Вашукевич Е.В.

Статистические методы в биологии

06.04.01 - Биология

Иркутск, 2021

Содержание

ВВЕДЕНИЕ

Важным показателем научно-исследовательской работы является её эффективность. Высокие значения могут быть достигнуты в тех случаях, когда результативность работы пригодны для обобщения и способствуют получению новой информации. Эффективность проводимых исследований во многом определяются качеством планирования, постановки и проведения эксперимента, а также глубинной анализа полученных результатов. В проведении исследований немаловажную роль играют все перечисленные этапы в равной степени.

В настоящее время уровень математической подготовки биолога может быть различным. И даже компьютерные программы, специально написанные на извлечение достоверной информации, в руках конкретного исследователя будут служить разным целям. Такая ситуация, когда одни и те же методы математической статистики, применяемые к исходным материалам, дают совершенно неодинаковые значения, отличающиеся по уровню достоверности и зависящая от личности «обработчика», является, не допустима.

Принципиальное значение в качестве анализа материалов имеют ошибки, выявить которые очень сложно. К такому типу ошибок можно отнести неправильное определение достоверности отдельных показателей из числа представляемых в таблицах, графиках, доверительных интервалах и т.д.

Информативность результатов экспериментальных исследований в немалой степени определяется глубиной статобработки и качеством их представления. Применение статистических приемов позволяет исследователю выявить и предотвратить издержки в постановке опытов и интерпретации результатов ещё до их публикации.

Математическая обработка данных, полученных в ходе биологического исследования – важное условие при выполнении курсовых и дипломных работ. Для того, чтобы сделать достоверное предположение об изучаемом явлении, используют только части – выборки, так как всю генеральную совокупность обследовать практически невозможно (тысячи, сотни тысяч, миллионы испытуемых).

Выборка должна отражать все свойства генеральной совокупности должна быть репрезентативной, для того чтобы любой исследователь мог понять результаты ваших исследований.

Основные способы достижения этого условия:

1) Случайная выборка;

2) Моделирование выборки по свойствам генеральной совокупности.

Существенным при организации выборки является вопрос о необходимом и достаточном числе испытуемых. Малое количество ($n<10$) испытуемых не обеспечит точности результатов. Большое количество ($n >30$) приведет к увеличению трудоемкости (времени и стоимости) исследования.

Статистическая достоверность, или статистическая значимость, результатов исследования определяется при помощи методов статистического вывода, которые предъявляют определенные требования к численности, или объему выборки.

Общие рекомендации по численности выборки:

1. Наибольший объем выборки необходим при разработке диагностической методики - от 200 до 1000-2500 человек.

2. Если необходимо сравнивать 2 выборки, их общая численность должна быть не менее 50 человек; численность сравниваемых выборок должна быть приблизительно одинаковой.

3. Если изучается взаимосвязь между какими-либо свойствами, то объем выборки должен быть не меньше 30-35 человек.

4. Чем больше изменчивость изучаемого свойства, тем больше должен быть объем выборки. Поэтому изменчивость можно уменьшить, увеличивая однородность выборки, например, по полу, возрасту и т. д.

Глава 1

Биологическая статистика: её роль и значение

Необходимость статистической обработки и представление экспериментальных данных возникли сразу, как только биологи перешли от описательного метода к анализу экспериментальных результатов. Действительно, исследователю, имеющему дело с измерениями и обработкой данных, постоянно приходится обращаться к элементарным основам математической статистики, чтобы извлечь максимально полезную информацию из результатов измерений.

Первые опыты использования математических знаний для анализа биологических явлений принадлежат Френсису Гальтону (1889), который использовал статистический анализ для оценки связи между отдельными признаками у людей и степени сходства между родственниками по такому сложному признаку как рост людей. Гальтон впервые ввел понятие **биометрика** в научную литературу.

Исследования Гальтона были продолжены Карлом Пирсоном, последователем, учеником, почитателем и пропагандистом Гальтона. Пирсон внес большой вклад в разработку теории и практики использования статистических показателей, применительно к описанию биологических явлений.

Несмотря на то, что ни Гальтон, ни Пирсон не смогли выделить генетические компоненты, контролирующие рост человека, своими работами эти авторы привлекли внимание биологов к статистике как к одному из методов анализа различных признаков. Позднее большой вклад именно в оценку отдельных факторов, вносящих вклад в формирование и развитие признаков, был сделан Робертом Фишером (1918).

С 60-х годов XX века математическая статистика стала обязательным условием анализа экспериментальных данных в области биологии, медицины и сельского хозяйства.

Одним из факторов, повлиявшим на необходимость использования статистических расчетов экспериментальных данных, является то, что все биологические объекты обладают противоположными свойствами.

Широкая амплитуда изменчивости признаков у различных объектов вынуждает экспериментаторов прибегать как к усреднению данных, так и к оценке границ изменчивости и силы связи между признаками.

Другим важным обстоятельством, повлиявшим на процесс внедрения статистических методов для анализа биологических явлений, явилось то, что практически все биологические явления и свойства подчиняются статистическим закономерностям, характерным не отдельным объектам, а целым совокупностям объектов. Оказалось, что если сгруппировать данные, полученные путем измерения любых биологических признаков, в единую совокупность, то эта совокупность будет иметь вид чисто статистической совокупности. Поэтому математическую статистику, используемую в приложении к биологии, стали называть **биологической статистикой** или **биометрией.**

Таким образом, биологической статистикой или биометрией называется область научного знания, охватывающая классификацию, систематизацию и обработку экспериментальных данных в биологии, медицине и сельском хозяйстве методами математической статистики.

Выборочные совокупности.

Цель большинства исследований состоит в сборе данных, которые впоследствии помогают получить информацию в какой-либо области. Обычно данные исследователям получаются из выборки индивидуумов. Цель любой исследовательской работы состоит в том, чтобы сгруппировать эти данные и извлечь из них достоверную информацию.

Всякое множество идентифицируемых объектов, отличающихся друг от друга незначительно по конкретному признаку, но сохраняющих сходство по некоторым существенным характеристикам, называется **совокупностью**. Совокупностью можно называть стадо животных, делянку или поле растений, породу животных или сорт культурных растений, клоны бактерий, штаммы вирусов и др. Совокупности представлены отдельными **членами** или **единицами**. Совокупность членов или единиц называется **объемом совокупности**. Само измеренное значение каждой единицы совокупности называется **вариантой** совокупности. Различия между вариантами совокупности называются **вариацией** или **дисперсией**.

