МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И
РАДИОЭЛЕКТРОНИКИ (ТУСУР)

Кафедра физической электроники (ФЭ)

Ю.В. Сахаров

УЧЕБНО-ИССЛЕДОВАТЕЛЬСКАЯ РАБОТА
СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Учебно-методическое пособие по практикам и
самостоятельной работе студентов

ТОМСК 2019

Рецензент: заведующий кафедрой ФГБОУ Томского государственного университета систем управления и радиоэлектроники, д.т.н., профессор Троян П.Е.

Сахаров, Юрий Владимирович

Учебно-исследовательская работа. Статистические методы обработки экспериментальных данных. Учебно-методическое пособие по практикам и самостоятельной работе студентов – Томск: Томск. гос. ун-т систем упр. и радиоэлектроники, 2019. – 38 с.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ 5

1.1 Программы статистического анализа 6

1.2 Методы статистического анализа данных 7

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №1 «Определение основных статистических показателей» 15

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №2 «Определение корреляции» 24

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №3 «Проверка гипотезы о нормальном распределении» 28

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №4 «Проверка гипотезы о равенстве дисперсий» 29

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №5 «Проверка гипотезы о равенстве средних» 31

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №6 «Регрессивный анализ» 37

Список использованных источников 38

ГЛАВА 1. СТАТИСТИЧЕКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

ПРЕДИСЛОВИЕ К ГЛАВЕ

Цель этой главы познакомить научных работников и выпускников технических специальностей с основными статистическими методами и научить применять эти методы. В главе изложены основные методы математической статистики и примеры их реализации с помощью программ статистического анализа. Для полного понимания материала необходимо хорошее знание прикладной математики и информатики. Главная задача, которую ставили перед собой авторы, состояла в том, чтобы дать четкое представление об основных статистических методах и способах их реализации с применением специализированных программ. Основной упор в главе делается на примеры статистической обработки данных, поступающих с различных технологических операций изготовления диэлектрических слоев при производстве ИМС. Основная задача главы научить читателя проводить предварительный анализ, сортировку и группировку данных, поступающих с технологических операций, осуществить выбор метода статистического анализа и реализовать его с помощью основных программ статистического анализа.

Полезную информацию может почерпнуть из книги и читатель, не знакомый с основами высшей математики. В главе отсутствует сложный математический аппарат, а изложены лишь основные методы математической статистики и способы их реализации с применением популярной электронной таблицы *Microsoft Excel*, входящей в пакет *Microsoft Office* [1, 2]. Чтобы читатели могли легко усвоить рассматриваемые статистические методы, в книгу включено большое число видеопримеров с комментариями по каждому методу статистического анализа. Для закрепления практических навыков включены исходные файлы примеров. Эти примеры приведены не для того, чтобы получился сборник готовых рецептов. Напротив, их назначение состоит в том, чтобы выделить и разъяснить тот или иной вопрос. Для закрепления приобретенных навыков и умений по окончании главы предусмотрена лабораторная работа.

ВВЕДЕНИЕ

Характерным для современного этапа развития естественных и технических наук является весьма широкое и плодотворное применение статистических методов во всех областях знания. Задача любой науки состоит в выявлении и исследовании закономерностей, которым подчиняются реальные процессы. Найденные закономерности имеют не только теоретическую ценность, они широко применяются на практике – в планировании, управлении и прогнозировании.

Математическая статистика – раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей. Математическая статистика по наблюденным значениям (выборке) оценивает вероятности событий либо осуществляет проверку предположений (гипотез) относительно этих вероятностей.

Изучение вероятностных моделей дает возможность понять различные свойства случайных явлений на абстрактном и обобщенном уровне, не прибегая к эксперименту. В математической статистике, наоборот, исследование связано с конкретными данными и идет от практики (наблюдения) к гипотезе и ее проверке.

При большом числе наблюдений случайные воздействия в значительной мере погашаются (нейтрализуются) и получаемый результат оказывается практически неслучайным, предсказуемым. Это утверждение (принцип) и является базой для практического использования вероятностных и математико-статистических методов исследования. Цель указанных методов состоит в том, чтобы, минуя сложное (а зачастую и невозможное) исследование отдельного случайного явления, изучить закономерности массовых случайных явлений, прогнозировать их характеристики, влиять на ход этих явлений, контролировать их, ограничивать область действия случайности.

Результаты эксперимента для инженера-исследователя были и остаются главным критерием при решении практических задач и при проверке теоретических гипотез. Однако при этом важно не только умело спланировать и поставить эксперимент, но и грамотно обработать его результаты. Этому вопросу часто не

уделяется должного внимания, и нередки случаи, когда результаты дорогостоящих экспериментов не подвергают даже простейшей обработке; при этом, как следствие, теряется огромное количество полезной информации.

Следует также подчеркнуть, что обработке экспериментальных данных с целью построения моделей «сложных систем» (эмпирических зависимостей) должна предшествовать предварительная обработка, содержание которой, в основном, состоит в отсеивании грубых погрешностей измерений и в проверке соответствия распределения результатов нормальному закону. Следует помнить, что только после выполнения предварительной обработки можно с наибольшей эффективностью, а главное корректно, использовать более сложные экспериментально-статистические методы, позволяющие получать математические модели даже таких процессов, строгое детерминированное описание которых вообще отсутствует.

1.1. ПРОГРАММЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

Обобщение и обработка экспериментальных данных представляет собой наиболее трудоемкий рутинный процесс, выполняемый в рамках статистических наблюдений, проводимых на различных уровнях, будь то предприятие, отрасль, регион или страна. Во всех случаях промежуточные и итоговые данные должны обеспечивать объективность и сопоставимость результатов наблюдений как внутри любого государства, так и на международном уровне. Программные продукты, разработанные для современной вычислительной техники, позволяют успешно решать эти задачи. Широкое внедрение компьютерной техники во все сферы деятельности человека создает благоприятные условия для автоматизации процессов обработки информации. Наиболее распространенным способом автоматизации является использование пакетов прикладных программ (ППП) общего и специального назначения на базе средств вычислительной техники. Статистические методы обработки данных включены в состав большинства электронных таблиц (таких, как ***Lotus 1-2-3, Quattro-Pro, Excel***), математических пакетов общего назначения (***MathCad, MatLab, Maple***), специализированных пакетов (***STATGRAPHICS, STATISTICA, SPSS, VORTEX***).

Из перечисленных программных продуктов наибольшее распространение получил табличный процессор ***Microsoft Excel*** [3, 4]. Это объясняется в первую очередь интеграцией табличного процессора в стандартный пакет ***Microsoft Office***. Подобный подход дает ряд преимуществ. Покупка пакета прикладных программ снижает себестоимость отдельно взятого программного продукта, входящего в него, а стандартный интерфейс позволяет значительно сократить время подготовки персонала. Это особо актуально для небольших фирм, не имеющих возможности приобрести специализированное программное обеспечение и провести переподготовку кадров. Еще одним немаловажным фактором является наличие в пакете прикладных программ значительного количества статистических функций (порядка 80), которые практически полностью удовлетворяют потребности большинства специалистов. Для решения специфических задач в ***Excel*** предусмотрена программная надстройка ***«Пакет анализа*** (гиперссылка1)***»***, реализованная еще в ***Excel 2003*** и с версиями практически не претерпевающая изменений. Наличие хорошо продуманной справочной системы и широкий выбор литературы способствуют популяризации данного программного продукта.

1.2. МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Объектом исследования в прикладной статистике являются статистические данные, полученные в результате наблюдений или экспериментов. *Статистические данные* – это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих.

Выборка – совокупность (массив) экспериментальных данных. В выборочном наблюдении используются понятия ***генеральная совокупность*** – изучаемая совокупность единиц, подлежащая изучению по интересующим исследователя признакам, и ***выборочная совокупность*** – случайно выбранная из генеральной совокупности **некоторая ее часть**. К данной выборке предъявляется требование репрезентативности, т.е. при изучении лишь части генеральной совокупности полученные выводы можно применять ко всей совокупности.

