МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное АВТОНОМНОЕ образовательное учреждение высшего профессионального образования
«Национальный исследовательский ядерный университет «МИФИ»

**Димитровградский инженерно-технологический институт –**филиал федерального государственного автономного образовательного учреждения высшего
профессионального образования «Национальный исследовательский ядерный университет «МИФИ»
**(ДИТИ НИЯУ МИФИ)**

**Лабораторная работа №1
по курсу «Основы математической статистики и
планирование эксперимента»**

Составил: доцент кафедры
 высшей математики
 канд. экон. наук
 Кожухова В.Н.

Димитровград 2017

**Лабораторная работа №1**

Элементы теории вероятностей. Основные виды распределений. Генерация случайных величин.

**Цель работы**

1)Повторение основных видов дискретных и непрерывных распределений: биномиального, Пуассона, равномерного, нормального, хи-квадрат, Фишера, Стьюдента.

2) Научиться разыгрывать случайные величины.

**Программные средства**

По желанию студента для выполнения лабораторной работы может использоваться Microsoft Excel, математические пакеты Mathcad, Maple, Matlab, Mathematica, статистические пакеты Statistica, R и др. В справочных сведениях по ходу выполнения работы будут приводиться примеры для Microsoft Excel, как наиболее доступного для большинства студентов пакета.

**Задание**

1. Для заданных значений параметров рассчитать теоретическую плотность распределения и интегральную функцию распределения для:

а) биномиального закона распределения;

б) распределения Пуассона;

в) нормального распределения;

г) равномерного распределения;

д) распределения хи-квадрат;

е) распределения Фишера;

ж) распределения Стьюдента.

Построить графики. Значения аргументов функций подобрать самостоятельно. Рассчитать математическое ожидание, дисперсию, среднее квадратическое отклонение, моду и медиану.

1. Сгенерировать *n* = 100 значений случайной величины с указанным законом распределения.

а) Построить вариационный ряд.

б) Разбить полученный ряд на *m* интервалов равной ширины *w* и подсчитать количество значений, попавших в каждый интервал (эмпирические абсолютные частоты).

в) Рассчитать высоту столбцов и построить гистограмму эмпирического распределения

г) Рассчитать накопленные относительные частоты попадания в интервалы и построить интегральную функцию эмпирического распределения.

1. Рассчитать среднее выборочное, дисперсию, среднее квадратическое отклонение (СКО), моду и медиану.
2. Проделать п. 2, 3 для выборки объемом 1000 значений. Как изменилась эмпирическая плотность распределения и выборочные числовые характеристики? Сравните их с теоретическими (теми, которые были заданы при генерации).

**Внимание! П. 2, 3, 4 проделать для равномерного и нормального законов распределения.**

**Справочные сведения**

1. Построение графиков плотностей и функций распределения.

Диапазон изменения аргументов функций выбирайте самостоятельно для получения «красивых» графиков функций.

а) **Биномиальный закон** распределения описывает вероятность количества «успехов» в последовательности из *n* испытаний.

Функция плотности распределения определяется 2 параметрами: вероятностью успеха в каждом испытании *p* и числом испытаний *n*.



|  |  |
| --- | --- |
|  | Используйте функцию =БИНОМ.РАСП(число\_успехов; число\_испытаний; вероятность\_успеха ; интегральная). Число успехов – аргумент функции, «Интегральная» – логический признак, может принимать значения 0 – ложь (плотность распределения) и 1 – истина (интегральная функция распределения). |

**б) Распределение Пуассона** (один параметр – ) имеет вид:

.

|  |  |
| --- | --- |
|  | Используйте функцию = ПУАССОН.РАСП(x;среднее;интегральная). Среднее – это параметр , x – это аргумент функции. «Интегральная» – аналогично предыдущему примеру. |

в) **Нормальный** закон распределения задается плотностью:

.

