МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра космического приборостроения и средств связи

УТВЕРЖДАЮ
Проректор по учебной работе
____________ О.Г. Локтионова
«_5_» _03_________2024 г.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ КОНСТРУИРОВАНИЯ, ТЕХНОЛОГИИ И НАДЕЖНОСТИ ЭЛЕКТРОННЫХ СРЕДСТВ

Методические указания к практическим занятиям

Курск – 2024

УДК 621.382

Составитель Е. О. Брежнева

Рецензент

Доктор технических наук, профессор Чернецкая И. Е.

Теоретические основы конструирования, технологии и надежности электронных средств: методические указания к практическим занятиям / Юго-Зап. гос. ун-т; сост. Е.О. Брежнева. Курск, 2024. - 52 с.

Приведены задания для практических занятий.

Методические указания соответствуют требованиям программы, утвержденной учебно-методическим объединением по специальностям автоматики и электроники (УМО АЭ).

Предназначены для студентов направления подготовки бакалавров 11.03.03 «Конструирование и технология электронных средств» всех форм обучения.

Текст печатается в авторской редакции

Подписано в печать 5.03. Формат 60×84 1/16.
Усл. печ. л. 3,02. Уч.- изд. л. 2,74. Тираж 100 экз. Заказ 145.
Бесплатно.
Юго-Западный государственный университет.
305040, г. Курск, ул. 50 лет Октября, 94

СОДЕРЖАНИЕ

Цель и задачи практических занятий

Целью практических занятий является формирование умений и навыков в области теоретических основ конструирования электронных средств.

Задачи практических занятий:

- изучить аналитические и экспериментальные математические методы, используемые при конструировании электронных средств.

Планируемые результаты обучения

В ходе выполнения заданий на практических занятиях формируются следующие компетенции:

ОПК-3 - способен применять методы поиска, хранения, обработки, анализа и представления в требуемом формате информации из различных источников и баз данных, соблюдая при этом основные требования информационной безопасности;

ОПК-4 - способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности.

Обучающийся должен

знать:

- основы статистической обработки данных и закона распределения случайной величины;
- методы регрессионного, дисперсионного и корреляционного анализа и группировки данных;

уметь:

- определять числовые характеристики случайной величины и осуществлять проверку статистических гипотез;
- использовать современные информационные технологии при решении практических задач в области теоретических основ конструирования электронных средств;

- применять методы регрессионного, дисперсионного и корреляционного анализа и группировки данных при решении практических задач;

владеть:

- навыками экспериментального исследования случайных величин;
- методами сбора, анализа и обработки данных;
- навыками использования современных информационных технологий для обработки данных и представления результатов.

Необходимые материально-техническое оборудование и материалы

1. Microsoft Windows Professional 7 Russian (Upgrade Academic OPEN1 License No Level № 60803556 - 13 копий);
2. LibreOffice (LGPL v3);
3. Антивирус Касперского (или ESETNOD);
4. ПК (Processor i5-2500, RAM DDR3 4 GB, HDD 320 GB, DVD RW, TFT-монитор 24” 1920x1080);
5. MatLab R2012b (лицензия №820456) – пакет прикладных программ.

1. Построение эмпирических функции и плотности распределения

1.1 Цель работы

1. Приобретение навыков построения статистических рядов.
2. Знакомство с графическим представлением статистических рядов.
3. Получение навыков использования современных информационных технологий при решении практических задач.

1.2 Основные теоретические положения

Формы представления эмпирических распределений

Статистическое исследование состоит из трех основных этапов. Первым этапом является наблюдение, при котором производится научно обоснованный сбор данных, характеризующих изучаемое явление или объект. Второй этап заключается в статистической сводке и группировке. На этом этапе данные систематизируются и определенным образом оформляются – чаще всего в виде статистических таблиц. Третьим этапом является анализ статистического материала.

Применение того или иного метода статистического анализа определяется математической моделью, описывающей свойства генеральной совокупности. Для корректного проведения эксперимента выбор и обоснование математической модели должны быть произведены до его начала. На практике при проведении обследований ограниченный объем предварительной информации не позволяет сделать обоснованное предположение о математической модели генеральной совокупности. В таких случаях ее выбор осуществляется на основе построения эмпирического распределения и анализа его характеристик. Под **эмпирическим распределением** принято понимать распределение элементов выборки по значениям изучаемого признака. Построение эмпирических распределений является необходимым этапом применения статистических методов. Основной задачей при построении эмпирического распределения является формулирование на основе его анализа предположения о форме распределения изучаемого признака в генеральной совокупности.

Построение статистических рядов

Выборка, полученная при проведении экспериментального исследования, представляет собой неупорядоченный набор чисел, записанных в той последовательности, в которой производились измерения. Обычно выборка оформляется в виде таблицы, в первой строке (или столбце) которой стоит номер опыта i, а во второй (втором) - зафиксированное значение случайной величины признака (таблица 1.1). В таком виде выборка представляет собой первичную форму записи статистического материала, который может быть обработан различными способами.

Таблица 1.1

№	1	2	3	4	5	6	7	8	9	10
x_i										

Проанализировать экспериментальные данные и тем более сделать какие-либо выводы на основе простой статистической совокупности весьма затруднительно. Исходя из этого, полученный статистический материал должен быть обработан для проведения дальнейшего исследования. Простейшим способом обработки выборки является ранжирование. Ранжированием называют расстановку вариант в порядке возрастания или убывания их значений.

Но и в таком виде полученные экспериментальные данные плохо обозримы и мало пригодны для непосредственного анализа. Именно поэтому для придания статистическому материалу большей компактности и наглядности он должен быть подвергнут дальнейшей обработке – строится статистический ряд. Построение статистического ряда начинается с группировки.

Группировкой называется процесс упорядочения и систематизации данных, полученных в ходе проведения эксперимента, направленный на извлечение содержащейся в них информации. В процессе группировки осуществляется распределение вариант выборки по группам или интервалам группировки, каждый из которых содержит некоторый диапазон значений изучаемого признака. Процесс группировки начинается с разбиения всего диапазона варьирования признака на интервалы группировки.

Ориентировочное значение оптимального числа интервалов k может быть определено, исходя из объема выборки n, с помощью формулы Стэрджесса:

$$k = 1 + 3{,}322\ \lg n.$$

Получаемое по формуле значение k почти всегда оказывается дробной величиной, которую необходимо округлить до целого числа, поскольку количество интервалов не может быть дробным. Практика показывает, что, как правило, лучше округлять в меньшую сторону.

После того, как число интервалов группировки определено, вычисляется ширина интервалов:

$$h = \frac{x_{\max} - x_{\min}}{k}.$$

Здесь h - ширина интервалов, а $x_{\max}$ и $x_{\min}$ - соответственно максимальное и минимальное значение признака в выборке. Величины $x_{\max}$ и $x_{\min}$ определяются непосредственно по таблице исходных данных.

Точность вычисленного значения h может либо совпадать с точностью проведения эксперимента, либо превышать ее. Особо отметим, что округление необходимо производить не в общепринятом математическом смысле, а в сторону увеличения, т.е. с избытком, чтобы не уменьшить общий диапазон варьирования признака - сумма ширины всех интервалов не должна быть меньше разности между максимальным и минимальным значениями признака.

После определения ширины интервалов группировки следует определить их границы. Нижнюю границу первого интервала целесообразно принять равной минимальному значению признака в выборке $x_{\min}$:

$$x_{\text{Н}1} = x_{\min}.$$

Для получения верхней границы первого интервала ($x_{\text{В}1}$) следует к значению нижней границы первого интервала прибавить значение ширины интервала:

$$x_{\text{В}1} = x_{\text{Н}1} + h.$$

Заметим, что верхняя граница каждого интервала (здесь – первого) будет являться одновременно и нижней границей следующего (в данном случае второго) интервала: $x_{Н2}= x_{В1}$.

Подобным образом определяются значения нижних и верхних границ всех оставшихся интервалов:

$$x_{Вi} = x_{Нi+1} = x_{Нi} + h.$$

Перед группировкой вариант введем понятие **срединного значения интервала** x_i, равного значению признака, равноудаленного от концов этого интервала. Учитывая, что оно отстоит от нижней границы на величину, равную половине ширины интервала, для его определения удобно воспользоваться соотношением:

$$x_i = x_{Нi} + h/2,$$

где $x_{Нi}$ - нижняя граница i-го интервала, а h - его ширина. Срединные значения интервалов будут использоваться в дальнейшем при обработке сгруппированных данных.

После определения границ всех интервалов следует распределить выборочные варианты по этим интервалам. Если варианта может быть отнесена к любому из двух соседних интервалов, для исключения неоднозначности при группировке, условимся относить варианты к верхнему интервалу.

Перейдем к рассмотрению статистической таблицы (таблица 1.2), которая состоит из семи столбцов.

Таблица 1.2- Статистическая таблица

№ интервала	Границы интервала	Срединное значение интервала	Частота	Накопленная частота	Частость	Накопленная частость
i	$x_{Hi} - x_{Bi}$	x_i	n_i	N_i	f_i	F_i
			$\sum$		$\sum$	

В первых трех столбцах статистической таблицы содержатся соответственно номера интервалов группировки i, их границы x_{Hi} - x_{Bi} и срединные значения интервалов x_i.