Характеристиками ***генеральной*** и ***выборочной*** совокупностей могут служить

средние значения изучаемых признаков, их дисперсии и средние квадратические отклонения, мода и медиана и др, которые будут рассмотрены ниже. Исследователя могут интересовать и распределение единиц по изучаемым признакам в генеральной и выборочной совокупностях. В этом случае частоты называются соответственно **генеральными** и **выборочными** [5, 6].

Система правил отбора и способов характеристики единиц изучаемой совокупности составляет содержание выборочного метода, суть которого состоит в получении первичных данных при наблюдении выборки с последующим обобщением, анализом и их распространением на всю генеральную совокупность с целью получения достоверной информации об исследуемом явлении.

Между признаками выборочной совокупности и признаками генеральной совокупности как правило, существует некоторое расхождение, которое называется ***ошибкой статистического наблюдения***. При массовом наблюдении ошибки неизбежны, но возникают они в результате действия различных причин. Величина возможной ошибки выборочного признака происходит из-за ошибок регистрации и ошибок репрезентативности.

Переменные – это величины, которые в результате измерения могут принимать различные значения. *Независимые переменные* – это переменные, значения которых в процессе эксперимента можно изменять, а *зависимые переменные* – это переменные, значения которых можно только измерять.

Переменные могут быть измерены в различных шкалах. Различие шкал определяется их информативностью.

Рассматривают следующие типы шкал, представленные в порядке возрастания их информативности: номинальная, *порядковая, интервальная, шкала отношений, абсолютная.* Эти шкалы отличаются друг от друга также и количеством допустимых математических действий. Самая «бедная» шкала – номинальная, так как не определена ни одна арифметическая операция, самая «богатая» – абсолютная.

Измерение в *номинальной (классификационной) шкале* означает определение принадлежности объекта (наблюдения) к тому или иному классу. Например: пол, род войск, профессия, континент и т.д. В этой шкале можно лишь посчитать

количество объектов в классах – частоту и относительную частоту.

Измерение в *порядковой (ранговой) шкале*, помимо определения класса принадлежности, позволяет упорядочить наблюдения, сравнив их между собой в каком-то отношении. Однако эта шкала не определяет дистанцию между классами, а только то, какое из двух наблюдений предпочтительнее. Поэтому порядковые экспериментальные данные, даже если они изображены цифрами, нельзя рассматривать как числа и выполнять над ними арифметические операции. В этой шкале дополнительно к подсчету частоты объекта можно вычислить ранг объекта. Примеры переменных, измеренных в порядковой шкале: оценки учащихся, призовые места на соревнованиях, воинские звания, место страны в списке по качеству жизни и т.д. Иногда номинальные и порядковые переменные называют категориальными, или группирующими, так как они позволяют произвести разделение объектов исследования на подгруппы.

При измерении в *интервальной шкале* упорядочивание наблюдений можно выполнить настолько точно, что известны расстояния между любыми двумя их них. Шкала интервалов единственна с точностью до линейных преобразований $(y=ax+b)$. Это означает, что шкала имеет произвольную точку отсчета – условный нуль. Примеры переменных, измеренных в интервальной шкале: температура, время, высота местности над уровнем моря. Над переменными в данной шкале можно выполнять операцию определения расстояния между наблюдениями. Расстояния являются полноправными числами и над ними можно выполнять любые арифметические операции.

Шкала отношений похожа на интервальную шкалу, но она единственна с точностью до преобразования вида $y=ax$. Это означает, что шкала имеет фиксированную точку отсчета – абсолютный нуль, но произвольный масштаб измерения. Примеры переменных, измеренных в шкале отношений: длина, вес, сила тока, количество денег, расходы общества на здравоохранение, образование, армию, средняя продолжительность жизни и т.д. Измерения в этой шкале – полноправные числа и над ними можно выполнять любые арифметические действия.

Абсолютная шкала имеет и абсолютный нуль, и абсолютную единицу

измерения (масштаб). Примером абсолютной шкалы является числовая прямая. Эта шкала безразмерна, поэтому измерения в ней могут быть использованы в качестве показателя степени или основания логарифма. Примеры измерений в абсолютной шкале: доля безработицы; доля безграмотных, индекс качества жизни и т.д.

Большинство статистических методов относятся к методам параметрической статистики, в основе которых лежит предположение, что случайный вектор переменных образует некоторое многомерное распределение, как правило, нормальное или преобразуется к нормальному распределению. Если это предположение не находит подтверждения, следует воспользоваться непараметрическими методами математической статистики.

В ходе обработки и анализа данных исследования **первым этапом** описание статистических показателей изучаемых признаков. Среди таковых основными можно отметить следующие показатели [7]:

Среднее (средняя арифметическая величина) - частное от деления суммы всех значений признака на их число. Оно определяется как сумма значений, деленное на их количество:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad (1.1)$$

где n – количество элементов в выборке.

Характеризует какую-либо совокупность в целом. Используется только для характеристики интервальных и порядковых шкал

Минимальное значение – это наименьшее значение переменной, встретившееся в массиве данных.

Максимальное значение – это наибольшее значение переменной, встретившееся в массиве данных.

Квадрат отклонений – это сумма квадратов отклонений величин от средней арифметической величины по выборке $\bar{x}$

$$s = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (1.2)$$

Дисперсия – величина, равная среднему значению квадрата отклонений s

отдельных значений признаков от средней арифметической величины $\bar{x}$. При этом различают **генеральную дисперсию** и **выборочную дисперсию**:

$$\sigma^2 = \frac{s}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1.3)$$

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, т.е. математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии. Для исправления выборочной дисперсии достаточно умножить ее на дробь:

$$\sigma_{\text{è}}^2 = \sigma\frac{n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1.4)$$

получим **исправленную дисперсию** $\sigma_{\text{и}}$. Исправленная дисперсия является несмещенной оценкой. **В качестве оценки генеральной дисперсии принимают исправленную дисперсию.**

При достаточно большой выборке $n \to \infty$ исправленная выборочная дисперсия стремится к генеральной $\sigma_{\text{и}}^2 \approx \sigma^2$.

Используется только для характеристики интервальных и порядковых шкал.

Среднее квадратическое отклонение (стандартное отклонение)– величина, равная квадратному корню из дисперсии σ^2. Это мера разброса измеренных величин. Также различается генеральное отклонение δ и выборочное (исправленное) $\delta_{\text{и}}$:

$$\delta_{\text{è}} = \sqrt{\sigma_{\text{è}}^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1.5)$$

Используется только для характеристики интервальных и порядковых шкал.

Среднее линейное отклонение – определяется как средняя арифметическая величина абсолютных значений отклонений отдельных вариантов от их средней арифметической величины $\bar{x}$.

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \qquad (1.6)$$

Медиана – значение переменной у той единицы совокупности, которая расположена в середине **ранжированного** ряда частотного распределения. Используется только для характеристики метрических шкал.

Мода – наиболее часто встречающееся значение переменной, т.е. значение, с которым наиболее вероятно можно встретиться в массиве.

Стандартная ошибка - есть величина, выражающая среднее квадратическое отклонение выборочной средней от математического ожидания.

$$m = \sqrt{\frac{\delta_{\text{è}}^2}{n}} \qquad (1.7)$$

Эта величина при соблюдении принципа случайного отбора зависит прежде всего от объема выборки n и от степени варьирования признака: чем больше n и чем меньше вариация признака (следовательно, и значение $\delta_{\text{и}}$), тем меньше величина средней ошибки выборки. Ошибки выборки свойственны только **выборочным наблюдениям**.

Частота – численное значение признака (количество значений попадающих в заданный интервал). Используется для всех видов шкал.

