть *динамическая статистика*. При использовании уравнений 2-1 и 2-2 для динамичной статистики необходимо для каждого расчета вводить все выборки сигнала, что приводит к неэффективному использованию мощности и памяти компьютера. Для решения этой проблемы можно преобразовать выражения 2-1 и 2-2 для альтернативного вычисления стандартного отклонения:

$$\sigma^2 = \frac{1}{N-1}\left[\sum_{i=0}^{N-1} x_i^{\,2} - \frac{1}{N}\left(\sum_{i=0}^{N-1} x_i\right)^2\right]$$

или в упрощенной форме:

$$\sigma^2 = \frac{1}{N-1}\left[\textit{сумма\_квадратов} - \frac{\textit{сумма}^2}{N}\right]$$

Уравнение 2-3. Расчет стандартного отклонения для динамической статистики.

При движении по сигналу динамический итог содержится в трех параметрах: (1) количество обработанных выборок, (2) сумма этих выборок, (3) сумма квадратов выборок. Среднее значение и среднеквадратичное отклонение могут быть рассчитаны с использованием этих трех параметров после любого количества обработанных выборок. Таблица 2-2 приводит программу, которая выводит значение среднего и стандартного отклонения после каждой обработанной точки сигнала.

```
100 'ДИНАМИЧНАЯ СТАТИСТИКА СРЕДНЕГО И СТАНДАРТНОГО ОТКЛОНЕНИЯ
110 '
120 DIM X[511]              'сигнал содержится в X[0] ... X[511]
130 '
140 GOSUB XXXX              'некоторая подпрограмма загрузки сигнала в X[]
150 '
160 N% = 0                  'обнуление всех трех параметров
170 SUM = 0
180 SUMSQUARES = 0
190 '
200 FOR I% = 0 TO 511       'цикл для каждой выборки сигнала
210 '
220     N% = N%+1           'обновление параметров
230     SUM = SUM + X(I%)
240     SUMSQUARES = SUMSQUARES + X(I%)^2
250 '
260     MEAN = SUM/N%       'расчет среднего значения и отклонения
270     VARIANCE = (SUMSQUARES - SUM^2/N%) / (N%-1)
280     SD = SQR(VARIANCE)
290 '
300     PRINT MEAN SD       'вывод результатов
310 '
320 NEXT I%
330 '
340 END
```

Таблица 2-2.

Перед завершением дискуссии о среднем значении и стандартном отклонении, следует упомянуть еще два термина. В некоторых случаях *среднее значение* описывает, что было измерено, в то время как *стандартное отклонение* представляет информацию о шумах и других помехах. При этом среднеквадратичное значение важно не само по себе, а только в *сравнении* со средним значением. Это обуславливает появление термина *отношение сигнал/шум* (SNR – signal-to-noise ratio), значение которого равно отношению среднего значения к стандартному отклонению. Также используется другой термин – *коэффициент изменчивости (вариации)*. Он определяется как стандартное отклонение, деленное на среднее значение и умноженное на 100%. Например, сигнал (или другая группа измеренных величин) с коэффициентом CV 2% имеет отношение сигнал/шум, равное 50. Лучшие значения данных *повышают* отношение сигнал/шум SNR и *понижают* коэффициент изменчивости CV.

Сигнал и основной процесс

Статистика – это наука об интерпретации цифровых данных, таких как полученный. В сравнении с ней, *вероятность* используется в ЦОС для понимания *процессов*, которые генерируют сигналы. Хотя эти понятия тесно связаны, различие между *полученным сигналом* и *основным* (лежащим в основе) *процессом* является ключевым во многих методах ЦОС.

Например, представьте сигнал из 1000 точек, полученный в результате выбрасывания монеты. Если монета выпадает орлом, соответствующей выборке присваивается единичное значение. При выпадении решки, соответственно, нулевое. *Процесс*, который создает этот сигнал, имеет среднее значение 0.5, определяемое относительной вероятностью получения каждого результата: 50% орел и 50% решка. Однако, маловероятно, что настоящий сигнал из 1000 точек будет иметь среднее значение 0.5. Случайность приводит к тому, что каждый раз при генерации сигнала число единиц и нулей будет различным. *Вероятность* основного процесса постоянна, но *статистика* полученного сигнала будет меняться при каждом повторном эксперименте. Эта случайная нерегулярность в фактических данных называется: *статистическая вариация*, *статистическая флуктуация* и *флуктуационный шум*.

Это создает некоторые трудности. Когда вы видите термины: *среднее значение* и *стандартное отклонение*, как вы поймете к чему обращается автор, к статистике фактического сигнала или к основному процессу, который создал сигнал? К сожалению, это можно понять только из содержания текста. Но не все термины в статистике и вероятности совпадают. Например, термины *гистограмма* и *функция вероятностной меры* обращаются к схожим понятиям, которым даны разные названия.

Вернемся к уравнению 2-2, вычисляющему стандартное отклонение. Как было замечено ранее, в этом выражении при расчете среднего значения квадратных отклонений происходит деление на N-1, а не на N. Для того чтобы понять, почему так происходит, представьте, что вы хотите найти среднее значение и стандартное отклонение некоторого *процесса*, который генерирует сигнал. Для этого вы получаете сигнал из N точек и рассчитываете среднее значение сигнала по уравнению 2-1. Вы можете использовать его для *оценки* среднего значения основного процесса, однако, существует ошибка, вызываемая статистической вариацией. В частности, для случайных сигналов, типичная ошибка между средним значением N точек и средним значением основного процесса составляет:

$$\textit{Типичная ошибка} = \frac{\sigma}{N^{1/2}}$$

Уравнение 2-4. Типичная ошибка в расчете среднего значения основного процесса при конечном размере выборки N. σ – стандартное отклонение.

Если N мало, статистический вариация рассчитанного среднего значения будет большой. Другими словами, не существует достаточного количества данных, чтобы правильно охарактеризовать процесс. При увеличении размера выборки N, будет получена меньшая ошибка. В теории вероятностей, *строгий закон больших чисел*, гарантирует, что ошибка становится нулевой, когда N достигает бесконечности.

Следующим шагом, мы хотели бы рассчитать стандартное отклонение полученного сигнала и оценить стандартное отклонение основного процесса. Здесь заключается проблема. Перед тем как рассчитать стандартное отклонение по уравнению 2-2, необходимо знать среднее значение. Однако, среднее значение основного процесса неизвестно, а существует среднее значение сигнала с выборкой размера N, *которое содержит ошибку вследствие статистической вариации*. Эта ошибка стремиться уменьшить вычисляемой значение стандартного отклонения. Для ее компенсации N заменяется на N-1. Если N большое, разница будет несущественной. Если N мало, замена обеспечит более точное значение стандартного отклонения основного процесса. Другими словами, уравнение 2-2 является *оценкой* стандартного отклонения *основного процесса*. При делении на N мы определим стандартное отклонение *полученного сигнала*.

