Министерство образования и науки Украины
Государственное учреждение
„Луганский национальный университет
имени Тараса Шевченко”

Л. Ф. Панченко

ПРАКТИКУМ ПО АНАЛИЗУ ДАННЫХ

Учебное пособие
для студентов высших учебных заведений

Луганск
ГУ „ЛНУ имени Тараса Шевченко”
2013

УДК 51-7:004.04(075.8)
ББК 32.973.26-018.2в6я73
П16

Р е ц е н з е н т ы :

Адаменко Е. В. – доктор педагогических наук, профессор, декан факультета допрофессиональной подготовки ГУ „Луганский национальный университет имени Тараса Шевченко”

Дымарский Я. М. – доктор физико-математических наук, професор, зав. каф. інформатики та інформаційних технологій ОВС Луганського державного університету внутрішніх справ імені Е.О.Дідоренка

Меняйленко А. С. – доктор технических наук, проректор, профессор кафедры информационных технологий и систем ГУ „Луганский национальный университет имени Тараса Шевченко”.

Панченко Л. Ф.

П16 Практикум по анализу данных : учебное пособие для студентов высших учебных заведений / Л. Ф. Панченко ; Гос. учрежд. „Луган. нац. ун-т имени Тараса Шевченко”. – Луганск : Изд-во ГУ „ЛНУ имени Тараса Шевченко”, 2013. – 269 с.

В учебном пособии рассматриваются теоретические и практические вопросы компьютерного анализа данных с использованием электронных таблиц MS Excel, статистического пакета SPSS, свободно распростарняемой среды R, включающие такие разделы анализа как описательная статистика, основы корреляционного и регрессионного анализа, проверки гипотез, дисперсионного анализа, методов многомерного анализа.

Учебное пособие предназначено для студентов и магистрантов специальностей „Информатика”, а также всех других специальностей, которые связаны с анализом данных на компьютере.

УДК 51-7:004.04(075.8)
ББК 32.973.26-018.2в6я73

Принято к печати Учебно-методическим советом Луганского национального университета имени Тараса Шевченко (протокол №10 от 05 июня 2013 года)

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Учебное пособие построено в соответствии с программой курса «Анализ данных» для студентов специальности «Информатика» и рекомендациями по преподаванию программной инженерии и информатики в университетах [46].

Пособие имеет следующую структуру: краткое изложение наиболее важных теоретических сведений по теме, практические работы, список рекомендованной литературы. Каждая практическая работа содержит подробные инструкции по решению типичной задачи, задачи для самостоятельного решения, требования к отчету и контрольные вопросы.

Для решения задач используются электронные таблицы Microsoft Excel, статистический пакет SPSS для Windows, среда R. R – эффективный инструмент разработки новых методов интерактивного анализа данных. Он свободно распространяется, быстро развивается и расширяется большой коллекцией пакетов, реализующих новейшие методы анализа данных.

Изложение материала сопровождается иллюстрациями – схемами, таблицами, изображением диалоговых окон – с тем, чтобы работающему с пособием было легче понять и освоить объективно нелегкие математические методы обработки данных с использованием компьютера.

Учебное пособие поможет студентам и магистрантам, которые будут использовать компьютерный анализ данных в своей профессиональной деятельности, освоить основные средства такого анализа с использованием Microsoft Excel, SPSS для Windows, R: описательную статистику, основы корреляционного анализа, линейной регрессии, проверку статистических гипотез, основы многомерного анализа.

Пособие может быть полезно не только студентам и магистрантам специальности «Информатика», а и всем тем, кто стремится овладеть математическими методами обработки данных с использованием компьютера: психологам, социологам, педагогам.

1. ЧАСТОТНОЕ РАСПРЕДЕЛЕНИЕ ДАННЫХ. МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ И МЕРЫ ИЗМЕНЧИВОСТИ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА В MS EXCEL, SPSS, R

Признаки и переменные – это измеряемые индикаторы тех объектов, процессов и явлений, которые изучаются в ходе маркетинговых, социологических, политологических, психолого-педагогических исследований, исследований в медицине, биологии, сельском хозяйстве и т.п.

Значения признака, полученные в результате осуществленных в процессе конкретного исследования измерений, называются ***наблюдениями, наблюдаемыми значениями, вариантами***.

Абсолютная частота – это число объектов, обладающих данным значением признака.

Относительная частота – это доля объектов, обладающих данным значением признака, в общем объеме выборки. **Относительная частота** получается путем деления абсолютной частоты на общее число измерений.

Относительные частоты измеряются в долях единицы. Сумма относительных частот для данного ряда измерений всегда равна единице.

Процентная частота – это процент объектов, обладающих данным значением признака, в общем объеме выборки**;** это относительная частота, выраженная в процентах.

Сумма процентных частот для данного ряда измерений всегда равна ста процентам.

Накопленная (кумулятивная) частота – это «накопленная встречаемость»; для шкал, выше ранговых, – это сумма частот значений, не превосходящих данное значение.

Табулирование данных – это сведение их в таблицу.

Основные требования к таблице:

- наличие четкого и ясного заголовка;

- наличие четких и ясных подписей колонок (столбцов) и строк.

Частотное распределение – упорядоченный подсчет количества признаков по каждому значению переменной. **Частотное распределение** показывает, сколько раз каждое значение переменной встречается в совокупности наблюдений.

Таблица частотного распределения – один из самых простых способов представления собранных данных. Она состоит, по крайней мере, из двух столбцов: левый содержит значения, которые может принимать переменная, а правый – абсолютную частоту (число раз, которое каждое значение встречается). Дополнительные столбцы, отражающие относительное и (или) процентное частотное распределение.

График частотного распределения – чаще всего это столбиковая диаграмма (гистограмма). Данные на гистограмме представляются в виде параллельных прямоугольников (столбиков) одинаковой ширины. Каждый столбик показывает один класс качественных данных (одну варианту). Высота столбика пропорциональна величине или частоте измеряемого параметра.

Мода – наиболее часто встречающееся значение в ряду измерений.

Медиана – серединное значение в упорядоченном ряду измерений.

Среднее (среднее арифметическое значение) – сумма всех значений в ряду измерений, разделенная на их общее количество.

Размах – разность между наибольшим и наименьшим наблюденными значениями в ряду измерений.

Дисперсия – отношение суммы квадратов отклонений от среднего арифметического к величине (n - 1), где n – общее число измерений.

Стандартное отклонение – положительное значение квадратного корня из дисперсии; измеряет "средний" разброс значений переменной относительно ее среднего арифметического в тех же единицах измерения, что и сама переменная.

Практическая работа 1.1

Построение частотного распределения данных и построение гистограммы с помощью MS Excel

Цель работы: научиться использовать возможности электронных таблиц MS Excel для расчета относительных, процентных, накопленных частот, строить графики частотного распределения.

Постановка задачи

В таблице 1.1.1 представлены данные о распределении пользователей сети Facebook по континентам.

Таблица 1.1.1

Исходные данные

Континент	Число пользователей Facebook (чел)
1. Северная Америка	186126740
2. Европа	169718660
3. Азия	117151400
4. Южная Америка	52870200
5. Африка	19649500
6. Австралия и Океания	11656460

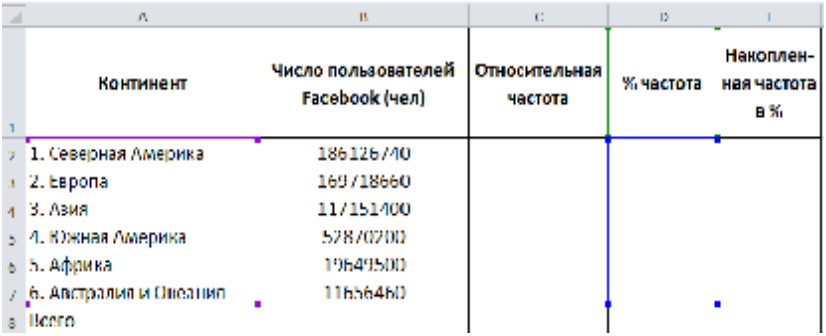
Вычислить с помощью MS Excel относительные, процентные и накопленные частоты и, используя, графические возможности электронных таблиц, представить графически частотное распределение пользователей сети Facebook по континентам.

