# 1 Статистика и вероятность. Случайные величины и распределения

## 1.1 Краткие теоретические сведения

**Случайной величиной** называется величина $X$, если все ее возможные значения образуют конечную или бесконечную последовательность чисел $x_1, \ldots, x_N$, и если принятие ею каждого из этих значений есть случайное событие. **Вероятностью** наступления данного случайного события $x_k$ называется предел относительной частоты наступления данного события $n_k/N$:

$$P(x_k) = \lim_{N\to\infty} \frac{n_k}{N}.$$

Случайная величина может быть не только дискретной, но и **непрерывной**. В этом случае она может принимать любое значение из заданной области действительных чисел. Непрерывная случайная величина характеризуется **плотностью вероятности**:

$$\varphi(x) = \lim_{\Delta x\to 0} \frac{P(x < X < x + \Delta x)}{\Delta x} = \frac{dP}{dx}.$$

Вероятность попадания значений $X$ в интервал $(x_1, x_2)$ можно вычислить по формуле:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \varphi(x)\, dx.$$

Вероятность того, что значения случайной величины не превышают заданного числа $x$ называют **функцией распределения**: $F(x) \equiv P(X \leqslant x) = \int_{-\infty}^{x} \varphi(x)\, dx$. Так как вероятность попадания значений $(X)$ в промежуток $(-\infty, \infty)$ равна единице, плотность вероятности должна быть **нормирована**:

$$\int_{-\infty}^{\infty} \varphi(x)\, dx = 1.$$

Две случайные величины $X$ и $Y$ называются **независимыми**, если наступление одного из событий $x_n$ никак не сказывается на результате события $y_n$.

В этом случае вероятность совместного наступления событий $x_n$ и $y_n$ равна $P(x_n y_n) = P(x_n)P(y_n)$. Дальнейшие формулы в этом разделе подразумевают независимость входящих в них случайных величин, если зависимость не указана явно.

Набор случайных величин $X$ можно охарактеризовать **средним арифметическим**: $<X> = {}^1\!/_N \sum_{n=1}^{N} x_n$. Для малых $N$ данная величина будет отличаться от **математического ожидания**, определяемого по формуле

$$M(X) \equiv \overline{X} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n \quad \text{и} \quad M(X) = \int_{-\infty}^{\infty} x\varphi(x)\,dx,$$

однако **закон больших чисел** говорит, что чем большей является величина $N$, тем ближе $<X>$ к $\overline{X}$.

Помимо математического ожидания набор случайных величин характеризуется медианой и модой. **Мода** — значение во множестве наблюдений, которое встречается наиболее часто. Иногда мод может быть больше одной. В этом случае можно сказать, что совокупность *мультимодальна*. Как правило мультимодальность указывает на то, что набор данных не подчиняется нормальному распределению. **Медиана** — возможное значение признака, которое делит отсортированную совокупность на две равные части: 50% «нижних» единиц ряда данных будут иметь значение признака не больше, чем медиана, а 50% «верхних» — значения признака не меньше, чем медиана.

*Свойства математического ожидания*: $M(\sum \mathfrak{C}_n \cdot X_n) = \sum \mathfrak{C}_n \cdot M(X_n)$, где $\mathfrak{C}_n$ – постоянная величина; $M(\prod X_n) = \prod M(X_n)$; $M(f(x)) = \int\limits_{-\infty}^{\infty} f(x)\varphi(x)\,dx$. Если $f(x) = (x - x_0)^n$, говорят, что $M(f(x))$ является **моментом** случайной величины порядка $n$. При $x_0 = 0$ момент называют **начальным**, а при $x_0 = \overline{X}$ — **центральным**.

Центральный момент второго порядка называют **дисперсией**: $D(X) = \overline{(x - \overline{x})^2}$. Разброс случайной величины относительно математического ожидания характеризуется **средним квадратичным отклонением** $\sigma = \sqrt{D}$. Дисперсию можно вычислить и по упрощенному выражению:

$$D = \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{2x\overline{x}} + \overline{\overline{x}^2} = \overline{x^2} - \overline{x}^2 \equiv M(x^2) - [M(x)]^2.$$

▶ Если функция плотности вероятности случайного события описывается аналитически, говорят, что она подчинена некоторому **распределению**. Наиболее известными видами распределения являются непрерывное, нормальное (гауссово), Пуассона, биномиальное, экспоненциальное и многие другие.

Говорят, что случайная величина имеет непрерывное **равномерное распределение** на отрезке $[a, b]$, где $a$, $b \in \mathbb{R}$, если ее плотность $\varphi(x)$ имеет вид

$$\varphi(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & x \notin [a, b] \end{cases}.$$

Интегрируя, получим, для $F(x)$:

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leqslant x < b \\ 1, & x \geqslant b \end{cases}.$$

Математическое ожидание и медиана для равномерного распределения совпадают с серединой отрезка $[a, b]$, модой же является любое значение из этого отрезка.

Важнейшую роль в физике имеет **нормальное** (гауссово) распределение. Физическая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех. Плотность вероятности нормального распределения имеет вид:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\overline{x})^2}{2\sigma^2}\right),$$

Функция распределения записывается через *интеграл Римана*:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left(-\frac{(t-\overline{x})^2}{2\sigma^2}\right) dt.$$

Особенностью нормального распределения является совпадение медианы, моды и математического ожидания.

Вероятность того, что нормальная случайная величина с параметрами $\overline{x}$ и $\sigma$ попадет в интервал $(\alpha, \beta)$ равна:

$$P(\alpha < X < \beta) = \Phi\Big(\frac{\beta-\overline{x}}{\sigma}\Big) - \Phi\Big(\frac{\alpha-\overline{x}}{\sigma}\Big),$$

где $\Phi(x)$ – **функция Лапласа**:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} \exp\left(-\frac{t^2}{2}\right) dt.$$

Для выяснения зависимости случайных величин $X$ и $Y$ используются такие функции, как корреляция и ковариация. **Ковариация** является мерой линейной зависимости случайных величин и определяется формулой: $\operatorname{cov}(X,Y) = \overline{(X-\overline{X})(Y-\overline{Y})}$. Понятно, что ковариация величины с самой собой есть ее дисперсия. *Ковариация независимых случайных величин равна нулю*, обратное неверно.

**Коэффициент корреляции** двух величин задается формулой

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)\cdot D(Y)}}.$$

*Коэффициент корреляции равен* $\pm 1$ *тогда и только тогда, когда* $X$ *и* $Y$ *линейно зависимы.* Если они независимы, $\rho_{X,Y} = 0$. Промежуточные значения коэффициента корреляции не позволяют однозначно судить о зависимости случайных величин, но позволяет предполагать степень их зависимости.

Для непрерывных величин аналогом коэффициента корреляции является *корреляционная функция.* Более часто применяют ее разновидность — **автокорреляционную функцию**:

$$\Psi(\tau) = \int f(t)f(t-\tau)\,dt,$$

показывающую связь сигнала (функции $f(t)$) с копией самого себя, смещенного на величину $\tau$. Автокорреляционная функция имеет обычно максимум при $t = 0$. Имея два сигнала, представляющих собой исходный и сдвинутый на неизвестную величину $z$, можно определить $z$ из корреляционной функции: эту величину однозначно покажет положение максимума корреляционной функции.

В теории сигналов вводят понятие шума. **Шум** — беспорядочные колебания различной физической природы, отличающиеся сложностью временной и спектральной структуры. Наиболее общим видом шума является **белый шум** — стационарный шум, спектральные составляющие которого равномерно распределены по всему диапазону задействованных частот. В природе и технике «чисто» белый шум (то есть белый шум, имеющий одинаковую спектральную мощность на всех частотах) не встречается (ввиду того, что *такой сигнал имел бы бесконечную мощность*), однако под категорию белых шумов попадают любые шумы, спектральная плотность которых одинакова (или слабо отличается) в рассматриваемом диапазоне частот. Белым шумом является любой шум с фиксированной дисперсией, нулевым математическим ожиданием и автокорреляционной функцией, имеющей вид *дельта-функции Дирака.* Чаще всего белый шум моделируют гауссовским распределением, т.к. такая модель хорошо подходит для математического описания многих природных процессов.

