Задания практикума на ЭВМ, вероятностноя группа, 6й семестр.
Мехмат МГУ
весна 2011/2012 уч. года
доц. Мусин М.М.

Семинар 2. Точечное оценивание параметров.

1. Открыть файл “C2 упр1.csv” с данным в формате csv, разделенными запятыми, выполнить шкалирование параметра flag.

2. Построить диаграмму рассеивания выборки X1 по шкале.

3. Записать в файл ”C2 упр1 результат [Фамилии].csv” новую выборку со шкалой.

4. Оценить моменты распределения, сделать выводы о нормальности выборок X1, X5 по данным точечных оценок, найти среди них логнормальную выборки, пользуясь свойствами ассиметрии и эксцесса нормального распределения.

5. Открыть файл ”C2 упр2.csv”. Получить из него выборки X1, …, X6.

Методом моментов определить параметры следующих распределений задачи 6 – 10

6. Выборка X1 Гамма

7. Выборка X2 Парето

8. Выборка X3 Вейбулла

9. Выборка X4 Отрицательная биномиальная

10. Выборка X5 Пуассона

11.* Методом численной максимизации правдоподобия найти параметр модели в выборке X6. Известно, что выборка подчиняется следующей модели: случайно выбирается целое число от 1 до 4 и равномерная случайная величина от 0 до theta.

12.* Запустить процедуру Expectaton Maximization для выборки X6

13. Сохранить листинг R и Gnumeric spreadsheet в файл “C2 [Фамилии].(R| gnumeric)”

Семинар 3. Точечное оценивание параметров.

1. Открыть файл “C3 упр1.csv”, получить из него выборки различного размера

X1000-X10 из распределения N(17,1). Для каждой выборки расчитать теоретическое среднеквадратическое отклонения среднего значения от истиного среднего и практическое отклонение. Полученные результаты в R записать в виде списка.

2. Получить значение медианы и усеченного среднего для k=2.

3. Сохранить полученные результаты в ”C3 упр1 реузльтат [Фамилии].csv” в виде таблицы: «Размер выборки, среднее, теоретич откл, практ откл, медиана, практ откл, усеченное среднее, практ откл».

4. * Расчитать доверительные интервалы (асимптотические если не получается непосредственно) уровня доверия 0.95 для среднего значения, медианы и усеченного среднего.

5. Нарисовать график зависимости ошибок всех оценок от размера выборки.

6. Открыть файл “C3 упр2 csv”, получить из него выборки X1_001, X1_005, X1_010, X2_001, X2_005, X2_010 из распределений коши и нормального соответственно с уровнем выбросов 0.01, 0.05, 0.1.

В заданиях 7-10 нужно реализовать указанную в задании оценку, применить её ко всем выборкам из упражнения 2 и получить таким образом оценку для характеристики сдвига данного семейства.

7. L оценка с функцией lambda(t)=sin(pi t).

8. M оценка с функцией потерь rho(x)=(1-exp(x-theta))^2.

9. R оценка с d1=...=dn=1 (медиану Средних Уолша).

10. * R оценку с ядром K(t)=cos(pi t)

Сохранить листинг R и Gnumeric spreadsheet в файл “C3 [Фамилии].(R|gnumeric)”

Семинар 4. Проверка статистических гипотез

Во всех задачах на проверку гипотез требуется писать подробный вывод в текстовой форме.

1. Имеются 10 мышей, на которые произвдено воздействие A, результаты смертности среди мышей приведены в файле “C4 упр1.csv” (1 мышь умерла, 0 выжила). Проверить гипотезу о безопасности воздействия A против альтернативы об опасности.
2. По закону игровые автоматы должны быть настроены таким образом, чтобы вероятность выигрыша игрока не была ниже 2/3. Игрок сыграл с игровым автоматом 10 раз и проиграл 7 раз (См. выборку ”C4 упр2.csv”) С каким

уровнем значимости игрок может утверждать, что произошло нарушение закона?

