Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра программного обеспечения информационных технологий

Дисциплина: Теория вероятностей и математическая статистика (ТВиМС)

ОТЧЕТ

По расчетной работе

Студент гр.751004 Мячков А.А.

Проверил: Лапицкая Н.В.

Петюкевич Н.С.

Минск 2019

СОДЕРЖАНИЕ

[1. Постановка задачи 3](#_Toc515670981)

[2. Используемые формулы 4](#_Toc515670982)

[3. Анализ данных 8](#_Toc515670983)

[3.1 Дискретная случайная величина 8](#_Toc515670984)

[3.2 Непрерывная случайная величина 14](#_Toc515670985)

[Вывод 21](#_Toc515670986)

**1 Постановка задачи**

В результате статистических наблюдений некоторой совокупности относительно изучаемого признака Х были получены выборочные данные (данные из открытых источников, размер выборки в диапазоне от 100 до 1000, 2 набора – дискретные и непрерывные).

ТРЕБУЕТСЯ:

1. составить дискретный или интервальный ряд распределения частот и частостей случайной величины Х;

2. построить полигон или гистограмму частот;

3. найти эмпирическую функцию распределения признака Х и построить её график;

4. вычислить все возможные числовые характеристики выборочных данных и проанализировать полученные результаты (близость дисперсии к нулю, соотношение между математическим ожиданием и дисперсией, близость оценок средней, моды и медианы, ассиметрия и эксцесс);

5. проверить качество группировки с помощью доли межгрупповой дисперсии в общей дисперсии выборки (правило сложения дисперсий)

6. выдвинуть гипотезу о виде распределения рассматриваемой случайной величины Х. Обосновать выбор вида распределения.

7. Написать аналитическое выражение теоретической функции распределения, найти теоретические (выравнивающие частоты), построить их графики и сравнить теоретические и эмпирические характеристики графически

8. приняв уровень значимости 0.05 по критериям согласия Пирсона, Романовского, Колмогорова, Ястремского, подтвердить или отвергнуть выдвинутую гипотезу о виде распределения.

**2 Использованные формулы**

В данном разделе приведены основные формулы и определение, согласно которым были произведены расчёты и построены соответствующие графики.

**Частота** - число единиц совокупности с данным значением признака Ni. Сумма Ni-тых должна быть равна количеству всех элементов ряда. Ниже приведены виды частот.

*Частота для дискретного ряда* – подсчет количества исходов, где варианта имела некое значения xi.

*Частота для интервального ряда* – подсчет количества исходов, где варианта оказалось на промежутке (x1…x2).

**Частотность**



fi – частота варианты;

N - объём совокупности.

**Средняя арифметическая (она же взвешенная)**



x – значения вариант;

f – частоты вариант.

**Формула Стерджесса**:

****

N – количество элементов в ряду.

h - величина интервала;

xmax – наибольшее значение группового признака;

xmin - наименьшее значение группового признака;

**Математическое ожидание**:

хі  - значение вариант

рі  - соответствующая вероятность

**Дисперсия**:

хі – значение вариант

хср – среднее значение

Мі – соответствующее математическое ожидание

2 формула:

хі – значение вариант

рі - вероятность

**Среднее квадратическое отклонение**:

D– дисперсия

**Мода**: Модой называется наиболее часто встречающееся значение признака у единиц совокупности мода в итерационных рядах.

– нижняя граница модального интервала

- величина интервала

– частота модального интервала

– частота интервала, предшествующего модальному

– частота интервала, следующего за модальным

**Медиана**: Медианой называют значение признака, делящего ранжированный ряд пополам. Медиана есть для дискретного и интервального ряда, что является довольно очевидным фактом.

*Медиана дискретного ряда* – зависит от накопленных частот.