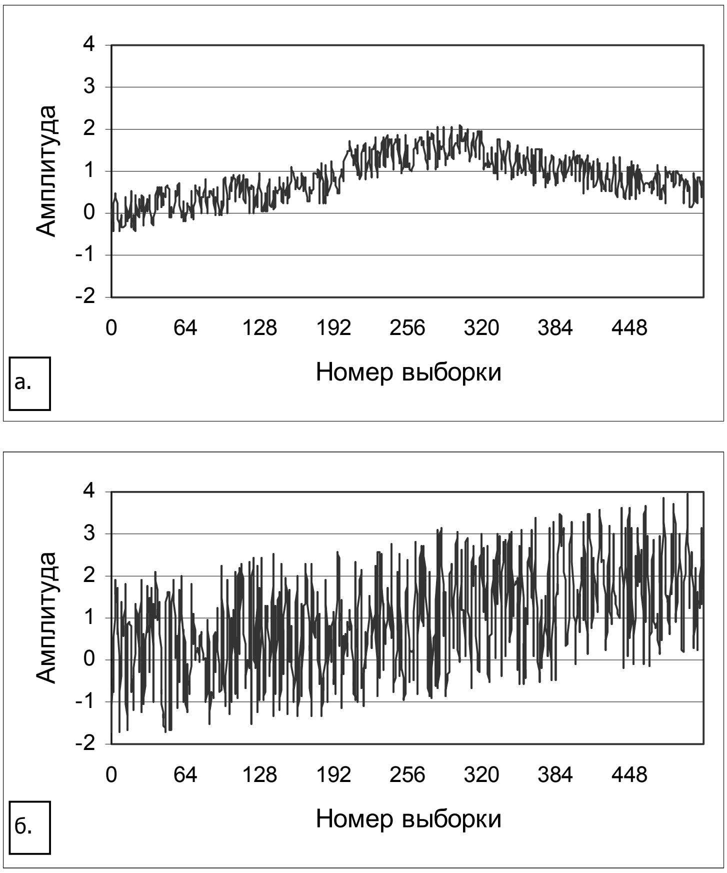
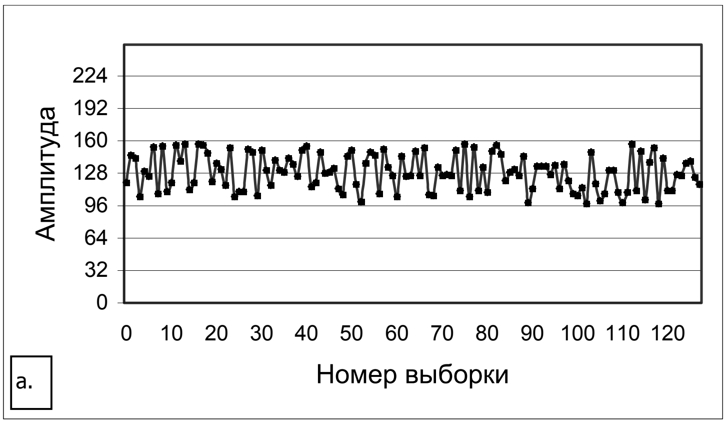


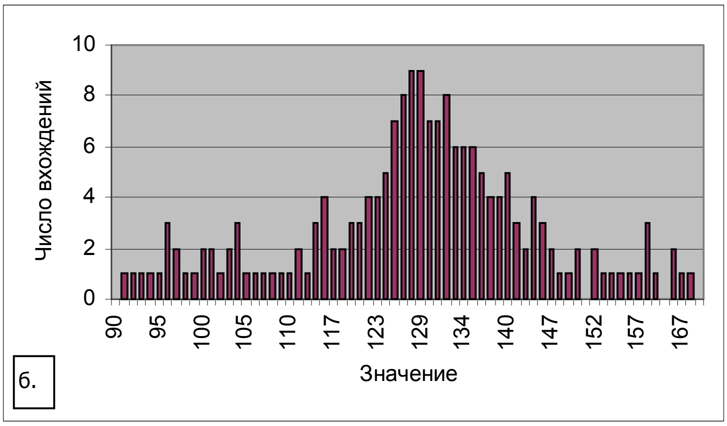
Рис.2-3. Пример сигналов, генерируемых нестационарными процессами. На (а) изменяются и среднее значение, и стандартное отклонение. На (б) стандартное отклонение постоянно, изменяется только среднее значение.

Иллюстрацией этих идей являются сигналы на Рис.2-3. Посмотрите на них и задайте вопрос: являются ли вариации этих сигналов результатом статистического шума или изменения основного процесса? Легко понять, что такие изменения слишком велики для простой случайности, и должны быть отнесены к основному процессу. Процессы, которые меняют свои характеристики, называются *нестационарными*. В сравнении с ним, сигналы, показанные ранее на Рис.2-1, были созданы стационарными процессами. Рис.2-3б показывает основную проблему нестационарных сигналов: медленно меняющееся *среднее значение* влияет на расчет *стандартного отклонения*. Эта ошибка может быть устранена разбиением сигнала на небольшие части и индивидуальным расчетом статистики для каждой части. Если необходимо, стандартное отклонение для каждой секции может быть усреднено для получения единственного значения.

Гистограмма, функция плотности вероятности (PDF) и функция вероятностной меры (PMF)

Предположим, что мы подключили 8-битный аналого-цифровой преобразователь к компьютеру и собрали 256000 выборок некоторого сигнала. Пример на Рис.2-4(а) показывает 128 точек сигнала, которые могли быть частью полученных данных. Значение каждой выборки может изменяться от 0 до 255. *Гистограмма* показывает *число выборок* в сигнале со всеми *возможными значениями*. Рис(б) показывает гистограмму для 128 выборок. Например, 2 точки имеют значение 110, 8 имеют значение 131, 0 точек имеет значение 170 и т.д. Мы представим гистограмму через H_I, где i является индексом, который лежит в диапазоне от 0 до M-1, где M равняется числу возможных значений. Рис(в) показывает гистограмму, для составления которой использовались все 256 тысяч точек сигнала. Как видно из рисунка, чем больше число используемых выборок, тем более сглаженной выглядит гистограмма. Как и в случае со средним значением, статистический шум (грубость) гистограммы обратно пропорционально квадрату числа используемых точек.





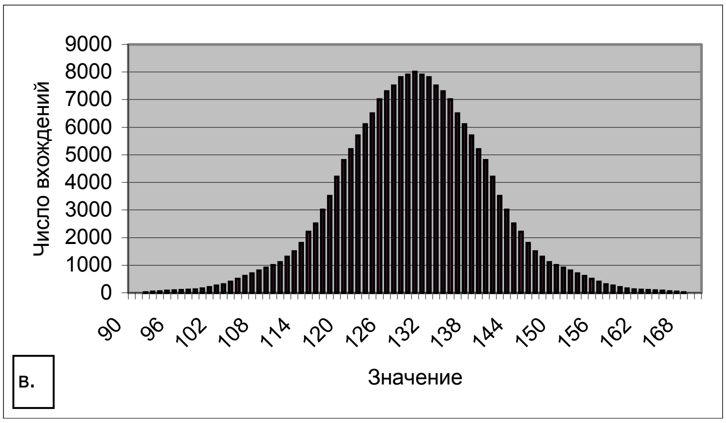


Рис.2-4. Примеры гистограмм.

Сумма всех значений гистограммы должно быть равно числу точек сигнала:

$$N = \sum_{i=0}^{M-1} H_i$$

Уравнение 2-5. Сумма значений гистограммы равна числу точек сигнала. Hi – гистограмма, N – число точек сигнала, M – число точек гистограммы.

Гистограмма может использоваться для эффективного расчета среднего значения и стандартного отклонения большого набора данных. Это особенно важно для *изображений*, которые могут содержать миллионы выборок. Гистограмма группирует выборки с одинаковым значением. Это позволяет рассчитать статистику при работе с несколькими группами, а не с огромным числом отдельных выборок. Среднее значение и стандартное отклонение могут быть вычислены из гистограммы с использованием следующих выражений:

$$\mu = \frac{1}{N} \sum_{i=0}^{M-1} iH_i$$

Уравнение 2-6. Вычисление среднего значения из гистограммы.

$$\sigma^2 = \frac{1}{N-1} \sum_{i=0}^{M-1} (i-\mu)^2 H_i$$

Уравнение 2-7. Вычисление стандартного отклонения из гистограммы.