Если бъем совокупности любой выборки обозначить через **n**, а каждую варианту совокупности через **x**, то в таком случае ряд вариант в совокупности можно записать, как x_1, x_2, x_3 … x_n. Следовательно, если исследовано влияние, какого либо вещества на деятельность сердца, то количество использованных животных можно назвать совокупностью, объем которой, например, равен n =20. Действие вещества на деятельность каждого животного - это варианта совокупности.

Если измеренный материал представляет собой только часть от какого-то общего большого материала, то такая совокупность будет называться **выборочной совокупностью**, а весь огромный материал, из которого взята выборочная совокупность, называется **генеральной совокупностью**. Например, если измеренные 100 зерен какого-то сорта пшеницы представляют собой выборочную совокупность с объемом n =100, то вес зерна целого сорта пшеницы можно считать генеральной совокупностью с объемом, равным $N \to \infty$.

Обработка данных и интерпретация полученных расчетов зависит от того, какие признаки были измерены в эксперименте. Признаки биологических объектов в основном представлены тремя категориями.

1. Признаки совокупности, различающиеся определенным качеством. Такие признаки носят название качественные или альтернативные. Например, совокупность представлена красными и белыми цветками гороха, как в опытах Г. Менделя.

2. Признаки, имеющие количественную меру, но различия между отдельными вариантами совокупности выражаются целыми числами. Например, число отложенных яиц у птиц. Такая количественная изменчивость и такие количественные признаки называются дискретными.

3. Признаки, имеющие количественную меру, но между отдельными вариантами совокупности нет четко выраженных (дискретных) границ. Например, живая масса животных, выраженная килограммах, но не в граммах. Такая изменчивость и такие признаки называются количественными признаками с непрерывной изменчивостью (мерные признаки, континьюальные).

Группировка данных выборочной совокупности по признакам с дискретной изменчивостью.

Если измерения какого-то признака у изучаемого биологического объекта произведены, они записаны в журналах в порядке измерения объектов, т.е. материал представляет собой набор цифр часто в хаотичном порядке. При рассмотрении этих цифровых значений невозможно обнаружить какие-либо закономерности. Существуют некоторые простые правила упорядоченья материала.

1. Можно переписать полученные данные в порядке нарастания их величины или в порядке уменьшения величин. Такой способ упорядоченья данных называется **ранжированием.** Ранжирование данных дает уже некоторое представление о характере экспериментального материала. Так, строго обозначены минимальное и максимальное значения вариант, которые носят название **лимитов изменчивости**. Самое маленькое значение варианты называется минимальным значением (min), а самое большое - максимальным значением (max). В центре такого упорядоченного (ранжированного) ряда цифр сосредоточено основное количество вариант со "средним значением" признака.

Пример. В университете проводили исследования по изучению функционирования работы сердца среди студентов первого курса на занятии физическая культура. Изучалась частота сердечных сокращений 30 студентов. Получены следующие данные:

65, 64, 63, 62, 68, 77, 68, 64, 73, 78, 65, 68, 62, 78, 62, 67, 69, 65, 71, 72, 68, 62, 63, 71, 68, 73, 67, 69, 68, 65.

Ранжируем варианты совокупности в порядке нарастания частоты сердечных сокращений у студентов первого курса;

62, 62, 62, 62, 63, 63, 64, 64, 65, 65, 65, 65, 67, 67, 68, 68, 68, 68, 68, 68, 69, 69, 71, 71, 72, 73, 73, 77, 78, 78.

Такой ранжированный ряд дает представление о том, что лимиты изменчивости такого признака, как частота сердечных сокращений, находится в границах min =62; max =78. Среднее значение частоты сердечных сокращений можно приблизительно оценить равным, 70.

2. Другой более наглядный способ упорядоченья вариант совокупности состоит в следующем. Необходимо выписать в столбик все типы вариантов, которые встретились в совокупности. Цифры могут быть записаны как в порядке нарастания от меньшего значения к большему, т.е. от 1 до 8, так и в порядке уменьшения, т.е. от 8 до 1. Запишем цифры в порядке нарастания, и определим сколько раз каждая цифра, встречается в распределении. Тогда различные типы вариант, которые встретились в распределении, будут называться **классами** и обозначаться x_i, а числа, соответствующие каждому классу - **численностями** или **частотами** и обозначаться будут буквой f.

Количество поросят ($x_{i)}$	Сколько раз встречается данное значение в распределении (f)
62	4
63	2
64	2
65	4
67	2
68	6
69	2
71	2
72	1
73	2
77	1
78	2
сумма	30

Такой ряд чисел, состоящий из классов и частот, называется **вариационным рядом**. Вариационный ряд дает более четкую картину закономерностей распределения по сравнению с простым ранжированием.

Для формирования выборки можно использовать следующие приемы:

Простая вероятностная выборка:

Простая повторная выборка. Использование такой выборки основывается на предположении, что каждый респондент с равной долей вероятности может попасть в выборку. На основе списка генеральной совокупности составляются карточки с номерами респондентов. Они помещаются в колоду, перемешиваются и из них наугад вынимается карточка, записывается номер, потом возвращается обратно. Далее процедура повторяется столько раз, какой объём выборки нам необходим. Минус: повторение единиц отбора.

Процедура построения простой случайной выборки включает в себя следующие шаги:

1. необходимо получить полный список членов генеральной совокупности и пронумеровать этот список. Такой список, напомним, называется основой выборки;

2. определить предполагаемый объем выборки, то есть ожидаемое число опрошенных;

3. извлечь из таблицы случайных чисел столько чисел, сколько нам требуется выборочных единиц. Если в выборке должно оказаться 100 человек, из таблицы берут 100 случайных чисел. Эти случайные числа могут генерироваться компьютерной программой.

4. выбрать из списка-основы те наблюдения, номера которых соответствуют выписанным случайным числам

Простая случайная выборка имеет очевидные преимущества. Этот метод крайне прост для понимания. Результаты исследования можно распространять на изучаемую совокупность. Большинство подходов к получению статистических выводов предусматривают сбор информации с помощью простой случайной выборки. Однако метод простой случайной выборки имеет как минимум четыре существенных ограничения:

1. зачастую сложно создать основу выборочного наблюдения, которая позволила бы провести простую случайную выборку.