Асимметричность (коэффициент асимметрии) - асимметрия характеризует степень несимметричности распределения относительно его среднего. Положительная асимметрия указывает на отклонение распределения в сторону положительных значений. Отрицательная асимметрия указывает на отклонение распределения в сторону отрицательных значений.

Если показатель асимметрии больше нуля, то есть преобладают положительные отклонения от среднего, то наблюдается правосторонняя асимметрия, то есть преобладание значений в выборке превышающих среднее арифметическое. Если же показатель асимметрии меньше нуля, налицо левосторонняя асимметрия, то есть превышение численности значений меньше чем среднее арифметическое (рис.1.1).

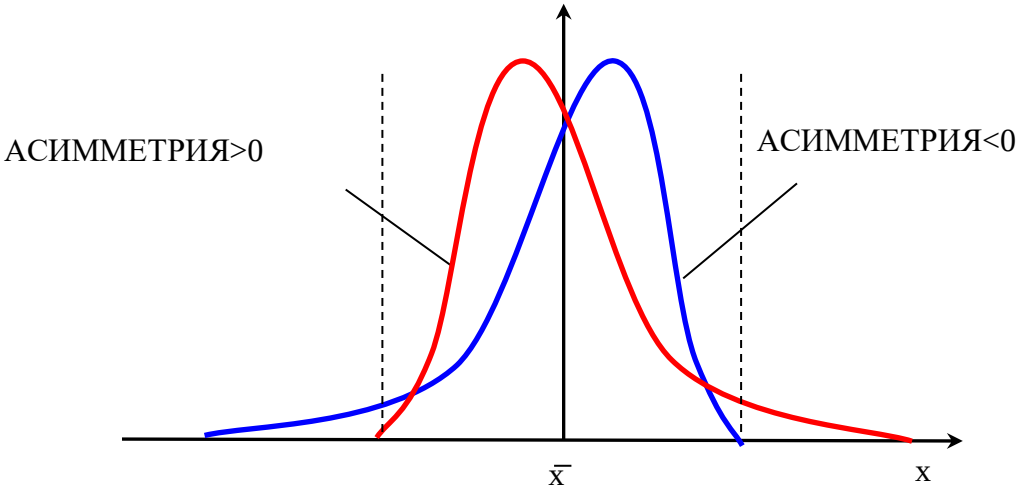

Рисунок 1.1 – Виды асимметрий

Коэффициент асимметрии определяется следующим образом:

$$\textbf{АСИММЕТРИЯ}=\frac{n}{(n-1)\cdot(n-2)}\cdot\sum_{i=1}^{n}\left(\frac{x_{\text{i}}-\bar{x}}{\delta_{\text{è}}}\right)^{3}, \qquad (1.8)$$

где $\delta_{и}$ – среднее квадратичное (стандартное) отклонение.

Эксцесс – эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение. Отрицательный эксцесс обозначает относительно сглаженное распределение (рис.1.2).

Показатель эксцесса характеризует степень колеблемости исходных данных, чем сильнее вариация, тем более пологой является кривая распределения и наоборот, чем однороднее совокупность, тем в большей степени варианты ряда сконцентрированы около средней и тем более островершинней будет кривая распределения. Эксцесс определяется следующим образом:

$$\textbf{ЭКСЦЕСС}=\left(\frac{n\cdot(n+1)}{(n-1)\cdot(n-2)\cdot(n-3)}\cdot\sum_{i=1}^{n}\left(\frac{x_{\text{i}}-\bar{x}}{\delta_{\text{è}}}\right)^{4}\right)-\frac{3\cdot(n-1)^{2}}{(n-2)\cdot(n-3)} \qquad (1.9)$$

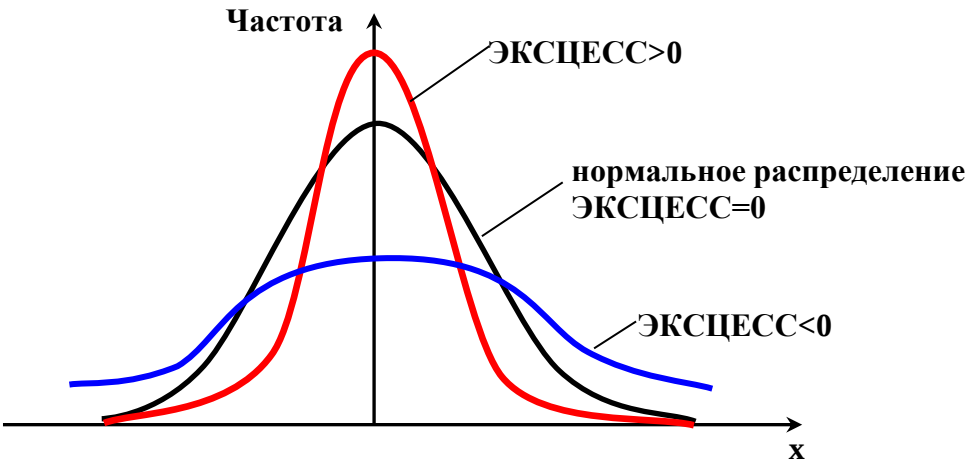

Рисунок 1.2 – Виды распределения и значения эксцесса

Коэффициент вариации – отношение среднего квадратического отклонения к среднему арифметическому.

$$Cv = \frac{\delta_{\grave{e}}}{\bar{x}} * 100\% \qquad (1.10)$$

При сильно асимметричных рядах распределения коэффициент вариации может достигать 100% и даже выше. Варьирование считается слабым, если Cv<10% средним, когда 10%<Cv<25%, и значительным при Cv>25%.

Используется только для характеристики метрических шкал.

Финалом перового этапа является **частотный анализ**.

Частотный анализ – заключается в оценке количества значений попадающих в заданный интервал, с целью наглядного представления вида распределения. Для частотного анализа необходимо построить таблицу частот, или как ее еще называют одновходовую таблицу, представляющую собой простейший метод анализа категориальных переменных. Таблицы частот могут быть с успехом использованы также для исследования количественных переменных, хотя при этом могут возникнуть трудности с интерпретацией результатов. Данный вид статистического исследования часто используют как одну из процедур разведочного анализа, чтобы посмотреть, каким образом различные группы наблюдений распределены в выборке, или как распределено значение признака на интервале от минимального до максимального значения. Как правило, таблицы частот графически иллюстрируются при помощи графиков и гистограмм.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №1 «Определение основных статистических показателей»

Задача: На предприятии, выпускающем ИМС, были отобраны по нескольку микросхем (4-6 штук) из разных партий. Всего было отобрано **97** микросхем. На отобранных микросхемах было определено напряжение пробоя подзатворного диэлектрика $U_{пр}$, *В* полевых транзисторов и данные сведены в таблицу 1.1. Исходя из допустимого диапазона пробивных напряжений $U_{пр}=60\pm15\%$ провести предварительный статистический анализ и оценить статистические показатели. На основе анализа сделать заключение.

Таблица 1.1.

№	1	2	3	4	5	6	….	97
$U_{пр}$, *В*	60,2	55,8	63,2	62,1	60,2	68,5	…..	61,2

Вторым этапом обработки и анализа данных исследования является описание связей между изучаемыми переменными. Если выборки являются **связанными**, т.е. зависящими друг от друга, то производят анализ **корреляционных связей**, если **не связанными**, то производят **проверку «нулевой» гипотезы** [8].

Корреляция – это связь между двумя (или более) переменными, при которой систематическое увеличение в значении одной переменной сопровождается систематическим увеличением или уменьшением в значении другой. Наличие такой статистической связи традиционно используется как основа для предположения относительно ожидаемого значения одной переменной при известном значении другой. Математической мерой корреляции двух случайных величин служит **коэффициент корреляции.** Существует несколько коэффициентов корреляции, указывающие на тесноту связи между исследуемыми переменными. Корреляционный анализ делится на параметрический и непараметрический. Наиболее известные параметрические методы анализа для связанных выборок: **корреляция Пирсона**, **ковариация.** Среди непараметрических наиболее часто используют коэффициент ранговой **корреляции Спирмена**, коэффициент ранговой **корреляции Кенделла**.