3. В файле “C4 упр3.csv” отражено последовательное изменение стоимости продукта при преодолении цепи посредников. Считая, что наценка каждого посредника является логнормальной случайной величиной с дисперсией показателя 0,01, проверить гипотезу о том, что логорифмическое среднее данной случайной величины не превышает ln(1.02).
4. В файле “C4 упр4.csv” приведены даты выплат по страховым полисам компании Б. Считая, что процесс выплат является пуассоновским с постоянной интенсивностью, проверить гипотезу о том, что данная интенсивность равна 3 выплатам в 30 дней.
5. В файле ”C4 упр5.csv” даны три выборки из гауссовского распределения, требуется проверить гипотезу о том, что среднее первой выборки равно 17.
6. Проверить гипотезу о том, что среднее третьей выборки превышает среднее второй более чем на 2 пункта.
7. В файле “C4 упр6.csv” даны две гауссовские выборки с неизвестной дисперсией, требуется проверить гипотезу об однородности.
8. В файле “C4 упр7.csv” даны 10 пар выборок с одинаковыми дисперсиями. Требуется проверить, что данные выборки попарно однородны. Записать в файл “C4 упр6 результат [фамилии].csv” в стоблик номера пары выборок, совпадение или различие (1/0), p-значение.
9. * В файле “C4 упр8.csv” даны 20 пар гауссовских выборок. Требуется проверить, что данные выборки попарно однородны. Записать в файл “C4 упр7 результат [фамилии].csv” в стоблик номера пары выборок, совпадение или различие (1/0), p-значение, применявшийся вид t-теста.
10. Сохранить листинг R и Gnumeric spreadsheet в файл “C4 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Семинар 5. Проверка статистических гипотез

Во всех задачах на проверку гипотез требуется писать подробный вывод в текстовой форме.

1. Провести непраметрический тест разности средних для парных выборок X1_1, X1_2.
2. Провести непраметрический тест разности средних для независимых выборок X2_1, X2_2.
3. Проверить выборку X3 на рормальность. Построить гистограмму, ядерную оценку и провести тест Колмогорова-Смирнова.
4. Сравнить форму выборки X4 с экспоненциальным.
5. Сравнить форму X5 с логномральным и гамма.
6. Классифицировать распределения выборок X6_1 – X6_5 по форме (искать среди Экспоненциального, Гамма, Вейбулла)
7. Предложить подходящее распределение для выборки X7.
8. Проверить гипотезу о равенстве распределений пар выборок X8_1,X8_2 и X8_3,X8_4.
9. Для пар выборок X9_1,X9_2 и X9_3,X9_4 проверить гипотезу о нормальности

и провести дисперсионный анализ при помощи t-тестов или критерия Мана-Уитни.

10. * Собрать данные о процентах различных кандидатов на президентских выборках (использовать данные центризберкома) по нескольким регионам или отдельным ТИК. Проверить полученные данные на форму распределения. Классифицировать регионы/ТИК по p-значению. Проверить соответствие полученых результатов здравому смыслу и статистике МВД о криминогенности.
11. Сохранить листинг R и Gnumeric spreadsheet в файл “C5 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Семинар 6. Корреляции

1. В файле «C6 упр1.csv» находятся выборки X1_1,Y1_1, X1_2,Y1_2, X1_3,Y1_3. Про выборки известно, что они гауссовские. Построить диаграммы рассеивания, найти корреляции и вычислить p-значения для данных трех пар выборок.
2. Для пары выборок из файла «C6 упр2.csv» нарисовать диаграмму рассеивания, рассчитать корреляцию, предложить преобразование для нарушения симметрии и рассчитать корреляцию по преобразованным данным. Написать вывод об исследовании зависимости при помощи данного теста.
3. Рассчитать для выборки из предыдущей задачи ранговые коэффициенты корреляции и получить данные о p-значении для них.
4. Для трех пар выборок X4_1,Y4_1, X4_2,Y4_2, X4_3,Y4_3 вычислить коэффициенты корреляции, определить уровень значимости и сделать вывод о зависимости наблюдений.
5. В файле «C6 упр4.csv» даны четыре пары выборок, проверить выборки на соответствие гауссовскому закону и на основании этого при помощи коэффициентов корреляции Пирсона или Спирмена сделать вывод о зависимости наблюдений в выборках.
6. *В файле «C6 упр5.csv» находятся данные о ценах трех акций входящих в портфель. Рассчитать матрицу корреляций данного портфеля, обнаружить зависимости, проверить уровень занчимости этих зависимостей.
7. *Для данных из предыдущей задачи рассчитать матрицу ковариаций акций, входящих в портфель. Предполагая, что портфель равномерно распределен между этими тремя акциями, оценить доходность (среднее значение) и риск (среднеквадратическое отклонение) данного портфеля.
8. *Рассчитать риск портфеля в случае, если он распределн между тремя акциями с весами ½, ¼ и ¼.
9. *В файле «C6 упр6.csv» находится выборка из четырехмерного случайного вектора. Оценить базис в пространстве такой, в котором данный вектор имеет независимые компоненты.
10. Сохранить листинг R и Gnumeric spreadsheet в файл “C6 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Семинар 7. Корреляции