В четвертом столбце располагаются частоты интервалов. **Частотой** интервала называется число, показывающее сколько вариант, т.е. результатов измерений попало в данный интервал. Для обозначения этой величины принято использовать символ n_i. Сумма всех частот всех интервалов всегда равна объему выборки n, что можно использовать для проверки правильности проведенной группировки.

Пятый столбец таблицы 1.2 предназначен для занесения в него **накопленной частоты** интервала - числа, полученного суммированием частоты текущего интервала с частотами всех предыдущих интервалов. Накопленную частоту принято обозначать латинской буквой N_i. Накопленная частота показывает, сколько вариант имеют значения не больше, чем верхняя граница интервала.

В шестой столбец таблицы помещается частость. **Частностью** называется частота, представленная в относительном выражении, т.е. отношение частоты к объему выборки. Сумма всех частостей всегда равна 1. Для обозначения частости используется символ f_i:

$$f_i=n_i/n.$$

И частота, и частость характеризуют повторяемость результатов в выборке. Следует отметить важное достоинство частости. Ее использование позволяет сопоставлять выборки различного объема. Частота для таких целей не применима.

В седьмом столбце таблицы расположена накопленная частость. **Накопленной частостью** является отношение накопленной частоты к объему выборки. Накопленная частость обозначается буквой F_i:

$$F_i = \frac{N_i}{n} .$$

Накопленная частость показывает, какая доля вариант выборки имеет значения, не превосходящие значения верхней границы интервала.

Последняя строка статистической таблицы используется для контроля над проведением группировки.

Статистическим рядом называется двойной числовой ряд, устанавливающий связь между численным значением исследуемого признака и его повторяемостью в выборке. Существенным достоинством статистических рядов является то, что они, в отличие от статистических совокупностей, дают наглядное представление о характерных особенностях варьирования признаков.

Графическое представление статистических рядов

В целях упрощения анализа статистических рядов и придания им большей наглядности используют графические представления. Основными видами графического представления статистических рядов являются гистограмма, полигон частостей и полигон накопленных частостей. Для визуального представления можно использовать как частости, так и частоты. Ограничимся рассмотрением частости, поскольку этот параметр более информативен.

Наиболее часто для анализа статистического ряда используется **гистограмма**, представляющая собой совокупность примыкающих друг к другу прямоугольников, основание каждого из которых равно ширине интервала группировки, а площадь - частости этого интервала.

Гистограмма строится в декартовой (прямоугольной) системе координат следующим образом. По оси абсцисс откладываются отрезки, отображающие интервалы группировки, а затем на каждом из них строится прямоугольник, площадь которого равна частости данного интервала. Для удовлетворения этому требованию высота прямоугольника выбирается равной частному от деления частости интервала на его ширину $H_i = f_i / h_i$. В случае, если все интервалы группировки имеют одинаковую ширину, высоты прямоугольников пропорциональны соответствующим частостям. Полная площадь гистограммы равна единице, что следует из способа ее построения. Действительно, площадь каждого из прямоугольников равна частости, а сумма всех частостей - единица.

В качестве примера на рисунке 1.1 приведена гистограмма распределения.

С увеличением числа экспериментальных данных можно использовать большее количество интервалов, имеющих меньшие ширины. Гистограмма при этом будет все более и более приближаться к некоторой кривой, ограничивающей площадь, равную единице. Эта кривая представляет собой не что иное как график плотности распределения (или, по-другому, плотности вероятности) исследуемой случайной величины. Таким образом, гистограмма является экспериментальным аналогом плотности распределения.

Другим распространенным способом графического представления статистических рядов является полигон частостей. **Полигон частостей** отображает зависимость частости от срединных значений интервалов. Полигон частостей строится в декартовой системе координат путем соединения прямыми линиями точек, абсциссы которых равны срединным значениям интервалов, а ординаты - частостям этих интервалов. Эти данные располагаются в третьем и шестом столбцах таблицы 1.2.

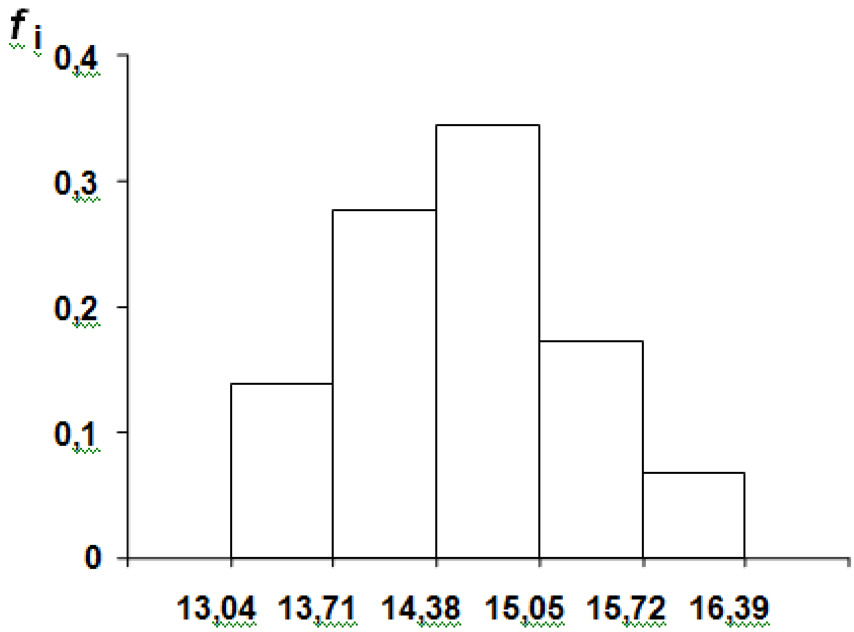

Рисунок 1.1 - Гистограмма

Полигон частостей может быть получен из гистограммы путем соединения середин верхних сторон прямоугольников гистограммы отрезками прямых. Полигон частостей изображен на рисунке 1.2.

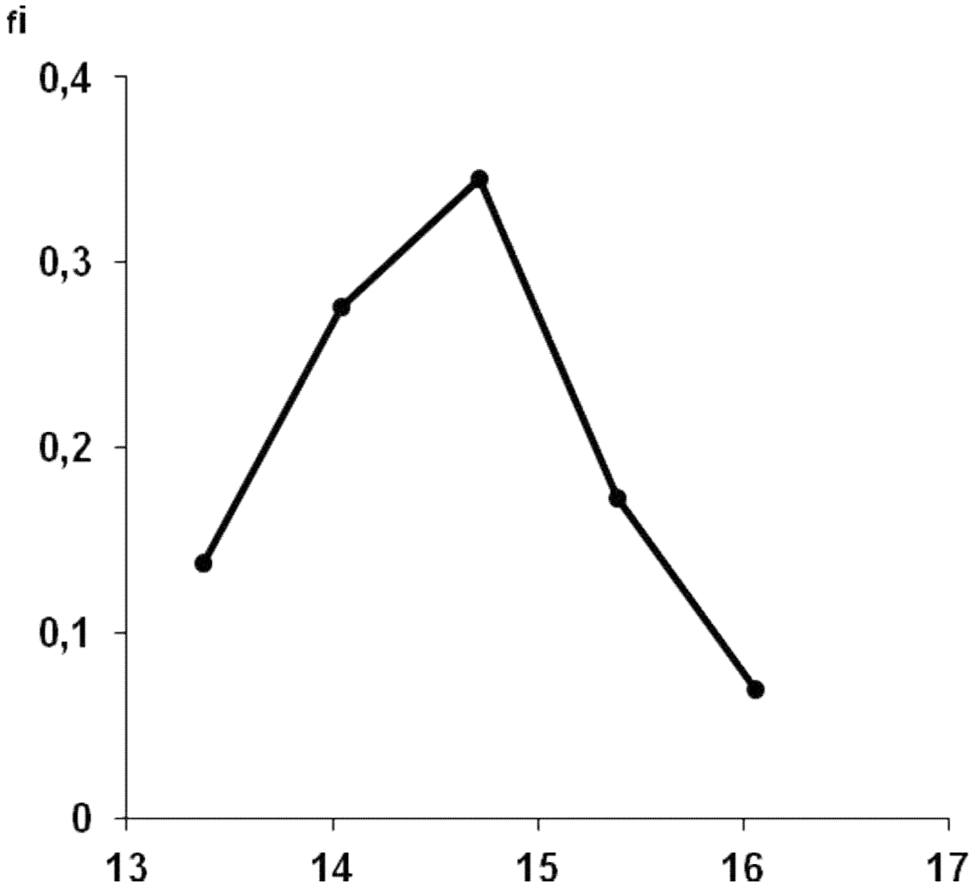

Рисунок 1.2 - Полигон частостей

Полигон частостей может оказаться более удобным и наглядным способом графического представления, чем гистограмма, в том случае, когда признак является непрерывным и его распределение описывается плавной зависимостью.