*Медиана в интервальном ряду*

– нижняя граница интервала, который содержит медиану

- величина интервала

– частота медианного интервала

– сумма частот

– сумма накопленных частот интервалов, предшествующих медианному

**Размах вариации:**

хmax– максимальное значение

хmin–минимальное значение

**Коэффициент вариации:**

хср – среднее значение

– среднее квадратическое отклонение

**Асимметрия:**

– момент третьего порядка

– среднее квадратическое отклонение

**Эксцесс:**

– момент четвертого порядка

– среднее квадратическое отклонение

**Центральный момент m-порядка:**

– частота

хі – значение

хср – среднее значение

**Критерий согласия Пирсона**:

mi – эмпирическая частость

m’i – теоретическая частость

**Критерий Романовского:**



 - наблюдаемое значение

k – количество степеней свободы

**Критерий Ястремского**:

k – число степеней свободы

n – число групп

Z = 0.6, при n < 20

**Теорема сложения дисперсий**

******

- общая дисперсия

- средняя из групповых дисперсий

- межгрупповая дисперсия

**Групповая средняя:**

xi – очередное значение в группе

n – объём группы

**Общая средняя:**

k – количество групп

ni – объём очередной группы

xji – очередное значение из группы ni

**Групповая дисперсия:**

xi – очередное значение в группе

n – объём группы

xгр.ср. – групповая средняя данной группы

**Внутригрупповая дисперсия:**

k – количество групп

Dгрi – групповая дисперсия очередной группы

ni – объём очередной группы

**Межгрупповая дисперсия:**

k – количество групп

ni – объём очередной группы

xсрi – групповая средняя очередной группы

xср – общая средняя

**3 Анализ данных**

3.1 Дискретная случайная величина

В качестве исходного дискретного ряда была взята статистика среднего количества потреблённого алкоголя в литрах чистого этанола на душу населения (в возрасте 15 лет и старше) в разных странах мира. В результате ряд состоит из 189 значений, что соответствует критериям задания. Ниже на рисунке 1 приведены скриншоты части таблицы с исходным рядом.

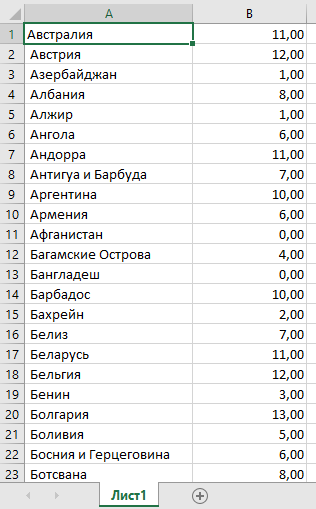


Рисунок 1 – Часть таблицы дискретного ряда

1) Далее на основе полученных данных была построена таблица частот и частостей, показанная на рисунке 2, где Ni – обозначает, сколько раз встречается xi, а Wi – отношение Ni общему количеству элементов ряда, в данном случае 189.

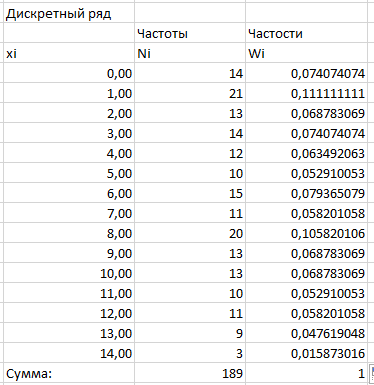


Рисунок 2 – Таблица частот и частостей

2) По данным из вышеприведённой таблицы был построен полигон относительных частот, что демонстрирует рисунок 3.



Рисунок 3 – Полигон относительных частот

3) Эмпирическая функция распределения составляется на основе значений вариант и накопленных ими частот, как видно на рисунке 4. На рисунке 5 также приведён график эмпирической функции.