Зашумленность сигнала характеризуют **отношением сигнал/шум** (SNR):

$$\mathrm{SNR} = \frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}} = \left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right)^2,$$

где $P$ и $A$ — соответственно, средняя мощность и среднеквадратичное значение амплитуды сигнала и шума. Зачастую SNR выражают в децибелах:

$$\mathrm{SNR(dB)} = 10 \lg \left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) = 20 \lg \left(\frac{A_{\mathrm{signal}}}{A_{\mathrm{noise}}}\right).$$

Причина множителя 10 становится понятной, исходя из приставки «деци», а множитель 20 возникает вследствие умножения на 2, появляющегося при логарифмировании амплитуд.

## 1.2 Примеры практических заданий

На этом занятии познакомимся со средой матричных вычислений Matlab и научимся выполнять в ней простейшие манипуляции с данными.

Название «Matlab» подразумевает работу с матрицами. Соответственно, все данные в Matlab представлены в матричной форме. *Этого не следует забывать при выполнении различных операций*! Чтобы присвоить переменной $A$ значение 10.5, достаточно написать: `A=10.5`. На экране тут же отобразится значение переменной $A$. *Для подавления вывода следует заканчивать команды точкой с запятой*. Попробуйте набрать теперь `A=10.5;` заметили разницу?

Если вы хотите инициализировать эту переменную вектором, напишите, например: `A=[1 2 3]` или же `A=[1,2,3]`. Вы получите вектор-строку с элементами 1, 2 и 3. Таким образом, перечислять элементы строки можно либо через запятую, либо через пробел. Чтобы получить вектор-столбец, следует разделять его элементы точкой с запятой: `A=[1;2;3]`. И, наконец, для ввода матриц: разделяйте элементы строки запятыми или пробелами, а строки — точкой с запятой. Введем единичную матрицу:

```
A=[1 0 0;0 1 0;0 0 1]
```

на экране отобразится:

```
A=
     1   0   0
     0   1   0
     0   0   1
```

Для обращения к элементу матрицы необходимо набрать его имя и в скобках указать либо абсолютный номер, либо *номер строки, а затем номер столбца.*

Наберите теперь `A(5)`, а после — `A(2,2)`. На экране отобразится один и тот же текст: `ans=1`. И действительно, пятый по счету элемент соответствует элементу второй строки второго столбца. Программистам следует обратить внимание, что *номера массивов начинаются с единицы*, в отличие от языков программирования, где первому элементу соответствует ноль.

Допустим, у вас есть матрица-строка, а вы хотите сделать матрицу-столбец. Для этого используйте операцию транспонирования: наберите `x=[1 2 3]`, а затем `x'`. Как видите, добавление после имени матрицы апострофа означает операцию транспонирования.

**Диапазоны данных** можно указывать, разделяя числа двоеточием. Так, `[1:10]` дает набор целых чисел от 1 до 10, а `[1:0.5:10]` — то же, но с шагом в 0.5.

Теперь создайте еще одну матрицу

```
a=[0 1 0;1 0 1;0 1 0]
```

Заодно обратите внимание: *в Matlab имена переменной зависят от регистра*! Для произведения простейших операций над матрицами используйте операторы `+`, `-`, `/` и `*`. Наберите поочередно `A+a`, `A-a`, `A*a`, `A/a`. Обратите внимание, что при этом производятся именно матричные операции. Поэтому необходимо согласовывать размеры матриц (вспомните курс алгебры). Еще одним, чисто матричным оператором, является оператор левого деления `\`, использующийся при решении систем уравнений. Кроме того, используется оператор возведения в степень, `^`.

Но, допустим, вы захотите произвести поэлементное умножение или деление матриц. В этом случае *до символа операции наберите точку*. Например, попробуйте набрать `A./a`. Попробуйте эти операции с более осмысленными векторами или матрицами.

Если вы присвоили переменной *одно* значение, все вычисления с ней выполняются как со скаляром.

Кроме чисел в Matlab есть понятия бесконечности (`Inf`) и не-числа (`NaN`). Вы можете присвоить эти величины переменным. Обратите внимание, что `NaN` соответствует неопределенным операциям, например, пределам вида $0/0$, $\infty/\infty$ и т.п. (однако, попробуйте операцию `Inf/0`). В отличие от `Inf`, `NaN` инвариантен относительно любых операций.

Еще одним удобством является дополнение история команд в стиле UNIX. Набирая команду и нажимая клавишу Tab вы упростите себе работу: если вариант дополнения только один, Matlab закончит команду, иначе — выведет окно выбора со списком возможных команд. Клавишами $\uparrow$ и $\downarrow$ можно перемещаться по истории команд, что полезно, если вам надо набирать много однотипных длинных команд, отличающихся незначительно.

В Matlab существует огромное количество функций. Обычно они имеют стандартный синтаксис типа `A=funk(b,...)`: аргументы функции помещаются в скобки, а ее значение приравнивается матрице. Помощь по функции `funk` можно получить командой `help funk` или нажав клавишу `F1` и воспользовавшись гипертекстовой системой помощи.

▶ Теперь познакомимся с командой `rand`. Эта команда *генерирует равномерно распределенные случайные числа из диапазона* $[0, 1]$. Создайте вектор-строку из ста случайных чисел командой `a=rand(1,100);`. Не забывайте писать в конце команд точку с запятой, иначе вывод (особенно для больших массивов) может занять длительное время. Командой `x=rand(100,100)` можно создать массив из $100 \times 100$ случайных чисел. Аналогично ведет себя команда `x=rand(100)`.

Для **отображения графиков** в Matlab есть команда `plot`. Если ей задан один аргумент, она отображает по оси X номер элемента, а по Y — его значение. Если задать аргументами массивы *одинаковой* длины, первый используется как значения оси X, второй — Y. Постройте график `plot(a)`. Убедились, что $a$ — массив совершенно случайных чисел?

Команда `mean(a)` вернет среднее арифметическое значение переменной $a$ *по столбцам*. Вторым аргументом команды можно указать, по какому именно столбцу мы хотим усреднить массив $a$. Найдите среднее значение своего вектора. Отобразите на графике все значения и их среднее командой

```
plot([1:100], a, [1:100], mean(a))
```

Если по точки двух графиков по оси X совпадают (как в предыдущем случае), можно укоротить запись:

```
plot([1:100],[a; mean(a)])
```

(команда `plot`, если задать ей в качестве аргумента матрицу, изображает разными цветами графики функций, соответствующих каждой строке матрицы).

Функция `normpdf(X,x,s)` позволяет построить график нормального распределения, соответствующий вектору координат X, для математического ожидания x и среднего квадратичного отклонения s. Наберите

```
x=[-70:30]; y=normpdf(x,-20,20);
```

а затем — `plot(x,y)`. На графике вы увидите плотность вероятности нормально распределенной случайной величины с математическим ожиданием -20 и среднеквадратичным отклонением 20.

Сгенерируем синусоидальный сигнал на участке $[0, 2\pi]$ командами

```
x=[0:pi/50:2*pi]; y=sin(x);
```

Теперь добавим к сигналу гауссов белый шум с амплитудой 10 дБ относительно амплитуды сигнала:

```
y1=awgn(y,10,'measured'); plot(x, [y; y1])
```

Третий параметр (measured) обязателен, т.к. без него процесс добавления шума будет несколько иным (мощность сигнала будет считаться равной 0 дБ), можете проверить на синусоиде с амплитудой 10.

▶ Для генерации синусоидального сигнала $y_0 = A\sin(t/T)$ с амплитудной модуляцией по закону $y_1 = f(t)$ необходимо перемножить эти две функции: $y = y_0 \cdot y_1$. Промодулируем синусоиду с периодом $\pi/5$ пилообразным сигналом с периодом 10 на интервале $x \in [0, 20]$. Для генерирования «пилы» используется функция `sawtooth`. Если задать ей один аргумент (вектор $x$), период будет равен $2\pi$, а сигнал будет изменяться в интервале $[-1, 1]$. Чтобы задать смещение максимума, равное $a{\cdot}2\pi$, необходимо указать: `y=sawtooth(x,a)`. Таким образом, чтобы получить «пилу» с интервалом сигнала в $[0, 1]$ и периодом 10, необходимо дать команду `y1=0.5+sawtooth(x*2*pi/10)/2;`. Следовательно, получить наш сигнал можно командой

```
y=sin(x*10).*(0.5+sawtooth(x*pi/5)/2);
```

(не забудьте про точку перед знаком умножения между функциями, иначе получите ошибку, т.к. Matlab попытается перемножить два вектора–строки).