2. результатом применения простой случайной выборки может стать большая совокупность, либо совокупность, распределенная по большой географической территории, что значительно увеличивает время и стоимость сбора данных.

3. результаты применения простой случайной выборки часто характеризуются низкой точностью и большей стандартной ошибкой, чем результаты применения других вероятностных методов.

4. в результате применения SRS может сформироваться нерепрезентативная выборка. Хотя выборки, полученные простым случайным отбором, в среднем адекватно представляют генеральную совокупность,

некоторые из них крайне некорректно представляют изучаемую совокупность. Вероятность этого особенно велика при небольшом объеме выборки.

Простая бесповторная выборка. Процедура построения выборки такая же, только карточки с номерами респондентов не возвращаются обратно в колоду.

Систематическая вероятностная выборка. Является упрощенным вариантом простой вероятностной выборки. На основе списка генеральной совокупности через определённый интервал (К) отбираются респонденты. Величина К определяется случайно. Наиболее достоверный результат достигается при однородной генеральной совокупности, иначе возможны совпадение величины шага и каких-то внутренних циклических закономерностей выборки (смешение выборки). Минусы: такие же как и в простой вероятностной выборке.

Серийная (гнездовая) выборка. Единицы отбора представляют собой статистические серии (семья, школа, бригада и т. п.). Отобранные элементы подвергаются сплошному обследованию. Отбор статистических единиц может быть организован по типу случайной или систематической выборки. Минус: Возможность большей однородности, чем в генеральной совокупности.

Районированная выборка. В случае неоднородной генеральной совокупности, прежде, чем использовать вероятностную выборку с любой техникой отбора, рекомендуется разделить генеральную совокупность на однородные части, такая выборка называется районированной. Группами районирования могут выступать как естественные образования (например, районы города), так и любой признак, заложенный в основу исследования. Признак, на основе которого осуществляется разделение, называется признаком расслоения и районирования.

«Удобная» выборка. Процедура «удобной» выборки состоит в установлении контактов с «удобными» единицами выборки — с группой студентов, спортивной командой, с друзьями и соседями. Если необходимо получить информацию о реакции людей на новую концепцию, такая выборка вполне обоснована. «Удобную» выборку часто используют для предварительного тестирования анкет.

Глава 2

Средняя арифметическая величина и её свойства.

Среднее арифметическое, которое очень часто называют просто «среднее значение», получают путем сложения всех значений и деления этой суммы на число значений в выборке. Это можно показать с помощью алгебраической формулы. Набор n наблюдений переменной x можно изобразить, как x_1 x_2 x_3 …, x_n. Формулу для определения средней арифметической величины обозначают следующим образом:

$$x = \frac{x_1+x_2+x_3+\ldots,+x_n}{n}$$

Средняя арифметическая величина обладает рядом свойств:

1. Если к каждому значению выборки прибавить или отнять одну и ту же величину, или умножить и разделить на одну и ту же величину, то средняя арифметическая увеличится или уменьшится на эту же величину.

2. Алгебраическая сумма отклонений отдельных вариант совокупности от средней арифметической этой совокупности равна нулю.

3. Сумма квадратов отклонений вариант совокупности от средней арифметической x меньше суммы квадратов отклонений от любой другой величины.

Средняя арифметическая величина является очень важным параметром, характеризующим выборочную совокупность. Она используется для характеристики любых совокупностей в технике, медицине и биологии.

Средняя арифметическая величина является обобщенной характеристикой совокупности. Часто значение средней арифметической величины реально не существует, например, 4,5 щенка и др. В этом смысле средняя арифметическая является абстрактной величиной, но в, то, же время она и конкретная величина, характеризующая типичное состояние признака в совокупности.

При предоставлении результатов в биологии средняя арифметическая величина обозначается, как М. Особо важно знать среднее арифметическое значение выборках с нечеткими границами и большим разбросам в значениях.

Пример.

Измерялась частота сердечных сокращений у 10 студентов первого курса на занятии физическая культура. Получены следующие данные:

62, 65, 72, 68, 69, 71, 63, 67, 64, 62

Средняя арифметическая в этом случае будет подсчитываться следующим образом.

62+65+72+68+69+71+63+67+64+62 / 10=66,3

Средняя арифметическая величина частоты сердечных сокращений у 10 студентов первого курса на занятии физическая культура составляет 66,3. Соответственно М=66,3.

Среднее квадратическое значение

Более точным показателем, характеризующим вариацию или рассеяние вариант вокруг среднего арифметического значения, является **среднее квадратическое значение**. Оно основано на рассмотрении отклонений значений признака отдельных единиц совокупности от средней арифметической. При этом используется способ усреднения отклонений вариантов от средней арифметической, позволяющий обойти трудность, обусловленную равенством нулю их алгебраической суммы. Данный способ сводится к расчету квадратов отклонений вариантов от средней с их последующим усреднением.

Среднее квадратическое отклонение *(σ)* представляет собой корень квадратный из дисперсии:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{\sum f_i}}$$

Среднее квадратическое отклонение показывает, насколько её среднее колеблется в исследуемой совокупности, и выражается в тех же единицах измерения, что и варианты.

Пример

62-66,3=-4,3

65-66,3=-1,3

72-66,3=5,7

68-66,3=1,7

69-66,3=2,7

71-66,3=4,7

63-66,3=-3,3

67-66,3=0,7

64-66,3=-2,3

62-66,3=-4,3

Возводим в квадрат каждое отклонение.

-4,3*-4,3=18,49

-1,3*-1,3=1,69

5,7*5,7=32,49

1,7*1,7=2,89

2,7*2,7=7,29

4,7*4,7=22,09

-3,3*-3,3=10,89

0,7*0,7=0,49

-2,3*-2,3=5,29

-4,3*-4,3=18,49

Находим сумму квадрата каждого отклонения

18,49+1,69+32,49+2,89+7,29+22,09+10,89+0,49+5,29+18,49=120,1

Извлекаем корень из 120,1. Таким образом, мы вычислили σ=10,96.

Ошибка средней арифметической

Для представления своих результатов вариантов выборки указывают значение средней арифметической (М) и ошибки средней в виде (м) М±м, а на графиках устанавливаются соответствующую ошибку в виде ⊥ ⊤.