Ход работы

1. Запустить Microsoft Excel.
2. Ввести данные в соответствии с табл. 1.1.2.

Таблица 1.1.2

Электронная таблица с исходными данными



	A	B	C	D	E
1	Континент	Число пользователей Facebook (чел)	Относительная частота	% частота	Накопленная частота в %
2	1. Северная Америка	186126740			
3	2. Европа	169718660			
4	3. Азия	117151400			
5	4. Южная Америка	52870200			
6	5. Африка	19649500			
7	6. Австралия и Океания	11656460			
8	Всего				

3. В клетке B8 подсчитать общее число пользователей по формуле суммы.

4. В клетках C2:C7, D2:D7, E2:E7 записать формулы для расчета относительных, процентных и накопленных частот (см. табл. 1.1.3).

Таблица 1.1.3

Электронная таблица с формулами

	A	B	C	D	E
1	Континент	Число пользователей Facebook (чел)	Относительная частота	% частота	Накопленная частота в %
2	1. Северная Америка	186126740	=B2/$B$8	=C2*100	=D2
3	2. Европа	169718660	=B3/$B$8	=C3*100	=E2+D3
4	3. Азия	117151400	=B4/$B$8	=C4*100	=E3+D4
5	4. Южная Америка	52870200	=B5/$B$8	=C5*100	=E4+D5
6	5. Африка	19649500	=B6/$B$8	=C6*100	=E5+D6
7	6. Австралия и Океания	11656460	=B7/$B$8	=C7*100	=E6+D7
8	Всего	=СУММ(B2:B7)	=СУММ(C2:C7)	=СУММ(D2:D7)	
9					

5. В ячейках C8 и D8 подсчитать с помощью формул суммы соответствующих частот. В результате должна получиться таблица 1.1.4.

6. Построить гистограмму (столбиковую диаграмму). Для этого, удерживая клавишу «*Ctrl*», выделить два несмежных блока данных A1:A7 и D1:D7 и выбрать *Вставка, Гистограмма*. В меню *Макет* добавить к построенной диаграмме заголовок, подписи горизонтальной и вертикальной осей, подписи данных. Должна получиться гистограмма, представленная на рис.1.1.1.

Таблица 1.1.4

Электронная таблица с результатами расчетов

	A	B	C	D	E
1	Континент	Число пользователей Facebook (чел)	Относительная частота	% частота	Накопленная частота в %
2	1. Северная Америка	186126710	0,33	33,4	33,4
3	2. Европа	169718660	0,30	30,5	63,9
4	3. Азия	117151400	0,21	21,0	84,9
5	4. Южная Америка	52870200	0,09	9,5	94,4
6	5. Африка	19649500	0,04	3,5	97,9
7	6. Австралия и Океания	11656460	0,02	2,1	100,0
8	Всего	557172960	1	100	

Рис. 1.1.1. Распределение пользователей Facebook по континентам

7. Постройте также полигон и кривую накопленных частот (см. рис. 1.1.2)

Рис. 1.1.2. Полигон и кривая накопленных частот

Задания для самостоятельного выполнения

1. В таблице 1.1.5 представлены данные о возрасте и численности украинских пользователей Facebook. Расчитайте процентную и накопленную частоту и постройте гистограмму, полигон и кривую накопленных частот для этих данных.

Таблица 1.1.5

Возраст пользователей Facebook в Украине

Возраст	Число пользователей (в тыс.чел.)	% пользователей	Накопленный %
13-17	159,60		
18-24	617,02		
25-34	836,92		
35-44	400,90		
45-54	188,52		
55-100	113,92		

2.Одной из первых систем с использованием естественного языка была система LUNAR, которая давала ответы на вопросы геологов о камнях, привезенных с Луны в ходе полета космического аппарата Апполон-11. На второй ежегодной лунной научной конференции в 1971 году системе LUNAR предъявили 111 запросов на естественном языке. Информация о них приведена в таблице 1.1.6. Рассчитайте относительные, процентные частоты и представьте результаты в виде графика. Прокомментируйте результаты.

Таблица 1.1.6

Обработка запросов на естественном языке в системе LUNAR

Запросы	Абсолютная частота	Относительная частота	Процентная частота
Невозможно разобрать и интерпретировать	11		

Запрос, оказавшийся некорректным вследствие опечаток	13		
Обработанные удовлетворительно	87		
Всего	111		

3. Пятнадцать профессиональных программистов предоставили коллегам для рецензирования свою лучшую программу. Каждый получил по четыре программы от коллег и оценивал их по шкале от 1 балла до 7. Результаты приведены в таблице 1.1.7.

Таблица 1.1.7

Совместные результаты трех общих исследований по рецензированию равными по рангу с участием профессиональных программистов

Разность между наивысшей и наинизшей из четырех оценок программы, усредненная по 15 программам	Число случаев	% случаев	Накопленный %
0 баллов	2		
1 балл	7		
2 балла	17		
3 балла	15		
4 балла	9		
5 баллов	7		
6 баллов	3		

Рассчитайте относительные, процентные частоты и представьте результаты в виде графика. Прокомментируйте результаты.

Требования к отчету:

Отчет должен содержать:

- ответы на контрольные вопросы;
- файлы с результатами расчетов.

Контрольные вопросы

1. Что такое абсолютная частота?
2. Что такое относительная частота? Как она вычисляется?
3. Что такое процентная частота? Как она вычисляется?
4. Что такое накопленная частота? Как она вычисляется?
5. Что показывает частотное распределение признака?
6. Охарактеризуйте гистограмму, полигон, кривую накопленных частот. Как они строятся и как редактируются?
7. Как выделить два несмежных блока в электронной таблице MS Excel?
8. Как построить диаграмму в MS Excel.
9. Как добавить необходимые подписи к диаграмме?
10. Дайте содержательную интерпретацию результатам расчетов в данной работе.

Практическая работа 1.2
Вычисление среднего и стандартного отклонения. Графическая интерпретация стандартного отклонения

***Цель работы**:* научиться использовать возможности электронных таблиц MS Excel для расчета мер центральной тенденции и мер изменчивости, графической интерпретации стандартного отклонения.

Постановка задачи

Двадцати респондентам было предложено дать оценку по 10-ти балльной шкале трем сортам мороженого. В таблице 1.2.1 приведены оценки, которые дали респонденты.

Таблица 1.2.1

Оценка респондентами трех сортов мороженого

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Оценка респондентами трех сортов мороженого																				
2		Номер респондента																			
3	Сорт мороженого	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
4	Пломбир	8	10	9	8	8	9	10	10	9	10	9	10	9	10	10	9	8	10	9	10
5	Сливочное	6	7	10	10	9	7	10	10	9	7	10	9	8	10	6	5	10	10	9	10
6	Фруктовое	6	8	10	3	7	4	10	9	4	5	7	4	7	10	9	10	6	5	7	10

Определить среднюю оценку респондентами каждого вида мороженого, а также стандартное отклонение для каждого вида. Построить графики, иллюстрирующие разброс оценок для каждого сорта мороженого.

Ход работы

1. Запустить Microsoft Excel.
2. Ввести данные в соответствии с табл. 1.2.1 (чтобы сделать маленькие клеточки, нужно выделить соответствующие столбцы, а затем один из столбцов уменьшить).
3. В клетке V3 написать слово «среднее», W3 – «станд. откл.», в клетках V4:V6 написать формулы для расчета среднего числа решенных задач (функция СРЗНАЧ), в клетках W4:W6 – формулы для расчета стандартного отклонения

(функция СТАНДОТКЛОН). На рис. 1.2.1 представлен фрагмент таблицы с формулами, а на рис. 1.2.2 – с результатами.

