И.М. Михалевич
С.П. Примина

Применение математических методов при анализе геологической информации (с использованием компьютерных технологий)

часть II

Министерство образования Российской Федерации

Государственное образовательное учреждение
высшего профессионального образования
ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

И.М. Михалевич
С.П. Примина

Применение математических методов при анализе геологической информации (с использованием компьютерных технологий)

Учебное пособие

ЧАСТЬ II

Иркутск, 2004

ББК 26.3:22.1
УДК 518.+550 (07)

**Рецензенты: канд.физ.-мат. наук, доц. А.В. Диогенов ;
канд. геол.-минерал.наук , доц. В.А. Бычинский**

Применение математических методов при анализе геологической информации (с использованием компьютерных технологий)/ Сост. И.М. Михалевич, С.П. Примина: Учеб. пособие. Ч.II. – Иркутск: Иркут. ун-т, 2004. – 120с.

Вторая часть учебного пособия является продолжением пособия, вышедшего в 2001 г./1/, и предназначена для дальнейшего изучения и применения распространенных математических методов при анализе данных, полученных при геолого-разведочных работах. Описание количественных методов сопровождается примерами и решением их с помощью известных компьютерных технологий (Excel, Biostat, Statistica).

Рассчитано на студентов геологических специальностей и других факультетов с естественным уклоном, также может быть использовано аспирантами, научными сотрудниками и практическими геологами.

Библиогр. 20 назв. Ил. 40. Табл. 29.

ВМЕСТО ВВЕДЕНИЯ

В первом выпуске учебного пособия по применению математических методов при анализе геоинформации /1/ были рассмотрены статистические приемы для работы с количественными данными и, прежде чем мы перейдем к рассмотрению других, более сложных статистических процедур, мы считаем необходимым показать приемы анализа качественных признаков, “сплошь и рядом” используемых в геологии.

Здесь, вместо введения, которое можно посмотреть в /1/, нам хотелось бы показать природу различных типов данных (в том числе и количественных), природу шкал, в которых они могут измеряться. Этот вопрос довольно хорошо рассмотрен в работе /2/.

Факторы и признаки

Значение фактора (переменной) описывает определенную характеристику в виде некоторого численного значения, причем это значение может меняться от объекта к объекту или для одного и того же объекта во времени.

Примеры – пористость в процентах, проницаемость в миллидарси.

Описательный признак – это некоторая категория (значение) характеристики, к которой объект принадлежит или не принадлежит, либо свойство или качество, которым объект обладает или не обладает.

Примеры – обеспеченность компьютерной техникой, категория запасов полезных ископаемых, высшее образование.

Некоторые характеристики можно выразить только одним способом, тогда как другие допускают представление обоими способами. Например, массу объекта можно рассматривать и как фактор (в килограммах), и как признак (есть “маленькая масса” объекта / нет “маленькой массы” объекта). При выборе формы представления важно учитывать причины проведения измерения, требования объективности, надежности и состоятельности, а также – свойства различных измерительных шкал. Эти соображения представлены ниже.

Непрерывные и дискретные факторы

Непрерывным называется фактор, который может принимать бесконечное число значений в любом интервале. Он допускает как сложение, так и деление значений и может измеряться с различной степенью точности при использовании более или менее совершенных методов измерения.

Примеры – длина (в метрах): 10,83; 154,74; масса объекта (в килограммах): 48,7; 970,65.

Дискретный фактор может иметь лишь конечное число значений в любом интервале. Эти значения обычно (но не всегда) целые числа.

Примеры – число анализов легкой фракции, число скважин глубокого бурения на площади.

Измерительные шкалы

Измерение – это процесс определения числового значения, противопоставленного сумме, степени, протяженности, размеру, количеству и т. п. для данного типа объектов в соответствии с определенными правилами. Последние устанавливаются по шкале измерений. Не все свойства объектов бывают изоморфны свойствам чисел, но тем не менее можно подобрать эмпирические операции для определения соответствия качеств объектов или для их упорядочения, а также для установления равенства между определенными свойствами объектов. Эти операции приводят к выбору четырех шкал измерения, которые рассматриваются ниже /2/.

Номинальная шкала

Номинальная шкала применяется для классификации объектов по признаку равенства их свойств. В данном случае безразлично, что определяют заданные классы – названия или числа. Так, цвет сланцев можно закодировать следующим образом: красный = 1, черный = 2, серый = 3, зеленый = 4. Такой набор чисел уже можно применять для различных целей геологических исследований. В данном случае нет необходимости придерживаться определенного порядка в расположении кодируемых понятий, так как безразлично, каким номером будет обозначен тот или иной класс. Например, группировку сланцев по цвету можно провести иначе: серый = 1, зеленый = 2, красный = 3, черный = 4. С числами номинальной шкалы можно проводить некоторые арифметические операции, такие, как подсчет числа индивидов, принадлежащих каждому классу, и выявление класса, соответствующего максимальному их количеству. Число индивидов каждого класса можно выразить в процентах от их общего количества.

Порядковая шкала

В тех случаях, когда объекты можно расположить в некотором порядке в зависимости от изменения какого-либо свойства, обычно применяется порядковая шкала. Упорядоченные классы обозначаются числами в последовательности от 1 до N, из которых каждое указывает, что изучаемое свойство объектов проявлено сильнее (слабее), чем в предыдущих классах. Порядковая шкала широко применяется в геологии. Классическими примерами такого применения может служить шкала твердости минералов и расположение слоев в толщах от наиболее древних до самых молодых.

Последовательное расположение чисел на порядковой шкале совсем не означает, что изучаемое свойство меняется равномерно; так, "длина шага", соответствующая переходу от 1 до 2, не обязательно должна быть равной "длине" перехода от 2 до 3. Например, число лет, соответствующее геологическим периодам в их упорядоченной последовательности, не обязательно будет одинаковым.

Число индивидов, попавших в каждый из классов порядковой шкалы, как и в случае номинальной шкалы, можно подсчитать с тем, чтобы выявить наиболее часто наблюдаемый класс. Кроме того, если все изучаемые индивиды можно расположить в упорядоченную последовательность, не прибегая к группировке в классы, то нетрудно определить и медиану, т. е. порядковый номер, соответствующий объекту, расположенному в середине последовательности, а также и другие квантили (медиана – 50-процентный квантиль). Хотя расположение объектов в последовательность от 1 до *N* в соответствии с изменениями некоторого количества и отличается от их группировки, например в четыре упорядоченных класса, обе эти процедуры являются однотипными операциями упорядочения. Предположим, что 11 образцов сланцев сгруппированы следующим образом: белые поставлены на первое место, светло-серые – на второе, темно-серые – на третье, черные – на четвертое. Можно поступить иначе, расположив все 11 образцов, начиная с самого светлого и кончая самым темным; тогда в результате на порядковой шкале будет получено 11 номеров вместо 4. .

Интервальная шкала

Интервальная шкала применяется в тех случаях, когда "длина шага", соответствующая переходу от одного класса к другому, эквивалентна длине интервала и может быть задана, но без указания точки абсолютного нуля. Так, измерение температуры проводится на интервальной шкале. В данном случае равенство единиц измерительной шкалы основано на равенстве приращений ртутного столбика по мере повышения температуры, хотя нулевую точку можно выбрать произвольно, например, по Фаренгейту или Цельсию.

Шкала отношений

Четвертая шкала измерений применяется в тех случаях, когда можно установить равенство отношений применительно к измеряемому количеству. Она называется *шкалой отношений,* или *относительной шкалой.* Такая шкала требует точного или по крайней мере подразумеваемого указания положения нулевой точки.

Примерами чисел относительной шкалы могут служить значения длины и массы.

Числа 1, 2, . . ., *N*, указывающие расположение объектов на порядковой шкале, а также результаты подсчета количеств объектов каждого типа, выраженные в процентах, являются числами шкалы отношений. Точно так же разность между двумя значениями интервальной шкалы представляет собой число шкалы отношений.

Требование равенства отношений заключается в том, что любое число X на относительной шкале можно заменить другим числом X', связанным с первым следующим соотношением: $X' = aX$ при $a > 0$. Так, переход от дюймов к сантиметрам осуществляется с помощью равенства X' *см* = **2,51** X дюймов, где отношение $X' / X = 2{,}51$ является константой. Такая зависимость не выполняется в случае перехода от температуры, измеренной по Фаренгейту, к соответствующему значению на шкале Цельсия.

Очень многие показатели, обычно применяемые в геологических исследованиях, выражаются числами относительной шкалы. Замеры мощности слоев, углов простирания и падения, подсчет числа слоев в стратиграфической единице, а также определение отметки слоя над уровнем моря – все это примеры чисел относительной шкалы. Точно так же измерения размеров частиц, характеристик их формы и ориентации, коэффициентов пористости и проницаемости породы бывают выражены на шкале отношений. Скорость течения реки, высота волны и ее период, а также глубина воды – примеры чисел относительной шкалы. Эта шкала самая многогранная из всех четырех шкал. Именно поэтому задача получения в результате измерения объективных данных сводится к их выражению с помощью шкалы отношений.

В таблице, приведенной ниже /2/, даны некоторые свойства четырех рассмотренных шкал с соответствующими геологическими примерами. В ней содержатся сведения о математической групповой структуре чисел каждой шкалы, а также перечень достаточных статистик.

Количественные и качественные данные

Прежде, чем мы перейдем к дальнейшему изложению, еще раз напомним, что данные можно разделить на две больших категории.

Качественные (категоризованные) данные – это измерения, для которых количественных значений нет, либо они скрыты. Такие переменные измеряются в номинальной шкале или в порядковой шкале. Эти данные называют еще атрибутивными или качественными.

Количественные данные имеют численные значения. Они измеряются в шкале интервалов или в шкале отношений.

Таблица

Шкала измерений

Шкала	**Основные операции**	**Примеры из стратиграфии и седиментологии**	**Примеры применяемых статистик**
Номинальная	Определение сходства свойств	Классификация осадочных пород на два класса или более; применяется любая группировка, основанная на наличии или отсутствии эквивалентности свойств	Частота для каждого класса; модальный класс (класс с наибольшей частотой)
Порядковая	Определение различия в свойствах по принципу «больше или меньше»	Классификация объектов в порядке увеличения (или уменьшения) некоторого свойства: например зерен – от угловатых до окатанных или песков – от плотных до рыхлых	Медиана, квантили
Интервальная	Определение принадлежности значения к данному интервалу	Определение эквипотенциала песка с помощью произвольной системы параллельных линий на электрокаротажной диаграмме	Среднее арифметическое, оценки стандартного отклонения, смешанных моментов и коэффициентов корреляции
Относительная	Определение равенства отношений	Измерение мощности четко определенных стратиграфических единиц и проницаемости песчаников	Все предыдущие характеристики, а также среднее геометрическое и коэффициент вариации

1. СТАТИСТИКИ, ГИПОТЕЗЫ, КРИТЕРИИ

1.1. Статистический анализ качественных признаков

Статистические процедуры, с которыми мы знакомились ранее /1/, в основном предназначены для анализа количественных признаков.

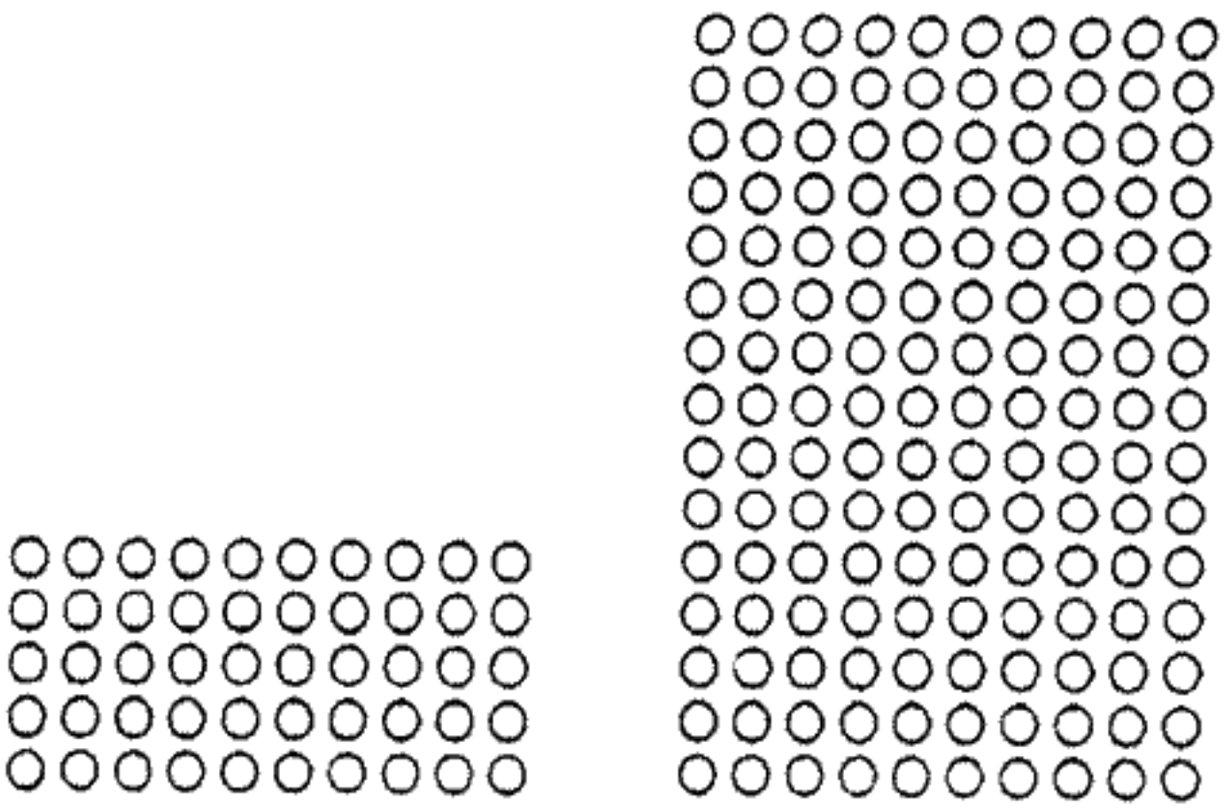
Как мы уже знаем из предисловия к этому выпуску, многие признаки невозможно измерить числом. Например, можно быть либо мужчиной, либо женщиной, либо мертвым, либо живым. Можно быть врачом, юристом, геологом, и так далее. Здесь мы имеем дело с качественными признаками. Эти признаки не связаны между собой никакими арифметическими соотношениями, упорядочить их нельзя. Единственный способ описания качественных признаков состоит в том, чтобы подсчитать *число* объектов, имеющих одно и то же значение. Кроме того, можно подсчитать, какая *доля* от общего числа объектов приходится на то или иное значение.

Одним из основных вопросов при работе с качественными признаками является вопрос подсчета долей, нахождения способа оценить точность, с которой доли, вычисленные по выборкам, соответствуют долям во всей совокупности. Здесь нам снова пригодится пример с марсианами /3/.

Вспомним, что экспедиция побывала на Марсе, где были измерены все его обитатели. Хотя ранее не говорилось об этом, но больше всего членов экспедиции поразило различие в пигментации марсиан: 50 марсиан были розового, а остальные 150 – зеленого цвета (рис. 1.1) .

Как описать совокупность марсиан по этому признаку? Ясно, что нужно указать долю, которую составляют марсиане каждого цвета во всей совокупности марсиан. В нашем случае доля розовых марсиан

$p_{роз.} = 50/200 = 0{,}25$ и зеленых $p_{зел.} = 150/200 = 0{,}75$.



Розовые Зеленые

Рис.1.1. Из 200 марсиан 150 имеют зеленую окраску, остальные 50 розовые. Если наугад извлечь марсианина, то вероятность, что он окажется розовым, составляет 50/200 = 0,25, то есть 25%.

Поскольку марсиане бывают только розовые и зеленые, справедливо тождество $p_{роз.} + p_{зел.} = 1$, или, $p_{зел.} = 1 - p_{роз.}$. Таким образом, для характеристики совокупности, которая состоит из двух классов, достаточно указать численность одного из них: если доля одного класса во всей совокупности равна p, то доля другого равна 1 – p. Заметим, что $p_{роз.}$ есть еще и вероятность того, что случайно выбранный марсианин окажется розовым.

Покажем, что доля p в некотором смысле аналогична среднему μ по совокупности. Введем числовой признак X, который принимает только два значения: 1 для розового цвета и 0 для зеленого. Среднее значение признака X равно:

$$\mu = \frac{\sum X}{N} = \frac{1+1+...+1+0+0+...+0}{200} = \frac{50*1+150*0}{200} = \frac{50}{200} = 0{,}25.$$

Как видим, полученное значение совпадает с долей розовых марсиан.

Повторим это рассуждение для общего случая. Пусть имеется совокупность из N членов. При этом M членов обладают каким-то качественным признаком, которого нет у остальных $N - M$ членов. Введем числовой признак X: у членов совокупности, обладающих качественным признаком, он будет равен 1, а у членов, не обладающих этим признаком, он будет равен 0. Тогда среднее значение X равно

$$\mu = \frac{\sum X}{N} = \frac{M*1+(N-M)*0}{N} = \frac{M}{N} = \mathrm{p},$$

то есть доле членов совокупности, обладающих качественным признаком.

Используя такой подход, легко рассчитать и показатель разброса – стандартное отклонение. Не совсем ясно, однако, что понимать под разбросом, если значений признака всего два: 0 и 1. На рис. 1.2 мы изобразили три совокупности по 200 членов в каждой. В первой из них (1.2,А) все члены принадлежат к одному классу. Разброс равен нулю. На рис. 1.2,Б разброс уже имеется, но он невелик. На рис. 1.2,В совокупность делится на два равных класса. В этом случае разброс максимален.

Итак, найдем стандартное отклонение. По определению оно равно

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}},$$

где для M членов совокупности значение $X = 1$, а для остальных N -M членов $X = 0$. Величина $\mu = p$.

Пропуская преобразования, приведенные в работе /3/, покажем, что стандартное отклонение для качественных признаков имеет вид:

$$\sigma = \sqrt{p(1-p)}.$$

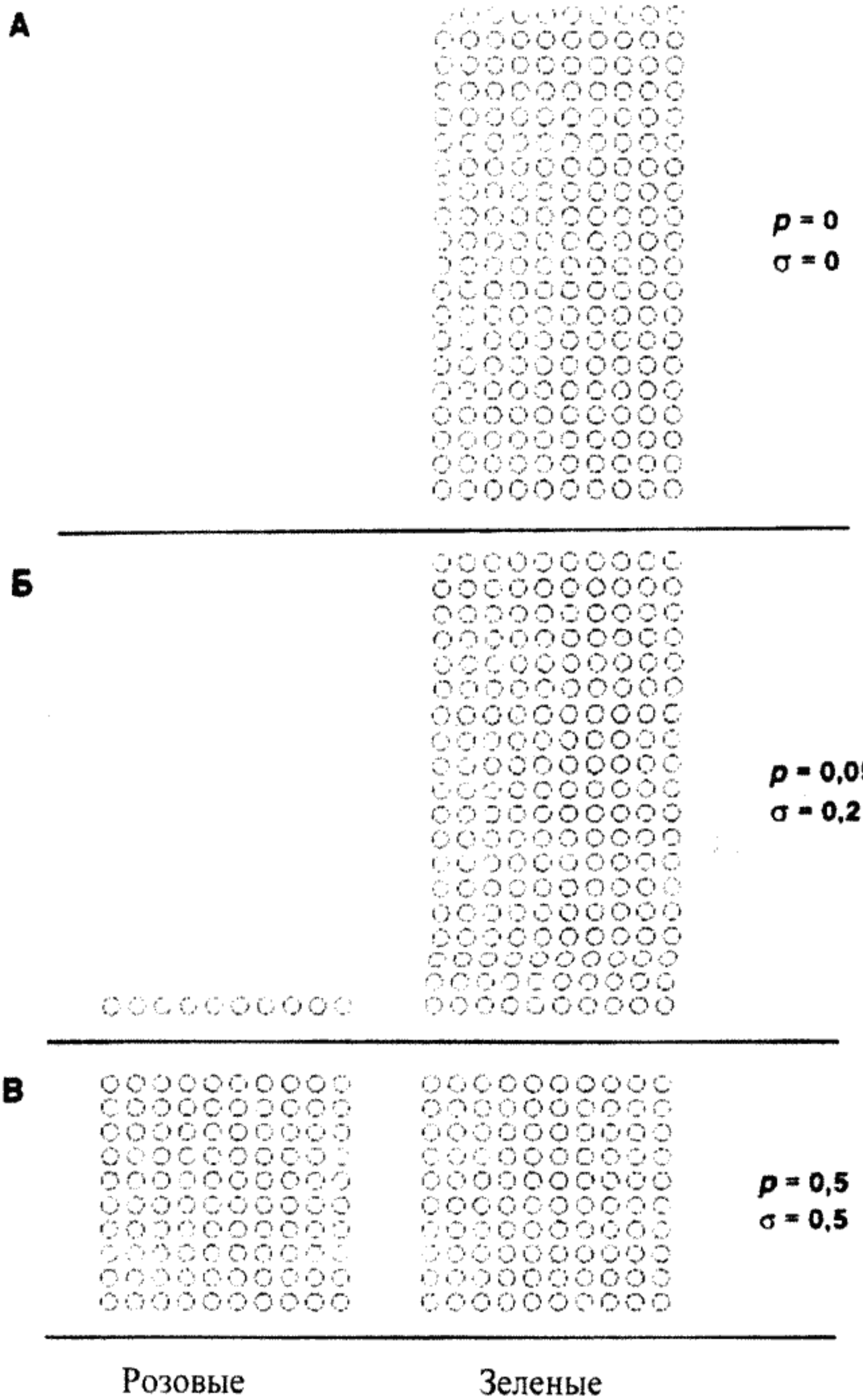


Рис. 1.2. Что такое разброс данных, если значений признака всего два? Возможно, это станет яснее, если вспомнить, что разброс – это отсутствие единства. Рассмотрим три совокупности из 200 марсиан. А. Все марсиане зеленые. Царит полное единство, разброс отсутствует, $\sigma = 0$. Б. Среди стройных рядов зеленых марсиан появилось 10 розовых. Единство немного нарушено, появился некоторый разброс, $\sigma = 0{,}2$. В. От единства марсиан не осталось и следа: они разделились поровну на зеленых и розовых. Разброс максимален, $\sigma = 0{,}5$.

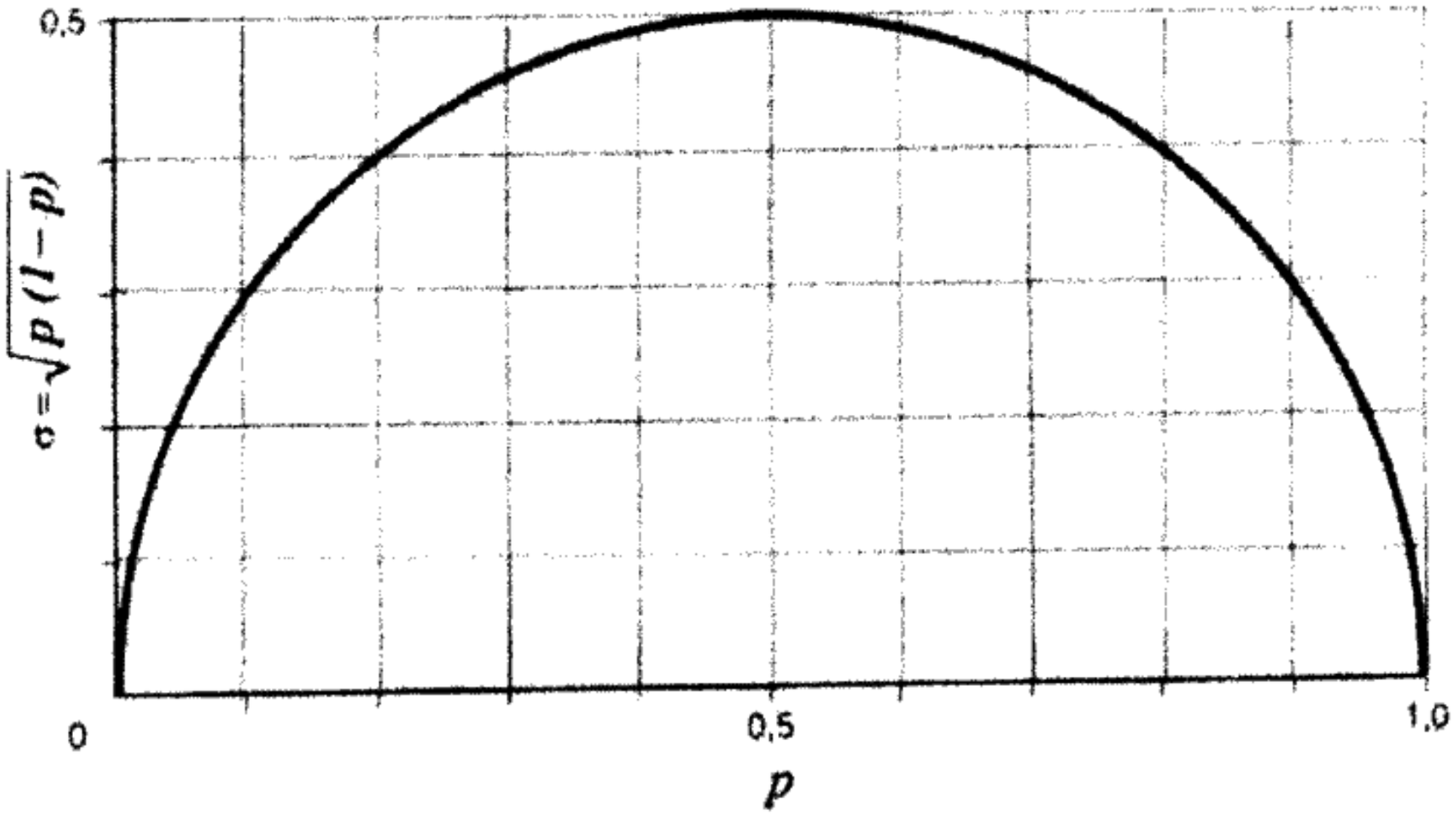


Рис. 1.3. Стандартное отклонение доли σ полностью определяется самой этой долей p. Когда доля равна 0 или 1, разброс отсутствует и σ = 0. Когда p = 0,5, разброс максимален, σ = 0,5.

Найденное стандартное отклонение σ полностью определяется величиной p. Этим оно принципиально отличается от стандартного отклонения для нормального распределения, которое не зависит от σ. На рис. 1.3 показана зависимость σ от p. Она вполне согласуется с теми впечатлениями, которые возникают при рассмотрении рис.1.3: стандартное отклонение достигает максимума при p = 0,5 и равно 0, когда p равно 0 или 1.

Зная стандартное отклонение σ, можно найти стандартную ошибку для выборочной оценки p. Посмотрим, как это делается.

1.1.1. Точность оценки долей

Если бы у нас были данные по всем членам совокупности, то не было бы никаких проблем, связанных с точностью оценок. Однако всегда приходится довольствоваться ограниченной выборкой. Поэтому возникает вопрос, насколько точно доли в выборке соответствуют долям в совокупности. Проделаем мысленный эксперимент наподобие того, который был проведен в работе /3/, когда рассматривали, насколько хорошей оценкой среднего по совокупности является выборочное среднее.

Предположим, что из всех 200 марсиан случайным образом выбрали 10. Распределение розовых и зеленых марсиан во всей совокупности, неизвестное исследователям, изображено в верхней части рис. 1.4. Закрашенные кружки соответствуют марсианам, попавшим в выборку. В нижней части рис. 1.4 показана информация, которой располагал бы исследователь, получивший такую выборку. Как видим, в выборке розовые и зеленые марсиане поделились поровну. Основываясь на этих данных, мы решили бы, что розовых марсиан столько же,

сколько и зеленых, то есть их доля составляет 50%.

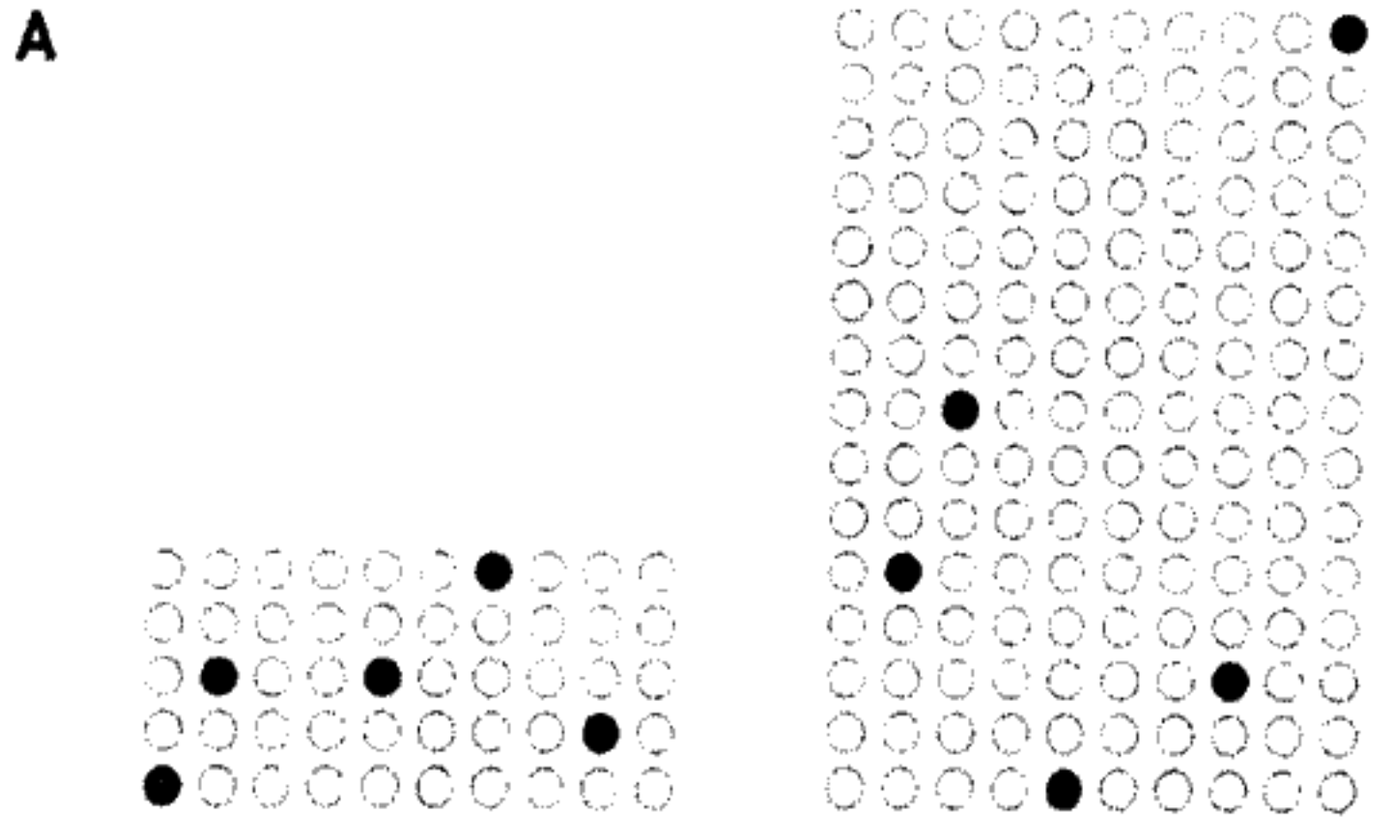


Рис. 1.4. А – из совокупности марсиан, среди которых 150 зеленых и 50 розовых, извлекли случайную выборку из 10 особей. В выборку попало 5 зеленых и 5 розовых марсиан, на рисунке они помечены черным. Б – в таком виде данные предстанут перед исследователем, который не может наблюдать всю совокупность и вынужден судить о ней по выборке. Оценка доли розовых марсиан p= 5 /10 = 0,5

Исследователь мог бы извлечь другую выборку, например одну из представленных на рис.1.5. Здесь выборочные доли розовых марсиан равны 30, 30, 10 и 20%. Как любая выборочная оценка, оценка доли (обозначим ее $\hat{p}$) отражает долю p в совокупности, но отклоняется от нее в силу случайности. Рассмотрим теперь не совокупность марсиан, а совокупность всех значений p, вычисленных по выборкам объемом 10 каждая. (Из совокупности в 200 членов можно получить более 10^{16} таких выборок.). На рис.1.6 приведены пять значений p, вычисленных по пяти выборкам с рис.1.4 и 1.5, и еще 20 значений, полученных на других случайных выборках того же объема. Среднее этих 25 значений составляет 30%. Это близко к истинной доле розовых марсиан 25%. По аналогии со стандартной ошибкой среднего найдем стандартную ошибку доли. Для этого нужно охарактеризовать разброс выборочных оценок доли, то есть рассчитать стандартное отклонение совокупности p. В данном случае оно равно примерно 14%; в общем случае $\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$,

где $\sigma_{\hat{p}}$ - стандартная ошибка доли, σ - стандартное отклонение, n – объем выборки. Поскольку $\sigma = \sqrt{p(1-p)}$, то

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Заменив в приведенной формуле истинное значение доли ее оценкой $\hat{p}$, получим оценку стандартной ошибки доли:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \qquad (1.1)$$

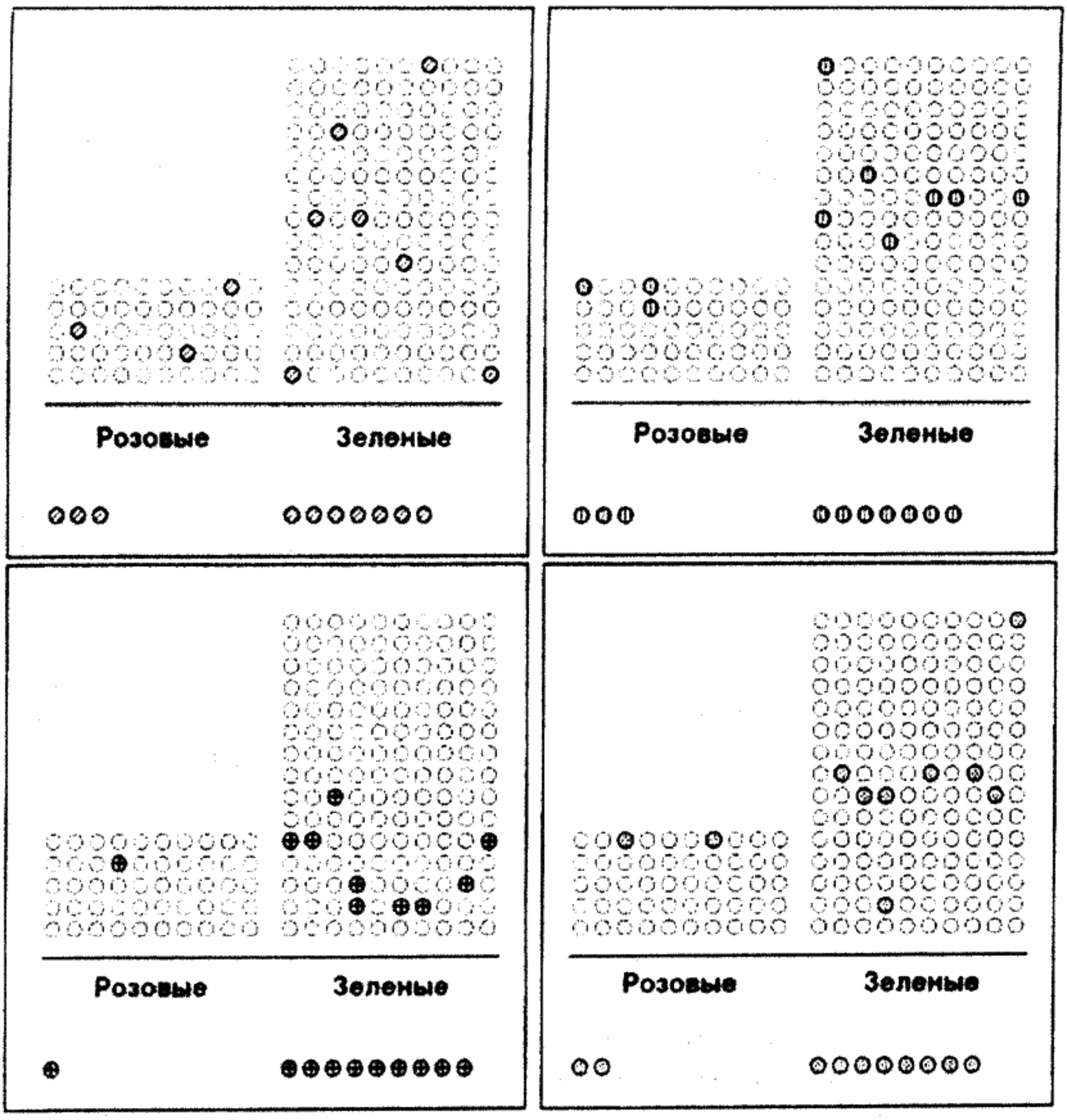


Рис.1.5. Еще 4 случайные выборки из совокупности марсиан. Оценки доли розовых: 30,30,10 и 20

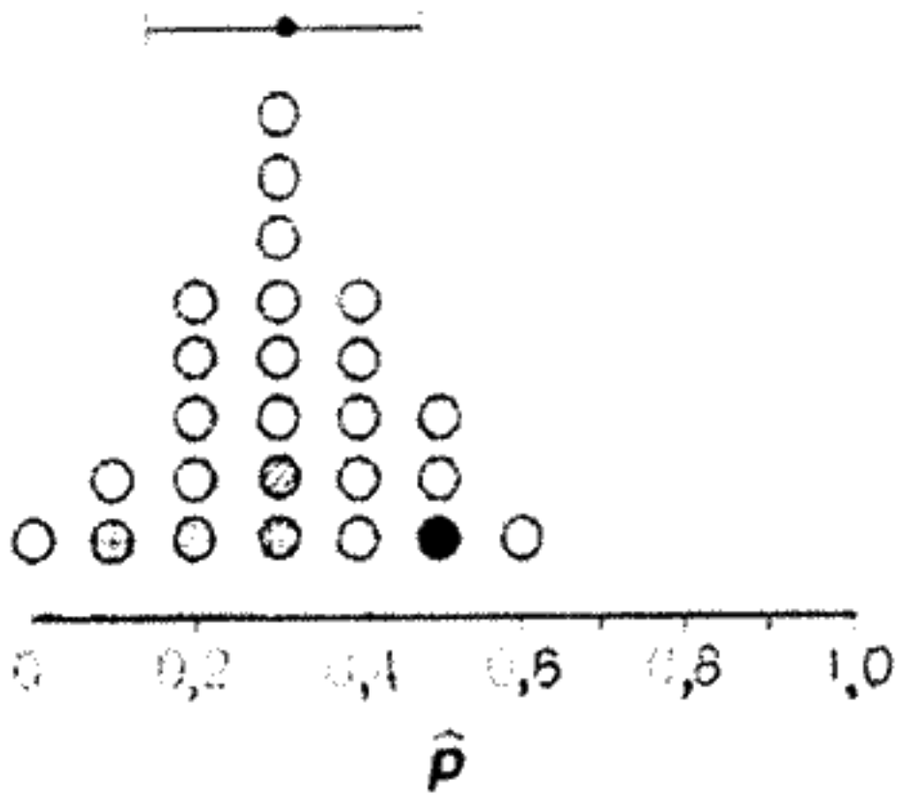


Рис.1.6. Нанесем на график оценки доли розовых марсиан, полученные по выборке с рис. 1.4 и четырем выборкам с рис.1.5. Добавим к ним еще 20 выборочных оценок. Получилось распределение выборочных оценок *p*. Стандартное отклонение совокупности средних — это стандартная ошибка доли

Из центральной предельной теоремы /1,3/ вытекает, что при достаточно большом объеме выборки выборочная оценка $\hat{P}$ приближенно подчиняется нормальному распределению, имеющему среднее *p* и стандартное отклонение $\sigma_{\hat{p}}$. Однако при значениях *p*, близких к 0 или 1, и при малом объеме выборки это не так. При какой численности выборки можно пользоваться приведенным способом оценки? Статистика утверждает, что нормальное распределение служит хорошим приближением, если $n \geq 30$ /1/ или, если и $n \cdot \hat{p}$, и $n \cdot (1-\hat{p})$ более 5 /3/.

Прежде чем "двигаться" дальше, перечислим те предпосылки, на которых основан излагаемый подход. Мы изучаем то, что в статистике принято называть независимыми испытаниями Бернулли. Эти испытания обладают следующими свойствами:

- Каждое отдельное испытание имеет ровно два возможных взаимоисключающих исхода.
- Вероятность данного исхода одна и та же в любом испытании.
- Все испытания независимы друг от друга.

В терминах совокупности и выборок эти свойства формулируются так:

- Каждый член совокупности принадлежит одному из двух классов.
- Доля членов совокупности, принадлежащих одному классу, неизменна.
- Каждый член выборки извлекается из совокупности независимо от остальных.

1.1.2. Сравнение долей

Критерий Стьюдента вычисляется на основе выборочных средних и стандартной ошибки:

$$t = \frac{\textit{разность выборочных средних}}{\textit{стандартная ошибка разности выборочных средних}}.$$

Выборочная доля $\bar{p}$ аналогична выборочному среднему. Выражение для стандартной ошибки показано в формуле (1.1). Теперь можно перейти к задаче сравнения долей, то есть к проверке нулевой гипотезы о равенстве долей. Для этого используется критерий *z*, аналогичный критерию Стьюдента *t*:

$$z = \frac{\textit{разность выборочных долей}}{\textit{стандартная ошибка разности выборочных долей}}.$$

Здесь мы пропустим выкладки, довольно подробно изложенные в /3/, и приведем критерий z для долей:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

О статистически значимом различии долей можно говорить, если значение *z* окажется «большим». С такой же ситуацией мы имели дело, рассматривая критерий Стьюдента. Отличие состоит в том, что *t* подчиняется распределению Стьюдента, а z – стандартному нормальному распределению. Соответственно, для нахождения «больших» значений *z* нужно воспользоваться стандартным нормальным распределением /1,3/. Однако, поскольку при увеличении числа степеней свободы распределение Стьюдента стремится к нормальному, критические значения *z* можно найти в последней строке критических значений *t* */1,3/.Для 5% уровня значимости оно составляет 1,96, для 1% – 2,58.*

1.1.3. Поправка Йейтса на непрерывность

Нормальное распределение служит лишь приближением для распределения *z*. При этом оценка *p* оказывается заниженной и нулевая гипотеза будет отвергаться слишком часто. Причина состоит в том, что *z* принимает только дискретные значения, тогда как приближающее его нормальное распределение непрерывно. Для компенсации излишнего «оптимизма» критерия *z* введена *поправка Йейтса,* называемая также поправкой на непрерывность. С учетом этой поправки выражение для z имеет следующий вид:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Поправка Йейтса уменьшает значение *z,* уменьшая тем самым расхождение с нормальным распределением.Рассмотренный выше метод хорошо работает, если качественный признак, который нас интересует, принимает два значения (марсианин зеленый или розовый).