▶ Создадим теперь две синусоиды, сдвинутые на три единицы: `x=[0:0.05:20]; y=sin(x); y1=sin(x+3);`. Попробуем определить, на сколько единиц сдвинут первый сигнал относительно второго. Для этого воспользуемся корреляционной функцией. Запишем `Corr=xcorr(y,y1)`. Корреляционная функция в данном случае имеет вдвое большую ширину, чем исходная, т.к. она получается путем последовательного сдвига второй функции относительно первой. Поэтому построим график командой `plot([-20:0.05:20],Corr)`. Воспользовавшись функцией увеличения можно увидеть, что ближайший к нулю максимум соответствует сдвигу одной функции относительно другой. В нашем случае сигнал был периодическим, поэтому при сдвигах на величины, превышающие половину периода, возникает ошибка, кратная периоду. Это необходимо учитывать в расчетах.

## 1.3 Задания для самостоятельного выполнения

1. Найдите сумму, разность, произведение и частное матриц

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{pmatrix}, \qquad B = \begin{pmatrix} 9 & 8 & 7 \\ 5 & 3 & 1 \\ 0 & 2 & 6 \end{pmatrix}.$$

   Найдите определители исходных и получившихся матриц (команда `det(A)`).
2. Найдите значение почленного, матричного и скалярного произведений векторов $a = (2, 5, 7)$ и $b = 11, 13, 17$. Скалярное произведение найдите двумя способами: путем перемножения векторов и при помощи функции `dot(a,b)`. Найдите векторное произведение $a \times b$ при помощи функции `cross(a,b)`.
3. Постройте график нормального распределения на интервале $[0, 100]$ с математическим ожиданием 50 и дисперсией 100.
4. Получите сигнал с амплитудной модуляции (из примера). Добавьте к нему гауссов белый шум с SNR 100, 50, 10 и 1 дБ. Постройте отдельно графики всех полученных сигналов. Можно ли сделать какой-либо вывод о виде сигнала при SNR=1? Как вы думаете, можно ли восстановить из него исходный сигнал?
5. Для полученного сигнала найдите следующие характеристики: математическое ожидание (`mean`), среднее квадратичное отклонение (`std`), медиану (`median`) и моду (`mode`). Найдите аналогичные величины для разности между зашумленным и оригинальным сигналом. Сравните полученные величины с теоретическими.
6. Попытайтесь определить сдвиг двух синусоид (из примера) при зашумлении:
   - только одной с уровнем сигнал/шум 1 дБ;
   - обеих с уровнем SNR=1 дБ;
   - одной с уровнем SNR=0.1 дБ;
   - обеих с уровнем SNR=0.1 дБ.

   Постройте один из сигналов с SNR=0.1 дБ. Можно ли определить его период? Можно ли определить период по автокорреляционной функции этого сигнала?

**Компьютерная обработка**

# 2 Теория физических измерений. Систематические и случайные погрешности

## 2.1 Краткие теоретические сведения

Развитие науки и техники неразрывно связано с точными измерениями, которые дают новую информацию об окружающем физическом мире. Эксперимент служит основной формой целенаправленного движения к познанию окружающего материального мира, то есть важнейшим инструментом науки. Естественным и наиболее простым способом количественного оценивания свойства является прямое сравнение двух вещей в определенном отношении (по степени проявления свойства). Для стандартизации измерений было разработано соглашение о единицах, используемых для измерений. В метрологии их называют **мерами**.

Результатом сравнения оцениваемой вещи с мерой является именованное число, называемое **значением величины**. Измерения одной и той же величины можно проводить как однократно, так и многократно.

Физические величины можно разделить на два типа: постоянные, изменяющиеся и случайные. К постоянным величинам относят инварианты, чье значение доподлинно известно, физические постоянные, а также величины, однозначно не изменяющие своего значения в процессе измерения. Изменяющиеся величины по определенному закону меняют со временем свое значение. Точное значение случайной величины определить невозможно, можно указать лишь ее дисперсию и среднее значение. Таким образом, эксперимент позволяет измерить с заданной точностью лишь постоянные и изменяющиеся величины. Однако, и в этом случае точность измерения будет ограничена случайными ошибками.

В отличие от классической механики, в квантовой механике многие физические величины являются связанными (например, импульс и координата), поэтому чем больше точность измерения одной из связанных величин, тем меньше точность измерения другой величины (согласно соотношению неопределенности).

**Представление** результатов физического измерения имеет немаловажное значение для понимания сути исследуемого процесса. *Результаты измерения можно представить в виде таблиц или графиков.* Наибольшую наглядность имеет графическое представление. При этом немаловажно правильно задать масштаб графика, направление осей и шкалы осей (линейную, экспоненциаль-

ную, логарифмическую или иную). Например, для отображения зависимости активности радиоактивного изотопа от времени, удобно использовать линейную шкалу для времени и логарифмическую — для активности препарата.

Количественной характеристикой неоднозначности результата измерения является **погрешность**. Ее оценивают, исходя из всей информации, накопленной при подготовке и выполнении измерений. Окончательный результат измерения нельзя расценивать как «истинное значение» измеряемой физической величины, так как в этом нет смысла из-за наличия погрешности.

Погрешность может быть выражена в единицах измеряемой величины $x$ — в таком случае она обозначается $\Delta x$ и носит название **абсолютной погрешности**. Однако абсолютная погрешность не отражает в полной мере качества измерений. Действительно, абсолютная погрешность 1 мм при измерении размера комнаты перед оклеиванием обоями свидетельствует о высоком качестве измерения, но та же погрешность неприемлема при измерении, например, расстояний между атомами в металле.

Критерием качества измерения является безразмерное отношение абсолютной погрешности к окончательному результату измерения, $\delta x = \Delta x/x$, которое называют **относительной погрешностью** и используют как в абсолютном, так и в процентном выражении.

Выделяют следующие виды погрешностей:

**промахи** возникают вследствие неисправности измерительных приборов или ошибок в эксперименте, сделанных по невнимательности;

**систематические погрешности** как правило, неизвестны и могут быть учтены лишь при выполнении измерений несколькими приборами. Данная погрешность возникает вследствие значительной величины влияния прибора на измеряемую величину, либо же из-за пренебрежения некоторыми величинами на этапе моделирования эксперимента;

**случайные погрешности** носят случайный характер (чаще всего они имеют гауссово распределение) и возникают из-за незначительных (флуктуационных) изменений хода эксперимента.

Рассмотрим мысленный эксперимент по изменению физической величины $x$. Пусть в результате $n$ измерений получен ряд значений $x_1, \ldots, x_n$. Для получения действительного значения величины $x$ необходимо, прежде всего выявить промахи, которые могут иметь вид неестественных значений результатов измерения. Далее учитывают систематические (приборные) погрешности, которые имеют вид поправок к результатам. Оставшиеся случайные погрешности обычно характеризуют средним квадратичным отклонением.

Приборные погрешности обычно указываются на лицевой панели прибора или в сопроводительной документации в виде **класса точности** — наибольшей погрешности в процентном соотношении от предела измерений.

В случае, когда $n$ имеет достаточно большую величину, погрешность результата $\sigma_{\overline{x}}$ уменьшается в $\sqrt{n}$ раз по сравнению с погрешностью отдельного измерения $\sigma$, и выражается формулой

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n(n-1)}}.$$

За оценку погрешности окончательного результата многократного измерения примем величину $\Delta x$, задающую симметричный относительно $\overline{x}$ интервал значений, называемый **доверительным интервалом**. Вероятность найти значение измеряемой величины в указанном интервале носит название **доверительной вероятности** $\alpha$. Для нормального распределения существуют таблицы доверительных вероятностей в относительных единицах $\Delta x/\sigma$. Случайную величину обычно записывают в виде

$$x = \overline{x} \pm \Delta x, \quad \alpha = \alpha_0.$$

Идеальным является бесконечное число измерений, однако, обычно ограничиваются пятью-десятью замерами, что приводит к искажению оценки погрешности. В этом случае погрешность рассчитывают по формуле

$$\Delta x_{\text{случ}} = t(\alpha, n)\sigma_{\overline{x}},$$

где $t(\alpha, n)$ – табличные **коэффициенты Стьюдента**.

