Практики 5 и 6. Сравнение МО двух нормальных случайных величин
Достоверность показаний зашумленных измерений

Практика является продолжением практики о зашумленных измерениях.

Код генератора generate_sensor_data **должен в коде фигурировать ровно один раз в этой написанной функции. Глобальные переменные в данной функции запрещены. В дальнейшем генерация измерений для обоих датчиков обязательно вызывает эту функцию, новые генераторы писать не надо. Код с нарушением этого требования не принимается.**

Части 2, 5 считаются отдельной практикой (№6). Их можно выполнять отдельно и позже, после изложения *t*-статистики на лекциях.

Часть 1. Функции проверки гипотез в случае известных дисперсий

1. Реализовать функцию расчета статистики *e* и *z* по двум выборкам P1, P2:

   function e = e_statistic(P1, P2)

   function z = z_statistic(P1, P2, D1, D2)

   Результатом обеих функций являются *__скалярные__* величины e, z, учитывающие усреднение по выборке. Формулы расчета обеих статистик должны в коде фигурировать ровно один раз в этих функциях. Глобальные переменные в данных функциях запрещены. В дальнейшем при расчете статистик e, z обязательно вызывать эти функции. Код с нарушением этого требования не принимается.
2. Осознать, почему на входе функции расчета *e*-статистики не нужны дисперсии случайных составляющих погрешности, а *z*-статистики нужны. Осознать, почему не передается объем выборки *n*.
3. Реализовать функцию проверки гипотезы о равенстве МО двух СВ на основе статистик *e* и *z*.

   function [h, border_low, border_high] = test_equal_means_e(e, D1, D2, n, alpha)

   function [h, border_low, border_high] = test_equal_means_z (z, alpha)

   h – результат проверки гипотезы

   border_low, border_high – границы области принятия основной гипотезы по e или z статистикам.

   Далее везде в коде для проверки гипотез использовать только эти функции. Код с нарушением этого требования не принимается.
4. Осознать, почему в функции расчета e-статистики не нужны дисперсии случайных составляющих погрешности, а для z-статистики нужны.

Часть 2. Функции проверки гипотез в случае неизвестных дисперсий на основе t-статистики

1. В дополнение к функциям e_statistic, z_statistic реализовать функцию расчета *t*-статистики *e* и *z* по двум выборкам P1, P2:

   function t = t_statistic(P1, P2)

   Результатом функции является скалярная величина t, учитывающая усреднение по выборке. Формулы расчета обеих статистик должны в коде фигурировать ровно один раз в этих функциях. Глобальные переменные в данных функциях запрещены. В дальнейшем при расчете статистик e, z обязательно вызывать эти функции. Код с нарушением этого требования не принимается.
2. Осознать, почему на входе функции расчета *e*-статистики не нужны дисперсии случайных составляющих погрешности, а *z*-статистики нужны. Осознать, почему не передается объем выборки *n*.
3. В дополнение к функциям test_equal_means_e, test_equal_means_z реализовать функцию проверки гипотезы о равенстве МО двух СВ на основе статистики *t*.

   function [h, border_low, border_high] = test_equal_means_t (t, n, alpha)

   n – объем выборки, по которому посчитана статистика t,

h – результат проверки гипотезы,
border_low, border_high – границы области принятия основной гипотезы по t-статистике.
Далее везде в коде для проверки гипотез использовать только эту функцию. Код с нарушением этого требования не принимается.

4. Осознать, чем отличается t-статистика от z-статистики.