1.2. Таблицы сопряженности: Критерий χ^2

Более того, поскольку метод является прямым аналогом критерия Стьюдента, число сравниваемых выборок также должно быть равно двум. Понятно, что и число значений признака, и число выборок может оказаться больше двух. Для анализа таких случаев нужен иной метод. С виду этот метод, который здесь будет изложен, сильно отличается от критерия *z,* но на самом деле между ними много общего.

Рассмотрим работу критерия на примере связи формы галек из отложений маршельской морены с их составом / 2 /. Данные занесены в таблицу 1.1. Проверим гипотезу о независимости форм галек из отложений маршельской морены от состава пород. Большая выборка галек из отложений ледниковой морены разделена по форме на угловатые и окатанные, а по составу на гранитные и метаморфические.

Таблица 1.1

Связь формы галек из отложений маршельской морены с их составом

Состав пород	Угловатые	Окатанные
Гранитные	41	170
Метаморфические	14	42

Для каждой из групп указано число угловатых и окатанных галек. У нас два признака: состав пород (гранитные – метаморфические) и форма галек (угловатые – окатанные); в табл.1.1 указаны все их возможные сочетания, поэтому такая таблица называется таблицей сопряженности. В данном случае размер таблицы 2x2.

Теперь посмотрим на табл. 1.2. Это таблица *ожидаемых* чисел, которые мы получили бы, если бы состав пород влиял на форму. Как рассчитать ожидаемые числа, будет показано чуть ниже, а пока обратим внимание на внешние особенности таблицы. Кроме дробных чисел в клетках можно заметить

еще одно отличие от табл. 1.1 – это суммарные данные по составу пород в правом столбце и по форме – в нижней строке. В правом нижнем углу – общее число галек. Обратите внимание, что, хотя числа в клетках в табл. 1.1 и 1.2 разные, суммы по строкам и по столбцам одинаковы.

Как же рассчитать ожидаемые числа? Гальки гранитной по составу – 211, метаморфической – 56. Угловатых галек 55 из 267, то есть в 20,60% случаев, окатанных – 212 из 267, то есть – 79,40% случаев.

Примем нулевую гипотезу о независимости двух признаков (состава и формы). Тогда угловатые гальки должны с равной частотой 20,60% наблюдаться в группах гранитных и метаморфических по составу галек. Рассчитав, сколько составляет 20,60% от 211 и 56, получим соответственно 43,46 и 11,54. Это и есть ожидаемые числа количества угловатых галек в группах гранитных и метаморфических по составу пород. Таким же образом можно получить ожидаемые числа окатанных галек: в группе гранитов – 79,40 % от 211, то есть 167,53, в группе метаморфических пород – 79,40 % от 56, то есть 44,47. Обратите внимание, что ожидаемые числа рассчитываются до второго знака после запятой – такая точность понадобится при дальнейших вычислениях.

Сравним табл. 1.1 и 1.2. Числа в клетках довольно слабо различаются. Следовательно, можно предположить, что форма галек не зависит от состава пород. Теперь осталось построить критерий, который бы характеризовал эти различия одним числом, и затем найти его критическое значение.

Таблица 1.2

Связь формы галек из отложений маршельской морены с их составом: ожидаемые числа

Состав пород	Угловатые	Окатанные	Сумма в строке
Гранитные	43.46	167.53	211
Метаморфические	11.54	44.47	56
Сумма в столбце	55	212	267

1.2.1. Критерий χ^2 для таблицы 2x2

Критерий χ^2 (читается «хи-квадрат») не требует никаких предположений относительно параметров совокупности, из которой извлечены выборки, – это один из *непараметрических* критериев, с которыми мы познакомимся. Займемся его построением. Во-первых, как и всегда, критерий должен давать одно число, которое служило бы мерой отличия наблюдаемых данных от ожидаемых, то есть в данном случае различия между таблицей наблюдаемых и ожидаемых чисел. Во-вторых, критерий должен учитывать, что различие, скажем, в

одну единицу имеет большее значение при малом ожидаемом числе, чем при большом. Определим критерий χ^2 следующим образом:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

где О – наблюдаемое число в клетке таблицы сопряженности, E – ожидаемое число в той же клетке. Суммирование проводится по всем клеткам таблицы. Как видно из формулы, чем больше разница наблюдаемого и ожидаемого числа, тем больший вклад вносит клетка в величину χ^2. При этом клетки с малым ожидаемым числом вносят больший вклад. Таким образом, критерий удовлетворяет обоим требованиям – во-первых, измеряет различия и, во-вторых, учитывает их величину относительно ожидаемых чисел.

Применим критерий χ^2 к данным по составу и форме пород. В табл. 1.1 приведены наблюдаемые числа, а в табл. 1.2 – ожидаемые.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(41-43.46)^2}{43.46} + \frac{(170-167.53)^2}{167.53} + \frac{(14-11.54)}{11.54} + \frac{(42-44.47)^2}{44.47} = 0.84$$

Много это или мало? Критическое значение χ^2 можно найти хорошо знакомым нам способом. На рис. 1.7 показано распределение возможных значений χ^2 для таблиц сопряженности размером 2x2 для случая, когда между изучаемыми признаками нет никакой связи. Величина χ^2 не превышает 3,84 только в 5% случаев. Таким образом, 3,84 – критическое значение для 5% уровня значимости. В примере с породами мы получили значение 0.84, поэтому мы принимаем гипотезу о независимости форм галек из отложений маршельской морены от состава пород.

Разумеется, как и все критерии значимости, χ^2 дает *вероятностную* оценку истинности той или иной гипотезы. На самом деле состав пород может и оказывать влияния на форму галек. Но, как показал критерий, это маловероятно.

Применение критерия χ^2 правомерно, если ожидаемое число в любой из клеток больше или равно 5 (в противном случае возможно использование точного критерия Фишера, подробное описание которого есть в работе /3/).

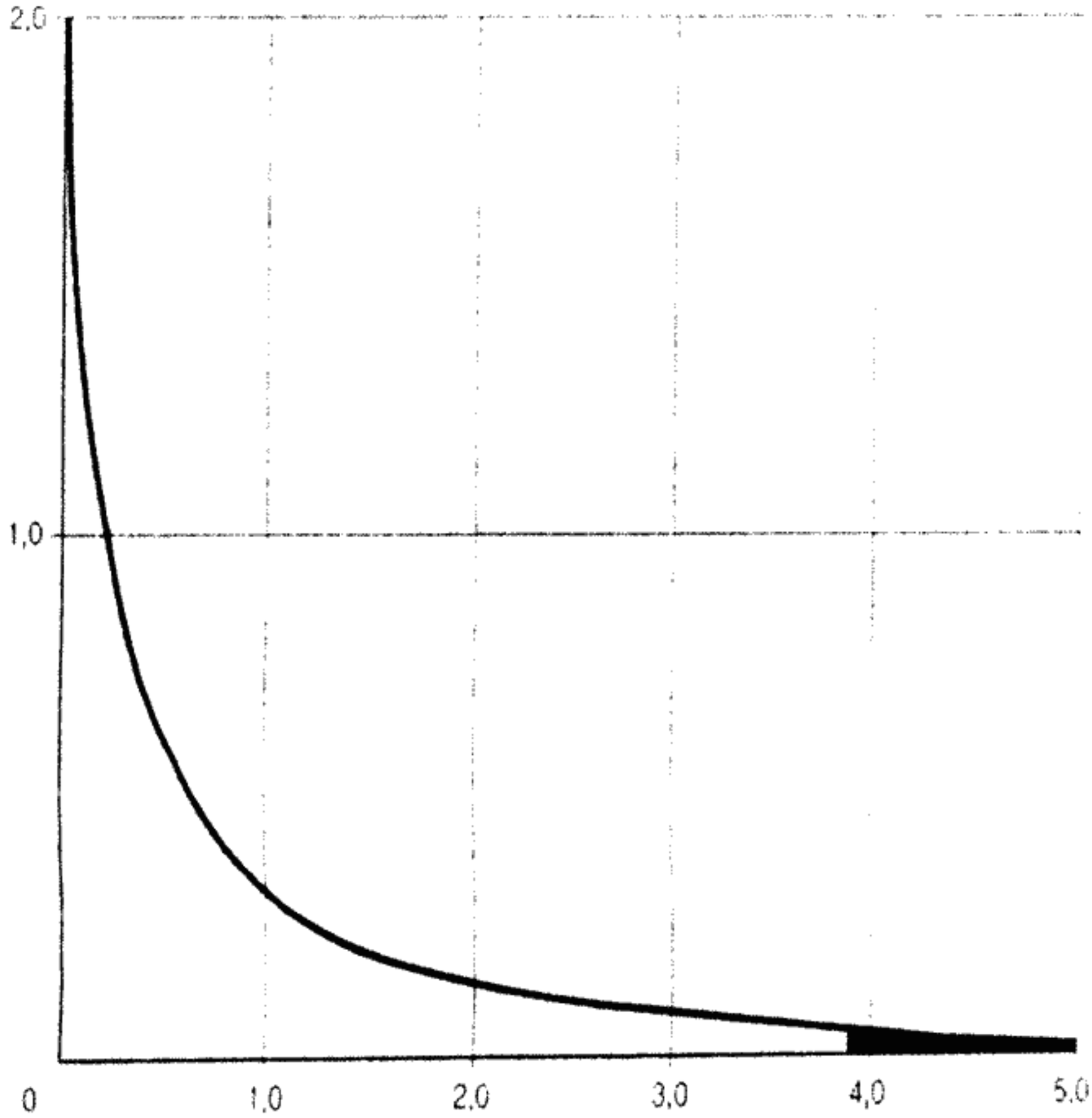


Рис. 1.7. Распределение χ^2 с 1 степенью свободы. Заштрихованная зона – это 5% наибольших значений

Критическое значение χ^2 зависит от размеров таблицы сопряженности, то есть от числа сравниваемых методов лечения (строк таблицы) и числа возможных исходов (столбцов таблицы). Размер таблицы выражается числом степеней свободы v:

$$v = (r - 1)(c - 1),$$

где r – число строк, а c – число столбцов. Для таблиц размером 2x2 имеем v = (2-1)(2-1) = 1. Критические значения χ^2 для разных v приведены в табл. 1.3

Приведенная ранее формула для χ^2 в случае таблицы 2x2 (то есть при 1 степени свободы) дает несколько завышенные значения (сходная ситуация была с критерием *z*). Это вызвано тем, что теоретическое распределение χ^2 непрерывно, тогда как набор вычисленных значений χ^2 дискретен. На практике это приведет к тому, что нулевая гипотеза будет отвергаться слишком часто.

Чтобы компенсировать этот эффект, в формулу вводят поправку Йейтса:

$$\chi^2 = \sum \frac{(|O-E| - \frac{1}{2})^2}{E}.$$

Заметим, поправка Йейтса применяется только при v = 1, то есть для таблиц 2x2.

Применим поправку Йейтса к изучению связи между составом пород и формой галек (табл. 1.1 и 1.2).Расчет проводился по программе **БИОСТАТ** /3/.

Решение.

χ^2 =0.53. $\alpha = 0.05$.В связи с тем, что вычисленное значение равное 0,53 меньше 3,84 (при $\alpha = 0.05$.), для отклонения гипотезы независимости изученных характеристик по-прежнему нет оснований, т.е. форму галек следует считать независящей от принадлежности образца к гранитам или метаморфическим породам.

Таблица 1. 3.

Критические значения χ^2 /3/

	Уровень значимости							
v	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
1	0,455	1,323	2,706	3,841	5,024	6,635	7,879	10,828
2	1,386	2,773	4,605	5,991	7,378	9,210	10,597	13,816
3	2,366	4,108	6,251	7,815	9,348	11,345	12,838	16,266
4	3,357	5,385	7,779	9,488	11,143	13,277	14,860	18,467
5	4,351	6,626	9,236	11,070	12,833	15,086	16,750	20,515
6	5,348	7,841	10,645	12,592	14,449	16,812	18,548	22,458
7	6,346	9,037	12,017	14,067	16,013	18,475	20,278	24,322
8	7,344	10,219	13,362	15,507	17,535	20,090	21,955	26,124
9	8,343	11,389	14,684	16,919	19,023	21,666	23,589	27,877
10	9,342	12,549	15,987	18,307	20,483	23,209	25,188	29,588
11	*10,341*	*13,701*	*17,275*	*19,675*	*21,920*	*24,725*	*26,757*	*31,264*
12	11,340	14,845	18,549	21,026	23,337	26,217	28,300	32,909
13	12,340	15,984	19,812	22,362	24,736	27,688	29,819	34,528
14	13,339	17,117	21,064	23,685	26,119	29,141	31,319	36,123
15	14,339	18,245	22,307	24,996	27,488	30,578	32,801	37,697
16	15,338	19,369	23,542	26,296	28,845	32,000	34,267	39,252
17	16,338	20,489	24,769	27,587	30,191	33,409	35,718	40,790
18	17,338	21,605	25,989	28,869	31,526	34,805	37,156	42,312
19	18,338	22,718	27,204	30,144	32,852	36,191	38,582	43,820
20	19,337	23,828	28,412	31,410	34,170	37,566	39,997	45,315
21	20,337	24,935	29,615	32,671	35,479	38,932	41,401	46,797
22	21,337	26,039	30,813	33,924	36,781	40,289	42,796	48,268
23	22,337	27,141	32,007	35,172	38,076	41,638	44,181	49,728
24	23,337	28,241	33,196	36,415	39,364	42,980	45,559	51,179
25	24,337	29,339	34,382	37,652	40,646	44,314	46,928	52,620
26	25,336	30,435	35,563	38,885	41,923	45,642	48,290	54,052
27	26,336	31,528	36,741	40,113	43,195	46,963	49,645	55,476
28	27,336	32,020	37,916	41,337	44,461	48,278	50,993	56,892
29	28,336	33,711	39,087	42,557	45,722	49,588	52,336	58,301
30	29,336	34,800	40,256	43,773	46,979	50,892	53,672	59,703
31	30,336	35,887	41,422	44,985	48,232	52,191	55,003	61,098
32	31,336	36,973	42,585	46,194	49,480	53,486	56,328	62,487
33	32,336	38,058	43,745	47,400	50,725	54,776	57,648	63,870
34	33,336	39,141	44,903	48,602	51,966	56,061	58,964	65,247
35	34,336	40,223	46,059	49,802	53,203	57,342	60,275	66,619
36	35,336	41,304	47,212	50,998	54,437	58,619	61,581	67,985
37	36,336	42,383	48,363	52,192	55,668	59,893	62,883	69,346
38	37,335	43,462	49,513	53,384	56,896	61,162	64,181	70,703
39	38,335	44,539	50,660	54,572	58,120	62,428	65,476	72,055
40	39,335	45,616	51,805	55,758	59,342	63,691	66,766	73,402

	Уровень значимости							
ν	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
41	40,335	46,692	52,949	56,942	60,561	64,950	68,053	74,745
42	41,335	47,766	54,090	58,124	61,777	66,206	69,336	76,084
43	42,335	48,840	55,230	59,304	62,990	67,459	70,616	77,419
44	43,335	49,913	56,369	60,481	64,201	68,710	71,893	78,750
45	44,335	50,985	57,505	61,656	65,410	69,957	73,166	80,077
46	45,335	52,056	58,641	62,830	66,617	71,201	74,437	81,400
47	46,335	53,127	59,774	64,001	67,821	72,443	75,704	82,720
48	47,335	54,196	60,907	65,171	69,023	73,683	76,969	84,037
49	48,335	55,265	62,038	66,339	70,222	74,919	78,231	85,351
50	49,335	56,334	63,167	67,505	71,420	76,154	79,490	86,661

1.2.2. Использование χ^2 – критерия (пример)

Распределение χ^2 имеет большое значение в практике, так как его можно использовать для проверки гипотез, содержащих как номинальные, так и порядковые данные.

Общеизвестная задача статистического анализа заключается в сравнении выборочного распределения с некоторым заранее заданным распределением. Так, статистические критерии можно применить для проверки гипотезы, заключающейся в том, что имеющиеся данные извлечены из совокупности с заранее известным распределением, возможно, нормальным. Для того чтобы убедиться в том, что это предположение не противоречит действительности, надо сравнить выборочное и теоретическое распределения. В большинстве случаев геолог задает соответствующие классы и затем проверяет, согласуется ли распределение с некоторым заданным теоретическим распределением. Т.е., требуется установить соответствие между формой двух распределений, одно из которых получено по выборке, а другое либо заранее известно, либо предполагается имеющим определенный тип. Требуется в вероятностных терминах получить ответ на вопрос: можно ли два указанных распределения отнести к одному типу?

Аналогичная задача возникла в заливе Уайтуотер (штат Флорида), где было проведено 48 измерений (табл. 1.4) солености поверхностных вод /6/. Предположим, что значения содержания соли в выборке с некоторой площади распределены нормально. Если эта гипотеза верна, то из нее следует, что происходит свободное перемешивание и обмен между открытыми морскими водами и пресной водой, втекающей в залив.

Таблица 1.4

Измерения солености вод в заливе Уайтуотер, штат ФЛОРИДА
С о л е н о с т ь, мкг/т

46	53	58	60	60	49	59	48	46	78
37	58	46	46	47	48	42	50	63	48
62	49	47	36	40	39	61	43	53	42
59	60	52	34	40	36	67	44	40	
40	56	51	51	35	47	53	49	50	

С другой стороны, если бы существовал какой-либо механизм, который стремился бы разделить соленую и пресную воду в заливе, то распределение содержания соли позволило бы его обнаружить. Это дало бы возможность получить представление о циркуляции воды и предсказать тип распределения донных осадков. С помощью соответствующего критерия согласия можно проверить, насколько хорошо выборочное распределение согласуется с нормальным.

Предположим, что совокупность определений солености, из которой взята наша выборка (табл.1.4), характеризуется нормальным распределением с неизвестным средним значением μ и дисперсией σ^2. Альтернативой этой гипотезе является предположение, что это распределение не согласуется с нормальным законом. Значение статистического критерия можно вычислить путем подразделения области определения стандартного нормального распределения на некоторое число отрезков. Вероятность того, что одно случайное наблюдение, извлеченное из стандартного нормального распределения, попадает в один из отрезков, равна площади под кривой в пределах отрезка. Используя эти вероятности, можно вычислить ожидаемое число наблюдений в каждом отрезке. Ожидаемые частоты в каждом отрезке затем сравниваются с соответствующими выборочными частотами. Если эти числа значительно отклоняются от ожидаемых, то маловероятно, чтобы выборка была извлечена из нормальной совокупности. Используем χ^2 – критерий.

Указанный статистический критерий в рассматриваемом примере вычисляется по формуле:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

где O_j – число наблюдений в j-м классе, E_j – ожидаемое число наблюдений в этом классе. Предполагается, что имеется k различных классов (или интервалов).

В этой задаче значение статистического критерия вычисляется с помощью подразделения области определения наблюдаемых значений на некоторое число отрезков, например четыре, но так, чтобы им соответствовали равные площади под нор-

мальной кривой, которым, следовательно, отвечают равные вероятности попадания в соответствующий интервал. Границами этих интервалов, соответствующих равным вероятностям, будут – ∞; – 0,67; 0,0; +0,67 и ∞. ("Правильность" границ интервалов можно проверить по таблице 3.1 в/1/).

Если данные стандартизованы, то можно ожидать, что приблизительно одна четверть значений попадает в каждый из интервалов. Далее подсчитывается число проб, попадающих в каждый из этих интервалов, находится разность между ожидаемыми и реальными числами, а результат возводится в квадрат. Квадрат разности делится на ожидаемое число попаданий в данный интервал. Полученные значения суммируются по всем четырем интервалам. Если сумма превышает критическое значение, то нулевая гипотеза отклоняется и делается вывод, что распределение значений солености не согласуется с нормальным.

Критическое значение при использовании χ^2 - критерия зависит от числа степеней свободы. В нашем примере число степеней свободы определяется следующим образом: "выборки" – это четыре категории, сравниваемые с соответствующими категориями стандартного нормального распределения. У нас число степеней свободы равно числу категорий без трех, т.е. единице. Мы потеряли две степени свободы потому, что для неизвестных параметров μ и σ^2 использовали их оценки $\bar{X}$ и s^2 соответственно, а еще одну степень свободы за счет того, что сумма частот по интервалам равна единице. Критическое значение χ^2 - распределения, соответствующее 5%-ному уровню значимости (α=0.05) и одной степени свободы, равно 3.84 (см. табл. 1.3).

Напомним, что χ^2 - критерий всегда является односторонним, и область отклонения гипотезы расположена справа (см. рис.1.7).

В нашем примере область наблюдаемых значений должна быть разбита на четыре части с равными вероятностями. Если значения солености распределены нормально, то приблизительно 12 нормализованных значений должно попасть в каждую из четырех категорий.

По выборке вычисляем действительное число наблюдений (частот попадания), содержащихся в каждой из этих групп. Так как групп всего четыре, то ожидаемые значения числа наблюдений равны 12.

Первый шаг – стандартизация данных по формуле

$$Z_i = (X_i - \bar{X})/s .$$

Выборка данных, полученных при опробовании в заливе Уайтуотер, имеет оценку среднего $\bar{X}$ = 49,54 и оценку стандартного отклонения s = 9,27.

Поэтому нормализация наблюдений осуществляется как

$$Z_i = (X_i - 49{,}54)/9{,}27$$

Таблица 1.5

Стандартизованные значения солености в заливе Уайтуотер

Номер образца	Выборочные значения	Стандартизованные значения	Номер образца	Выборочные значения	Стандартизованные значения
1	46,00	—0,38	25	35,00	—1,57
2	37,00	—1,35	26	49,00	—0,06
3	62,00	1,34	27	48,00	—0.17
4	59,00	1,02	28	39,00	—1,14
5	40,00	—1,03	29	36,00	—1,46
6	53,00	0,37	30	47,00	—0,27
7	58,00	0,91	31	59,00	1,02
8	49,00	—0,06	32	42,00	—0,81
9	60,00	1,13	33	61,00	1,24
10	56,00	0,70	34	67,00	1,88
11	58,00	0,91	35	53,00	0,37
12	46,00	—0,38	36	48,00	—0,17
13	47,00	—0,27	37	59,00	0,05
14	52,00	0,27	38	43,00	—0,71
15	51,00	0,16	39	44,00	—0,60
16	60,00	1,13	40	49,00	—0,06
17	46.00	—0,38	41	46,00	—0,38
18	36,00	—1,46	42	63,00	1,45
19	34,00	—1,68	43	53,00	0,37
20	51,00	0,16	44	40,00	—1,03
21	60,00	1,13	45	50,00	0,05
22	47,00	—0,27	46	78,00	3,07
23	40,00	—1,03	47	48.00	—0,17
24	40,00	—1,03	48	42,00	—0,81

Таблицы 1.6 и 1.7

Стандартизованные значения солености, сгруппированные для проверки гипотезы о нормальном распределении таким образом, что каждой группе соответствует вероятность 0,25

Категории от –∞ до —0,67		Категория от —0,67 до 0,0	
—1,35	—1,14	—0,38	—0,17
—1,03	—1,46	—0,06	—0,27
—1,46	—0,81	—0,38	—0,17
—1,68	—0,71	—0,27	—0,60
—1,03	—1,03	—0,38	—0,06
—1,03	—0,81	—0,27	—0,38
—1,47		—0,06	—0,17
Общее число наблюдений 13		Общее число наблюдений 14	

Категории от 0,0 до +0,67		Категория от +0,67 до ∞	
0,37	0,37	1,34	1,13
0,27	0,05	1,02	1,02
0,16	0,37	0,91	1,24
0,16	0,05	1,13	1,88
		0,70	1,45
		0,91	3,07
		1,13	
Общее число наблюдений 8		Общее число наблюдений 13	

Стандартизованные значения приведены в табл. 1.5. В табл. 1.6 и 1.7 приведены результаты разбиения всей выборки на четыре категории. Если выборку можно считать извлеченной из нормальной совокупности, то следует ожидать приблизительно 12 наблюдений на категорию.

Вычисляя значения критерия χ^2, получим следующий результат:

$$\chi^2 = \frac{(13-12)^2}{12} + \frac{(14-12)^2}{12} + \cdot\frac{(8-12)^2}{12} + \frac{(13-12)^2}{12} = \frac{22}{12} = 1{,}83.$$

Вычисленное значение χ^2 меньше критического 3.84 5%-ного уровня значимости и одной степени свободы. Поэтому нет оснований считать, что распределение значений солености в поверхностных водах существенно отклоняется от нормального закона. (Теперь можно делать определенные геологические выводы).

Статистика χ^2 позволяет проверить гипотезу не только о нормальном распределении. Можно применить этот критерий для проверки гипотезы о любом другом законе распределения /6/.

1.3. Наличие связи (корреляции) между признаками (коэффициенты Пирсона и Спирмена)

Одним из важных разделов статистики является корреляционный анализ, используемый как для количественных, так и качественных данных. Понятие корреляции отражает, главным образом, степень выраженности связи между вариационными рядами. Наглядно эта связь может быть отражена графически. На координатной плоскости по оси абсцисс откладывают значения одного вариационного ряда, а по оси ординат – другого. Совокупность таких точек на координатной плоскости (их число равно числу наблюдений) создает общую картину корреляции и обычно позволяет построить некоторую усредненную кривую взаимозависимости параметров, составляющих оба вариационных ряда

(регрессионный анализ). На практике исследователя часто может интересовать не сама зависимость одной переменной от другой, а именно характеристика тесноты связи между этими переменными, которую можно было бы выразить одним числом. Эта характеристика называется *коэффициентом корреляции.* Выраженность линейной связи между двумя случайными величинами X и Y, имеющими нормальное распределение, обычно оценивают коэффициентом корреляции Пирсона, рассчитываемым по следующей формуле:

$$r = \frac{n \cdot \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{\sqrt{\left(n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right) \cdot \left(n \cdot \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2 \right)}},$$

где X_i и Y_i — соответствующие значения параметра в i-ом наблюдении $(i = \overline{1,n})$, n – количество наблюдений.

Вычисленный коэффициент корреляции является выборочной оценкой генерального коэффициента корреляции совокупности, а значит, как и любая случайная величина, имеет ошибку σ. Отношение выборочного коэффициента корреляции к своей ошибке является критерием для проверки нулевой гипотезы о равенстве нулю генерального коэффициента корреляции совокупности (или соответственно о независимости случайных величин X и Y):

$$t = r \sqrt{\frac{n-2}{1-r^2}} \text{ – статистика Стьюдента с n-2 степенями свободы .} \qquad (1.3)$$

Число степеней свободы для проверки критерия равно n-2, гипотезу проверяют по таблицам распределения Стьюдента /1,3,6/ в соответствии с выбранным уровнем значимости. Если вычисленное значение превзойдет или окажется равным соответствующему табличному, нулевую гипотезу отвергают.

Приведенная формула для вычисления коэффициента корреляции является параметрической, т.е. предполагает, что анализируемые переменные распределены по нормальному закону.

Существуют и другие формулы для вычисления коэффициента корреляции, кроме этого формулы могут уточняться по отношению к большим и малым выборкам /2, 4, 6,7,14/.

Однако, независимо от способа вычисления, коэффициент корреляции обладает определенными свойствами.

Величина коэффициента корреляции всегда заключена в пределах $-1 < r < 1$. Если $r < 0$, то это означает, что с увеличением в вариационном ряду наблюдаемых величин X соответствующие им значения Y второго вариационного ряда в среднем уменьшаются. Если $r > 0$, то с увеличением значений одного параметра другой параметр так же в среднем возрастает. Если $r = 0$, то это означает, что параметры X и Y абсолютно независимы. При $r = 1$ между параметрами существует прямо пропорциональная функциональная зависимость.

Чем больше абсолютная величина коэффициента корреляции, тем при данном объеме выборки больше доверительная вероятность того, что характер

характер связи действительно соответствует полученному коэффициенту корреляции. На рис.1.8. показаны некоторые типичные варианты зависимостей и соответствующие им значения коэффициентов корреляции.

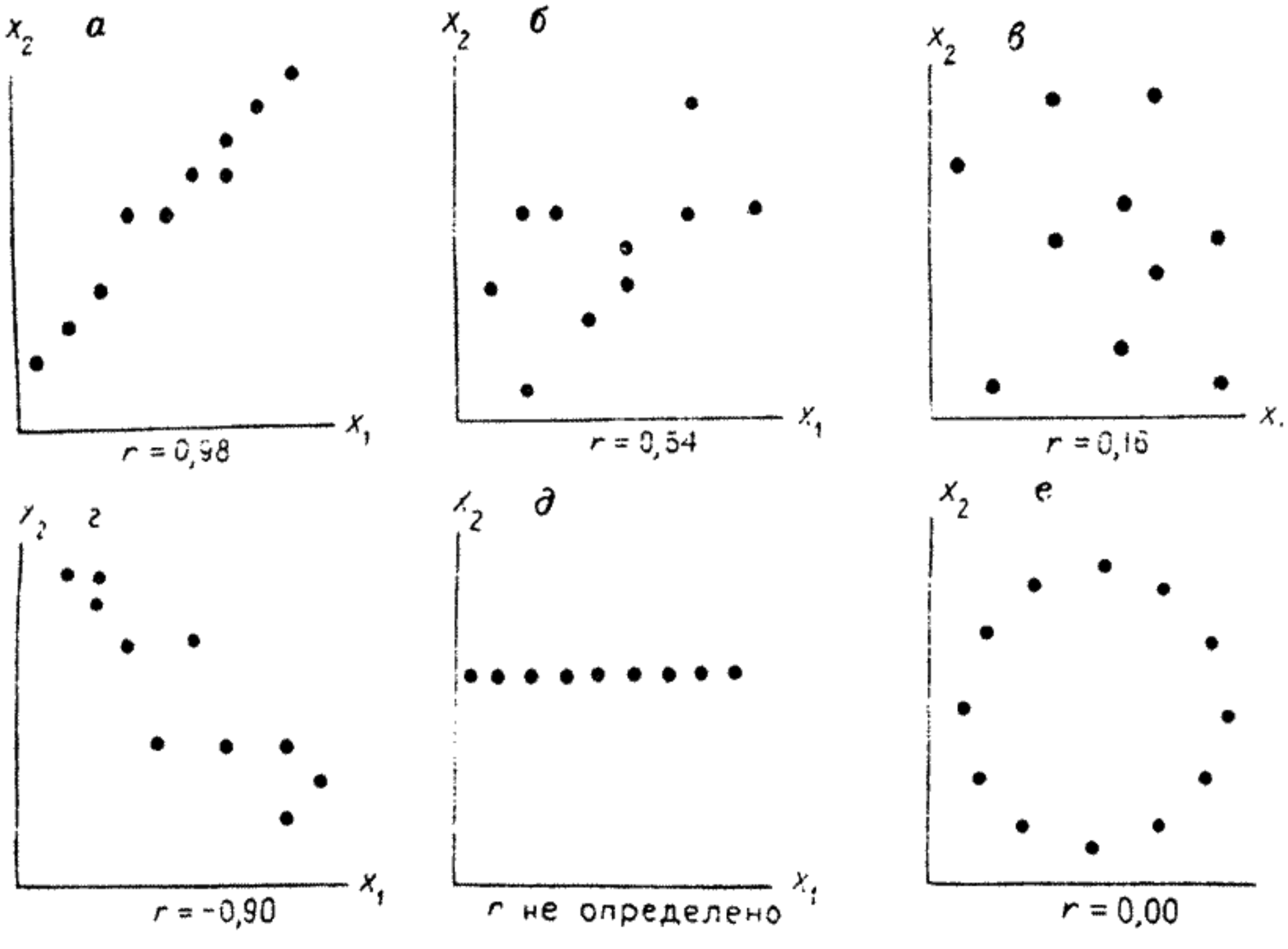


Рис.1.8 . Схематичное изображение различных вариантов зависимостей между переменными X и У (X_1 и X_2) и соответствующие значения коэффициента корреляции Пирсона

Известно, что часто встречаются случаи, когда данные представлены качественными признаками, в том числе порядковыми переменными. При этом приходится оперировать так называемыми ранговыми коэффициентами корреляции /3,5,6,7/. Кроме того, такой непараметрический подход применяется в случае малых выборок и если изучаемые выборки не распределены по нормальному закону. К таким коэффициентам, например, относится коэффициент корреляции рангов, предложенный К. Спирменом и вычисляемый по формуле:

$$r = 1 - \frac{6\sum_{i=1}^{n} di^2}{n(n^2-1)}, \qquad (1.4)$$

где *di* — разность между рангами сопряженных признаков, *n* – число парных членов ряда.

При полной связи ранги признаков совпадут, и разность между ними будет равна 0, соответственно коэффициент корреляции будет равен 1. Если же признаки варьируются независимо, коэффициент корреляции получится равным 0.

Для оценки значимости коэффициента ранговой корреляции Спирмена для разных уровней значимости и объемов выборки обычно используют соответствующую таблицу (табл.1.10) или, при n > 50, пользуются формулой (1.3).

Расчет рангового коэффициента корреляции Спирмена проиллюстрируем на вымышленном примере. Так, допустим, в ходе исследований изучалось влияние фактора А на содержание химического элемента В (в условных единицах) в породе С и концентрацию химического элемента D в воде (в у.е), пробы которой разделены по какому-то признаку Е на 3 группы (табл.1.8).

Оценим коэффициент корреляции рангов для параметров X1 и X2 (табл. 1.9), "взятых" из табл.1.8 (столбцы 2 и 7).

Если бы отдельные варианты ряда не повторялись, их рангами были бы натуральные числа от 1 в порядке возрастания. Но одинаковым значениям вариант присваиваются ранги, равные средним арифметическим их рангов. Величина d_i представляет собой попарные разности рангов изучаемых выборок. В качестве правила для проверки правильности ранжирования используют равенство 0 (нулю) суммы d_i (предпоследний столбец табл.1.9).

Таблица 1.8

Результаты вымышленного исследования

Содержание химического элемента **B** в породе **C**, (у.е)			Концентрация химического элемента **D**, (у.е)				
			Исходное содержание в воде			Прирост концентрации	
группа 1	группа 2	группа 3	группа 1	группа 2	группа 3	группа 1	группа 2
столбец 1	столбец 2	столбец 3	столбец 4	столбец 5	столбец 6	столбец 7	столбец 8
12	8	8	0.7	0.8	0.8	4	4
13	8	9	1.4	0.9	0.9	5	3
14	9	9	1.8	2.5	2.3	4	3.5
15	10	11	1.5	1.2	2	3.5	2
14	7	12	1.1	1.3	1.4	5	1
13	7	12	1.6	1.5	1.6	5	1.5
13	9	13	1.7	1.6	1.3	3.5	1
10	9	13	1.3	2.1	1.7	4	1.5
11	11	12	1.4	2	1.5	2	2
16	6	11	2.2	1	1.6	5	2

Таблица 1.9

Данные для расчета рангового коэффициента Спирмена

Параметр X1	Параметр X2	Ранг R_{X1}	Ранг R_{X2}	di= R_{X1} - R_{X2}	di^2
8	4	4,5	5	-0,5	0,25
8	5	4,5	8,5	-4,0	16,0
9	4	7	5	2,0	4,0
10	3,5	9	2,5	6,5	42,25
7	5	2,5	8,5	-6,0	36,0
7	5	2,5	8,5	-6,0	36,0
9	3,5	7	2,5	4,5	20,25
9	4	7	5	2,0	4,0
11	2	10	1	9	81,0
6	5	1	8,5	-7,5	56,25

В примере сумма d_i^2 равна 296, по формуле (1.4) для $n = 10$ получаем ранговый коэффициент корреляции $r = -0,79$.

Критическое значение для уровня значимости 5% равно 0,648 (табл.1.10). Так как значение рангового коэффициента корреляции по модулю превосходит соответствующее критическое значение, с вероятностью более 95% можно утверждать, что между сравниваемыми параметрами существует значимая отрицательная корреляционная связь.

Если рассчитать коэффициент корреляции Пирсона для параметров X1 и X2 (табл.1.9) , то получим значение $r = -0,90$, т.е. мы видим, что коэффициенты корреляции, рассчитанные по Спирмену (-0,79) и Пирсону (-0,90) близки.

С помощью известных статистических пакетов рассчитайте коэффициенты корреляции Пирсона и Спирмена для столбцов 1-3, 4-8 табл.1.8.

Таблица 1.10

Критические значения коэффициентов ранговой корреляции Спирмена /3/

n	Уровень значимости α								
	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
4	0,600	1,000	1,000						
5	0,500	0,800	0,900	1,000	1,000				
6	0,371	0,657	0,829	0,886	0,943	1,000	1,000		
7	0,321	0,571	0,714	0,786	0,893	0,929	0,964	1,000	1,000
8	0,310	0,524	0,643	0,738	0,833	0,881	0,905	0,952	0,976
9	0,267	0,483	0,600	0,700	0,783	0,833	0,867	0,917	0,933
10	0,248	0,455	0,564	0,648	0,745	0,794	0,830	0,879	0,903
11	0,236	0,427	0,536	0,618	0,709	0,755	0,800	0,845	0,873
12	0,217	0,406	0,503	0,587	0,678	0,727	0,769	0,818	0,846
13	0,209	0,385	0,484	0,560	0,648	0,703	0,747	0,791	0,824
14	0,200	0,367	0,464	0,538	0,626	0,679	0,723	0,771	0,802
15	0,189	0,354	0,446	0,521	0,604	0,654	0,700	0,750	0,779
16	0,182	0,341	0,429	0,503	0,582	0,635	0,679	0,729	0,762
17	0,176	0,328	0,414	0,485	0,566	0,615	0,662	0,713	0,748
18	0,170	0,317	0,401	0,472	0,550	0,600	0,643	0,695	0,728
19	0,165	0,309	0,391	0,460	0,535	0,584	0,628	0,677	0,712
20	0,161	0,299	0,380	0,447	0,520	0,570	0,612	0,662	0,696
21	0,156	0,292	0,370	0,435	0,508	0,556	0,599	0,648	0,681
22	0,152	0,284	0,361	0,425	0,496	0,544	0,586	0,634	0,667
23	0,148	0,278	0,353	0,415	0,486	0,532	0,573	0,622	0,654
24	0,144	0,271	0,344	0,406	0,476	0,521	0,562	0,610	0,642
25	0,142	0,265	0,337	0,398	0,466	0,511	0,551	0,598	0,630
26	0,138	0,259	0,331	0,390	0,457	0,501	0,541	0,587	0,619
27	0,136	0,255	0,324	0,382	0,448	0,491	0,531	0,577	0,608
28	0,133	0,250	0,317	0,375	0,440	0,483	0,522	0,567	0,598
29	0,130	0,245	0,312	0,368	0,433	0,475	0,513	0,558	0,589
30	0,128	0,240	0,306	0,362	0,425	0,467	0,504	0,549	0,580
31	0,126	0,236	0,301	0,356	0,418	0,459	0,496	0,541	0,571
32	0,124	0,232	0,296	0,350	0,412	0,452	0,489	0,533	0,563
33	0,121	0,229	0,291	0,345	0,405	0,446	0,482	0,525	0,554
34	0,120	0,225	0,287	0,340	0,399	0,439	0,475	0,517	0,547
35	0,118	0,222	0,283	0,335	0,394	0,433	0,468	0,510	0,539
36	0,116	0,219	0,279	0,330	0,388	0,427	0,462	0,504	0,533
37	0,114	0,216	0,275	0,325	0,383	0,421	0,456	0,497	0,526
38	0,113	0,212	0,271	0,321	0,378	0,415	0,450	0,491	0,519
39	0,111	0,210	0,267	0,317	0,373	0,410	0,444	0,485	0,513
40	0,110	0,207	0,264	0,313	0,368	0,405	0,439	0,479	0,507

Окончание табл. 1.10

n	Уровень значимости α								
	0,50	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
41	0,108	0,204	0,261	0,309	0,364	0,400	0,433	0,473	0,501
42	0,107	0,202	0,257	0,305	0,359	0,395	0,428	0,468	0,495
43	0,105	0,199	0,254	0,301	0,355	0,391	0,423	0,463	0,490
44	0,104	0,197	0,251	0,298	0,351	0,386	0,419	0,458	0,484
45	0,103	0,194	0,248	0,294	0,347	0,382	0,414	0,453	0,479
46	0,102	0,192	0,246	0,291	0,343	0,378	0,410	0,448	0,474
47	0,101	0,190	0,243	0,288	0,340	0,374	0,405	0,443	0,469
48	0,100	0,188	0,240	0,285	0,336	0,370	0,401	0,439	0,465
49	0,098	0,186	0,238	0,282	0,333	0,366	0,397	0,434	0,460
50	0,097	0,184	0,235	0,279	0,329	0,363	0,393	0,430	0,456

1.4. Непараметрические методы

Все статистические методы, с которыми мы знакомились, в основном, являются параметрическими, т. е. они основаны на характеристиках распределений, параметры которых известны. Все они строятся для выборок из нормальных совокупностей. Для обоснования возможности использования этих критериев в тех случаях, когда исследуемая совокупность не является нормальной, при условии, что объем выборки велик и совокупность не очень сильно отличается от нормальной, следует обратиться к центральной предельной теореме. Иногда, однако, исследуемая совокупность может сильно отличаться от нормальной, или же объем выборки нельзя увеличить. В таких случаях следует обратиться к категории критериев, называемых непараметрическими статистическими критериями. Их можно применять для обработки информации более низких шкал, таких, как номинальные и порядковые данные, в отличие от метрических данных, используемых в параметрической статистике.

Не требуется никаких допущений о виде исходного распределения, отсюда и название – непараметрические критерии. Вообще, в тех случаях, когда выборочная совокупность имеет характеристики, необходимые в параметрическом анализе, непараметрические критерии оказываются менее мощными, чем эквивалентные параметрические.

Для сравнения средних значений может применяться и целый ряд непараметрических критериев, среди которых важное место занимают так называемые ранговые критерии. Применение этих критериев основано на ранжировании членов сравниваемых групп. При этом сравниваются не сами члены ранжированного ряда, а их порядковые номера или ранги. Познакомиться с основными непараметрическими критериями можно, например, в работах /2,5,6,10,16/. Там же даны и основные таблицы для проверки этих критериев.

При решении конкретной задачи очень важно правильно выбрать критерий.