Таблица 2-3 содержит программу для расчета гистограммы, среднего значения и стандартного отклонения с использованием приведенных выражений.

```
100 'ВЫЧИСЛЕНИЕ ГИСТОГРАММЫ, СРЕДНЕГО ЗНАЧЕНИЯ И СТАНДАРТНОГО ОТКЛОНЕНИЯ
110 '
120 DIM X%[25000]          'X%[0] - X%[25000] содержит сигнал
130 DIM H%[255]            'H%[0] - H%[255] содержит гистограмму
140 N% = 25001             'число точек в сигнале
150 '
160 FOR I% = 0 TO 255      'обнуление гистограммы
170   H%[I%] = 0
180 NEXT I%
190 '
200 GOSUB XXXX             'некоторая подпрограмма загрузки сигнала в X%[ ]
210 '
220 FOR I% = 0 TO 25000    'расчет гистограммы для 25001 точек
230   H%[ X%[I%] ] = H%[ X%[I%] ] + 1
240 NEXT I%
250 '
260 MEAN = 0               'расчет среднего значения по уравнению 2-6
270 FOR I% = 0 TO 255
280   MEAN = MEAN + I% * H%[I%]
290 NEXT I%
300 MEAN = MEAN / N%
310 '
320 VARIANCE = 0           'расчет стандартного отклонения по уравнению 2-7
330 FOR I% = 0 TO 255
340   VARIANCE = VARIANCE + H[I%] * (I%-MEAN)^2
350 NEXT I%
360 VARIANCE = VARIANCE / (N%-1)
370 SD = SQR(VARIANCE)
380 '
390 PRINT MEAN SD          'вывод результатов
400 '
410 END
```

Таблица 2-3.

Вычисление гистограммы проходит быстро, поскольку требует только индексации и инкремента. В сравнении с этим, расчет среднего значения и стандартного отклонения требует времени, которое тратится на операции сложения и умножения. Стратегия такого метода заключается в том, что использование медленных операций проводится на нескольких точках гистограммы, а не на множестве отсчетов сигнала. Это делает алгоритм намного быстрее тех, которые были описаны ранее, почти в десять раз для длинных сигналов, если вычисление осуществляется на стандартных компьютерах.

Следует вспомнить, что принятый сигнал является «зашумленной версией» основного процесса, поэтому некоторым концепциям даны разные имена. Гистограмма формируется из полученного сигнала. Соответствующая функция строится и для основного процесса и называется *функцией вероятностной меры* (probability mass function – pmf). Гистограмма вычисляется для выборок конечного размера, в то время как функцию вероятностной меры (далее PMF) *хотелось бы* рассчитать для бесконечного числа отсчетов. PMF можно как найти из гистограммы, так и рассчитать с применением математических выражений.

Рис.2-5 показывает пример PMF, а также гистограмму, которая может быть с ней связана. Для того чтобы понять эти концепции, обратите внимание на размерность вертикальной оси. Как было описано ранее, вертикальная ось гистограмм определяет число вхождений, т.е. число раз, сколько определенная величина встретилась в сигнале. Вертикальная ось PMF содержит подобную информацию, за исключением того, что величины не фрагментарны. Другими словами, каждое значение в гистограмме должно быть поделено на число точек для приближения функции к PMF. Это подразумевает, что каждое значение PMF лежит между нулем и единицей, а сумма всех значений равна единице.

Важность функции вероятностной меры состоит в том, что она описывает *вероятность* генерации определенного значения. Например, представьте, что сигнал с PMF, изображенной на Рис.2-5б. Какова вероятность того, что отсчет, взятый из такого сигнала, будет иметь величину 120? Рис.2-5б дает ответ – 0.03, или 1 шанс из 34. Какова вероятность того, что значение отсчета будет больше 150? Суммируя величины функции для 151, 152, …, 255, получим ответ – 0.0122, или 1 шанс из 82. Таким образом, каждые 82 отсчета сигнала можно ожидать значение больше 150. Какова вероятность, что значение отсчета будет между 0 и 2555. Суммируя все величины PMF, получим вероятность 1.0, что является бесспорным.

Гистограмма и функция вероятностной меры могут использоваться только с дискретными данными, такими как дискретные сигналы, записанные на компьютер. Подобная концепция применима и для непрерывных сигналов, таким как напряжение для аналоговой электронике. *Функция плотности вероятности* (probability density function – pdf) называемая также *функцией распределения вероятности* (далее PDF) является для непрерывных сигналов тем же самым, чем является функция вероятностной меры для дискретных сигналов. Например, представьте, что есть аналоговый сигнал, который проходит через аналого-цифровой преобразователь, в результате чего получается дискретный сигнал, изображенный на Рис.2-4а. Для простоты будем считать, что напряжение между 0 и 255 мВ преобразуется в значения от 0 до 255. PMF такого сигнала показана *маркерами* на Рис.2-5б. Подобная PDF аналогового сигнала представлена на Рис.2-5в *сплошной линией*, показывая сигнал, состоящий из континуальных значений, каким является напряжение в аналоговой электронике.