Эта ошибка определяется по формуле:

$$m = \frac{\sigma}{\sqrt{n}},$$

$$m = \frac{\sigma}{\sqrt{n-1}}$$

Где m- ошибка средней арифметической,

где σ - среднее квадратическое отклонение выборочной совокупности;

n - объем выборки (число измерений или испытуемых).

Из формулы видно, что чем больше разнообразие признака (величина σ), тем больше ошибка, и чем больше численность выборки, тем меньше ошибка

Если бы все объекты были одинаковы, то есть разнообразие было бы равно нулю, то и ошибка была бы равна нулю (m = 0). В этом случае даже один экземпляр точно характеризовал бы всю генеральную совокупность.

Например:

Мы вычислили значение σ в предыдущей главе (см. выше). Для выборки менее 30 используется 2 формула подсчета средней арифметической. Если n более 30 используется первая формула. Подставляем значения:

m=10,96/3=3,65

Мы получаем значение m=3,65.

Глава 3

Критерий Стьюдента.

Обычна ситуация исследования, когда интересующее исследователя свойство изучается на двух или более выборках с целью их дальнейшего сравнения. Эти выборки могут находиться в различных соотношениях. Независимые выборки характеризуются тем, что вероятность отбора любого испытуемого одной выборки не зависит от отбора любого из испытуемых другой выборки. Количество испытуемых может не совпадать в первой и второй выборках. Зависимые выборки характеризуются тем, что каждому испытуемому одной выборки поставлен в соответствие по определенному критерию испытуемый из другой выборки.

Результаты лабораторных исследований - излюбленный объект для статистической обработки в научных трудах по самым различным специальностям. Специалисты очень часто не могут из огромного количества материала вычленить наиболее важные результаты. Очень часто такая картина связанна с не желанием использовать статистический анализ и не правильное использование различных критериев. Прежде всего, необходима только идея: что и с чем будем сравнивать. Затем правильно выбрать критерий для определения достоверности и только затем грамотное использование математических формул. Наиболее часто в настоящее время в научной литературе можно увидеть t-критерий Стьюдента и критерий Вилкаксона (Мани-Уитни).

t-критерий Стьюдента — общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Наиболее частые случаи применения t-критерия связаны с проверкой равенства средних значений в двух выборках.

Данный критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннесса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

Данный критерий наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности». В этом случае можно применить критерий Стьюдента (при условии достаточно больших объёмов выборок (n≥30), или убедившись, что статистические ряды близки к нормальному закону распределения).

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.

Рассмотрим пример применения двухвыборочного t-критерия.

Исследовали частоту сердечных сокращений у студентов 1 курса на занятии физическая культура. Первая группа составила студенты, входящие по медицинскому осмотру в основную группу. Ко второй группе относились студенты, по медицинскому осмотру в дополнительную группу (студентам имели отклонения в здоровье). Количество студентов в первой группе равно 10, во второй 9.

Алгоритм расчета параметров распределения:

1. Размещаем полученные данные в таблице в первой вертикальной колонке;

№	ЧСС	№	ЧСС

1	62	1	65
2	65	2	64
3	72	3	63
4	68	4	62
5	69	5	68
6	71	6	77
7	63	7	68
8	67	8	64
9	64	9	73
10	62		

2. Рассчитываем среднее арифметическое имеющихся данных:

1	66,3	2	67,2

3. Вторую колонку заполняем отклонениями данных от среднего значения: X - M (из каждого значения вычитается среднее арифметическое);

	66,3±3,6		67,2±5,01

Таким образом, в результате несложных расчетов получаем два основных параметра – среднее значение и стандартное отклонение, характеризующих распределение признака в совокупности данных (в нашем случае – значения частоты сердечных сокращений у студентов первого курса основной и дополнительной группы).

Полученные результаты мы можем записать в формате М±м.

Наиболее простая формула расчета критерия Стьюдента выглядит так:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}$$

(в числителе – разность средних значений двух групп, в знаменателе – квадратный корень из суммы квадратов стандартных ошибок этих средних).

$t=66{,}3-67{,}2/\sqrt{3{,}6^2+5{,}01^2}=-0{,}9/6{,}17=0{,}15$

Сравниваем полученное в эксперименте значение t с табличным значением с учетом степеней свободы.

ν	Уровень значимости α		
	0,05	0,01	0,001
15	2,131	2,947	4,073
16	2,120	2,921	4,015
17	2,110	2,898	3,965
18	2,101	2,878	3,922
19	2,093	2,861	3,883
20	2,086	2,845	3,850
21	2,080	2,831	3,819
22	2,074	2,819	3,792
23	2,069	2,807	3,768
24	2,064	2,797	3,745
25	2,060	2,787	3,725
26	2,056	2,779	3,707
27	2,052	2,771	3,690
28	2,048	2,763	3,674
29	2,045	2,756	3,659
30	2,042	2,750	3,646
31	2,040	2,744	3,633
32	2,037	2,738	3,622
33	2,035	2,733	3,611
34	2,032	2,728	3,601
35	2,030	2,724	3,591

Табличное значение t крит равняется 2,093 при допущении возможности риска сделать ошибочное суждение в пяти случаях из ста (уровень значимости=5 % или 0,05).

Если полученное в эксперименте эмпирическое значение t превышает табличное, то есть основания принять альтернативную гипотезу (Н1). В эксперименте t=0,15, табличное t=2,093, 2,093>0,15, откуда следует вывод о том, что частота сердечных сокращений у студентов 1 курса не зависит от медицинской группы.

Данный метод используется только в случаях, когда имеется 2 группы. Одна контрольная (интактная), а вторая экспериментальная. Количество индивидуальных показателей в первой и второй группе может быть не одинаково.

Пример 2.

Рассмотрим пример, 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 случае связанных выборок с равным числом измерений в каждой можно использовать более простую формулу t-критерия Стьюдента.

Вычисление значения t осуществляется по формуле:

$$t_{эмп} = \frac{\bar{d}}{Sd}$$

Где —d разности между соответствующими значениями переменной M_1 и переменной M_2, а d - среднее этих разностей;

Sd вычисляется по следующей формуле:

$$Sd = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n \cdot (n-1)}}$$

Если $t_{эмп} < t_{крит}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная.