где принять  - параметры распределения (математическое ожидание и стандартное отклонение соответственно).

|  |  |
| --- | --- |
|  | Можно воспользоваться функцией **НОРМ.РАСП** (**x**;**среднее**;**стандартное\_откл**;**интегральная**), где x – аргумент; среднее – M; **стандартное\_откл ­–** σ; интегральная – логический признак, принять равным **0** для плотности распределения (и **1** – для интегральной функции распределения). |

г) **Равномерный** закон распределения задается плотностью:



Функция равномерного распределения:



|  |  |
| --- | --- |
|  | В Excel нет встроенной функции равномерного распределения. Построить самостоятельно. |

д) **Хи-квадрат** распределение с *k* степенями свободы – это распределение **суммы** **квадратов** *k* независимых стандартных **нормальных** случайных величин. Нормальная СВ называется стандартной, если имеет
*M* = 0, *σ* = 1.

Плотность распределения хи-квадрат:

,

где  – гамма-функция.

|  |  |
| --- | --- |
|  | Используйте функцию = ХИ2.РАСП(x;степени\_свободы;интегральная). Степени\_свободы – это параметр *k*, x – это аргумент функции. «Интегральная» – аналогично предыдущим примерам. |

**е) Распределение Фишера**

Пусть  — две независимые случайные величины, имеющие распределение хи-квадрат:, где . Тогда распределение случайной величины  называется распределением Фишера (распределением Снедекора) со степенями свободы  и .

|  |  |
| --- | --- |
|  | Используйте функцию = F.РАСП(x; степени\_свободы1; степени\_свободы2; интегральная). Степени\_свободы – это параметры  и , x – это аргумент функции. «Интегральная» – аналогично предыдущим примерам. |

**ж) Распределение Стьюдента**

Пусть  — независимые стандартные нормальные случайные величины, такие что . Тогда распределение случайной величины t, где , называется распределением Стьюдента с *k* степенями свободы. Её распределение абсолютно непрерывно и имеет плотность

,

где Г — гамма-функция Эйлера.

|  |  |
| --- | --- |
|  | Используйте функцию = СТЬЮДЕНТ.РАСП(x;степени\_свободы;интегральная). Степени\_свободы – это параметр , x – это аргумент функции. «Интегральная» – аналогично предыдущим примерам. |

**Пример выполнения п.1:**

Для каждого закона распределения нужно получить подобный график, как на рис. 1. Всего должно быть 7 графиков плотностей и 7 графиков функций распределений.

*Графики плотности и функции не обязательно строить на одной диаграмме, можно на разных.*



Рисунок – Примеры графиков распределений

Для распределений Фишера и Стьюдента шаг аргумента лучше брать меньше (см. рис. 2 как пример подбора аргумента для исходных параметров)!



Рисунок – Примеры расчета значений функций

**Квантиль** распределения – значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

Пример: Квантиль порядка 0,5 называется **медианой*:*** , где

 - значение медианы распределения.

**Медианой *Me***  случайной величины X называют то ее значение, для которого выполняются равенство вероятностей событий, то есть, плотность вероятностей справа и слева одинаковы и равны половине (0,5).







**Модой *Mo*** дискретной случайной величины X называют те ее возможное значение, которые соответствует наибольшей вероятности появления (т.е. такое значение величины X, которое случается чаще всего при проведении экспериментов, опытов, наблюдений). В случае непрерывной случайной величины модой называют то ее возможное значение, которому соответствует максимальное значение плотности вероятностей



В зависимости от вида функции f(x) случайная величина X может иметь разное количество мод. Если случайная величина имеет одну моду, то такое распределение вероятностей называют одномодальным; если распределение имеет две моды – двухмодальным и более – мультимодальным.

Существуют и такие распределения, которые не имеют моды, их называют антимодальными.

Графически мода и медиана изображены на рис. 3.