Корреляция Пирсона (далее называемая просто корреляцией) предполагает, что две рассматриваемые переменные измерены, по крайней мере, в интервальной шкале. Она определяет степень, с которой значения двух переменных "пропорциональны" друг другу, при этом **пропорциональность** означает просто **линейную зависимость**. Математическую меру пропорциональности отражает **коэффициент корреляции** или **линейный коэффициент корреляции** r_{xy}. **Линейный коэффициент корреляции** — параметр, который характеризует степень линейной взаимосвязи между двумя выборками, рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}, \qquad (1.11)$$

где x_i – значения, принимаемые в выборке x;

y_i – значения, принимаемые в выборке y;

$\overline{x}$ – среднее арифметическое по x;

$\overline{y}$ – среднее арифметическое по y;

n – количество элементов в выборке.

Коэффициент корреляции изменяется от -1 до 1. Знак коэффициента корреляции очень важен для интерпретации полученной связи. Если знак коэффициента линейной корреляции – плюс, то связь между коррелирующими признаками такова, что большей величине одного признака (переменной) соответствует большая величина другого признака (другой переменной). Иными словами, если один показатель (переменная) увеличивается, то соответственно увеличивается и другой показатель (переменная). Такая зависимость носит название **прямо пропорциональной зависимости**.

Если же получен знак минус, то большей величине одного признака соответствует меньшая величина другого. Иначе говоря, при наличии знака минус, увеличению одной переменной (признака, значения) соответствует уменьшение другой переменной. Такая зависимость носит название **обратно пропорциональной**

зависимости. Абсолютное значение коэффициента корреляции показывает меру линейной зависимости. Корреляция высокая, если на графике зависимость "можно представить" прямой линией (с положительным или отрицательным углом наклона). При значении 0 линейной зависимости между двумя выборками нет.

Важно, что значение коэффициента корреляции не зависит от масштаба измерения. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в дюймах и фунтах или в сантиметрах и килограммах.

Если возвести коэффициент корреляции в квадрат, то получим значение **коэффициента детерминации** $\boldsymbol{R^2}$, представляющего долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных).

Другим способом оценки пропорциональности между двумя переменными (выборками) является **ковариация**. Мерой пропорциональности служит коэффициент ковариации S_{xy}

Коэффициент ковариации также характеризует степень линейной зависимости двух случайных величин x и y , однако в отличии от коэффициента линейной корреляции Пирсона, он зависит от единиц измерения переменной (т.е. является абсолютной мерой корреляции). Он подсчитывается как среднее произведений отклонений каждой переменной:

$$S_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)}{n-1} \qquad (1.12)$$

Коэффициенты ковариации и корреляции связаны между собой:

$$r_{xy} = \frac{S_{xy}}{\sigma_x \cdot \sigma_y}, \qquad (1.13)$$

где σ_x и σ_y – исправленные значения выборочной дисперсии.

Рассмотренные выше методы корреляционного анализа являются обоснованным лишь в условиях **нормального или близкого к нормальному** распределению признаков в изучаемой совокупности. Как видно из формулы (1.11)

для определения линейного коэффициента корреляции необходимо знать значения факторного *x* и результативного *y* признаков.

В некоторых случаях можно встретиться с такими качествами, которые не поддаются выражению числом единиц. В этом случае прибегают к непараметрическим методам, позволяющим измерить интенсивность связи как между количественными признаками, форма распределения которых отличается от нормальной, так и между качественными признаками.

В основу непараметрических методов положен принцип нумерации значений статистического ряда. Каждой единице совокупности присваивается порядковый номер в ряду, который будет упорядочен по уровню признака. Таким образом, ряд значений ранжируется, а номер каждой отдельной единицы будет ее рангом.

Можно получить предварительное представление о наличии связи между признаками, если сопоставить последовательность взаимного расположения рангов факторного и результативного признаков. Для этого ранги индивидуальных значений факторного признака располагают в порядке возрастания, и если ранги результативного признака обнаруживают тенденцию к увеличению, можно предполагать наличие прямой связи. Если же с увеличением рангов факторного признака ранги результативного признака уменьшаются, то это говорит о возможном наличии между изучаемыми признаками обратной связи.

Коэффициент ранговой корреляции Спирмена является непараметрическим аналогом коэффициента корреляции Пирсона и основан на рассмотрении разности рангов значений факторного и результативного признаков. В этом случае определяется фактическая степень параллелизма между двумя количественными рядами изучаемых признаков и дается оценка тесноты установленной связи с помощью количественно выраженного коэффициента. Это тот же самый коэффициент корреляции Пирсона, только рассчитанный не для самих результатов измерений случайных величин, а для их ранговых значений. Только в отличие от коэффициента корреляции Пирсона, который может выявить только линейную зависимость одной переменной от другой, коэффициент корреляции

Спирмена может выявить монотонную зависимость, там, где непосредственная линейная связь не выявляется.

Предположим, что мы исследуем функцию *y=10/x*. У нас есть следующие результаты измерений *x* и *y* {{1,10}, {5,2}, {10,1}, {20,0.5}, {100,0.1}}

Для этих данных коэффициент корреляции Пирсона равен -0.4686, то есть связь слабая либо отсутствует. А вот коэффициент корреляции Спирмена строго равен -1, что говорит исследователю, что *y* имеет строгую отрицательную монотонную зависимость от *x*.

Практический расчет коэффициента ранговой корреляции Спирмена включает следующие этапы:

1) Сопоставить каждому из признаков их порядковый номер (ранг) по возрастанию (или убыванию).
2) Определить разности рангов каждой пары сопоставляемых значений.
3) Возвести в квадрат каждую разность и суммировать полученные результаты.
4) Вычислить коэффициент корреляции рангов по формуле:.

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad (1.14)$$

где $d_i = x_i - y_i$, - разности между рангами переменных *x* и *y*;

n – количество признаков.

Правило ранжирования:

1. Меньшему значению начисляется меньший ранг.

Наименьшему значению начисляется ранг 1.

Наибольшему значению начисляется ранг, соответствующий количеству ранжируемых значений. Например, если n=7, то наибольшее значение получит ранг 7, за возможным исключением для тех случаев, которые предусмотрены правилом 2.

2. В случае, если несколько значений равны, им начисляется ранг, представляющий собой среднее значение из тех рангов, которые они получили

бы, если бы не были равны. Ранги для одинаковых значений переменных называются **связанными рангами**.

Например, производятся временные исследования, в результате 3 наименьших значения равны 10 секундам (таблица 1.2).

Если бы мы измеряли время более точно, то эти значения могли бы различаться и составили бы, скажем, 10.2 сек; 10.5 сек; 10.7 сек. В этом случае они получили бы ранги, соответственно, 1, 2 и 3. Но поскольку полученные нами значения равны, каждое из них получает средний ранг:

$$\frac{1+2+3}{3}=2$$

Далее, допустим, идет время 11 с, соответственно оно получит ранг 3. Далее Допустим, следующие два (5 и 6 значения времени) значения времени равны 12 сек. Они должны были бы получить ранги 4 и 5, но, поскольку они равны, то получают средний ранг

$$\frac{5+6}{2}=5{,}5$$

Пусть дальше идут времена 13 с и 14 с, они получат ранги 7 и 8 соответственно. Затем снова предположим, что идут три подряд одинаковых значения времени 15 с. Соответственно их ранг будет

$$\frac{9+10+11}{3}=10$$

Таблица 1.2. Пример ранжирования:

Время, с	**10**	**10**	**10**	11	**12**	**12**	13	14	**15**	**15**	**15**
Ранг, **d**	**2**	**2**	2	4	**5,5**	**5,5**	7	8	**10**	**10**	**10**

Коэффициент корреляции Спирмена может также принимать значения от -1 до +1. При этом отрицательный коэффициент корреляции Спирмена позволяет принять гипотезу о наличии монотонной отрицательной связи, т.е. увеличение значения одной переменной ведет к уменьшению значения коррелирующей с ней переменной. Положительный коэффициент корреляции свидетельствует о

положительной связи между переменными: увеличение одной переменной соответствует увеличению другой.