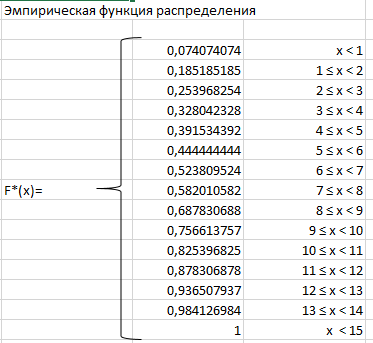


Рисунок 4 – Эмпирическая функция распределения

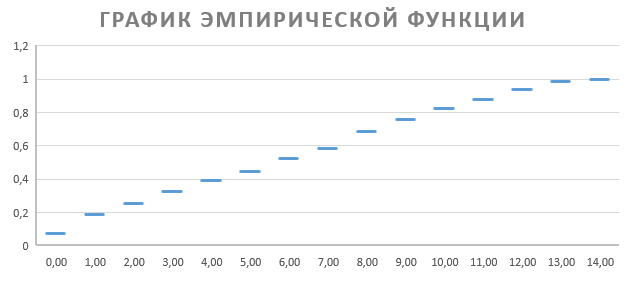


Рисунок 5 – График эмпирической функции распределения

4) Для исходных данных были вычислены все возможные числовые характеристики: математическое ожидание, дисперсия (2 различными способами), средняя взвешенная величина, среднее квадратическое отклонение, мода, медиана, размах вариации, коэффициент вариации, асимметрия и эксцесс. Для нахождения асимметрии и эксцесса также необходимо было дополнительно вычислить центральные моменты 3 и 4 порядков соответственно. Результаты вычислений представлены на рисунке 6. Формулы, по которым производились вычисления, представлены в разделе «Используемые формулы».

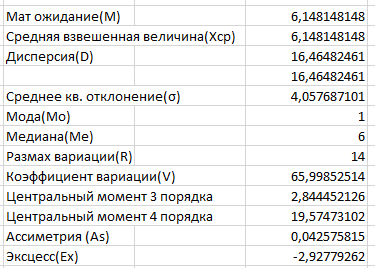


Рисунок 6 – Расчеты числовых характеристик

5) Так как математическое ожидание и дисперсия не равны, то сделан вывод, что исходные данные не подходят под данный вид распределения Пуассона. Но в качестве эксперимента попробуем проверить этот вид распределения, таким образом функция распределения Пуассона будет выглядеть следующим образом:

, где λ – средняя взвешенная величина.

Теоретические (выравнивающие) частоты находятся по формуле:

, где N – число испытаний. В качестве используется функция Пуассона, где .

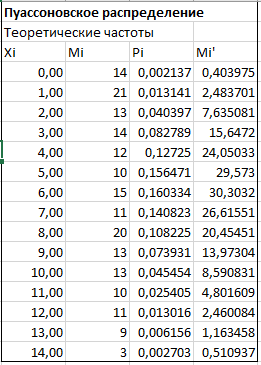


Рисунок 7 – Расчет Пуассоновского распределения

6) По критерию согласия Пирсона были посчитаны χ2 наблюдаемое и χ2 критическое (с использованием соответствующих формул). На основании полученных результатов был сделан следующий вывод: Поскольку неравенство χ2 наблюдаемое ≤ χ2 критическое не выполняется, то выдвинутая гипотеза отвергается.

Критерий согласия Пирсона высчитывается по формуле  и сравнивается по таблице с критическим значением , значением где d – число степеней свободы d = n – r - 1 = 14 – 1 – 1 = 12. Вычисленный критерий меньше критического значения, значит, гипотеза не опровергается и она может быть принята. Критерий Романовского высчитывается по формуле . Если вычисленное значение меньше 3, следовательно, гипотеза снова подтверждается.

Как видно из рисунка 8, ни в одном из критериев не выполнились условия подтверждения гипотезы.

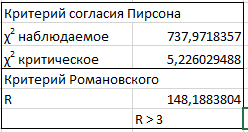


Рисунок 8 – Критерии Пирсона и Романовского

3.2 Непрерывная случайная величина

В качестве исходного непрерывного ряда были взяты результаты замеров температуры в Париже на декабрь 2017г. На рисунке 9 показана таблица значений.