Пример. Изучался уровень ориентации учащихся на художественно-эстетические ценности. С целью активизации формирования этой ориентации в экспериментальной группе проводились беседы, выставки детских рисунков, были организованы посещения музеев и картинных галерей, проведены встречи с музыкантами, художниками и др. Закономерно встает вопрос: какова эффективность проведенной работы? С целью проверки эффективности этой работы до начала эксперимента и после давался тест. Из методических соображений в таблице приводятся результаты небольшого числа испытуемых.

Результаты эксперимента

Ученики (n=10)	**Баллы**		**Вспомогательные расчеты**	
	до начала эксперимента (X)	**в конце эксперимента (У)**	**d**	**d^2**
Иванов	14	18	4	16
Новиков	20	19	-1	1
Сидоров	15	22	7	49
Пирогов	11	17	6	36
Агапов	16	24	8	64
Суворов	13	21	8	64
Рыжиков	16	25	9	81
Серов	19	26	7	49
Топоров	15	24	9	81
Быстров	9	15	6	36
$\sum$	148	211	63	477
Среднее	14,8	21,1		

Вначале произведем расчет по формуле:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{63}{10} = 6,3$$

Затем применим формулу, получим:

$$Sd = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n \cdot (n-1)}} = \sqrt{\frac{477-(63 \cdot 63)/10}{10 \cdot (10-1)}} = \sqrt{\frac{477-396,9}{90}} = \sqrt{0,890} = 0,943$$

И, наконец, следует применить формулу. Получим:

$$t_{эмп} = \frac{\bar{d}}{Sd} = \frac{6,3}{0,943} = 6,678$$

Число степеней свободы: k=10-1=9 и по таблице находим $t_{крит}$ =2.262, экспериментальное t=6,678, откуда следует возможность принятия альтернативной гипотезы (H_1) о достоверных различиях средних арифметических, т. е. делается вывод об эффективности экспериментального воздействия.

В терминах статистических гипотез полученный результат будет звучать так: на 5% уровне гипотеза H_0 отклоняется и принимается гипотеза H_1. и обозначаться как $p<0,05$ и соответствующим символом (*), двумя символами (**) обозначается $p<0,01$. такое обозначение в таблицах и на графиках принято, как типовое и не требует дополнительной расшифровки.

В настоящее время применяются оба подхода к сравнению двух групп по количественному признаку: путем проверки статистических гипотез и с использованием доверительного интервала. Эти два подхода дают ответы на различные вопросы: в первом случае, – «В какой степени можно быть уверенным, что различия между генеральными совокупностями действительно существуют?», во втором – «Насколько велики различия генеральных совокупностей?»

Оба подхода основаны на одних и тех же допущениях и статистических принципах, поэтому скорее не конкурируют, а дополняют друг друга. Тем не менее, значение подхода с построением доверительного интервала все более возрастает. Знакомясь с работой, содержащей лишь средние значения, стандартные ошибки средних и число наблюдений, мы можем самостоятельно рассчитать доверительный интервал, чтобы понять, имеет исследование сугубо академическую либо еще и практическую ценность.

Глава 4

Критерий Вилкаксона-Манни-Уитни.

U-критерий Манна — Уитни (англ. Mann — Whitney U-test) — статистический критерий, используемый для оценки различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно. Позволяет выявлять различия в значении параметра между малыми выборками ($n<3$).

Другие названия: критерий Манна — Уитни — Уилкоксона (англ. Mann — Whitney — Wilcoxon, MWW), критерий суммы рангов Уилкоксона (англ. Wilcoxon rank-sum test) или критерий Уилкоксона — Манна — Уитни (англ. Wilcoxon — Mann — Whitney test). Реже: критерий числа инверсий.

Данный метод выявления различий между выборками был предложен в 1945 году Фрэнком Уилкоксоном (F. Wilcoxon). В 1947 году он был существенно переработан и расширен Х. Б. Манном (H. B. Mann) и Д. Р. Уитни (D. R. Whitney), по именам, которых сегодня обычно и называется.

Критерий принимает во внимание не только знаки разностей, но и их величину, и поэтому является более мощным критерием. Индивидуальную разность рассчитывают для каждой пары результатов. Нулевые разности классифицируют как положительные или отрицательные. Разности располагают в порядке их модуля без учета их знаков и соответственно ранжируют.

U-критерий Манна-Уитни используется для оценки различий между двумя малыми выборками ($n_1,n_2 \geq 3$ или $n_1=2$, $n_2 \geq 5$) по уровню количественно измеряемого признака. При этом первой выборкой принято считать ту, где значение признака больше.

Существует несколько способов использования критерия и несколько вариантов таблиц критических значений, соответствующих этим способам (Гублер Е. В., 1978; Рунион Р., 1982; Захаров В. П., 1985; McCall R., 1970; Krauth J., 1988).

Этот метод определяет, достаточно ли мала зона перекрещивающихся значений между двумя рядами. Мы помним, что 1-м рядом (выборкой, группой) мы называем тот ряд значений, в котором значения, по предварительной оценке, выше, а 2-м рядом - тот, где они предположительно ниже.

Чем меньше область перекрещивающихся значений, тем более вероятно, что различия достоверны. Иногда эти различия называют различиями в расположении двух выборок.

Рассмотрим алгоритм применения U-критерия Манна-Уитни:

1.Перенести все данные испытуемых на индивидуальные карточки, пометив карточки 1-й выборки одним цветом, а 2-й – другим.

2.Разложить все карточки в единый ряд по степени возрастания признака и проранжировать в таком порядке.

3.Вновь разложить карточки по цвету на две группы.

4.Подсчитать сумму рангов отдельно по группам и проверить, совпадает ли общая сумма рангов с расчетной.

5. Определить большую из двух ранговых сумм .

6.Вычислить эмпирическое значение *U*:

$$U = n_1 \cdot n_2 + \frac{n_x \cdot (n_x + 1)}{2} - T_x ,$$

где - количество испытуемых в - выборке (i = 1, 2), - количество испытуемых в группе с большей суммой рангов.

7. Задать уровень значимости α и, используя специальную таблицу, определить критическое значение $U_{кр}(\alpha)$. Если , то H_0 на выбранном уровне значимости принимается.

Рассмотрим использование U критерия Манна-Уитни на примере.