Часть 3. Распределения статистик в условиях H0 и H1

1. Задаться объемом измерительной выборки 5 < *n* < 20.
2. Генерируя e-статистику по **выборкам из generate_sensor_data**, **на одном графике** построить:
    a. теоретическую плотность распределения статистики Критерия в условиях основной гипотезы (показания датчиков «в среднем» равны, отличия обусловлены случайной составляющей погрешности и носят чисто случайный характер);
    b. теоретическая плотность распределения статистики Критерия в условиях альтернативной гипотезы;
    c. в тех же условиях методом статистического моделирования Монте-Карло (N ≥ 1000) построить полигон относительных частот статистики Критерия (hist_density);
    d. методом статистического моделирования (N ≥ 1000) построить полигон относительных частот статистики Критерия при условии альтернативной гипотезы;
    e. осознать, чем отличается *n* от *N*.
3. Построить то же самое для статистики z.
4. Построить то же самое для статистики t.
5. На построенные в этой части задания графики (их должно быть три) нанести нижнее и верхнее пороговые значения для принятия или отклонения основной гипотезы. Пороги брать из функций test_equal_means_e, test_equal_means_z, test_equal_means_t.
6. Подобрать *n*, на котором визуально видна разница между t-статистикой и z-статистикой. В чем она заключается?

Часть 4. Исследование статистических свойств критериев e, z

1. Задаться объемом измерительной выборки 5 < *n* < 20.
2. Методом Монте-Карло оценить ошибку **первого** рода при принятии гипотезы о равенстве МО нормальных СВ на основе e-статистики.
   Для этого сгенерировать N ≥ 1000 реализаций e-статистики в условиях основной гипотезы и посчитать долю случаев, в которых функция test_equal_means_e принимает альтернативную гипотезу.
3. Аналогичным образом оценить ошибку **второго** рода при принятии гипотезы о равенстве МО нормальных СВ на основе e-статистики
4. Посчитать ошибки первого и второго рода для z-статистики.
5. Увеличить объем измерительной выборки n в два раза. Проверить ошибки первого и второго рода. Обосновать их изменение или, наоборот, неизменность.

Часть 5. Исследование ошибок первого и второго рода для статистик t и z

1. Методом статистического моделирования исследовать ошибки первого и второго рода в зависимости от объема выборки для z-статистики и t-статистики.
    Для этого построить графики ошибок 1-го, 2-го рода в зависимости от объема выборки n для z-статистики и t-статистики. Все 4 графика построить на одном, объяснить, чем вызвано их отличие.
2. Методом статистического моделирования исследовать ошибку первого и второго рода в зависимости от близости систематических погрешностей датчиков.

Для этого построить графики ошибок 1-го, 2-го рода в зависимости от разницы систематических погрешностей для z-статистики и t-статистики. Все 4 графика построить на одном, объяснить, чем вызвано их отличие.

3. Задаться требуемыми ошибками первого и второго рода (не обязательно одинаковыми), определить по построенным графикам
   a. объемы выборки, при котором эти ошибки будут обеспечены.
   b. разницу систематических погрешностей, которую можно выявить с такими ошибками

Вопросы к защите

1. Вероятностная модель измерения двумя датчиками одного физического параметра.
   a. Записать вероятностную модель генератора сигнала датчика «до измерений» с учетом истинного значения параметра, аналитически вывести его МО, дисперсию. Вывести МО, дисперсию для разницы показаний датчиков.
   b. Эскиз графика матожидания показаний датчика от времени.
   c. Определение случайного процесса. Почему сигналы датчиков являются случайным процессами? Определение стационарного случайного процесса. Являются ли они стационарными? Те же вопросы для разницы показаний датчиков.
2. Распределения статистик критерия
   a. e-статистика
      i. Формула расчета усредненной статистики $\bar{e}_n$ по скользящему окну по двум временным рядам показаний датчиков.
      ii. Вывести МО, дисперсию $\bar{e}_n$ в общем случае.
      iii. Вывести распределения статистики $\bar{e}_n$ в условиях основной и альтернативной гипотезы. Как связана дисперсия средневыборочного с дисперсией исходной СВ?
      iv. Привести эскизы графиков плотности распределения $\bar{e}_n$-статистики в условиях основной и альтернативной гипотезы.
      v. В каком случае график плотности распределения в условиях альтернативной гипотезы слева и в каком случае справа?
      vi. Как меняются графики плотностей распределения при изменении объема выборки? Как выглядят эти график для e-статистики без усреднения.
   b. z-статистика
      i. Формула расчета статистики z_n.
      ii. Вывести МО, дисперсию z_n в общем случае.
      iii. Вывести распределения статистики z_n в условиях основной и альтернативной гипотезы.
      iv. Привести эскизы графиков плотности распределения z_n-статистики в условиях основной и альтернативной гипотезы.
      v. В каком случае график плотности распределения в условиях альтернативной гипотезы слева и в каком случае справа?
      vi. Как меняются графики плотностей распределения при изменении объема выборки?
   c. Почему в функции расчета e-статистики не нужны дисперсии случайных составляющих погрешности, а для z-статистики нужны?
3. Проверка гипотез
   a. Определения
      i. Что такое статистическая гипотеза?
      ii. Формулировка основной и альтернативной гипотезы данной работы.
      iii. Что такое статистика критерия? Какие статистики использовались в данной работе?
      iv. Что такое критическая область и область принятия гипотезы?

