Методические указания
по выполнению лабораторной работы № 6
по курсу «Обработка данных в точных измерительных системах»

Суть обработки данных заключается в получении количественных характеристик обрабатываемых данных (возможно, неточных), либо результатов таких их преобразований, что оказываются удобными для дальнейшего принятия решений. Следовательно, суть обработки данных – в выполняемых вычислениях.

Если исходные данные, подвергаемые обработке, неточны (например, являются результатами измерений), то неточны и любые результаты их обработки. Назначение данного курса заключается в том, чтобы научиться с помощью программных средств автоматически определять характеристики погрешности результатов вычислений с неточными данными, унаследованные ими от неточных исходных данных.

Программное обеспечение

Основной рекомендуемый инструмент для выполнения заданий – пакет математического моделирования Matlab, с которым Вы знакомы. При этом при желании можно воспользоваться любым другим инструментом программирования (C++, Python и др.). Дальнейшие примеры программирования относятся к Matlab.

Чтобы воспользоваться программным обеспечением Matlab в условиях дистанционного образовательного процесса, можно использовать один из трех взаимозаменяемых вариантов.

1. Получить студенческую лицензию на сайте разработчика.

https://ch.mathworks.com/campaigns/products/trials.html

2. Воспользоваться сервисами удаленного доступа СПбПУ, позволяющими получить доступ к университетскому серверу Matlab.

https://www.spbstu.ru/upload/it/VDI_for_users.pdf

3. Выполнять задания в свободно распространяемом пакете математического моделирования Octave, язык программирования которого в рамках предлагаемых заданий полностью идентичен языку программирования Matlab.

https://www.gnu.org/software/octave/

https://octave.sourceforge.io/

Для Octave есть online-версия для браузера:

https://octave-online.net/

Лабораторная работа № 6

Лабораторная работа посвящена сравнительному исследованию методов оценки разброса обрабатываемой совокупности относительно центрального значения с точки зрения робастности получаемых оценок (устойчивости к наличию выбросов в данных).

Необходимо сравнить следующие статистические оценки центра совокупности:

– среднеквадратическое отклонение,

– среднее значение абсолютного отклонения,

– статистика критерия Граббса для определения выбросов,

– медиана абсолютного отклонения от медианы (MAD).

Работа включает: программирование аналитических оценок (см. лекции) наследственной погрешности для каждой оценки центра обрабатываемой совокупности, вызванной неточностью исходных данных (т.е. элементов совокупности); проведение статистического эксперимента методом Монте-Карло с целью получения стохастических оценок наследственной погрешности; сравнение аналитических оценок с численными оценками, полученными методом Монте-Карло; интерпретация результатов и выводы.

Постановка задачи

Пусть имеется совокупность значений $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, полученных в результате выполнения измерений.

Пусть Δx_i – абсолютная погрешность величины x_i, т.е $\Delta x_i = x_i - x_{\text{ист}\,i}$, где $x_{\text{ист}\,i}$ – истинное значение измеряемой величины x_i. Пусть Δ_i – предельное (наибольшее возможное) значение модуля абсолютной погрешности величины x_i, то есть $|\Delta x_{\mathrm{i}}| < \Delta_i$.

Необходимо оценить наследственную погрешность результата оценки центра совокупности $\mathbf{x}$ указанными выше методами.

Пусть алгоритм обработки неточных данных задан функцией вида

$$y_{\text{ист}} = f(x_{\text{ист}\,1}, x_{\text{ист}\,2}, \ldots, x_{\text{ист}\,n}),$$

где f – некоторая функция от элементов совокупности (в рамках лабораторной работы – вычисление центра совокупности действительных значений $\mathbf{x}_{\text{ист}}$), $y_{\text{ист}}$ – значения данной функции f.

Поскольку элементы совокупности $\mathbf{x}$ известны с погрешностью, то точно определить значение центра совокупности нельзя. Получаемая оценка неизбежно содержит наследственную погрешность Δy, обусловленную наличием погрешностей $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$. Кроме того, значение Δy будет зависеть от вида функции f, то есть от выбранного метода вычисления среднего.

Исходные данные

Принять в качестве исследуемой совокупности вектор значений из 99 случайных значений в диапазоне от (–1) до (+1), а последнее (сотое) значение – равным 1.687.

Принять следующие пределы относительных погрешностей: 1% – для всех значений, кроме последнего, (1, 2, …, 20)% – для последнего значения в указанной выше совокупности.

Таким образом, моделируется ситуация, когда одно значение в совокупности получено с существенно большей погрешностью, нежели остальные члены ряда, т.е. представляет собой статистический выброс.

Число итераций статистического эксперимента методов Монте-Карло задать порядка $N = 10^5 \div 10^6$.

Замечание: на этапе отладки программы полезно задать меньшее значение N, чтобы ускорить вычисления.

Задание к лабораторной работе № 6

Оценить значение центра для заданной совокупности $\mathbf{x}$: с помощью среднеквадратического отклонения, среднего значения абсолютного отклонения, статистики критерия Граббса для определения выбросов, медианы абсолютного отклонения от медианы (MAD).

Для каждой из оценок оценить аналитически и методом Монте-Карло наибольшее возможное значение Δ_y модуля отклонения Δy оценки значения центра совокупности **x** для заданных пределов погрешности исходных данных, приняв для усеченного среднего коэффициент $k = 0{,}2$.

Построить графики (ниже Δ_n – это предел абсолютной погрешности последнего значения в обрабатываемой совокупности, $n = 100$):

– зависимости Δ_y от Δ_n, где Δ_y оценивается по аналитическим формулам,

– зависимости Δ_y от Δ_n, где Δ_y оценивается методом Монте-Карло.