Проведение контрольной работы по математике (алгебра и геометрия) в средней общеобразовательной школе дало следующие результаты по 10-балльной шкале для класса, обучающегося по программе «Развивающего обучения» (7 «Б»), и класса, обучающегося по традиционной системе (7 «А»):

Ученик \ Класс	7 «А» (баллы)	7 «Б» (баллы)
1	9	5
2	7	10
3	7	7
4	8	8
5	6	8
6	4	4
7	4	6
8	8	8
9	6	8
10	6	9
11	5	7
12	-	10

Определите, превосходят ли учащиеся 7 «Б» учащихся 7 «А» по уровню знаний по математике.

Сравнение результатов показывает, что баллы, полученный за контрольную работу, в 7 «Б» классе несколько выше, поэтому первой считаем выборку результатов 7 «Б» класса. Таким образом, нам требуется определить, можно ли считать имеющуюся разницу между баллами существенной. Если можно, то это будет означать, что класс, обучающийся по системе «развивающего обучения» имеет более качественные знания по математике. В противном случае, на выбранном уровне значимости различие окажется несущественным.

Для оценки различий между двумя малыми выборками (в данном примере их объёмы равны: $n_1=12$, $n_2=11$) используем критерий Манна-Уитни. Проранжируем представленную таблицу:

7 «Б» (баллы)	ранг	7 «А» (баллы)	ранг

10	22,5		
10	22,5	9	20.5
9	20.5	8	16.5
8	16.5	8	16.5
8	16.5	7	11.5
8	16.5	7	11.5
8	16.5	6	7.5
7	11.5	6	7.5
7	11.5	6	7.5
6	7.5	5	4.5
5	4.5	4	2
4	2	4	2
Сумма:	168.5	Сумма:	107.5

При ранжировании объединяем две выборки в одну. Ранги присваиваются в порядке возрастания значения измеряемой величины, т.е. наименьшему рангу соответствует наименьший балл. Заметим, что в случае совпадения баллов для нескольких учеников ранг такого балла следует считать, как среднее арифметическое тех позиций, которые занимают данные баллы при их расположении в порядке возрастания. Например, 4 балла получили 3 ученика (см. таблицу). Значит, первые 3 позиции в расположении займёт балл, равный 4. Поэтому ранг для 4 баллов – это среднее

арифметическое для позиций 1, 2 и 3, или: .

Аналогично рассуждаем при вычислении ранга для балла, равного 5. Такой балл получили двое учащихся. Значит, при распределении по возрастанию первые три позиции занимает балл, равный 4, а четвёртую и пятую позиции займёт балл, равный 5. Поэтому его ранг будет равен среднему арифметическому между числами 4 и 5, т.е. 4.5.

Используя предложенный принцип ранжирования, получим таблицу рангов. Заметим, что выбор среднего арифметического в качестве ранга применяется при любом ранжировании, в том числе необходимого и для вычисления других критериев достоверности или же коэффициента корреляции Спирмена.

Чтобы использовать критерий Манна-Уитни, рассчитаем суммы рангов рассматриваемых выборок. Сумма для первой выборки равна 168,5, для второй – 107,5. Обозначим наибольшую из этих сумм через T_x (T_x=168.5). Среди объёмов n_1 и n_2 выборок наибольший обозначим n_x. Этих данных достаточно, чтобы воспользоваться формулой расчёта эмпирического значения критерия:

$$u_{эмп} = n_1 n_2 + \frac{n_x(n_x+1)}{2} - T_x$$

T_x=168,5, n_x=12>11=n_2. Тогда:

$$u_{эмп} = 11 \cdot 12 + \frac{12 \cdot 13}{2} - 168.5 = 41.5$$

Критическое значение критерия находим по специальной таблице. Пусть уровень значимости равен 0.05.

Гипотеза H_0 о незначительности различий между баллами двух классов принимается, если $u_{кр}<u_{эмп}$. В противном случае H_0 отвергается и различие определяется как существенное.

$$u_{кр}(0.05) = 35.$$
$$u_{кр} = 35 < 41.5 = u_{эмп}$$

Следовательно, различия в уровне знаний по математике среди учащихся можно считать несущественными.

Схема использования критерия Манна-Уитни выглядит следующим образом

1. Пометить данные двух выборок X и Y и представить их в виде единого упорядоченного ряда.

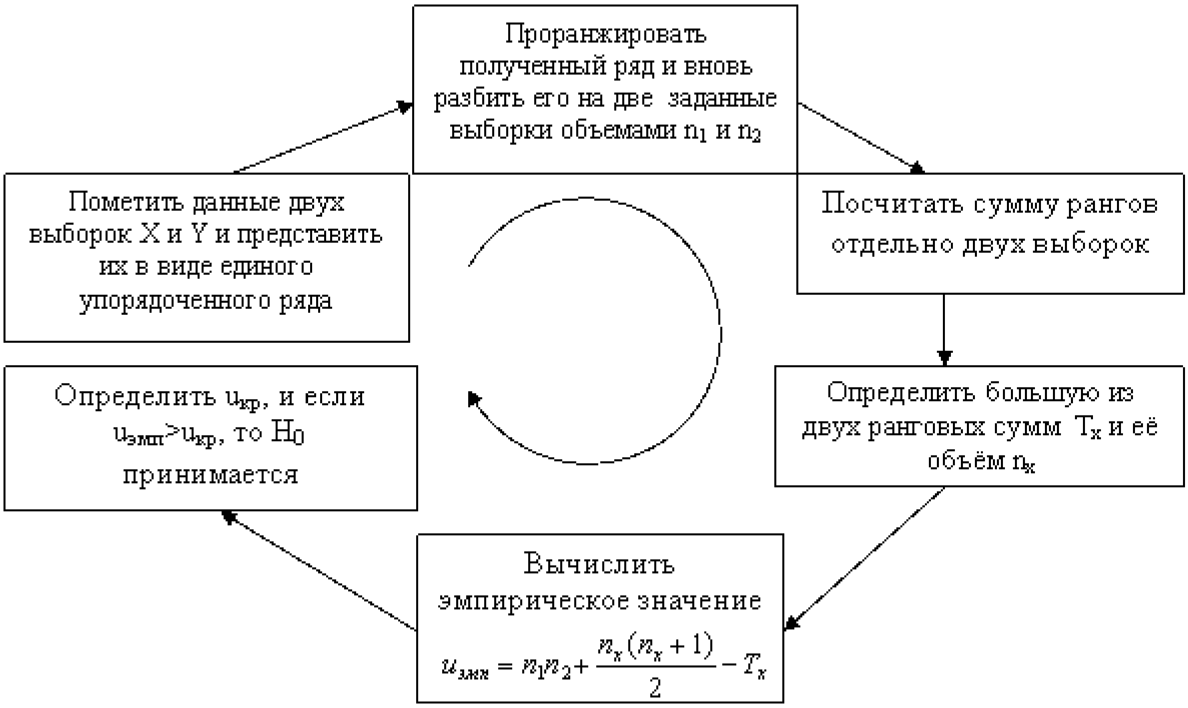
2. Посчитать сумму рангов отдельно двух выборок.