1.4.1. Критерий Манна — Уитни (Уилкоксона)

Критерий Манна – Уитни можно использовать как непараметрический эквивалент t – критерия для проверки гипотезы о равенстве средних двух выборок. Предположим, что мы имеем две выборки объема n_1 и n_2 и хотим проверить гипотезу о том, что они являются выборками из одной и той же совокупности. Объединим обе выборки и расположим значения наблюдений в порядке возрастания от меньшего к большему. Каждому наблюдению припишем его ранг, т. е. наименьшему значению припишем ранг 1, следующему по величине – ранг 2 и так далее, до наибольшего наблюдения, которое будет иметь ранг $(n_1+ n_2)$. Если обе выборки были взяты из одной и той же совокупности наудачу, то можно ожидать, что наблюдения одной из выборок будут более или менее равномерно рассеяны в последовательности рангов.

Пусть X_i - *i-е* наблюдение первой выборки, а Y_i – *i-е* наблюдение второй выборки. Ранг наблюдения X_i будет обозначаться через $R(X_i)$, а ранг Y_i – через $R(Y_i)$. Критерий Манна—Уитни имеет вид

$$U_i = \sum_{i=1}^{n} R(X_i) - \frac{n(n+1)}{2} \quad . \tag{1.5}$$

Первый член – просто сумма рангов наблюдений из первой выборки. Критические значения U приведены в табл. 1.11.

В качестве тестовой статистики выбирают минимальную величину U и сравнивают ее с табличным значением для принятого уровня значимости. Гипотеза принимается, и различия считаются недостоверными, если рассчитанное значение больше соответствующего табличного (табл.1.11).

В качестве примера использования критерия Манна – Уитни рассмотрим данные табл. 1.8. Проверим гипотезу о принадлежности сравниваемых независимых выборок к одной и той же генеральной совокупности с помощью непараметрического критерия Манна–Уитни. Для расчета критерия расположим варианты сравниваемых выборок в порядке возрастания в один обобщенный ряд и присвоим вариантам обобщенного ряда ранги от 1 до $n_1 + n_2$.

Первая строка – упорядоченные значения столбца 2 (табл.1.8), вторая – упорядоченные значения столбца 3 (табл.1.8), третья – соответствующие ранги в обобщенном ряду:

6	**7**	**7**	**8**	**8**		**9**	**9**	**9**			**10**	**11**							
					8				**9**	**9**			**11**	**11**	**12**	**12**	**12**	**13**	**13**
1	**2,5**	**2,5**	**5**	**5**	**5**	**9**	**9**	**9**	**9**	**9**	**12**	**14**	**14**	**14**	**17**	**17**	**17**	**19,5**	**19,5**

Надо обратить внимание, что если имеются одинаковые варианты, им присваивается средний ранг, однако, значение последнего ранга должно быть равно $n_1 + n_2$ (в нашем случае 20). Это правило используют для проверки правильности ранжирования.

Отдельно для каждой выборки рассчитываем суммы рангов их вариант R1 ($R(X_i)$) и R2 ($R(Y_i)$). В нашем случае:

R1 = 1 + 2.5 + 2.5 + 5 + 5 + 9 + 9 + 9 + 12 + 14 = 69

R2 = 5 + 9 + 9 + 14 + 14 + 17 + 17 + 17 + 19.5 + 19.5 = 141

Для проверки правильности вычислений можно воспользоваться правилом: $R1 + R2 = 0.5 * (n_1 + n_2) * (n_1 + n_2 + 1)$, т.е. $R1 + R2 = 69 + 141 = 0.5 * 20 * 21 = 210$.

Статистика $U_1 = 69 - 10*11/2 = 14$, $U_2 = 141 - 10*11/2 = 86$. Для проверки одностороннего критерия выбираем минимальную статистику $U_1 = 14$ и сравниваем ее с табличным значением (табл.1.11) для $n_1 = n_2 = 10$ и уровня значимости 1%, равным 19. Так как вычисленное значение критерия меньше табличного, нулевая гипотеза отвергается на выбранном уровне значимости, и различия между выборками признаются статистически значимыми.

Таблица 1.11.

Критические значения статистики U-критерия Манна – Уитни /5/

Односторонний критерий, α=0,01

n2/n1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_2
3	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5	3
4	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10	4
5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	5
6		3	4	6	7	8	9	11	12	14	15	16	18	19	20	22	6
7			6	7	9	11	12	14	16	18	19	21	23	24	26	28	7
8				9	11	13	15	17	20	22	24	26	28	30	32	34	8
9					14	16	19	21	23	25	28	31	33	36	38	40	9
10						19	22	24	27	30	33	36	38	41	44	47	10
11							25	28	31	34	37	41	44	47	50	53	11
12								31	35	38	42	46	49	53	56	60	12
13									39	43	47	51	55	59	63	67	13
14										47	51	56	60	65	69	73	14
15											56	61	66	70	75	80	15
16												66	71	76	82	87	16
17													77	82	88	94	17
18														88	94	100	18
19															101	107	19
20																114	20

Продолжение табл.1.11

Двусторонний критерий, α=0,01

n2/n1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	n_2
5	0	0	0																5
6	0	0	1	2															6
7	0	0	1	3	4														7
8	0	1	2	4	6	7													8
9	0	1	3	5	7	9	11												9
10	0	2	4	6	9	11	13	16											10
11	0	2	5	7	10	13	16	19	21										11
12	1	3	6	9	12	15	18	21	24	23									12
13	1	4	7	10	13	17	20	24	27	31	34								13
14	1	4	7	11	15	18	22	26	30	34	38	42							14
15	2	5	8	12	16	20	25	29	33	37	42	46	51						15
16	2	5	9	13	18	22	27	31	36	41	46	50	55	60					16
17	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70				17
18	2	6	11	16	21	26	31	37	42	47	53	59	64	70	75	77	81		18
19	3	7	12	17	22	28	34	39	45	51	57	63	69	75	81	87	93		19
20	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105	20
21	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98	105	112	21
22	4	9	14	21	27	34	40	47	54	61	68	75	82	89	97	104	111	118	22
23	4	9	15	22	29	36	43	50	57	64	72	79	87	94	102	109	117	125	23
24	4	10	16	23	30	37	45	52	60	68	76	83	91	99	107	115	123	131	24
25	5	10	17	24	32	39	47	55	63	71	79	88	96	104	113	121	129	138	25

Критерий Манна – Уитни появляется в немного различающихся формах в литературе. Среди них можно назвать критерий Зигеля – Тьюки, критерий Фестинджера и т. д. Вариант Зигеля – Тьюки наиболее интересен, так как он может быть использован для проверки равенства дисперсий в двух выборках и является, таким образом, непараметрическим аналогом F-критерия.

1.4.2. Т-критерий Уилкоксона

В случае попарно связанных выборок применяется Т-критерий Уилкоксона. При этом ранжируют попарные разности – положительные и отрицательные (кроме нулевых) в один ряд так, чтобы наименьшая абсолютная разница (без учета знака) получила первый ранг, одинаковым величинам присваивают один ранг. Отдельно вычисляют сумму рангов положительных (Т+) и отрицательных разностей (Т-), меньшую из двух таких сумм без учета знака считают тестовой статистикой данного критерия. Нулевую гипотезу принимают на данном уровне значимости, если вычисленная статистика превзойдет табличное значение (табл.1.12) (число парных наблюдений уменьшают на число исключенных нулевых разностей). Таким образом, можно сказать, что если нулевая гипотеза верна, статистики Т+ и Т – примерно равны, сравнительно малые или большие значения Т-статистик заставят нас отклонить нулевую гипотезу об отсутствии различий.

Посмотрим пример из /5/. Допустим, в результате проведения исследования был вычислен ряд попарных разностей между показателем эффекта в двух попарно связанных группах ($n_1 + n_2 = 10$). Например, так называемая задача "до и после":

0,2 -0,4 0,7 -0,9 1,3 1,5 -0,1 0,8 -1,0 1,1.

Ранжируем попарные разности в один ряд, независимо от знака разности, получаем следующий ранжированный ряд:

−0,1	0,2	−0,4	0,7	0,8	−0,9	−1,0	1,1	1,3	1,5
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10.**

Рассчитаем отдельно сумму рангов положительных (Т+) и отрицательных (Т -) разностей, в нашем случае Т+ = 2 + 4 + 5 + 8 + 9 + 10 = 38, Т- = 1 + 3 + 6 + 7 = 17. Для проверки двустороннего Т- критерия используем меньшую статистику Т – = 17 и сравним ее с табличным значением (табл.1.12) для числа попарных разностей n = 10 и уровня значимости 5%. Такое табличное критическое значение равно 9. Рассчитанное минимальное значение Т статистики превосходит соответствующее табличное значение, а значит нулевая гипотеза остается в силе (нулевая гипотеза об отсутствии различия).

Таблица 1.12

Критические значения статистики парного Т-критерия /5/

Число парных наблюдений *n*	Уровни значимости α, %		Число парных наблюдений *n*	Уровни значимости α, %	
	5	1		5	1
Односторонний критерий					
5	0	—	14	25	16
6	2	0	15	30	19
7	3	0	16	35	23
8	5	1	17	41	28
9	8	3	18	47	33
10	10	5	19	53	38
11	13	7	20	60	42
12	17	10	21	67	50
13	21	12	22	74	56
Двусторонний критерий					
6	1	—	16	31	21
7	3	—	17	36	24
8	5	1	18	41	29
9	7	3	19	47	33
10	9	4	20	53	39
11	12	6	21	60	44
12	15	8	22	67	50
13	18	11	23	74	56
14	22	14	24	82	62
15	26	17	25	90	69

2. ДИСПЕРСИОННЫЙ АНАЛИЗ

Практическое значение дисперсионного анализа заключается в том, что с его помощью из целой группы факторов, предположительно оказывающих влияние на исследуемый признак, можно выделить те, которые действительно на него влияют / 8 /.

Суть дисперсионного анализа состоит в разложении общей дисперсии результативного признака на части, обусловленные влиянием контролируемых факторов, и остаточную дисперсию, объясняемую неконтролируемым влиянием или случайными обстоятельствами. Выводы о существенности влияния контролируемых факторов на результат производятся путем сравнения частей общей дисперсии при выполнении требования нормальности распределения результативного признака.

Существует ряд моделей дисперсионного анализа /3,6,8,9,16/ – они могут классифицироваться по природе факторов, по их числу. Рассмотрим некоторые, наиболее часто используемые модели.

2.1. Однофакторный дисперсионный анализ

При однофакторном дисперсионном анализе проверяемая гипотеза и альтернатива имеют следующий вид:

$H_0: \mu_1 = \mu_2 = ... \mu_m$,где μ_i (i=1,m) – среднее значение выборки фактора

H_1: по крайней мере, одно среднее значение отлично от остальных.

В однофакторном дисперсионном анализе общая дисперсия разбивается на две составляющие: дисперсию внутри каждого множества повторений выборки фактора (внутривыборочную дисперсию) и дисперсию между сравниваемыми выборками фактора (межвыборочную дисперсию).

В математической статистике разработана формализованная процедура однофакторного дисперсионного анализа которая приведена в табл.2.1 /6/.

Последняя содержит перечень источников изменчивости, столбец сумм квадратов, соответствующих различным источникам, число степеней свободы для каждой из них, столбец средних квадратов, который содержит выборочные оценки дисперсий и значений F - критерия.

Таблица 2.1

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	F-критерий
Между выборками	SS_A	$m-1$	MS_A	MS_A/MS_W
Внутри выборок	SS_W	$N-m$	MS_W	
Общая изменчивость	SS_T	$N-1$		

Общая изменчивость по всем наблюдениям (по всем повторениям и по всем выборкам фактора) SS_T характеризуется формулой

$$SS_T = \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij}^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2 , \qquad (2.1)$$

где X_{ij} i-е повторение (i=1, 2,..., n) в j- ой выборке фактора (j= 1,2,..., m).

В двойной сумме первая указывает, что суммирование проводится по каждой выборке фактора, содержащей n повторений, а затем складываются полученные результаты всех m выборок фактора. Общее число наблюдений N равно сумме повторений по выборкам фактора.

Последний член в правой части выражения (2.1) называется поправочным. Отметим, что такие же члены имеются и в других аналогичных суммах.

Сумму, характеризующуюся межвыборочной изменчивостью, находят по следующей формуле:

$$SS_A = \sum_{j=1}^{m} \frac{1}{n}\left(\sum_{i=1}^{n} X_{ij}\right)^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2 , \qquad (2.2)$$

где суммирование проводится по всем повторениям в каждой выборке фактора $\sum_{i=1}^{n} X_{ij}$ а затем каждая из полученных сумм возводится в квадрат и полученный результат делится на число повторений n в каждой выборке; далее полученные результаты суммируются по всем выборкам фактора и, наконец, вычитается поправочный член. Межвыборочный источник изменчивости относится к контролируемым факторам, влияющим на результат.

Величина, характеризующая второй источник изменчивости, имеет вид

$$SS_w = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} X_{ij}\right)^2 . \qquad (2.3)$$

Заметим, что первый член в правой части здесь такой же как и первый член формулы (2.1) для SS_T а последний член совпадает с первым членом формулы (2.2) для SS_A. Поэтому SS_W можно вычислить по формуле

$$SS_W = SS_T - SS_A . \qquad (2.4)$$

Число степеней свободы по всем данным равно N-1. Число степеней свободы для величины SS_A равно m-1, так как мы оцениваем ее по средним значениям каждой выборки фактора. Этот источник изменчивости (внутри выборок) объясняет неконтролируемое влияние фактора или случайными обстоятельствами.

Разность между этими числами степеней свободы равна числу степеней свободы для величины SS_W.

С целью иллюстрации этого метода дисперсионного анализа приведем пример из работы /8/.

Для выяснения влияния заработной платы на производительность труда шести однородным группам разного объема были предложены задачи одинаковой сложности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличались между собой величиной денежного вознаграждения за решаемую задачу. В табл.2.2. сведено количество решенных задач членами каждой группы.

Проверим гипотезу об отсутствии влияния денежного вознаграждения на работоспособность или, как в работе /8/ , – "нулевая гипотеза утверждает, что разница значений показателя разных уровней фактора равна нулю".

(Часто реализацию фактора называют уровнем).

Таблица 2.2

Влияние вознаграждения (от меньшего к большему)					
группа 1	**группа 2**	**группа 3**	**группа 4**	**группа 5**	**группа 6**
10	8	12	12	16	19
11	10	17	15	22	18
9	16	14	16	18	24
13	13	9	16	20	23
7	12	13	13		27
8		16	19		25
9					24
					22

Используя формулу (2.1) для SS_T, получим SS_T = 1024,89.

Далее мы можем подсчитать величину SS_A по (2.2). SS_A=216,35.

Наконец, вычитая SS_A из SS_T , получаем внутривыборочную сумму квадратов SS_W = 808,54.

Общее число степеней свободы равно N-1, или 35 (см. табл.2.2).Так как мы оцениваем межвыборочную изменчивость по шести измерениям (по средним значениям шести столбцов), то число степеней свободы для SS_A равно m-1, т. е. 5. Разность чисел степеней свободы должна соответствовать остатку сумм квадратов. Эта разность чисел степеней свободы равна N-m или 30. Теперь вычисленные исправленные суммы квадратов SS_A, SS_a, SS_W нужно разделить на соответствующие им числа степеней свободы. В результате мы получаем оценки дисперсий.

Оценка межвыборочной дисперсии 808,54/5 = 161,7.

Оценка внутривыборочной дисперсии 216,35/30 = 7,21.

Вычислив F - критерий, мы получаем значение равное 22.42. Выбрав критическую область, соответствующую заданному уровню значимости и заданному числу степеней свободы, можно теперь принять или отвергнуть прове-

ряемую гипотезу – в этом примере $F>F$ *кр. (22,42>2,53)* . *Fкр.* =2.53 при α =0.05 и степенях свободы равным 5 и 30 /1,6/.

Следовательно, гипотезу об отсутствии влияния денег на работоспособность подлежит отклонению и принимается альтернативная гипотеза, т.е. гипотеза о значимом варьировании средних по выборкам фактора и, следовательно, фактор оказывает существенное влияние на результат – в примере на количество решенных задач представителями шести групп существенно влияет материальный стимул.

В табл.2.3 по данным табл.2.2 показан результат однофакторного дисперсионного анализа, рассчитанного на Excel для примера о влиянии вознагражде-ния на производительность труда (*df* – степени свободы).

Таблица 2.3

Однофакторный дисперсионный анализ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
1	7	67	9,571428571	3,95238095		
2	5	59	11,8	9,2		
3	6	81	13,5	8,3		
4	6	91	15,16666667	6,16666666		
5	4	76	19	6,66666666		
6	8	182	22,75	9,07142857		
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Знач.*	*F критич.*
Между группами	808,541269	5	161,708254	22,4233926	2,606E-09	2,533553811
Внутри групп	216,347619	30	7,211587302			
Итого	1024,88888	35				

Помимо основного заключения (влияние материального стимулирования на результат) можно сделать следующие выводы /8/:

– отношение SS_A / SS_T = 808,54/ 1024,88 * 100% = 78,89% дает возможность утверждать, что фактор оплаты труда имеет свыше 78%-ный вклад в производительность труда;

– отношение SS_w / SS_T = 216,34/1024,88 *100% = 21,11 % "говорит" о том, что свыше 21% относится на вклад неконтролируемых факторов, влияющих на производительность труда (например, соображения карьерного роста, страх потери работы и.т.д.).

Здесь же можно утверждать, что 78,11% работников на изменение заработной платы будут реагировать с вероятностью $1 - \alpha = 1 - 2.6*10^{-9} = 99,9\%$. P – значение, равное 2,606E – 09 (см. табл.2.3) – есть вероятность нашей ошибки.

2.2. Двухфакторный дисперсионный анализ

В задачах этого анализа предполагается, что на результат могут влиять

два фактора, каждый из которых имеет конечное число значений и дает возможность определения степени влияния этих факторов на конечный результат, если такое влияние есть.

Имеется две разновидности двухфакторного дисперсионного анализа в зависимости от того, производились ли повторные измерения при каждом сочетании значений двух исследуемых факторов или нет.

2.2.1. Двухфакторный дисперсионный анализ без повторений

В двухфакторном дисперсионном анализе без повторных измерений исходные данные должны представлять собой матрицу размером $n*m$, в которой столбцы отвечают различным уровням первого фактора j= 1,...,*m*, строки отвечают различным уровням второго фактора (эффектам) i= 1,...,*n*, а каждая ячейка содержит один результат (отклик), измеренный при соответствующем сочетании уровней исследуемых факторов.

В этом случае проверяются две нулевые гипотезы:

$$H_0^1 : \mu_1^1 = \mu_2^1 = ... = \mu_m^1,$$

$$H_0^2 : \mu_1^2 = \mu_2^2 = ... = \mu_n^2,$$

где μ_i^1 (i=1,m) – средние значения первого фактора, μ_i^2 (i=1,n) – средние значения второго фактора.

Соответствующие альтернативные гипотезы заключаются в том, что средние значения для групп откликов, измеренные при различных уровнях (значениях) факторов имеют различия. (Результирующий признак, на который факторы оказывают влияние, часто называют откликом.)

Схема двухфакторного дисперсионного анализа без повторений приведена в табл.2.4.

Таблица 2.4

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	Значения *F*-критерия
Между выборками	SS_A	$m-1$	MS_A	MS_A/MS_e[а]
Между эффектами	SS_B	$n-1$	MS_B	MS_B/MS_e[б]
Ошибка	SS_e	$(m-1)(n-1)$	MS_e	
Общая изменчивость	SS_T	$N-1$		

а – значения F-критерия “между выборками” (строками, фактор 1)
б – значения F-критерия “между эффектами” (столбцами, фактор 2)
ошибка – неучтенные (неконтролируемые) факторы.

Величина SS_T вычисляется по формуле (2.1); SS_A по формуле (2.2). Величина SS_B является суммой квадратов по эффектам (столбцам):

$$SS_B = \sum_{i=1}^{n} \frac{1}{m}\left(\sum_{j=1}^{m} X_{ij}\right)^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2, \qquad (2.5)$$

где m , n- значения размерности факторов.

$$SS_e = SS_T - (SS_A + SS_B). \qquad (2.6)$$

Ошибка суммы квадратов SS_E находится по формуле:

Работу двухфакторного дисперсионного анализа без повторений рассмотрим на примере из /7/.

Переменные, представленные в табл.2.5, представляют урожайность пяти сортов картофеля (ц/га), выращенных на шести участках одинакового размера и почвенного состава, причем каждый из этих участков обрабатывался одним из шести сортов удобрений.

Таблица 2.5

	Сорта картофеля				
Участки	1	2	3	4	5
1	6	9	6	2	6
2	4	7	8	3	5
3	9	3	10	7	4
4	8	4	14	4	10
5	15	11	13	9	14
6	12	14	15	11	9

Необходимо выяснить, различна ли в среднем урожайность разных сортов картофеля независимо от применяемого удобрения и различна ли эффективность используемых удобрений независимо от сорта.

Проверим гипотезы об отсутствии влияния сорта картофеля и видов удобрения на урожайность с использованием данных табл.2 .5.

Решение задачи проведено с помощью Excel, результат представлен в табл. 2.6.

Таблица 2.6

Двухфакторный дисперсионный анализ без повторений

ИТОГИ	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Строка 1	5	29	5,8	6,2
Строка 2	5	27	5,4	4,3
Строка 3	5	33	6,6	9,3
Строка 4	5	40	8	18
Строка 5	5	62	12,4	5,8
Строка 6	5	61	12,2	5,7
Столбец 1	6	54	9	16
Столбец 2	6	48	8	17,6
Столбец 3	6	66	11	12,8
Столбец 4	6	36	6	12,8
Столбец 5	6	48	8	14

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Строки	248	5	49,6	8,406779	0,0002046	2,710891
Столбцы	79,2	4	19,8	3,355932	0,0295208	2,866080
Погрешность	118	20	5,9			
Итого	445,2	29				

Результаты расчета, представленные в табл.2.6 позволяют сделать следующие выводы:

- влияние видов удобрений (“строки” в табл. 2.6) значимо отражается на урожайности $F > F_{кр.}$ (8.406>2.710) при вероятности ошибки 0.0002<0.05 (α= 0.05 – заданный уровень значимости);

- влияние сорта картофеля (“столбцы” в табл. 2.6) также значимо отражается на урожайности . $F > F_{кр.}$ (3.335 >2.866) при вероятности ошибки 0.029<0.05) (α= 0.05 – заданный уровень значимости);

Используя данные табл.2.6 рассчитаем степень влияния факторов на урожайность по схеме, аналогичной при подобных вычислениях в однофакторном анализе.

Степень влияния факторов:

- вид удобрения (строки) – SS_A / SS_T *100%=248/445,2*100% =55,705%.

- Сорт картофеля (столбцы) – SS_B / SS_T *100%=79,2/445,2*100%=17,789%.

- неконтролируемые факторы (ошибка) – SS_e / SS_T *100%=118/445,2*100% =26,504.

2.2.2. Двухфакторный дисперсионный анализ с повторениями

Двухфакторный дисперсионный анализ с повторениями дает возможность исследовать влияние на результат (отклик) не только контролируемых факторов, но и их взаимодействия (иногда говорят наложения).

Схема двухфакторного дисперсионного анализа с повторениями показана в табл.2.7 и в /8/.

Таблица 2.7

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	Значения F – критерия
Между выборками	SS_A	m-1	MS_A	MS_A / MS_e
Между эффектами	SS	n-1	MS_B	MS_B / MS_e
Взаимодействие	SS_{AB}	(m-1)(n-1)	MS_{AB}	MS_{AB} / MS_e
Ошибка	SS_e	N-m*n	MS_e	
Общая изменчивость	SS	N-1		

Здесь мы не будем показывать формулы расчета сумм квадратов SS и их средних значений MS (табл.2.7) и сразу перейдем к иллюстрации работы двухфакторного дисперсионного анализа с повторениями на примере. В качестве примера рассмотрим задачу проверки влияния возраста и стажа работников определенной специальности на производительность труда. Это результаты обследования 60 работников производства, у которых фиксировалась средняя часовая выработка в натуральных единицах продукции /9/. Данные обследования показаны в табл.2.8.

Таблица 2.8

Стаж	**Возраст (фактор А)**		
(фактор В)	**от 25 до 35**	**от 35 до 45**	**от 45 до 55**
от 1 до 4 лет	19	19	18
	20	20	19
	20	20	20
	20	23	21
	22	25	23
от 4 до 7 лет	30	20	19
	31	29	25
	32	30	25
	32	31	26
	34	31	26
от 7 до 10 лет	35	36	24

	35	40	24
	39	41	24
	40	42	25
	41	45	25
Свыше 10 лет	40	28	20
	40	31	24
	41	35	25
	41	36	31
	42	40	32

Результат расчета двухфакторного дисперсионного анализа с повторениями по данным табл.2.8 приведен в табл.2.9.

Таблица 2.9

ИТОГИ	от 25 до 35	от 35 до 45	от 45 до 55	Итого
от 1 до 4 лет				
Счет	5	5	5	15
Сумма	101	107	101	309
Среднее	20,2	21,4	20,2	20,6
Дисперсия	1,2	6,3	3,7	3,5428571
от 4 до 7 лет				
Счет	5	5	5	15
Сумма	159	141	121	421
Среднее	31,8	28,2	24,2	28,06666
Дисперсия	2,2	21,7	8,7	19,63809
от 7 до 10 лет				
Счет	5	5	5	15
Сумма	190	204	122	516
Среднее	38	40,8	24,4	34,4
Дисперсия	8	10,7	0,3	60,4
свыше 10 лет				
Счет	5	5	5	15
Сумма	204	170	132	506
Среднее	40,8	34	26,4	33,73333
Дисперсия	0,7	21,5	25,3	50,63809
Итого				
Счет	20	20	20	
Сумма	654	622	476	
Среднее	32,7	31,1	23,8	
Дисперсия	68,5368	66,6210	13,32631	

Дисперсионный анализ				
Источник вариации	**SS**	**df**	**MS**	**F**
Выборка	1842,53	3	614,1777	66,81897
Столбцы	900,4	2	450,2	48,97914
Взаимодействие	537,467	6	89,57777	9,745542
Внутри	441,2	48	9,191666	
Итого	3721,6	59		

Рассмотрим “выдачу” результатов расчетов при влиянии стажа (фактор В, “выборка” в табл.2.9) и возраста (фактор А, “столбцы ” табл.2.9) на уровне α = 0.05.

Фактор “стаж” значимо влияет на производительность, т.к.

$$F > F_{крит.} (66,81>2,79),$$

фактор “возраст ” имеет влияние (48,97 > 3,19) , взаимодействие стажа и возраста (факторы А и В) также влияет на конечный результат (9,74>2,29).

Влияние факторов в процентном отношении на производительность труда следующее (см. табл. 2.9):

- стаж (1842,53/3721.6*100%) = 49,51% ,
- возраст (900,4/3721,6 *100%) = 24,19% ,
- взаимодействие (537,46/3721,6 * 100%) = 14,44% ,
- неконтролируемые факторы (“внутри”) (441,2/3721,6 * 100%) = 11,85%.

Влияние это по всем факторам практически имеет доверительную вероятность сто процентов (Р – значения равны $3,702*10^{-17}$, $2,560*10^{-12}$, $5,197*10^{-07}$).

В этом примере мы выяснили, что факторы “стаж” и ”возраст” влияют на производительность труда, но мы не выяснили “как” влияют эти факторы – положительно или отрицательно.

Используя данные табл.2.9, попробуем решить эту задачу. Выберем из табл.2.9 средние значения выработки продукции по категориям стажа для каждого возрастного интервала. Эти данные внесем в табл.2.10.

Таблица 2.10

Стаж/ Возраст	**от 25 до 35 лет**	**от 35 до 45 лет**	**от 45 до 55 лет**
От 1 до 4 лет	20.2	21.4	20.2
От 4 до 7 лет	31.8	28.2	24.2
От 7 до 10 лет	38.0	40.8	24.4
Свыше 10 лет	40.8	34.0	26.4

На основании этих данных построим графики. График (рис.2.1) отражает взаимодействие возраста и стажа. Из графика видно, что средняя часовая выработка увеличивается с ростом стажа у молодых работников (25–35 лет), для возрастной группы 35 – 45 лет производительность труда растет, если стаж не превышает 10 лет, далее производительность падает. Для третьей возрастной группы (45 –55) лет самая низкая производительность труда независимо от стажа работы.

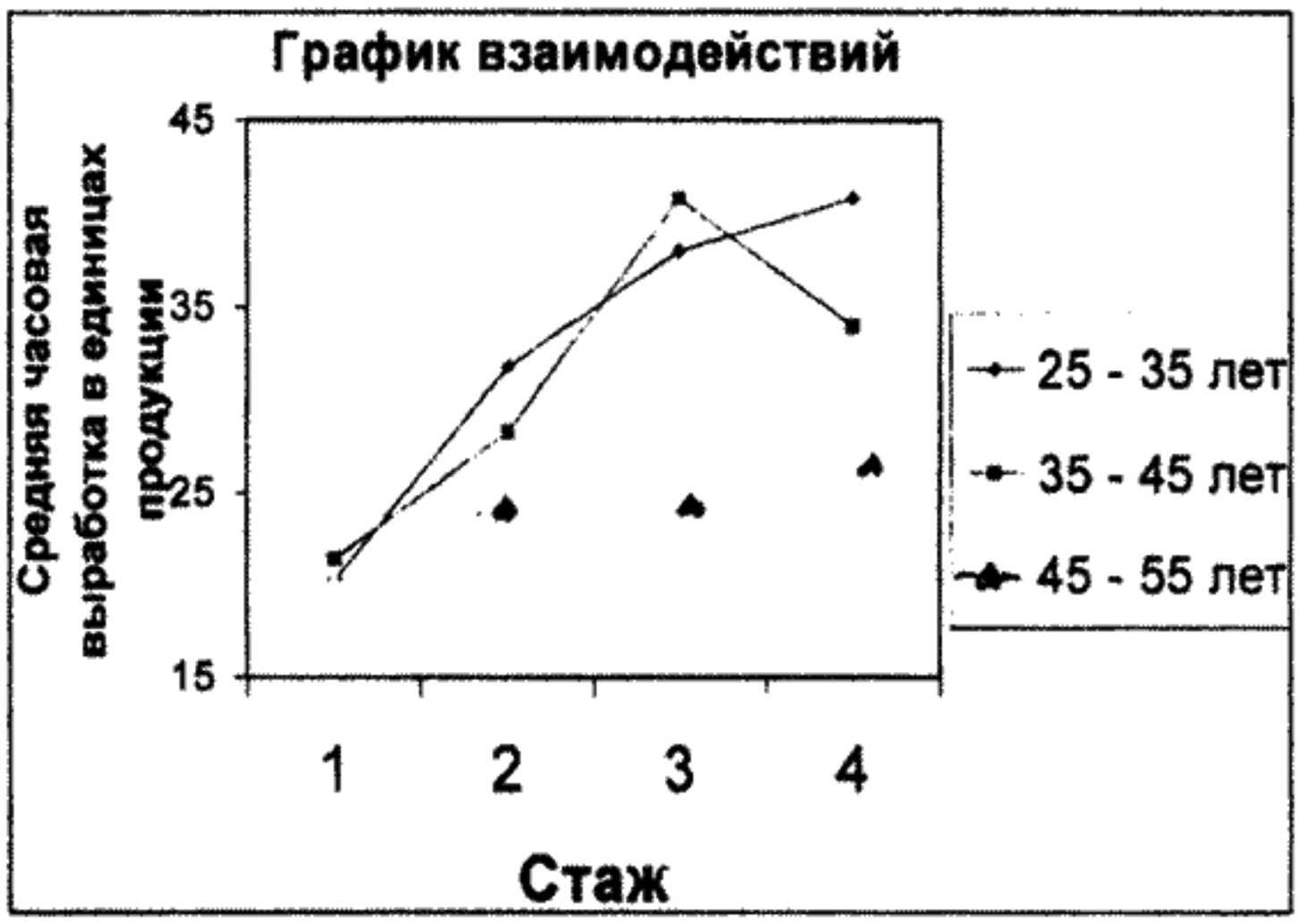


Рис. 2.1

На графике (рис 2.2) видно, что наибольшая производительность наблюдается у молодых людей со стажем работы свыше 10 лет и людей среднего возраста со стажем от 7 до 10 лет, и что при маленьком стаже (от 1 до 4 лет) производительность труда самая низкая независимо от возраста.

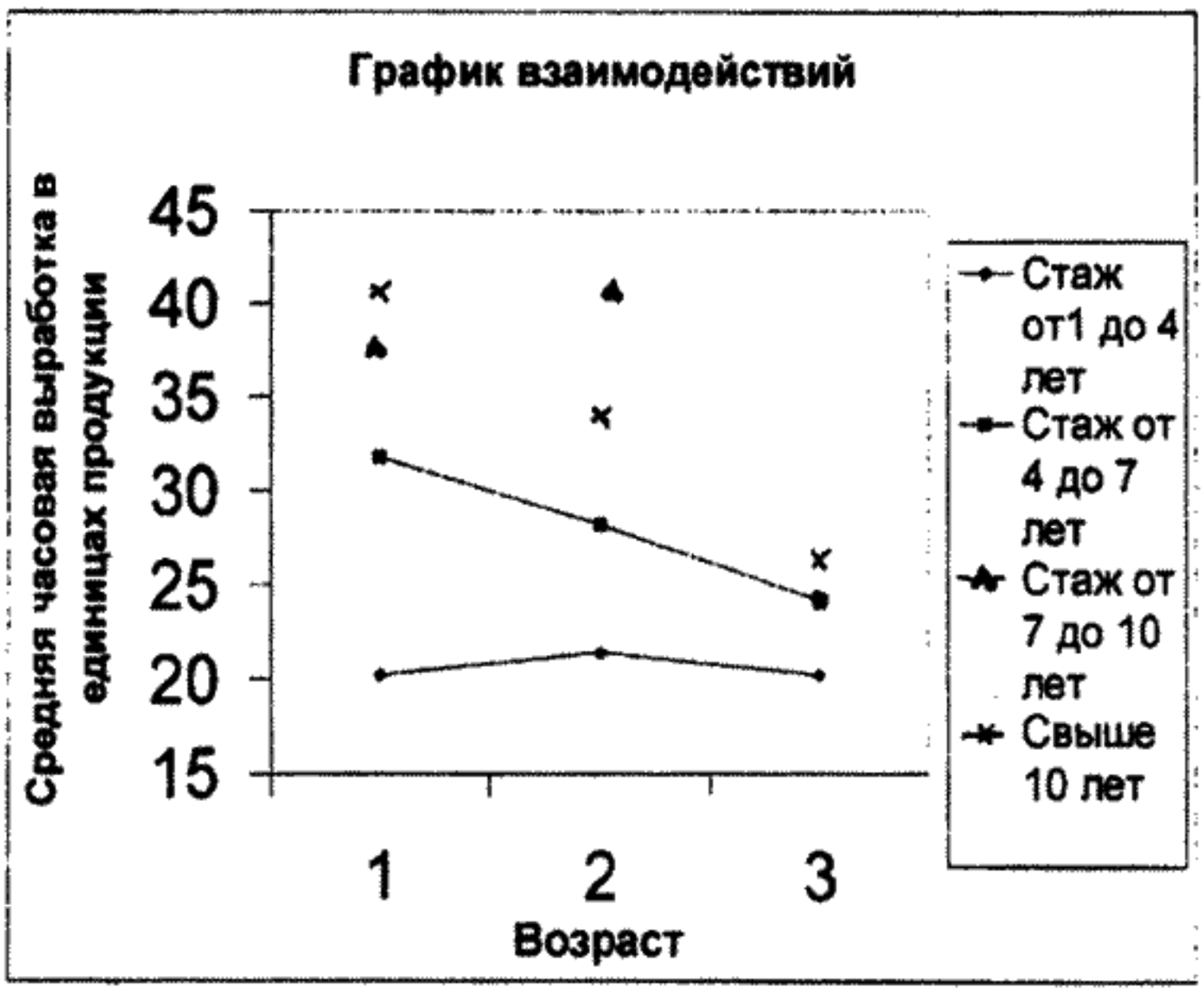


Рис 2.2

(Выводы справедливы к исследованной генеральной совокупности работников определенной специальности /9/).

3. РЕГРЕССИОННЫЙ АНАЛИЗ

Как мы уже знаем, установить наличие связи между признаками возможно с помощью коэффициента корреляции. Однако корреляционный анализ не дает информации о характере связи. Связь между признаками можно выяснить с помощью регрессионного анализа.

Для многих задач с определяемыми количественными переменными представляет интерес исследования влияния некоторых переменных на другие.

Обычно существующая функциональная связь слишком сложна для описания, задача регрессионного анализа при этом состоит в подборе упрощенной аппроксимации этой связи с помощью математической модели. Регрессионный анализ имеет в своем распоряжении специальные процедуры проверки, является ли выбранная математическая модель адекватной для описания имеющихся данных. При исследовании такой приближенной математической модели можно больше узнать об изучаемой истинной зависимости. Даже если по физическому смыслу между переменными не существует реальной связи, отражение ее с помощью математического уравнения может быть полезно (например, для уменьшения пространства исходных признаков). Таким образом, можно сказать, что регрессия часто используется при попытках установить причинную связь. Еще одно возможное использование регрессии – количественное измерение эффекта с помощью коэффициента регрессии. Однако чаще всего регрессионный анализ используется для прогноза, т.е. предсказания значений ряда зависимых переменных по известным значениям других переменных. Коротко суть основной задачи регрессионного исчисления можно сформулировать следующим образом: как по величине переменной *X* можно судить о величине переменной *Y /5/.*

Простейшим примером процедуры регрессионного анализа является часто возникающая практическая задача — подбор прямой по парам наблюдений (*Xi , Yi, i=1,n).*

Если задача включает большее число переменных-предикторов (или регрессоров, или параметров прогноза), то она называется многофакторной. Относительно регрессионного анализа говорят о линейности или нелинейности модели. Величина наивысшей степени регрессора (переменной) в модели называется порядком модели.

(Пример: $Y = b_0 + b_1 X + b_2 X^3$ – уравнение модели третьей степени).

3.1. Прямолинейная связь между двумя переменными

Итак, связь между зависимой случайной величиной *Y* и величиной *X,* которая является переменной (но не случайной переменной), выражается уравнением регрессии *Y* относительно *X.* Мы не случайно оговорили, что переменная-предиктор *X* не подвержена случайной вариации, тогда как переменная отклика *Y* подвержена. В практическом смысле такое предположение редко выполняется, однако, если это не так, то требуются более сложные математические мето-

ды построения зависимостей, даже в случае однофакторной модели. Поэтому всегда полезно по возможности организовать эксперимент так, чтобы разброс истинного значения предикторной переменной (или диапазон ее изменения) существенно превышал разброс случайных ошибок, содержащихся, вероятно, в этой переменной. Тогда ошибками, содержащимися в предикторной переменной, можно будет пренебречь и пользоваться обычными методами регрессионного анализа.

Наиболее простой вариант линии регрессии переменной Y от переменной X имеет вид:

$$Y = b_0 + b_1 X + \varepsilon.$$

Это уравнение представляет собой линейную по регрессору X однофакторную математическую модель, ε – случайная ошибка модели. Обычно с помощью метода наименьших квадратов /6/ на основе имеющихся данных идентифицируются коэффициенты полинома b , b_1, являющиеся выборочными оценками соответствующих параметров модели β_0, β_1. Тогда в качестве предсказывающего можно использовать уравнение: $\bar{\bar{Y}} = b_0 + b_1 X$, две черты над символом Y означают предсказанное значение Y для данного X при определенных значениях регрессионных параметров.

Метод наименьших квадратов предполагает идентификацию неизвестных параметров модели в соответствии с минимизацией функционала качества приближения. Слагаемые такого функционала представляют собой квадрат отклонений реальных значений переменной отклика Y от соответствующего модельного значения /5,6, 11/.

Не вдаваясь в вычислительные тонкости, скажем только, что метод наименьших квадратов дает оценки b_0, b_1, коэффициентов полинома β_0, β_1. Обычно получение таких оценок проводят с помощью соответствующей компьютерной программы. В наиболее простом случае линейной однофакторной модели оценки коэффициентов могут быть рассчитаны по следующим формулам /5/:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \left[\sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i\right]/n}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n} = \frac{\sum_{i=1}^{n} (X_i - Xcp)(Y_i - Ycp)}{\sum_{i=1}^{n} (X_i - Xcp)^2},$$

$$b_0 = Ycp - b_1 \cdot Xcp,$$

где $Xcp., Ycp.$ – средние арифметические значения наблюдений X и Y соответственно.

Для графического изображения задачи используется прямоугольная система координат, любой паре значений (Xi , Yi) соответствует точка в регрессионной области. Через скопление точек на регрессионном графике нужно провести прямую так, чтобы, исходя из значений X, можно было бы как можно точнее оценить значения Y (см. рис.3.1).

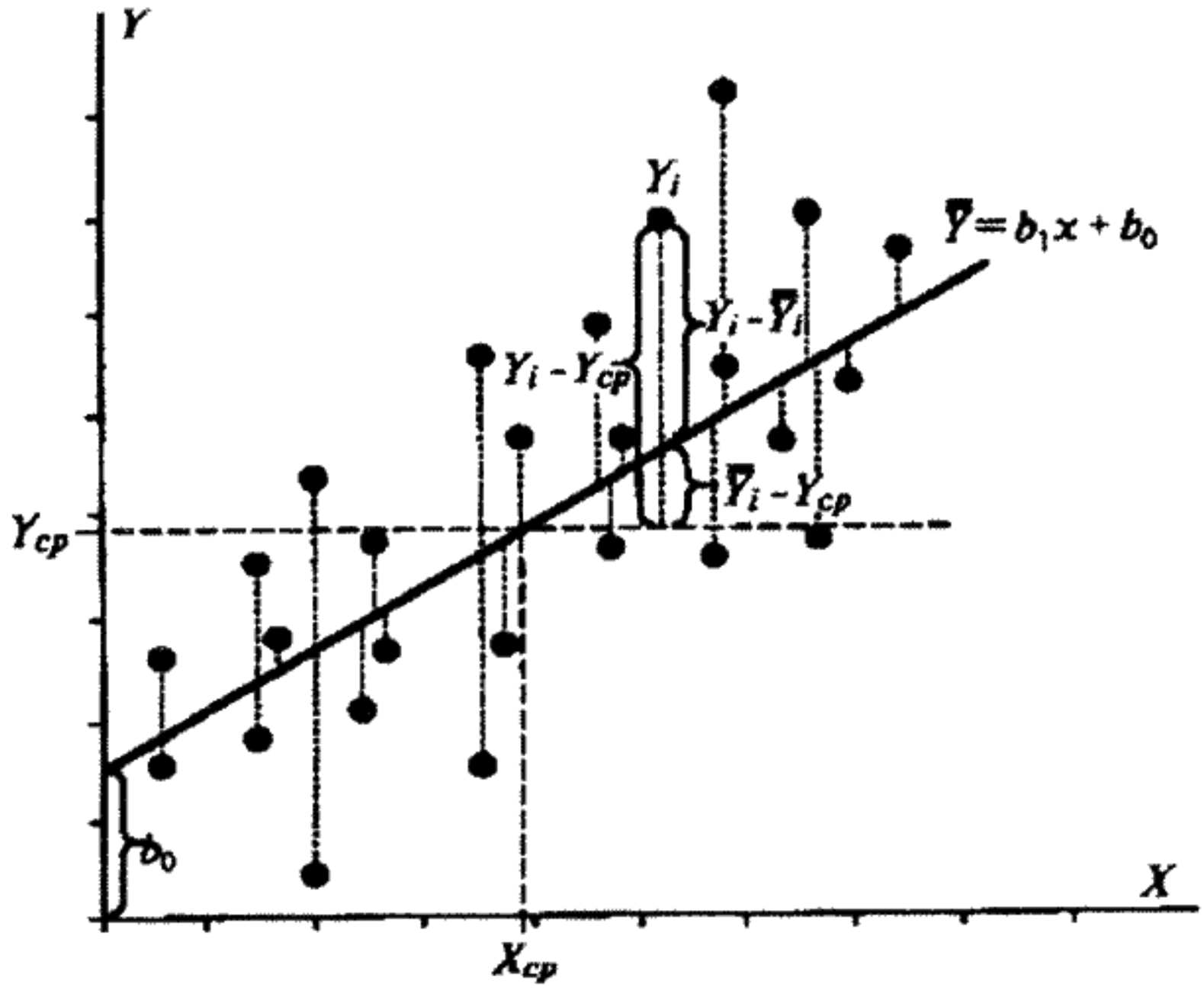


Рис.3.1. Прямая регрессия *Y* по *X* /5/

Пунктиром обозначены отклонения наблюдаемых значений от линии регрессии *(Yi – $\overline{Y}$ i)*. Величина ($\overline{Y}$ *i – Ycp)* является отклонением предсказанного по модели значения от среднего значения; величина *Yi— Ycp* является отклонением измеренного значения отклика от среднего.