	V	W
3	среднее	станд. откл.
4	=СРЗНАЧ(B4:U4)	=СТАНДОТКЛОН(B4:U4)
5	=СРЗНАЧ(B5:U5)	=СТАНДОТКЛОН(B5:U5)
6	=СРЗНАЧ(B6:U6)	=СТАНДОТКЛОН(B6:U6)

Рис. 1.2.1. Фрагмент таблицы с формулами

	V	W
3	среднее	станд. откл.
4	9,25	0,786397516
5	8,6	1,667017507
6	7,05	2,372540278

Рис. 1.2.2. Фрагмент таблицы с результатами

4. Проанализировать полученные данные.

5. Построить точечную диаграмму для каждой группы данных и сравнить их.

Для этого выделить данные первой группы B4:U4 и нажать на кнопку *Мастер диаграмм* на панели инструментов. В первом окне мастера выбрать тип диаграммы («точечная»), в последующих окнах задать заголовки диаграммы и ее размещение (на текущем листе). Повторить те же действия для данных второй и третьей группы.

Должны получиться три диаграммы, подобные приведенной рис. 1.2.3. Необходимо расположить их рядом на том же листе, что и данные.

Рис. 1.2.3. Графическое представление оценок пломбира

6. Сохранить результаты в личной папке.

Требования к отчету:

Отчет должен содержать:

– ответы на контрольные вопросы;
– файлы с результатами расчетов.

Контрольные вопросы

1. Дайте определения известным вам мерам центральной тенденции.
2. Дайте определения известным вам мерам вариации.
3. Назовите функции электронных таблиц для вычисления мер центральной тенденции.
4. Назовите функции электронных таблиц для вычисления мер изменчивости.
5. В чем состоит графическая интерпретация стандартного отклонения? Дайте содержательную интерпретацию результатам расчетов в данной работе.

Практическая работа 1.3
Вычисление мер центральной тенденции и мер изменчивости в R

Ход работы

1. Запустить R (подробнее об R см. [42-46]).
2. Определить три вектора, представляющие оценки респондентами трех сортов мороженого:

```
> plomb <- c(8,10,9,8,8,9,10,10,9,10,9,10,9,10,10,9,8,10,9,10)
> sliv  <- c(6,7,10,10,9,7,10,10,9,7,10,9,8,10,6,5,10,10,9,10)
> fruct <- c(6,8,10,3,7,4,10,9,4,5,7,4,7,10,9,10,6,5,7,10)
```

3. Подсчитать средние значения оценок трех сортов мороженого:

```
> mean(plomb); mean(sliv); mean(fruct)
[1] 9.25
[1] 8.6
[1] 7.05
```

```
> median(plomb); median(sliv); median(fruct)
[1] 9
[1] 9
[1] 7
```

```
> summary(plomb)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
   8.00    9.00    9.00    9.25   10.00   10.00
```

4. Подсчитать стандартные отклонения оценок трех сортов мороженого:

```
> sd(plomb);sd(sliv);sd(fruct)
```

```
[1] 0.7863975
[1] 1.667018
[1] 2.372540
```

5. Построить три графика (см. рис. 1.3.1–1.3.2), иллюстрирующие разброс оценок респондентов (поскольку одновременно по умолчанию чертится один график, копируйте их в текстовый редактор).

```
> plot(plomb)
> title(main="Пломбир")
> plot(sliv)
> title(main="Сливочное мороженое")
> plot(fruct)
> title(main="Фруктовое мороженое")
```

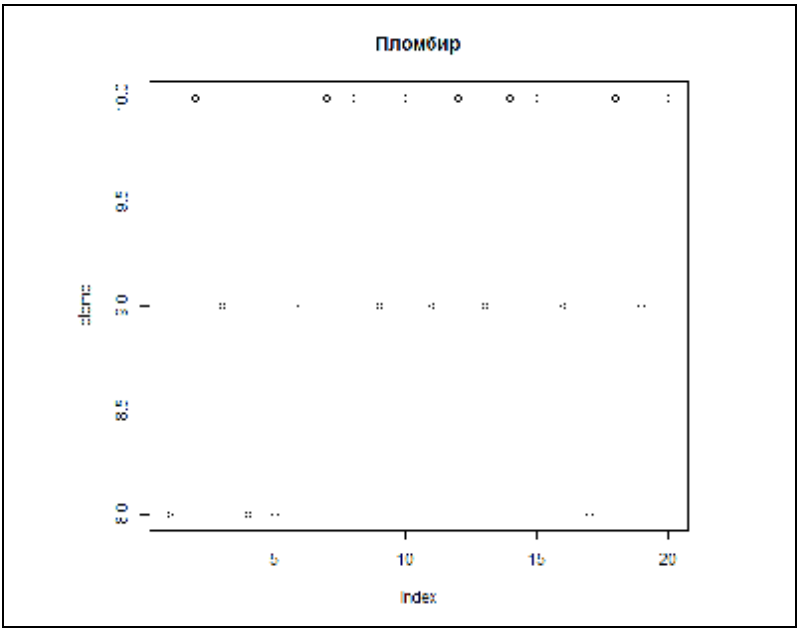


Рис. 1.3.1. Оценка респондентами пломбира

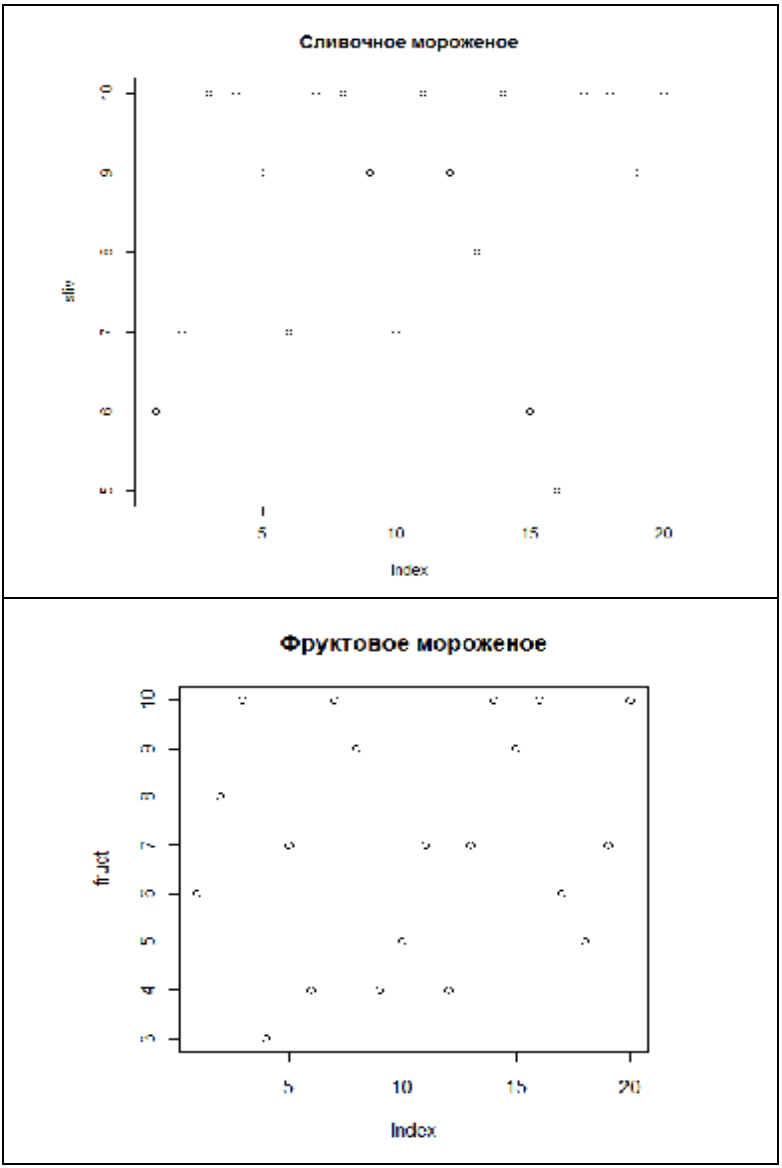


Рис. 1.3.2. Графическое представление оценок сортов мороженого

6. Вызвать справку для команд plot и title, перевести и изучить ее.

```
>help(title)
```