   v. Что такое простая и сложная гипотеза? Какая из двух рассматриваемых гипотез является простой, а какая сложной?

b. Ошибки 1-го и 2-го рода
   i. Что такое ошибка первого и второго рода? Как посчитать ошибку первого и второго рода для e и z-статистик, зная плотности распределения статистик в условиях основной и альтернативных гипотез? Что такое мощность критерия?
   ii. Как влияет объем выборки на ошибку первого рода?
   iii. Как влияет объем выборки на ошибку второго рода? Ответ обосновать по эскизам распределений e и z-статистик, нарисованных ранее.
   iv. У какой из статистик e, z ошибка второго рода меньше?
   v. Как посчитать ошибку второго рода для e-статистики без усреднения? В какую сторону она будет отличаться от ошибки усредненной e-статистики?
   vi. Зачем применяется усреднение при расчете e-статистики? В чем выигрыш от применения усреднения.

c. Расчет порогов при заданном уровне значимости
   i. Как связаны плотность распределения вероятности и функция распределения вероятности.
   ii. Что такое инверсная функция распределения вероятностей? Как с ее помощью посчитать пороги для e?

4. Вопросы по t-статистике
   a. Формула расчета t-статистики
   b. Почему расчет t-статистики корректен, несмотря на нестационарность измерительных временных рядов P1, P2?
   c. Сформулировать основную и альтернативную гипотезы, которые проверялись с помощью t-статистики.
   d. Какое распределение у t-статистики в условиях основной и альтернативной гипотезы? Какие распределения у z-статистики в этих же условиях. Нарисовать эскиз, демонстрирующий отличия между этими статистиками?
   e. Что будет, если пороги для t-статистики считать так, как будто, это z-статистика? Что в этом случае произойдет с ошибками 1-го и 2-го рода?
   f. Какие факторы влияют на ошибки первого и второго рода? В какую сторону изменяются ошибки 1-го, 2-го рода при увеличении/уменьшении этих факторов? Чтобы подготовить ответ, вспомните эксперименты с e, z, t-статистиками из двух ДЗ. Должно быть 3 фактора.
   g. Как соотносятся ошибки второго рода z- и t-статистик при одном и том же объеме выборки?
   h. Какой показатель (показатели) самый важный для работы системы контроля достоверности датчиков?
   i. Как выбрать объем выборки, учитывая, что с технической точки зрения лучше ее взять меньше, чтобы уменьшить время сбора показаний датчиков для анализа.
   j. Можно ли воспользоваться t-статистикой, если известны дисперсии случайных составляющих погрешности?
   k. В каких случаях невозможно использовать z-статистику, но можно t-статистику?
   l. В каких случаях нельзя использовать z-статистику и t-статистку?
   m. Что лучше: z-статистика или t-статистика?