После измерения случайной погрешности результат записывают в виде $x = \overline{x} \pm \Delta x$, где

$$\Delta x = \sqrt{(\Delta x_{\text{случ}})^2 + \sigma^2_{\text{приб}}}.$$

Следует учесть при выполнении измерений, что *если результаты измерений не выходят за рамки приборной погрешности, можно уверенно считать* $\Delta x = \sigma_{приб}$. Повышение количества измерений в этом случае не способствует повышению точности эксперимента.

Для записи результата следует согласовывать точность измеренной величины и погрешности: количество значащих цифр в них должно совпадать. Для определения количества значащих цифр следует учесть, что *относительная неточность оценивания величины* $\sigma$ *составляет примерно* $(n-1)^{-1/2}$. Таким образом, точность оценки погрешности при выполнении десяти измерений

составляет около 30%, что делает бессмысленным писать более одной значащей цифры (т.е. правильной будет запись в виде $x = 154 \pm 2$ и неправильной: $x = 154.3 \pm 2.1$). Для более наглядного представления результата можно вынести за скобки общий множитель. Например, получив значение $\overline{x} = 3954.2 \cdot 10^3$ при погрешности $\Delta x = 126 \cdot 10^2$ на основании десяти измерений, следует записать результат в виде $x = (395 \pm 1) \cdot 10^4$.

Для определения погрешности в случае, когда *физическая величина определяется аналитически исходя из измерения нескольких физических величин*, необходимо придерживаться следующих правил.

1. Предельная **абсолютная погрешность** суммы или разности равна сумме абсолютных погрешностей. Пусть $Y = X_1 + X_2$, $X_1 = \overline{x_1} \pm \Delta x_1$, $X_2 = \overline{x_2} \pm \Delta x_2$, тогда $\Delta y = \Delta x_1 + \Delta x_2$. В общем виде:

$$\Delta\big(\pm \sum a_n\big) = \sum \Delta a_n.$$

2. Предельная **относительная погрешность** произведения или частного равна сумме относительных погрешностей (при пренебрежении величинами второго и бóльших порядков малости). Как следствие: относительная погрешность $n$-й степени случайной величины равна произведению относительной погрешности этой величины на $n$.

$$\prod(a_i \pm \Delta a_i) = \prod a_i \prod(1 \pm \delta a_i) \approx \prod a_i (1 \pm \sum \delta a_i),$$

$$\big(a[1 \pm \delta a]\big)^n \approx a^n (1 \pm n\delta a).$$

3. **В сложных функциях** вида $y = f(x_1, \ldots, x_n)$ можно оценить погрешность, воспользовавшись приближением:

$$\delta y \approx \frac{dy}{y} = \frac{df(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)}, \tag{2.1}$$

в котором следует заменить $dx_i/x_i = \delta x_i$ – относительная погрешность измерения величины $x_i$, $dx_i = \Delta x_i$ – абсолютная погрешность. Все слагаемые, возникающие при расчете по формуле (2.1) необходимо суммировать по абсолютной величине.

▶ Многие изменяющиеся физические величины имеют линейную зависимость. Оценить параметры такой зависимости можно графически, либо аналитически. Графическая оценка заключается в следующем алгоритме. На график в виде точек наносятся значения измеренных величин. Вокруг каждой точки строится прямоугольник со сторонами, соответствующими погрешностям измеренных величин. Далее между точками проводится прямая так, чтобы по возможности

пересечь все прямоугольники. Если на протяжении всей прямой наблюдается примерно одинаковое количество точек над и под ней, и при этом точки располагаются хаотически (т.е. не образуют групп, последовательно расположенных над или под прямой), данную зависимость можно интерпретировать как линейную. Из коэффициента наклона полученной прямой и точки пересечения ординаты определяются параметры зависимости $X = AY + B$.

Одним из наиболее распространенных приемов статистической обработки экспериментальных данных, относящихся к различным функциональным зависимостям физических величин друг от друга является **метод наименьших квадратов**. В своей простейшей форме он применим к линейной зависимости $y = ax + b$ и позволяет получить достоверные оценки ее параметров, $a$ и $b$, а также оценить их погрешности.

Пусть проведено $n$ парных измерений величин $x$ и $y$. По экспериментальным данным необходимо найти оценки параметров $a$ и $b$, а также оценки их дисперсий $\sigma_a^2$ и $\sigma_b^2$. Предположим, что значения $x_i$ известны без погрешностей, а величины $y_i$ распределены нормально относительно величины $\overline{y_i}$ с одинаковыми дисперсиями $\sigma_y^2$. Математические ожидания $\overline{y_i}$ удовлетворяют выражению $\overline{y_i} = ax_i + b$.

В этом случае критерий минимизации суммы квадратов отклонения $Y = \sum(y_i - \overline{y_i})^2$ позволяет с достаточно большой вероятностью оценить величины $a$ и $b$. Минимизация $Y$ производится по обеим переменным, т.е. необходимо, чтобы выполнялись равенства: $dY/da = 0$ и $dY/db = 0$. Эти два условия образуют систему уравнений, из которых и находятся искомые неизвестные величины:

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2}, \tag{2.2}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - (\overline{x})^2}. \tag{2.3}$$

Соответствующие дисперсии равны:

$$\sigma_y^2 = \frac{n}{n-2}\Big(\overline{y^2} - (\overline{y})^2 - a^2\big[\overline{x^2} - (\overline{x})^2\big]\Big), \qquad \sigma_a^2 = \frac{\sigma^2}{n\big(\overline{x^2} - (\overline{x})^2\big)}, \quad \sigma_b^2 = \sigma_a^2\overline{x^2}.$$

Найденные величины позволяют **интерполировать** (отыскать промежуточные значения функции) и **экстраполировать** («продлить» функцию за пределы табличных значений) заданную таблично функцию $y(x)$.

▶ В случае функции двух и более переменных $z \equiv z(x, y) = ax + by + c$ коэффициенты $a$, $b$ и $c$ также можно отыскать методом наименьших квадратов.

Единственным отличием в данном случае будет то, что система будет состоять из трех уравнений для трех неизвестных.

Теперь предположим, что у нас имеется нелинейная функция $y = f(x)$. Даже в этом случае возможно использовать метод наименьших квадратов, но формулы (2.2) и (2.3) могут значительно видоизмениться. Некоторые зависимости, однако, можно свести к линейным. Например, функция $y = \mathrm{e}^{ax+b}$ после логарифмирования примет вид $\ln y = ax + b$. Заменяя переменную $y_1 = \ln y$ (т.е. прологарифмировав измеренные величины $y_i$) можно определить коэффициенты $a$ и $b$ по тем же формулам, что и для линейной зависимости. Однако, следует иметь в виду, что в этом случае дисперсии $\sigma$, $\sigma_a$ и $\sigma_b$ будут вычисляться немного иначе.

## 2.2 Примеры практических заданий

Случайная погрешность физического измерения имеет природу, аналогичную белому шуму, поэтому для начала рассмотрим простейшие методы очистки одномерных сигналов вида $y = y(t)$ от шумов.

Используем сигнал из задания 4 лабораторной работы №1, но создадим массив из десяти сигналов, зашумленных с одинаковым уровнем SNR:

```
x=[0:0.05:20];
y=sin(x*10).*(0.5+sawtooth(x*pi/5)/2);
for a=[1:10]
y1(a,:)=awgn(y,1,'measured');
end
```

Вид цикла `for` отличается от языков программирования вроде C: цикл поочередно перебирает все значения переменной `a`. Если бы мы заранее инициализировали ее массивом, можно было бы просто написать `for a`. Цикл `for` заканчивается командой `end` (ею же заканчиваются и многие другие циклы). Двоеточие в адресации `y(a,:)` означает, что мы выбираем **все** элементы по второй координате (т.е. приравнивание производится к целой строке). Еще одним отличием от языков программирования является динамическое расширение матриц: нет необходимости в начале работы с ней сообщать ее предельный размер.

Итак, мы получили массив `y1`, в строках которого содержатся зашумленные варианты одного и того же сигнала. Можете отобразить их все графически командой `plot(x,y1)`. Оценить зашумленность сигнала можно командой `plot(y,y1)`. Если бы сигналы в `y1` совпадали с `y`, мы увидели бы отрезок с коэффициентом наклона 1. Чем дальше форма полученной фигуры от такого отрезка, тем больше зашумленность сигнала.