Описание

This function can be used to add labels to a plot. Its first four principal arguments can also be used as arguments in most high-level plotting functions. They must be of type character or expression. In the latter case, quite a bit of mathematical notation is available such as sub- and superscripts, greek letters, fractions, etc: see plotmath

Использование

```
title(main = NULL, sub = NULL, xlab = NULL, ylab = NULL,
    line = NA, outer = FALSE, ...)
```

Аргументы

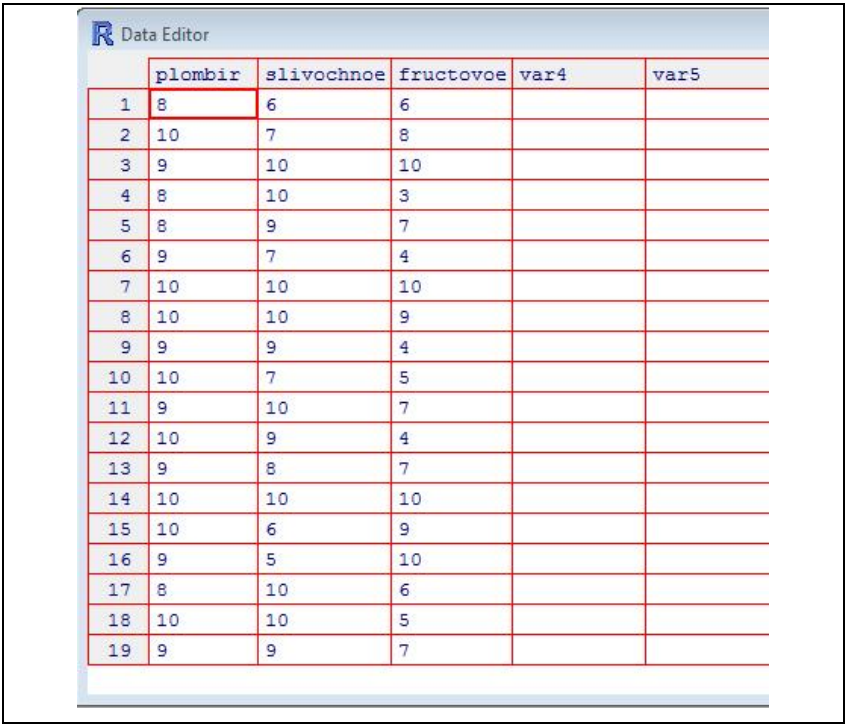
main	The main title (on top) using font and size (character expansion) par("font.main") and color par("col.main").
sub	Sub-title (at bottom) using font and size par("font.sub") and color par("col.sub").
xlab	X axis label using font and character expansion par("font.lab") and color par("col.lab").
ylab	Y axis label, same font attributes as xlab.
line	specifying a value for line overrides the default placement of labels, and places them this many lines outwards from the plot edge.
outer	a logical value. If TRUE, the titles are placed in the outer margins of the plot.

7. Объединить все переменные в одну таблицу данных и просмотреть ее содержимое

```
d <- data.frame(plombir=plomb, slivochnoe=sliv, fructovoe=fruct)
> d
   plombir slivochnoe fructovoe
1        8         6         6
2       10         7         8
3        9        10        10
4        8        10         3
5        8         9         7
6        9         7         4
7       10        10        10
8       10        10         9
9        9         9         4
10      10         7         5
11       9        10         7
12      10         9         4
13       9         8         7
14      10        10        10
15      10         6         9
16       9         5        10
17       8        10         6
18      10        10         5
19       9         9         7
20      10        10        10
```

8. Просмотреть созданную таблицу во встроенном редакторе (см. рис.1.3.3):

```
> fix(d)
```



	plombir	slivochnoe	fructovoe	var4	var5
1	8	6	6		
2	10	7	8		
3	9	10	10		
4	8	10	3		
5	8	9	7		
6	9	7	4		
7	10	10	10		
8	10	10	9		
9	9	9	4		
10	10	7	5		
11	9	10	7		
12	10	9	4		
13	9	8	7		
14	10	10	10		
15	10	6	9		
16	9	5	10		
17	8	10	6		
18	10	10	5		
19	9	9	7		

Рис. 1.3.3. Просмотр созданной таблицы в редакторе данных

9. Нарисовать график для пломбира с подписями осей и ломаной красного цвета (рис.1.3.4):

```
>plot(d$plombir, xlab="номер респондента", ylab="оценка", main="Пломбир", col="red", type="l")
```

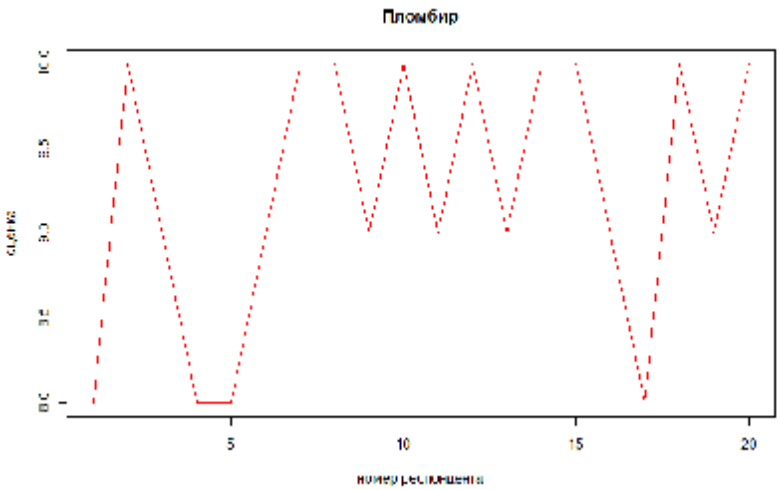


Рис.1.3.4. График с подписями осей

10. Сохранить рабочее пространство: File, Save Workplace

Контрольные вопросы

1. Как определяется числовой вектор в языке R?
2. С помощью какой команды чертится график?
3. Как дать ему название?
4. Как объединить несколько векторов в одну таблицу?
5. Как вызвать встроенный редактор для просмотра или редактирования таблицы?
6. Для чего сохраняют рабочее пространство и каким образом это осуществляется?
7. Дайте содержательную интерпретацию результатам расчетов в данной работе.
8. Какие возможности R для описательной статистики мы изучили?
9. Пользуясь справкой, встроенной в R , заполните следующую таблицу:

Практическая работа 1.4
Построение и расчет параметров распределения для несгруппированных данных с помощью пакета анализа

Цель работы: научиться давать имена диапазонам, строить частотное распределение и гистограмму для несгруппированных данных с помощью средства «Гистограмма» из *Пакета анализа* данных, рассчитывать итоговые статистики с помощью средства «Описательная статистика».

Ход работы

1. Ввести в диапазон *B2:B76* столбца *B* следующие данные, представляющие собой оценки 75-ти взрослых людей в тесте на определение коэффициента интеллектуальности Стенфорда-Бине (табл. 1.3.1).

Таблица 1.4.1

Исходные данные

141	104	101	130	148
92	87	115	91	96
100	133	124	92	123
132	118	98	101	107
97	124	118	146	107
110	111	138	121	129
106	135	97	108	108
107	110	101	129	105
105	110	116	113	123
83	127	112	114	105
127	114	113	106	139
95	105	95	105	106
109	102	102	102	89
108	92	131	86	134

104	94	121	107	103

2. Дать интервалу с данными имя **IQ** (выделить интервал, выбрать меню ***Вставка***, ***Имя***, ***Присвоить***).

3. Построить гистограмму с помощью средства «**Анализ данных**»: меню ***Данные, Анализ данных***, ***Гистограмма***. В окне ***Гистограмма*** задать необходимые параметры (см. рис. 1.3.1). Указание: если в меню ***Данные***, нет пункта Анализ данных, зайти в ***Файл, Параметры, Надстройки, Перейти***, поставить флажок у опции *Пакет анализа*.