Полигон накопленных частостей представляет собой зависимость накопленных частостей от значений верхних границ интервалов. Полигон накопленных частостей строится в декартовой системе координат посредством соединения прямыми линиями точек, абсциссы которых равны значениям верхних границ интервалов, а ординаты - накопленным частостям этих интервалов. Эти данные располагаются во втором и седьмом столбцах таблицы 1.2. Полигон накопленных частостей приведен на рисунке 1.3.

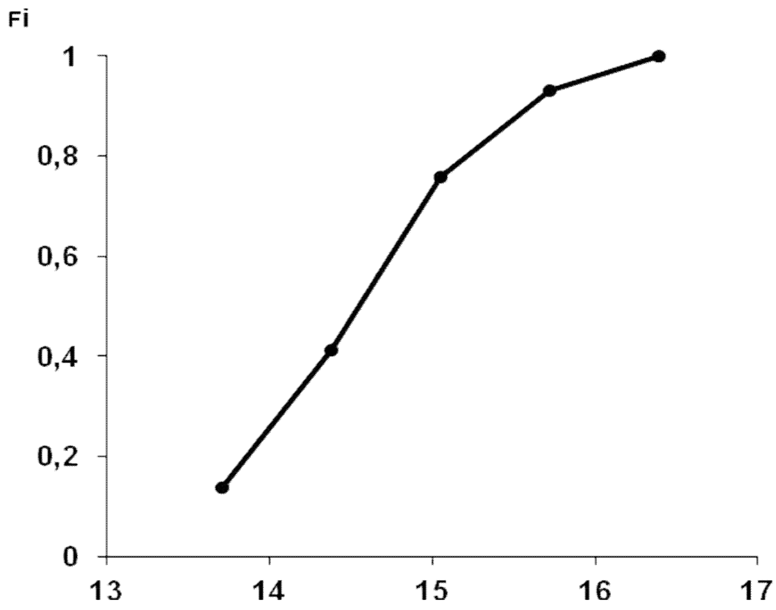

Рисунок 1.3 - Полигон накопленных частостей

С увеличением числа опытных данных в выборке и соответственно увеличением числа используемых интервалов полигон накопленных частостей будет приближаться к кривой, являющейся графиком функции распределения исследуемой случайной величины. Таким образом, он является экспериментальным аналогом функции распределения.

1.3 Задачи для решения

Для экспериментальных данных, в соответствии с номером варианта, построить статистический ряд, гистограмму, полигон частостей и полигон накопленных частостей. Сделать выводы.

Таблица 1.3- Вариант №1

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0,53	1,83	-2,25	0,86	0,31	-1,30	-0,43	0,34	3,57	2,76
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	-1,34	3,03	0,72	-0,06	0,71	-0,20	-0,12	1,48	1,40	1,41
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	0,67	-1,20	0,71	1,63	0,48	1,03	0,72	-0,30	0,29	-0,78

Таблица 1.4- Вариант №2

№	1	2	3	4	5	6	7	8	9	10
x_i	101	99	99	100	98	102	101	100	102	99
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	100	100	101	101	100	100	100	101	102	102
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	100	101	99	99	100	102	100	101	100	102

Таблица 1.5 - Вариант №3

№	1	2	3	4	5	6	7	8	9	10
x_i	29,2	10,0	25,4	26,3	27,3	11,6	17,9	15,1	26,0	18,6
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	28,2	13,6	15,2	12,9	12,7	27,3	21,5	20,9	12,8	27,0
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	22,4	17,0	20,2	18,0	11,5	14,7	12,4	13,6	14,7	18,3

Таблица 1.6 - Вариант №4

№	1	2	3	4	5	6	7	8	9	10
x_i	12	47	48	30	30	24	47	25	15	42
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	26	20	27	14	16	48	49	34	13	20
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	25	43	11	12	17	36	40	36	29	32

Таблица 1.7 - Вариант №5

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	14	19	13	18	13	15	17	19	11	21
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	19	16	15	15	14	16	16	19	19	18
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	15	19	16	14	21	20	17	17	17	13

Таблица 1.8 - Вариант №6

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0,52	0,15	0,29	0,18	0,57	0,69	0,07	0,10	0,02	0,89
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	0,68	0,92	0,28	0,65	0,04	0,92	1,24	0,23	0,10	0,46
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	0,68	0,43	0,62	0,26	0,03	0,18	0,91	0,37	1,16	0,27

Таблица 1.9 - Вариант №7

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	12	11	15	11	12	11	26	15	12	16
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	17	12	13	13	12	24	15	14	17	12
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	11	16	12	19	11	19	13	15	11	14

Таблица 1.10 - Вариант №8

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	14	11	15	14	11	24	11	18	12	13
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	12	15	13	11	28	21	11	12	17	14
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	20	16	16	15	19	15	20	11	21	11

Таблица 1.11- Вариант №9

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	5,68	4,14	3,92	4,90	4,74	6,19	5,60	5,54	3,55	4,03
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	5,20	4,65	6,29	6,34	4,41	5,87	6,39	5,32	6,62	6,06
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	5,21	5,87	5,19	4,58	5,35	5,03	4,63	6,77	5,22	7,73

Таблица 1.12- Вариант №10

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	15	17	20	21	18	13	18	11	15	18
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	21	19	16	19	15	21	21	20	15	16
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	13	19	20	21	17	17	12	20	15	13

1.4 Контрольные вопросы

1. Дать определение случайной величины.
2. Генеральная и выборочная совокупности.
3. Формы представления эмпирических распределений.
4. Последовательность построения статистических рядов.
5. Определение числа и ширины интервалов.
6. Частота и накопленная частота интервалов: определения, расчет.
7. Частость и накопленная частость: определения, расчет.
8. Графическое представление статистических рядов. Гистограмма.
9. Графическое представление статистических рядов. Полигон частостей.
10. Графическое представление статистических рядов. Полигон накопленных частостей.
11. Что понимается под эмпирическим распределением?
12. Перечислите этапы статистического исследования/
13. Назовите формы представления данных.
14. Запишите формулу Стэрджесса.
15. Как округляется, полученное по формуле Стэрджесса число интервалов?
16. Запишите формулу для расчета ширины интервалов.
17. Как округляется результат, полученный по формуле для расчета ширины интервала?
18. Как определяется нижняя граница первого интервала?
19. К какому интервалу присваивают значение, попавшее на границу интервалов?
20. Как определяют верхнюю границу интервала?

2 Вычисление и анализ выборочных числовых характеристик

2.1 Цель работы

Целью работы является формирование навыков определения числовых характеристик выборки и анализа полученных результатов.

2.2 Основные теоретические положения

Числовые характеристики выборки

В практических целях для получения информации о поведении признака часто используют отдельные числовые параметры, до некоторой степени, характеризующие существенные черты распределения. Использование таких характеристик позволяет компактно выразить все существенные сведения с помощью минимального количества числовых параметров. Такие характеристики, назначение которых - выразить в сжатой форме наиболее существенные особенности распределения, называются числовыми характеристиками. Среди числовых характеристик наибольшее практическое значение имеют характеристики положения, рассеяния и формы распределений.

Характеристики положения

Рассмотрение числовых характеристик выборки необходимо начать с тех из них, которые характеризуют положение значений исследуемого признака на числовой оси, т. е. указывают некоторое среднее, ориентировочное значение, около которого группируются экспериментальные данные. К ним относятся среднее арифметическое, мода и медиана.

Среднее арифметическое равно сумме значений всех вариант выборки, деленное на объем выборки:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; .$$

Здесь n - объем выборки, а x_i - варианты выборки.

Среднее арифметическое является наиболее важной характеристикой положения, поскольку при его определении используется вся имеющаяся информация о выборке. Для обозначения среднего арифметического используется та же буква, что и для вариант выборки, с той лишь разницей, что над буквой ставится черта - символ усреднения. В рассматриваемом случае исследуемый признак обозначен через X, его числовые значения - через x_i, а среднее арифметическое имеет обозначение $\bar{x}$.

Из определения среднего арифметического следует, что сумма отклонений выборочных значений признака от него равна нулю.

Вычислять среднее арифметическое исходя из его определения при большом объеме выборки становится затруднительным и можно применить следующий прием: воспользоваться результатами группировки и считать приближенно значения вариант в каждом интервале постоянными и равными срединному значению, которое выступает в роли «представителя» интервала. Число вариант в интервале равно частоте интервала, поэтому среднее арифметическое для сгруппированных данных будет выражаться следующей приближенной формулой:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \; ,$$

где п - объем выборки; k - число интервалов группировки; n_i - частоты интервалов; x_i - срединные значения интервалов.

Среди других характеристик положения наиболее важны мода и медиана. Они характеризуют величину варианты, занимающей определенное положение в статистической совокупности.

Модой случайной величины называется значение признака, встречающееся в выборке наиболее часто. Условимся использовать для обозначения моды символы *Мо*. Геометрически мода соответствует максимуму кривой эмпирического распределения.

Прежде чем приступить к вычислению значения моды в случае сгруппированных данных, необходимо определить модальный интервал. **Модальным** называется интервал группировки, содержащий наибольшее число вариант, т.е. имеющий максимальную частоту (частость).

Значение моды определяется по результатам группировки с помощью следующего соотношения:

$$Mo = x_{MoH} + h\frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})},$$

где x_{MoH} - нижняя граница модального интервала; h -ширина интервала группировки; n_{Mo} - частота модального интервала; n_{Mo-1} - частота интервала, предшествующего модальному; $n_{мo+1}$ - частота интервала, следующего за модальным.