3. Определить $u_{кр}$, и если $u_{эмп}>u_{кр}$, то H_0 принимается.

4. Определить большую из двух ранговых сумм T_x и её объём n_x.

5. Вычислить эмпирическое значение.

6. Проранжировать полученный ряд и вновь разбить его на две заданные выборки объемами n_1 и n_2



Критерий Вилкоксона - один из самых известных инструментов непараметрической статистики (наряду со статистиками типа Колмогорова-Смирнова и коэффициентами ранговой корреляции).

4.2 Критерий Вилкоксона для сопряженных распределений

В экспериментальных исследованиях возникает необходимость для каждого варианта опыта иметь и вариант контроля. Часто опыт и контроль осуществляется на одних и тех же особях. Например, измеряется кровяное давление у нескольких человек в покое, а затем испытуемым предлагается определенная одинаковая физическая нагрузка, после чего вновь измеряется кровяное давление. Поскольку два показателя кровяного давления измерены у одного и того же человека, то такие варианты следует считать попарно сопряженными и в таком случае для выявления сходства или различий характера распределения 2-х совокупностей используется критерий Вилкоксона для сопряженных пар.

Пример

Воспользуемся Т-критерием для проверки полученных через Т-критерий расчетов. Рассмотрим пример, полагая, что данные попарно сопряжены, тогда

(n=10)	Показатели систолического кровяного давления			
	До начала эксперимента (Х)	В конце эксперимента (У)	Разность d	d^2
Иванов	120	130	10	100
Новиков	118	125	7	49
Сидоров	115	127	12	144
Пирогов	123	130	7	49
Агапов	130	131	1	1
Суворов	127	136	9	81
Рыжиков	124	134	10	100
Серов	127	136	9	81
Топоров	126	131	5	25
Быстров	127	131	4	16
$\sum$	$\sum x=1237$	$\sum y=1311$	$\sum d=74$	$\sum d^2=646$
Среднее	$x=123{,}7$	$y=131{,}1$	$d=7{,}4$	

Для вычисления t используем следующие формулы:

$t_d=d/S_d$

$S_d=\sqrt{S/n(n-1)}$

$S=\sum d^2-\sum d/n$

Вставляем в формулы полученные результаты:

$S=646-74^2/10= 646-5476/10=646-547{,}6=98{,}4$

$S_d=\sqrt{98{,}4/10*9}= 1{,}0455$

$t_d=7{,}4/1{,}0455=7{,}078$

Сравниваем полученные значения t_d с табличными значениями d_f (см. приложение).

Для $d_f=n - 1=10-1=9$ Из таблицы 1 (см. приложения)

имеем:

t_{st}	2,3 -	P = 0,95	или	P = 0,05
	3,3 -	P= 0,997		P = 0,003
	4,8	P = 0,999		P=0,001

Поскольку $t_d > t_{st}$ при p = 0,05 нулевая гипотеза принимается, т.е. после нагрузки систолическое давление существенно увеличивается.

Глава 5. Корреляционный анализ

Корреляционный анализ– один из методов исследования взаимосвязи между двумя или более переменными. Многие многомерные модели основаны на анализе корреляционных или связанных с ними ковариационных таблиц, построенных для множества переменных.

Термин «корреляция», буквально означающий «соотношение» или «взаимосвязь», имеет следующий смысл: корреляция есть наличие взаимной согласованности в изменчивости двух или нескольких признаков, явлений. Корреляционный анализ изучает сопряженную изменчивость двух или нескольких признаков. Работа по проведению корреляционного анализа начинается с построения корреляционных решеток. Пусть имеется выборка наблюдений (x, y) из популяции W. Причем x и y– случайные величины. Корреляционная решетка для двух переменных (признаков) представляет собой таблицу из $m \times к$ клеток, где m и $к$ – число значений признаков x и y. Значения признаков удобно располагать в возрастающем порядке слева направо для x и сверху вниз для y. В клетках таблицы производят разноску сопряженных частот f_{xy} (число наблюдений со значениями признаков х и у) в зависимости от значений двух признаков одновременно.

Схема корреляционной решетки

x У	X_1	X_2	X_3	…	X_k	f_y
Y_1 Y_2 Y_3 … y_m	F_{11} F_{21} F_{31} F_{m1}	F_{12} F_{22} F_{32} F_{m2}	F_{13} F_{23} F_{33} F_{m3}	… … … …	F_{1k} F_{2k} F_{3k} F_{mk}	F_{y1} F_{y2} F_{y3} F_{ym}
f_x	F_{x1}	F_{x2}	F_{x3}	…	F_{xk}	n

f_x , ,f_y– частоты вариационных рядов, f_{ij}– сопряженные частоты, n– объем выборки, x, y– значения признаков:

$$f_{xj} = \sum_{i=1}^{m} f_{ij}, \quad j = 1 \div k$$

$$f_{yi} = \sum_{j=1}^{k} f_{ij}, \quad i = 1 \div m$$

$$n = \sum_{j=1}^{k} f_{xj} = \sum_{i=1}^{m} f_{yi}$$

Для расчета коэффициента корреляции необходимо предварительно рассчитать некоторые статистики (средние значения, дисперсию, среднеквадратичное отклонение, корреляционный момент) по следующим формулам:

$$\bar{x} = \frac{\sum_{j=1}^{k} f_{xj} \cdot x_j}{n}, \qquad \bar{y} = \frac{\sum_{i=1}^{m} f_{yi} \cdot y_i}{n}$$ – средние