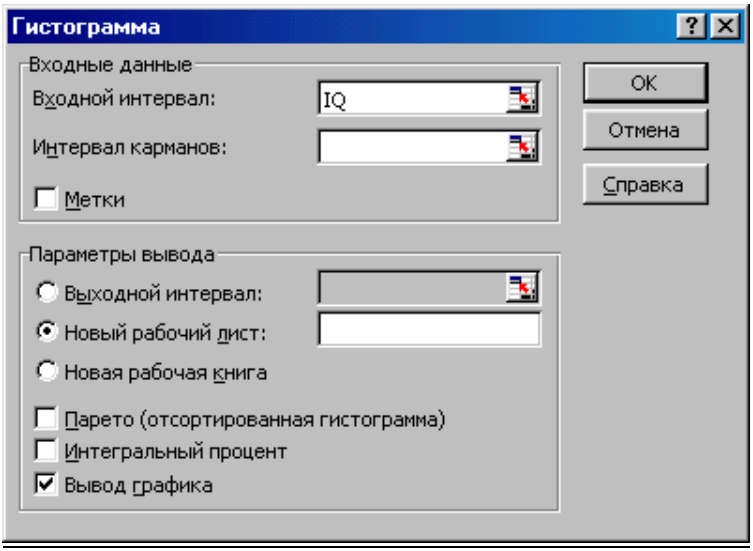


Рис. 1.4.1. Окно «Гистограмма»

Интервал карманов – это необязательный набор граничных значений, определяющих отрезки (карманы). Эти значения должны быть введены в возрастающем порядке. В Microsoft Excel вычисляется число попаданий данных между текущим началом отрезка и соседним, большим по порядку, если такой есть. При этом включаются значения на нижней границе отрезка и не включаются значения на верхней границе.

Если диапазон карманов не был введен, то автоматически создается набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных.

В результате на новом листе получится таблица (см. рис. 1.4.2) и график (рис. 1.4.3).

Карман	*Частота*
83,00	1
91,13	4
99,25	10
107,38	22
115,50	14
123,63	7
131,75	8
139,88	6
Еще	3

Рис. 1.4.2. Таблица вывода «Карманы и частота»

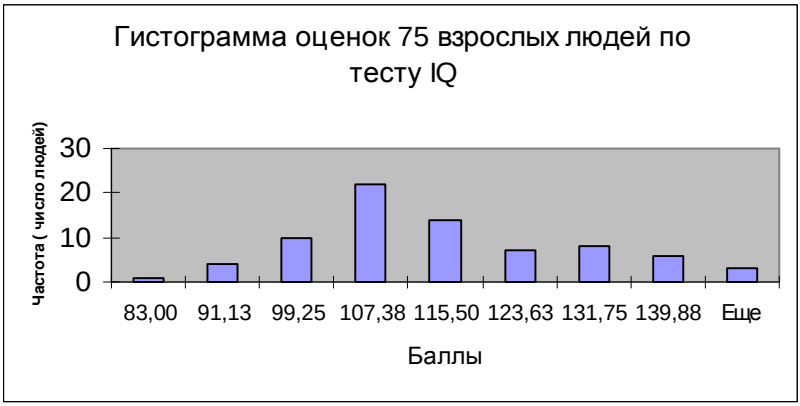


Рис. 1.4.3. Гистограмма оценок по тесту IQ

4. Рассчитать показатели описательной статистики для данных теста: меню ***Данные, Анализ данных, Описательная статистика***. В итоге получится таблица с комплексом показателей описательной статистики, включающим среднее, стандартную ошибку, медиану, моду и т.д.

5. Сохранить результаты в личной папке.

Требования к отчету:

Отчет должен содержать:

– ответы на контрольные вопросы;
– файлы с результатами расчетов.

Контрольные вопросы

1. Как дается имя диапазону данных? Для чего нужны такие имена?

2. Как построить гистограмму для несгруппированных данных?

3. Как быстро рассчитать показатели описательной статистики?

4. Что такое карман?

5. Можно ли рассчитать накопленные частоты и построить кривую накопленных частот, пользуясь *Пакетом анализа*?

6. Дайте содержательную интерпретацию результатам расчетов в данной работе.

7. Что означает каждый из показателей в таблице вывода процедуры «Описательная статистика»?

Практическая работа 1.5

Построение частотного распределения для несгруппированных данных в R

Ход работы

1. Запустите R
2. Определите таблицу данных, представляющую оценки IQ 75 респондентов.

141	104	101	130	148
92	87	115	91	96
100	133	124	92	123
132	118	98	101	107
97	124	118	146	107
110	111	138	121	129
106	135	97	108	108
107	110	101	129	105
105	110	116	113	123
83	127	112	114	105
127	114	113	106	139
95	105	95	105	106
109	102	102	102	89
108	92	131	86	134
104	94	121	107	103

Для этого введите эти данные в область A1:A75 листа Excel, затем выделите их и скопируйте в буфер обмена (в R используется десятичный разделитель «точка»). Чтобы создать в R таблицу данных из буфера обмена используйте команду:

```
iq <- data.frame(read.table("clipboard"))
```

3. Проверьте себя, введя команды:

```
>iq
```

Выведется содержимое таблицы iq.

```
>fix(iq)
```

Откроется встроенный редактор данных (см. рис. ниже), в котором можно вносить изменения или просматривать объект.

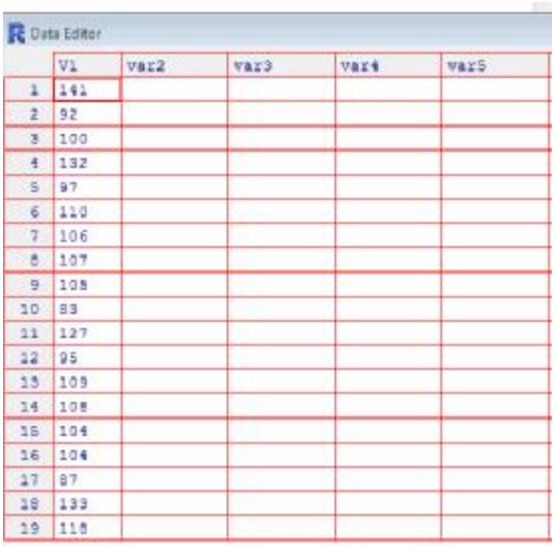


Рис.1.5.1.Встроенный редактор

4. Просмотрите структуру объекта iq

```
> str(iq)
>'data.frame':   75 obs. of  1 variable:
 $ V1: num  141 92 100 132 97 110 106 107 105 83 ...
```

5. Рассмотрим еще один способ обмена данными из Excel и R. Можно использовать функцию scan, чтобы скопировать колонку с числами из Excel в R. Выполните команду x <- scan(), введите Ctrl-v чтобы вставить в R, and нажмите enter чтобы обозначить конец ввода для команды scan. Переменная x будет содержать числа из Excel. Отметим, что scan работает с колонками, а не строками в Excel. Чтобы скопировать строку, ее надо предварительно транспонировать в колонку, а затем копировать уже эту колонку. Функция scan() не ограничивается работой с Excel. Ее можно использовать также, чтобы вставить колонку чисел из таких приложений как Notepad или Word.