Рисунок – Мода и медиана распределения

В таблице 1 приведены значения основных характеристик распределений:

Таблица 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Характерис-тика | Бином. | Пуассона | Нормаль-ное | Равномер-ное | Хи-ква-драт | Фишера | Стьюдента |
| Математичес-кое ожидание |  |  | M |  | *k* |  если k2 > 2 | 0 если k>1 |
| Дисперсия |  |  |  |  | *2k* |  если k2 > 4 |  если k>2 |
| Среднее квадратическое отклонение |  |  |  |  |  |  если k2 > 4 |  если k>2 |
| Мода |  |  | M | любое число из отрезка  | k–2, если  |  если k1 > 2 | 0 |
| Медиана |  |  | M |  | ≈  k – 2/3 | – | 0 |

Здесь  – округление до ближайшего целого в меньшую сторону, «пол».

|  |  |
| --- | --- |
|  | Генерация случайных чисел вызывается командой: *Данные – Анализ данных – Генерация случайных чисел*. Если *Анализ данных* на вкладке *Данные* отсутствует, его необходимо установить через *Параметры Excel* – *Надстройки – Перейти – Пакет анализа*. |

Генерировать случайные числа необходимо с теми параметрами, которые использовались при построении графиков равномерного и нормального распределения (см. рис. 4).



Рисунок – Генерация 100 значений равномерно распределенной СВ с параметрами 14 и 17.

а) Построить вариационный ряд – значит упорядочить полученные значения по возрастанию.

б) Разбить полученный ряд на *m* интервалов равной ширины *h* и подсчитать количество значений, попавших в каждый интервал (эмпирические абсолютные частоты).

**Необходимо рассчитать:**

Минимальное число столбцов (интервалов) гистограммы распределения (формула Стерджесса): .

Ширина интервала .

|  |  |
| --- | --- |
|  | Минимальное и максимальное значения в выборке: МИН(число1;число2; ...) и МАКС(число1;число2; ...)Количество наблюдений, попавших в *i*-тый интервал, определяется с помощью функции ЧАСТОТА(массив\_данных;массив\_интервалов). Массив\_интервалов – правые границы каждого интервала! Формулу необходимо ввести как формулу массива. После ввода формулы в первую ячейку (например, А1) выделите весь нужный диапазон (например, A1:A6 – для 6 интервалов), нажмите клавишу F2, а затем нажмите клавиши CTRL+SHIFT+ENTER. Если формула не будет введена как формула массива, отобразится только одно ее значение в ячейке A1. **Не вводите фигурные скобки с клавиатуры! См. рис. 5 и 6.** |



Рисунок – как это должно быть в цифрах



Рисунок – как это должно быть в формулах

в) Рассчитать высоту столбцов и построить гистограмму эмпирического распределения (рис.7)

Высота *i*-того столбца: ,

где  – количество наблюдений, попавших в *i*-тый интервал (столбец частоты), *h* – ширина интервала, *n* – число наблюдений в выборке.



Рисунок – Гистограмма по выборке равномерного распределения

г) Рассчитать накопленные относительные частоты попадания в интервалы и построить интегральную функцию эмпирического распределения (э.ф.р. – рис. 8).



Рисунок – Эмпирическая функция распределения

Гистограмму и функцию распределения необходимо строить по серединам интервалов (ось X). Эмпирическая функция распределения представит собой диапазон накопленных частот .

Таким образом, должна получиться таблица 2 со столбцами:

Таблица 2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| № интервала | Начало интервала | Конец интервала | Середина интервала | Частота  | Относ. частота  | Высота  | э.ф.р. |
| 1 |  |  |  | … | … | … | … |
| 2 |  |  | … |  |  |  |  |

1. Расчет выборочного среднего, выборочной дисперсии и выборочного стандартного отклонения:

 - выборочное среднее,

 - выборочная дисперсия,

 - выборочное СКО,

где  – сгенерированный ряд наблюдений в выборке.

|  |  |
| --- | --- |
|  | выборочное среднее СРЗНАЧ(число1;число2; ...)выборочная дисперсия ДИСП.В (число1;число2; ...)выборочное СКО СТАНДОТКЛОН.В (число1;число2; ...) |

**Мода** для выборки из *дискретного* распределения ­– значение во множестве наблюдений, которое встречается наиболее часто. Иногда в совокупности встречается более чем одна мода (например: 6, 2, 6, 6, 8, 9, 9, 9, 10; мода = 6 и 9).