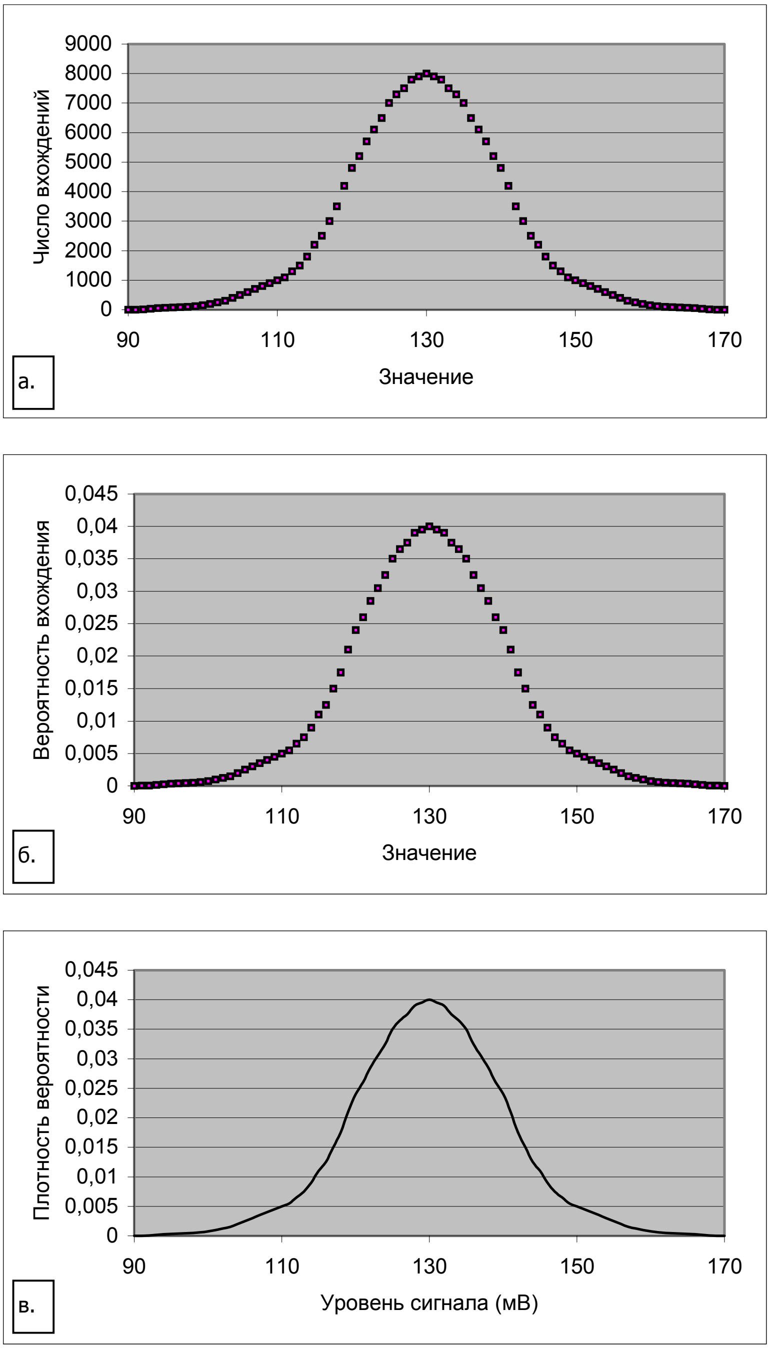
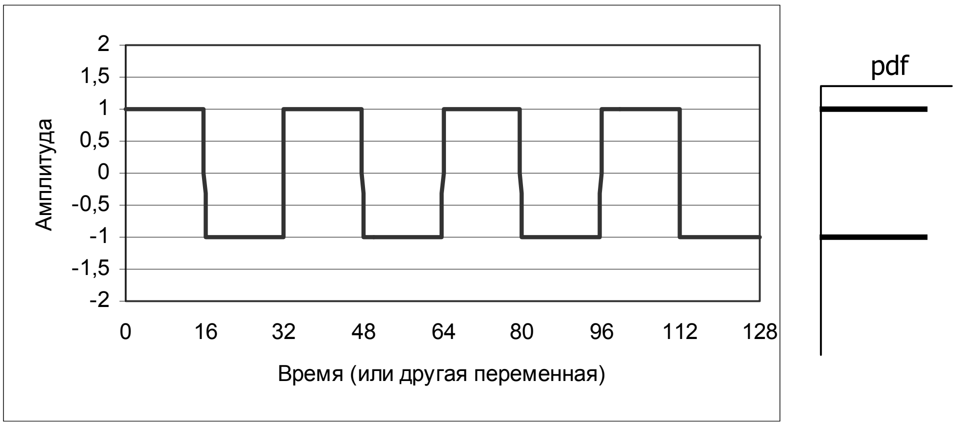


Рис.2-5. Взаимосвязь гистограммы (а), функции вероятностной меры, PMF (б) и функции плотности вероятности, PDF (в).

Вертикальная ось PDF измеряется в плотности вероятности, а не в просто вероятности. Например, PDF выше 0.03 при значении 120.5 не означает, что напряжение 120.5 мВ будет встречаться в сигнале с вероятностью 3%. Фактически, вероятность того, что в аналоговом сигнале встретится значение 120.5 мВ, бесконечно мала, поскольку число вероятных значений включает в себя и значения 120.9997, 120.9998, 120.9999 мВ и т.д.

Чтобы вычислить вероятность, необходимо умножить *плотность вероятности* на *диапазон* величин. Например, вероятность того, что сигнал будет находиться между 120.4 и 120.5 мВ равна (120.5 – 120.4)*0.03 = 0.003 и т.п. Если PDF в интересующем диапазоне непостоянна, то умножение сводится к интегрированию по этому интервалу. Другими словами, площадь под PDF ограничена определенными значениями. Поскольку сигнал должен всегда принимать *некоторое значение*, общая площадь под кривой PDF, интеграл от -∞ до +∞, будет равна единице.

а. меандр



б. сигнал треугольной формы

в. случайный сигнал

Рис.2-6. Три непрерывных сигнала и их функции плотности вероятности.

Гистограмма, PDF и PMF очень схожи. Математики всегда применяют их по назначению, но часто можно встретить их взаимную противоречивую замену в применении учеными и инженерами, что приводит к ошибке. Рис.2-6 показывает три *непрерывных* сигнала и их плотности вероятности. Если бы они были *дискретными*, то обозначение по горизонтальной оси выражалось бы в «количестве отсчетов», и использовалась функция вероятностной меры.

При вычислении гистограмм встречается проблема, когда число уровней каждой выборки может принимать большее количество значений, чем число выборок в сигнале. Это особенно справедливо для сигналов, представленных в виде чисел с *плавающей запятой*, где каждая выборка хранится в виде дробной величины. Например, целочисленное представление может требовать значение 3 или 4, в то время как число с плавающей запятой может представить биллионы чисел от 3 до 4. Ранее был описан способ расчета гистограмм, считающий число возможных уровней квантизации. Он не подходит для работы с дробными числами, потому что существует биллион возможных уровней, которые должны быть приняты в расчет, даже если они не будут встречены в сигнале. Например, в сигнале изображения, состоящего из 10000 выборок, каждая выборка может иметь один биллион возможных величин. Преобразование в гистограмму будет содержать один биллион точек данных, и только 10000 из них будут отличны от нуля.

Решение таких проблем в технике назвали элементная дискретизация (binning). Это осуществляется произвольным выбором длины гистограммы, состоящей из подходящего количества точек, например, 1000, часто называемых элементами дискретизации (bins). Величина каждого элемента представляет общее количество выборок сигнала, значение которых лежит в *определенном диапазоне*. Например, представьте сигнал, представленный числами с плавающей запятой, содержащий величины между 0.0 и 10.0, и гистограмму из 1000 точек. Элемент 0 в гистограмме показывает число выборок, значения которых лежат в пределах от 0 и 0.01, элемент 1 – это число выборок с величинами между 0.01 и 0.02, и т.д. до элемента 999, показывающего число выборок с величинами от 9.99 до 10.0. Таблица 2-4 показывает программу для расчета гистограммы таким способом.

```
100 'РАСЧЕТ ЭЛЕМЕНТНО ДИСКРЕТИЗИРОВАННОЙ ГИСТОГРАММЫ
110 '
120 DIM X[25000]          'X[0] до X[25000] содержит сигнал,
130 '                     'представленный дробными числами от 0.0 до 10.0
140 DIM H%[999]           'H%[0] до H%[999] содержит гистограмму
150 '
160 FOR I% = 0 TO 999     'обнуление элементов гистограммы
170 H%[I%] = 0
180 NEXT I%
190 '
200 GOSUB XXXX            'некоторая подпрограмма получения сигнала X%[ ]
210 '
220 FOR I% = 0 TO 25000   'расчет элементно дискретизированной гистограммы
230 BINNUM% = INT( X[I%] * .01 )
240 H%[ BINNUM%] = H%[ BINNUM%] + 1
250 NEXT I%
260 '
270 END
```

Таблица 2-4.

Сколько элементов необходимо использовать? Это должен быть компромисс между двумя проблемами. Как показано на Рис.2-7 много элементов создают трудность в оценке *амплитуды* PMF основного процесса. Это происходит из-за того, что в каждый элемент попадает только несколько выборок, и статистический шум очень высок. С другой стороны, небольшое число элементов создает трудность оценки PMF основного процесса по горизонтальной оси. Другими словами, число элементов определяется компромиссом между точностями по вертикальной и горизонтальной осям.