3.2. Точность оценки регрессии

Рассмотрим вопрос о том, какая точность может быть приписана нашей оценке линии регрессии. Точность аппроксимации данных регрессионной моделью оценивается с помощью анализа остатков, т.е. разностей между наблюдаемыми и предсказанными по модели значениями. Для этого представим остаток ε_i = Yi – $\overline{Y}$ i в виде разности двух величин:

1) отклонение наблюдаемого значения отклика Y_i от среднего откликов *Y*ср.;
2) отклонение предсказанного значения отклика $\overline{Y}$ i от того же самого среднего значения *Ycp.*

Если рассмотреть все *n* наблюдений, то можно выразить сумму квадратов отклонений наблюдений Y_i от среднего в виде двух основных слагаемых: суммы квадратов отклонений наблюдаемых значений *Yi* относительно регрессии и суммы квадратов отклонений регрессионных значений относительно среднего. Второй член в правой части характеризует вариацию, связанную с регрессией, и объясняет разброс за счет исследуемого фактора. Первое слагаемое в этой сумме является «необъяснимой» вариацией, отклонения отражают влияние случайных факторов, и эта вариа-

ция обычно называется остаточной (рис.3.1). Пригодность линии регрессии зависит от соотношения этих слагаемых. Тогда, воспользовавшись методами дисперсионного анализа, строим таблицу дисперсионного анализа /3,5,6/ (табл.3.1).

При проведении регрессионного анализа с помощью различных программ часто можно встретиться с *величиной* R^2, ее обычно называют *коэффициентом детерминации.*

Эта величина измеряет долю общего разброса относительно среднего (ss^2), объясняемую регрессией $(\sum_{I=1}^{N}(\overline{\overline{Y}}\text{-}Y\ cp.)^2)$. Таблица дисперсионного анализа может помочь вычислить искомую величину R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(\overline{\overline{Y}}_i - Ycp)^2}{\sum_{i=1}^{n}(Y_i - Ycp)^2}.$$

Таблица 3.1

Общий вид таблицы дисперсионного анализа для оценки точности регрессии

Источник вариации	Число степеней свободы	Сумма квадратов *SS*	Средние квадраты
Обусловленный регрессией	1	$\sum_{i=1}^{n}(\overline{\overline{Y}}_i - Ycp)^2$	$\sigma_1^2 = \sum_{i=1}^{n}(\overline{\overline{Y}}_2 - Ycp)^2/1$
Относительно регрессии (остаток)	$n-2$	$\sum_{i=1}^{n}(Y_i - \overline{\overline{Y}}_i)^2$	$\sigma^2 = \sum_{i=1}^{n}(Y_i - \overline{\overline{Y}}_i)^2/(n-2)$
Общий	$n-1$	$ss^2 = \sum_{i=1}^{n}(Y_i - Ycp)^2$	

Фактически R – это корреляция между наблюдаемыми значениями Y и предсказанными (рассчитанными) значениями $\overline{\overline{Y}}$. Коэффициенты регрессионного уравнения являются случайными величинами и, по имеющимся данным, мы находим лишь их выборочные оценки.

С помощью таблицы дисперсионного анализа можно оценить дисперсию коэффициентов β_0 и β_1 регрессионной модели /5,6,11,12/ . Так, оценка дисперсии β_0 равна /5/:

$$D[b_0] = (\sigma^2 \cdot \sum_{i=1}^{n} X_i^2) / (n \cdot \sum_{i=1}^{n}(X_i - Xcp)^2).$$

Оценка дисперсии β_1 равна:

$$D[b_1] = \sigma^2 / \sum_{i=1}^{n} (X_i - Xcp)^2.$$

Оценка величины остаточного стандартного отклонения σ содержится в таблице дисперсионного анализа (см. табл.2.11.). Корень квадратный из дисперсии $D[\beta_0]$ и $D[\beta_1]$ задает соответствующие стандартные ошибки оценок регрессионных коэффициентов.

С помощью дисперсионного анализа можно, кроме того, проверить гипотезу о равенстве нулю коэффициента β_1 в уравнении регрессии. Для проверки справедливости нулевой гипотезы $H_0 : \beta_1 = 0$ нужно по таблице дисперсионного анализа вычислить F-отношение $F = \sigma_1^2 / \sigma^2$ и проверить по таблице значений F - критерия Фишера /1,5,6/ для выбранного уровня значимости . Если рассчитанное значение F превосходит табличное, нулевая гипотеза отвергается на выбранном уровне α. Таким образом проверяется значимость выбранного уравнения регрессии. Проверить значимость рассчитанного коэффициента β_1 можно и с помощью t -критерия Стьюдента. Для этого формулируется нулевая гипотеза $H : \beta_1 = 0$ и альтернатива к ней.

Рассчитываем статистику критерия как отношение оценки коэффициента β_1 к оценке его стандартной ошибки:

$$t = \frac{b_1 * (\sum_{i=1}^{n} (X_i - X_{cp})^2)^{1/2}}{\sigma},$$

и сравниваем полученную величину с табличным значением для выбранного уровня значимости и числом степеней свободы $n - 2$. Если рассчитанное значение превосходит табличное (таблица критических значений t -критерия Стьюдента в /1,5,6/) , нулевая гипотеза отвергается на выбранном уровне значимости.

Аналогично можно проверить и значимость оценки свободного члена β_0 в уравнении регрессии. Даже если априори известно, что данная линия регрессии должна проходить через начало координат, лучше исходить из того, что модель содержит "ненулевой" свободный член. Получив выборочную оценку для коэффициента β_0, нужно проверить гипотезу о его значимости (другими словами, проходит ли данное уравнение регрессии через начало координат). Для этого рассчитывается t – статистика критерия Стьюдента в виде:

$$t = \frac{b_0}{\left\{\dfrac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - Xcp)^2}\right\}^{1/2} \cdot \sigma}.$$

Для значений коэффициентов регрессии β_0, β_1 по их выборочным оценкам b_0 и b_1 могут быть рассчитаны доверительные интервалы по формуле $b_j \pm t(n-2,\ 1\text{-}0.5^*\alpha)^*\sqrt{D(b_j)}$, где $t(n\text{-}2,\ 1\text{-}\ 0{,}5^*\alpha)$ – коэффициент Стьюдента, определяемый по таблице Стьюдента /1,5,6/, $(n\ \text{-}2)$ – число степеней свободы, индекс j может принимать значения 0 и 1 для обозначения коэффициентов уравнения регрессии, табличное значение t выбирается с учетом двустороннего доверительного интервала α, обычно равно 5%. Кроме того, возможно построение совместной доверительной области для параметров β_0 и β_1 /2,5,8,11,12/.

3.3. Доверительные интервалы уравнения регрессии

Выражения для дисперсий коэффициентов β_0 и β_1 используются для построения доверительных интервалов уравнения регрессии. Для любого фиксированного значения x имеет место равенство, дающее оценку дисперсии соответствующего $\bar{\bar{Y}}$:

$$D\left[\bar{\bar{Y}}\right] = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x - Xcp)^2}{\sum_{i=1}^{n}(X_i - Xcp)^2}\right].$$

При любом значении переменной X соответствующие значения переменной отклика Y распределены нормально со средним значением $\bar{\bar{Y}}$. Поэтому по заданному значению x можно построить 95% доверительный интервал для «истинного» среднего значения уравнения регрессии $\bar{\bar{Y}}$: $\bar{\bar{Y}} \pm t^*\sqrt{D\left[\bar{\bar{Y}}\right]}$, где величина коэффициента Стьюдента t определяется для 95% доверительной вероятности и числа степеней свободы, равного n- 2 (таблица в /1,5,6/).

Нужно еще раз подчеркнуть, что таким образом мы задаем доверительные границы для линии регрессии, и построение такой доверительной области производится в связи с тем, что уравнение регрессии строится по выборке значений. Доверительные границы представляют собой кривые – гиперболы, лежащие по обе стороны от линии регрессии. Наименьшую ширину область имеет вблизи значений X, равных $X_{cp.}$, и расширяется при удалении от среднего значения. По мере удаления от «центра» значений X и, тем более за пределами

нашего наблюдения, точность предсказания ухудшается, соответственно и доверительная область становится шире (рис.3.2). С заданной вероятностью, обычно 95%, можно утверждать, что «истинная» линия регрессии находится в границах полученной доверительной области. Не удивительно, что некоторые наблюдаемые значения Y_i лежат вне построенного доверительного интервала (см. рис.3.2). Дело в том, что мы строили доверительную область для линии регрессии, а не для значений переменной отклика, которая получилась бы шире построенной нами доверительной области.

Однако исследователя может интересовать задача построения доверительной области не для уравнения регрессии, а для значений зависимой переменной Y. Такая доверительная область также может быть построена. Ее границы задаются соотношением $\overline{\overline{Y}} \pm t * \sqrt{D[Y]}$, а оценка $D[Y]$ вычисляется по следующей формуле для любого значения переменной X:

$$D[Y] = \sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(x - Xcp)^2}{\sum_{i=1}^{n} (X_i - Xcp)^2} \right].$$

Таким образом, определяется доверительная область, в которую попадает определенный процент (например, 95% при соответствующем выборе коэффициента Стьюдента t) всех значений переменной отклика. Данный интервал называется также интервалом прогноза, поскольку он задает доверительные пределы, между которыми с заданной вероятностью будет находиться новое наблюдение Y, отвечающее заданному значению переменной X /5,6,11,12/.

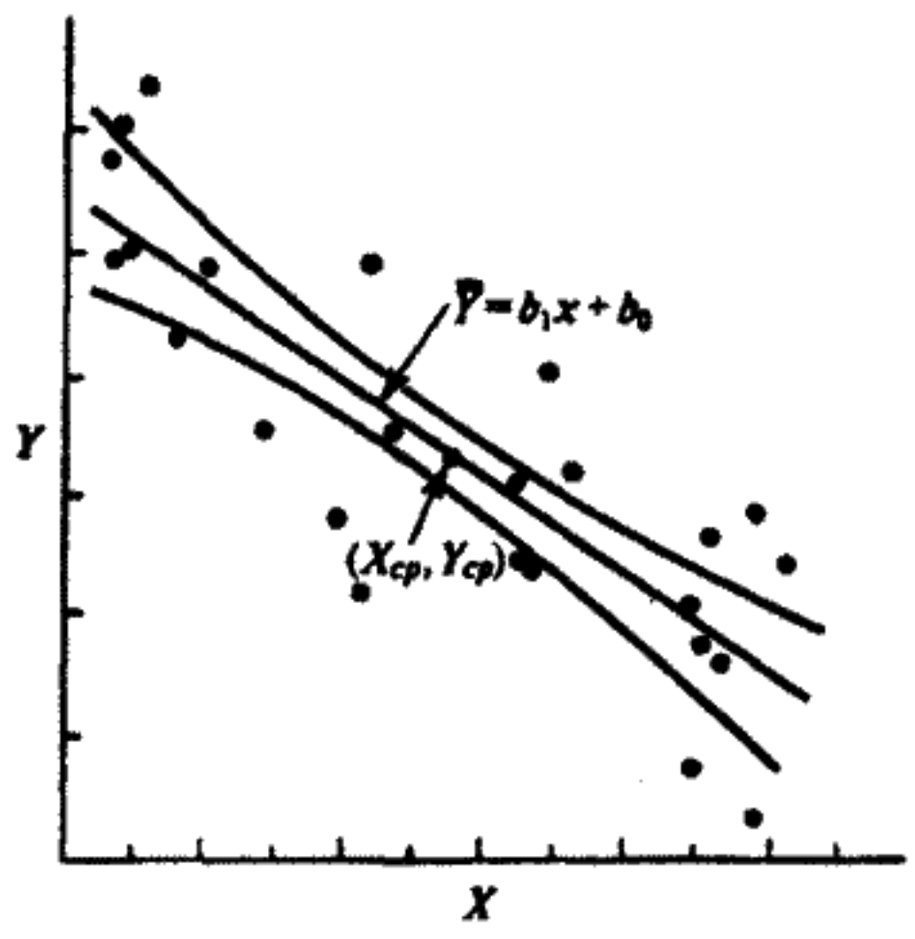


Рис. 3.2 Линия регрессии и соответствующая 95% доверительная область для данной линии регрессии

Таким образом, определяется доверительная область, в которую попадает определенный процент (например, 95% при соответствующем выборе коэффициента Стьюдента t) всех значений переменной отклика. Данный интервал называется также интервалом прогноза, поскольку он задает доверительные пре-

делы, между которыми с заданной вероятностью будет находиться новое наблюдение Y, отвечающее заданному значению переменной X /5,11,12,16/.

Надо обратить особое внимание на связь корреляции и регрессии, поскольку часто эти понятия путают. Корреляционный коэффициент Пирсона r_{xy} учитывает меру линейной зависимости между двумя переменными X и Y. В то время как оценка коэффициента регрессии b_1 измеряет величину изменения переменной Y, которую можно предсказать, если изменение X равно единице. При этом справедливо соотношение:

$$b1 = \left\{ \frac{\sum_{i=1}^{n}(Y_i - Ycp)^2}{\sum_{i=1}^{n}(X_i - Xcp)^2} \right\}^{1/2} \cdot r_{XY} = \frac{\sigma_Y}{\sigma_X} \cdot r_{XY},$$

где σ_Y и σ_X — выборочные оценки средних квадратичных отклонений для выборок X и Y. В корреляционном анализе X и Y – случайные переменные, распределенные по нормальному закону, В случае же регрессионного анализа мы различаем независимую переменную X и зависимую переменную Y.

В заключение, перед рассмотрением примеров использования регрессионного анализа, отметим, что подробно с ним можно ознакомиться по работам / 2,5,6,8,9,11,12,16/.

3.4. Примеры использования регрессионного анализа

Рассмотрим примеры по использованию регрессионного анализа. Простая регрессия (линейная, однофакторная) – нахождение аналитического вида зависимости двух переменных X и Y.

В качестве примера используем данные по исследованию связи между данными бурения (H_m) и данными МОВ (S_m). (МОВ-данные сейсмики)/4/. Данные приведены в табл. 3.2.

Таблица 3.2

Глубины залегания (в метрах) подошвы соленосного комплекса по данным глубокого бурения и сейсморазведки (МОВ).

№	Номер скважины	Данные бурения H (метры)	Данные МОВ S (метры)
1	П-13	2300	2350
2	Г-1	2395	2000
3	П- 11	3103	3050
4	П-38	2861	2640
5	П-88	3750	3400
6	Г-90	3733	3790
7	Г-91	3856	3580
8	Г-93	3762	3400
9	Г-4	4384	3870
10	П-15	4220	3920
11	П-1	4687	4000
12	Г-4-1	5390	4600
13	Г-3	5390	4600
14	Г-7	925	950
15	СГ-2	4880	4840
16	Г-1-1	3140	3290
17	Г-2	3670	3370
18	Г-10	3662	3320
19	П-89	3706	3600
20	Г-1-2	3920	3400
21	Г-1-3	4818	4500

По данным табл.3.2 в Excel строим регрессионную линейную модель. Результат на рис. 3.3.

На рис. 3.3 видно, что получена модель при следующих основных параметрах – R – квадрате равном 0.94, при средней абсолютной ошибке равной 191.6, при "разбросе" значений остатков (исходное значение минус рассчитанное, например, 2300–2496.47 = -196.47) от -430.2 до 350.9 ("размах остатков" 781.1). Некоторые сомнения вызывают, в общем-то неплохой модели, разброс остатков (781.1) и среднее значение остатков (191.6), превышающее 5%(среднее значение $H = 3740.5$ /1/).

Все расчеты проведены с помощью Excel.

Поэтому была построена полиномиальная модель второй степени (рис.3.4).

Эта модель не обладает лучшими статистическими свойствами, чем модель первой степени. Об этом говорят сравнения регрессионных статистик обеих моделей (например, R – квадрат 0,94 в модели второго порядка (второй степени) против значения R – квадрата равного 0.94 в модели первого порядка). Средняя абсолютная ошибка равна 194.4 (против 191.6 в первой модели). Размах ошибки 791.7 (781.1 в первой модели) при разбросе значений остатков от –465.7 до 328.2.

Более подробно результаты регрессионного анализа рассмотрим на примере многофакторного регрессионного анализа (его могут называть еще множественной регрессией).

В предлагаемом учебном примере (табл.3.3) задача состоит в построении модели прогноза глубины залегания кровли келловейских отложений (H2). В качестве параметров прогноза использованы данные залегания кровли (H0) и подошвы (H1) меловых отложений. (Пример заимствован из работы /4/). Модель построена с помощью Excel (рис.3.5).

Таблица 3.3

Глубины залегания стратиграфических границ, км

№	H2	H1	H0
1	2.11	1.7	0.91
2	1.87	1.51	0.78
3	1.91	1.52	0.72
4	1.81	1.42	0.66
5	1.7	1.33	0.52
6	1.28	1.04	0.31
7	1.06	0.93	0.24
8	1.6	1.31	0.56

В исследованиях уровень значимости, на котором отвергается нулевая гипотеза, должен быть равен или меньше α – вероятности ошибки первого рода. Наиболее часто при геологических исследованиях α выбирают равное либо 0.05, либо 0.1, что говорит о бесспорной адекватности модели.

Статистики коэффициентов модели позволяют сделать следующие выводы:

- Y – пересечение (коэффициент равен 0.660) – гипотеза равенства коэффициента нулю отвергается на уровне значимости *0.0503.*

- H1 (коэффициент 2.098*)* отличен от нуля на уровне значимости 0.*002.*

- H0 (коэффициент – 0.841*)* отличен от нуля на уровне значимости 0.*094* .

При α=0.05 коэффициент -0.841 (при H0) гипотеза равенства нулю коэффициента не может быть подвержена отклонению, он не отличается от нуля на уровне значимости 0.05 (0.094).

Свободный член (Y – пересечение) на этом уровне значимости также может быть исключен из модели.

При α=0.1 модель прогноза будет выглядеть следующим образом

$$H2 = -0.660 + 2.098H1 - 0.841H0 .$$

По данным табл.3.3, в качестве примера была построена многофакторная полиномиальная модель. (Листинг решения в Приложении, рис 4.9)

3.5.Тренд–анализ

При изучении регрессионного анализа нельзя не рассмотреть тренд-анализ – геологическое название статистического метода выделения двух компонент (систематической и случайной). Рассмотрим *"в самом первом приближении "что это такое?"*. Графическое изображение пространственных изменений геологических параметров в виде графиков, профилей, карт широко распространено в геологической практике. На этих геологических документах обычно выделяют как направление изменения (возрастания или убывания) изучаемого признака, так и положение аномальных зон (в разрезе или на площади).

Особого внимания заслуживает проблема выделения региональных направлений изменения геологического параметра. Такие направления изменения, например, гранулометрического состава, указывают на положение области денудации – источника сноса; направление регионального увеличения продуктивности нефтеносных структур может быть связано с положением области генерации углеводородов и т. д. В условиях сравнительно простого геологического строения (или слабой изученности) такие региональные направления достаточно уверенно выделяются на соответствующих картах. Однако в более сложных условиях при мозаичном характере распределения локальных аномалий изучаемого геологического признака выделение направлений региональной тенденции его изменения часто представляет трудную задачу, в решение которой обычно вносятся субъективные представления априорных геологических концепций /2,4,6/.

В наиболее общей форме пространственные изменения изучаемого геологического признака могут быть представлены в виде суммы

$$h(x,y) = P(x,y) + e(x,y), \qquad (3.1)$$

где h (x,y) — функция изучаемого геологического параметра; P (x,y) – полином некоторой степени n, приближенно описывающий изменения изучаемого

признака в системе координат; e (x, y) — остаток изменений признака, который не может быть описан многочленом степени меньше n.

Из смысла слагаемых (3.1) следует, что P (x,y) отображает лишь наиболее общие региональные тенденции изменения изучаемого геологического параметра, его регулярную компоненту, остаток e (x,y) – местные изменения параметра под действием локальных факторов, его нерегулярную компоненту.

Выявление региональной тенденции (регулярной компоненты) изменений изучаемого признака и носит название тренд-анализа.

Тренд-анализ заслуживает отдельного рассмотрения, довольно полно он описан в работах /2,4,6,16/.Технику расчета трендов с помощью программы Excel смотрите в задаче 13 (раздел 4.3, Задание б).

№	Скваж.	H (метры)	S (метры)
1	П-13	2300	2350
2	Г-1	2395	2000
3	П- 11	3103	3050
4	П-38	2861	2640
5	П-88	3750	3400
6	Г-90	3733	3790
7	Г-91	3856	3580
8	Г-93	3762	3400
9	Г-4	4384	3870
10	П-15	4220	3920
11	П-1	4687	4000
12	Г-4-1	5390	4600
13	Г-3	5390	4600
14	Г-7	925	950
15	СГ-2	4880	4840
16	Г-1-1	3140	3290
17	Г-2	3670	3370
18	Г-10	3662	3320
19	П-89	3706	3600
20	Г-1-2	3920	3400
21	Г-1-3	4818	4500

ВЫВОД ИТОГОВ

Регрессионная статистика

Множ. R	0.973758429
R-квадрат	0.948205477
Норм.R-квад.	0.94547945
Станд.ошибка	249.5928162
Наблюдения	21

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значим. F*
Регрессия	1	21668876.24	21668876.24	347.8341565	1.12973E-13
Остаток	19	1183634.904	62296.57391		
Итого	20	22852511.14			

	Коэффициенты	*Стандарт.ошибка*	*t-статист.*	*P-Знач.*	*Нижн. 95%*	*Верхн. 95%*
Y-пересечен.	-159.0653836	216.0696534	-0.736176419	0.470615466	-611.3045061	293.1737389
S (метры)	1.130017567	0.060589739	18.6503125	1.12973E-13	1.003201746	1.256833387

ВЫВОД ОСТАТКА

	Наблюд.	*Предсказ. H (м)*	*Остатки*
	1	2496.475898	-196.475898
	2	2100.96975	294.0302503
	3	3287.488195	-184.4881947
	4	2824.180992	36.81900762
	5	3682.994343	67.00565697
	6	4123.701194	-390.701194
	7	3886.397505	-30.39750503
	8	3682.994343	79.00565697
	9	4214.102599	169.8974006
	10	4270.603478	-50.60347769
	11	4361.004883	325.995117
	12	5039.015423	350.984577
	13	5039.015423	350.984577
	14	914.4513047	10.54869527
	15	5310.219639	-430.219639
	16	3558.692411	-418.6924107
	17	3649.093816	20.90618397
	18	3592.592938	69.4070623
	19	3908.997856	-202.9978564
	20	3682.994343	237.005657
	21	4926.013666	-108.0136664

Рис. 3.3.Исходные данные и результаты регрессивного анализа для однофакторной линейной модели

H (м)	S (м)	S^2(м)
2300	2350	5 522 500.00
2395	2000	4 000 000.00
3103	3050	9 302 500.00
2861	2640	6 969 600.00
3750	3400	11 560 000.00
3733	3790	14 364 100.00
3856	3580	12 816 400.00
3762	3400	11 560 000.00
4384	3870	14 976 900.00
4220	3920	15 366 400.00
4687	4000	16 000 000.00
5390	4600	21 160 000.00
5390	4600	21 160 000.00
925	950	902 500.00
4880	4840	23 425 600.00
3140	3290	10 824 100.00
3670	3370	11 356 900.00
3662	3320	11 022 400.00
3706	3600	12 960 000.00
3920	3400	11 560 000.00
4818	4500	20 250 000.00

ВЫВОД ИТОГОВ

Регрессионная статистика

Множ. R	0.973930399
R-квадрат	0.948540423
Норм. R-квадрат	0.942822692
Станд.ошибка	255.6017479
Наблюдения	21

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	21676530.58	10838265.29	165.8945575	2.53044E-12
Остаток	18	1175980.563	65332.25352		
Итого	20	22852511.14			

	Коэффициенты	*Станд. ошибка*	*t-статист.*	*P-Знач.*	*Нижн. 95%*	*Верхн. 95%*
Y-пересеч.	-24.81677791	450.3226688	-0.05510888	0.956658804	-970.9103303	921.2767745
S (метры)	1.032077534	0.292784797	3.525038	0.002418433	0.416959026	1.647196043
S^2(м)	1.60207E-05	4.6805E-05	0.342286744	0.736099371	-8.2313E-05	0.000114354

ВЫВОД ОСТАТКА

Наблюд.	*Предсказ.H (м)*	*Остатки*
1	2489.039886	-189.0398857
2	2103.421194	291.5788063
3	3272.052503	-169.0525032
4	2811.525963	49.47403707
5	3669.446429	80.55357129
6	4116.880384	-383.8803843
7	3875.348825	-19.34882478
8	3669.446429	92.55357129
9	4209.264088	174.7359122
10	4267.108037	-47.10803725
11	4359.824972	327.1750281
12	5061.738438	328.2615624
13	5061.738438	328.2615624
14	970.1155846	-45.11558458
15	5345.733602	-465.7336022
16	3544.128248	-404.1282478
17	3635.230293	34.76970673
18	3578.267484	83.73251622
19	3898.290952	-192.2909517
20	3669.446429	250.5535713
21	4943.951824	-125.9518237

Рис. 3.4.Исходные данные и результаты регрессивного анализа для однофакторной линейной модели второй степени

	H2	H1	H0							
1	2.11	1.7	0.91		**ВЫВОД ИТОГОВ**					
2	1.87	1.51	0.78							
3	1.91	1.52	0.72		*Регрессионная статистика*					
4	1.81	1.42	0.66		Множественный R	0.996930279				
5	1.7	1.33	0.52		R-квадрат	0.993869981				
6	1.28	1.04	0.31		Нормирован.R-квад.	0.991417974				
7	1.06	0.93	0.24		Стандарт. ошибка	0.032105492				
8	1.6	1.31	0.56		Наблюдения	8				
					Дисперс. анализ					
						df	*SS*	*MS*	*F*	*Значимость F*
					Регрессия	2	0.835596187	0.417798093	405.3290995	2.94208E-06
					Остаток	5	0.005153813	0.001030763		
					Итого	7	0.84075			
						Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	
					Y-пересечение	-0.660119711	0.257292501	-2.565639136	0.050300888	
					H1	2.098090815	0.366878189	5.718766824	0.002286081	
					H0	-0.841382869	0.409068547	-2.056826113	0.094815387	
					ВЫВОД ОСТАТКА					
					Наблюдение	*Предсказанное H2*	*Остатки*			
					1	2.140976264	-0.030976264			
					2	1.851718782	0.018281218			
					3	1.923182663	-0.013182663			
					4	1.763856553	0.046143447			
					5	1.692821981	0.007178019			
					6	1.261066047	0.018933953			
					7	1.089172859	-0.029172859			
					8	1.61720485	-0.01720485			

Рис 3.5. Исходные данные и результаты регрессионного анализа для многофакторной линейной модели

4. ПРИЛОЖЕНИЯ

4.1. Практическое применение MS EXCEL

4.1.1. Расчет χ^2 – критерия

1. Запуск программы **EXCEL**
2. Ввод данных из таблиц 1.1 и 1.2 (см. рис. 4.1)
3. Щелчок по кнопке "**вставка функций**" на стандартной панели инструментов. Открывается диалоговое окно (рис.4.1). Выбираем в Мастере функций из категорий – "**Статистические**", из функций – "**ХИ2ТЕСТ**".

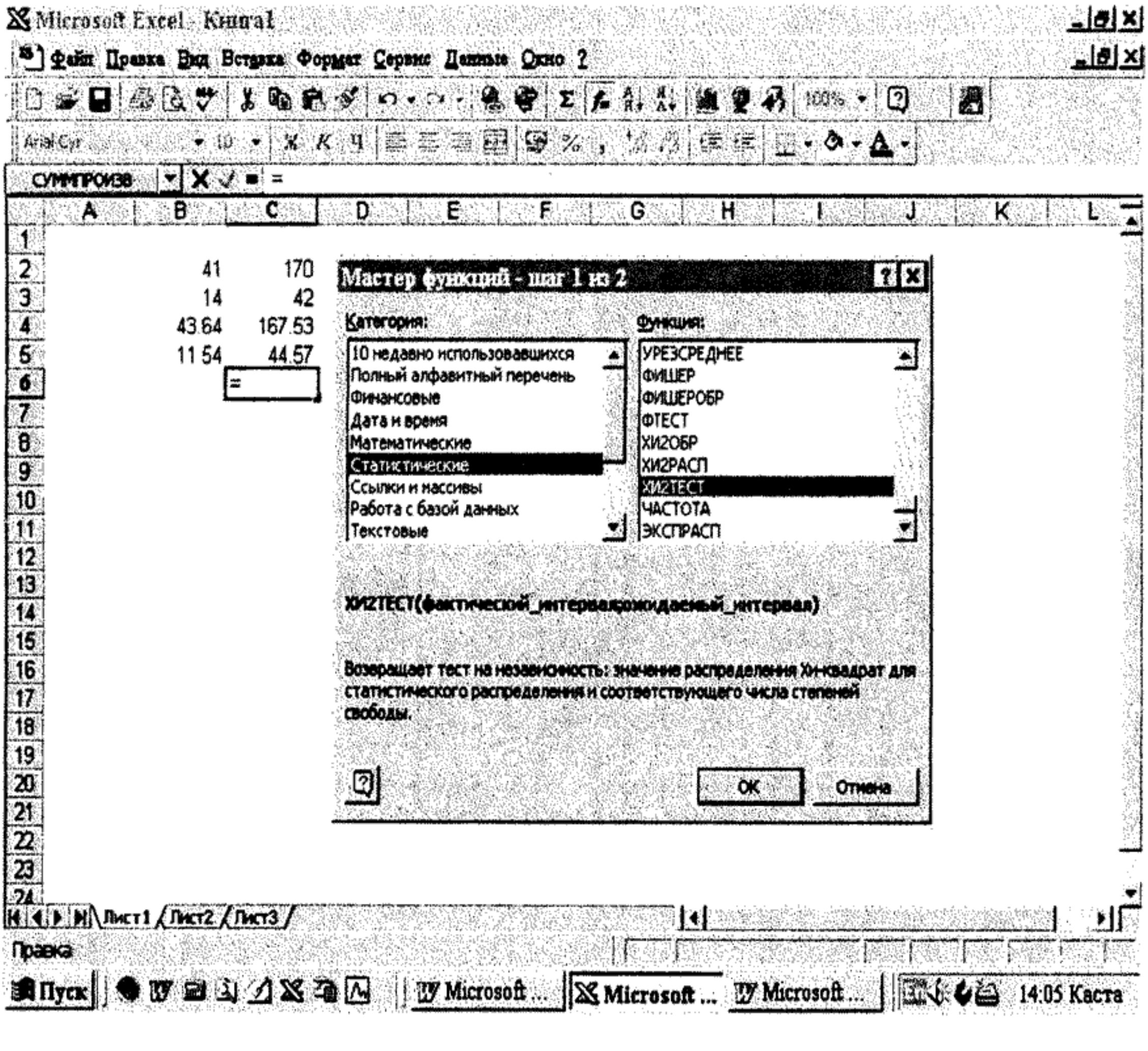


Рис.4.1

4. В диалоговое окно "**ХИ2ТЕСТ**" вносим данные по фактическим и ожидаемым числам. (Ожидаемые числа должны быть предварительно рассчитаны). Получаем значение вероятности равное 0.351 (рис.4.2).

5. Выбираем из категорий – "**Статистические**", из функций – "**ХИ2ОБР**" (рис. 4.3).

6. В диалоговое окно "**ХИ2ОБР**" (рис. 4.4) вводим данные о рассчитанной вероятности (в примере – ячейка С6) и степень свободы (в примере степень свободы равна 1).

Получаем значение χ^2-критерия (в примере результат в ячейке D6, χ^2=0.86).

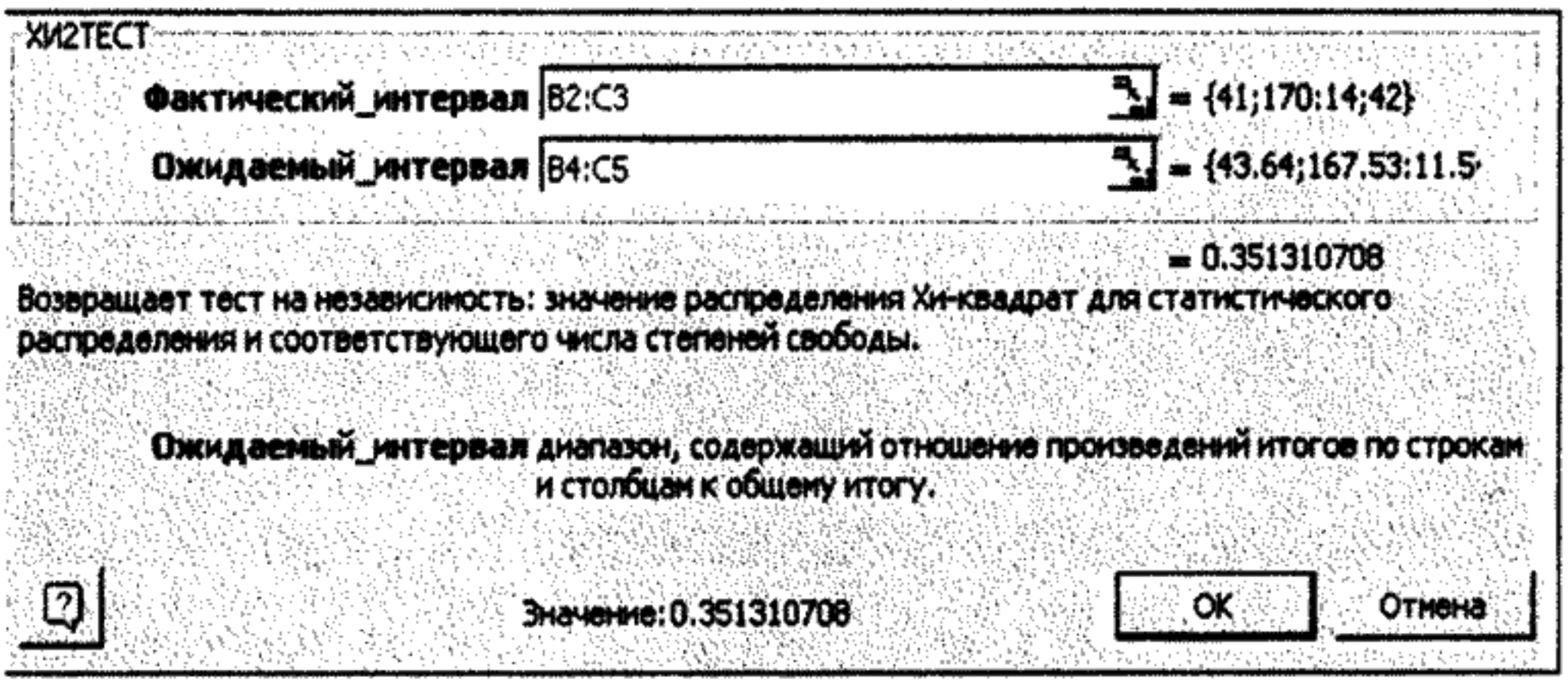


Рис. 4.2

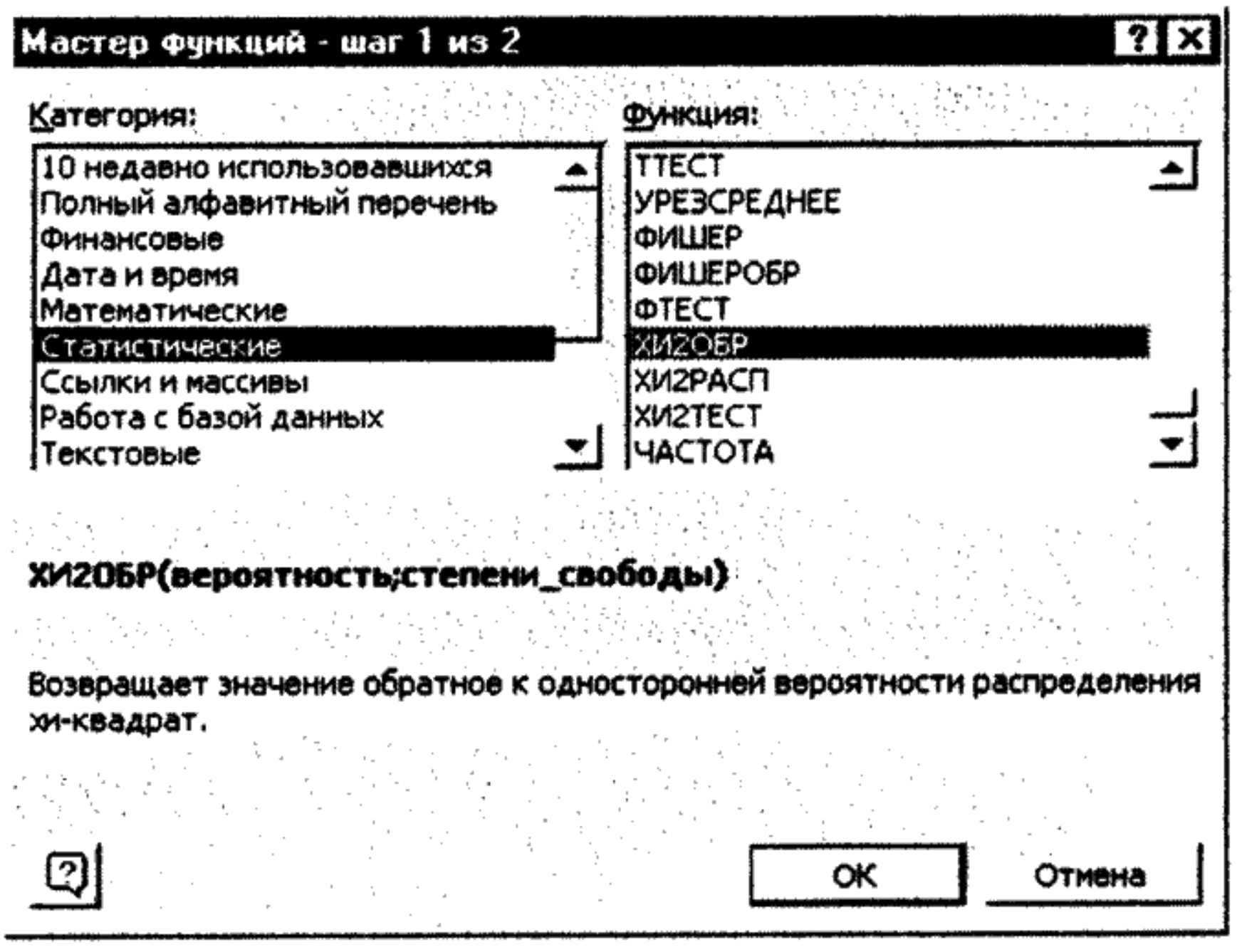


Рис. 4.3

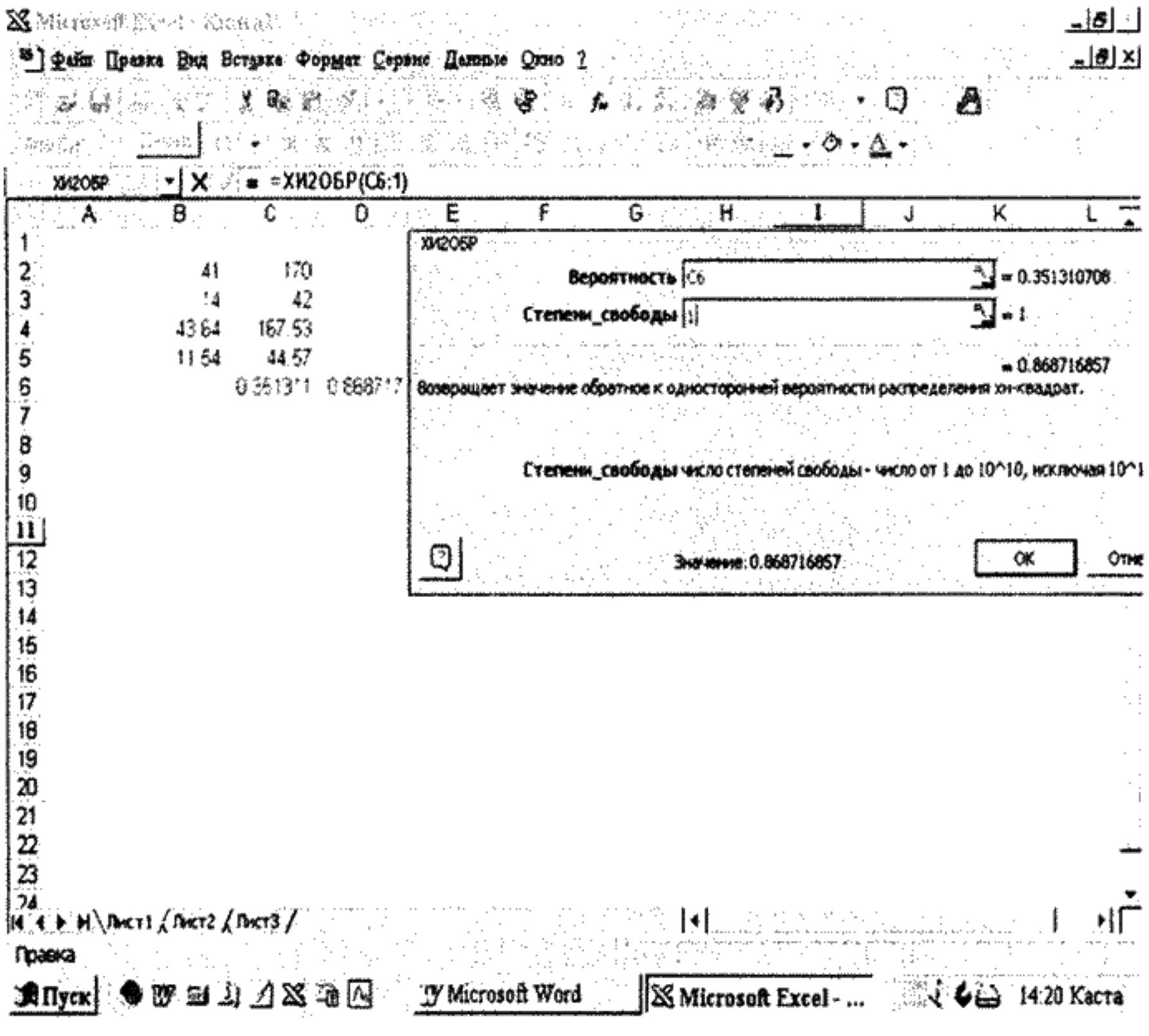


Рис 4.4

К сожалению, в электронной таблице Excel нет процедуры определения ожидаемых чисел (ожидаемого интервала на рис. 4.2) , поэтому использование возможностей Excel для расчета χ^2 – критерия несколько трудоемко, но возможно. Пример такого расчета подробно показан в работе /17/. Другой вариант – с помощью Excel определять ожидаемые числа по схеме, предложенной в главе 1 и далее по схеме как показано выше. Возможно также использование других статистических пакетов, например, пакета **БИОСТАТ** (биостатистика) /3/, оглавление которого приведено в конце параграфа 4.2 ПРИЛОЖЕНИЯ. Наиболее полный набор статистических процедур представлен в системе **STATISTICA** /19/, успешно используется пакет **STADIA** /7/ и т.д.