Немного сложнее с *интервальными* данными, когда вместо конкретных значений в выборке имеются интервалы. В этом случае говорят о **модальном интервале**, то есть интервале, частота которого максимальна относительно других интервалов. Однако и здесь можно отыскать конкретное модальное значение, хотя оно будет условным и примерным, так как нет точных исходных данных. Есть некоторое общее правило, по которому рассчитывается мода в интервальных данных. Представим, что у нас есть набор данных, как в табличке ниже.



Для наглядности изобразим соответствующую диаграмму (рис. 9).



Рисунок – Диаграмма распределения

Требуется найти модальное значение цены.

Вначале нужно определить модальный интервал, который соответствует интервалу с наибольшей частотой. В нашем примере это третий интервал с ценой от 301 до 400 руб. На графике – самый высокий столбец. Теперь нужно определить конкретное значение цены, которое соответствует максимальному количеству. Делается допущение о том, что интервалы выше и ниже модального в зависимости от своей частоты имеют разные вес и «перетягивают» моду в свою сторону. Если частота интервала, следующего за модальным, больше, чем частота интервала перед модальным, то мода будет правее середины модального интервала и наоборот (рис. 10).



Рисунок – Модальный интервал

На рисунке 10 отчетливо видно, что соотношение высоты столбцов, расположенных слева и справа от модального определяет близость моды к левому или правому краю модального интервала. Задача по расчету модального значения состоит в том, чтобы найти точку пересечения линий, соединяющих модальный столбец с соседними (как показано на рисунке пунктирными линиями) и нахождении соответствующего значения признака (в нашем примере цены). По данному рисунку нетрудно вывести формулу расчета моды в интервальном ряду.

Формула моды имеет следующий вид.



где Мо – мода,

xMo – значение начала модального интервала,

hMo – ширина модального интервала,

nМо – частота модального интервала,

nМо-1 – частота интервала, находящего перед модальным,

nМо+1 – частота интервала, находящего после модального.

Второе слагаемое формулы моды соответствует длине красной линии на рисунке выше. Рассчитаем моду для нашего примера.



Таким образом, мода интервального ряда представляет собой сумму, состоящую из значения начального уровня модального интервала и отрезка, который определяется соотношением частоты ближайших интервалов от модального.

|  |  |
| --- | --- |
|  | = МОДА.ОДН – рассчитывает моду по заданным значениям. = МОДА.НСК – позволяет рассчитать сразу несколько модальных значений (одинаковых максимальных частот) для одного ряда данных, если они есть. Функцию нужно вводить как формулу массива, перед этим выделив количество ячеек равное количеству требуемых модальных значений. Иногда действительно модальных значений может быть несколько. Однако для этих целей предварительно лучше посмотреть на диаграмму распределения.  |

Моду для интервальных данных одной функцией в Excel рассчитать **нельзя** – нужно забивать формулу вручную.

**Медиана** для выборки из дискретного распределения – срединное значение для ранжированного ряда, половина чисел имеют значения большие, чем медиана, а половина чисел — меньшие (например: 1, 2, 3, 4, 5, то медиана = 3; если число значений в ряду четно, например: 1, 2, 3, 4, 5, 6, то медиана = (3+4) / 2).

|  |  |
| --- | --- |
|  | МЕДИАНА(число1;число2;...). |

Так происходит поиск или расчет медианы в дискретных данных. Однако данные, как в нашем случае, могут быть еще и интервальными, где выбрать конкретное значение не представляется возможным, так как конкретных значений просто нет. Как и в моде, медиану в таком случае рассчитывают по некоторому общепринятому правилу.