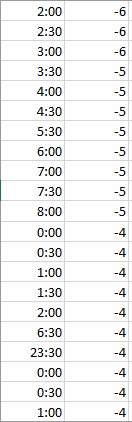


Рисунок 9 – Данные темриратур Парижа

Далее было произведено разбиение на интервалы в соответствии с формулой Стерджесса. Скриншоты таблицы интервалов приведены ниже на рисунке 10.

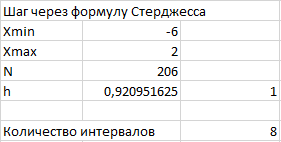


Рисунок 10 – Формулы Стерджесса для данных о погоде

На рисунке 11 представлены интервальные ряд.

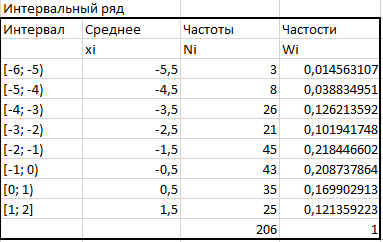


Рисунок 11 – Интервальный ряд данных

На следующем шаге, проанализировав интервалы разбиения, был построена гистограмма частости, график которой приведён на рисунке 12.

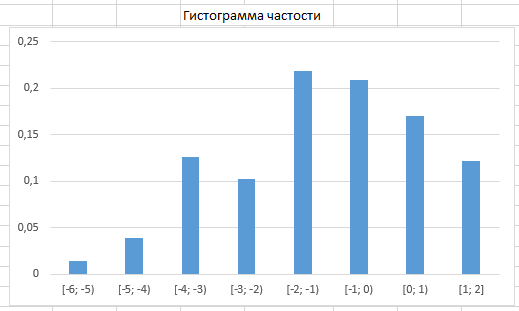


Рисунок 12 – Гистограмма частости данных о погоде

Эмпирическая функция распределения составляется на основе значений вариант и накопленных ими частот. Функция показана в таблице «Эмпирическая функция распределения» Excel документа. На основе полученной функции был построен график.

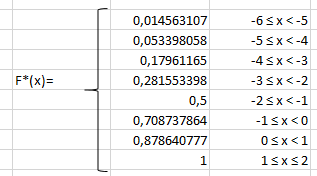
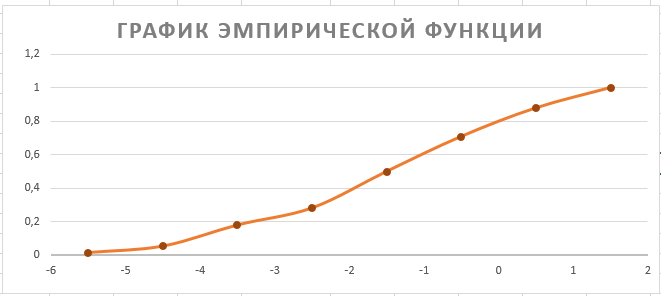
 

Рисунок 13 – Эмпирическая функция распределения

Исходя из внешнего вида гистограммы частот и эмпирической функции распределения, было сформировано предположение, что ряд распределён по закону по нормальному распределению для данных о температуре города героя Парижа.

Для исходных данных были вычислены все возможные числовые характеристики, а именно: математическое ожидание, дисперсия (2 различными способами), средняя взвешенная величина, среднее квадратическое отклонение, среднее линейное отклонение, мода, медиана, размах вариации, коэффициент вариации, асимметрия и эксцесс. Для нахождения асимметрии и эксцесса также необходимо было дополнительно вычислить центральные моменты 3 и 4 порядков соответственно. Результаты вычислений находятся в таблице «4. Числовые характеристики выборочных данных». Все используемые формулы описаны в разделе «Используемые формулы».



Рисунок 14 – Числовые характеристики

Исходя из предположений о том, что полученное распределение данных о температуре – нормальное распределение, то функция распределения будет выглядеть следующим образом:

*,* где µ - математическое ожидание, σ – среднеквадратическое отклонение, .