4.1.2. Расчет коэффициентов корреляции

Для иллюстрации расчетов коэффициентов корреляции Пирсона и Спирмена воспользуемся табл.1.9. Введем данные этой таблицы в Excel и подсчитаем значения коэффициентов (рис.4.5).

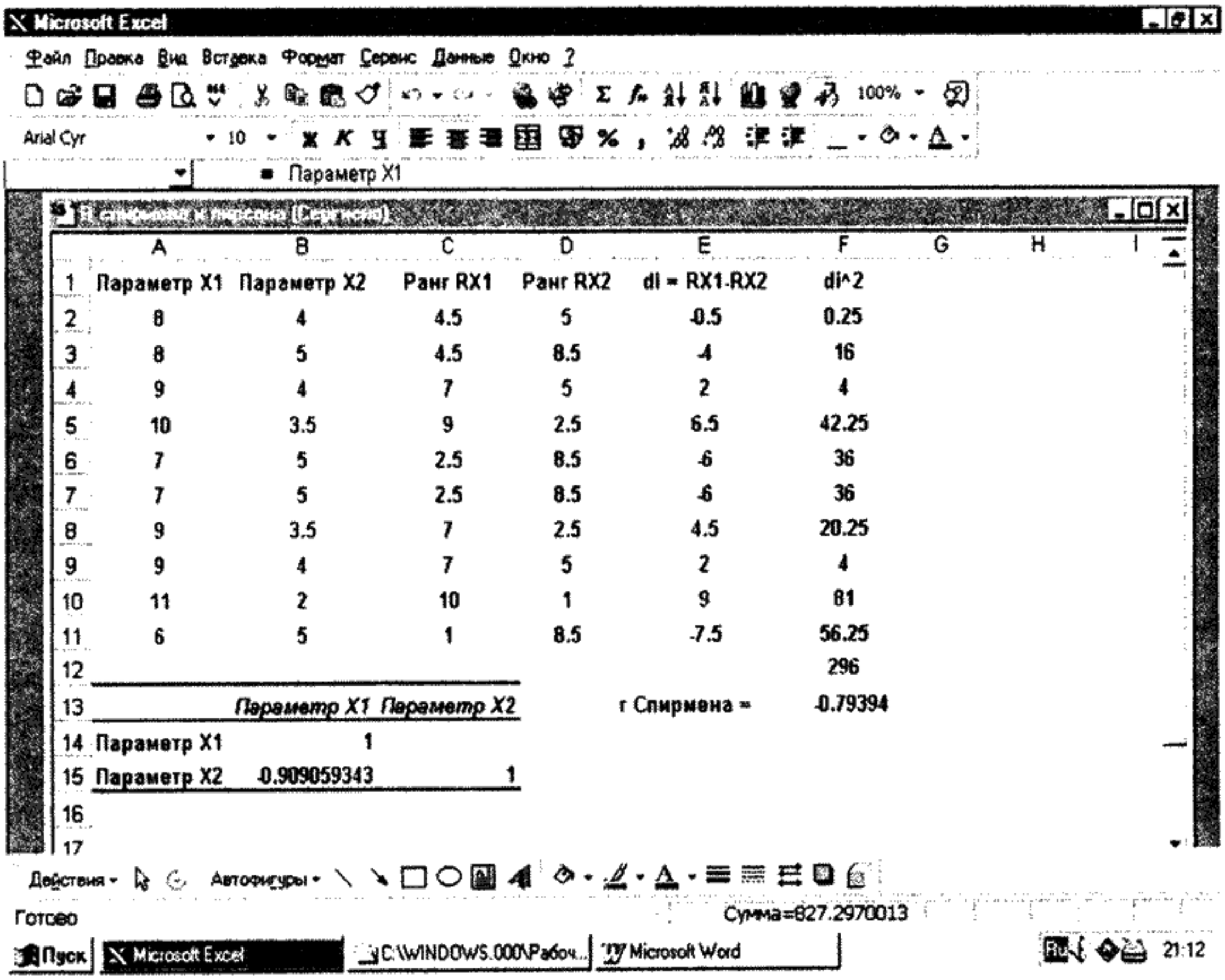
1. Коэффициент корреляции Пирсона рассчитывается по схеме: **пункт меню Сервис => Анализ данных => Инструменты анализа => Корреляция**. В открывшемся окне (рис 4.6) в поле **Входной интервал** вводим диапазон данных (если есть названия признаков, то в поле **Метки** устанавливаем флажок).

Выходной интервал – адрес ячейки области вывода. В нашем примере коэффициент корреляции Пирсона между признаками X1 и X2 в ячейке B15 (рис.4.5). (Коэффициент корреляции Пирсона можно рассчитать для двух признаков с помощью **Мастера функций**).

2. Так как коэффициент корреляции Спирмена не предусмотрен для расчетов в **Анализе данных** и при помощи **Мастера функций**, его не сложно определить самостоятельно с использованием простейших функций Excel.

Для этого, зная ранги параметров X1 и X2 (табл.1.9 или рис. 4.5), легко рассчитать *di* для формулы Спирмена $r = 1 - \frac{6\sum^{n} di^2}{n(n^2-1)}$ средствами Excel /15/.Для примера (рис.4.5) формула будет выглядеть следующим образом:

= 1 – (6 * F12) / (10 * (10 ^ 2-1)).



	A	B	C	D	E	F	G	H	I
1	Параметр X1	Параметр X2	Ранг RX1	Ранг RX2	dI = RX1-RX2	di^2			
2	8	4	4.5	5	-0.5	0.25			
3	8	5	4.5	8.5	-4	16			
4	9	4	7	5	2	4			
5	10	3.5	9	2.5	6.5	42.25			
6	7	5	2.5	8.5	-6	36			
7	7	5	2.5	8.5	-6	36			
8	9	3.5	7	2.5	4.5	20.25			
9	9	4	7	5	2	4			
10	11	2	10	1	9	81			
11	6	5	1	8.5	-7.5	56.25			
12						296			
13			Параметр X1	Параметр X2		r Спирмена =	-0.79394		
14	Параметр X1		1						
15	Параметр X2		-0.909059343	1					
16									
17									

Рис.4.5

Здесь в ячейке F12 – сумма в квадрате разностей между рангами сопряженных признаков. В ячейке F13 значение коэффициента Спирмена (рис. 4.5).

Процесс ранжирования в Excel проводится с помощью функции **РАНГ( )**, которая работает с ограничениями – не проводит осреднение рангов и поэтому нужно либо дополнять Excel модулем ранжирования с осреднением /17/, либо использовать статистические пакеты /3,7,9,18,19/.

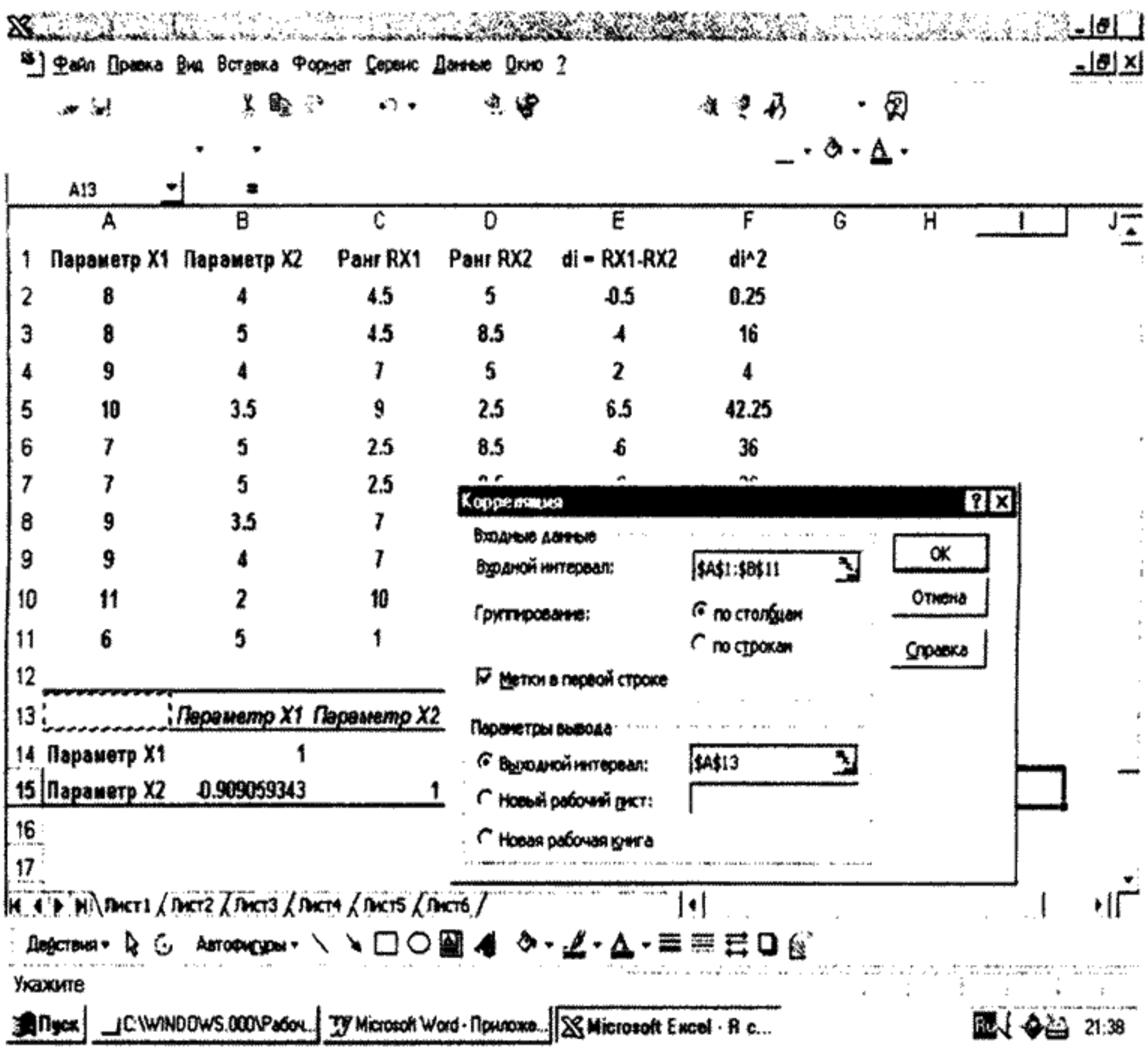


Рис. 4.6

4.1.3. Использование EXCEL при дисперсионном анализе

1. Запуск программы EXCEL

2. Вводим данные по примеру (гл. 2) влияния денежного вознаграждения на результаты труда (рис. 4.7).

3. Для проведения анализа открываем пункт меню **Сервис**, выбираем **Анализ данных** и в открывшемся окне **Инструменты анализа** указываем пункт **Однофакторный дисперсионный анализ** (рис. 4.8).

Microsoft Excel - однофакторный дисперсионный (Сарапченко)

Файл Правка Вид Вставка Формат Сервис Данные Окно ?

H31

	A	B	C	D	E	F	G	H	I	J
1	ВЛИЯНИЕ ВОЗНАГРАЖДЕНИЯ (от MIN К MAX)									
2	1	2	3	4	5	6				
3	10	8	12	12	16	19				
4	11	10	17	15	22	18				
5	9	16	14	16	18	24				
6	13	13	9	16	20	23				
7	7	12	13	13		27				
8	8		16	19		25				
9	9					24				
10						22				
11	Однофакторный дисперсионный анализ									
12	ИТОГИ									
13	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*					
14	1	7	67	9.57143	3.952381					
15	2	5	59	11.8	9.2					
16	3	6	81	13.5	8.3					
17	4	6	91	15.1667	6.166667					
18	5	4	76	19	6.666667					
19	6	8	182	22.75	9.071429					
20	Дисперсионный анализ									
21	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F крит*			
22	Между группами	808.5413	5	161.708	**22.42339**	2.6063E-09	**2.53355**			
23	Внутри групп	216.3476	30	7.21159						
24										
25	Итого	1024.889	35							
26										
27										

Лист1 / Лист2 / Лист3 /

Готово NUM

Пуск | Microsoft Word | C:\WINDOWS.000\Рабоч... | Microsoft Excel - одн... | 14:02

Рис. 4.7

В диалоговом окне **Однофакторный дисперсионный анализ** задается область исходных данных **(A2 :F10), Альфа** (“по умолчанию” – 0.05, меняется в зависимости от требований задачи), метка – если есть названия признаков, то в поле **Метки** устанавливаем флажок), выходной интервал – в примере **A11** (рис 4.7).

Двухфакторный дисперсионный анализ (с повторениями и без) рассчитывается аналогично однофакторному. Листинги с итогами расчета и интерпретация результатов показаны в главе 2.

4.1.4. Использование EXCEL при регрессионном анализе

1. Запуск программы EXCEL

2. Вводим данные для построения многофакторной полиномиальной модели (рис. 4.9).

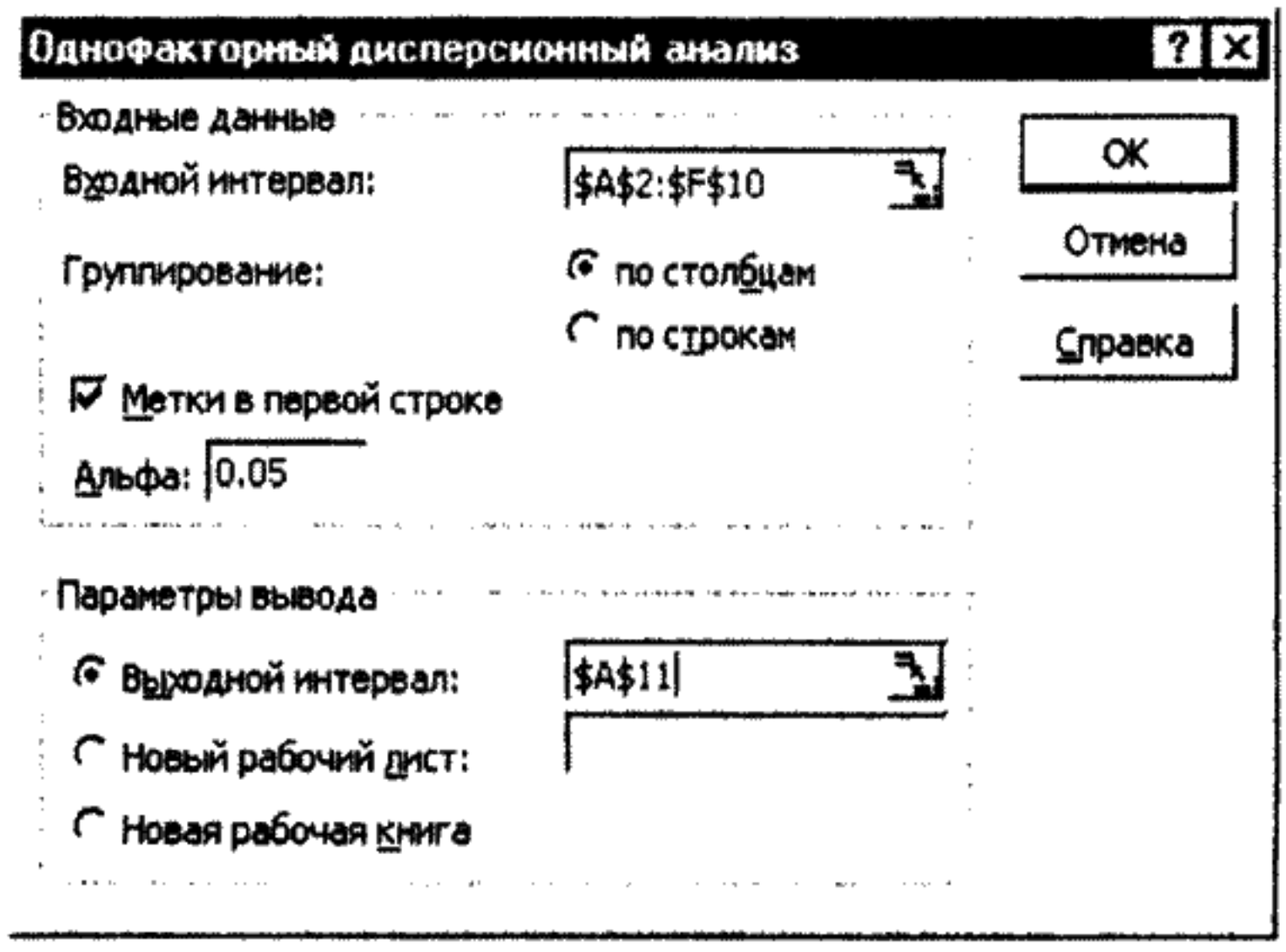


Рис. 4.8

3. Для проведения анализа открываем пункт меню **Сервис**, выбираем **Анализ данных** и в открывшемся окне **Инструменты анализа** указываем пункт **Регрессия**.

В диалоговом окне **Регрессия** задается область исходных данных (рис. 4.10):

входной интервал Y (в примере H2)

входной интервал X (в примере H0, H1 и т.д.)

уровень надежности (по умолчанию 0.05, меняется в зависимости от требований задачи),

метка – если есть названия признаков, то в поле Метки устанавливаем флажок,

выходной интервал – в примере F1 (или любой другой).

Составляющие области вывода (листинг результатов) /8/ .

По окончанию расчета на рабочий лист выводится три группы результатов (рис.4.9). Пример их интерпретации показан в разделе 3.4.

Первая группа – *регрессионная статистика*, включает в свой состав:

- *Множественный R* – коэффициент множественной корреляции;
- *R – квадрат* – множественный коэффициент детерминации;
- *Нормированный R-квадрат* – корректированный коэффициент детерминации;
- *Стандартная ошибка* – стандартная ошибка регрессии;
- *Наблюдения* – количество наблюдений.

Вторая группа результатов – *дисперсионный анализ*, здесь использован ряд общепринятых сокращений.

Вот их расшифровки:

- *df* - степени свободы;
- *SS* – сумма квадратов отклонений;
- *MS* – средний квадрат отклонения;
- *F* – отношение дисперсий;
- *Значимость F* – критическое значение квантиля распределения Фишера, на котором отвергается нулевая гипотеза отсутствия влияния фактора.

Построчно в таблице выводятся показатели, характеризующие изменчивости: присущую модели и случайную.

- *Регрессия* – здесь выводятся характеристики, связанные с закономерной изменчивостью: сумма квадратов отклонений между группами *SS*, соответствующее число степеней свободы *df*, на основании количества которых определяется *Значимость F* и частное от этих величин – средний квадрат отклонений *MS*. Так же в данной строке выведен собственно результат анализа: *F - отношение.*
- *Остаток* – тут представлены показатели, характеризующие действие случайных факторов – те же самые, что и для предыдущей строки, только, разумеется, без окончательных результатов анализа.
- *Итого* – представлены суммы квадратов отклонений от среднего *SS* и количество степеней свободы *SS* значениям регрессии и остатка. В данном случае *SS* – характеристика полной изменчивости.

Следующая группа результатов включает в свой состав значения коэффициентов уравнения регрессии, а также статистики, на основании которых проверяется значимость влияния фактора для каждого коэффициента, включенного в модель.

- *Коэффициенты* – значение коэффициентов;
- *Стандартная ошибка* — стандартная ошибка коэффициентов;
- *t – статистика* – значение статистики критерия, на основании которого определяется уровень значимости отклонения гипотезы равенства коэффициенов нулю *(P-значение);*
- *P-значение* – уровень значимости, на котором отвергается гипотеза равенства коэффициентов нулю;
- *Нижние 95%* – нижняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности;
- *Верхние 95%* – верхняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности;
- *Нижние ...* % – нижняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности (значение задается при определении параметров анализа);

Глубины залегания стратиграфических границ

	H2	H1	H0	H1^2	H0^2	H1*H0
1	2.11	1.7	0.91	2.89	0.83	1.55
2	1.87	1.51	0.78	2.28	0.61	1.18
3	1.91	1.52	0.72	2.31	0.52	1.09
4	1.81	1.42	0.66	2.02	0.44	0.94
5	1.7	1.33	0.52	1.77	0.27	0.69
6	1.28	1.04	0.31	1.08	0.10	0.32
7	1.06	0.93	0.24	0.86	0.06	0.22
8	1.6	1.31	0.56	1.72	0.31	0.73

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	5	0.838355848	0.16767117	140.0672824	0.007103898
Остаток	2	0.002394152	0.001197076		
Итого	7	0.84075			

	Коэф - ты	*Стандарт. Ош.*	*t-статис.*	*P-Значение*
Y-пересечение	-4.39007517	6.841624209	-0.641671486	0.586812389
H1	10.95084395	18.36211466	0.596382506	0.611431602
H0	-8.192235785	19.06666657	-0.429662718	0.709302894
H1^2	-5.088170859	12.28663427	-0.414122431	0.718972319
H0^2	-2.941459535	13.54150524	-0.217218063	0.848184005
H1*H0	8.070631493	25.62398258	0.314963978	0.782612915

ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	0.998575166
R-квадрат	0.997152362
Норм. R-квадрат	0.990033266
Стандарт. ошибка	0.034598785
Наблюдения	8

Вывод остатка

Наблюдение	*Предсказан. H2*	*Остатки*
1	2.116055483	-0.006055483
2	1.870222703	-0.000222703
3	1.908734404	0.001265596
4	1.775955967	0.034044033
5	1.70039733	-0.00039733
6	1.275141179	0.004858821
7	1.059251022	0.000748978
8	1.634241911	-0.034241911

Рис.4.9.**Исходные данные и результаты регрессионного анализа для многофакторной полиномиальной модели**

– *Верхние* % – верхняя граница доверительного интервала, в котором находится значение коэффициентов генеральной совокупности (значение задается при определении параметров анализа).

При необходимости есть возможность вывести таблицу стандартных и простых остатков, где для каждого значения ряда выводится предсказанное значение, с которым сопоставляется остаток, представляющий собой разность между прогнозным и реальным значением ряда. Кроме вывода табличной информации, есть возможность просмотреть графики остатков, а также ряд других полезных графиков /8/.

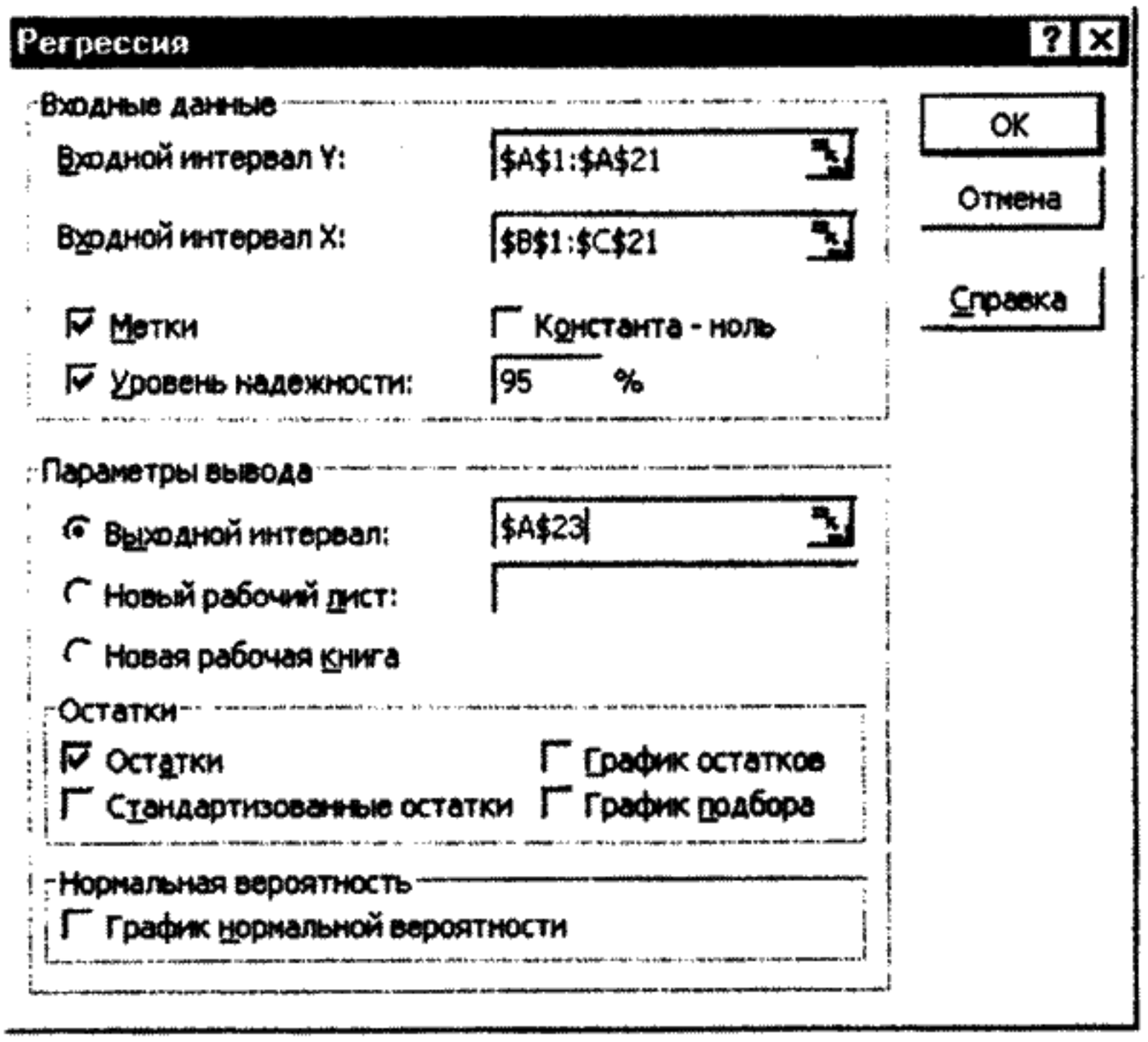


Рис. 4.10

4.2. Использование программ БИОСТАТ и STATISTICA при расчете непараметрических критериев

4.2.1. Статистический пакет БИОСТАТ

В Excel не предусмотрены процедуры расчета непараметрических критериев, поэтому мы рассчитаем их с помощью программы БИОСТАТ.

В этой программе используется Т – критерий Манна – Уитни вместо рассмотренного в главе 1 U – критерия ($U=T-n_m(n_m-1)/2$).

Критические значения Т – критерия Манна – Уитни приведены в табл.4.1 Столбец критических значений содержит пары чисел. Различия статистически значимы, если Т не больше первого из них или не меньше второго. Например,

когда в одной группе 3 человека, а в другой 6, различия статистически значимы, если T < *7* или T > 23.

Порядок вычисления T – критерий Манна – Уитни следующий /3/:

– Данные обеих групп объединяют и упорядочивают по возрастанию. Ранг 1 присваивают наименьшему из всех значений, ранг 2 — следующему и так далее. Наибольший ранг присваивают самому большому среди значений в обеих группах. Если значения совпадают, им присваивают один и тот же средний ранг (например, если два значения поделили 3-е и 4-е места, обоим присваивают ранг 3,5).

– Для меньшей группы вычисляют T – сумму рангов ее членов. Если численность групп одинакова, T можно вычислить для любой из них.

– Полученное значение T сравнивают с критическими значениями. Если T меньше или равно первому из них либо больше или равно второму, то нулевая гипотеза отвергается (различия статистически значимы).

Что делать, если нужной численности групп в таблице не оказалось?

В таком случае лучше воспользоваться тем, что при численности групп, большей 8, распределение T приближается к нормальному со средним

$$\mu_T = \frac{n_м (n_м + n_б + 1)}{2}$$

и стандартным отклонением

$$\sigma_T = \sqrt{\frac{n_м n_б (n_м + n_б + 1)}{12}},$$

где n_m и $n_б$ – объемы меньшей и большей выборок. В таком случае величина

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

имеет стандартное нормальное распределение. Это позволяет сравнить z_T с критическими значениями нормального распределения (последняя строка табл. 3.2 /1/). Более точный результат обеспечивает поправка Йейтса:

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T}.$$

По вышеизложенной схеме по программе БИОСТАТ был рассчитан пример (табл.1.8) по T – критерию Манна – Уитни.

Результат: T = 141; $z_T = 2.7$; p = 0.007.

Критическое значение t – критерия при $\alpha = 0.05$ равно 2.02.

Вывод: различия между выборками статистически значимы ($z_T > t_{крит}$).

Этот вывод не противоречит выводу, полученному по U – критерию Манна – Уитни в главе 1.

Далее, по причине отсутствия расчетного аппарата критерия Уилкоксона (для попарно связанных выборок, рассмотренного в главе 1) в Excel, покажем алгоритм определения этого критерия по С.Гланцу /3/ и вычисления его с помощью программы БИОСТАТ.

Таблица 4.1.

Критические значения Т – критерия Манна – Уитни (двусторонний вариант) /3/

Численность группы		Приблизительный уровень значимости α					
		0,05			0,01		
меньшей	большей	Критические значения		Точное значение α	Критические значения		Точное значение α
3	4	6	18	0,057			
	5	6	21	0,036			
	5	7	20	0,071			
	6	7	23	0,048	6	24	0,024
	7	7	26	0,033	6	27	0,017
	7	8	25	0,067			
	8	8	28	0,042	6	30	0,012
4	4	11	25	0,057	10	26	0,026
	5	11	29	0,032	10	30	0,016
	5	12	28	0,063			
	6	12	32	0,038	10	34	0,010
	7	13	35	0,042	10	38	0,012
	8	14	38	0,048	11	41	0,008
	8				12	40	0,016
5	5	17	38	0,032	15	40	0,008
	5	18	37	0,056	16	39	0,016
	6	19	41	0,052	16	44	0,010
	7	20	45	0,048	17	48	0,010
	8	21	49	0,045	18	52	0,011
6	6	26	52	0,041	23	55	0,009
	6				24	54	0,015
	7	28	56	0,051	24	60	0,008
	7				25	59	0,014
	8	29	61	0,043	25	65	0,008
	8	30	60	0,059	26	64	0,013
7	7	37	68	0,053	33	72	0,011
	8	39	73	0,054	34	78	0,009
8	8	49	87	0,050	44	92	0,010

Последовательность шагов, позволяющая по наблюдениям, выполненным "до и после", вычислить критерий:

- Вычислите величины изменений наблюдаемого признака. Отбросьте пары наблюдений, которым соответствует нулевое изменение.
- Упорядочите изменения по возрастанию их абсолютной величины и присвойте соответствующие ранги. Рангами одинаковых величин назначьте средние тех мест, которые они делят в упорядоченном ряду.
- Присвойте каждому рангу знак в соответствии с направлением изменения: если значение увеличилось – «+», если уменьшилось – «-».
- Вычислите сумму знаковых рангов *W* (Существует вариант критерия Уилкоксона, в котором суммируют только положительные или только отрицательные знаковые ранги (гл. 1). На выводе это никак не сказывается, однако значение *W*, естественно, получается другим. Поэтому важно знать, на какой вариант критерия рассчитана имеющаяся в вашем распоряжении таблица критических значений).
- Сравните полученную величину *W* с критическим значением (табл.4.2). Если она больше критического значения, изменение показателя статистически значимо.

Таблица 4.2

Критические значения *W* (двусторонний вариант) /3/

n	*W*	*P*	*n*	*W*	*P*
5	15	0,062	13	65	0,022
6	21	0,032		57	0,048
	19	0,062	14	73	0,020
7	28	0,016		63	0,050
	24	0,046	15	80	0,022
8	32	0,024		70	0,048
	28	0,054	16	88	0,022
9	39	0,020		76	0,050
	33	0,054	17	97	0,020
10	45	0,020		83	0,050
	39	0,048	18	105	0,020
11	52	0,018		91	0,048
	44	0,054	19	114	0,020
12	58	0,020		98	0,050
	50	0,052	20	124	0,020
				106	0,048

Просчитаем пример из главы 1 (1.4.2) по программе **БИОСТАТ**.
Результат: W = 21,0 ; p>0,06; n=10.
Критическое значение $W_{крит.}$ = 39; p=0,048; (табл.4.2).

Вывод: различия между выборками статистически не значимы.

Этот вывод не противоречит выводу, полученному в главе 1 – нулевая гипотеза остается в силе (нулевая гипотеза об отсутствии различия).

Биостатистика

Процедуры, которые возможно просчитать в статистическом прикладном пакете «Биостатистика»:

Описательная статистика
Дисперсионный анализ
Ранговая корреляция по Спирмену
Критерий t (Стьюдента)
Критерий Уилкоксона
Критерий Данна
Критерий Фридмана
Линейная регрессия и корреляция
Стандартная ошибка доли
Точный критерий Фишера
Критерий хи-квадрат
Критерий Мак-Нимара
Критерий Манна-Уитни
Критерий Крускала-Уоллиса
Критерий Ньюмена-Кейлса
Критерий z
Критерий Даннета

4.2.2. Статистический пакет STATISTICA

Для расчета критерия Манна – Уитни возможно использование статистического пакета STATISTICA /19/. (Пример из табл.1.8 рассчитан с применением пакета STATISTICA 6.0). Здесь необходимо:

1. Запуск программы STATISTICA 6.0
2. В меню программы STATISTICA 6.0 выбираем пункт меню **Файл => Новый.** В вызванном диалоговым окне “Создание нового документа” определяем *количество* (переменные, факторы и.т.п) и *число регистров* (анализов, вариант и.т.д). В нашем примере (табл.1.8) *количество* будет равно двум: 1 – будет иметь название “КОД”, 2 – упорядоченные значения столбца 2 (табл.1.8) и упорядоченные значения столбца 3 (табл.1.8). Название этой переменной “Выборки 1 – 2”. Переменная КОД имеет два значения: 1- для значений столбца 2, 2 – для столбца 3 (табл.1.8). Количество вариант равняется в сумме 20. Результат ввода данных на рис.4.11
3. В меню программы STATISTICA 6.0 выбираем пункт **Статистика => Не параметрический => Compring two independet samples (groups).**
4. OK.

   Открывается диалоговое окно “Сравнение дух групп: М – У.

   Рис. 4.12

   Выбираем кнопку **Variables.** Щелчок.

5. Далее – как показано на рис. 4.12 .
6. OK.
7. Производим щелчок на кнопке **U тест Манна – Уитни**. Результат на рис. 4.13.

Результаты, полученные с помощью статистического пакета STATISTICA, аналогичны расчетам по примеру из табл.1.8, гл. 1.