При проведении исследования может оказаться, что модальным оказывается первый или последний интервал группировки. В этом случае предыдущий или последующий интервал не существует и возникает вопрос о пути применения последней формулы. Если один из интервалов не существует, то при проведении вычисления моды значение частоты, соответствующее этому интервалу, следует принять равным нулю.

Часто для характеристики распределения применяется еще одна характеристика положения - медиана. **Медианой** называется такое значение признака, при котором половина значений экспериментальных данных оказывается меньше его, а вторая половина — больше. Для обозначения медианы принято использовать символы *Me*. Геометрический смысл медианы – это абсцисса точки, в которой площадь, ограниченная кривой распределения, делится пополам.

В случае несгруппированных данных для нахождения медианы необходимо ранжировать выборку, т. е. расположить данные в порядке их возрастания или убывания. Медианой будет являться значение признака, находящееся в середине ранжированного ряда. В ранжированной выборке, содержащей *n* членов, ранг R_{Me}, т.е. порядковый номер, медианы равен:

$$R_{Me} = \frac{n+1}{2},$$

а сама медиана совпадает с членом выборки, имеющим номер R_{Me}. Описанное правило дает однозначный результат, если выборка содержит нечетное число членов.

Если же выборка содержит четное число членов, то медиана не может быть определена столь однозначно. Действительно, R_{Me} оказывается дробным. В этом случае берут два члена выборки с номерами большим и меньшим R_{Me} и считают медиану, равной их среднему значению.

Для определения медианы в случае сгруппированных данных необходимо найти медианный интервал. Интервал группировки, содержащий медиану, называется медианным. **Медианным** является интервал, в котором накопленная частота впервые окажется больше половины объема выборки (либо накопленная частость - больше 0,5). Значение медианы определяется по следующей формуле:

$$Me = x_{MeH} + h\frac{0{,}5n - N_{Me-1}}{n_{Me}},$$

где $x_{Me\text{Н}}$ - нижняя граница медианного интервала; n - объем выборки; h - ширина интервалов группировки; $N_{Me\text{-}1}$ - накопленная частота интервала, предшествующего медианному; n_{Me} - частота медианного интервала.

Значения среднего арифметического, моды и медианы совпадают только для симметричных одномодальных распределений. Значение медианы наиболее важно при исследовании сильно асимметричных эмпирических распределений.

Характеристики рассеяния

Для описания распределения необходимо охарактеризовать диапазон изменения значений признака. Для описания диапазона варьирования признака используются характеристики рассеяния.

Наиболее широкое применение нашли размах вариации, дисперсия, стандартное отклонение и коэффициент вариации.

Размах вариации определяется как разность между максимальным и минимальным значением признака в изучаемой совокупности:

$$R=x_{\max}-x_{\min}.$$

Второй характеристикой рассеяния является **дисперсия.** Дисперсия представляет собой средний квадрат отклонения значения случайной величины от ее среднего значения. Дисперсия есть характеристика рассеяния, разбросанности значений величины около ее среднего значения.

При проведении выборочных исследований необходимо установить оценку для дисперсии. Дисперсия, вычисляемая по выборочным данным, называется выборочной дисперсией и обозначается S^2.

На первый взгляд наиболее естественной оценкой для дисперсии является статистическая дисперсия, вычисленная, исходя из определения, по формуле:

$$S^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}.$$

В этой формуле $\sum_{i=1}^{n}(x_i-\bar{x})^2$ - сумма квадратов отклонений значений признака x_i от среднего арифметического $\bar{x}$. Для получения среднего квадрата отклонений эта сумма поделена на объем выборки n.

Однако такая оценка не является несмещенной. Можно показать, что сумма квадратов отклонений значений признака для выборочного среднего арифметического меньше, чем сумма квадратов отклонений от любой другой величины, в том числе от истинного среднего (математического ожидания). Поэтому результат, получаемый по приведенной выше формуле, будет содержать систематическую ошибку, и оценочное значение дисперсии окажется заниженным. Для ликвидации смещения достаточно ввести поправочный коэффициент $\frac{n}{n-1}$. В результате получается следующее соотношение для оценочной дисперсии:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

При больших значениях *n*, естественно, обе оценки - смещенная и несмещенная – будут различаться очень мало и введение поправочного множителя теряет смысл. Как правило, уточнение формулы для оценки дисперсии следует производить при $n<30$.

В случае сгруппированных данных последнюю формулу для упрощения вычислений можно привести к следующему виду:

$$S^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{n-1},$$

где k - число интервалов группировки;

n_i - частота интервала с номером i;

x_i - срединное значение интервала с номером i.

Дисперсия случайной величины имеет размерность квадрата размерности случайной величины, что затрудняет ее интерпретацию и делает не очень наглядной. Для более наглядного описания рассеяния удобнее пользоваться характеристикой, размерность которой совпадает с размерностью исследуемого признака. С этой целью вводится понятие **стандартного отклонения** (или **среднего квадратического отклонения**).

Стандартным отклонением называется положительный корень квадратный из дисперсии:

$$S = \sqrt{S^2}.$$

Стандартное отклонение характеризует степень отклонения признака от среднего арифметического.

Помимо абсолютных показателей вариации, которыми являются дисперсия и стандартное отклонение, в статистике вводятся относительные. Наиболее часто применяется коэффициент вариации. **Коэффициент вариации** равен отношению стандартного отклонения к среднему арифметическому, выраженному в процентах:

$$V = \frac{S}{\bar{x}} 100\% .$$

Из определения ясно, что по своему смыслу коэффициент вариации представляет собой относительную меру рассеяния признака.

Характеристики формы

При проведении статистических исследований встречаются распределения, имеющие самые разнообразные формы. Для характеристики отклонения формы распределения от симметричной используется **коэффициент** асимметрии или просто **асимметрия,** обозначаемая *As* и вычисляемая по формуле:

$$As = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{nS^3},$$

где x_i - значение *i*-й варианты; $\bar{x}$ - среднее арифметическое; S - среднее квадратическое отклонение; n - объем выборки.

Для симметричной формы распределения коэффициент асимметрии равен нулю. На рисунке 2.1 и 2.2 показано два асимметричных распределения. Одно из них (рисунок 2.1) имеет положительную асимметрию ($As>0$), а другое (рисунок 2.2) – отрицательную ($As<0$). Иногда положительную асимметрию называют левосторонней, а отрицательную – правосторонней.

Для сгруппированных данных формула для вычисления коэффициента асимметрии имеет вид:

$$As = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{nS^3}.$$

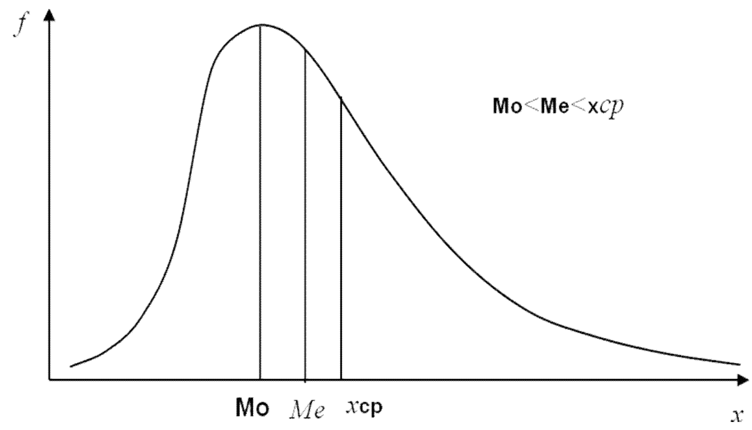

Рисунок 2.1 - Положительная (левосторонняя) асимметрия

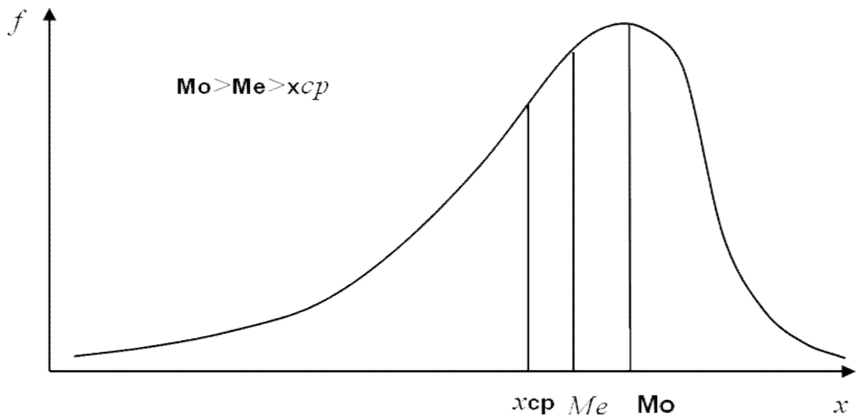

Рисунок 2.2 - Отрицательная (правосторонняя) асимметрия

Здесь n_i –частота интервала с номером i; x_i - его срединное значение; k - число интервалов группировки.