Для восстановления сигнала из десяти измерений попробуем усреднить наборы сигналов и найти их медиану:

```
y_mean = mean(y1);
y_med = median(y1);
plot(x,[y;y_mean;y_med])
```

Как видите, оба восстановленных сигнала имеют примерно одинаковые величины и довольно близки к реальной функции (особенно на участках с большой амплитудой сигнала). Однако, как мы увидим впоследствии, если к сигналу добавлен шум типа «соль/перец», медианная фильтрация будет работать намного эффективнее фильтрации по среднему арифметическому.

▶ Теперь допустим, что мы имеем линейную зависимость $y = ax + b$, заданную таблично в виде $y(x)$. Для определения методом наименьших квадратов коэффициентов линейной (а также высших степеней) зависимости служит функция `polyfit(x,y,n)`. Она содержит три аргумента: $x$ – вектор аргумента, $y$ – вектор функции, $n$ – степень аппроксимирующего полинома. Ее результат в простейшем случае представляет собой вектор коэффициентов (начиная со старшей степени). Если функцию вызвать как `[p,S] = polyfit(x,y,n)`, вектор $p$ будет содержать коэффициенты, а в структуре $S$ будут содержаться такие данные, как степени свободы (df) и норма отклонений данных от аппроксимирующей кривой (normr). Для восстановления полученной зависимости используется функция `polyval(p,x)`, где $p$ – полученный функцией `polyfit` вектор коэффициентов, $x$ – вектор аргумента. В таком виде функция возвращает вектор восстановленной функции. В виде `[y, delta] = polyval(p,x,S)` функция возвращает массив погрешностей (т.е. в каждой точке восстановленные значения функции можно представить в виде $y = y \pm delta$, т.е. оценить абсолютную погрешность восстановления можно при помощи команды `mean(delta)`.

Найдем коэффициенты модельной зависимости. Пусть $y = 7.15x + 4.22$. Построим вектора, соответствующие аргументу и функции:

```
x = [0:100]; y = 7.15*x + 4.22;
```

Зашумим сигнал для получения разброса точек $y_i$:

```
y1 = awgn(y,10,'measured');
```

Отобразим на экране оба ряда: `plot(x,y,x,y1,'.')` (запись `'.'` означает, что график будет отображаться точками). Как видите, разброс данных достаточно велик. Определим коэффициент корреляции: `corrcoef(x,y1)`. Он довольно близок к единице, следовательно, мы можем попытаться получить коэффициенты линейной зависимости и восстановить функцию:

```
[p,S] = polyfit(x,y1,1);          % коэффициенты a и b
[y2, delta] = polyval(p,x,S);     % восстановленный вектор
plot(x,y1,'.',x,[y;y2])           % все три графика, зеленый - наш
mean(delta)                       % абсолютная ошибка
mean(delta)/mean(y)               % относительная ошибка
```

▶ Можно найти приближение методом наименьших квадратов и другим способом. Пусть $Y$ – вектор–столбец значений функции, $A = (ab)$ – вектор–столбец коэффициентов разложения. Тогда условие $y_i = ax_i + b$ можно представить в виде матричного произведения $Y = XA$. Второй столбец матрицы $X$ целиком состоит из единиц, а в первом находится последовательность значений $x_i$. В этом случае нахождение коэффициентов сводится к решению системы линейных уравнений $y_i = ax_i + b$, дающему минимальную невязку. Такое решение находится при помощи операции левостороннего матричного деления: $X\backslash Y$. Решим предыдущий пример таким способом.

```
X = [x' ones(size(x'))]; % создаем матрицу аргумента
                         % (т.к. x и y1 - строки, транспонируем их)
A = X\y1'                % находим коэффициенты
                         % и отображаем их на экране
```

Функция `ones` создает матрицу с единичными элементами размером, соответствующим значению аргументов (или вектор с единичными элементами, если аргумент один). Функция `size` возвращает вектор, каждый элемент которого соответствует количеству элементов аргумента в данной размерности (а количество элементов вектора равно размерности матрицы–аргумента). Все, что следует за знаком % Matlab игнорирует, считая это комментарием.

Полученные значения должны быть примерно равны найденным предыдущим способом. Как мы увидим далее, такой способ нахождения корней аппроксимации пригоден не только для полиномиальных, но и для многих других функций.

▶ Попробуем создать квадратичную зависимость и аппроксимировать ее методом наименьших квадратов. Пусть зависимость на отрезке $[0, 100]$ имеет вид $y = 2.4x^2 - 0.87x + 2.13$. Создадим соответствующие массивы данных, добавим шум с SNR=20 дБ и отобразим оба сигнала на графике:

```
x=[1:100];
y=2.4*x.^2-0.87*x+2.13;
y1=awgn(y,20,'measured');
plot(x,[y;y1]);
```

Теперь создадим вектор коэффициентов аппроксимации полиномом второй степени восстановим функцию и отобразим на графике:

```
[p, S] = polyfit(x, y1, 2);
[y2, DELTA] = polyval(p, x, S);
plot(x,[y1;y2]);
```

Сравните функции $y$ и $y_2$: `plot(x,[y;y2])`. Отобразите на экране найденные коэффициенты: `p`. Рассчитайте среднее квадратичное отклонение аппроксимации (`mean(DELTA)`). Также рассчитайте относительную ошибку аппроксимации `mean(DELTA)/mean(y1)`.

▶ Однако, чаще всего функциональные зависимости имеют иные виды зависимости. Допустим, нам известно, что измеряемая величина изменяется по закону

$$y = a_0 + a_1 \mathrm{e}^{-t} + a_2 t e^{-t}. \tag{2.4}$$

Для аппроксимации такой функцией можно представить уравнение (2.4) в матричном виде $Y = TA$, где $T$ – функциональная матрица, у которой в первом столбце размещены единицы (соответствует умножению на $a_0$), во втором — функция $\mathrm{e}^{-t}$, а в третьем — $t\,\mathrm{e}^{-t}$. Найти коэффициенты $A$ можно при помощи оператора левого деления: $A = T\backslash Y$. Для закрепления материала выполните следующий пример.

```
t = [0 0.3 0.8 1.1 1.6 2.3]';      % сразу вводим данные в столбцах
y = [0.6 0.67 1.01 1.35 1.47 1.25]';
T = [ones(size(t))  exp(-t)  t.*exp(-t)];
A = T\y
```

Теперь отобразим данные на графике:

```
x = (0:0.1:2.5)';
Y = [ones(size(x))  exp(-x)  x.*exp(-x)]*A;
plot(x,Y,'-',t,y,'o')
```

Указание `'-'` означает, что график построится в виде непрерывной линии (его можно опустить); `'o'` выведет график в виде точек, каждая из которых будет обозначена кружками.

## 2.3 Задания для самостоятельного выполнения

1. На интервале `[0:0.05:20]` задайте функцию `y=sin(x)`. Создайте три массива: на 10, 100 и 1000 зашумленных копий функции `y` с SNR=-10. Отобразите одну из этих функций. Найдите для каждого массива медиану. Постройте графики всех трех медиан. Для отображения трех разных графиков на одном листе можно воспользоваться функцией `subplot(abc)`, где `a` – количество графиков по вертикали, `b` – количество графиков по горизонтали, `c` – номер текущего графика. Например, чтобы построить три графика один над другим, надо дать последовательность команд:

   ```
   subplot(311); plot(график1)
   subplot(312); plot(график2)
   subplot(313); plot(график3)
   ```

   Как видите, даже когда амплитуда шума в три раза превышает амплитуду сигнала, можно выделить сигнал чисто статистическими методами, однако, при этом необходимо провести огромное количество измерений.
2. Найдите *обоими способами* коэффициенты $a$ и $b$ для таблично представленной зависимости $y(x)$, предполагая, что она имеет линейный вид. Найдите матрицу коэффициентов корреляции $x$ и $y$ (командой `corrcoef(x,y)`. Элемент матрицы с координатами $(i, j)$ соответствует корреляции $i$-го и $j$-го столбцов. Таким образом, следует обратить внимание на элементы, расположенные на побочной диагонали. Если они близки к 1, можно говорить о линейной зависимости данных. Данные представлены в таблице:

| **x** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **y** | 7.7 | 13.7 | 22.0 | 23.1 | 23.7 | 36.7 | 35.6 | 47.8 | 50.2 | 52.1 |

   ($a = 5.1$, $b = 3.4$).
3. Известно, что некоторая зависимость (см. таблицу ниже) имеет вид $y = ax\sin(x) - b\ln(x)$. Определите коэффициенты $a$ и $b$ и постройте данную кривую с более детальным отображением (на векторе `[1:0.05:10]`). Подсказка: сразу же задайте вектора $x$ и $y$ как столбцы; матрица $X$ задается командой `X=[x.*sin(x)  -log(x)]`.

| **x** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **y** | -0.68 | 8.41 | -23.0 | -37.2 | -73.2 | -39.7 | 9.14 | 21.0 | 7.97 | -72.5 |

   ($a = 7.72$, $b = 14.8$).
4. Промоделируйте эксперимент измерения ста значений функции $y = 1.7x^3 + 3.4x^2 - 2.9x + 9.2$ и восстановления коэффициентов зависимости. Для этого создайте вектор аргумента `x=[1:100]`, получите по формуле соответствующий вектор функции `y_ideal`, а из него — зашумленный результат `y` с

SNR=25 дБ.