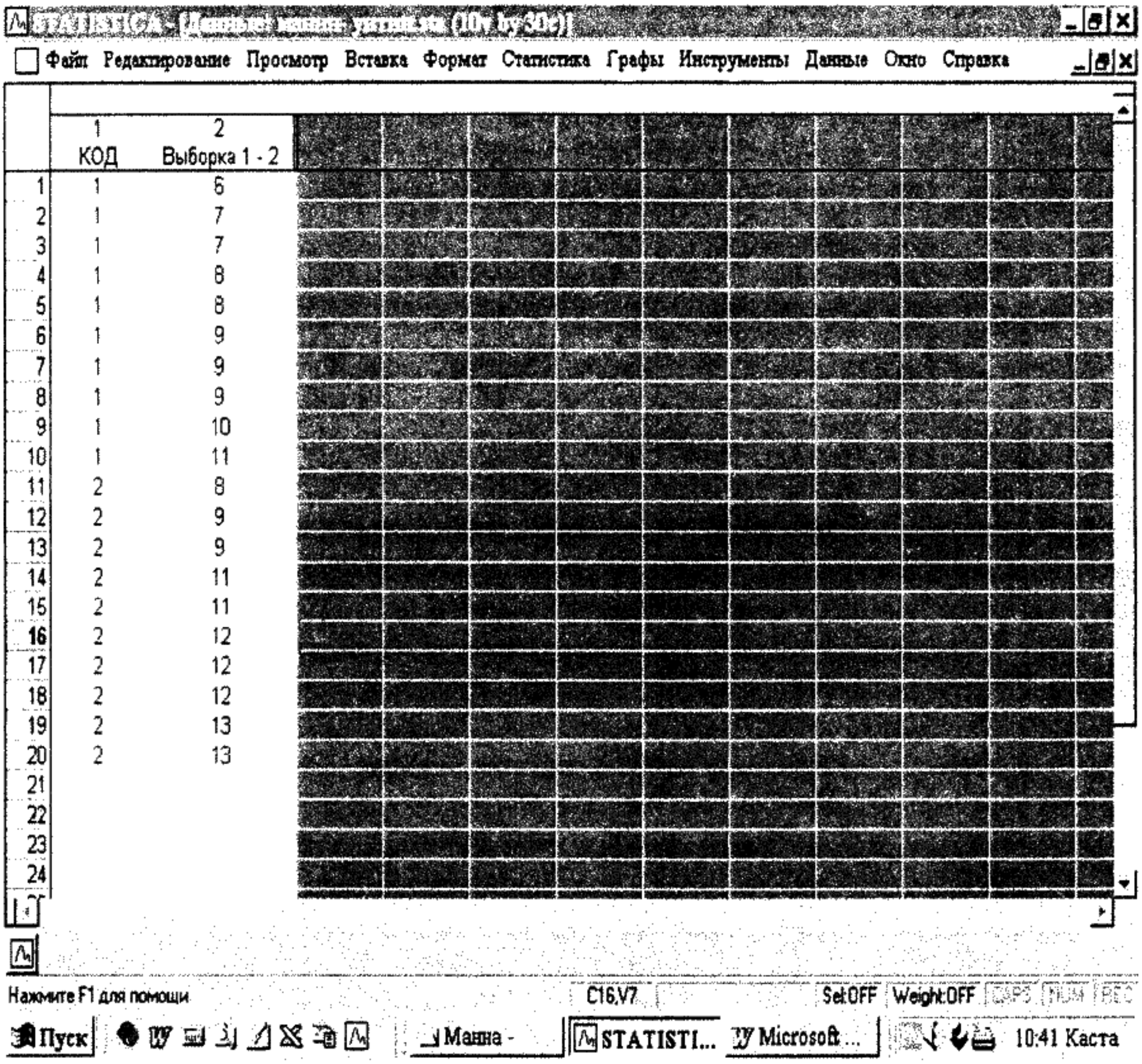


Рис 4.11

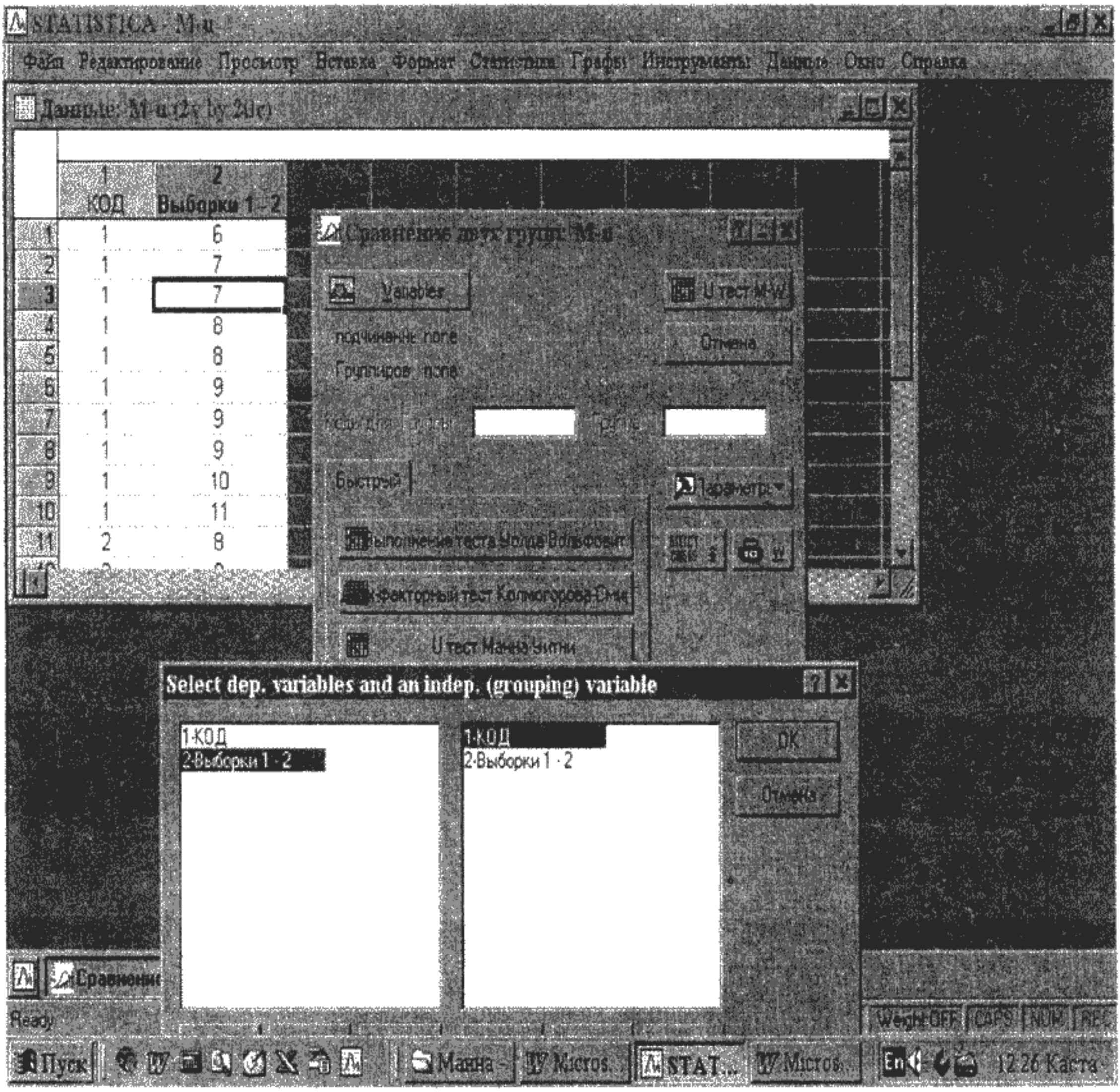

STATISTICA - M-u
Файл Редактирование Просмотр Вставка Формат Статистика Графы Инструменты Данные Окно Справка
Данные: M-u (2v by 20c)
1 КОД
2 Выборки 1 - 2
1 6
1 7
1 7
1 8
1 8
1 9
1 9
1 9
1 10
1 11
2 8
Сравнение двух групп: M-u
Variables
U тест M-W
подчинен. поле
Отмена
Группиров. поле
Быстрый
Параметры
Выполнение теста Уилка Вольфовит
Факторный тест Колмогорова-Сми
U тест Манна-Уитни
Select dep. variables and an indep. (grouping) variable
1-КОД
2-Выборки 1 - 2
1-КОД
2-Выборки 1 - 2
OK
Отмена
Сравнение
Ready
Weight:OFF
CAPS
NUM
REC
Пуск
Макна...
Micros...
STAT...
Micros...
12:26 Каста

Рис. 4.12

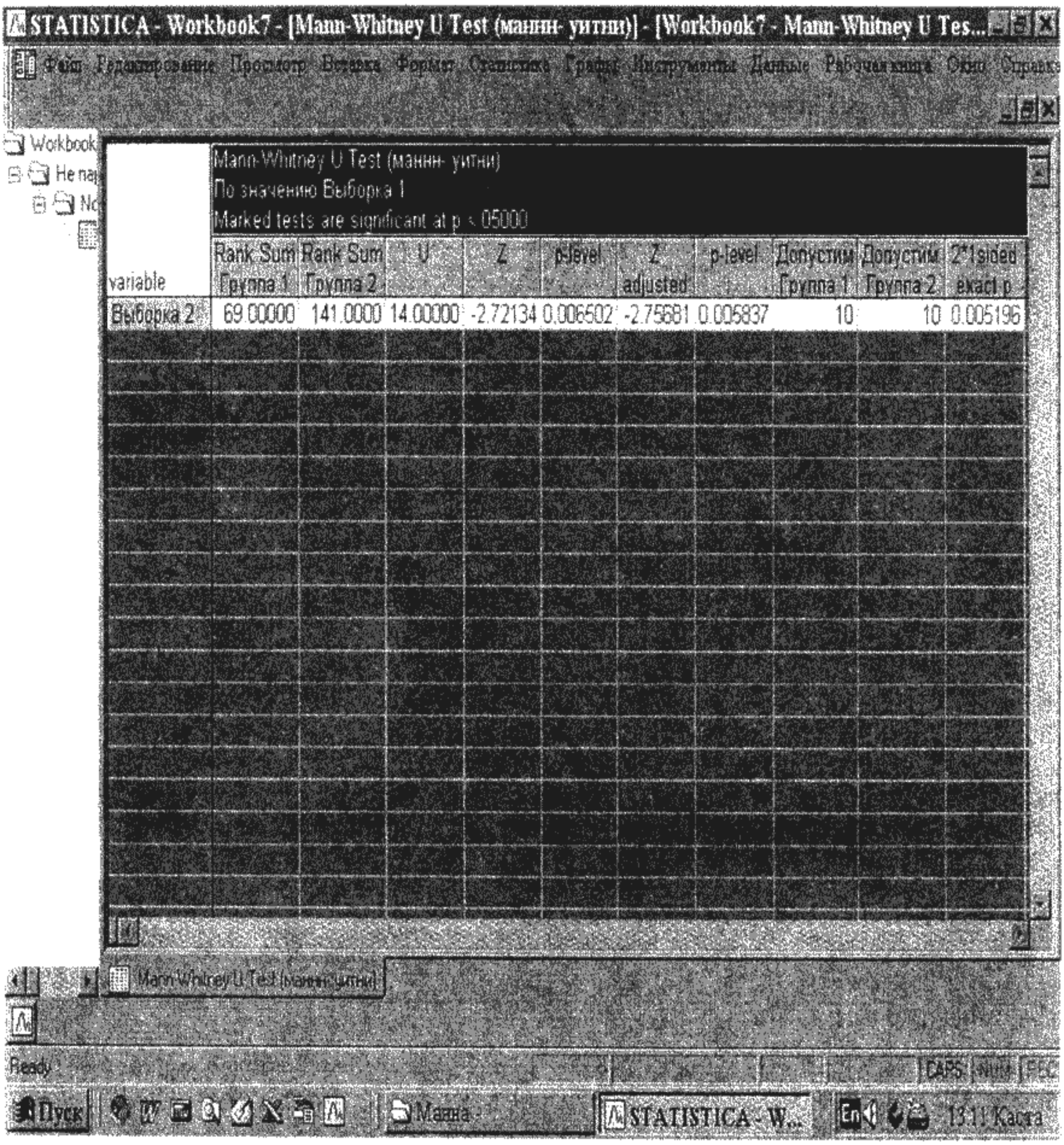

STATISTICA - Workbook7 - [Mann-Whitney U Test (маннн- уитни)] - [Workbook7 - Mann-Whitney U Tes...
Файл Редактирование Просмотр Вставка Формат Статистика График Инструменты Данные Рабочая книга Окно Справка
Workbook
Mann-Whitney U Test (маннн- уитни)
По значению Выборка 1
Marked tests are significant at p < 05000
variable
Rank Sum Группа 1
Rank Sum Группа 2
U
Z
p-level
Z adjusted
p-level
Допустим Группа 1
Допустим Группа 2
2*1sided exact p
Выборка 2
69.00000
141.0000
14.00000
-2.72134
0.006502
-2.75681
0.005837
10
10
0.005196
Ready
CAPS
NUM
SCRL
Пуск
Маина
STATISTICA - W...
13:11 Каста

Рис. 4.13

4.3. Задачи и упражнения

Теоретические и практические вопросы решения задач и упражнений, предложенных ниже, можно найти в /1,3,4,5,6,8,13,15,19,20/ и в данном учебном пособии.

1. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующей выборки: 0; 0; 0; 1; 1; 1; 1; 1; 1;1; 1; 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 4; 4; 5; 5; 5; 5; 6; 7; 9; 10; 11. Можно ли считать, что выборка извлечена из совокупности с нормальным распределением? Обоснуйте ответ.

2. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующих данных: 289; 203; 359; 243;232; 210; 251; 246; 224; 239; 220; 211. Можно ли считать, что выборка извлечена из совокупности с нормальным распределением? Обоснуйте ответ.

3. Найдите среднее, стандартное отклонение, медиану, 25-й и 75-й процентили для следующих данных: 1,2; 1,4; 1,6; 1,7; 1,7;1,8; 2,2; 2,3; 2,4; 6,4; 19,0; 23,6. Можно ли считать, что это — выборка из совокупности с нормальным распределением? Обоснуйте ответ.

4. Постройте графики для двух наборов данных. Найдите для каждого линию регрессии и коэффициент корреляции. Посмотрите результаты.

X	Y
15	19
15	29
20	25
20	35
25	31
25	41
30	37
30	47
60	40

X	Y
20	21
20	31
30	18
30	28
40	15
40	25
40	75
40	85
50	65
50	75
60	55
60	65

5. Курение считают основным фактором, предрасполагающим к хроническим заболеваниям легких. Что касается пассивного курения, оно таким фактором обычно не считается. Ученые усомнились в безвредности пассивного курения и исследовали проходимость дыхательных путей у некурящих, пассивных и активных курильщиков Для характеристики состояния дыхательных путей взяли один из показателей функции внешнего дыхания – максимальную объемную скорость середины выдоха, которую измеряли во время профилактического осмотра сотрудников Калифорнийского университета в Сан-Диего. Уменьшение этого показателя – признак нарушения проходимости дыхательных путей. Данные обследования представлены в таблице.

Максимальная объемная группа	Число обследованных	Максимальная объемная скорость средины выдоха, л/с	
		Среднее	Стандартное откло-
Некурящие			
Работающие в помещении, где не курят	200	3,17	0,74
Работающие в накуренном помещении	200	2,72	0,71
Курящие			
Выкуривающие небольшое число сигарет	200	2,63	0,73
Выкуривающие среднее число сигарет	200	2,29	0,70
Выкуривающие большое число сигарет	200	2,12	0,72

Можно ли считать максимальную объемную скорость середины выдоха одинаковой во всех группах? /3/.

6. Ниже таблица содержит данные о группировке 680 человек по двум признакам:

1- цвет глаз (по строкам): синий, серый, коричневый;

2- цвет волос (по столбцам): светлый, русый, черный, коричневый.

Необходимо оценить степень связанности этих признаков.

177	71	17	14
95	119	75	25
12	44	23	8

7. Предположим, что 5% всех студентов носят очки. Чему равна вероятность того, что в группе из 25 человек не будет ни одного, будут 1, 2, 3, носящих очки? /10/.

8. У 12 работающих на ультразвуковых установках изучалось содержание сахара в крови натощак до работы и через три часа после работы.

Есть ли различия? Исходные данные приведены в таблице /18/.

Содержание сахара в крови обследованных натощак до работы и после 3 часов работы на ультразвуковых установках

№ пп	САХ_ДО	САХ_ПОС	№ пп	САХ_ДО	САХ_ПОС
1	112	54	**7**	64	64
2	82	67	**8**	70	66
3	101	96	**9**	88	48
4	72	59	**10**	81	50
5	79	79	**11**	66	61
6	82	76	**12**	88	61

9. Проверьте гипотезу о связи форм галек из отложений маршельской морены с их составом, где большая выборка галек из отложений ледниковой морены разделена по форме на угловатые и окатанные, а по составу на гранитные и метаморфические. Данные приведены в табл.4.3 и в /2 /.

Таблица 4.3

Связь формы галек из отложений маршельской морены с их составом

	Угловатые	**Окатанные**	**Сумма в строке**
Гранитные	41	170	211
Метаморфические	14	42	56
Сумма в столбце	55	212	267

10. (Несколько иной подход к использованию χ^2-критерия. При расчетах используйте EXCEL.)

Равномерность расположения точек является важным условием, необходимым для применения многих видов анализа карт. Достоверность карты находится в прямой зависимости от плотности и равномерности расположения точек наблюдения. Однако большинство геологов оценивают распределение точек наблюдения лишь с качественных позиций. Даже несмотря на то, что часто подчеркивается желание получить равномерное распределение точек наблюдения, степень равномерности крайне редко измеряется. χ^2-критерий можно использовать для количественной оценки расположения точек на двумерной поверхности (карте) /2 /.

Самостоятельное задание.

Проведите количественный анализ равномерности расположения скважин на рис 4.14 при большем 12 числе квадратов и меньшем 12.

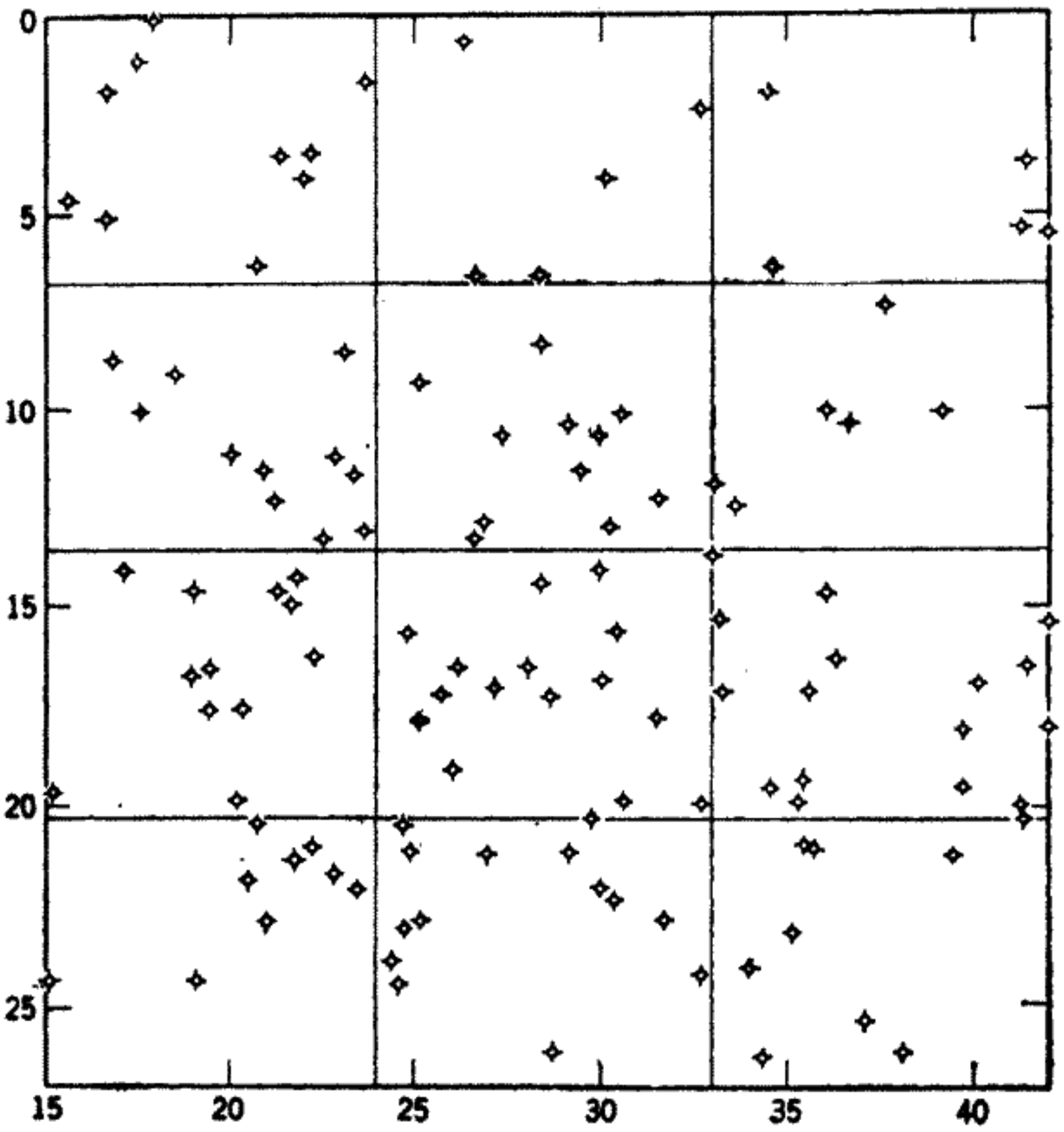


Рис 4.14. Расположение 123 скважин, вскрывающих кровлю ордовикских пород: Карта разделена на 12 клеток одинаковых размеров

11. Предположим, например, что получено пять образцов песчаника с кальцитовым цементом. Каждый из них обладает своими литологическими особенностями: в одном бросается в глаза его крупнозернистость, другой характеризуется наличием глинистых частиц, третий слабо ожелезнен и т.д. Нужно определить, одинаковы ли в них содержания карбоната.

Экспериментальный подход к этой задаче заключается в дроблении образцов на более мелкие части и определении содержания карбоната в каждой из них путем взвешивания после обработки кислотой. Каждая мелкая часть называется повторением. Цель, которая преследуется, при разбивании первоначального куска на части, — определение изменчивости, вызванной погрешностями взвешивания. Очевидно, что если изменчивость между повторными определениями для одного образца велика по сравнению с различиями между образцами, то последние трудно обнаружить.

Предположим, что мы разбили исходный образец на шесть частей и собираемся проанализировать каждую из них. Наблюдаемые изменения возникают по ряду причин – из-за колебаний состава внутри исходного образца, из-за небрежности в получении повторных наблюдений (остатки одного травления могут быть промыты более тщательно, чем остатки другого), из-за изменения условий взвешивания (повторные образцы могут содержать различные количества влаги либо на результаты взвешивания может повлиять зависимость положения нулевой точки на весах от изменений температуры в течение дня и т. д.) и благодаря влиянию других более тонких факторов. Комбинация всех этих источников изменчивости приводит к возникновению так называемой экспериментальной ошибки или изменчивости, не учитываемой только различиями между образцами.

Для того чтобы избежать возможности появления систематической ошибки в статистическом анализе, повторные наблюдения должны быть отобраны наудачу. Это так называемая рандомизация наблюдений. Необходимость этой процедуры станет очевидной, если имеется некоторый фактор, который непрерывно изменяется во время эксперимента, например продолжающееся высыхание проб, ожидающих своей очереди взвешивания. Если взвесить все шесть проб, полученных из образца 1, а затем все пробы, полученные из образца 2 и т.д., то при последнем взвешивании могут быть зарегистрированы большие весовые потери лишь по той причине, что пробы высыхали в течение более продолжительного периода времени. Один из способов решения этой задачи— последовательная нумерация каждой повторной процедуры и выбор этих номеров в процессе анализа по таблице случайных чисел. Действительно, если процесс протекает поэтапно, то целесообразно приписать наудачу номера каждому образцу на каждом шаге. Тогда различные источники ошибок перемешиваются или совмещаются для всех повторных проб, а не концентрируются в нескольких из них.

Проверьте гипотезу эквивалентности пяти образцов с помощью однофакторного дисперсионного анализа, при котором проверяемая гипотеза и альтернатива имеют следующий вид:

$$H_0: \mu_1 = \ldots\ldots = \mu_5 .$$

H_1: по крайней мере одно среднее значение отлично от остальных.

Таблица 4.5

Содержание карбонатного цемента в пяти образцах песчаника, % (числа в скобках обозначают порядковый номер пробы в процессе анализа)

Номер поэтапной пробы	Номер образца				
	1	2	3	4	5
1	19,2(11)	18,7(04)	12,5(28)	20,3(12)	19,9(21)
2	18,7(08)	14,3(19)	14,3(16)	22,5(30)	24,3(06)
3	21,3(09)	20,2(14)	8,7(20)	17,6(24)	17,6(18)
4	16,5(17)	17,6(07)	11,4(29)	18,4(03)	20,2(22)
5	17,3(26)	19,3(05)	9,5(27)	15,9(13)	18,4(12)
6	22,4(15)	16,1(25)	16,5(01)	19,0(02)	19,1(10)

Данные для рассматриваемой нами задачи приведены в табл. 4.5 /6/. Просчитайте пример с использованием Excel. Сделайте интерпретацию.

12. Ордовикские песчаники Сент-Питер представлены очень чистыми ортокварцитами, которые распространены в верховьях р. Миссисипи.

Так как зерна этих пород хорошо окатаны и отсортированы, то они необыкновенно однородны по своему строению. В связи с этим нефтяные месторождения, приуроченные к песчаникам, при добыче нефти путем откачки ведут себя так, как можно в точности предсказать с помощью теоретических моделей их поведения, хотя последние построены на основе идеализации условий. Отклонения поведения модели от действительности могут указать на ошибочность допущений в структуре модели.

Небольшой нефтяной район в южном Иллинойсе представляется идеально приспособленным для исследования совпадения в поведении модели и реального нефтяного месторождения. Так как этот район арендовался только одной компанией, тщательно хранившей документацию, то данные о добыче нефти из этого месторождения оказались доступными для исследования. Однако, прежде чем выполнить исчерпывающий анализ поведения месторождения, целесообразно проверить на примере вышеупомянутого песчаника предположение об однородности его свойств.

Из множества скважин, пробуренных в процессе разработки, десять были выбраны случайным образом для проведения анализа. В каждой пробе наудачу был высечен 1 куб породы объемом 16 $см^3$ таким образом, чтобы вертикальная ориентация пробы сохранялась. С помощью соответствующего прибора были сделаны два измерения скорости движения флюида сквозь высеченные кубы: в вертикальном направлении по отношению к слоистости и в горизонтальном, параллельно слоистости. Используя эти измерения, вычислили проницаемость образца в квадратных микрометрах.

Двадцать вычисленных значений проницаемости приведены в табл.4.6. По этим двадцати значениям требуется получить ответ на вопрос: имеются ли значимые различия в проницаемости, зависящие от положения образца в изучаемом районе (т. е. от расположения скважин) или от выбранных направлений измерения? /6/.

Таблица 4.6

Проницаемость случайно отобранных образцов песчаников Сент-Питер (штат Иллинойс), измеренная в различных направлениях, мкм^2

Направления			
Вертикальное	Горизонтальное	Вертикальное	Горизонтальное
1,037	1,124	0,928	0,943
0,963	0,960	1,108	1,165
0,842	0,921	0,821	0,803
1,121	1,202	0,797	0,792
1,043	1,028	0,949	1,004

Просчитайте пример с использованием Excel. Сделайте интерпретацию.

13. Рассмотрим одну из распространенных ситуаций прогноза слабоизученного геологического параметра с помощью сведений об изменениях другого, хорошо изученного параметра /4/. Эта ситуация, характеризующая ухудшение геологической изученности с глубиной, в особенности типична для изучения структурных планов глубокозалегающих границ (например, кровля триасовых отложений в табл.4.7). Расположение скважин показано на рис. 4.15, где параметрические скважины обозначены как р вместо П – .. как в табл.4.7.

Допустим, что по данным структурно-поискового бурения на площади работ построена структурная карта по неглубоко залегающей поверхности нижнемеловых отложений (табл. 4.8). Кроме того, на площади пробурено пять глубоких (параметрических) скважин, вскрывающих погруженные, перспективные в нефтегазоносном отношении триасовые отложения (таб.4.7). Положение параметрических скважин на площади и их количество не позволяют построить достоверную структурную карту поверхности триасовых отложений традиционным методом интерполяции.

Воспользуемся для решения этой задачи статистической простой линейной моделью, описывающей связь глубин залегания прогнозного горизонта (триаса) и хорошо изученной поверхности (нижнего мела) в виде

$$H_2 = a + b * H_1,$$

где H_2 _ прогнозная глубина поверхности триасовых отложений; H_1 - фактическая глубина кровли нижнемеловых отложений; **a** и **b** – постоянные коэффициенты.

Таблица 4.7

Исходные данные

Параметрические скважины	Глубина, м	
	кровли нижнемеловых отложений	кровли триасовых отложений
П-18	614	1539
П-19	626	1554
П-20	633	1568
П-21	646	1581
П-22	657	1598

Для расчета постоянных коэффициентов воспользуемся программой регрессионного анализа EXCEL. Результат показан на рис. 4.16

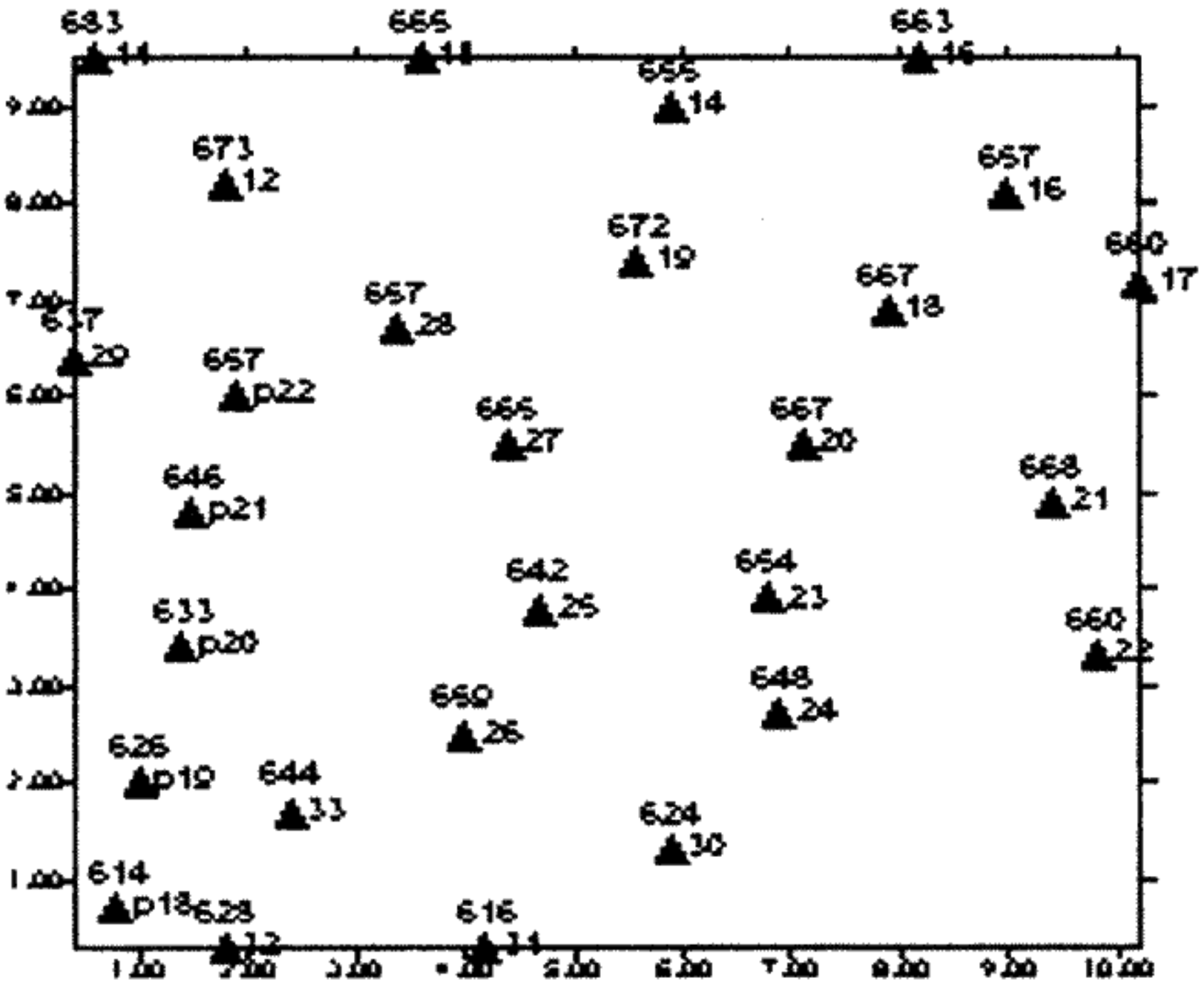


Рис.4.15

		H2	**H1**
1	**П-18**	1539	614
2	**П-19**	1554	626
3	**П-20**	1568	633
4	**П-21**	1581	646
5	**П-22**	1598	657

Регрессион. статистика	
Множествен.R	0.9966
R-квадрат	0.9932
Нормир.R-квад	0.9910
Стандартная ошибка	2.1731
Наблюдения	5

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	2091.832	2091.832331	442.94	0.000235
Остаток	3	14.16767	4.722556302		
Итого	4	2106			

	Коэф-фици-енты	*Станд. ошибка*	*t-статистика*	*P-Значе-ние*	*Нижние 95%*	*Верх-ние 95%*
Y-пересечение	**704.065**	41.06084	17.1468743	0.0004	573.391	834.73
Переменная H1	**1.36009**	0.064624	21.0462567	0.0002	1.154436	1.5657

Наблюдение	*Пред-сказан-ное H2*	*Остатки*
1	1539.165	-0.1659
2	1555.487	-1.48709
3	1565.007	2.99221
4	1582.689	- 1.68907
5	1597.650	0.34984

Рис. 4.16

По данным решения (рис.4.16),уравнение связи будет иметь вид

$$\mathbf{H_2 = 704.065 + 1.360\ H_1} \qquad (4.2)$$

Результаты этого прогноза приведены в табл.4.8 (столбец H_2) .

Таблица 4.8

№ пп.	Номер скважины	X	Y	H_1	H_2
1	11	0.6	9.5	683	1633.012
2	12	1.8	8.2	673	1619.411
3	13	3.6	9.5	665	1608.530
4	14	5.9	9	655	1594.929
5	15	8.2	9.5	663	1605.810
6	16	9	8.1	657	1597.650
7	17	10.2	7.2	660	1601.730
8	18	7.9	6.9	667	1611.251
9	19	5.6	7.4	672	1618.051
10	20	7.1	5.5	667	1611.251
11	21	9.4	4.9	668	1612.611
12	22	9.8	3.3	660	1601.730
13	23	6.8	3.9	654	1593.569
14	24	6.9	2.7	648	1585.409
15	25	4.7	3.8	642	1577.248
16	26	4	2.5	659	1600.370
17	27	4.4	5.5	665	1608.530
18	28	3.4	6.7	657	1597.650
19	29	0.4	6.4	637	1570.448
20	30	5.9	1.3	624	1552.766
21	31	4.2	0.3	616	1541.886
22	32	1.8	0.3	628	1558.207
23	33	2.4	1.7	644	1579.968

Задания.

а. С использованием уравнения связи (4.2) и программы SURFER/ 20/ построить прогнозную структурную карту триасовых отложений (**H_2**). Шаг изолиний задать 5 м. На карту вынести скважины с их номерами и отметками в них (координаты X и Y внесены в табл. 4.8 с карты расположения скважин – рис 4.15).

б. Построить тренды кровли нижнемеловых отложений (H1) по формулам:

H1 = a0 + a1*X + a1*Y – *тренд первой степени*

H1 = a0 + a1*X + a2*Y + a3*X^2 + a4*X*Y + a5*Y^2 –

– *тренд второй степени*

Использовать данные табл.4.8. Предварительно самостоятельно подгото-

вить регулярную координатную сеть (X = от 0 до 11, от Y = 0 до 10). Использовать программы EXCEL и SURFER. (^ – знак возведения в степень, * – знак умножения, данные для расчета тренда в Excel формируются аналогично данным в табл.4.9).

Таблица 4.9

X	Y
0	0
0	1
0	2
0	3
0	4
0	5
0	6
0	7
0	8
0	9
0	10
1	0
1	1
.....	.....
3	0
3	1
3	2
3	3
.....	.....
11	4
11	5
11	6
11	7
11	8
11	9
11	10

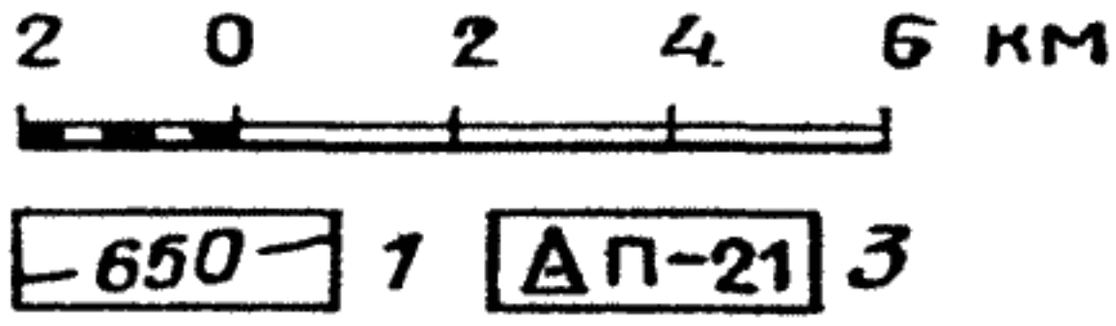
X изменяется от 0 до 11
Y изменяется от 0 до 10

в) В самостоятельных работах /4/, представленных ниже, приведены данные для построения прогнозных карт.

Проведите по этим данным построения, начиная с получения модели типа (4.2) с помощью регрессионного анализа (EXCEL) и заканчивая “рисовкой” карты по программе SURFER. Первоначально подготовьте

таблицу (без Н2) аналогичную табл.4.8 .

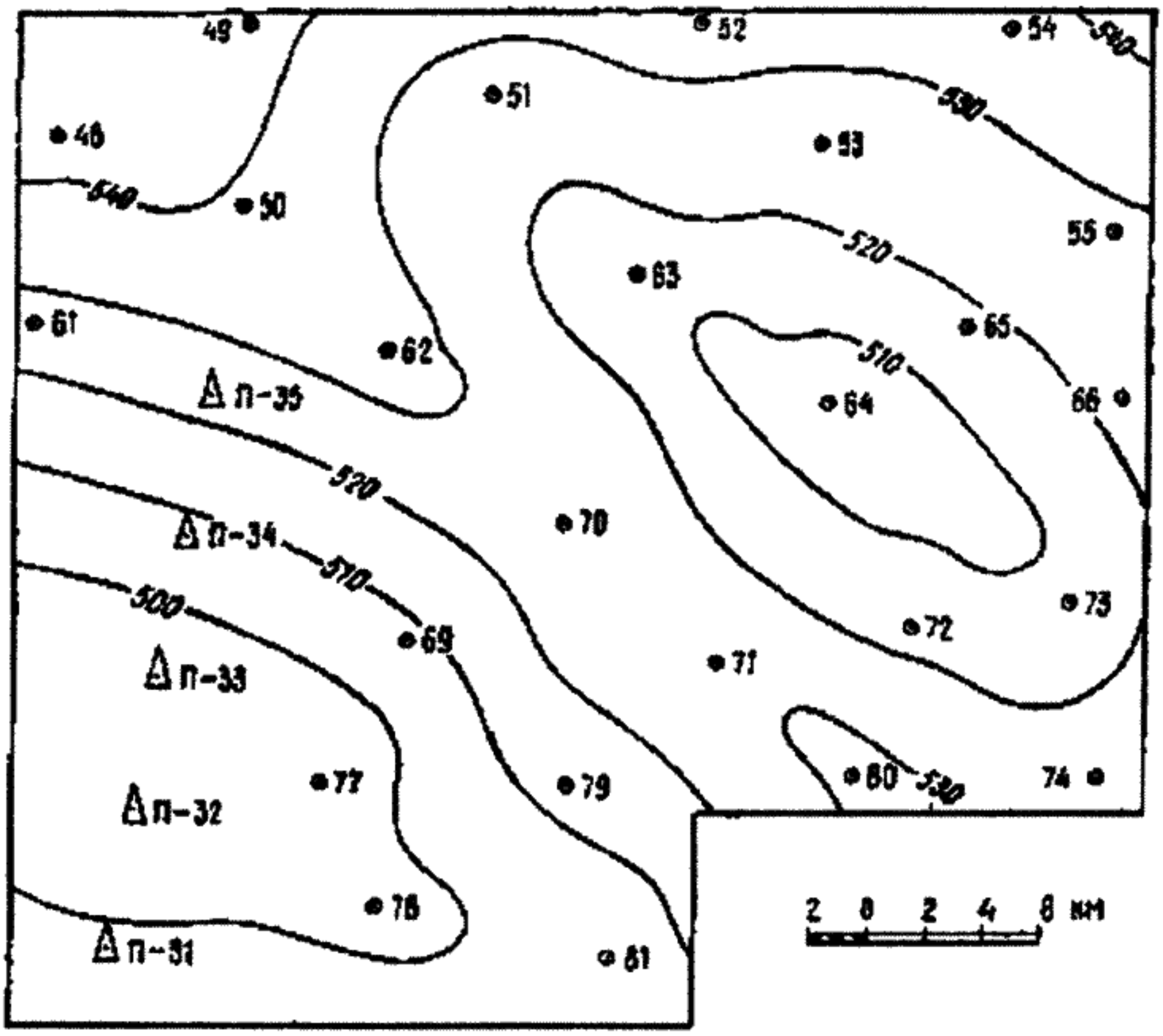
Условные обозначения к самостоятельным лабораторным работам.



Изогипсы кровли отложений: 1 – не глубоко залегающих поверхностей, 2 – глубоко залегающих и малоизученных поверхностей. Скважины: 3 – параметрические; 4 – структурно-поисковые.

САМОСТОЯТЕЛЬНАЯ РАБОТА 1

Структурная карта по подошве эоценовых отложений площади 1



Исходные данные

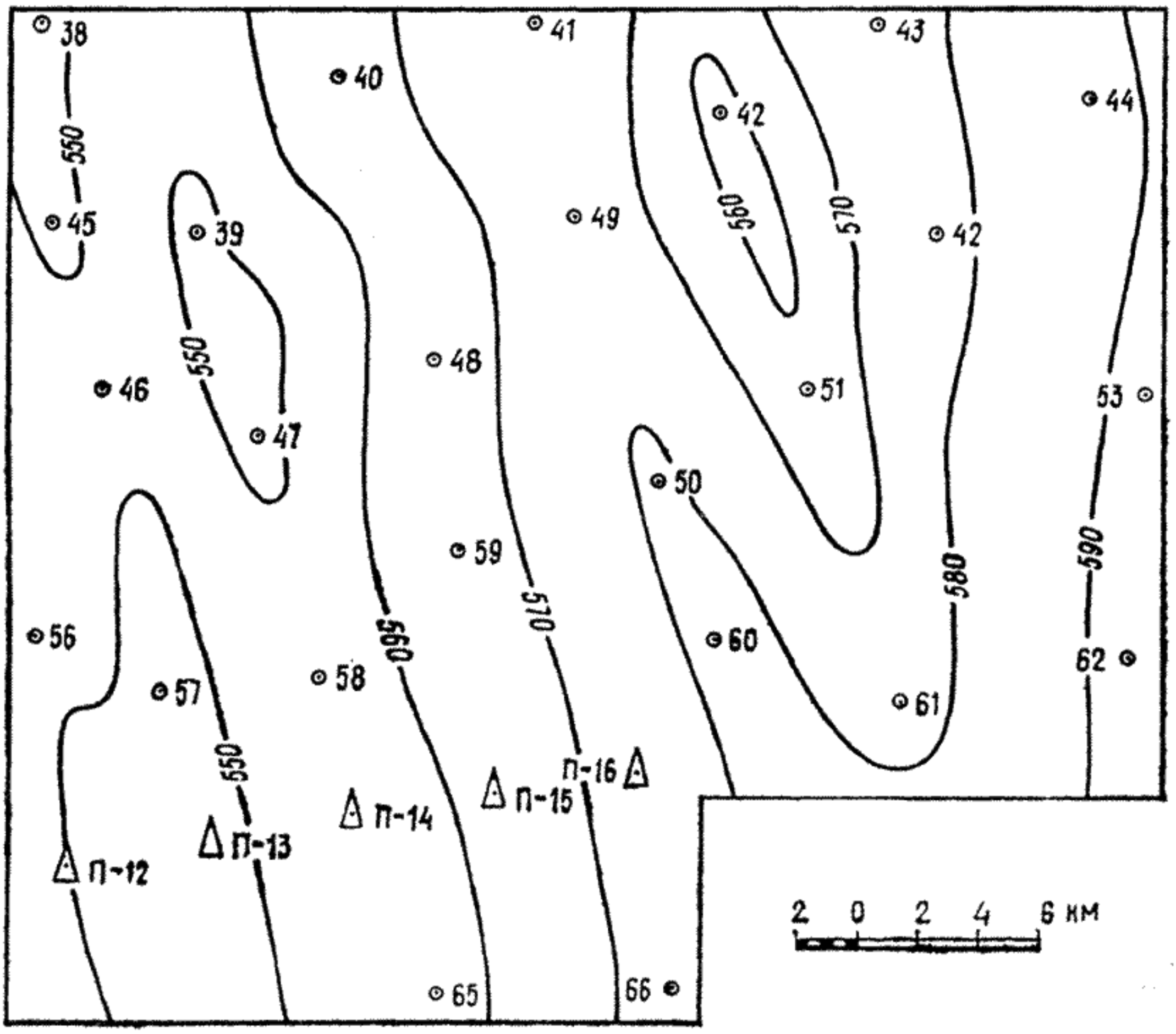
Параметрические скважины	Глубина, м	
	подошвы эоценовых отложений	кровли оксфордских отложений
П-31	503	1456
П-32	494	1443
П-33	497	1448
П-34	508	1472
П-35	528	1507

САМОСТОЯТЕЛЬНАЯ РАБОТА 2

Исходные данные

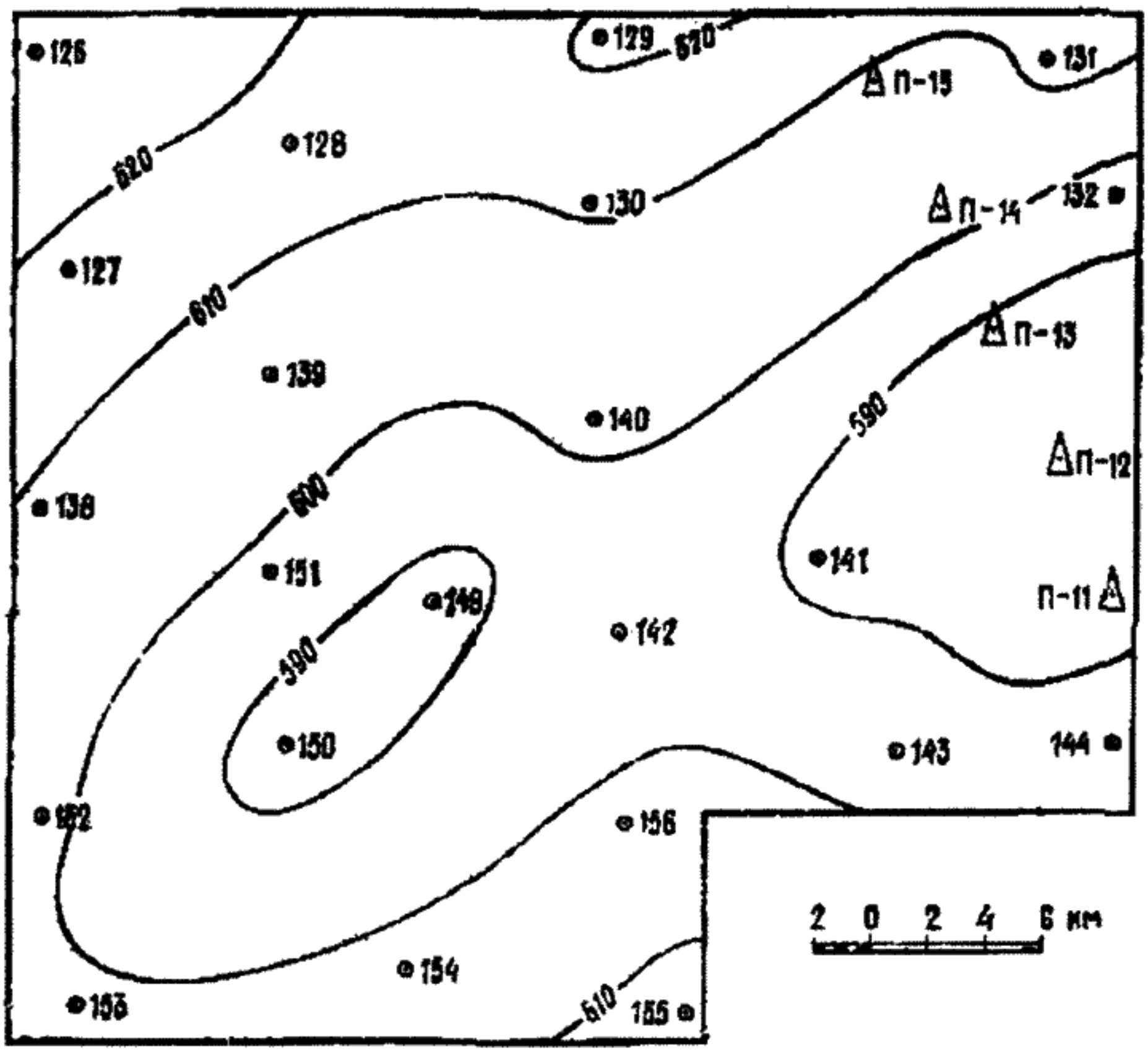
Параметрические скважины	Глубина, м	
	кровли нижнемеловых отложений	кровли нижнеюрских отложений
П-12	550	1379
П-13	548	1373
П-14	556	1387
П-15	563	1408
П-16	576	1433

Структурная карта по кровле нижнемеловых отложений площади 2.