Для начала находят медианный интервал. Это такой интервал, через который проходит искомое медианное значение. Определяется с помощью накопленной доли ранжированных интервалов. Где накопленная доля впервые перевалила через 50% всех значений, там и медианный интервал со скрытой внутри медианой.

Исходим из предположения, что распределение данных внутри медианного интервала равномерное (т.е. 30% ширины интервала – это 30% значений, 80% ширины – 80% значений и т.д.). Отсюда, зная количество значений от начала медианного интервала до 50% всех значений совокупности (разница между половиной количества всех значений и накопленной частотой предмедианного интервала), можно найти, какую долю они занимают во всем медианном интервале. Эта доля переносится на ширину медианного интервала, указывая на конкретное значение, именуемое медианой.

Для примера рассчитаем медиану по следующим данным (рис. 11).





Рисунок – Расчет медианы по интервальным данным

Формула медианы имеет следующий вид:



где xMe — начало медианного интервала;

hMe — ширина медианного интервала;

n/2 — количество всех значений, деленное на два;

S(Me-1)— суммарное количество наблюдений, которое было накоплено до начала медианного интервала, т.е. накопленная частота предмедианного интервала;

nMe — частота медианного интервала.

Требуется найти медианную цену, то есть ту цену, дешевле и дороже которой по половине количества товаров. Для начала произведем вспомогательные расчеты накопленной частоты, накопленной доли, общего количества товаров.



По последней колонке определяем медианный интервал – 300-400 руб (накопленная доля впервые более 50%). Ширина интервала – 100 руб. Теперь остается подставить данные в приведенную выше формулу и рассчитать медиану.



То есть у одной половины товаров цена ниже, чем 350 руб., у другой половины – выше.

Медиану для интервальных данных одной функцией в Excel рассчитать **нельзя** – нужно забивать формулу вручную.

На рис. 12 представлены расчеты для нашей примерной выборки.



Рисунок – Расчет числовых характеристик выборочного распределения

4. Отсутствие комментариев к п.4 лабораторной работы не свидетельствует о том, что его выполнять не нужно. Не забудьте также, что п. 2, 3, 4 нужно проделать не только для равномерного закона, но и для нормального, а также сгенерировать не только 100 значений выборки, но и 1000, и сравнить результаты расчетов.

**Список литературы**

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2003. – 479 с.
2. <http://statanaliz.info/>
3. <https://ru.wikipedia.org/>

**Отчет по лабораторной работе**

Лабораторная работа сдается в печатном виде. Отчет должен содержать:

1. Аккуратно оформленный титульный лист с названием вуза, дисциплины, № и названием лабораторной работы, ФИО студента, № группы, городом и годом.
2. Все пункты лабораторной работы, а именно:

а) В п. 1 нужно привести заполненную табл. 1 и построенные 14 графиков функций. Не нужно приводить данные для построения графиков. На графике должны быть указаны: название (какой закон распределения, какие у него параметры), подписаны оси абсцисс и ординат. Линии сетки – по желанию. Легенда нужна, если на одной диаграмме несколько графиков.

б) В п. 2-4 для каждого из 4 случаев (100 значений, равномерное; 1000 значений, равномерное; 100 значений нормальное; 1000 значений нормальное) нужно привести заполненную табл. 2, графики гистограммы и функции распределения, и рассчитанные выборочные характеристики. Графики должны быть оформлены аналогично п. 2а), т.е. с подписями названия, осей и т.д.

в) В п. 4 нужно указать, как изменились выборочные характеристики выборки в 1000 наблюдений от выборки в 100 наблюдений, похожи ли они на теоретические, рассчитанные в табл. 1, стали ли они точнее? Постараться понять, что и почему с ними произошло и пояснить.

г) Не нужно распечатывать сгенерированные выборки! Пожалейте свои принтер и бумагу.

3. Если Вы пользовались какой-то литературой помимо этих указаний, стоит в конце привести и ее. Сюда же входят и сайты.