1. В файле «C7 упр1.csv» имеется выборки X1_1,Y1_1. Построить по данной выборке таблицу сопряженности. Проверить зависимость критерием хи-квадрат.
2. Выборки X2_1,...,Y2_3 преобразовать в таблицы сопряженности и проверить на зависимость.
3. В файлах «C7 упр2 T1.txt» ...«C7 упр2 T3.txt» приведены таблицы сопряженности. Прочитать их и проверить на зависимость.
4. Програмные продукты планеты Плюк оцениваются по шкале от 1 до 4 по качеству, кроме того имеется два способа написания програмных продуктов быстрый и медленный. Известно, что среди быстро написанных программных продуктов оценку 1 имеют 120 продуктов оценку 2 – 124 продукта, 3 — 133 продукта, 4 – 106 продуктов. Среди медленно написанных продуктов оценку 1 имеют 97, 2 – 142, 3 – 129 и 4 – 149 продуктов. Выяснить, имеется ли на планете Плюк статистически значимая связь между скоростью написания программных продуктов и их качеством.
5. В файле «C7 упр3.csv» приведены три выборки. Вычислить корреляции первой и второй, после этого вычислить частные корреляции первой и второй при условии третьей. Для расчета функции частной корреляции использовать явную формулу.
6. По выборкам X6_1,...,X6_6 из файла «C7 упр4.scv», построить корреляционную матрицу. Составить список с частными корреляциями i-й по j-й при условии k-й. Выделить тройки, где данные частные корреляции существенны.
7. В файле «C7 упр5.csv» выборки X7_1 и Y7_1 из гамма распределения с параметрами 2 и 3. Сделать замену переменных в маргинальных распределениях и вычислить p-значение в тесте независимости при помощи коэффициента Пирсона. Применить к исходным выборками коэффицент Пирсона и Спирмена.
8. Выборки X8_1 и Y8_1 из сингулярного распределения. Преобразовать их к таблице истинности по целочисленным интервалам, применить критерий хи-квадрат.
9. *Написать функцию, получающую на вход пару выборок, опционально уровень надежности и тип распределения (дискретное, непрерывное, сингулярное). Функция должна определять подходящий для данной ситуации статистический критерий, применять его и в выводе выдавать p-значение и решение о зависимости между выборками.
10. *Обработать с помощью функции из предыдущей задачи набор пар выборок из файла “C7 упр7.csv”
11. Сохранить листинг R и Gnumeric spreadsheet в файл “C7 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Семинар 8. Регрессии

1. В файле «C8 упр1.csv» содержатся выборки X1 и Y1. Построить регрессию Y

по X. Сделать вывод об адекватности построенной регресии по оценке дисперси остатков и коэффициенту детерминации.
2. Нарисовать график наблюденных и предсказанных значений.
3. Проанализировать остатки на нормальность.
4. Построить коэффициенты корреляции Y и X.
5. Построить многомерную регрессию Y5 по X5_1,...X5_4. Проверить адекватность модели регрессии.
6. Построить матрицу корреляций величин, сравнить с полученными данным регрессионного анализа.
7. Найти незначимые переменные. Незначимость их обосновать.
8. Построить новую регрессию по значимым переменным. Проверить адекватность.
9. Подобрать полином не более чем 5й степени Y9 по X9 по данным из файла «C8 упр3.csv».
10. В файле «C8 упр4.csv» содержится 5 пар выборок, попарная зависимость между выборками нелинейная. Подобрать наиболее подходящую для каждой пары выборок.
11. В файле «C8 упр5.csv» содержатся выборки Y11,X11_1,...,X11_3. Связь Y с регрессорами нелинейная. Найти форму зависимости.
12. * В файле «C8 упр6.csv» содержатся выборка из нелинейной зависимости Y по набору регрессоров. Найти форму зависимости, отбросить незначимые регрессоры.
13. Сохранить листинг R и Gnumeric spreadsheet в файл “C8 [Фамилии].(R| gnumeric)” и прислать на почту maxim@musin.cc

Все задания выполняются в R и Gnumeric.