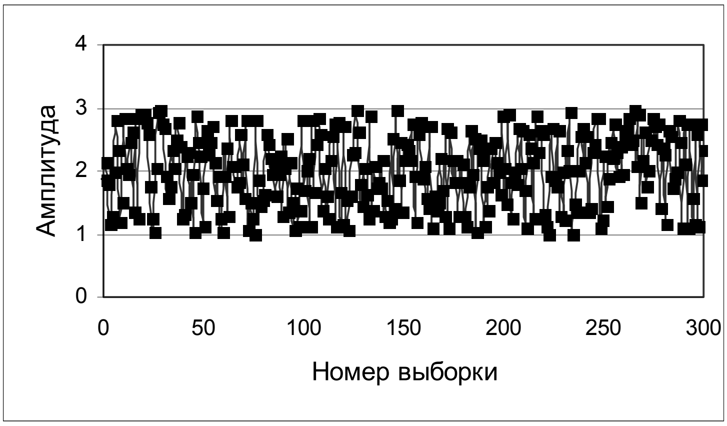


Рис.2-7а. Сигнал для примера элементно дискретизированных гистограмм. Для представления используются дробные числа. Число выборок – 300. Значения распределены между 1 и 3.

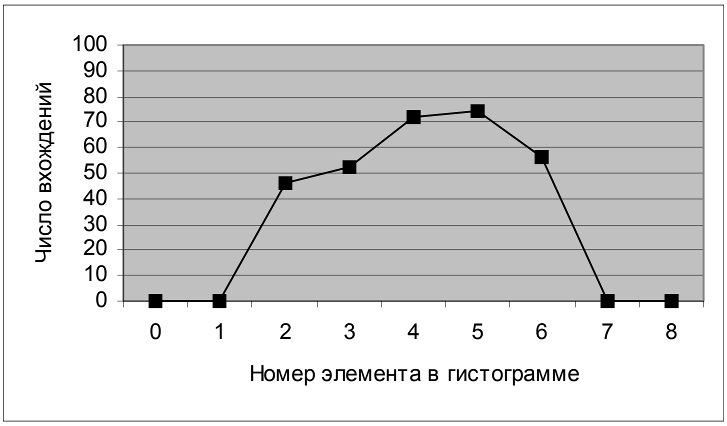


Рис.2-7б. Пример элементно дискретизированных гистограмм с различным числом элементов. На верхней их число составляет 601, на нижней 9.

Нормальное распределение

Сигналы, сформированные от случайных процессов, обычно имеют колоколообразную PDF. Такая функция называется *нормальным распределением*, *гауссово распределение* или *распределение Гаусса*, по имени немецкого математика Карла Фридриха Гаусса (1777-1885). Причина, по которой, такая кривая широко распространена в природе, будет обсуждаться позже вместе с *цифровой генерацией помех*. Основная форма кривой генерируется от экспоненты:

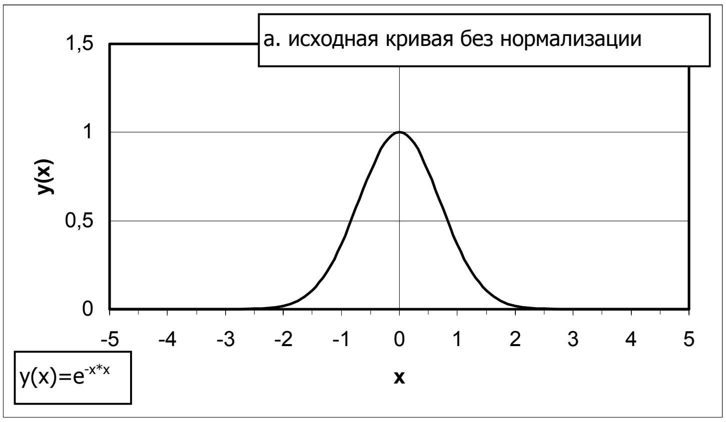
$$y(x) = e^{-x^2}$$

При добавлении регулируемых среднего значения и стандартного отклонения эта кривая может быть преобразована в законченное распределение Гаусса. Вдобавок, выражение должно быть нормализовано так, чтобы площадь под кривой равнялась единице. Это стандартное требование всех функций распределения вероятности. Общая форма нормального распределения выражается через:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-u)^2 / 2\sigma^2}$$

Уравнение 2-8. Выражение для нормального распределения. P(x) – функция распределения вероятности, u – среднее значение, σ – стандартное отклонение.

Рис.2-8. показывает примеры кривых нормального распределения с различными средними значениями и стандартными отклонениями. *Среднее значение* центрирует кривую над определенным значением, в то время как стандартное отклонение оказывает влияние на ширину колокола.



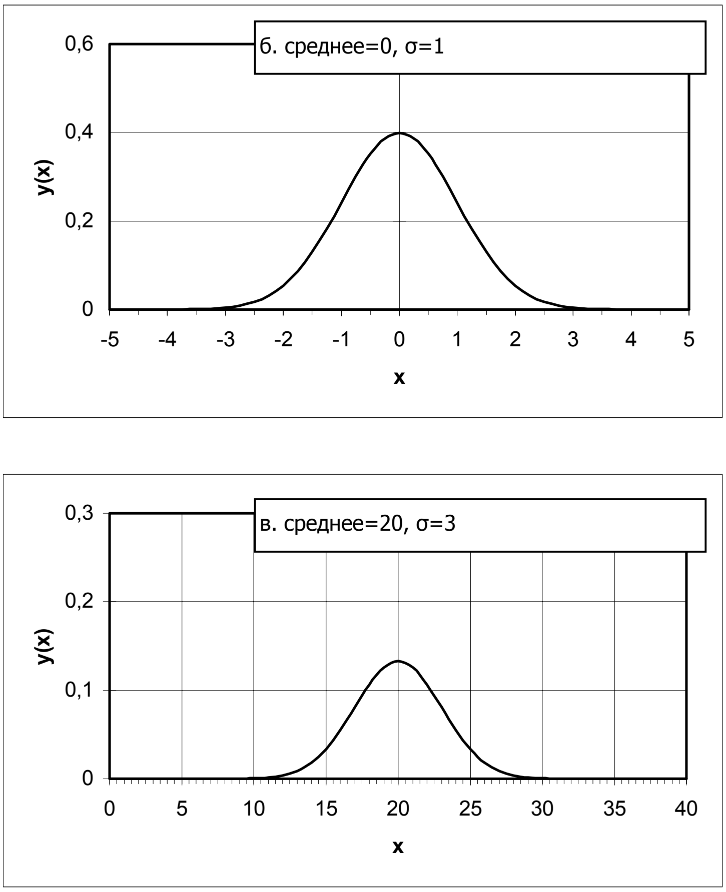


Рис.2-8. Примеры кривых нормального распределения.