Методом `polyfit` — `polyval` получите значения коэффициентов. Отобразите на графике точками исходные данные и прямой линией полученный интерполяцией результат.

5. Аналогично предыдущему заданию составьте модель эксперимента по измерению амплитуды напряжения в контуре, испытывающем колебания с основной частотой $\Omega = 1000\,\text{Гц}$ и двумя гармониками $\Omega \pm \omega$, где $\omega = 74\,\text{Гц}$. Известно, что суммарное колебание описывается приближенной формулой $U = a\sin(\Omega t) + b\sin(\omega t) - c\cos(\omega t)$. Создайте интервал времен `t=[0: 0.06: 120]`. Для получения идеальных значений $U$ положите $a = 361$, $b = 117$, $c = 92$. Отношение сигнал/шум при получении зашумленного сигнала выберите равным 20 дБ.

   Восстановите значения коэффициентов $a$, $b$ и $c$.

# 3 Теория оценок

## 3.1 Краткие теоретические сведения

Для вычисления вероятности попадания случайной величины, имеющей нормальное распределение, в заданный интервал $[a, b]$, используются функции Лапласа. Вероятность того, что случайная величина $x$ лежит в интервале $[\overline{x} - \delta, \overline{x} + \delta]$ равна

$$P(|x - \overline{x}| < \delta) = 2\Phi(\delta/\sigma).$$

Если $\delta = 3\sigma$, вероятность попадания случайной величины в данный отрезок будет равна 0.9973%. Другими словами, лишь 0.27% наблюдаемых могут располагаться вне этого интервала. Этот вывод называют **правилом трех сигм**: *если случайная величина распределена нормально, то абсолютная величина ее отклонения от математического ожидания не превосходит утроенного среднего квадратичного отклонения.*

Согласно **теореме Ляпунова**, *случайная величина, являющаяся суммой большого числа взаимно независимых случайных величин, имеет нормальное распределение.* Так как результат физического измерения определяется огромным количеством случайных величин (давление, температура, магнитное поле Земли и т.п.), любой результат физического измерения имеет нормальное распределение.

Так как при измерении реальных (эмпирических) случайных величин ограничиваются конечным (и обычно небольшим) количеством измерений, вид распределения данной величины может отличаться от гауссова распределения. Пусть $x_i$, $i = \overline{1, n}$ — нормальные независимые случайные величины с $<x> = 0$ и $\sigma_x = 1$. Тогда сумма квадратов этих величин $\chi^2 = \sum_{i=1}^{n} x_i^2$ распределена по закону «**хи квадрат**» с $k = n$ степенями свободы. Каждое линейное соотношение между этими величинами уменьшает количество степеней свободы распределения на единицу. Плотность этого распределения:

$$f(x) = \begin{cases} 0, & x \leqslant 0, \\ \frac{1}{2^{k/2}\Gamma(k/2)}\, \mathrm{e}^{-x/2}\, x^{k/2-1}, & x > 0, \end{cases}$$

где $\Gamma(x) = \int\limits_0^\infty t^{x-1} e^{-t} dt$ – гамма-функция, в частности, $\Gamma(n+1) = n!$. Отсюда видно, что с увеличением степеней свободы распределение «хи квадрат» при-

ближается к нормальному распределению (что соответствует закону больших чисел).

Если $z$ – нормальная случайная величина с $\overline{z} = 0$ и $\sigma_z = 1$, а $v$ – независимая от $z$ величина, распределенная по закону $\chi^2$ с $k$ степенями свободы, то величина $T = Z/\sqrt{V/k}$ имеет распределение, которое называют **t-распределением** или **распределением Стьюдента**. Данное распределение при увеличении $k$ тоже приближается к нормальному распределению.

Из-за того, что количество экспериментов (т.е. наблюдаемых величин) при измерении физической величины ограничено конечным (небольшим) числом, для определения как можно более точного значения математического ожидания данной величины используется *теория оценок*.

**Несмещенной** называют оценку величины $x$, совпадающую с $\overline{x}$, однако, действительные оценки являются смещенными. **Эффективной** называют оценку с наименьшей дисперсией. **Состоятельной** называют оценку, стремящуюся при увеличении количества наблюдаемых к $\overline{x}$.

Аналогично характеристикам случайных величин, для выборок вводят понятие группового и общего среднего, генеральной и выборочной дисперсии. В качестве оценки дисперсии выборки обычно принимают **исправленную дисперсию** $D = \frac{n}{n-1} D_{\text{выб}}$. Т.е. при расчете дисперсии суммы квадратов отклонений наблюдаемых от их среднего значения необходимо делить на число наблюдаемых без единицы.

**Точностью** оценки (или доверительным интервалом) $\delta$ называют длину полуинтервала, в котором с определенной вероятностью находится измеряемая величина. Соответствующую вероятность $\gamma$ называют **надежностью** оценки. Т.о., получим соотношение:

$$P(|x - \overline{x}| < \delta) = \gamma.$$

Примем, что если случайная величина $x$ распределена нормально, ее выборочная средняя $<x>$ также распределена нормально. В этом случае $\overline{<x>} = a$ и $\sigma_{<x>} = \sigma/\sqrt{n}$, где $a$ и $\sigma$ – математическое ожидание и среднеквадратичное отклонение случайной величины $x$. В этом случае для надежности оценки получим:

$$\gamma \equiv P(<x> - \delta < a < <x> + \delta) = 2\Phi(\delta\sqrt{n}/\sigma).$$

Однако, чаще всего для данной наблюдаемой $\sigma$ неизвестно. В этом случае можно оценить доверительный интервал при помощи распределения Стьюдента. В этом случае параметры распределения определяются так:

$$\gamma \equiv P(\Big|\frac{<x> - a}{S/\sqrt{n}}\Big| < t_\gamma) = 2\int_0^{t_\gamma} S(t,n)\,dt, \qquad \delta = \frac{t_\gamma S}{\sqrt{n}},$$

где $S = \sqrt{D}$ – «исправленное» среднеквадратичное отклонение, параметр $t_\gamma$ для данных $n$ и $\gamma$ можно найти в соответствующих таблицах.

Итак, оценку истинного значения измеряемой величины производят при помощи доверительного интервала. При этом параметр $a$ считают равным выборочному среднему $<x>$.

Оценку среднеквадратичного отклонения $\sigma$ случайной величины $x$ производят при помощи критерия «хи квадрат». Вероятность $\gamma = P(|\sigma - S| < \delta)$ в этом случае равна

$$\gamma = \int\limits_{\frac{\sqrt{n-1}}{1+\delta/S}}^{\frac{\sqrt{n-1}}{1-\delta/S}} R(\chi, n)\, d\chi \equiv \mathcal{X}(n, S).$$

Вычислив по выборке $S$ и найдя по специальной таблице для $S$ и заданной $\gamma$ величину $q = \delta/S$, получим искомый доверительный интервал.