Теоретические (выравнивающие) частоты находятся по формуле:

Mi' = φ = =

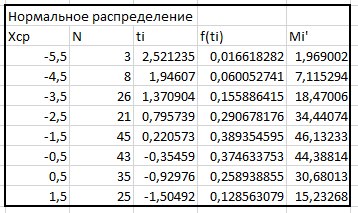


Рисунок 15 – Таблица расчетов нормального распределения

Для подтверждения гипотезы был найден наблюдаемый критерий согласия Пирсона. Формула нахождения критерия находится в разделе «Использованные формулы». Найденное число степеней свободы и принятый уровень значимости 0.05 был найден критический критерий. После сравнения, наблюдаемого и критического критериев между собой, мы не можем принять выдвинутую ранее гипотезу, так как наблюдаемый критерий больше критического. Также были посчитаны критерии Романовского и Ястремского. Поскольку критерий Романовского R < 3, и критерий Ястремкого J < 3, то это доказывает то, что гипотеза может быть принята. Результаты подсчета показаны на рисунке 16.

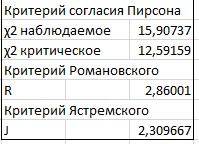


Рисунок 16 – Критерий Пирсона, Романовского и Ястремского для нормального распределения

На примере данных о погоде проведена проверка качества разбиения на интервалы. Для этого используем правило сложения дисперсий:

******

На основе этих данных рассчитываем общую ****** и межгрупповую дисперсию****** , среднюю из внутригрупповых дисперсий***.***

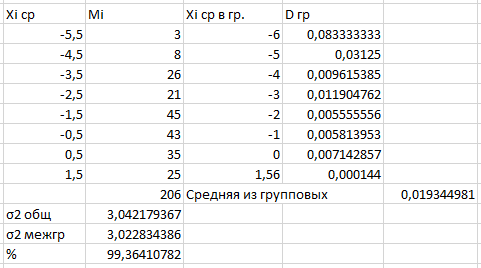


Рисунок 21. Таблица проверки выполнения формулы

Отношение межгрупповой дисперсии к общей составило 99%, а также сумма межгрупповой и средней внутригрупповой дисперсии приблизительно равна общей дисперсии, следовательно, разбиение на группы проведено успешно.

**Вывод**

В ходе проделанной работы были найдены следующие данные: данные о температуре в Париже в качестве непрерывной случайной величины и данные о потреблении алкоголя на душу населения в год в единицах измерения литр в разных странах мира. Эти данные были загружены в Excel для дальнейшей работы с ними.

Для непрерывных случайных величин было произведено разбиение на интервалы, построены гистограммы частостей, найдены эмпирическую функцию распределения и построены её графики, вычислены всевозможные числовые характеристики, проведён анализ качества разбиения на группы с помощью правила сложения дисперсий; выдвинуто предположение о виде распределения, найден критерий согласия Пирсона, а также критерии Ястремского и Романовского, был сделан вывод о нормальном распределениях. Предположение о нормальном распределении подтвердилось двумя из критериев: Романовского и Ястремского. Это значит, что данные последовательности скорее всего распределяются по этим законам.

Для дискретных случайных величин была сформирована таблицу частот и частостей, на основе которых построен полигон относительных частот, найдена эмпирическая функция распределения и простроен её график, вычислены всевозможные числовые характеристики, произведена проверка на пуассоновское распределение. Можно заметить, что в распределении Пуассона дисперсия равна математическому ожиданию (среднему взвешенному). В данном случае величины не равны между собой. Приведённые значения могут означать только одно - это не распределение Пуассона. Но тем не менее была проверена теория распределения Пуассона, которая не оправдалась. Принадлежность к геометрическому и биномиальному взятые данные проверены быть не могут, потому что ничего не известно о вероятности появления события P для данной выборки.