САМОСТОЯТЕЛЬНАЯ РАБОТА 3

Структурная карта по подошве плиоценовых отложений площади 3



Исходные данные

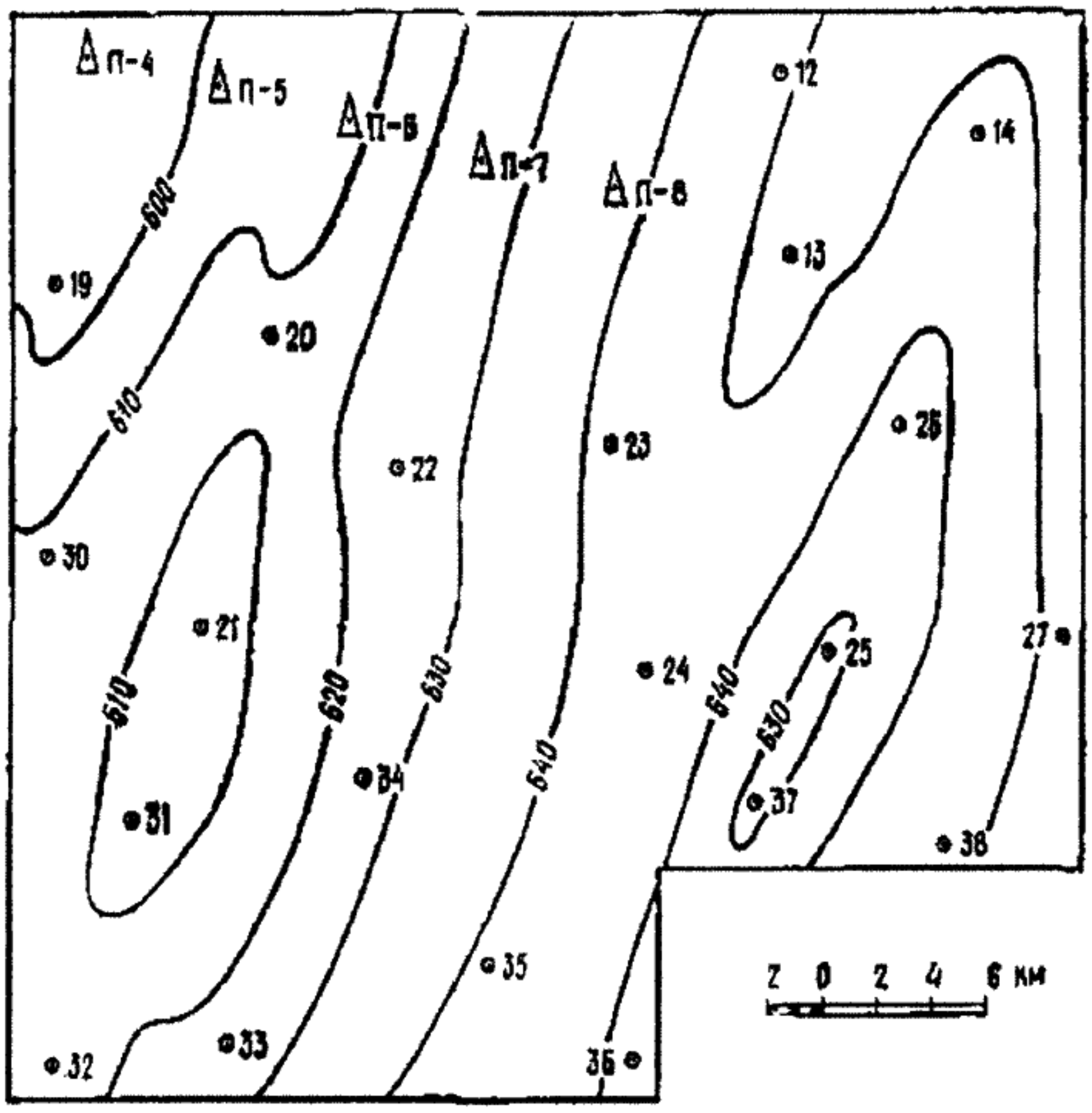
Параметрические скважины	Глубина, м	
	подошвы плиоценовых отложений	кровли юрских отложений
П-11	583	1742
П-12	580	1733
П-13	587	1753
П-14	603	1778
П-15	609	1791

САМОСТОЯТЕЛЬНАЯ РАБОТА 4

Исходные данные

Параметрические скважины	Глубина, м	
	подошвы меловых отложений	кровли триасовых отложений
П-4	598	1632
П-5	601	1639
П-6	608	1647
П-7	627	1686
П-8	637	1698

Структурная карта по подошве меловых отложений площади 4

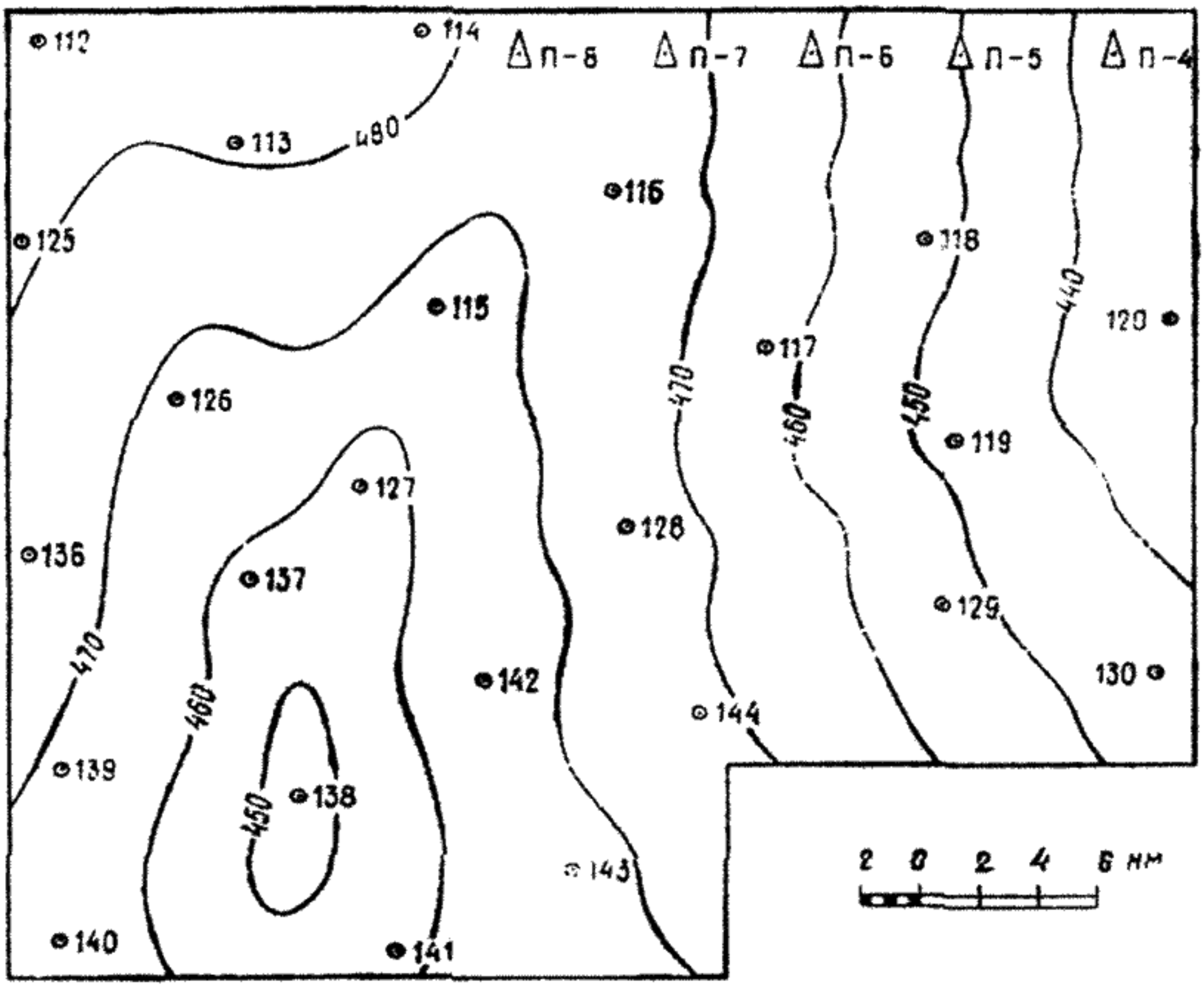


САМОСТОЯТЕЛЬНАЯ РАБОТА 5

Исходные данные

Параметрические скважины	Глубина, м	
	кровли неогеновых отложений	кровли барремских отложений
П-4	434	1603
П-5	450	1626
П-6	462	1656
П-7	472	1670
П-8	478	1684

Структурная карта по кровле неогеновых отложений площади 5

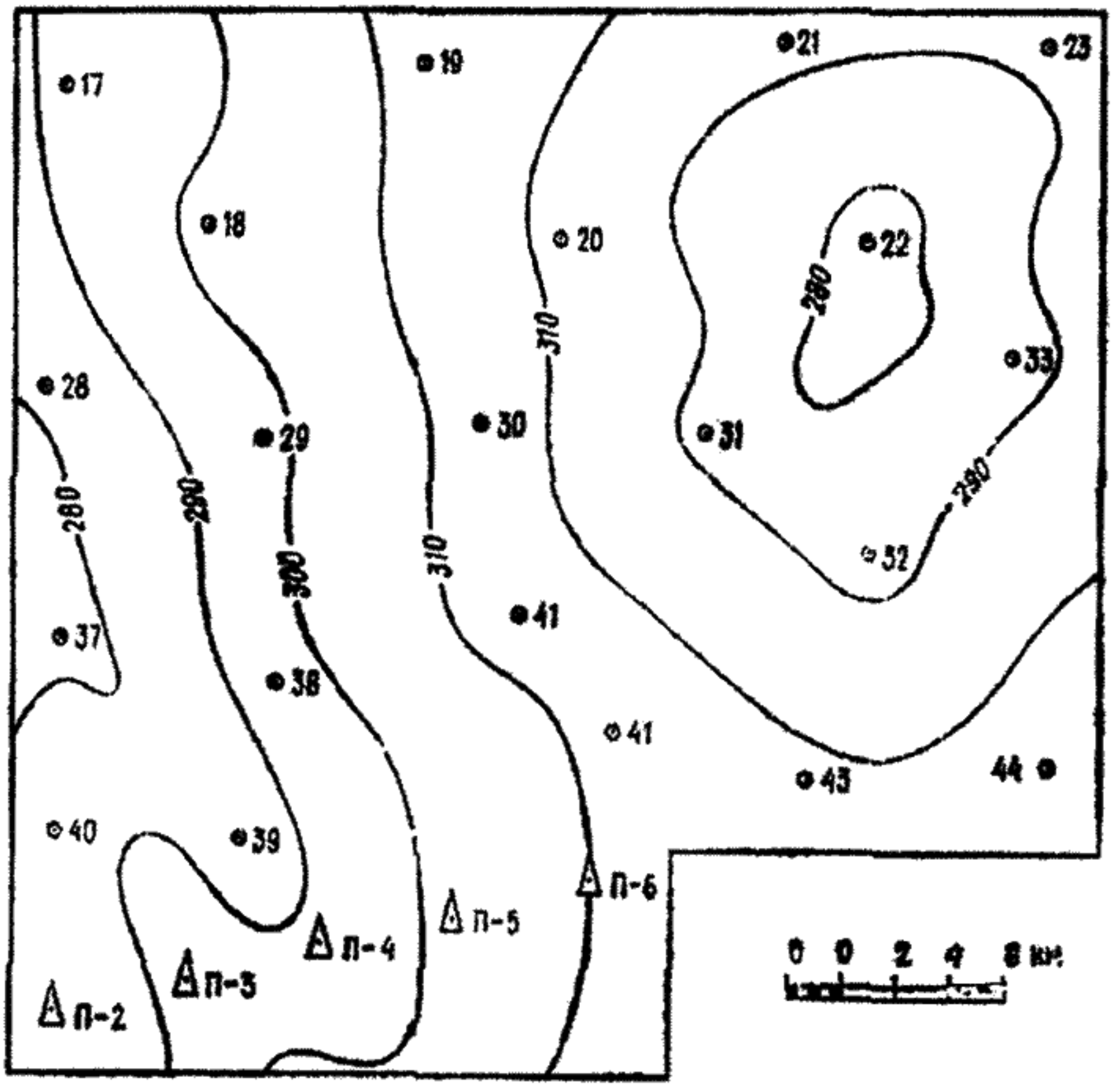


САМОСТОЯТЕЛЬНАЯ РАБОТА 6

Исходные данные

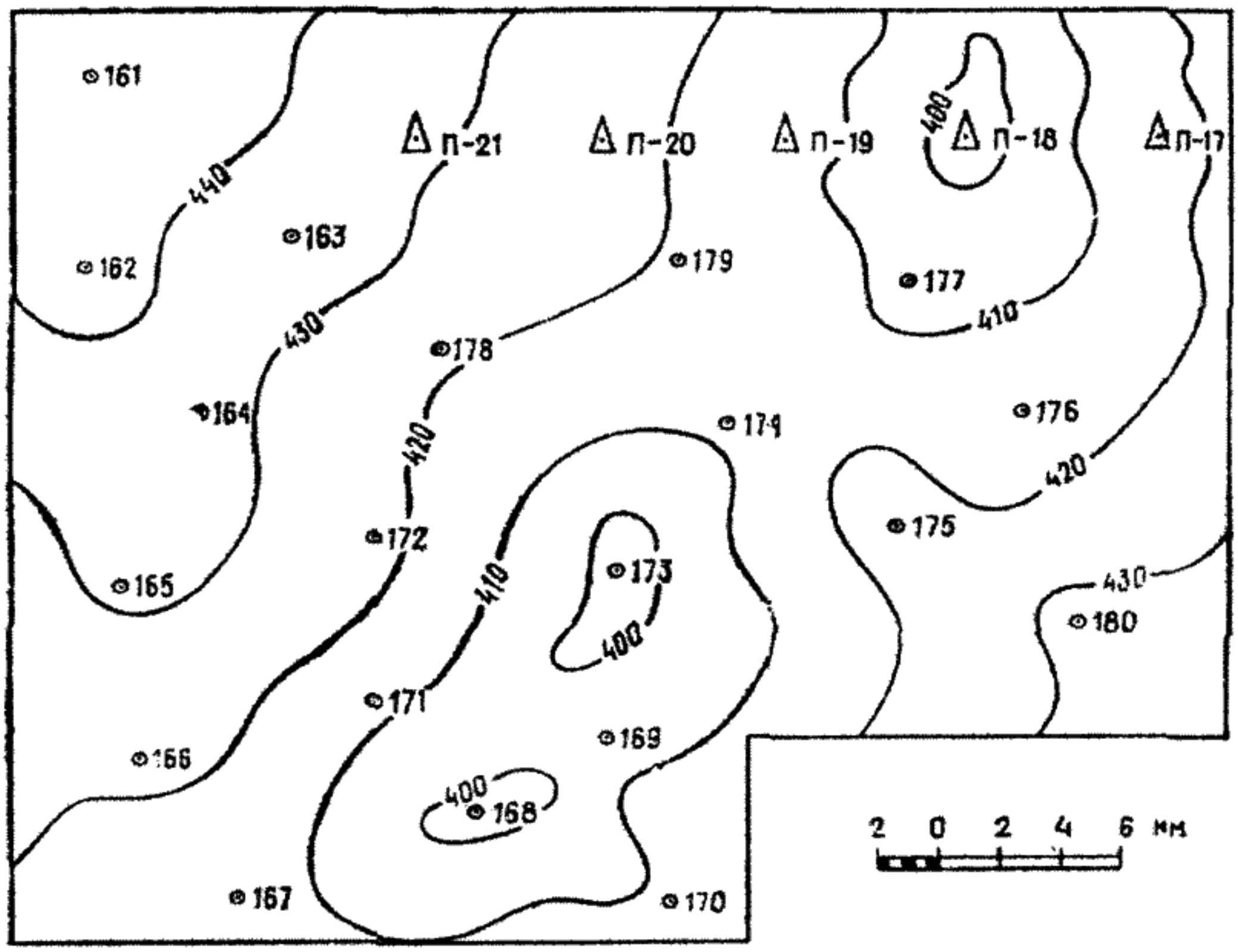
Параметрические скважины	Глубина, м	
	подошвы палеоценовых отложений	кровли сеноманских отложений
П-2	283	1081
П-3	292	1092
П-4	292	1101
П-5	305	1123
П-6	319	1127

Структурная карта по кровле отложений палеоцена площади 6



САМОСТОЯТЕЛЬНАЯ РАБОТА 7

Структурная карта по кровле отложений майкопской свиты площади 7



Исходные данные

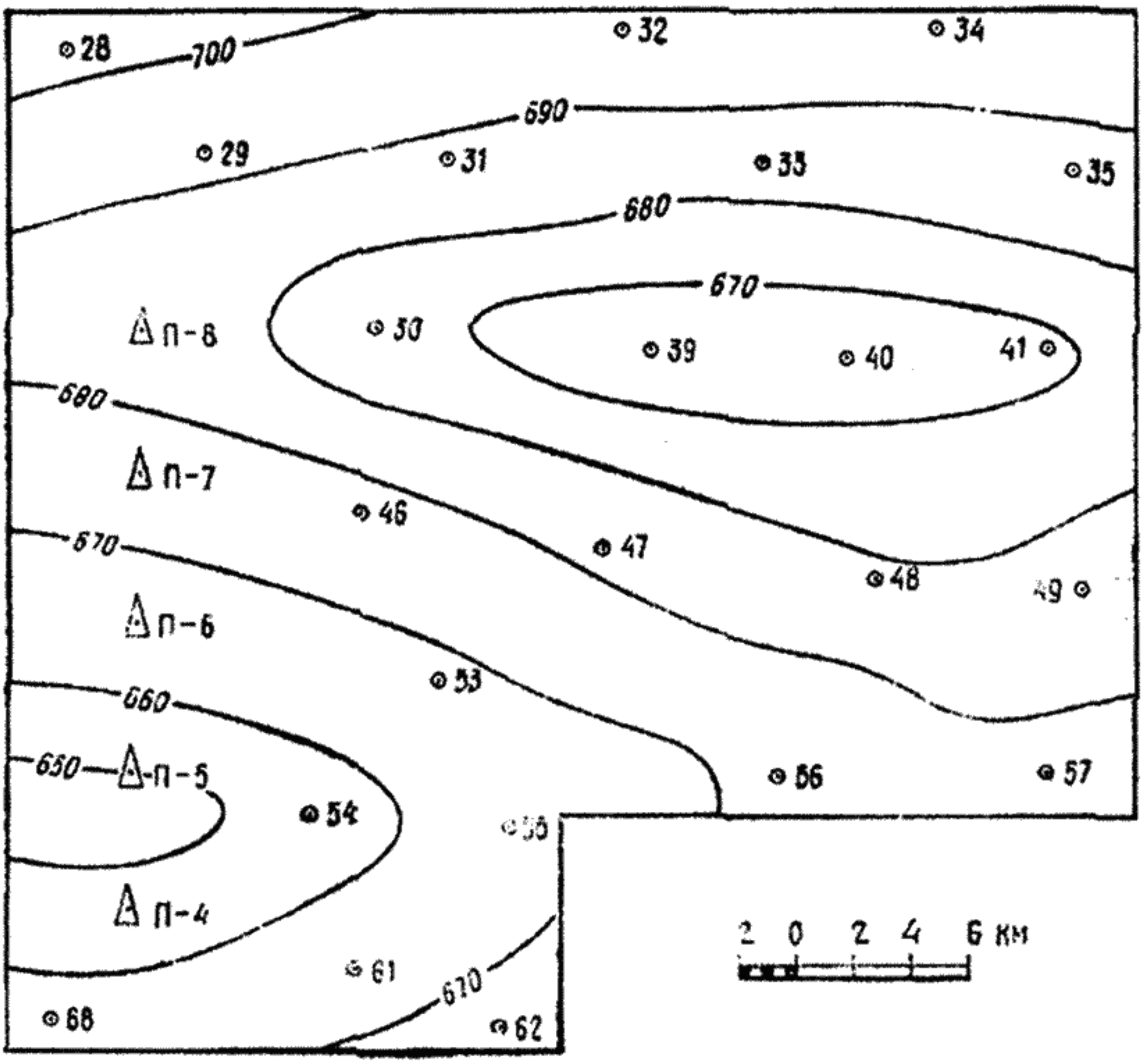
Параметрические скважины	Глубина, м	
	подошвы отложений майкопской свиты	кровли сеноманских отложений
П-17	418	1234
П-18	399	1194
П-19	414	1221
П-20	426	1246
П-21	431	1269

САМОСТОЯТЕЛЬНАЯ РАБОТА 8

Исходные данные

Параметрические скважины	Глубина, м	
	подошвы меловых отложений	кровли байосских отложений
П-4	655	1308
П-5	650	1306
П-6	667	1339
П-7	673	1332
П-8	685	1350

Структурная карта по кровле меловых отложений площади 8

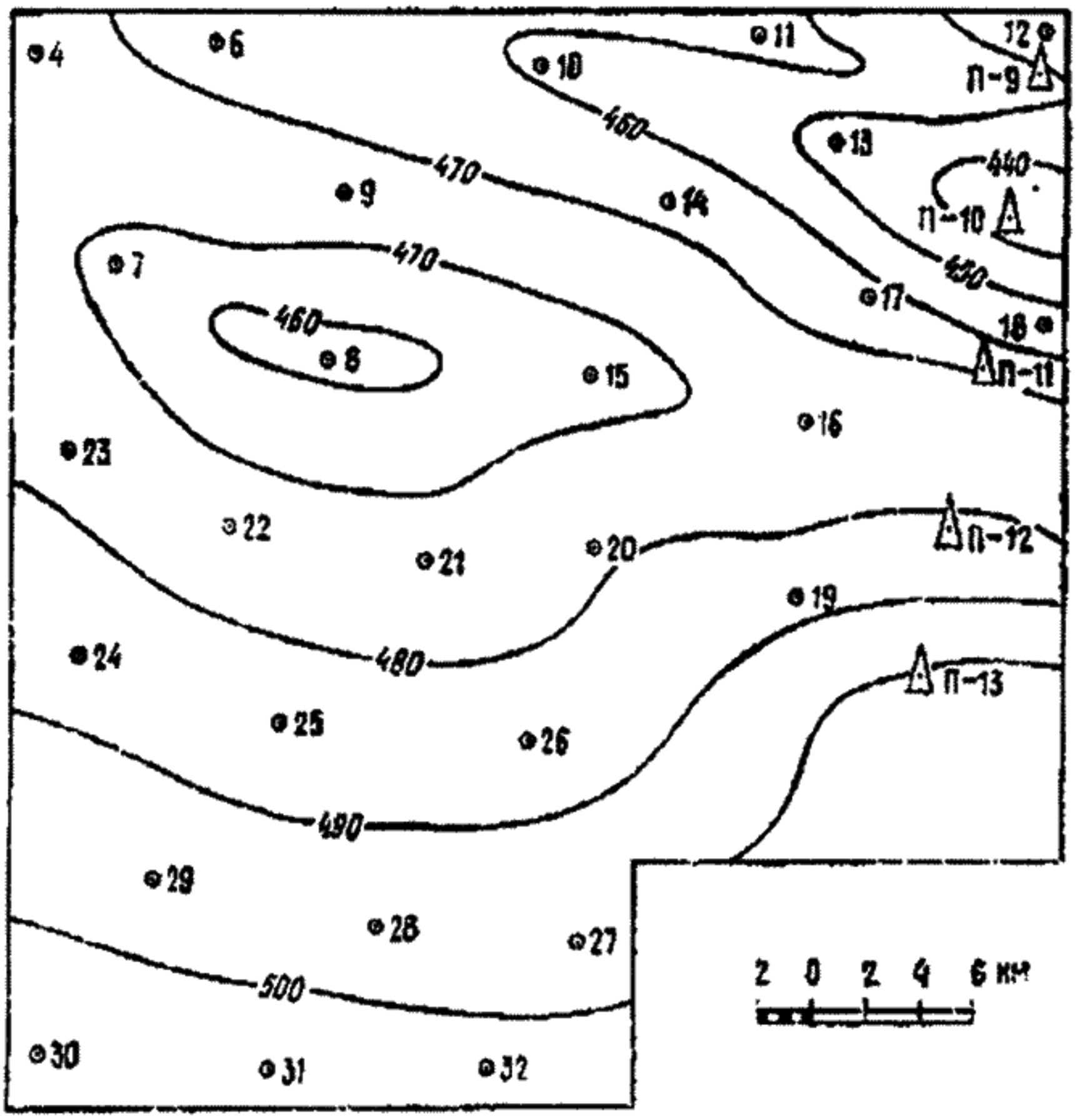


САМОСТОЯТЕЛЬНАЯ РАБОТА 9

Исходные данные

Параметрические скважины	Глубина, м	
	кровли миоценовых отложений	кровли юрских отложений
П-9	452	1232
П-10	438	1214
П-11	474	1273
П-12	481	1286
П-13	506	1325

Структурная карта по кровле отложений миоцена площади 9

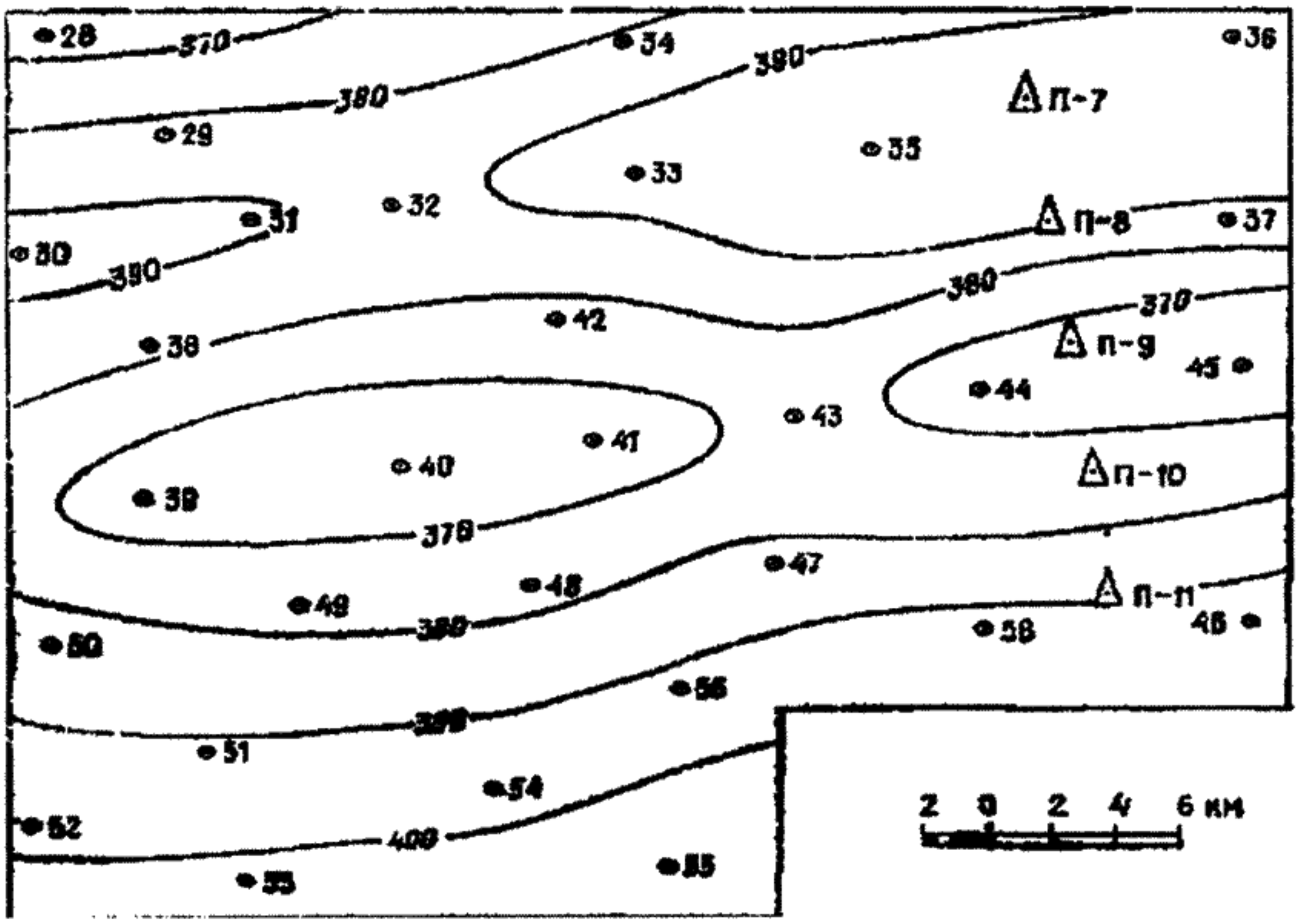


САМОСТОЯТЕЛЬНОЕ ЗАДАНИЕ 10

Исходные данные

Параметрические скважины	Глубина, м	
	подошвы неогеновых отложений	подошвы меловых отложений
П-7	398	1624
П-8	390	1612
П-9	362	1569
П-10	375	1582
П-11	390	1606

Структурная карта по подошве отложений неогена площади 10



14. Проведите построение многофакторной полиномиальной модели по данным (рис. 4.9) с помощью пакета Statistica 6.0 (если пакет "стоит" на Вашем **PC**). Сравните результаты (рис. 4.9).

4.4. Ответы и решения

1. Среднее – 3,09 (см ниже таблицу результатов); стандартное отклонение – 2,89; медиана – 2; 25-й процентиль – 1; 75-й процентиль – 5. Вряд ли данные извлечены из совокупности с нормальным распределением:

среднее довольно сильно отличается от медианы, медиана гораздо ближе к 25-му процентилю, чем к 75-му, а значит, распределение асимметрично. Поскольку среднее почти равно стандартному отклонению, в случае нормального распределения примерно 15% значений было бы меньше нуля. Поэтому отсутствие отрицательных значений также говорит против нормальности распределения.

Таблица результатов

Столбец 1			Точка	Столбец 1	Ранг	Процент
0			33	11	1	100.00%
0			32	10	2	96.80%
0	Столбец 1		31	9	3	93.70%
1			30	7	4	90.60%
1	Среднее	3.09090909	29	6	5	87.50%
1	Стандартная ошибка	0.50257600	25	5	6	75.00%
1	Медиана	2	26	5	6	75.00%
1	Мода	1	27	5	6	75.00%
1	Стандартное отклонение	2.88707936	28	5	6	75.00%
1	Дисперсия выборки	8.33522727	23	4	10	68.70%
1	Эксцесс	1.17924356	24	4	10	68.70%
1	Асимметричность	1.31357682	19	3	12	56.20%
1	Интервал	11	20	3	12	56.20%
1	Минимум	0	21	3	12	56.20%
2	Максимум	11	22	3	12	56.20%
2	Сумма	102	15	2	16	43.70%
2	Счет	33	16	2	16	43.70%
2	Наибольший(1)	11	17	2	16	43.70%
3	Наименьший(1)	0	18	2	16	**43.70%**
3			4	1	20	**9.30%**
3			5	1	20	9.30%
3			6	1	20	9.30%
4			7	1	20	9.30%
4			8	1	20	9.30%

5			9	1	20	9.30%
5			10	1	20	9.30%
5			11	1	20	9.30%
5			12	1	20	9.30%
6			13	1	20	9.30%
7			14	1	20	9.30%
9			1	0	31	.00%
10			2	0	31	.00%
11			3	0	31	.00%

2.Среднее — 244; стандартное отклонение — 43; медиана — 235,5; 25-й процентиль — 211; 75-й процентиль — 246. Выборка вполне может быть извлечена из совокупности с нормальным распределением: медиана близка к среднему и находится примерно посредине между 25-м и 75-м процентилями. Сравните с предыдущей задачей.

3.Среднее – 5,4; стандартное отклонение – 7,6; медиана – 2,0; 25-й процентиль – 1,6; 75-й процентиль – 2,4. Выборку нельзя считать извлеченной из нормально распределенной совокупности: среднее не только не равно медиане, но даже превышает 75-й процентиль. Стандартное отклонение превышает среднее, при этом среди данных нет отрицательных значений (и не может быть по самой природе данных). Высокие значения среднего и стандартного отклонения обусловлены главным образом двумя «выпадающими» значениями — 19,0 и 23,6.

4. а) a =24,3; b = 0,36; r = 0,561; б) a = 0,5; b = 1,15; r = 0,599. Первый пример показывает, сколь большое влияние может иметь единственная точка. Второй пример показывает, как важно нанести данные на график, прежде чем приступить к регрессионному анализу: здесь выборка явно разнородна и может быть описана двумя различными зависимостями. Условия применимости регрессионного анализа не соблюдены, и попытка выразить связь единственной линией регрессии несостоятельна.

5. F=64,18; $v_{меж.} = 4$; $v_{вну.} = 995$. Различия статистически значимы (максимальную объемную скорость середины выдоха нельзя считать одинаковой во всех группах, $F_{кр.} = 3,32$ для $\alpha = 0,01$).

6. $\chi^2 = 107,485$; v = 6; ($\chi^2_{критич.}$ для $\alpha = 0,05$ равно 18,548). Гипотеза о независимости признаков отвергается, т.е. признаки связаны между собой.

7. Ни одного – P=0,21; один – P=0.33; двое – P= 0,25; трое – P=0,12.

8.W = 64; n = 11; p < 0,06; $W_{крит.} = 44$ (p = 0,054). Снижение уровня содержания сахара в крови через три часа работы на ультразвуковых установках по сравнению с его уровнем натощак существенное.

9. χ^2=0,53. $\alpha = 0.05$. В связи с тем, что вычисленное значение = 0,53 меньше 3,84 (при $\alpha = 0.05.$), для отклонения гипотезы независимости изученных характеристик нет оснований, т.е. форму галек следует считать независящей от принадлежности образца к гранитам или метаморфическим породам.

10. Всю карту можно разделить на множество подобластей равного размера (иногда их называют *квадратами)* так, что каждая подобласть будет содержать некоторое множество точек. Если точки наблюдения расположены равномерно, то следует ожидать, что каждая подобласть будет содержать одно и то же число точек. Эту гипотезу об отсутствии существенных различий в числе точек для каждой подобласти можно проверить с помощью критерия χ^2, который теоретически не зависит от формы или ориентировки подобластей. Однако критерий будет наиболее эффективным, если число подобластей сделать по возможности большим (что приводит к увеличению числа степеней свободы), при условии, что все подобласти содержат не менее пяти точек. Ожидаемое число точек для каждой подобласти будет равно

$$E = \frac{\text{общее число точек наблюдения}}{\text{число подобластей}}, \qquad (4.1)$$

Критерий χ^2 для проверки гипотезы о равномерном распределении точек будет определен следующим образом:

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i},$$

где O_i – наблюдаемое число точек в подобласти с номером i, а E_j – ожидаемое число, определяемое выражением (4.1). Критерию χ^2 соответствует v = m – 2 степеней свободы, где m – число подобластей.

В качестве примера применения этого критерия рассмотрим данные, приведенные на рис 4.14, которые показывают расположение 123 нефтяных скважин в нефтегазоносном районе США. На рис 4.14 вся площадь карты разделена на 12 равных участков и число точек для каждого участка равно приблизительно 10. В табл.4.4 приведены наблюдаемые значения числа точек в каждом участке, а также показана процедура вычисления критерия χ^2. Так как в данном случае число степеней свободы v = 10, то критическое значение χ^2, соответствующее 5%-ному уровню значимости, равно 18,3. Вычисленное значение критерия равно 15,2, которое не превышает 18,3, что дает основание сделать вывод о несущественном отклонении распределения точек от равномерного.

Заметим, что этот вывод касается только однородности распределения точек по участкам определенного размера. Вполне возможно, что существует такой вариант размера квадратов (особенно меньший, чем выбранный), при котором гипотеза о равномерности будет отклонена.

Таблица 4.4

Число скважин по 12 клеткам карты

Наблюдаемое число точек	$\frac{(O-E)^2}{E}$	Наблюдаемое число точек	$\frac{(O-E)^2}{E}$
10	0,00	16	3,30
5	2,60	15	2,26
5	2,60	9	0,14
11	0,06	14	1,42
12	0,32	8	0,48
6	1,73	Сумма 123	$\chi^2 = 15,23$ [a]
12	0,32		

[a]-Значение критерия несущественно при уровне значимости $\alpha = 0,05$.

11.F=10.4. Проверьте.

12.

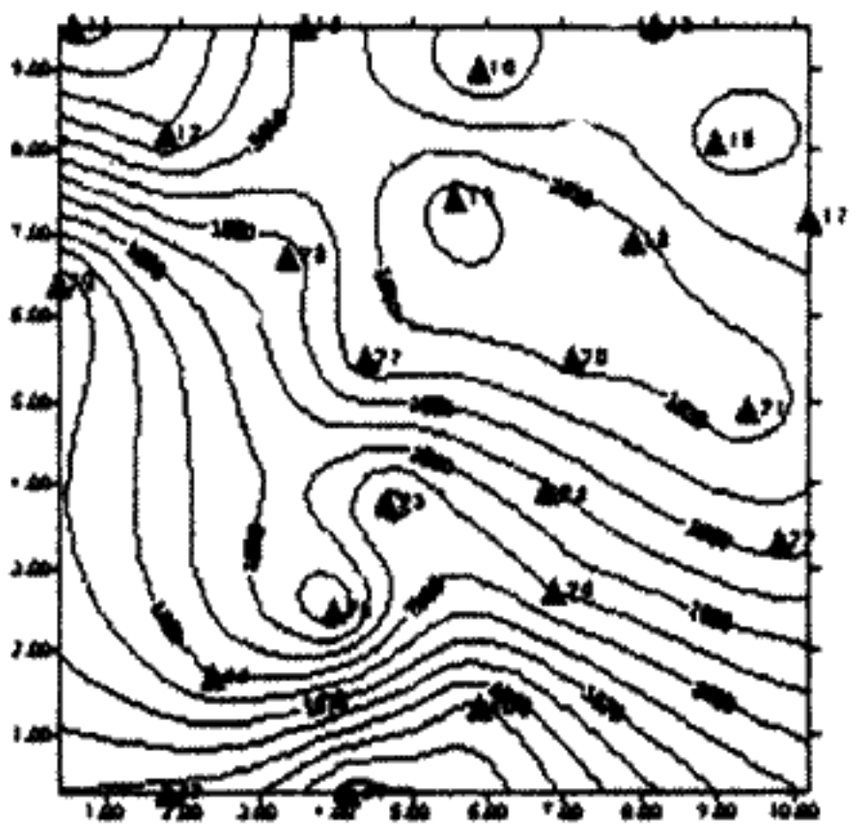
1.037	1.124	**Двухфакторный дисперсионный анализ без повторений**				
0.963	0.96					
0.842	0.921	***ИТОГИ***	***Счет***	***Сумма***	***Среднее***	***Дисперсия***
1.121	1.202	Строка 1	2	2.161	1.0805	0.0037845
1.043	1.028	Строка 2	2	1.923	0.9615	4.5E-06
0.928	0.943	Строка 3	2	1.763	0.8815	0.0031205
1.108	1.165	Строка 4	2	2.323	1.1615	0.0032805
0.821	0.803	Строка 5	2	2.071	1.0355	0.0001125
0.797	0.792	Строка 6	2	1.871	0.9355	0.0001125
0.949	1.004	Строка 7	2	2.273	1.1365	0.0016245
		Строка 8	2	1.624	0.812	0.000162
		Строка 9	2	1.589	0.7945	1.25E-05
		Строка 10	2	1.953	0.9765	0.0015125
		Столбец 1	10	9.609	0.9609	0.013511433
		Столбец 2	10	9.942	0.9942	0.019670178

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Строки	0.290452	9	0.032272	**35.49873809**	5.70613E-06	3.178897146
Столбцы	0.005544	1	0.005544	**6.098722203**	0.03559601	5.117357205
Погрешность	0.008182	9	0.000909			
Итого	0.304179	19				

13.