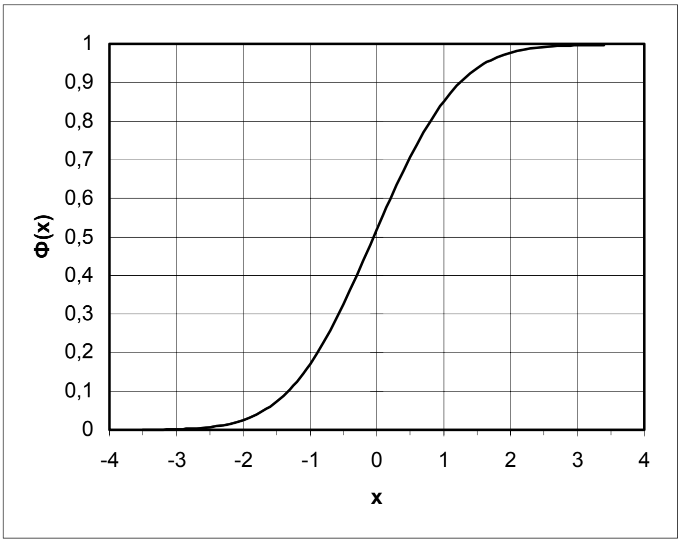
Интересной характеристикой нормального распределения является то, что шлейфы (экспоненциально затухающие части функции) очень быстро спадают до нуля, гораздо быстрее, чем у других функций, например, затухающие экспоненты или 1/x. Например, на двух, четырех и шести стандартных отклонениях от среднего, значение гауссовой кривой спадает до 1/19, 1/7563 и 1/166.666.666, соответственно. Поэтому нормально распределенные сигналы как, например, на Рис.2-6в, кажутся аппроксимированными до величин полного размаха. В принципе, сигналы такого типа могут иметь отклонения до бесконечных амплитуд. На практике, резкий спад PDF нормального распределения предписывает, что такие экстремумы никогда не встретятся. Это выражается в том, что сигналы обладают относительно ограниченный вид с различимой амплитудой около 6-8σ.

Как было сказано ранее, интеграл от функции распределения вероятности (PDF) используется для нахождения вероятности того, что сигнал будет находиться в определенном диапазоне значений. Это делает интеграл от PDF настолько важным, что он получил собственное название – *функция интегрального распределения* (CDF – cumulative distribution function). Неприятной проблемой, связанной с функцией нормального распределения, является то, что она не может быть интегрировано обычными методами, поэтому интеграл по гауссовой функции считается *численным интегрированием*. Это достигается тонкой дискретизацией непрерывной функции, например, на миллион точек в диапазоне от -10σ до $+10\sigma$. Выборки в этом дискретном сигнале *складываются* для имитации интегрирования. Дискретизированная кривая, полученная в результате такого интегрирования, запоминается в таблице для последующего использования при расчете вероятностей.

CDF нормального распределения показана на Рис.2-9, а ее численные значения приведены в Таблице 2-5. Поскольку эта кривая часто используется в вероятности, она имеет свой символ: $\Phi(x)$ (строчная греческая фи). Например, $\Phi(-2)$ имеет значение 0.0228. Это означает, что существует 2.28% вероятность того, что величина сигнала в любое случайное время будет в пределах от $-\infty$ до 2 стандартных отклонений ниже среднего значения. Подобно этому, значение $\Phi(1)$ равное 0.8413 подразумевает, что существует 84.13% вероятность того, что величина сигнала будет в пределах от $-\infty$ до одного стандартного отклонения выше среднего значения. Для расчета вероятности того, что величина сигнал будет *между* двумя значениями, необходимо произвести вычитание соответствующих чисел из таблицы. Например, вероятность того, что величина сигнала в любое случайное время будет находиться между двумя стандартными отклонениями ниже среднего и одним отклонением выше среднего, рассчитывается как $\Phi(1) - \Phi(-2) = 0.8185$, или 81.85%.

При использовании этого метода, выборки, взятые из нормального распределенного сигнала, будут в пределах $\pm 1\sigma$ около 68% всего времени, $\pm 2\ \sigma$ около 95%, 3σ около 99.75%. Вероятность того, что сигнал будет дальше 10 стандартных отклонений от среднего значения настолько мала, что это может встретиться всего на несколько *микросекунд* за все время существования вселенной!

Уравнение 2-8 может использоваться для расчета функции вероятностной меры нормально распределенных *дискретных* сигналов. В этом случае, x соотносится с одним из квантованных уровней, которые может принимать сигнал, например, 4096 бинарными значениями, существующие в 12-разрядных аналого-цифровых преобразователях. Параметр $1/\sqrt{2\pi}\sigma$ используется для нормализации площади под кривой PDF к *единице*. Чтобы сумма всех значений PMF была равна единице, можно использовать другой параметр. В большинстве случаев, для этого генерируют кривую без нормализации, суммируют все значения, а затем делят каждое из них на общую сумму.



x	Φ(x)	x	Φ(x)
-3.4	0.003	0.0	0.5000
-3.3	0.005	0.1	0.5398
-3.2	0.007	0.2	0.5793
-3.1	0.0010	0.3	0.6179
-3.0	0.0013	0.4	0.6554
-2.9	0.0019	0.5	0.6915
-2.8	0.0026	0.6	0.7257
-2.7	0.0035	0.7	0.7580
-2.6	0.0047	0.8	0.7881
-2.5	0.0062	0.9	0.8159
-2.4	0.0082	1.0	0.8413
-2.3	0.0107	1.1	0.8643
-2.2	0.0139	1.2	0.8849
-2.1	0.0179	1.3	0.9032
-2.0	0.0228	1.4	0.9192
-1.9	0.0287	1.5	0.9332
-1.8	0.0359	1.6	0.9452
-1.7	0.0446	1.7	0.9554
-1.6	0.0548	1.8	0.9641
-1.5	0.0668	1.9	0.9713
-1.4	0.0808	2.0	0.9772
-1.3	0.0968	2.1	0.9821
-1.2	0.1151	2.2	0.9861
-1.1	0.1357	2.3	0.9893
-1.0	0.1587	2.4	0.9918
-0.9	0.1841	2.5	0.9938
-0.8	0.2119	2.6	0.9953
-0.7	0.2420	2.7	0.9965
-0.6	0.2743	2.8	0.9974
-0.5	0.3085	2.9	0.9981
-0.4	0.3446	3.0	0.9987
-0.3	0.3821	3.1	0.9990
-0.2	0.4207	3.2	0.9993
-0.1	0.4602	3.3	0.9995
0.0	0.5000	3.4	0.9997

Рис.2-9. и Таблица 2-5.
Функция интегральной вероятности нормального распределения, среднее значение = 0, стандартное отклонение = 1.

Цифровая генерация помех

Случайная помеха является важной темой как в электронике, так и в ЦОС. Например, она ограничивает насколько маленький сигнал можно измерить прибором, дистанцию, на которой может работать радиостанция, какое количество радиоизлучения необходимо для выработки рентгеновского изображения. Распространенной потребностью в ЦОС является генерация сигналов, которые похожи на различные типы случайных помех. Это необходимо для проверки эффективности алгоритмов, которые должны *работать* в условиях наличия помех.

Сердцем цифровой генерации помех является *генератор случайных чисел*. Большинство языков программирования содержат эту стандартную функцию. Оператор BASIC: X=RND, загружает в переменную X новое случайное число, каждый раз когда он встречается в программе. Каждое случайное число имеет значение между нулем и единицей, с одинаковой вероятностью быть между двумя этими значениями. Рис.2-10а показывает сигнал, который сформирован из 128 выборок, взятых по принципу такого генератора случайных чисел. Среднее значение основного процесса, который генерирует этот сигнал, составляет 0.5. Стандартное отклонение составляет $1/\sqrt{12} = 0.29$, распределение равномерно между нулем и единицей.

Алгоритмы должны проверяться с использованием такого типа данных, которые они встречают в реальных условиях. Это создает необходимость цифровой генерации помехи с гауссовой функцией распределения вероятности. Существует два способа для генерации таких сигналов, используя генератор случайных чисел. Рис.2-10 иллюстрирует первый метод. Рис(б) показывает сигнал, состоящий из выборок, полученных сложением двух случайных чисел, т.е. X=RND+RND. Поскольку каждое случайное число может принимать значение от нуля до единицы, их сумма будет находиться в пределах от нуля до двух. Среднее значение будет составлять *единицу*, а стандартное отклонение $1/\sqrt{6}$ (запомните, когда складываются два независимых сигнала, дисперсия также складывается). Как показано, PDF сигнала изменилась от равномерного до треугольного распределения. Поэтому сигнал находится большее свое время вокруг *единичного* значения, и меньше около *нуля* и *двух*.