6. Постройте гистограмму оценок 75 взрослых людей:

```
>  hist(iq$V1)
```

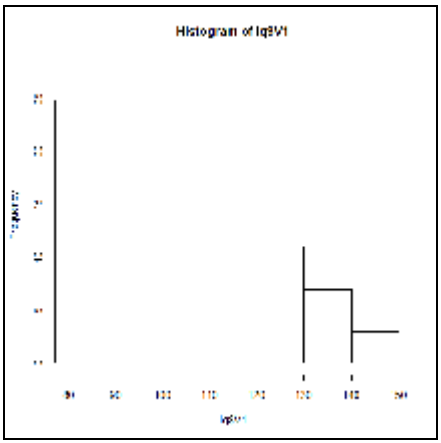


Рис. 1.5.2. Гистограмма оценок 75 взрослых людей по тесту IQ

7. Уточните аргументы команды title и задайте все заголовки на диаграмме:

```
> args(title)
function (main = NULL, sub = NULL, xlab = NULL, ylab =
NULL,
    line = NA, outer = FALSE, ...)
```

8. Рассчитайте показатели описательной статистики для данных теста IQ.

```
> summary(iq$V1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   83.0   101.5     108.0     111.2   122.0   148.0
```

9. Рассчитайте показатели изменчивости: дисперсию и стандартное отклонение

```
> var(iq$V1); sd(iq$V1)
[1] 225.2173
[1] 15.00724
```

10. Рассчитайте квартили:

```
 > quantile(iq$V1)
   0%   25%   50%   75%  100%
 83.0 101.5 108.0 122.0 148.0
```

11. Найдите 10-й и 90-й процентили:

```
>quantile(iq$V1, probs=c(0.25, 0.75))
```

12. Сохраните рабочее пространство: File, Save Workplace

Контрольные вопросы

1. Как скопировать данные из Excel в среду R?
2. С помощью какой команды чертится гистограмма частотного распределения?
3. Какие показатели описательной статистики рассчитывает функция summary? Почему в качестве аргумента у нее используется выражение **iq$V1**? Можно ли было ввести в качестве аргумента только **iq**?
4. Каковы правила выбора имен для объектов в R?
5. Как рассчитать первый и девятый дециль?
6. Дайте содержательную интерпретацию результатам расчетов в данной работе.
7. Заполните следующую таблицу

Мера	**Функция R**
Минимальное значение	
Максимальной значение	
1-й квартиль	
3-й квартиль	
Сумма	
Асимметрия, эксцесс	

Практическая работа 1.6
Подготовка данных анкетного опроса к обработке с помощью SPSS для Windows

Цель работы: на учебном примере опроса респондентов-покупателей магазина «Демпинг» изучить процесс подготовки данных в SPSS. Научиться описывать переменные, вводить и редактировать данные и сохранять их в файле.

Постановка задачи

В таблице 1.6.1 приведена выборка из ответов покупателей магазина «Демпинг» на вопросы анкеты. Эти данные об 11-ти респондентах включают:

- номер анкеты;
- пол (1 – «мужской», 2 – «женский»);
- возраст;
- удовлетворенность покупкой (1 – «полностью удовлетворен», 4 – «абсолютно неудовлетворен»).

Таблица 1.6.1

Данные опроса покупателей магазина «Демпинг»

Номер анкеты	Пол	Возраст	Удовлетворенность покупками в магазине «Демпинг»
1	1	50	3
2	2	31	2
3	1	21	3
4	1	18	4
5	2	17	1
6	2	34	1
7	2	32	1
8	2	25	1
9	2	22	2
10	2	21	2
11	2	19	4

Таблица 1.6.1 называется **матрицей данных**. Данные, предназначенные для обработки в SPSS для Windows, должны быть представлены в виде такой матрицы. Матрица данных состоит из определенного числа строк и столбцов. Строки и столбцы образуют прямоугольную таблицу. При этом каждая строка соответствует одной анкете, а каждый столбец – одной переменной. Так как в нашем небольшом опросе участвовало 11 респондентов, матрица содержит 11 строк. Каждая строка включает четыре столбца для переменных **number** (номер анкеты), **sex** (пол), **age** (возраст) и **udovl** (удовлетворенность покупкой).

Ход работы

1. Запустите SPSS для Windows.
2. Чтобы описать переменные, в редакторе данных дважды щелкните на ячейке с надписью ***Var*** или на ярлычке ***Variable View*** в левом нижнем углу экрана.

Определите переменные, как показано на рисунке 1.6.1. Ниже описаны подробно необходимые шаги.

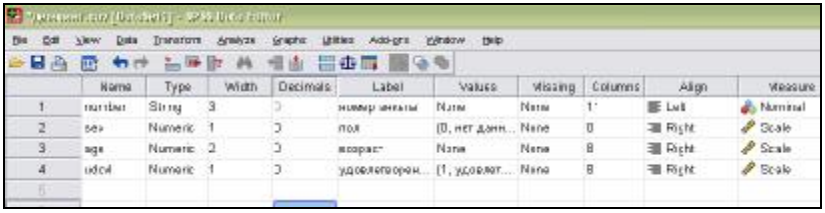


Рис. 1.6.1. Определение переменных в редакторе данных SPSS

Определение переменной «номер анкеты»

- Введите в поле ***Name*** текст "number" и подтвердите ввод нажатием на клавишу <Enter> или <Tab>.
- Чтобы задать тип переменной, щелкните в поле ***Type*** на кнопке с тремя точками. Откроется диалоговое окно ***Define***

Variable Type (Определение типа переменной). Выберите тип ***String*** (строковый) и установите число символов "3" (Мы предполагаем, что у нас не более 999 анкет). Подтвердите настройку кнопкой ОК и перейдите к следующему полю клавишей <Tab>.

- В полях ***Width*** и ***Decimals*** примите настройки, предлагаемые по умолчанию.
- Для метки переменной задайте текст "Номер анкеты".
- Для меток значений примите предлагаемую настройку None, нажав <Enter>.
- Примите предлагаемые настройки в поле ***Missing***, установите "8" в поле ***Columns, "Right***" – в поле ***Alignment*** и выберите настройку ***"Nominal***" в поле ***Measure.***

Определение переменной «пол»

- Введите в поле **Name** текст "sex" и подтвердите ввод нажатием на клавишу <Enter> или <Tab>.
- Чтобы задать тип переменной, щелкните в поле **Type** на кнопке с тремя точками. Откроется диалоговое окно **Define Variable Type** (Определение типа переменной). Примите предлагаемую настройку **Numeric** (Численный) и установите длину "1" и количество десятичных разрядов "0", так как в этой переменной будут храниться только значения 1, 2 или 0. Подтвердите настройку кнопкой ОК и перейдите к следующему полю клавишей <Tab>.
- Для формата столбца примите без изменений предлагаемые значения формата "1" и количества десятичных разрядов "0".
- Для метки переменной задайте текст "Пол".
- Щелкните в поле **Value Labels** на кнопке  . Откроется диалоговое окно **Define Value Labels** (Определение меток значений).

Метки значений определяются следующим образом:

Определение меток значений переменной «пол»

(см. рис. 1.6.2)

- Вначале введите в поле **Value** (Значение) число "1". Нажмите клавишу <Tab>.
- Введите в поле **Value label** (Метка значения) текст "женский".
- Щелкните на кнопке **Add** (Добавить). Метка значения будет добавлена в список.
- Повторите эти действия для значений "2" — "мужской" и "0" — "нет данных".

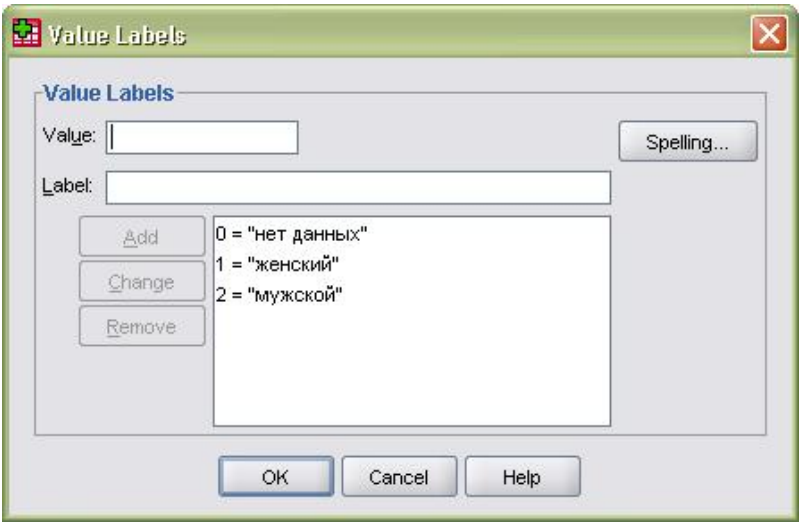


Рис. 1.6.2. Определение меток значений переменной «пол» в окне «Value Labels»

Определение меток значений переменной «возраст»

- Введите в поле ***Name*** текст "age" и подтвердите ввод.
- Чтобы задать тип переменной, щелкните в поле Type на кнопке с тремя точками ... .Откроется диалоговое окно ***Define Variable Type***. Примите предлагаемую настройку

Numeric и установите длину "2" (мы предполагаем, что все респонденты не старше 99 лет) и количество десятичных разрядов "0". Подтвердите настройку кнопкой ОК и перейдите к следующему полю клавишей <Tab>.