Для оценки других характеристик распределения используется **метод моментов**, согласно которому *начальные и центральные эмпирические моменты являются состоятельными оценками соответствующих начальных и центральных теоретических моментов.*

## 3.2 Примеры практических заданий

Найдем общую среднюю совокупности, состоящей из следующих трех групп:

| Группа | I | | II | | III | |
|---|---|---|---|---|---|---|
| Значение признака | 1 | 3 | 2 | 4 | 3 | 6 |
| Частота признака | 11 | 34 | 22 | 28 | 31 | 14 |
| Объем выборки | $11 + 34 = 45$ | | $22 + 28 = 50$ | | $31 + 14 = 45$ | |

Для начала найдем групповые средние: $<x_1>$, $<x_2>$ и $<x_3>$:

```
>> x1 = (11*1 + 34*3)/45
x1 =
    2.5111
>> x2 = (22*2 + 28*4)/50
x2 =
    3.1200
>> x3 = (31*3 + 14*6)/45
x3 =
    3.9333
```

Теперь найдем общую среднюю по групповым средним:

```
>> X = (x1*45 + x2*50 + x3*45)/(45 + 50 + 45)
X =
    3.1857
```

Однако, при работе с большими массивами данных лучше использовать преимущества матричной алгебры:

```
>> xi = [1 3 2 4 3 6];
>> ni = [11 34 22 28 31 14 ] ;
>> N = sum(ni)
N =
   140
>> X = sum(xi.*ni/N)
X =
    3.1857
```

Найдем *генеральную дисперсию* и генеральное среднеквадратичное отклонение данной выборки:

```
>> D = sum(ni.*(xi-X).^2)/N
D =
    1.5369
>> sigma=sqrt(D)
sigma =
    1.2397
```

Кроме того, определить среднеквадратичное отклонение ряда $x$ можно при помощи команды `std(x)`.

▶ Теперь рассмотрим ряд измерений некоторой физической величины $x$. Результаты серии измерений заданы таблицей ($\nu_i$ – частота соответствующего значения $x_i$):

| $x_i$ | 31 | 28 | 34 | 26 | 35 | 30 | 34 | 32 | 40 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu_i$ | 20 | 12 | 10 | 5 | 7 | 20 | 12 | 19 | 4 | 2 |

Известно, что некоторые результаты могут быть заведомо ошибочными. Нам необходимо оценить среднее значение данной величины, исключив ошибочные результаты. Составим массивы величины $x$ и соответствующих частот $n$:

```
>> x=[ 31 28 34 26 35 30 34 32 40 20];
>> n=[ 20 12 10  5 7 20 12 19 4 2];
```

Отобразив данные на графике (`plot(x,n,'o')`) можно заметить, что действительно некоторые значения сильно отклоняются от положения, которое они занимали бы при нормальном распределении.

Найдем среднее значение величины $x$ и ее среднеквадратичное отклонение:

```
>> X=sum(x.*n)/sum(n)
X =
   31.4144
>> sigma=sqrt(sum(n.*(x-X).^2)/sum(n))
sigma =
    3.1891
```

Определим границы доверительного интервала $[a, b]$ в пределах трех $\sigma$:

```
>> a=X-3*sigma
a =
   21.8470
>> b=X+3*sigma
b =
   40.9818
```

Теперь исключим из выборки значения, выходящие за пределы интервала. При помощи функции `find` можно найти индексы членов массива, удовлетворяющих заданному условию. Наберите `find(x<a)`. На экране отобразится индекс единственного элемента, выходящего за границы данного интервала: 10. Исключить его из дальнейших расчетов можно, приравняв соответствующую ему частоту нулю: `n(find(x<a))=0;`.

Теперь повторим вычисление `X` и `sigma`, `a` и `b`. *Для того, чтобы вызвать из истории команд строку, начинающуюся с определенных символов, наберите один–два первых символа и нажмите клавишу «вверх».* Таким образом можно быстро вызвать из истории команд нужную вам команду, не перебирая все промежуточные.

Теперь проверим, не влияет ли «подозрительное» значение $x = 40$ на точность измерения. Найдем медиану нашего ряда и оценим доверительный интервал по медиане. Для этого нам необходимо построить новый вектор `newx`, в котором значения величины $x$ будут содержаться столько раз, какова их частота:

```
>> for a = [1:length(n)]
newx = [newx ones(1,n(a)).*x(a)];
end
med = median(newx)
```

```
med =
    31
>> b=med+3*sigma
b =
   39.4442
>> a=med-3*sigma
a =
   22.5558
```

Действительно, значение $x = 40$ выбивается из доверительного интервала. Приравняем его к нулю, и найдем $<x>$, близкое к истинному:

```
>> n(find(x>b))=0;
>> X=sum(x.*n)/sum(n)
X =
   31.3048
>> sigma=sqrt(sum(n.*(x-X).^2)/sum(n))
sigma =
    2.3345
>> a=X-3*sigma
a =
   24.3012
>> b=X+3*sigma
b =
   38.3083
>> find(x>b)
ans =
     9
>> find(x<a)
ans =
    10
```

Итак, кроме уже исключенных значений, все значения $x_i$ удовлетворяют критерию «трех сигм», следовательно, можно записать ответ: $x = 31.3 \pm 2.3$.

▶ Теперь оценим значения величины $<x>$ при помощи доверительного интервала с надежностью 95% при помощи распределения Стьюдента. Для этого в MATLAB существует функция `ttest`. В простейшем случае вида `h=ttest(x)` она возвращает вероятность отклонения гипотезы о нормальном распределении величины $x$ с математическим ожиданием $\overline{x} = 0$. Проверка даст результат: 1. Действительно, математическое ожидание нашей величины далеко не равно нулю. Второй аргумент функции `ttest` задает предполагаемое математическое

ожидание. Введите `h=ttest(x,X)`. Вы получите ответ: `h=0`. Т.е., можно принять гипотезу о гауссовой форме распределения величины $x$ около ее среднего значения. Оценить 95%-й доверительный интервал величины $x$ можно при помощи расширенного вывода функции `ttest` в форме `[h,p,ci]=ttest(x,X)`. В этом случае параметр `h` сообщает о степени ненадежности гипотезы, `p` равен вероятности совпадения величины `X` с математическим ожиданием ряда `x`, `ci` сообщает границы 95%-го доверительного интервала. Определим доверительный интервал для нашего ряда без исключения заведомо ложных результатов и с их исключением:

```
>> [h,p,ci]=ttest(x,X)
h =
     0
p =
    0.8647
ci =
   27.0673   34.9327

>> [h,p,ci]=ttest(x(1:8),X)
h =
     0
p =
    0.9622
ci =
   28.6157   33.8843
```

Итак, в обоих случаях гипотеза о соответствии распределения величины $x$ нормальному распределению принимается, однако, во втором случае вероятность определения математического ожидания $\overline{x}$ выше, и доверительный интервал уже, что явно свидетельствует о большей надежности вычислений. *Середина доверительного интервала $\overline{x} = 31$ совпадает с медианой ряда, что лишний раз подтверждает бóльшую степень надежности медианного усреднения по сравнению со средним арифметическим.*

Matlab предоставляет огромный набор инструментальных средств. Однако, как вы уже могли заметить, при работе с большим количеством однообразных данных приходится много раз повторять одни и те же команды. Эту задачу можно упростить, создав **скрипт** (или m-файл). Скрипт представляет собой описание и реализацию пользовательской функции, которая вызывается из командной строки Matlab аналогично любой команде, однако может содержать значительное количество инструкций, облегчающих работу пользователя.

M-файл может содержать любые инструкции. Если он не начинается со слова `function`, выполняется все его содержимое, однако если вы хотите использовать какие-либо переменные в этом файле, их придется инициализировать заранее, причем их имена должны быть абсолютно такими же, как и в m-файле. Удобнее, однако, создать m-файл в виде функции, принимающей в качестве аргументов необходимые переменные и возвращающей определенные величины.

Заголовок файла функции имеет вид

```
%
% Комментарий, отображающийся при введении команды help имя_функции
%
function [возвращаемые величины] = имя_функции(входные, аргументы)
```

Далее следуют операторы, выполняемые в теле функции. Если после команды вы пропустите символ точки с запятой, ее вывод будет отображен на экране.