По абсолютным значениям коэффициента ранговой корреляции Спирмена условно оценивают тесноту связи между признаками, считая значения коэффициента равные 0,3 и менее, показателями слабой тесноты связи; значения более 0,4, но менее 0,7 - показателями умеренной тесноты связи, а значения 0,7 и более - показателями высокой тесноты связи.

Для нахождения **уровня значимости** корреляции обращаемся к таблице «Критические значения коэффициента корреляции рангов Спирмена,» в которой приведены критические значения для коэффициентов ранговой корреляции таблица (таблица 1.3). Коэффициенты в таблице приводятся для разных вероятностей: 0,05 – для вероятности 95% и 0,01 – для вероятности 99%, коэффициенты могут быть рассчитаны и для других вероятностей с помощью стандартных функций, заложенных в некоторых программных продуктах, к примеру, *STATISTICA.*

Таблица 1.3. Критические значения коэффициента корреляции рангов Спирмена

n	$\rho_{кр}$		*n*	$\rho_{кр}$		*n*	$\rho_{кр}$	
	0,05	**0,01**		**0,05**	**0,01**		**0,05**	**0,01**
5	0,94	-	**17**	0,48	0,62	**29**	0,37	0,48
6	0,85	-	**18**	0,47	0,60	**30**	0,36	0,47
7	0,78	0,94	**19**	0,46	0,58	**31**	0,36	0,46
8	0,72	0,88	**20**	0,45	0,57	**32**	0,36	0,45
9	0,68	0,83	**21**	0,44	0,56	**33**	0,34	0,45
10	0,64	0,79	**22**	0,43	0,54	**34**	0,34	0,44
11	0,61	0,76	**23**	0,42	0,53	**35**	0,33	0,43
12	0,58	0,73	**24**	0,41	0,52	**36**	0,33	0,43
13	0,56	0,70	**25**	0,49	0,51	**37**	0,33	0,43
14	0,54	0,68	**26**	0,39	0,50	**38**	0.32	0,41
15	0,52	0,66	**27**	0,38	0,49	**39**	0,32	0,41
16	0,50	0,64	**28**	0,38	0,48	**40**	0,31	0,40

Если для данной выборки *n*, рассчитанное значение ρ оказывается больше критического $\rho_{кр}(0{,}01)$, то отвергается «нулевая гипотеза» (об отсутствии корреляции между рядами), т.е. корреляция между переменными *x* и *y* значима. Если рассчитанное значение ρ оказывается меньше $\rho_{кр}(0{,}05)$, то «нулевая гипотеза» подтверждается, т.е. корреляция между переменными *x* и *y* отсутствует.

При наличии одинаковых рангов формула расчета коэффициента линейной корреляции Спирмена (1.14) будет несколько иной. В этом случае в формулу вычисления коэффициентов корреляции добавляются два новых члена, учитывающие одинаковые ранги. Они называются поправками на одинаковые ранги и добавляются в числитель расчетной формулы.

$$D1 = \frac{k1^3 - k1}{12},$$

$$D2 = \frac{k2^3 - k2}{12},$$

где *k1* – число одинаковых рангов в первом столбце;

k2 - число одинаковых рангов во втором столбце.

Если имеется две группы одинаковых рангов, в каком-либо столбце то формула поправки несколько усложняется:

$$D3 = \frac{(k3^3 - k3) + (k4^3 - k4)}{12},$$

где *k3* – число одинаковых рангов в первой группе ранжируемого столбца;

k4 - число одинаковых рангов во второй группе ранжируемого столбца

В общем виде модификация формулы такова:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2 + D1 + D2 + D3}{n(n^2 - 1)} \qquad \textbf{(1.15)}$$

Коэффициент ранговой корреляции Кенделла также является мерой связи между переменными *X* и *Y*. Формула коэффициента ранговой корреляции Спирмена имеет вид:

$$\tau = \frac{2S}{n(n-1)} \qquad (1.15)$$

где $S = P + Q$

Для вычисления τ нужно упорядочить ряд рангов переменной *x*, приведя его к ряду натуральных чисел. Затем рассматривают последовательность рангов переменной *y*.

Первое слагаемое P – это мера соответствия последовательности рангов переменной *y* последовательности рангов переменной *x* При определении слагаемого **P** надо установить, сколько числе, находящихся справа от каждого из элементов последовательности рангов переменной *y*, имеет величину ранга, превышающую ранг рассматриваемого элемента.

Второе слагаемое Q – это мера несоответствия последовательности рангов переменной *y* последовательности рангов переменной *x*. Для определения Q подсчитывают, сколько чисел, находящихся справа от каждого из членов последовательности рангов переменной *y* имею ранг меньше, чем эта единица. Такие величины берутся со знаком минус.

При достаточно большом числе наблюдений между коэффициентами корреляции Спирмена и Кенделла существует соотношение $\rho = \frac{3}{2}\tau$.

Корреляции Спирмена и Кенделла относятся к ранговым корреляциям. Коэффициенты ранговой корреляции весьма близки к соответствующим значениям коэффициентов Пирсона. Эти корреляции отсутствуют в пакете анализа Excel, поэтому могут быть вычислены только с помощью стандартных математических функций.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №2 «Определение корреляции»

Задача №2. На предприятии выпускающем ИМС были проведены исследования по влиянию температуры отжига ИМС на тангенс угла диэлектрических потерь подзатворного диэлектрика для трех ИМС (соответственно *tg δ1, tg δ2, tg δ3*). Данные занесены в таблицу 1.4. Задача:

1) Определить взаимосвязь между температурой отжига и тангенсом угла потерь для ИМС.
2) Определить корреляцию между тангенсами угла потерь подзатворного диэлектрика в выбранных ИМС в процессе отжига.
3) Сделать вывод.

Таблица 1.4.

T, °C	100	110	120	130	140	150	…	250
tg δ1	0,005	0,0045	0,0042	0,004			…	0,0001
tg δ2	0,008	0,0072	0,0065					
tg δ3	0,002							

В случае если анализируются **несвязанные выборки** (то есть между ними отсутствует корреляция), то проводят **проверку гипотез**. Цель этого исследования заключается в том, чтобы подтвердить или опровергнуть **статистическую гипотезу** о влиянии (отсутствии влияния) на одну из выборок одного или нескольких факторов, по сравнению с соседней выборкой. К примеру это могут быть пациенты в больнице, часть из которых лечили стандартным препаратом **А**, а часть экспериментальным препаратом **Б**, задача состоит в том чтобы подтвердить или опровергнуть эффективность лечения препаратом **Б**, используя некоторые критерии эффективности. Другим примером, может быть определение более точного измерительного прибора из двух имеющихся, путем измерения величин прибором **А** и прибором **Б**. Применительно к нашей специфике это может быть, к примеру, оценка эффективности влияния (отсутствии влияния) температурного отжига на диэлектрические свойства диэлектриков или оценка влияния (не влияния) состава атмосферы при отжиге и т.д.

Статистической гипотезой называется любое предположение относительно вида или параметров распределения генеральной совокупности. Чаще всего исследуются гипотезы о предполагаемом законе распределения выборочной совокупности, об ожидаемых значениях параметров известного распределения, о принадлежности нескольких выборочных совокупностей одной и той же генеральной совокупности и т.п.