Для быстрой предварительной оценки асимметрии распределения можно воспользоваться ее простейшим показателем - мерой скошенности**. Мера скошенности** (*Sk*) определяется как отклонение среднего арифметического ($\bar{x}$) от моды (*Mo*):

$$Sk = \frac{\bar{x} - Mo}{S}.$$

Нормировка на среднее квадратическое отклонение S производится для обезразмеривания, что необходимо для сравнительного анализа степени асимметрии различных распределений.

Следующий показатель - *эксцесс* - служит для характеристики так называемой крутости, т.е. островершинности или плосковершинности распределения. **Эксцессом** называется случайная величина, определяемая соотношением:

$$Ex = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{nS^4} - 3 \ .$$

Кривые, более **островершинные** по сравнению с кривой нормального распределения, обладают положительным эксцессом, а кривые более **плосковершинные** – отрицательным эксцессом. Таким образом, нормальное распределение служит эталоном, а эксцесс показывает крутизну эмпирического распределения относительно крутизны кривой нормального распределения.

2.3 Задачи для решения

Для экспериментальных данных, в соответствии с номером варианта, определить характеристики положения, рассеяния и формы распределений. Сделать выводы.

Таблица 2.1- Вариант №1

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0,53	1,83	-2,25	0,86	0,31	-1,30	-0,43	0,34	3,57	2,76
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	-1,34	3,03	0,72	-0,06	0,71	-0,20	-0,12	1,48	1,40	1,41
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	0,67	-1,20	0,71	1,63	0,48	1,03	0,72	-0,30	0,29	-0,78

Таблица 2.2- Вариант №2

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	101	99	99	100	98	102	101	100	102	99
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	100	100	101	101	100	100	100	101	102	102
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	100	101	99	99	100	102	100	101	100	102

Таблица 2.3- Вариант №3

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	29,2	10,0	25,4	26,3	27,3	11,6	17,9	15,1	26,0	18,6
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	28,2	13,6	15,2	12,9	12,7	27,3	21,5	20,9	12,8	27,0
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	22,4	17,0	20,2	18,0	11,5	14,7	12,4	13,6	14,7	18,3

Таблица 2.4- Вариант №4

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	12	47	48	30	30	24	47	25	15	42
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	26	20	27	14	16	48	49	34	13	20
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	25	43	11	12	17	36	40	36	29	32

Таблица 2.5- Вариант №5

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	14	19	13	18	13	15	17	19	11	21
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	19	16	15	15	14	16	16	19	19	18
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	15	19	16	14	21	20	17	17	17	13

Таблица 2.6- Вариант №6

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	0,52	0,15	0,29	0,18	0,57	0,69	0,07	0,10	0,02	0,89
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	0,68	0,92	0,28	0,65	0,04	0,92	1,24	0,23	0,10	0,46
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	0,68	0,43	0,62	0,26	0,03	0,18	0,91	0,37	1,16	0,27

Таблица 2.7- Вариант №7

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	12	11	15	11	12	11	26	15	12	16
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	17	12	13	13	12	24	15	14	17	12
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	11	16	12	19	11	19	13	15	11	14

Таблица 2.8- Вариант №8

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	14	11	15	14	11	24	11	18	12	13
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	12	15	13	11	28	21	11	12	17	14
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	20	16	16	15	19	15	20	11	21	11

Таблица 2.9- Вариант №9

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	5,68	4,14	3,92	4,90	4,74	6,19	5,60	5,54	3,55	4,03
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	5,20	4,65	6,29	6,34	4,41	5,87	6,39	5,32	6,62	6,06
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	5,21	5,87	5,19	4,58	5,35	5,03	4,63	6,77	5,22	7,73

Таблица 2.10- Вариант №10

№	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x_i	15	17	20	21	18	13	18	11	15	18
№	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
x_i	21	19	16	19	15	21	21	20	15	16
№	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	13	19	20	21	17	17	12	20	15	13

2.4 Контрольные вопросы

1. Дайте определения: вариационного ряда; статистического ряда; гистограммы; диаграммы; выборочной числовой характеристики (среднее, дисперсия, мода, медиана).
2. Перечислите функции MATLAB, используемые при статистической обработке результатов измерений.
3. Назовите требования к числовым характеристикам выборки.
4. Перечислите характеристики положения.
5. Перечислите характеристики рассеяния.
6. Перечислите характеристики формы.
7. Запишите формулу для расчета среднего арифметического.
8. В каких случаях при расчетах используют формулы для сгруппированных данных? К каким последствиям это приводит?
9. Как определяется модальный интервал группировки?
10. Как рассчитывается мода по результатам группировки?
11. Как поступают в случае, если модальным оказывается первый или последний интервал группировки?
12. Какое значение признака называется медианой?
13. Каков геометрический смысл медианы?
14. Приведите формулу для расчета медианы для несгрупированных данных.
15. Как поступают при расчете медианы, если выборка содержит четное число членов?
16. Какой интервал группировки называется медианным?
17. Характеристики положения (среднее арифметическое, мода и медиана): расчет для несгруппированных данных и сгруппированных данных, интерпретация результатов.
18. Характеристики рассеяния (размах вариации, дисперсия, стандартное отклонение и коэффициент вариации): расчет для несгруппированных данных и сгруппированных данных, интерпретация результатов.
19. Чему равен коэффициент ассиметрии для симметричной формы?
20. Каким эксцессом обладают островершинные кривые?

3 Проверка статистических гипотез

3.1 Цель работы

1. Целью работы является формирование навыков решения задач, связанных с проверкой статистических гипотез.

3.2 Основные теоретические положения

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известного распределения.

Постановка задачи начинается с выдвижения основного утверждения (нулевой гипотезы H_0), причем наряду с выдвинутой гипотезой всегда рассматривают и противоречащую ей гипотезу, которую называют конкурирующей (альтернативной) гипотезой H_1.

Пример: 1) H_0: $p=0.5$ 2) H_0: $m=3$
H_1: $p\neq0.5$ H_1: $m>3$

Далее на основе экспериментальной информации конструируется специально подобранная из разумных соображений случайная величина, распределение которой известно при выполнении гипотезы H_0. Именно эта случайная величина К, которую называют статистическим критерием или просто критерием служит для проверки справедливости нулевой гипотезы H_0.

После выбора определенного критерия К множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза принимается, а другое, при которых она отвергается.

Областью принятия гипотезы (областью допустимых значений критерия) называют совокупность значений критерия, при которых нулевую гипотезу принимают. Это такие значения критерия, которые характерны для известного при справедливости нулевой гипотезы распределения критерия *К,* т.е. возникающие с большой вероятностью.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают. Это такие значения критерия, которые не характерны для данного распределения, т.е. возникающие с малой вероятностью.

Критическими точками (границами упомянутых областей) $K_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Процедура проверки простой параметрической гипотезы выглядит так:

1. Формируют нулевую гипотезу H_0 и альтернативную гипотезу H_1 на основе выборочных данных.
2. Конструируют, исходя из логики задачи, случайную величину на основе результатов выборки, которую в данном разделе называют критерием; распределение критерия в случае истинности гипотезы H_0 известно.
3. Вся область возможных значений критерия разбивается на две подобласти (или два подмножества). Одно подмножество – это совокупность естественных (правдоподобных), т.е. наиболее вероятных для данного распределения значений. В это подмножество критерий попадает с высокой вероятностью γ. Эта вероятность содержится в условиях задачи. Она носит название «доверительная вероятность» (иначе «уровень доверия»). Обычно для γ задают следующие стандартные значения: $\gamma = 0.90; 0.95; 0.99$).

Другое подмножество – это область редко возникающих для данного закона распределения значений (неправдоподобных значений). Вероятность попадания критерия K в эту область мала и равна $\alpha = 1-\gamma$; α носит название «уровень значимости» ($\alpha = 0.10;0.05;0.01$). Критерий K принято обозначать через t.

4. Вычисляют значение критерия $K_{набл}$ на основе выборочных значений изучаемого признака. Если $K_{набл}$ попадает в область правдоподобных значений для данного закона распределения, то с вероятностью γ утверждают, что гипотеза H_0 не противоречит экспериментальным данным, а поэтому принимают основную гипотезу. Если значения $K_{набл}$ попадает в область неправдоподобных для данного закона распределения значений, то гипотезу H_0 отвергают и принимают альтернативную гипотезу H_1.

5. Если при проверке гипотезы H_0 эта нулевая гипотеза принимается, то данный факт не означает, что высказанное в нулевой гипотезе утверждение является единственно верным. Просто оно не противоречит имеющимся выборочным данным. Возможно, что и другое утверждение также не будет противоречить выборочным данным.

6. Если наблюдаемое значение критерия $K_{набл}$ попадает в область неестественных значений и мы, следовательно, отвергаем гипотезу H_0 и принимаем гипотезу H_1, то не можем ли мы при этом совершить ошибку - отвергнуть верную гипотезу H_0 и принять ложную гипотезу H_1? Да, можем, но вероятность этой ошибки мала. Уровень значимости α – это вероятность отклонить нулевую гипотезу, когда она верна (иначе $P(H_1/H_0) = \alpha$).