$$\begin{cases} D_x^2 = \dfrac{(\sum_{j=1}^{k} f_{xj} x_j^2 - \frac{1}{n}(\sum_{j=1}^{k} f_{xj} \cdot x_j)^2)}{n-1}, \\ D_y^2 = \dfrac{(\sum_{i=1}^{m} f_{yi} \cdot y_i^2 - \frac{1}{n}(\sum_{i=1}^{m} f_{yi} \cdot y_i)^2)}{n-1} \end{cases}$$ – дисперсии

$$\begin{cases} \sigma_x = \sqrt{D_x^2}, \\ \sigma_y = \sqrt{D_y^2}, \end{cases}$$ – среднеквадратичные отклонения

$$m_{xy} = \frac{1}{n-1}(\sum_{i=1}^{m}\sum_{j=1}^{k} f_{ij} \cdot x_j \cdot y_i - \frac{1}{n}(\sum_{j=1}^{k} f_{xj} \cdot x_j \cdot \sum_{i=1}^{m} f_{yi} \cdot y_i)),$$

m_{xy} – корреляционный момент.

Для удобства расчетов рекомендуется заполнить табл.:

Таблица для расчета коэффициента линейной корреляции

X y	X_1	X_2	…	$X_к$	f_{yi}	$f_{yi}y_i$	$\sum_{j=1}^{k} f_{ij} \cdot x_j$	$Y^2_i * f_{yi}$	$y_i \sum_{j=1}^{k} f_{ij} x_j$
1	2	3	4	5	6	7	8	9	10
Y_1	f_{11}	f_{12}	…	f_{1k}					
Y_2	f_{21}	f_{22}	…	f_{2k}					
…	…	…	…	…					
y_m	f_{m1}	f_{m2}	…	f_{mk}					
f_{xj}									
$f_{xj} \cdot x_j$									
1	2	3	4	5	6	7	8	9	10
$\sum_{i=1}^{m} f_{ij}$									
$f_{xj} \cdot x^2_j$									
$x_j \sum_{i=1}^{m}$									

Коэффициент линейной корреляции рассчитывается по формулам:

$$r = \frac{m_{xy}}{\sigma_x \sigma_y}$$ или $$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} f_{ij}(x_j - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{k}(x_j - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}}.$$

Вычисления по второй формуле более громоздки, т.к. отклонения от средней находятся для каждой сопряженной частоты.

Значение коэффициента корреляции заключено в пределах от -1 до 1. Положительное значение коэффициента указывает, что *У* имеет тенденцию возрастать совместно с *X*, отрицательное наоборот – *У* уменьшается с возрастанием *X*. Экстремальное значение (-1 или +1) соответствует полной линейной зависимости между *X* и *У*. Чем ближе коэффициент корреляции к 1, тем ближе зависимость к линейной.

ЗАКЛЮЧЕНИЕ

В заключении необходимо высказать следующее замечание. Использование того или иного метода математической статистики определяется, во-первых, тем, что каждый из методов имеет определенные границы применимости. Использование же методов вне границ их возможностей приводит к ошибочным выводам. Поэтому важным моментом при обработке экспериментального материала является выбор адекватной математической модели, а для этого экспериментатору необходимо ясно представлять, с одной стороны, цели и задачи исследования, а, с другой стороны, смысл применяемых методов математической обработки.

В настоящем учебно-методическом пособии изложены базовые методы статистического анализа биологического экспериментального материала. Конечно, приведенные методы анализа не исчерпывают всего математического арсенала, который может быть использован при статистическом анализе данных. Так, в биологии часто используются частные коэффициенты корреляции и регрессии, ранговые коэффициенты корреляции и другие. В непараметрической статистике, кроме метода х2, метода Колмогорова-Смирнова и метода Вилкоксона также существуют и другие методы.

Однако овладение основными, наиболее используемыми методами статистической обработки экспериментального материала, является обязательным условием для освоения других менее общих методов. Ниже приводится список книг по математической и биологической статистике, в которых более полно представлены методы статистической обработки экспериментального материала.

Список использованной литературы:

1. Васильева, Л.А. Статистические методы в биологии, медицине и сельском хозяйстве: учеб. пособие / Л.А. Васильева. – Новосибирск.: Институт цитологии и генетики СО РАН, 2007.-124 с.
2. Красильников, В.В. Высшая математика. Вероятность. Статистика. Исследование операций: учеб. пособие / В.В. Красильников. - Набережные Челны.: Печатный двор, 1996. - 225 с.
3. Петри, А. Наглядная статистика в медицине / А. Петри, К. Сэбин - М. : Издательский дом Гэотар-мед, 2003. - 143 с.
4. Тукшаитов, Р.Х. Основы оптимального представления статистических показателей на графиках, диаграммах и в таблицах. / Р.Х. Тукшаитов. - К.: Типография КГЭУ, 2006. - 227 с.
5. Википедия – универсальная энциклопедия / URL: http://ru.wikipedia.org/wiki/T-критерий_Стьюдента (дата обращения: 20.02.13)
6. Википедия – универсальная энциклопедия / http://ru.wikipedia.org/wiki/критерий_Уилкоксона (дата обращения: 20.02.13)

Приложение

Таблица 1 Стандартные значения критерия Стьюдента

	t_1 0,95	t_1 0,99	t_1 0,999		t_1 0,95	t_1 0,99	t_1 0,999
1	12.7	63.7	636.6	13	2.2	3.0	4.2
2	4.3	9.9	31.6	14-15	2.1	3.0	4.1
3	3.2	5.8	12.9	16-17	2.1	2.9	4.0
4	2.8	4.6	8.6	18-20	2.1	2.9	3.9
5	2.6	4.0	6.9	21-24	2.1	2.8	3.8
6	2.4	3.7	6.0	25-28	2.1	2.8	3.7
7	2.4	3.5	5.4	29-30	2.0	2.8	3.7
8	2.3	3.4	5.0	31-34	2.0	2.7	3.7
9	2.3	3.3	4.8	35-42	2.0	2.7	3.6
10	2.2	3.2	4.6	43-62	2.0	2.7	3.5
11	2.2	3.1	4.4	63-175	2.0	2.6	3.4
12	2.2	3.1	4.3	176 и более	2.0	2.6	3.3