Рис(в) продолжает эту идею, генерируя каждую выборку путем суммирования двенадцати случайных чисел. Среднее значение в этом случае составляет *шесть*, а стандартное отклонение *единицу*. Важно, что функция распределения вероятности становится виртуально похожа на гауссову. Этот метод может использоваться для генерации нормально распределенного сигнала с произвольными средним значением и стандартным отклонением. Для каждой выборки сигнала: (1) сложить двенадцать случайных чисел, (2) вычесть шесть, чтобы получить среднее равное нулю, (3) умножить на желаемое стандартное отклонение и (4) добавить желаемое среднее значение.

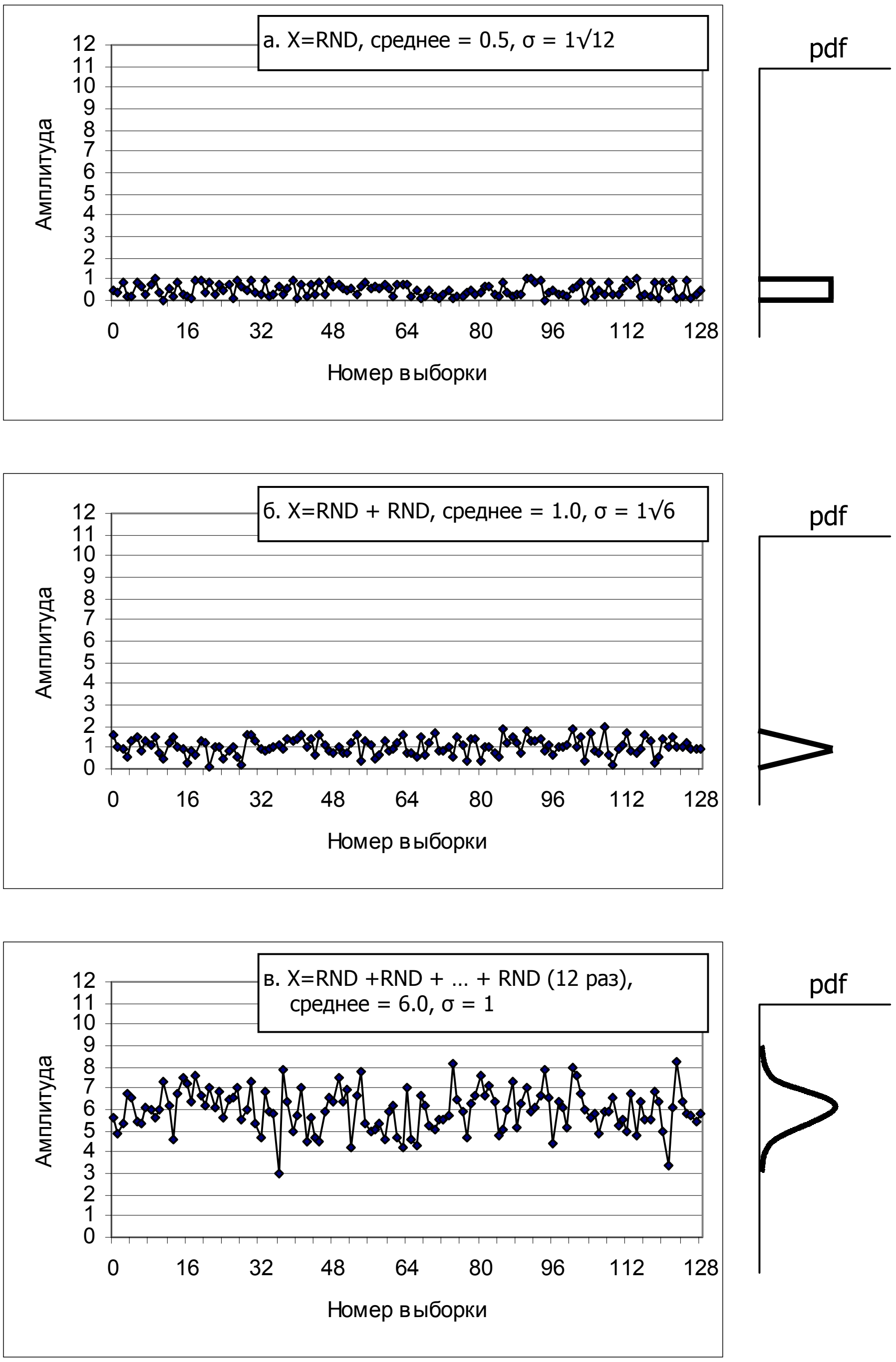


Рис.2-10. Преобразование равномерного распределения в нормальное.

Математическая основа этого алгоритма содержится в *теореме о центральном пределе*, одной из важных концепций в вероятности. В простейшем форме, теорема о центральном пределе говорит, что *сумма* случайных чисел становится нормально распределенной, чем больше чисел участвует в сложении. Теорема о центральном пределе *не требует*, чтобы отдельные случайные числа были получены из какого-либо распределения, или были получены из *одинакового* распределения. Теорема о центральном пределе показывает причину, почему нормально распределенные сигналы настолько часто встречаются в природе. Всякий раз при взаимодействии различных случайных воздействий, результирующая PDF будет нормально распределена.

Второй метод генерации нормально распределенных чисел состоит в том, что реализуются *два* генератора случайных чисел для получения R_1 и R_2. Нормально распределенное случайное число X может быть тогда найдено как:

$$X = \left(-2\log R_1\right)^{1/2}\cos\left(2\pi R_2\right)$$

Уравнение 2-9. Генерация нормально распределенных чисел. R_1 и R_2 являются случайными числами с равномерным распределением между 0 и 1. Результат, X, является нормально распределенным случайным числом со средним значением «0» и стандартным отклонением «1». Логарифм является натуральным, косинус считается в радианах.

Как и в предыдущем случае, этот метод позволяет генерировать нормально распределенное случайное число с произвольными средним значением и стандартным отклонением. Для этого необходимо: рассчитать число по выражению 2-9, умножить на желаемое стандартное отклонение и добавить желаемое среднее значение.

Генератор случайных чисел начинает работу с некоторого *начального числа* (seed), между нулем и единицей. При запуске генератора случайных чисел начальное число передается через постоянный алгоритм, в результате чего получается новое число между нулем и единицей. Это новое число называют *случайным числом*, которое сохраняется для использования на следующий раз в качестве начального числа. Алгоритм, который преобразует начальное число в новое случайное число, часто выражается в виде:

$$R = \left(aS + b\right)\operatorname{mod}ulo\ c$$

Уравнение 2-10. Распространенный алгоритм генерации равномерно распределенного случайного числа между нулем и единицей. В этом методе S – начальное число, R – новое случайное число, a, b, c - соответствующим образом выбранные константы.

Таким способом может быть сгенерирована последовательность случайных чисел, начиная с некоторого начального числа, что позволяет при многократном запуске программы получать одинаковые последовательности случайных чисел. Чтобы изменить эту последовательность необходимо изменить начальное число. Большинство языков программирования предоставляют функцию *изменения начального числа* (reseeding), позволяя выбирать его значение. Распространенным способом является использование в качестве начального числа *системное время*, что позволяет получить новые последовательности случайных чисел при каждом запуске программы.