Вид альтернативной гипотезы (для исходной простой параметрической гипотезы $H_0: \theta = \theta_0$) может быть таким:

1. H_1: $\theta \neq \theta_0; \gamma + \alpha = 1$
2. H_1: $\theta < \theta_0$
3. H_1: $\theta > \theta_0$.

Проверка гипотезы о числовом значении математического ожидания *m* нормально распределенной величины

Постановка задачи:

H_0: $m = m_0$

H_1: $m \neq m_0$ …………………. (1);

$m < m_0$ *или* $m > m_0$ …………..(2).

Вводим критерий

$$K = t = \frac{(x - m_0)\sqrt{n}}{S}.$$

1) Если объем выборки $n \leq 30$, то при справедливости нулевой гипотезы случайная величина t имеет распределение Стьюдента с $k = n-1$ степенями свободы. Находим $t_{кр}$ (по значениям k и α) на основе таблицы «Критические точки распределения Стьюдента», причем для альтернативной гипотезы вида (1) используем двустороннюю критическую область, а для альтернативной гипотезы вида (2) используем одностороннюю критическую область;

2) Если n>30, то случайная величина t имеет стандартный нормальный закон распределения, поэтому находим $t_{кр}$ по таблице функции Лапласа на основе решения уравнения:

$\Phi_0(t_{кр}) = \frac{\gamma}{2}$ для случая (1) и

$\Phi_0(t_{кр}) = \gamma - 0{,}5$ для случая (2).

3) Для повышения точности вычислений переход от одного закона распределения (Стьюдента) к другому закону распределения (нормальному) следует делать не при n=30, а при n=120.

Проверка гипотезы о равенстве математических ожиданий (о равенстве генеральных средних) двух нормально распределенных генеральных совокупностей

Пусть имеются две нормально распределенные генеральные совокупности, причем в первой совокупности изучаемый признак $X \sim N(m_1; \sigma_1)$, во второй совокупности изучаемый признак $Y \sim N(m_2; \sigma_2)$.

Постановка задачи:

H_0: $m_1 = m_2$; (иначе: $\overline{X} = \overline{Y}$)

H_1: $m_1 \neq m_2$; (иначе: $\overline{X} \neq \overline{Y}$) или $m_1 > m_2$, $m_1 < m_2$.

Так как здесь рассматривается случай больших выборок n>30, то $\sigma_1^2 \approx S_1^2$; $\sigma_2^2 \approx S_2^2$.

Для решения задачи используется критерий $t = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$.

Проверка гипотезы о значимости выборочного коэффициента корреляции Пирсона

Рассматривается двумерная нормально распределенная генеральная совокупность (X, Y), т.е. случайные величины X и Y в ней распределены нормально. Из этой совокупности извлечена выборка объемом *n пар* (x_i, y_i) и по ней вычислен выборочный коэффициент корреляции Пирсона, который оказался отличным от нуля. Возникает вопрос, объясняется ли это действительно существующей линейной связью между случайными величинами X и Y в генеральной совокупности или является следствием случайности отбора переменных в выборку. Можно ли при этом заключить, что и коэффициент корреляции ρ между случайными величинами X и Y во всей генеральной совокупности также отличен от нуля?

Постановка задачи:

H_0: $\rho=0$

H_1: $\rho\neq0$.

Если нулевая гипотеза отвергается, то это означает, что коэффициент корреляции в генеральной совокупности значимо отличается от нуля (кратно говоря «значим»), и, следовательно, в генеральной совокупности признаки X и Y связаны линейной зависимостью. Если же принимается нулевая гипотеза, то генеральный коэффициент корреляции незначим, и, следовательно, признаки X и Y в генеральной совокупности не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы используется случайная величина

$$t = \frac{r \times \sqrt{n-2}}{1-r^2}.$$

Показано, что эта случайная величина при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Число степеней свободы на две единицы меньше объема выборки, поскольку в выражении для r задействованы две связи, заданные формулами для вычисления средних значений по выборке:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}; \ \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

При больших объемах выборки ($n>30$) можно вместо распределения Стьюдента использовать стандартный нормальный закон распределения.

Поскольку конкурирующая гипотеза имеет вид $\rho\neq0$, то следует строить двустороннюю критическую область.

Определив, куда попадает вычисленное значение $t_{набл}$, делаем вывод о справедливости нулевой или же альтернативной гипотезы: если $|t_{набл}| < t_{кр}$, то принимается гипотеза H_0, если $|t_{набл}| \geq t_{кр}$, то принимается гипотеза H_1.

Проверка статистических гипотез о дисперсиях

Проверка статистических гипотез о дисперсиях также представляет интерес для практики, так как дисперсии характеризуют разброс значений, а значит и такие величины как точность изготовления, измерения и т. д. Для проверки гипотез о равенстве дисперсий нужно знать такую функцию распределения, которая не зависела бы ни от каких неизвестных параметров выборки. Для этого используется распределением Фишера. По этому закону распределена случайная величина

$$F = \frac{D^* X}{D^* Y} = \frac{\sigma_x^{*2}}{\sigma_y^{*2}},$$

где случайные величины X и Y независимы и нормально распределены.

Если объем первой выборки равен n_1, а объем второй равен n_2, то распределение Фишера зависит от числа степеней свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$. По распределению Фишера можно построить критические области для проверки гипотез о равенстве дисперсий. Так как плотность распределения обладает асимметрией, то выберем критическую область так, чтобы вероятность попадания в правую и левую часть критической области были равны $\alpha/2$. Доверительная вероятность определяет два значения f_1 и f_2 из равенства

$$P(f_1 < F < f_2) = \beta.$$

Распределение Фишера $F=D^*X/D^*Y$ с k_1 и k_2 степенями свободы и распределение Фишера $F_1=D^*Y/D^*X$ с k_2 и k_1 степенями свободы фактически дают одно и тоже распределение. Поэтому при построении критической области указывают только верхнюю границу и в числителе указывают большую из дисперсий. Проверка гипотезы о равенстве дисперсий производится так: по заданному уровню значимости находят критическое значение f и вычисляют значение F. Если значение $F>f$, то гипотеза о равенстве дисперсий противоречит данным наблюдений и ее следует отклонить. Если нет, то гипотеза согласуется с данными наблюдений и ее следует принять.

3.3 Задачи для решения

1. Установка для напыления должна быть настроена на величину сопротивления напыляемых резисторов 20кОм. В результате измерений параметра пятнадцати резисторов среднее значение составило 13,5 кОм, стандартное отклонение равно 1,89 кОм^2. Определить правильность настройки. Сформулируйте постановку задачи для двух различных альтернативных гипотез. Представьте графическое решение задачи (для односторонней и двухсторонней критической области).

2. Установка для металлизации полипропиленовой пленки настроена на формирование токопроводящего слоя толщиной 20нм. При замере получились следующие значения: 15,2 нм, 18,3 нм, 21,2 нм, 25,2 нм, 23,4 нм, 16,5 нм, 19,5 нм, 20,7 нм, 21,5 нм, 22,5 нм. Определить правильность настройки. Сформулируйте постановку задачи для двух различных альтернативных гипотез. Представьте графическое решение задачи (для односторонней и двухсторонней критической области).

3. Напыление резисторов осуществляется на двух установках. С первой установки отобрано 10 резистор, среднее значение величины сопротивления которых составило 987,7 Ом, стандартное отклонение – 2587,12 Ом^2. На второй установке напыление нанесено на 16 резисторов. Среднее значение величины сопротивления - 1005,0 Ом, стандартное отклонение – 3605,73 Ом^2.

4. По выборке объема n=7, извлеченной из нормальной двумерной генеральной совокупности, был вычислен коэффициент корреляции Пирсона r=0,57. При уровне значимости α=10% проверить гипотезу H_0 о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе $\rho \neq 0$. Представьте графическое решение задачи.

5. По выборке объема n=112, извлеченных их нормальной двумерной генеральной совокупности, был вычислен коэффициент

корреляции Пирсона r=0,57. При уровне значимости α=10% проверить гипотезу H_0 о равенстве генерального коэффициента корреляции нулю (ρ=0) при конкурирующей гипотезе ρ≠0. Представьте графическое решение задачи.

6. Два однотипных станка обрабатывают одинаковые детали. Отобраны две пробы и подсчитаны дисперсии отклонения от заданных размеров: n_1=10; $\sigma^{*2}{}_1$=9,6мк^2 и n_2=15; $\sigma^{*2}{}_2$=5,7мк^2.

7. Определить, одинакова или различна точность замеров i-ой и j-ой партий резисторов (таблица 3.1).