Гипотезу, утверждающую, что различие между сравниваемыми характеристиками отсутствует, а наблюдаемые отклонения объясняются лишь случайными колебаниями в выборках, на основании которых производится сравнение, называют **нулевой (основной) гипотезой** и обозначают H_0. Наряду с основной гипотезой рассматривают и **альтернативную** (конкурирующую, противоречащую) ей гипотезу H_1,. И если нулевая гипотеза будет отвергнута, то нет оснований отвергать альтернативную.

Задача проверки статистической гипотезы заключается в принятии одного из двух взаимоисключающих решений: отклонения или неотклонения выдвинутой гипотезы. Любое правило, позволяющее однозначно принять решение, называется **критерием**.

Проверка статистической гипотезы осуществляется с помощью статистического критерия в соответствии со следующим алгоритмом:

- сформулировать две взаимоисключающие гипотезы;
- установить или постулировать закон распределения;
- вычислить тестовую статистику;

Из тестовых статистик наиболее известными являются: ***t*-статистика Стьюдента** (проверка гипотезы о равенстве средних), **статистика Фишера** (проверка гипотезы о равенстве дисперсий), **статистика хи-квадрат** (проверка соответствия результатов измерений установленным допускам).

Проверка гипотезы о законе распределения. Распределение непрерывной случайной величины x называют нормальным, если соответствующая ей плотность распределения выражается формулой:

$$y(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\delta}\right)^2}, \qquad (1.16)$$

где δ - среднеквадратичное отклонение;

$x - \bar{x}$ – отклонение значения x от среднего.

Для проверки гипотезы о соответствии, экспериментального закона распределения случайной величины нормальному применяют **критерий Пирсона** или, как его иначе называют, критерий χ^2 (хи-квадрат), так как принятие и отклонение гипотезы основаны на χ^2-распределении.

Использование критерия Пирсона основано на сравнении эмпирических n_i (наблюдаемых) и теоретических n_{Ti} (вычисленных в предположении нормального распределения) частот.

В критерии согласия Пирсона в качестве показателя, по которому судят о соответствии фактического распределения предполагаемому теоретическому, берется случайная величина, значение которой рассчитывается по формуле:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_{Ti})^2}{n_{Ti}}, \qquad (1.17)$$

где k – число интервалов ряда частоты которых сравнивают.

Установлено, что поведение этой случайной величины при $n \to \infty$ подчиняется χ^2 распределению, критические значения которого в зависимости от числа степеней свободы r при уровне значимости α =0,05 (т.е. с 5% - ной вероятностью ошибки) принимают значения, представленные в табл. 1.5.

Таблица 1.5. Критические значения согласия Пирсона при α=0,05

Число степеней свободы r	1	2	3	4	5	6	7	8	9	10	11
Критические значения χ_q^2	3,8	6,0	7,8	9,5	11,6	12,6	14,1	15,5	16,9	18,3	19,7

Гипотеза не отвергается, если соблюдается условие $\chi^2 < \chi_q^2$ Значение χ_q^2

принимают по числу степеней свободы $r = m\text{-}c\text{-}1$ (где m - число интервалов ряда, с число параметров предполагаемого распределения, например для нормального распределения $c = 2$).

Порядок расчета критерия Пирсона:

1) проводят ЧАСТОТНЫЙ анализ, выбирая количество интервалов $m \approx \sqrt{n}$, получая информацию о фактическом распределении частот n_i. При этом рекомендуется выбирать не менее 10 интервалов.
2) производится нормировка полученных распределений частот, переходя тем самым к безразмерным значениям:

$$t_i = \frac{z_i - \bar{x}}{\delta_{\text{è}}},$$

где $\delta_{\text{и}}$ - стандартное отклонение в выборке.

$\bar{x}$ - среднее значение в выборке.

z_i – значение интервала.

3) определяются интегральные вероятности, т.е. вероятности значений $f(t_i)$, попадающих в отдельный интервал **при нормальном распределении**. Они рассчитываются как разница значений вероятностей верхней и нижней границы интервала.
4) Вычисляются теоретические частоты n_{Ti}

$$n_{Ti} = \frac{f(t_i)}{\sum_{i=1}^{k} t_i} \cdot n,$$

где k – число значений интегральной вероятности (на одно меньше чем количество интервалов);

n – количество элементов в выборке.

При этом сумма теоретических и фактических частот должна совпасть.

5) по формуле (1.17) рассчитывается χ^2 и сравнивается с χ_q^2 для количества степеней свободы $r=m\text{-}3$.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №3 «Проверка гипотезы о нормальном распределении»

Задача: На предприятии, выпускающем ИМС, были отобраны по нескольку микросхем (4-6 штук) из разных партий. Всего было отобрано **100** микросхем. На отобранных микросхемах было определено напряжение пробоя подзатворного диэлектрика $U_{пр}$, *В* полевых транзисторов и данные сведены в таблицу 1.6. **Проверить гипотезу о нормальном распределении в выборке**.

Таблица 1.6.

№	1	2	3	4	5	6	….	97
$U_{пр}$, *В*	60,2	55,8	63,2	62,1	60,2	68,5	…..	61,2

Статистика Фишера (проверка гипотезы о равенстве дисперсий) - это статистический метод обработки данных, основанный на отношении дисперсий двух выборок, разработанный Р. Фишером.

Точность измерений, степень изменчивости экономических показателей, в том числе и показателей качества, оцениваются по степени рассеивания отдельных результатов относительно их средних значений. Мерой рассеивания является дисперсия. Сравнивая между собой дисперсии, можно решить задачу оценки однородности результатов измерений, процессов или явлений. Например, если даны результаты тестирования для частных и общественных школ, то можно определить, имеют ли эти школы различные уровни разнородности учащихся по результатам тестирования.

Таким образом, если необходимо решить вопрос о принадлежности двух выборок одной генеральной совокупности, проверяют гипотезу о равенстве дисперсий. Это выполняется с помощью **критерия Фишера** (***F-критерия***), который формулируется следующим образом: если значение $F_{\delta} = \frac{\sigma_x^2}{\sigma_y^2}$ превышает F_q , то расхождение между двумя дисперсиями считается значимым, т.е. гипотеза H_0 (гипотеза о сходстве) может быть отвергнута на уровне заданной вероятности (обычно 1 или 5 %), а принимается гипотеза H_1.

При расчете в числителе должна находиться большая из двух независимо определенных выборочных дисперсий, σ_x^2 и σ_y^2. Значение F_q получают исходя из уровня значимости (обычно уровень значимости принимают равным 0,05, что соответствует 5% погрешности) и степеней свободы $r_1 = n_1-1$ *и* $r_2 = n_2-1$, где n_1 и n_2 объемы первой (большей) и второй (меньшей) выборки.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №4 «Проверка гипотезы о равенстве дисперсий»

При исследовании емкости МДП-конденсаторов были проведены замеры емкости эталонным прибором (прибором1) и прибором без класса точности (прибор2) и данные сведены в таблицу 1.7. Проверить гипотезу о равенстве дисперсий (т.е., что приборы меряют одинаково).

Таблица 1.7.

№	1	2	3	….	….33
Прибор1 Емкость, С, пФ	30	32	31	….	…
Прибор2 Емкость, С, пФ	25	29	23	….	…

***t*-статистика Стьюдента** (проверка гипотезы о равенстве средних) - это отношение стандартной ошибки оценки коэффициента к его абсолютной величине. Для проверки гипотезы вычисляют *t-критерий Стьюдента*, который позволяет найти вероятность того, что оба средних значения в выборке относятся к одной и той же совокупности. Данный критерий наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности».