- В полях ***Column format*** и ***Decimals*** примите настройки, предлагаемые по умолчанию.
- Для метки переменной введите текст "Возраст", а для меток значений примите предлагаемую настройку **None**, нажав <Enter>.
- Чтобы задать пропущенные значения, щелкните в поле ***Missing values*** на кнопке с тремя точками  . Откроется диалоговое окно ***Define Missing Values***. По умолчанию предлагается вариант ***No missing values*** (Нет пропущенных значений), то есть все значения рассматриваются как допустимые. Введите единичное отсутствующее значение "0" и закройте диалоговое окно кнопкой ОК.
- Примите предлагаемые настройки "8" в поле Columns, "Right" в поле Alignment и "Scale" в поле Measure.

Определение меток значений переменной «удовлетворенность» (см. рис. 1.6.3)

- Введите в поле ***Name*** текст "***udovl***" и подтвердите ввод нажатием клавиши <Tab>.
- Чтобы задать тип переменной, щелкните в поле Type на кнопке с тремя точками. Откроется диалоговое окно ***Define Variable Type***. Примите предлагаемую настройку ***Numeric*** и установите длину "1" и количество десятичных разрядов "0", так как в этой переменной будут храниться только значения от 1 до 4 и 0 как отсутствующее значение. Подтвердите настройку кнопкой ОК и перейдите к следующему полю клавишей <Tab>.
- Для формата столбца примите значение "1" и количество десятичных разрядов "0".
- Для метки переменной задайте текст "Удовлетворенность".

o Щелкните в поле ***Value label*** на кнопке ... .Откроется диалоговое окно ***Define Value Labels***.

o Вначале введите в поле Value (Значение) число "1". Нажмите клавишу <Tab>.

o Введите в поле Value label (Метка значения) текст "удовлетворен полностью".

o Щелкните на кнопке Add (Добавить). Метка значения будет добавлена в список.

o Повторите эти действия для значений "2" – "скорее да, чем нет", "3" – "скорее нет, чем да", 4 – "полностью не удовлетворен" и "0" – "нет данных".

3. Перейдите на вкладку **Data View** и введите данные 11 анкет (согласно таб.1.6.1).

4. Сохраните файл с данными в личной папке.

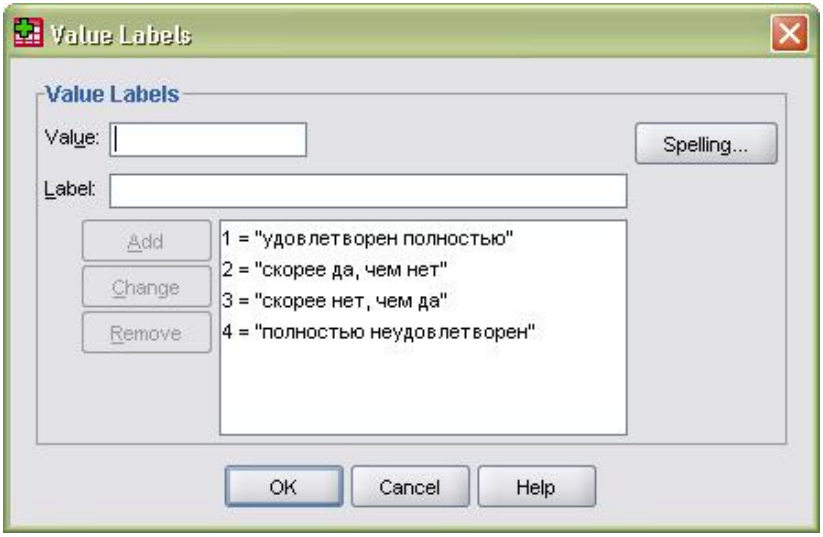


Рис. 1.6.3. Определение меток значений переменной «удовлетворенность» в окне «Value Labels»

Требования к отчету:

Отчет должен содержать:

- ответы на контрольные вопросы;
- файл с введенными данными.

Контрольные вопросы

1. Как запустить SPSS?
2. Как перейти к описанию переменных в SPSS?
3. Что означает тип переменной «***String***»?
4. Что устанавливают в колонке «***Decimals***» при описании переменных?
5. Что такое метки значений?

Практическая работа 1.7
Построение частотного распределения и вычисление статистических показателей с помощью SPSS для Windows

Цель работы: рассмотреть основные технические приемы при работе с SPSS для Windows; научиться строить частотное распределение, вычислять основные описательные статистики.

Ход работы

1. Открыть файл, содержащий итоги опроса покупателей магазина «Демпинг», выбрав меню: ***File, Open***

2. Построить частотное распределение для переменной **«пол».**

Для этого (см. рис. 1.7.1):

- Выбрать меню ***Analyze*** (Анализ), ***Descriptive statistics*** (Описательная статистика), ***Frequencies***… (Частотное распределение)
- Выделить переменную «пол»**,** щелкнув на ней.
- Нажать кнопку с треугольной стрелкой, перенеся выделенную переменную из списка исходных переменных в список выбранных переменных.
- Нажать кнопку OK.

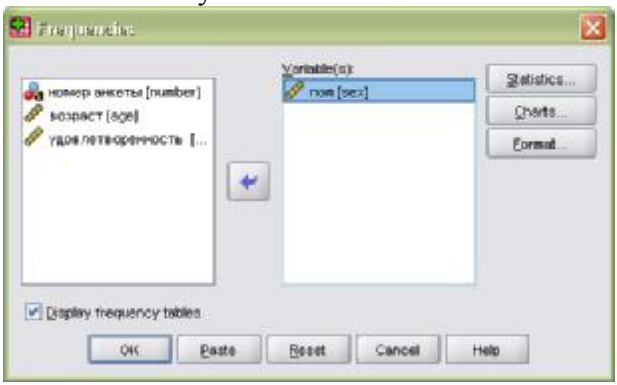


Рис. 1.7.1. Диалоговое окно «Frequencies». Первый этап построения частотного распределения для переменной «пол»

o Ознакомиться с полученными результатами в окне просмотра (рис. 1.7.2).

пол

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	женский	3	27,3	27,3	27,3
	мужской	8	72,7	72,7	100,0
	Total	11	100,0	100,0	

Рис. 1.7.2. Таблица частот для переменной «пол» в окне просмотра

3. Вернуться в редактор данных
o Либо выбрав меню ***Windows***
o Либо щелкнув на панели инструментов на символе редактора данных.

4. Аналогично рассчитать частотное распределение для переменной «удовлетворенность покупками» (см. рис. 1.7.3).

удовлетворенность

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	удовлетворен полностью	4	36,4	36,4	36,4
	скорее да, чем нет	3	27,3	27,3	63,6
	скорее нет, чем да	2	18,2	18,2	81,8
	полностью не-удовлетворен	2	18,2	18,2	100,0
	Total	11	100,0	100,0	

Рис. 1.7.3. Таблица частот для переменной «удовлетворенность покупками» в окне просмотра

5. Определить наибольшее, наименьшее и среднее значение переменной «возраст» а также моду, медиану и стандартное отклонение (см. рис. 1.7.4).