Таблица 3.1 – Варианты заданий

Номер варианта	Номер партий из таблицы 3.2	
1	1	2
2	1	3
3	1	4
4	1	5
5	1	6
6	1	7
7	1	8
8	1	9
9	1	10
10	2	3

Таблица 3.2

Номер партии	Результаты измерений, Ом									
1	895	962	992	1109	1061	1045	1028	1057	1063	929
2	1044	998	987	995	983	1013	1025	1113	1069	983
3	1067	1052	976	985	988	987	1019	1142	1047	984
4	1044	962	992	985	983	1045	1025	1113	1063	929
5	1067	998	976	995	988	987	1025	1112	1065	984
6	1044	964	992	997	1061	1045	1018	1115	1065	929
7	991	1001	999	1112	1040	1115	1021	1034	970	999
8	972	993	987	983	1033	1039	1219	1031	1037	999
9	1067	962	987	985	983	1013	1028	1057	1037	984
10	967	988	981	985	1017	1026	1137	985	1006	999

3.4 КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Статистический критерий, область принятия гипотезы, критическая область, критические точки.
2. Процедура проверки простой параметрической гипотезы.
3. Процедура проверки гипотезы о числовом значении математического ожидания.
4. Процедура проверки гипотезы о равенстве математических ожиданий двух нормально распределенных генеральных совокупностей.
5. Процедура проверки гипотезы о значимости выборочного коэффициента корреляции Пирсона.
6. Процедура проверки статистических гипотез о дисперсиях.
7. Пример постановки задачи при проверке статистических гипотез.
8. Общие принципы и алгоритмы проверки статистических гипотез.
9. Ошибки при проверке статистических гипотез.
10. Критерии согласия и однородности: χ^2-Пирсона.
11. Критерии согласия и однородности: Колмогорова.
12. Критерии согласия и однородности: Шапиро – Уилка.
13. Критерии согласия и однородности: χ^2-Пирсона, Смирнова, Андерсона.
14. Области применения критериев.
15. Ошибка первого рода.
16. Ошибка второго рода.
17. Мощность критерия.
18. Критические значения критерия и критические области.
19. Классификация гипотез.
20. Интервальное и точечное оценивание.
21. Порядок проверки критериев.
22. Как и какие факторы нужно изменить, чтобы повысить мощность критерия?
23. Какой способ предпочтительнее при проверке гипотезы о распределении при малом объеме экспериментальных данных. Почему?

4 Коэффициенты корреляции и дисперсионный анализ

4.1 Цель работы

1. Целью работы является формирование навыков применения методов корреляционного и дисперсионного анализа для решения практических задач.

4.2 Основные теоретические положения

Нормированный коэффициент корреляции Браве-Пирсона

В качестве оценки генерального коэффициента корреляции используется коэффициент корреляции r Браве-Пирсона. Для его определения принимается предположение о двумерном нормальном распределении генеральной совокупности, из которой получены экспериментальные данные. Это предположение может быть проверено с помощью соответствующих критериев значимости. Следует отметить, что если по отдельности одномерные эмпирические распределения значений x_i и y_i согласуются с нормальным распределением, то из этого еще не следует, что двумерное распределение будет нормальным. Для такого заключения необходимо еще проверить предположение о линейности связи между случайными величинами X и Y. Строго говоря, для вычисления коэффициента корреляции достаточно только принять предположение о линейности связи между случайными величинами, и вычисленный коэффициент корреляции будет мерой этой линейной связи.

Коэффициент корреляции Браве–Пирсона относится к параметрическим коэффициентам и для практических расчетов вычисляется по формуле:

$$r_{xy}^{p} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}} .$$

Из формулы видно, что для вычисления необходимо найти средние значения признаков X и Y, а также отклонения каждого статистического данного от его среднего. Зная эти значения, находятся суммы. Затем, вычислив значение, необходимо определить достоверность найденного коэффициента корреляции, сравнив его фактическое значение с табличным для $f = n - 2$. Если $r_ф \geq r_{st}$, то можно говорить о том, что между признаками наблюдается достоверная взаимосвязь.

Коэффициент ранговой корреляции Спирмена

Если потребуется установить связь между двумя признаками, значения которых в генеральной совокупности распределены не по нормальному закону, т. е. предположение о том, что двумерная выборка получена из двумерной нормальной генеральной совокупности, не принимается, то можно воспользоваться коэффициентом ранговой корреляции Спирмена:

$$r_{xy}^S = 1 - \frac{6\sum(d_x - d_y)^2}{n(n^2 - 1)},$$

где d_x и d_y – ранги показателей x_i и y_i; n – число коррелируемых пар.

Коэффициент ранговой корреляции также имеет пределы 1 и –1. Если ранги одинаковы для всех значений x_i и y_i, то все разности рангов (d_x - d_y) = 0 и коэффициент ранговой корреляции = 1. Если ранги x_i и y_i расположены в обратном порядке, то коэффициент ранговой корреляции = -1. Таким образом, коэффициент ранговой корреляции является мерой совпадения рангов значений x_i и y_i.

Когда ранги всех значений x_i и y_i строго совпадают или расположены в обратном порядке, между случайными величинами X и Y существует функциональная зависимость, причем эта зависимость не обязательно линейная, как в случае с коэффициентом линейной корреляции Браве-Пирсона, а может быть любой монотонной зависимостью (т. е. постоянно возрастающей или

постоянно убывающей зависимостью). Если зависимость монотонно возрастающая, то ранги значений x_i и y_i совпадают и коэффициент = 1; если зависимость монотонно убывающая, то ранги обратны и коэффициент = –1. Следовательно, коэффициент ранговой корреляции является мерой любой монотонной зависимости между случайными величинами X и Y.
Из формулы видно, что для вычисления необходимо сначала проставить ранги показателей x_i и y_i, найти разности рангов для каждой пары показателей и квадраты этих разностей. Зная эти значения, находятся суммы. Затем, вычислив значение коэффициента, необходимо определить достоверность найденного коэффициента корреляции, сравнив его фактическое значение с табличным. Если $r_ф \geq r_{st}$, то можно говорить о том, что между признаками наблюдается достоверная взаимосвязь.

Коэффициент ранговой корреляции целесообразно использовать в следующих случаях:

- если экспериментальные данные представляют собой точно измеренные значения признаков X и Y и требуется быстро найти приближенную оценку коэффициента корреляции;
- когда значения x_i и y_i заданы в порядковой шкале, т. е. когда признаки не могут быть точно измерены, но их наблюдаемые значения могут быть расставлены в определенном порядке.

Дисперсионный анализ

Однофакторный дисперсионный анализ используется для сравнения средних значений для трех и более выборок. Фактором называется независимая переменная, влияние которой изучается на зависимую переменную.

Анализ основан на расчете F – статистики (статистика Фишера), которая представляет собой отношение двух дисперсий: межгрупповой и внутригрупповой. F – тест в однофакторном дисперсионном анализе устанавливает, значимо ли отличаются средние нескольких независимых выборок.

Процедура выполнения однофакторного дисперсионного анализа:

1) определение независимых и зависимых переменных;
2) разложение полной дисперсии;

3) измерение эффекта η^2;

4) проверка значимости;

5) представление результата.

Необходимым условием для проведения дисперсионного анализа является то, чтобы независимая переменная была категориальной, а зависимая – метрической.

Набор данных в однофакторном дисперсионном анализ состоит из k – независимых одномерных выборок, элементы которых измерены в одинаковых единицах. Допустимы различные объемы (размеры) выборок.

Таблица 4.1 - 1 этап. Подготовка данных для анализа

	Независимая переменная – фактор *(количество выборок k = 4)*			
	Выборка 1	Выборка 2	Выборка 3	Выборка k
Зависимая:	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{k,1}$
Зависимая:	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{k,2}$
Зависимая:	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{k,3}$
Зависимая:	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{k,4}$
Зависимая:	$X_{1,5}$	$X_{2,5}$		$X_{2,5}$
Зависимая:		$X_{2,6}$		$X_{k,6}$
Зависимая:		$X_{2,7}$		
Объем $n=n_1+n_2+n_3+\ldots+n_k$	$n_1 = 5$	$n_2 = 7$	$n_3 = 4$	$n_k = 6$
Среднее	X_1	X_2	X_3	X_k
Ст. отклонение	σ_1	σ_2	σ_3	σ_k

Нулевая гипотеза в однофакторном дисперсионном анализе утверждает, что все средние значения из различных генеральных совокупностей (которые представлены выборочными средними) равны между собой.

H_0 : $\mu_1 = \mu_k$ (или $X_1 = X_2 = \ldots = X_k$).

Альтернативная гипотеза утверждает, что хотя бы два любых средних не равны между собой.

H_1 : $\mu_1 \neq \mu_k$ (или $X_1 \neq X_k$).

F – тест состоит в расчете F – *статистики* и сравнении ее с табличным значением.

Поскольку нулевая гипотеза утверждает, что средние всех генеральных совокупностей равны, необходимо оценить это среднее значение по всем выборкам, т.е. рассчитать *общее среднее.*

Общее среднее представляет собой среднее всех значений из всех выборок.

Если размеры выборок не равны, то среднее рассчитывается как средневзвешенное с учетом размера выборок:

$$X = \frac{\sum_{i=1}^{k} n_i * X_i}{n}$$

2 этап. Для изучения различий между зависимыми переменными проводится разложение полной дисперсии:

$$SS = SS_{between} + SS_{within},$$

где $\boldsymbol{SS_{between}}$ - межгрупповая вариация и $\boldsymbol{SS_{within}}$ - внутригрупповая вариация.

Межгрупповая вариация ($SS_{between}$) показывает, насколько выборочные средние отличаются между собой. Она равна нулю, если средние равны и тем больше, чем сильнее различаются средние. Расчет межгрупповой дисперсии (вариации):

$$MS_b = \frac{SS_{between}}{k-1}$$

и средний квадрат:

$$SS_{between} = \sum_{i=1}^{k} n_i * (X_i - X)^2$$

k – количество выборок.