На практике часто возникает необходимость сравнить два различных технологических процесса или два разных способа обработки (измерения, изготовления). В этом случае для установления сходства или различия методов используются средние значения показателей. Однако следует различать случаи **зависимых** и **независимых** выборок. К примеру, если определенную партию

болтов измерили двумя различными микромерами, имеют место **зависимые (связанные)** выборки, так как диаметр каждого из болтов измерялся и первым и вторым прибором, а следовательно, значения попарно взаимосвязаны. И напротив, если имело место сравнение двух различных марок стали, говорят о **независимых** выборках (протекание технологического процесса изготовления стали в каждом случае уникально).

При использовании критерия можно выделить два случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух **независимых, несвязанных выборок** (так называемый **двухвыборочный *t* - *критерий***). В этом случае есть контрольная группа и экспериментальная (опытная) группа, количество испытуемых в группах может быть различно.

Во втором случае, когда **одна и та же группа** объектов порождает числовой материал для проверки гипотез о средних, используется так называемый ***парный t - критерий***. Выборки при этом называют **зависимыми, связанными**. В этом случае количество значений в выборках должно быть одинаково.

Случай независимых выборок (двухвыборочный *t* - *критерий*)

Статистика критерия для случая несвязанных, независимых выборок равна:

$$t_{\text{ô}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}}}, \qquad (1.18)$$

где $\bar{x}$, $\bar{y}$ - средние арифметические в выборке *x* и *y* соответственно

σ_1 и σ_2 – исправленная выборочная дисперсия первой и второй выборки.

n1 и *n2* – количество элементов в первой и второй выборках соответственно.

Подсчет числа степеней свободы осуществляется по формуле:

$$k = n1 + n2 - 2 \qquad (1.21)$$

Далее необходимо сравнить полученное значение t_s с критическим значением *t* - *распределения* Стьюдента t_q., которое является табличной величиной (может также определено с помощью функций заложенных в статистические пакеты) в зависимости от степени свободы и вероятности ошибки (обычно берется 0,05 или 0,01, т.е. 5% и 1% вероятности ошибки). Если$|t_{\text{ф}}| < t_q$, то гипотеза H_0 принимается, в

противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза H_1 .

Случай связанных выборок (парный *t - критерий*)

В случае связанных выборок с равным числом измерений в каждой можно использовать более простую формулу *t-критерия* Стьюдента.

Вычисление значения *t -критерия* осуществляется по формуле:

$$t_{ô} = \frac{\overline{d}}{Sd}, \qquad (1.22)$$

где d – среднее арифметическое разностей $d_i = x_i - y_i$

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \qquad (1.23)$$

$$Sd = \sqrt{\frac{1}{1-n}\sum_{i=1}^{n}(d_i - \overline{d})^2} \qquad (1.24)$$

Число степеней свободы *k* определяется по формуле $k=n-1$.

Если $|t_ф| < t_q$ то гипотеза H_0 принимается, в противном случае нулевая гипотеза отвергается и принимается альтернативная гипотеза H_1 . При этом критическое значением *t - распределения* Стьюдента t_q., определяется с помощью таблиц (или функций заложенных в статистические пакеты) в зависимости от степени свободы и вероятности ошибки (обычно берется 0,05 или 0,01, т.е. 5% и 1% вероятности ошибки).

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №5 «Проверка гипотезы о равенстве средних»

На предприятии выпускающем ИМС есть две установки термического окисления (установка 1 и установка 2). Для исследования были отобраны две партии по 50 ИМС и исследованы напряжение пробоя МДП-конденсаторов Uпр и данные занесены в таблицу 1.8. При этом одна партия ИМС (50 шт.) изготавливалась с использованием установки 1, а вторая (50 шт) – с использованием установки 2. Определить есть ли сходства между работой этих двух установок.

Таблица 1.8.

	№	1	2	3	4	….	50
Установка1	Uпр, В	45	42	45	52		54
Установка2	Uпр, В	48	41	47	45		54

Финальным этапом статистического анализа является **регрессивный анализ**.

Регрессионный анализ – статистический метод установления зависимости между независимыми и зависимыми переменными. Регрессионный анализ на основе построенного уравнения регрессии определяет вклад каждой независимой переменной в изменение изучаемой (прогнозируемой) зависимой переменной величины. В регрессионном анализе моделируется взаимосвязь одной случайной переменной от одной или нескольких других случайных переменных. При этом, первая переменная называется зависимой, а остальные – независимыми. Выбор или назначение зависимой и независимых переменных является произвольным (условным) и осуществляется исследователем в зависимости от решаемой им задачи. Независимые переменные называются факторами, регрессорами **или предикторами**, а зависимая переменная – **результативным признаком, или откликом**.

Если число предикторов равно 1, регрессию называют простой, или однофакторной, если число предикторов больше 1 – множественной или многофакторной. В общем случае регрессионную модель можно записать следующим образом:

$$y = f(x_1, x_2, x_3, ..., x_n),$$

где y – зависимая переменная (отклик), x_i ($i = 1, \ldots, n$) – предикторы (факторы), n – число предикторов.

Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является **среднеквадратичная ошибка**: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента. Для проверки значимости

отдельных коэффициентов регрессии, т.е. гипотезы H_0 используют t-критерий Стьюдента. Значимость всего уравнения регрессии, т.е. гипотеза H_0 определяется по F-критерию (критерию Фишера).

Разности между фактическими значениями зависимой переменной и восстановленными называются **регрессионными остатками** (residuals). В литературе используются также синонимы: **невязки** и **ошибки.** Одной из важных оценок критерия качества полученной зависимости является сумма квадратов остатков ***SSE (Sum of Squared Errors)***. Еще один критерий – дисперсия остатков MSE $\sigma^2=SSE/(N-2)$.

Линейная регрессия, предполагает, что функция y линейно зависит от x. при этом зависимость от свободной переменной b не обязательна: $y=kx+b$.

Посредством регрессионного анализа можно решать ряд важных для исследуемой проблемы задач:

1) Уменьшение размерности пространства анализируемых переменных (факторного пространства), за счет замены части факторов одной переменной – откликом. Более полно такая задача решается факторным анализом.

2) Количественное измерение эффекта каждого фактора, т.е. множественная регрессия, позволяет исследователю задать вопрос (и, вероятно, получить ответ) о том, «что является лучшим предиктором для...». При этом, становится более ясным воздействие отдельных факторов на отклик, и исследователь лучше понимает природу изучаемого явления.

3) Вычисление прогнозных значений отклика при определенных значениях факторов, т.е. регрессионный анализ, создает базу для вычислительного эксперимента с целью получения ответов на вопросы типа «Что будет, если… ».

4) В регрессионном анализе в более явной форме выступает причинно-следственный механизм. Прогноз при этом лучше поддается содержательной интерпретации.

5) Регрессивный анализ позволяет моделировать эмпирические формулы путем установления взаимосвязи между переменными.

Формулы используемые при регрессивном анализе достаточно сложны и

рассматриваться в этом разделе не будут. Подробно остановимся на представлении данных при регрессивном анализе в MS Excel:

R*-коэффициент множественной корреляции R** - выражает степень зависимости независимых переменных *x* и зависимой переменной *y*. **Множественный *R равен квадратному корню из коэффициента детерминации, эта величина принимает значения в интервале от нуля до единицы. В простом линейном регрессионном анализе множественный *R* равен коэффициенту корреляции Пирсона.

***R*-квадрат (коэффициент детерминации)**, называемый также мерой определенности, характеризует качество полученной регрессионной прямой. Это качество выражается степенью соответствия между исходными данными и регрессионной моделью (расчетными данными). Мера определенности всегда находится в пределах интервала [0;1].

В большинстве случаев значение *R*-квадрат находится между этими значениями, называемыми экстремальными, т.е. между нулем и единицей, Если значение R-квадрата близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных, И наоборот, значение *R*-квадрата, близкое к нулю, означает плохое качество построенной модели.

Нормированный *R*-квадрат - коэффициент детерминации, скорректированный на величину выборки. Применяется как альтернатива коэффициенту детерминации при малой выборке.