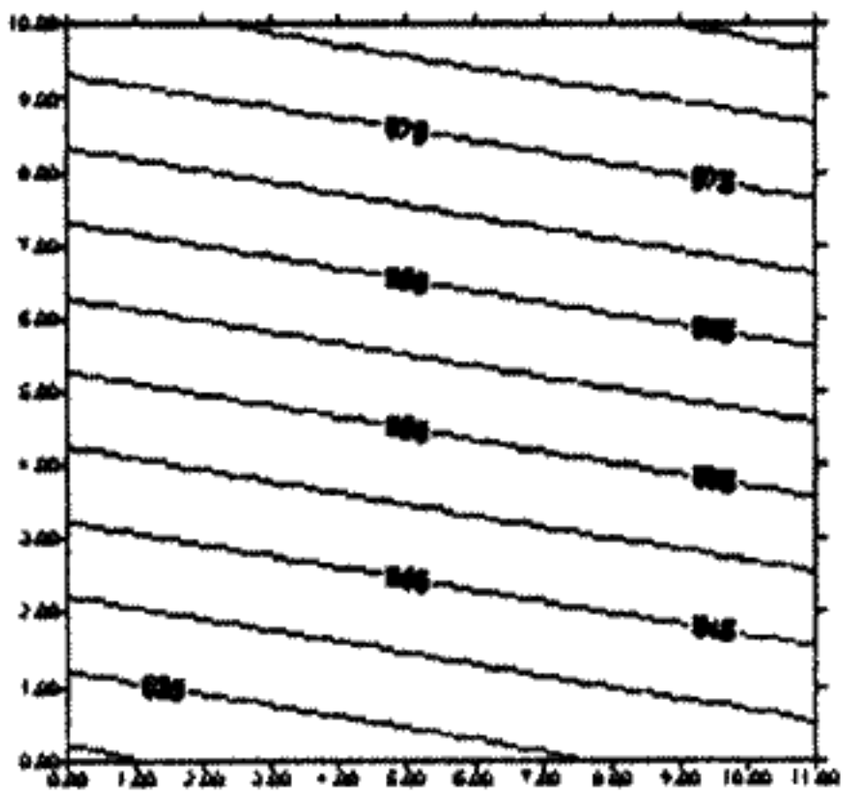
а.



Прогнозная структурная карта кровли триасовых отложений

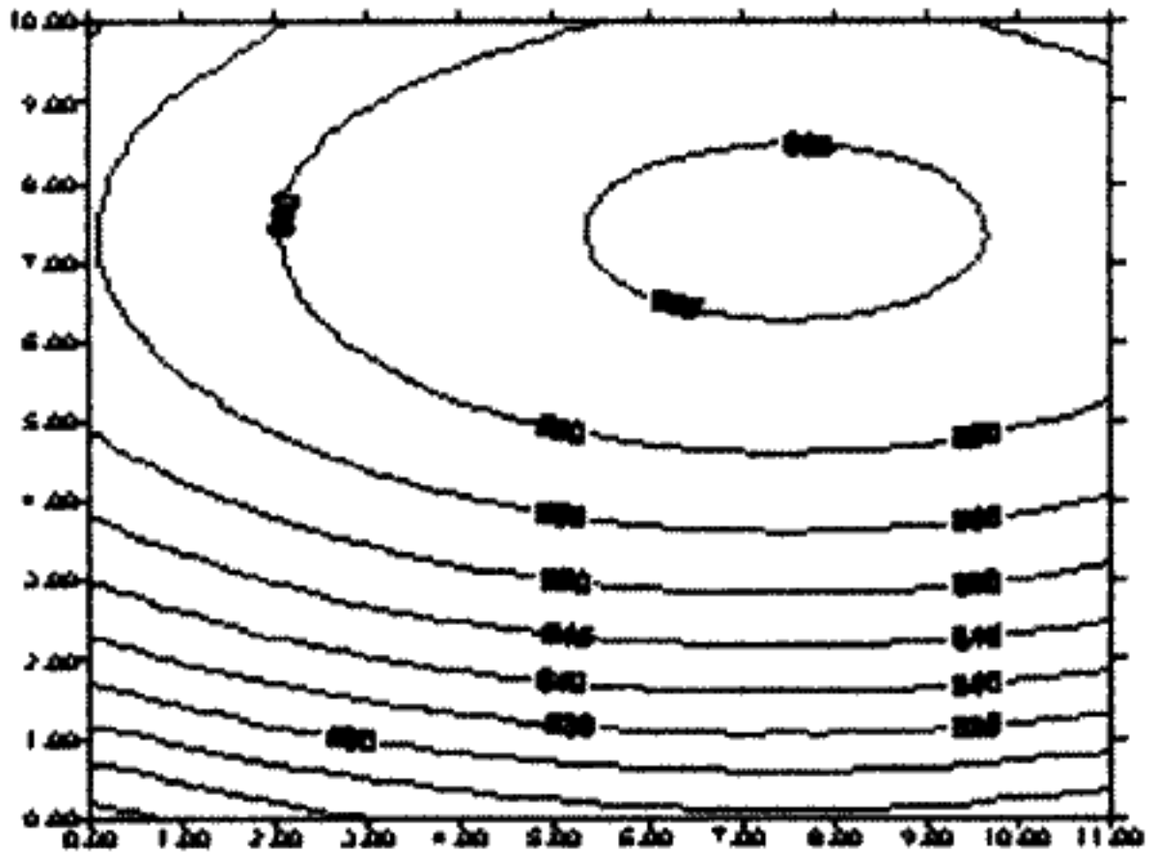
б.

$$H1 = 629.1 + 0.76*X + 4.91*Y$$



Тренд первой степени

$$H1 = 612.73 + 3.01*X + 11.37*Y - 0.20*X^2 - 0.71*Y^2$$



Тренд второй степени

в. *Самостоятельные работы (1–10) решаются по схеме построения прогнозной структурной карты кровли триасовых отложений.*

14. Запуск программы **STATISTICA 6.0.**

После ввода данных, которые возможно ввести либо "вставкой" их из программы Excel, где была раннее введена информация для расчета модели, либо как это было сделано при расчете критерия Манна – Уитни (4.2.2). Результат на рис.4.17.

	1 H2	2 H1	3 H0	4 H1^2	5 H0^2	6 H1*H0
1	2.11	1.7	0.91	2.89	0.83	1.55
2	1.87	1.51	0.78	2.28	0.61	1.18
3	1.91	1.52	0.72	2.31	0.52	1.09
4	1.81	1.42	0.66	2.02	0.44	0.94
5	1.7	1.33	0.52	1.77	0.27	0.69
6	1.28	1.04	0.31	1.08	0.10	0.32
7	1.06	0.93	0.24	0.86	0.06	0.22
8	1.6	1.31	0.56	1.72	0.31	0.73

Рис.4.17.

Далее в меню STATISTICA 6.0. выбираем **Статистика => Множественная** регрессия. Высвечивается окно **Составная линейная регрессия: 1** (рис. 4.18). Щелкаем по кнопке **Variables.**

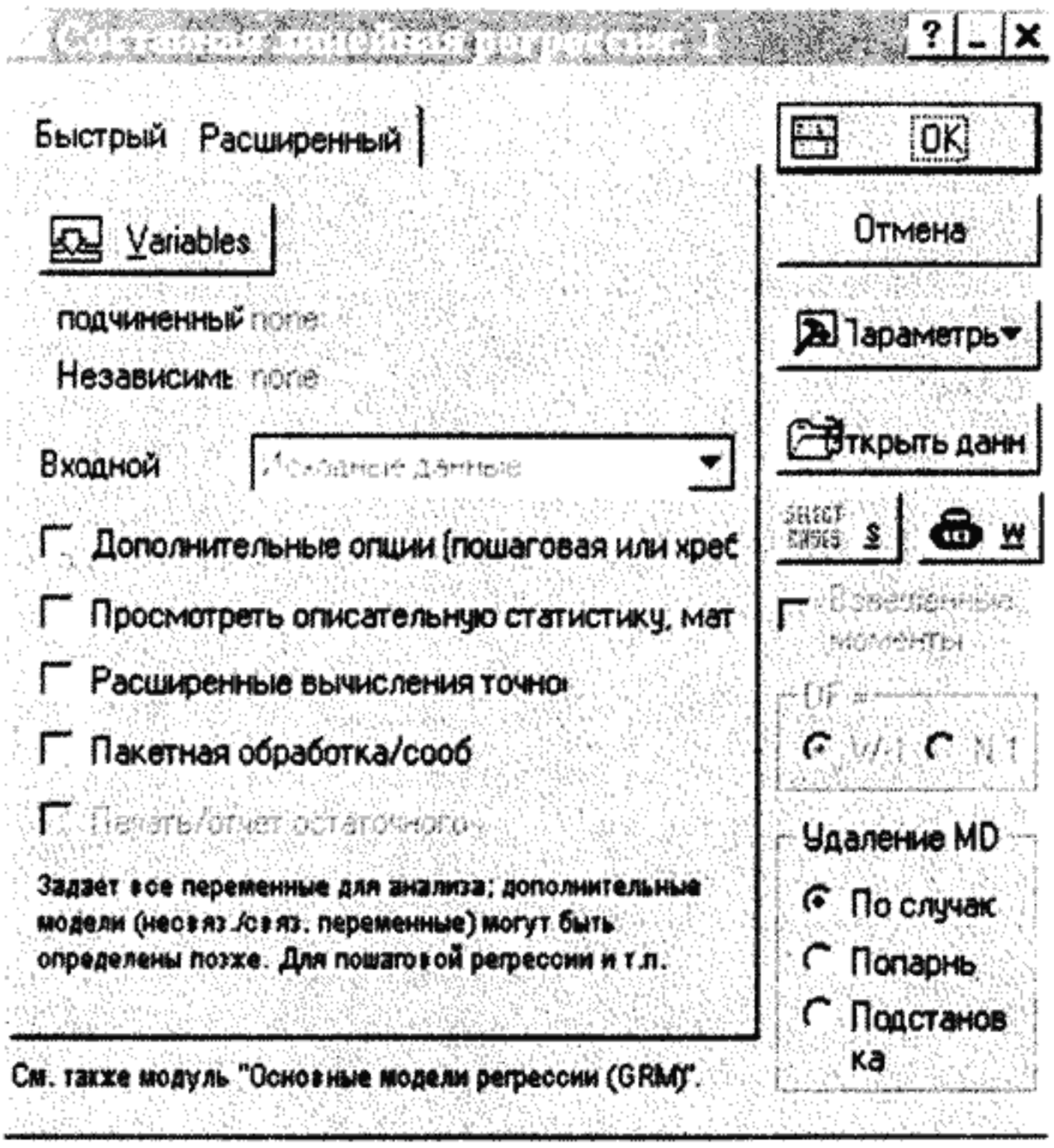


Рис. 4.18

На экране появляется форма (рис. 4.19), в которой в качестве подчиненной (зависимой) переменной выбираем H2, независимых: H1 – H1*H0. Нажимаем **OK**.

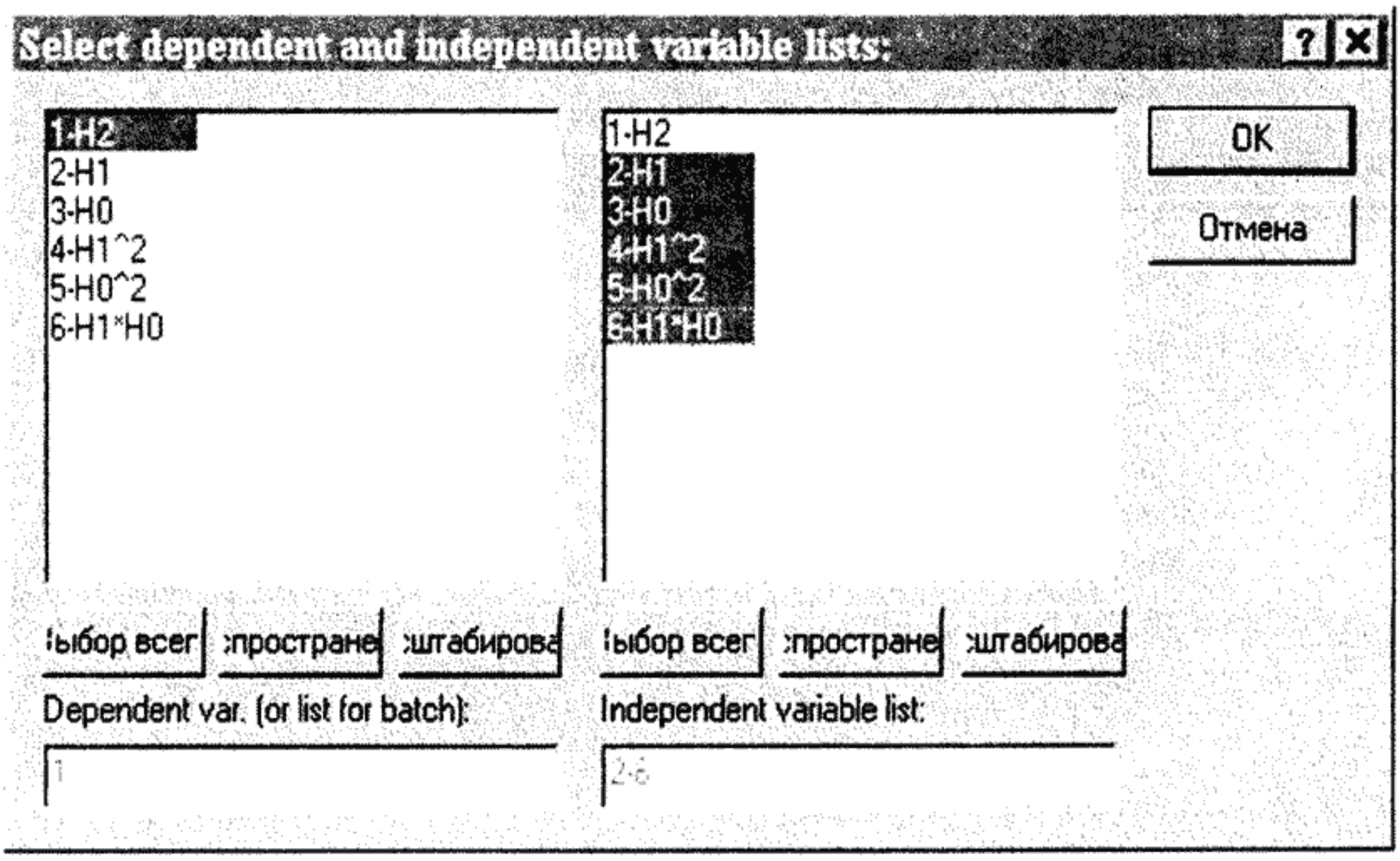


Рис .4.19

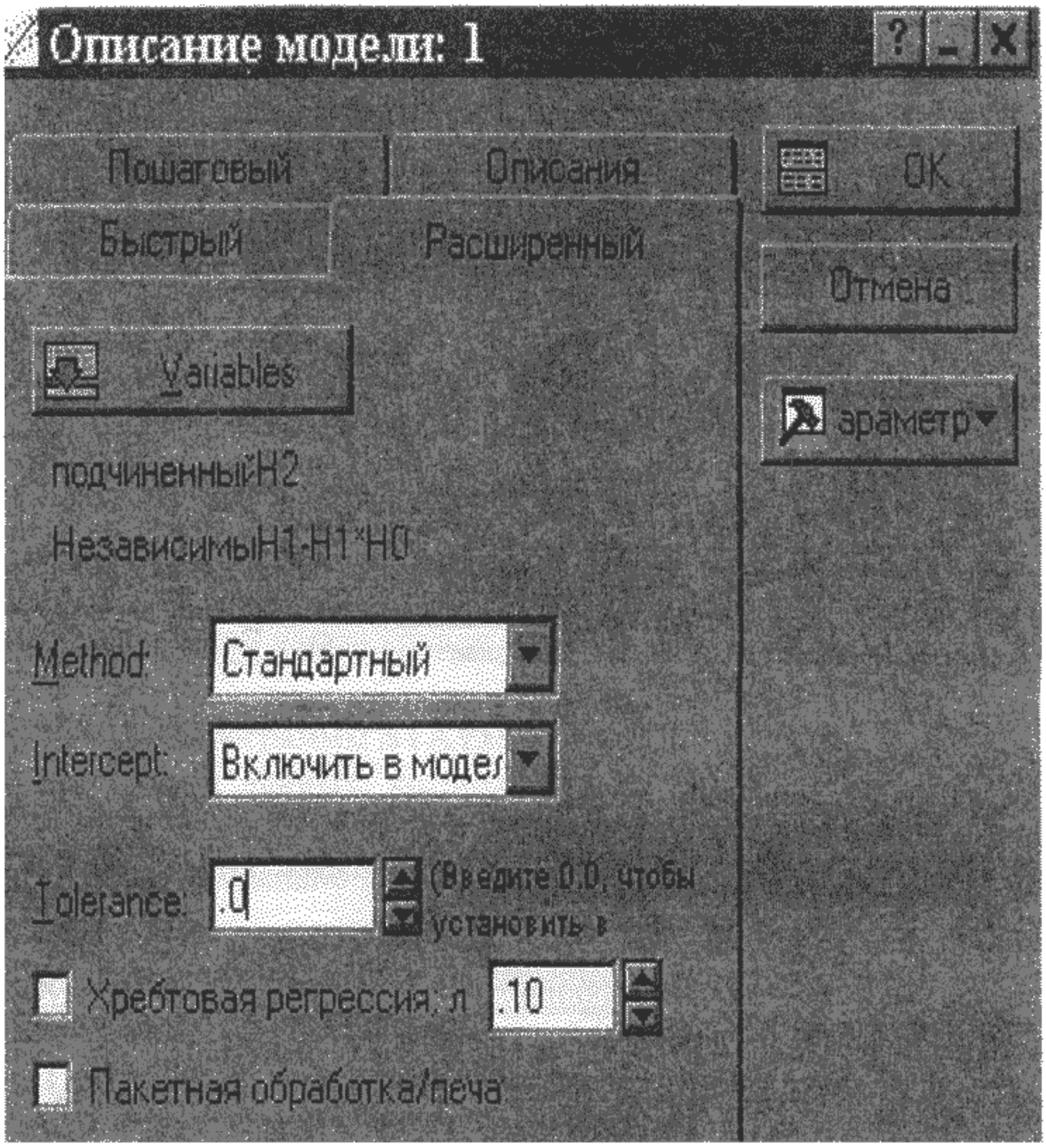


Рис.4.20

В вызванном окне **Описание модели:1** (рис. 4.20) в пункте **Tolerance** вводим значение **0.0.** Щелкаем на поле **Расширенный.** Результат на рис. 4.21

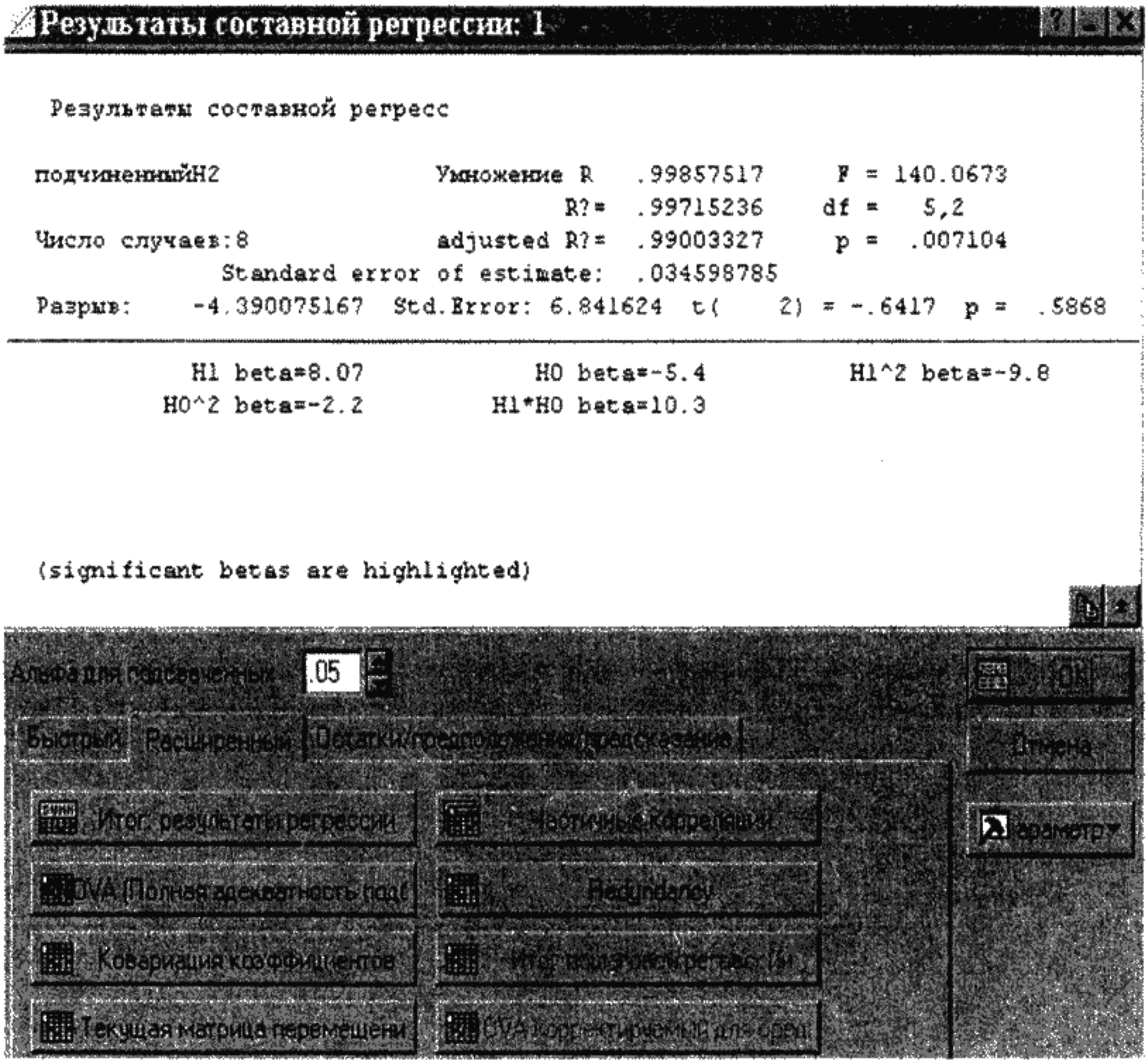


Рис. 4.21

Более подробно результаты можно посмотреть при просмотре полей типа **Итог: результаты регрессии** и т. п. формы **Результаты составной регрессии: 1** (рис. 4.21).

Результаты регрессии по данным, показанным на рис 4.17, приведены на рис.4.22 и рис.4.23.

Более подробно результаты можно посмотреть при просмотре полей типа **Итог: результаты регрессии** и т. п. формы **Результаты составной регрессии: 1** (рис. 4.21).

Результаты регрессии по данным, показанным на рис 4.17, приведены на рис.4.22 и рис.4.23.

Regression Summary for Dependent Variable: H2 (1)
R= .99857517 R?= .99715236 Adjusted R?= 99003327
F(5,2)=140.07 p<.00710 Std.Error of estimate: .03460

N=8	**Бета**	Std.Err. of Beta	B	Std.Err. of B	t(2)	p-level
ОТРЕЗОК			-4.39008	6.84162	-0.641671	0.586812
H1	8.06662	13.52592	10.95084	18.36211	0.596383	0.611432
H0	-5.41218	12.59635	-8.19224	19.06667	-0.429663	0.709303
H1^2	-9.75384	23.55304	-5.08817	12.28663	-0.414122	0.718972
H0^2	-2.21139	10.18052	-2.94146	13.54151	-0.217218	0.848184
H1*H0	10.30160	32.70725	8.07063	25.62398	0.314964	0.782613

Рис.4.22

Дисперсионный анализ; DV: H2 (1)

Эффект	Sums of Squares	df	Средст Squares	F	**p-level**
Regress.	0.838356	5	0.167671	140.0673	0.007104
Резидент	0.002394	2	0.001197		
Итог	0.840750				

Рис. 4.23

Сравните с результатом на рис. 4.9

4.5. Активизация надстройки "Анализ Данных" /8/

Для подключения пакета анализа нужно войти в пункт корневого меню Excel **Сервис** (рис.4.24) и, выбрав подпункт **Надстройки,** вызываем одноименное окно диалога (рис.4.25). В нем необходимо включить **Пакет анализа,** щелчком мыши установив маркер его активности.

Если в списке надстроек отсутствует данный пункт (Пакет анализа), необходимо запустить программу установки Microsoft Office, где выбрав продолжение **Добавить/Удалить** получаем доступ к **Сопровождению**, в котором, **выделив Microsoft Excel,** нажимаем кнопку **Состав.** В появившемся окне выбираем **Надстройки** и, нажатием на **Состав,** получаем возможность активизировать **Пакет анализа**. По завершению работы программы установки появляется возможность включить надстройку. (Более подробно об активизации надстройки **"Анализ Данных"** написано в /8/).

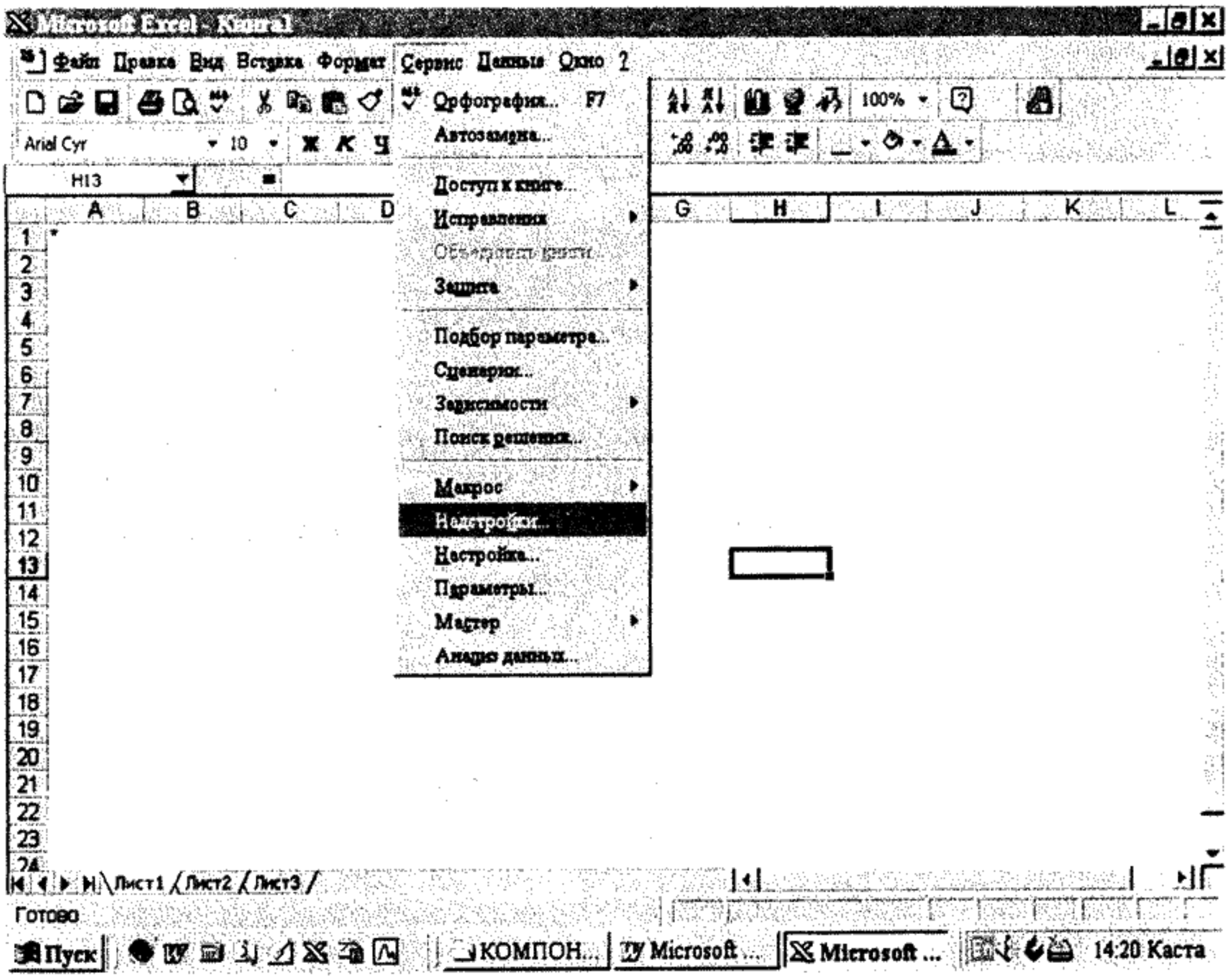


Рис. 4.24. В меню “Сервис” выбираем пункт “Надстройки”

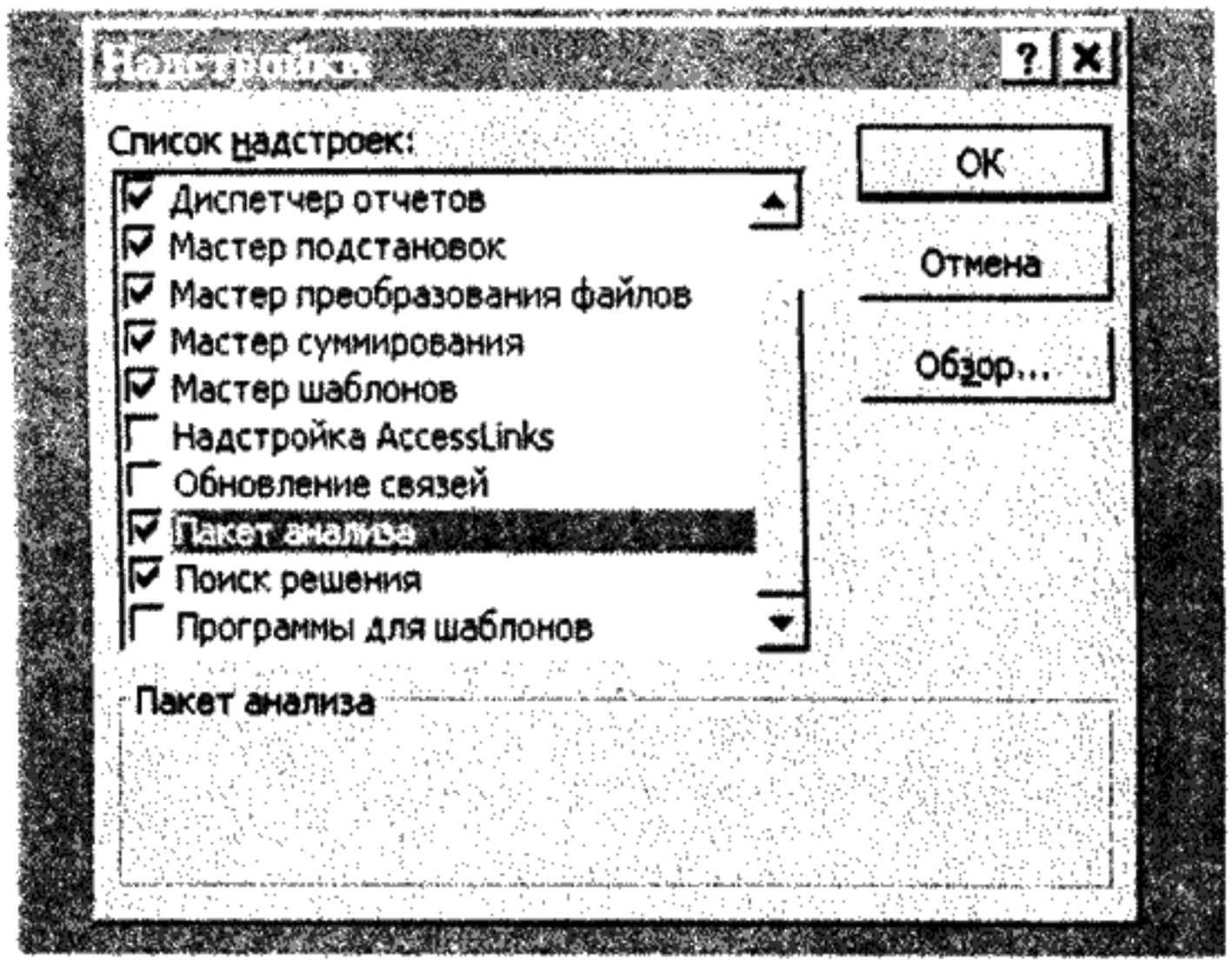


Рис 4.25. После активизации пункта меню “Пакет анализа” диалог завершается щелчком по ОК

4.6. Проверка данных на соответствие нормальному распределению

В /17/ показан один из способов проверки, является ли закон распределения выборки нормальным. Предлагается набор функций, которые позволяют при их вызове получить ответ в виде надписи **NORM** или **NO_NORM**. Ниже приводится текст, который необходимо записать в рабочую книгу. Для этого нужно открыть меню **Сервис** пункт **Макрос** и вызвать **Редактор Visual Basic**. Откроется окно **Microsoft Visual Basic**. В меню этого окна нужно выбрать последовательно **Insert**, **Macro**, **Module**, откроется лист **Module**, где следует набрать следующий текст:

```
Option Base 1
'Функция вычисления дисперсии выборки
Function VAR_1(R_1 As Object) As Double
'Функция возвращает значение дисперсии выборки
'R_1 – массив (анализируемая выборка)
 Dim mas1() As Double
 num_elem = R_1.Count 'вычисление количества элементов в массиве
'Инициализация временных переменных
 Mean_Tmp = 0#
 Var_Tmp = 0#
'Вычисление среднего значения выборки
 For i = 1 To num_elem
 Mean_Tmp = Mean_Tmp + R_1.Cells(i)
 Next i
```

```
 Mean1 = Mean_Tmp / num_elem
'Вычисление стандартного отклонения
 For i = 1 To num_elem
 Var_Tmp = Var_Tmp + (Mean1 – R_1.Cells(i)) ^ 2
 Next i
'Вычисление дисперсии выборки
 VAR_1 = (Var_Tmp / (num_elem – 1))
End Function
Function STD_2(R_1 As Object) As Double
'Функция возвращает значение стандартного отклонения выборки
'R_1 – массив (анализируемая выборка)
'вычисление среднеквадратичного отклонения
 STD_2 = VAR_1(R_1) ^ (1 / 2)
End Function
Function NORMSAMP_1(R_1 As Object) As String
'R_1 – массив (анализируемая выборка)
 num_elem = R_1.Count 'вычисление количества элементов в массиве
 Mean_Tmp = 0#
 Abs_Tmp = 0#
'вычисление среднего
 For i = 1 To num_elem
 Mean_Tmp = Mean_Tmp + R_1.Cells(i)
 Next i
 Mean1 = Mean_Tmp / num_elem
'Вычисление абсолютного среднего отклонения
 For i = 1 To num_elem
 Abs_Tmp = Abs_Tmp + Abs(R_1.Cells(i) – Mean1)
 Next i
 Abs_1 = Abs_Tmp / num_elem
 S_1 = STD_2(R_1)
 x_1 = Abs(Abs_1 / S_1 – 0.7979)
 y_1 = 0.4 / (num_elem ^ 1 / 2)
 If x_1 < y_1 Then NORMSAMP_1 = "NORM"
 If y_1 <= x_1 Then NORMSAMP_1 = "NO_NORM"
End Function
```

После этого можно в любом месте книги вызывать функцию NORM-SAMP_1(), параметрами которой передаются интервалы ячеек, где находятся данные. На рис. 4.26 показан пример проверки, является ли закон распределения нормальным.

Для практического использования проверку данных на соответствие нормальному распределению подготовила Рожкова Н.Ю.(кафедра информатики Иркутского института усовершенствования врачей).

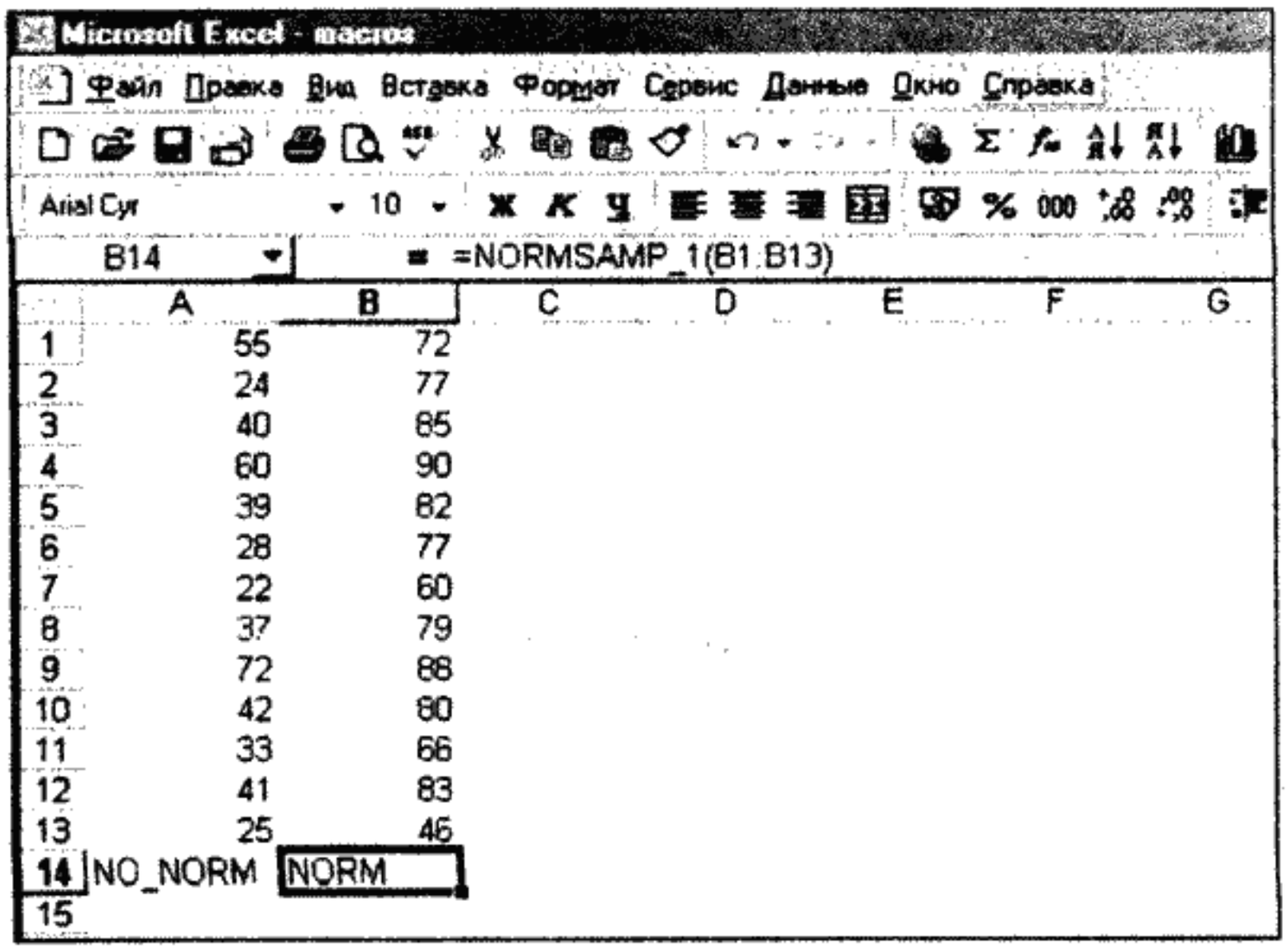

Microsoft Excel - макрос
Файл Правка Вид Вставка Формат Сервис Данные Окно Справка
Arial Cyr
10
B14
=NORMSAMP_1(B1:B13)
A B C D E F G
1 55 72
2 24 77
3 40 85
4 60 90
5 39 82
6 28 77
7 22 60
8 37 79
9 72 88
10 42 80
11 33 66
12 41 83
13 25 46
14 NO_NORM NORM
15

Рис. 4.26

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Михалевич И.М., Примина С.П. Применение математических методов при анализе геологической информации (с использованием Excel): Учеб. пособие. Ч. 1. – Иркутск: Иркут. ун -т, 2001. – 60 с.
2. Крамбейн У., Грейбилл Ф. Статистические модели в геологии. – М.: Мир, 1969 . – 397 с.
3. Гланц С. Медико-биологическая статистика. – М.: Практика, 1999. – 459 с.
4. Арабаджи М.С., Бакиров Э.А., Мильничук В.С., Сенюков Р.В. Математические методы и ЭВМ в поисково-разведочных работах. – М.: Недра, 1984. – 264 с.
5. Сергиенко В.И., Бондарева И.Б. Математическая статистика в клинических исследованиях. – М.: ГЕОТАР–МЕД, 2001. – 256 с.
6. Девис Дж.С. Статистический анализ данных в геологии . – М.: Недра, 1990; Т.1. – 319с. ; Т.2. – 427 с.
7. Кулаичев А.П. Методы и средства анализа данных в среде Windows и STADIA. Т. 1. Изд. 3-е, перераб. и доп. – М: Информатика и компьютеры, 1999. – 341 с.
8. Скрипченко Н.А. Анализ данных в MICROSOFT EXCEL. – М.: Изд-во ИГТУ, 1998. – 60 с.
9. Дюк В. Обработка данных на ПК в примерах. – М.: Питер Паблишинг, 1997. – 231 с.
10. Закс Л. Статистическое оценивание. – М.: Статистика, 1976 . – 598 с.
11. Алгоритмы и программы восстановления зависимостей/ Под ред. Вапника В.Н. – М.: Наука, 1984. – 816 с.
12. Драйпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Статистика , 1973. – 392 с.
13. Дементьев Л.Ф., Жданов М.А., Кирсанов А.Н. Применение математической статистики в нефтегазопромысловой геологии. – М.:Недра, 1977. – 255 с.
14. Чини Р.Ф. Статистические методы в геологии/ Пер. с англ. – М.: Мир, 1986. – 189 с.
15. Алферова М.А., Михалевич И.М., Рожкова Н.Ю., Сыклен А.Е. Примеры практической работы с Excel: Учеб.-метод. пособие. Вып. 2. – Изд. 3-е. – Иркутск: ИГИУВ, 2003 . – 41 с.
16. Миллер Р., Кан Дж. Статистический анализ в геологических науках. – М.: Мир,1965. – 482 с.

17.Лавач С.М., Чубенко А.В., Бабич П.М. Статистические методы в медико-биологических исследованиях с использование Excel. – Киев: Морион, 2000. – 320 с.

18.Юнкеров В.И., Григорьев С.Г. Математико-статистические методы обработки данных медицинских исследований. – СПб.: ВМедА, 2002. – 266 с.

19.Боровиков В. STATISTIKA: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Питер, 2001. – 656 с.

20.Примина С.П., Михалевич И.М., Шипунова И.Б., Лузин В.Ф. Компьютерная обработка данных нефтяной геологии (на примере построения структурной карты): Метод.указания. – Иркутск: Изд. Иркут. ун-т, 2001. – 15 с.

ОГЛАВЛЕНИЕ

И.М. Михалевич
С.П. Примина

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКИХ МЕТОДОВ ПРИ АНАЛИЗЕ ГЕОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

(с использованием компьютерных технологий)

Учебное пособие
Часть П

Редактор Э.А. Невзорова

Темплан 2004 . Поз. 1.
Подписано в печать 9.04.04. Формат 60x90 1/16 .
Печать трафаретная. Усл.-печ. л. 7,5. Уч.-изд. л. 7,0. Тираж 100 экз.

Редакционно-издательский отдел
Иркутского государственного университета
664003, г. Иркутск, бульвар Гагарина, 36