Повторить весь эксперимент и построить графики, введя в обрабатываемую совокупность дополнительно 1, 2, 3 выброса (их значения – на Ваше усмотрение). Повторить весь эксперимент и построить графики, изменив значение коэффициента k для оценки усеченным средним на значение, равное 0,1.

$$\left[\begin{array}{l} y_{\mathrm{i}} = |x_{\mathrm{i}} - x_{\mathrm{med}}|, \\ \mathrm{MAD} = y_{\mathrm{med}}. \end{array}\right.$$

В разделе «Выводы» ответить на следующие вопросы (опираясь, в том числе, на результаты выполненного эксперимента).

а) Какая из оценок центра совокупности наименее робастная? Какая – наиболее робастная?

б) Сравните преимущества и недостатки оценки разброса в совокупности при использовании среднеквадратического отклонения и MAD.

Теоретические сведения, необходимые для выполнения работы

При выполнении лабораторной работы используйте следующие формулы для вычисления среднего по совокупности:

– среднеквадратическое отклонение

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2},$$

– среднее значение абсолютного отклонения

$$m = \frac{1}{n-1} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|,$$

– статистика критерия Граббса для определения выбросов

$$G = \max_i \; |x_i - \bar{x}|/s,$$

– медиана абсолютного отклонения от медианы (MAD)

$$\left[\begin{array}{l} y_i = |x_i - x_{\mathrm{med}}|, \\ \mathrm{MAD} = y_{\mathrm{med}} \end{array}\right.$$

или

$$\mathrm{MAD} = \mathrm{median}\left\{ |x_i - x_{\mathrm{med}}| \right\},$$

где $x_{(\mathrm{i})}$ – i-ый элемент вариационного ряда (упорядоченной совокупности), n – число элементов исследуемой совокупности, median $\{\mathbf{x}\}$ – оператор вычисления выборочной медианы по совокупности **x**.

Используйте следующие формулы для вычисления интервалов возможных значений погрешности оценки центра совокупнисти, унаследованной от исходных данных:

– для среднеквадратического отклонения

:

$$\Delta_s = \frac{1}{(n-1)\cdot s}\cdot\sum_{i=1}^{n}|x_i-\bar{x}|\cdot\Delta_i,$$

$$s\pm\Delta_s = \left[s-\frac{1}{(n-1)\cdot s}\cdot\sum_{i=1}^{n}|x_i-\bar{x}|\cdot\Delta_i,\ s+\frac{1}{(n-1)\cdot s}\cdot\sum_{i=1}^{n}|x_i-\bar{x}|\cdot\Delta_i\right],$$

– для среднего значения абсолютного отклонения:

$$\Delta_m = \frac{1}{n-1}\cdot\sum_{i=1}^{n}\left|\operatorname{sign}(x_i-\bar{x})-\overline{\operatorname{sign}}\right|\cdot\Delta_i,$$

$$m\pm\Delta_m = \left[m-\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left|\operatorname{sign}(x_i-\bar{x})-\overline{\operatorname{sign}}\right|\cdot\Delta_i,\ m+\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left|\operatorname{sign}(x_i-\bar{x})-\overline{\operatorname{sign}}\right|\cdot\Delta_i\right],$$

– для статистики критерия Граббса для определения выбросов:

$$\Delta_G = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(1+\frac{n}{n-1}\cdot\frac{(x_j-\bar{x})^2}{s^2}\right)\cdot\frac{\Delta_i}{s}+\left(1-\frac{2}{n}-2\cdot\frac{(x_k-\bar{x})^2}{s^2\cdot(n-1)}\right)\cdot\frac{\Delta_k}{s},$$

$$G\pm\Delta_G = [G-\Delta_G,\ G+\Delta_G],$$

где k – индекс значения, наиболее далеко отстоящего от среднего арифметического (т.е. $k=\arg\max_i |x_i-\bar{x}|/s$),

– для медианы абсолютного отклонения от медианы (MAD):

$$l_1=(x_1-\Delta_1), l_2=(x_2-\Delta_2), \ldots, l_n=(x_n-\Delta_n),$$

$$r_1=(x_1+\Delta_1), r_2=(x_2+\Delta_2), \ldots, r_n=(x_n+\Delta_n),$$

$$x_{\text{med}}\in\left[L_{x_{\text{med}}}, R_{x_{\text{med}}}\right]$$

$$ly_i = \min\left\{\left|l_i-R_{x_{\text{med}}}\right|,\ \left|r_i-L_{x_{\text{med}}}\right|\right\}$$

$$ry_i = \max\left\{\left|l_i-R_{x_{\text{med}}}\right|,\ \left|r_i-L_{x_{\text{med}}}\right|\right\}$$

$$ly_{\text{med}} = \text{median}\{ly_{\text{i}}\},$$

$$ry_{\text{med}} = \text{median}\{ry_{\text{i}}\}.$$

$$\text{MAD}\in[ly_{\text{med}}, ry_{\text{med}}].$$

Вычислив границы интервалов, нетрудно найти предельное значение Δ_y модуля $|\Delta y|$ отклонения результата y оценки среднего по совокупности **x** для каждого из методов.

Среда реализации: любая на Ваш выбор.

По итогам реализации составьте краткий отчет, содержащий: титульный лист, текст задания, текст составленной Вами программы, полученные результаты, выводы.

В отчет прошу включить краткое описание задания, полученный результат, выводы.