Стандартная ошибка (отклонение результата)– это оценка стандартного отклонения распределения коэффициента регрессии вокруг его истинного значения. Определяется как:

$$\delta_{\text{ðåã}} = \sqrt{\frac{1}{(n-k-1)}}\text{, где}$$

n – число уравнений (наблюдений);

k – число оцениваемых параметров регрессии (число регрессоров).

df - число степеней свободы. Для регрессии *df=k*, для остатков *df= n-k-1*

SS – сумма квадратов отклонений от среднего арифметического. Приводится отдельно для регрессии и для остатков ***SSE***, являющейся из важных характеристик качества полученной зависимости. Если ***SSE*** близко к нулю, то это говорит о хорошем качестве полученной зависимости.

MS – дисперсии представляют собой несмещенные оценки дисперсий зависимости переменной, обусловленных соответственно регрессией и воздействием неучтенных случайных факторов ошибок. Также приводистя отдельно для регрессии и отдельно для остатков MSE $\sigma^2=SSE/(n-2)$.

F (критерий значимости уравнения регрессии) - значение критерия Фишера для регрессии**.** Рассчитывается по формуле:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

Если расчетное значение *F* оказывается большим чем критическое значение F_q , то есть уравнение регрессии значимо, следовательно исследуемая зависимая переменная *y* очень близко описывается включенными в регрессионную модель переменными. Критическое значение критерия Фишера определяется по таблицам (или с помощью функций в статистических пакетах) с учетом следующих параметров (вероятность ошибки α (обычно 0,05), *k, n-k-1*).

Значимость F – показывает абсолютное значение вероятности ошибки уравнения регрессии, при рассчитанном значении критерия Фишера *F* (не должно превышать вероятность ошибки 0,05 или 0,01). К примеру, если значимость ***F*** оказалась равна 0,03**,** то это значит, что уравнение регрессии значимо лишь при 5 %-ном уровне значимости, но не при 1 %-ном уровне значимости. Значение 0,005 показывает, что уравнение значимо не только при 5 %-ном уровне значимости, но и при 1 %-ном уровне значимости. Если уравнение регрессии значимо, то значение **значимости *F*** стремится к нулю.

Коэффициенты – значения коэффициентов в уравнении регрессии. В итоге уравнение регрессии примет вид: $y = b + k_1x_1 + k_2x_2 + \sum k_i x_i$

Стандартная ошибка - показывает отклонение фактических значений результирующего показателя от теоретической расчетной величины на удалении *σ*

при распределении Гаусса, или (грубо) какой разброс данных присущ выборке.

Если стандартная ошибка больше абсолютной величины коэффициента, это **коэффициент незначимый**. Этот коэффициент (свободный член или регрессор) нужно исключить из уравнения регрессии и пересчитать таблицы. Но это грубый анализ. Столбец ***t-статистика*** дает более точную оценку значимости коэффициентов.

t-статистика - оценка коэффициента, деленная на его стандартную ошибку *tp=(Коэффициент)/(Стандартная ошибка)*. Этот критерий имеет закон распределения Стьюдента с числом степеней свободы $n-(k+1)$: число исходных точек, минус число регрессоров, минус свободный член, если есть.

Очень часто при построении регрессионной модели неизвестно, влияет тот или иной фактор x_i на y. Включение в модель факторов, которые не влияют на выходную величину, ухудшает качество модели. Вычисление *t-статистики* помогает обнаружить такие факторы. Приближенную оценку можно сделать так: если при $n>>k$ величина *t-статистики* по абсолютному значению существенно **больше трех**, соответствующий коэффициент следует считать **значимым**, **а фактор включить в модель**, иначе исключить из модели. Более точно критическое значение коэффициента Стьюдента можно определить по таблице (или с помощью функций в статистических пакетах) с учетом вероятности ошибки (обычно 0,05) и степени свободы $(n-k-1)$. Таким образом, можно предложить технологию построения регрессионной модели, состоящую из двух этапов:

1) обработать статистическим пакетом (к примеру *Excel*) все имеющиеся данные, проанализировать значения *t-статистики*;

2) удалить из таблицы исходных данных столбцы с теми факторами, для которых коэффициенты незначимы и обработать статистическим пакетом (к примеру, *Excel*) новую таблицу.

P*-Значение** - фактически достигнутые уровни значимости, соответствующие значениям *t-статистики*. К, примеру ***P-значение 0,03 показывает, что этот коэффициент значим лишь при 5 %-ном уровне значимости, но не при 1 %-ном уровне значимости. Значение 0,005 показывает, что этот коэффициент значим не

только при 5 %-ном уровне значимости, но и при 1 %-ном уровне значимости. Если значение оказывается больше 0,05, то такой коэффициент следует **исключить из модели** и пересчитать коэффициенты. Его оставление только ухудшит качество модели. Для значимых коэффициентов *P*-Значение **близко к нулю.**

Столбцы **НИЖНИЕ 95 %** и **ВЕРХНИЕ 95 %** показывают соответственно нижние и верхние интервалы значений коэффициентов при 95 %-ном уровне значимости [8, 9].

Значение столбца НИЖНИЕ 95 % = КОЭФФИЦИЕНТ – СТАНДАРТНАЯ ОШИБКА*t-критерий

Значение столбца НИЖНИЕ 95 % = КОЭФФИЦИЕНТ + СТАНДАРТНАЯ ОШИБКА*t-критерий.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ №6 «Регрессивный анализ»

На предприятии выпускающем ИМС, собрали показатели параметров внешней среды, температуру травителя (KOH/спирт/H_2O) для травления Si и оценили глубину травления при этих условиях, полученные данные свели в таблицу 1.9.

<u>Задание:</u> построить математическую модель описывающую влияние этих факторов на глубину травления Si. Определить факторы, наиболее сильно оказывающие влияние на процесс травления. Сделать выводы. Оценить глубину травления при температуре травителя 120 С и нормальных внешних условиях (давление 760 мм рт.ст., температура окружающей среды 25 С, абсолютная влажность 27 г/м^3) за время 35 минут.

Таблица 1.9.

температура травителя, С	температура окружающей среды, С	влажность г/м3	атмосферное давление, мм рт.ст.	время травления, мин	глубина травления, нм
30	20	21	768	15	171
…	…	…	…	…	…
100	18	16,8	764	17	254

Список использованных источников:

1. Статистика: теория и практика в Excel: учеб. пособие / В.С. Лялин, И.Г. Зверева, Н.Г. Никифорова. - М.: Финансы и статистика; ИНФРА-М, 2010. - 448 е.: ил.
2. Анализ данных в Excel. Просто как дважды дна / П. Корнелл; пер. с англ. – М.: Эксмо, 2007. — 224 с.: ил.
3. К. Берк, П. Кэйри. Анализ данных с помощью Microsoft Excel. : Пер. с англ. — М. : Издательский дом "Вильяме", 2005. — 560 с.
4. Ларсен, Рональд У. Инженерные расчеты в Excel.: Пер. с англ. – М.: Издательский дом «Вильямс», 2004. – 544 с.
5. Эконометрика. Под редакцией И. И. Елисеевой - М.: Финансы и статистика., 2007. - 575 с.
6. Воскобойников Ю.Е., Тимошенко Е.И. Математическая статистика с примерами в Excel: Учебное пособие. - Новосибирск: Изд.НГАСУ, 2006. - 154 с.
7. Н. Джонсон, Ф. Лион. Статистика и планирование эксперимента в технике и науке. Методы обработки данных. - М.: Мир, 1980. - 511 с.
8. Макарова Н. В., Трофимец В. Я. Статистика в Excel: Учеб. пособие. - М.: Финансы и статистика, 2002. - 368 е.: ил.
9. Хаушильд В., Mom В. Статистика для электротехников в приложении к технике высоких напряжений / Пер. с нем.— Л.: Эиергоатомиздат. Леиингр. отд-ние, 1989.– 312 с.: ил.