Внутригрупповая вариация (SS_{within}) показывает, насколько отличаются между собой значения по каждой выборке:

$$SS_{within} = \sum_{i=1}^{k} (n_i - 1) * \sigma_i^2$$

и средний квадрат:

$$MS_w = \frac{SS_{within}}{n-k}$$

3 этап. Эффект влияния независимой переменной на зависимую переменную рассчитывается через корреляционное отношение η^2 *(эта-квадрат)*, которое рассчитывается по формуле:

$$\eta^2 = \frac{SS_{between}}{SS}$$

Значение корреляционного отношения находится в пределах от 0 до 1. Оно равно 0, когда все выборочные средние равны, т.е. независимая переменная не влияет на зависимую, и, наоборот, влияние увеличивается с ростом этого значения. Другими словами, величина η^2 представляет собой меру вариации зависимой переменной, вызванную влиянием на нее независимой переменной.

4 этап фактически сводится к процедуре статистической проверки гипотезы о равенстве средних (наличии различий) путем расчета *F – статистики:*

$$F = \frac{MS_{between}}{MS_{within}}$$

5 этап. Для того, чтобы сделать окончательный вывод, необходимо обратиться к *F* – таблице, содержащей критические значения *F* - статистики при истинной нулевой гипотезе. Чтобы найти критическое значение, необходимо учесть количество степеней свободы (*df – degree freedom*) и соответствующий уровень проверки (по умолчанию 5%).

Степень свободы для межгрупповой вариации составляет «*k* – 1» (значение по столбцу), а для внутригрупповой вариации «*n – k*».

F – тест заключается в сравнении *F* – статистики, рассчитанной по имеющимся данным с критическим значением *F* – таблицы. Результат является значимым, если $F_{стат} > F_{критич}$, поскольку это говорит о наличии существенных различий между средними значениями по группам.

4.3 Задачи для решения

1. Имеются данные о квалификации и месячной выработке пяти рабочих цеха (таблица 4.2):

Таблица 4.2

Табельный номер	Разряд, *X*	Выработка продукции за смену, шт., *У*	X^2	$У^2$	*XУ*
1	6	130			
2	2	60			
3	3	70			
4	5	110			
5	4	90			
Итого					

Для изучения связи между квалификацией рабочих и их выработкой определить коэффициент корреляции.

2. 10 студентам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Исследователя интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная X — обозначает среднее время решения наглядно-образных, а переменная Y— среднее время решения вербальных заданий тестов. Критическое значение коэффициента корреляции равно 0,72.

3. Определить достоверность взаимосвязи между показателями износа оборудования и количества бракованных деталей десяти станков с помощью расчета рангового коэффициента корреляции, если данные выборок таковы:

x_i, срок службы в годах ~ 15; 5; 3; 7; 7; 11; 8; 20; 13;10

y_i, кол-во бракованных деталей ~ 26; 20; 25; 22; 27; 28; 16; 15; 18; 24.

Данные тестирования занести в рабочую таблицу и сделать необходимые расчеты (таблица 4.4). Табличное значение коэффициента равно 0,64.

Таблица 4.3

№ испытуемых	X	$У$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	19	17					
2	32	7					
3	33	17					
4	44	28					
5	28	27					
6	35	31					
7	39	20					
8	39	17					
№ испытуемых	X	$У$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
9	44	35					
10	44	43					
Сумма							
Среднее							

Таблица 4.4

x_i	d_x	y_i	d_y	$(d_x - d_y)$	$(d_x - d_y)^2$
15		26			
5		20			
3		25			
7		22			
7		27			
11		28			
8		16			
20		15			
13		18			
10		24			
				$\sum(d_x - d_y)$	$\sum(d_x - d_y)^2$

4. Требуется оценить влияние степени износа технологического оборудования на качество выпускаемой продукции (количество выпускаемых деталей в смену, соответствующих нормативной документации). Имеются следующие данные по 30 станкам (таблица 4.5):

Таблица 4.5

	Уровень износа технологического оборудования		
	Низкий	Средний	Высокий
	Кол-во деталей за смену		
1	10	8	5
2	9	8	7
3	10	7	6
4	8	9	4
5	9	6	5
6	8	4	2
7	9	5	3
8	7	5	2
9	7	6	1
10	6	4	2

Заполнить таблицу 4.6.

Таблица 4.6

Однофакторный дисперсионный анализ					
Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*	
Столбец 1					
Столбец 2					
Столбец 3					
…					
…					
Дисперсионный анализ					
Источник	*SS*	*df*	*MS*	***F***	*F-*

вариации					*критич.*
Между группами					
Внутри групп					
Итого					

5. Поставки продукции для вашей компании осуществляются тремя поставщиками («Мега+», «Коста» и «Трамп») в разное время: дневные часы, ночные смены и даже в пересменки. Естественно, с вашей стороны, контроль за качеством продукции в дневное время выше, чем в другое время. Вами собраны данные с оценками качества (в баллах), представленные в таблице 4.7.

Таблица 4.7

	Дневная смена	Ночная смена	Пересменка
«Мега+»	77,06	93,12	77,05
«Коста»	81,14	88,13	78,11
«Трамп»	82,02	81,18	79,91

Определить, есть ли отличие в качестве продукции, которая поставляется в разное время. Результаты занести в таблицу 4.6.

4.4 Контрольные вопросы

1. Сформулируйте задачи корреляционного анализа.
2. Перечислите свойства коэффициентов корреляции.
3. Дайте определение выборочной ковариации.
4. Дайте определение прямой и обратной корреляции.
5. Границы и область применения нормированного коэффициента корреляции Браве-Пирсона.
6. Границы и область применения коэффициента ранговой корреляции Спирмена.
7. Как осуществляется проверка достоверности коэффициента корреляции?
8. Сформулируйте задачи дисперсионного анализа.

9. Перечислите основные этапы однофакторного дисперсионного анализа.

10. Сформулируйте нулевую гипотезу в однофакторном дисперсионном анализе.

11. В чем заключается F – тест?

12. Как определить степень свободы для межгрупповой и внутригрупповой вариации?

13. Коэффициент корреляции. Свойства коэффициента корреляции.

14. Парная и частная корреляция.

15. Алгоритм вычисления коэффициентов корреляции.

16. Оценка значимости коэффициентов корреляции.

17. Какого типа практические задачи обычно решают методом дисперсионного анализа?

18. Как математически формулируется задача однофакторного дисперсионного анализа?

19. В чем заключается основная идея метода дисперсионного анализа?

20. Каким образом производится оценивание существенности влияния фактора в однофакторном дисперсионном анализе?

21. Как производится оценивание влияния двух факторов и их взаимодействий в двухфакторном дисперсионном анализе?

22. Однофакторный дисперсионный анализ для несвязанных выборок.

23. Дисперсионный анализ для связанных выборок.

24. Двухфакторный дисперсионный анализ.

Шкала оценивания и критерии оценивания

Шкала оценивания: **4-х** балльная.

Критерии оценивания:

4 балла (или оценка **«отлично»)** выставляется обучающемуся, если он самостоятельно выполнил практическую работу, оформил отчет в соответствии с предъявляемыми требованиями; полно ответил на все вопросы по практической работе.

3 балла (или оценка **«хорошо»)** выставляется обучающемуся, если он самостоятельно выполнил практическую работу, оформил отчет в соответствии с предъявляемыми требованиями; испытывает затруднения при ответе менее чем на 10% вопросов по практической работе.

2 балла (или оценка **«удовлетворительно»)** выставляется обучающемуся, если он самостоятельно выполнил практическую работу, оформил отчет в соответствии с предъявляемыми требованиями; испытывает затруднения при ответе на 20-30% вопросов по практической работе.

0-1 балл (или оценка **«неудовлетворительно»)** выставляется обучающемуся, если он не проявлял самостоятельности при выполнении практической работы; оформление отчета не соответствует предъявляемым требованиям (не соответствует ГОСТ, содержит не все пункты); испытывает затруднения при ответе более чем на50% вопросов по практической работе.

Список рекомендуемой литературы

1. Дьяконов В.П. MATLAB 6/6.1/6.5 + SIMULINK 4/5 в математике и моделировании. – М.: СОЛОН-Пресс, 2003 – 565 с.
2. Советов Б.Я., Яковлев С.А. Моделирование систем: Учеб. пособие. – М.: Высш. шк., 2001. – 344 с.
3. Круглов В. В., Дли М.И., Голунов Р. Ю. Нечёткая логика и искусственные нейронные сети: Учеб. Пособие. – М.: Издательство Физико-математической литературы, 2001. – 224 с.
4. Скляр Б. Цифровая связь. Теоретические основы и практическое применение. – М., С-Пб,К., изд.дом «Вильямс», 2003. – 1092 с.
5. С.В. Поршнев MATLAB 7. Основы работы и программирования. Учебник. М.: Бином. Лаборатория знаний, 2006. - 320 с.
6. Дьяконов В. П. Simulink 5/6/7: Самоучитель. – М.: ДМК_Пресс, 2008. – 784 с.