Итак, создадим m-файл, осуществляющий проверку выборки на корректность при помощи критерия «трех сигм». Зайдите в меню File $\Rightarrow$ New $\Rightarrow$ M-File. Появится окно редактора. Сохраните файл в рабочей папке под именем `three_s.m`. Наберите содержимое файла.

```
1  % three_s.m
2  % [ X sigma ] = three_s(x, n)
3  % Производит отбор выборки x с соответствующими частотами n
4  % при помощи критерия "трех сигм"
5  % результат: среднее значение X и его среднеквадратичное отклонение, sigma
6
7  function [ X sigma ] = three_s(x, n)
8  newx = []; % вспомогательный массив
9  Data = [x ; n]; % совмещенный массив данных
10 X = sum(x.*n)/sum(n); % среднее арифметическое
11 sigma = sqrt(sum(n.*(x-X).^2)/sum(n)); % среднеквадратичное отклонение
12 down = X-3*sigma; % нижняя граница доверительного интервала
13 up = X+3*sigma;   % верхняя граница -//-
14 a=find(x < down); % a и b - массив координат, выходящих за границы
15 b=find(x > up);
16 while (length(a)>0) || (length(b)>0) % пока есть неверные значения
17 Data = Data(:, find(Data(1, find(Data(1,:) >= down)) <= up)); % выбрасываем их
18 x = Data(1,:);
19 n = Data(2,:);
20 X = sum(x.*n)/sum(n);
21 for a = [1:length(n)]
22 newx = [newx ones(1,n(a)).*x(a)];
23 end
24 X = median(newx);
```

```
25 sigma = sqrt(sum(n.*(x-X).^2)/sum(n));
26 down = X-3*sigma;
27 up = X+3*sigma;
28 a = find(x < down);
29 b = find(x > up);
30 end
```

Обратите внимание, что при запуске скрипта текущей папкой Matlab должна быть папка, в которую вы его сохранили.

Попробуйте вначале для того, чтобы лучше разобраться с принципом работы данного скрипта, не ставить точку с запятой в конце строк 17, 24, 25, 28 и 29. Запустить скрипт можно командой `[X sigma] = three_s(x,n)`.

▶ Зачастую физику-экспериментатору приходится проверять нулевую гипотезу о равенстве средних двух независимых наборов данных. Пусть в результате одного измерения некоторой физической величины $x$ был получен ряд данных:

```
x1 =  47.78 36.40 35.66 8.93 40.42 54.16 51.76 44.32 46.19 50.75
```

Затем было произведено независимое измерение этой же физической величины при других условиях эксперимента. При этом был получен ряд:

```
x2 =  44.09 46.75 44.20 7.99 47.74 75.07 62.48 44.43 34.73 55.26
```

Требуется проверить нулевую гипотезу о равенстве математических ожиданий данных величин.

Для проверки данной гипотезы существует функция Matlab `ttest2`. Наберите `ttest2(x1,x2)`. В ответ вы получите: `ans=0`, т.е. гипотеза о неравенстве математических ожиданий наших двух рядов отклонена на 95%-м уровне. Для определения доверительного интервала и вероятности равенства математических ожиданий воспользуемся расширенным выводом команды:

```
>> [h p ci] = ttest2(x1, x2)
h =
     0
p =
    0.5122
ci =
  -19.2074    9.9334
```

Таким образом, вероятность того, что математические ожидания выборок равны, составляет лишь $p = 51\%$, при этом доверительный интервал математического ожидания разности $x_1 - x_2$ достаточно широк: $c_i = [-19.2, 9.9]$, т.е.

математические ожидания данных рядов могут разниться на 4.6 со среднеквадратичным отклонением $\sigma = 14.6$.

Большая ширина доверительного интервала говорит о том, что данные в рядах $x_1$ и $x_2$ получены с низкой надежностью. Однако, найдя медианы рядов $x_1$, $x_2$ и совмещенного ряда $(x_1; x_2)$ можно попытаться с достаточно высокой степенью вероятности оценить математическое ожидание величины $x$.

## 3.3 Задания для самостоятельного выполнения

1. Некоторая совокупность состоит из трех групп: $X_1$, $X_2$, и $X_3$. Группы имеют следующие значения:

```
X1 =  35.04 35.45 35.01 34.94 34.63 35.11 34.41 35.29 35.69
      34.69 35.36 35.53 34.30 34.36 35.23
X2 =  34.30 34.80 34.86 34.81 35.08 34.79 35.04 33.93 34.48
      34.41 33.74 34.60 34.00
X3 =  35.17 34.21 34.78 34.65 34.16 33.62 34.53 34.12 34.82
      34.77 35.29 34.81 34.28 34.72 34.12 34.55 34.53 34.55
```

   Найдите: групповые средние (35,00, 34.53, 34.54), общее среднее (34.69), групповые дисперсии (0.19, 0.19, 0.16), генеральную дисперсию (0.22).
2. Усовершенствуйте скрипт `three_s.m` так, чтобы помимо основных вычислений на экране отображались среднее арифметическое значение массива с данными, а также 95%-й доверительный интервал по критерию Стьюдента.
3. Определите математическое ожидание и среднеквадратичное отклонение рядов данных $x_1$ и $x_2$ (пример со стр. 29) по-отдельности, а затем вместе. Сравните результаты со значениями, полученными при помощи функции `three_s` (вторым аргументом функции должен быть массив единиц: `ones(1,10)` для проверки рядов по-отдельности). Для совместной проверки рядов формат вызова функции следующий:

```
>> [mean, sigma] = three_s([x1 x2],ones(1,20))
```

   Аналогично следует вызывать функцию `std` для определения среднеквадратичного отклонения рядов:

```
>> std(x1)
>> std(x2)
>> std([x1 x2])
```

4. Определите давление в цилиндре с газом, исходя из закона Менделеева–Клапейрона: $pV = mRT/\mu$, если известно, что масса газа $m = 2$ грамма, $\mu = 29$ г/моль, $R = 8.31$, а объем и температуру газа измеряли в течение

минуты, получив следующие значения:

| Величина | Значение | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $V$, л | 2.27 | 2.27 | 2.26 | 2.25 | 2.26 | 2.27 | 2.29 | 2.28 | 2.25 | 2.28 |
| $T$, К | 399.4 | 399.1 | 399.3 | 396.8 | 399.5 | 400.2 | 400.6 | 403.0 | 399.2 | 401.3 |

Считайте, что за это время давление газа не успело сколь-нибудь значительно измениться. Определите погрешности измерения величин $V$ и $T$. Считая, что остальные величины являются постоянными, определите косвенную погрешность измерения $p$.

Для удобства вычислений *создайте скрипт, позволяющий для заданного ряда данных получить математическое ожидание, среднеквадратичное отклонение и относительную ошибку.*

Запишите результат в виде $p = \overline{p} \pm \sigma_p$ ($p = 101 \pm 1\,\text{кПа}$).

5. Для определения емкости $C$ неизвестного конденсатора при помощи осциллографа исследовали затухающий импульс, возникающий при разрядке конденсатора через резистор $R = 3\,\text{кОм}$. По показаниям осциллографа были записаны следующие значения тока:

| $t$, с | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $I$, А | 1.00 | 0.72 | 0.52 | 0.37 | 0.26 | 0.19 | 0.14 | 0.10 | 0.07 | 0.04 | 0.03 |

Известно, что погрешность считывания значений тока с экрана осциллографа составляет $\sigma_I = 0.01\,\text{А}$. Кроме того, известно что сопротивление резистора известно с точностью 3%. Из формулы $I = I_0 \exp(-t/[RC])$ определите погрешность измерения емкости конденсатора.

Методом наименьших квадратов определите значение емкости конденсатора, исходя из уравнения $t = -RC \ln I$ (составьте матрицу `X=-R*log(I')` и найдите решение: `C=X\t` (97 мкФ). Запишите ответ в виде $C = \overline{C} \pm \sigma_C$.

Для увеличения точности эксперимента было проведено еще одно измерение, результаты которого несколько отличались от предыдущих:

| $t$, с | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $I$, А | 1.00 | 0.75 | 0.56 | 0.41 | 0.30 | 0.23 | 0.17 | 0.12 | 0.10 | 0.07 | 0.05 |

Проверьте нулевую гипотезу о равенстве средних в обоих опытах. Определите величину емкости во втором случае (112 мкФ).

Столь большое различие емкостей, полученных в результате двух независимых экспериментов, заставило предположить, что в результате длительной эксплуатации резистор $R$ нагрелся, что вызвало увеличение его сопротивления. Считая емкость конденсатора прежней, определите сопротивление резистора во втором случае (3.5 кОм).