Семинар 9. Регрессии

1. В файле «C9 упр1.csv» содержатся выборки негауссовской линейной регрессии Y1 по X1. Оценить регрессионные коэффициенты. Вывести график предсказанных и наблюдаемых значений Y1. Оценить дисперсию остатков.
2. Зависимость Y2 от X2 носит нелинейный характер. Пользуясь ядерными оценками Надарая-Ватсона построить аппроксимацию функции зависимости. Вывести график предсказанных и наблюдаемых значений Y2. Сделать предположение о виде функции E(Y2|X2=x)=f(x).
3. В файле «C9 упр3.csv» лежат данные о принятии или непринятии решения Y3 по страховой выплате в зависимости от уровня ущерба X3. Оценить зависимость принятия решений от уровня ущерба при помощи логистической регрессии.
4. Там же имеются данные о зависимости наступлении страхового случая Y4 от трех различных параметров X4_1,X4_2,X4_3. Оценить вклад параметров в принятие решений, пользуясь логистической регрессией. В комментариях к решению выписать уравнение оцененной разделяющей плоскости.
5. Произвести тот же анализ при помощи функции probit.
6. В файле «C9 упр4.csv» имеются данные зависимости Y6 от количественных

факторов X6_1,...,X6_3, а также от качественных факторов D6_1,D6_2. Найти зависимость.

7. В файле «C9 упр5.csv» имеются данные о зависимости Y7 от количественных факторов X7_1, X7_2, и параметра Z7, принимающего значения A,B и C. Проанализировать зависимость пользуясь многомерной регрессией с dummy переменными.
8. Установить зависимость Y8 от X8 если известно, что на положительной и отрицательных полуосях зависимости разные, но имеют вид полинома степени не более 2.
9. * Выполнить предыдущее задание для выборок из файла «C9 упр7.csv» при неизвестной границе между зависимостями.
10. * В файле «C9 упр8.csv» найти зависимость между значением Y10, переменными X10_1,X10_2 и факторной переменной X10_3 при помощи многомерной логистической регрессии с dummy переменными.
11. Сохранить листинг R и Gnumeric spreadsheet в файл “C9 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Задания 1 и 2 выполняются только в R, все остальные задания выполняются в R и Gnumeric.

Семинар 10. Факторный анализ

1. Провести анализ методом главных компонент данных из «C10 упр1.csv»
2. Отобрать существенные компоненты пользуясь процентом объясненной дисперсии.
3. Провести анализ методом главных компонент нецентрированной выборки из «C10 упр2.csv».
4. Интерпретировать данные метода главных компонент.
5. Провести анализ методом главных компонент, избежав нелинейных закономерностей данные из «C10 упр3.csv».
6. Провести факторный анализ данных из «C10 упр4.csv», считая что искомое количество факторов равно 4.
7. Провести факторный анализ, подобрав оптимальное количество факторов.
8. *Провести анализ главных компонент и факторный анализ многомерной выборки из «C10 упр5.csv».
9. Сохранить листинг R и Gnumeric spreadsheet в файл “C10 [Фамилии].(R|gnumeric)” и прислать на почту maxim@musin.cc

Задания 1 - 5 выполняются только в R, все остальные задания выполняются в R и Gnumeric.

Семинар 11. Временные ряды

1. Открыть данные из файла «C11 упр1.csv», создать по столбцу x1 временной ряд. Вывести график ряда, автокорреляций и частных автокорреляций.
2. Подогнать под созданный временной ряд модель MA(5). Произвести анализ остатков на гауссовость.
3. Для ряда x3. Посмотреть автокорреляции и частные корреляции. Подогнать под него модель AR(3). Сделать анали остатков.
4. Для данных в x4 подогнать AR модель.
5. Вычесть из ряда x5 линейный тренд. Подогнать под него модель AR(1).
6. Подобрать для стационарного ряда в x6 правильную модель ARMA(p,q), базируясь на автокорреляциях и частных автокорреляциях. Провести анализ остатков.
7. Проверить полученный результат на стационарность ряда и независимость остатков.
8. Построить прогноз временного ряда. Отобразить среднее значение прогноза, верхний и нижний доверительный интервал на 10 единиц времени вперед.
9. Подобрать для стационарного ряда x9 правильную модель ARIMA(p,d,q).
10. Подобрать для нестационарного временного ряда x10 правильную модель ARIMA(p,d,q).
11. Построить прогноз на 10 единиц времени вперед для последнего временного ряда.
12. *Для данных в «C11 упр2.csv» подогнать ARMA модель.
13. Сохранить листинг R и Gnumeric spreadsheet в файл “C11 [Фамилии].R” и прислать на почту maxim@musin.cc

Все задания выполняются только в R.