С чисто математической точки зрения, числа, генерируемые таким способом, не могут быть абсолютно случайными, так как каждое из них полностью зависит от *предыдущего* числа. Для описания такой ситуации часто используется термин *псевдослучайный*. Однако вы не должны беспокоится по этому поводу. Последовательности, получаемые от генератора случайных чисел, являются статистически случайными с высокой точностью. Маловероятно, что вы встретите ситуацию, где они будут непригодными.

Точность и погрешность

Точность (accuracy) и погрешность (precision) являются терминами описания систем, которые *измеряют*, *оценивают* или *предсказывают*. Во всех случаях, имеется некоторый параметр, значение которого хотелось бы узнать. Он называется *истинное значение* или просто *истина*. Метод предоставляет *измеренное значение*, которое бы хотелось иметь как можно ближе к истинному значению. *Точность* и *погрешность* являются способами описания ошибки, которая возникает между этими значениями.

К сожалению, точность и погрешность часто взаимно заменяются в нетехнической литературе. Фактически, их описание в словаре часто ссылается друг на друга. Назло этому, ученые и инженеры имеют специфическое определение каждому из этих терминов. Вы обязательно должны использовать эти термины корректно!

Например, представьте, что океанограф измеряет глубину, используя систему гидролокации. Короткие звуковые импульсы передаются с корабля, отражаются от дна океана и принимаются на поверхности в виде эха. Звуковые волны распространяются в воде с приблизительно одинаковой скоростью, позволяя найти глубину из прошедшего времени между переданным и принятым импульсом. Как и при всех эмпирических измерениях, существует некоторая ошибка между измеренным и истинным значением. Это может быть вызвано несколькими факторами: случайной помехой в электронике, сигналами на поверхности воды, растения на дне океана, изменения температуры воды, вызывающие изменение скорости распространения звуковых волн и т.п.

Для устранения этих эффектов, океанограф проводит несколько последовательных измерений на участке, где глубина составляет *ровно* 1000 метров (истинное значение). Эти измерения упорядочиваются в гистограмме, показанной на Рис.2-11. Как следует из теоремы о центральном пределе, собранные данные имеют нормальное распределение. *Среднее значение* попадает на центр распределения и представляет собой лучшую оценку глубины, полученную на основе собранных данных. Стандартное отклонение определяет ширину распределения, описывая вариацию, произошедшую между последовательными измерениями.

Эта ситуация приводит к двум основным типам ошибок, с которыми может столкнуться система. Во-первых, среднее значение может быть сдвинуто от истинного значения. Величину этого сдвига называют *точностью* измерения. Во-вторых, отдельные измерения могут не сходиться друг с другом, что демонстрирует ширина распределения. Она называется *погрешностью* измерения и выражается через вычисление квадратического отклонения, отношения сигнал-шум или коэффициента вариации.

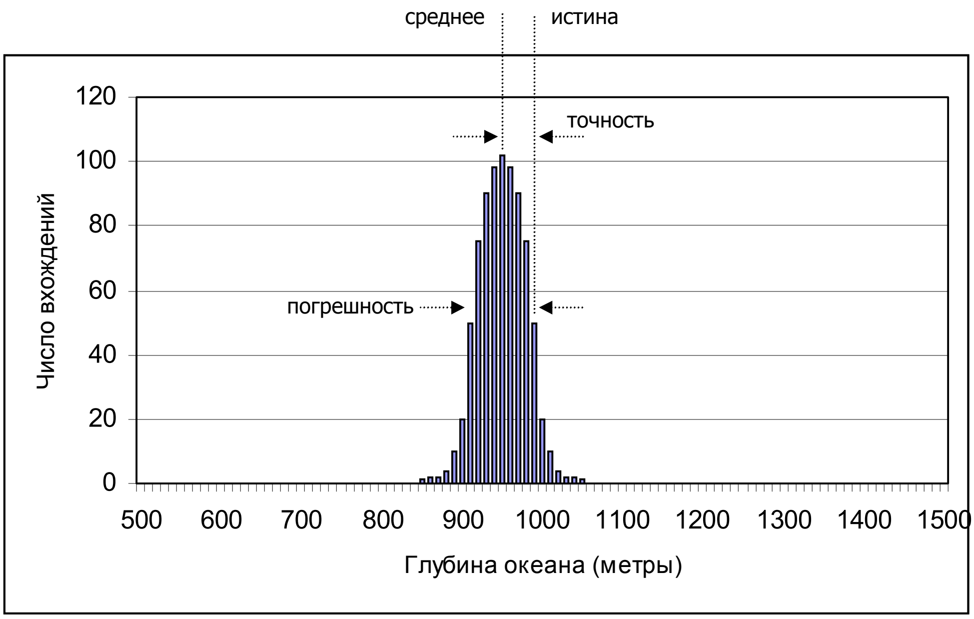


Рис.2-11. Определение точности и погрешности.

Предположим, что измерение имеет хорошую точность, но большую погрешность; гистограмма центрируется около истинного значения, но она очень широкая. Хотя измерения корректны в целом, каждое отдельное из них, является плохим измерением истинного значения. В такой ситуации говорят, что имеется плохая *повторяемость*, измерения сделанные последовательно, не сходятся друг с другом. Большая погрешность вызвана *случайными ошибками*. Это название дано ошибкам, которые изменяются при каждом повторении измерения. Усреднение нескольких измерений *всегда* снижает погрешность. Вкратце, *погрешность является мерой случайных помех*.

Теперь представьте, что измерение имеет маленькую погрешность, но плохую точность. Это приводит к тому, что гистограмма узкая, но не центрируется у истинного значения. Последовательные измерения очень близки по значению, но *все* имеют большую ошибку. Низкая точность вызывается *систематическими ошибками*. Это ошибки, которые повторяются каждый раз, когда проводится измерение одним и тем же методом. Точность зависит от того, как хорошо *откалиброванa* система. Например, при измерении глубины океана, непосредственно измеряемым параметром является прошедшее время. Оно переводится в значение глубины методом поверки, который связывает *миллисекунды* и *метры*. Он может состоять как из простого умножения на постоянную скорость, так и осложнятся дюжиной корректировок второго порядка. Вкратце, *точность является мерой калибровки*.

На практике существует множество ситуаций, когда точность и погрешность взаимосвязаны. Например, представьте проектирование электронного усилителя на 1% резисторах. Допустимость показывает, что величина каждого из резисторов может быть в 1% диапазоне от указанного значения при различных условиях: температура, влажность и т.п. Эта ошибка в сопротивлении вызывает соответствующую ошибку в коэффициенте усиления. Является ли эта ошибка проблемой точности или погрешности?

Ответ зависит от того, как были проведены измерения. Например, представьте, что вы сделали *один* усилитель и проверили его в течение нескольких минут несколько раз. Ошибка в усилении остается постоянной при каждой проверке, и вы делаете вывод, что проблема заключается в *точности*. Напротив, представьте, что вы сделали *тысячу* усилителей. Усиление от прибора к прибору изменяется случайно, значит проблема в *погрешности*. Также, каждый из этих усилителей показывает колебание коэффициента усиления при различной температуре и других изменениях окружающей среды. Снова, проблема вызвана *погрешностью*.

При определении, чем вызвана ошибка, задайте себе два вопроса. Первый: Улучшит ли измерения усреднение последовательных показаний? Если да, ошибка вызвана погрешностью, если нет, точностью. Второй: Калибровка устраняет ошибку? Если да, проблема в точности, если нет, в погрешности. Особенное внимание следует делить тому, как калибруется устройство, и как часто это должно проводиться.