- Выбрать меню ***Analyze*** (Анализ), ***Descriptive statistics*** (Описательная статистика), ***Frequencies***… (Частотное распределение).
- В диалоговом окне ***Frequencies*** щелкнуть на кнопке ***Reset*** (Сброс), а потом перенести переменную «возраст» в список выбранных переменных.
- Щелкнуть кнопку ***Statistics…***
- В окне ***Frequencies: Statistics*** установить флажки ***Min (***минимальное***), Max (***максимальное***), Mean*** (среднее арифметическое), а также флажки для моды, медианы и стандартного отклонения (***Std. deviation***).
- Щелкнуть на кнопке ***Continue.***
- Снять флажок ***Display Frequencies tables*** (Показывать частотные таблицы).
- Щелкнуть на кнопке ОК.

Рис. 1.7.4. Диалоговое окно «Frequencies: Statistics»

o Ознакомиться с результатами в окне просмотра (рис. 1.5.5).

6. Сохранить результаты расчетов в личной папке.

Statistics

возраст

N	Valid	11
	Missing	0
Mean		26,36
Median		22,00
Mode		21
Std. Deviation		9,801
Minimum		17
Maximum		50

Рис. 1.7.5. Окно просмотра с результатами расчета

Требования к отчету:

Отчет должен содержать:

– ответы на контрольные вопросы;
– файл с введенными данными.

Контрольные вопросы

1. Как перенести переменную из списка исходных переменных в список выбранных переменных в окне «***Frequencies***»?

2. Описать параметры окна **«Frequencies: Statistics».**

3. Каково назначение пяти нижеприведенных стандартных кнопок в главном диалоговом окне?

o ***OK***
o ***Paste***
o ***Reset***
o ***Cancel***
o ***Help***

Практическая работа 1.8
Построение и редактирование графиков в SPSS для Windows

Цель работы: научиться строить и редактировать графики при работе со SPSS для Windows.

Ход работы

1. Открыть файл, содержащий итоги опроса покупателей магазина «Демпинг». Файл данных содержит 4 переменных: «номер анкеты», «пол», «возраст» респондента и «удовлетворенность» сделанными в магазине покупками. Меню: ***File, Open.***

2. Построить график частотного распределения для переменной «удовлетворенность»**.**

Для этого:

- Выбрать меню ***Analyze*** (Анализ), ***Descriptive statistics*** (Описательная статистика), ***Frequencies***… (Частотное распределение)
- Выделить переменную «удовлетворенность»**,** щелкнув на ней.
- Нажать кнопку с треугольной стрелкой, перенеся выделенную переменную из списка исходных переменных в список выбранных переменных.
- Щелкнуть на кнопке ***Charts***… (Диаграммы).
- В диалоговом окне ***Frequencies: Charts*** … (Частоты: Диаграммы) щелкнуть на опции ***Bar Charts*** (Столбиковые диаграммы), в области ***Chart Values*** (Значения диаграммы) щелкнуть на опции ***Percentage*** (Проценты) и затем на кнопке ***Continue*** (Далее) (см. рис. 1.8.1).

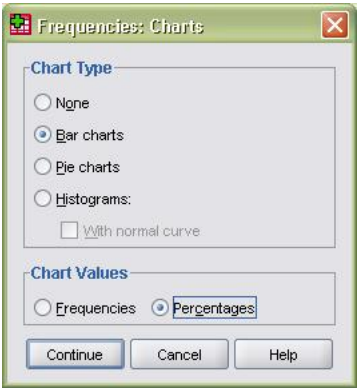


Рис. 1.8.1. Диалоговое окно «Frequencies: Charts»

- Снять флажок ***Display Frequency tables*** (Показывать частотные таблицы)
- Нажать кнопку OK.
- Ознакомиться с полученными результатами в окне просмотра.

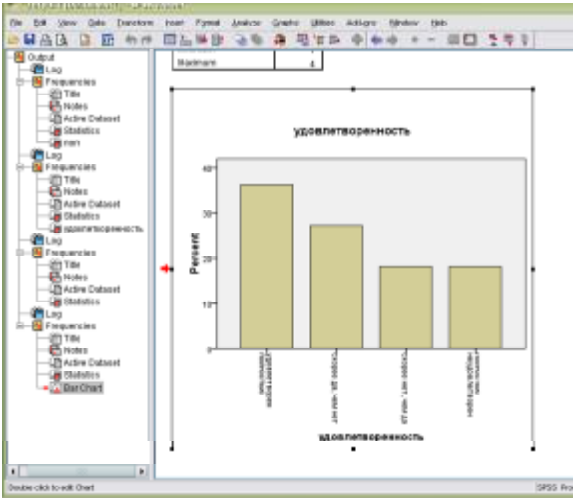


Рис. 1.8.2. Окно просмотра.
Гистограмма для признака «удовлетворенность»

Чтобы отредактировать график, нужно выполнить следующее:

- Щелкнуть дважды на какой-либо точке в пределах графика, после чего он будет помещен в редактор диаграмм.
- Задать объемный вид столбцов на графике:
  - В меню редактора диаграмм выбрать:

Format (Формат), ***Bar Style*** (Вид столбца)

Откроется диалоговое окно ***Bar styles***

  - Щелкнуть на области ***3-D effects***
  - В поле Depth (Глубина) ввести число “40”
  - Щелкнуть **Apply All (**Применить для всех**)** и затем на выключателе **Close.**
- Дать графику название
  - Меню ***Chart*** (Диаграмма), ***Title*** (Заголовок)
  - В поле ***Title1*** (Заголовок 1) введите текст “Опрос респондентов», а в поле ***Title2*** (Заголовок 2) – “Отношение к магазину Демпинг”.
  - Выделить график с помощью рамки

Chart (Диаграмма)

Outer Frame (Рамка снаружи)

  - Пометить столбцы процентными показателями.
  - В меню редактора диаграмм выберите:

Format (Формат)

Bar Label Style (Метки столбца)

Откроется диалоговое окно ***Bar Label Style*** (Метки столбцов)

  - Щелкнуть на области ***Framed*** (В рамке), затем на ***Apply All*** (Применить ко всем) и в заключение на ***Close.***
  - Закрыть редактор диаграмм.
  - Сохранить отредактированный график ***File, Save As***

Задание для самостоятельного выполнения

В таблице 1.8.1 приведены данные об использовании респондентами Internet для личных целей.

Эти данные о 30-ти респондентах включают:

- пол (1 – «мужской», 2 – «женский»);

- степень знакомства с Internet (1 – «почти не знаком», 7 – «хорошо знаком»);
- использование Internet (в часах в неделю);
- отношение к Internet и Internet-технологиям (измеренные по семибалльной шкале: 1 – «неблагосклонное», 7 – «максимально благосклонное»),
- использование Internet для приобретения товаров и банковских операций (1 – «да», 2 – «нет»).

Таблица 1.8.1

Данные об использовании Internet

Номер	Пол	Знакомство с Интернет	Использование Internet (в часах в неделю)	Отношение к Internet	Отношение к Internet-технологиям	Использование для покупок	Использование для банковских операций
1	1	7	14	7	6	1	1
2	2	2	2	3	3	2	2
3	2	3	3	4	3	1	2
4	2	3	3	7	5	1	2
5	1	7	13	7	7	1	1
6	2	4	6	5	4	1	2
7	2	2	2	4	5	2	2
8	2	3	6	5	4	2	2
9	2	3	6	6	4	1	2
10	1	7	15	7	6	1	2
11	2	4	3	4	3	2	2
12	2	5	4	6	4	2	1
13	1	6	9	6	5	2	1
14	1	6	8	3	2	2	2
15	1	0	5	5	4	2	2
16	2	4	3	4	3	2	2
17	1	6	9	5	3	1	1
18	1	4	4	5	4	1	2
19	1	7	14	6	6	1	1
20	2	6	6	6	4	2	2
21	1	6	9	4	2	2	2
22	1	5	5	5	4	2	1
23	2	3	2	4	2	2	2
24	1	7	15	6	6	1	1
25	2	6	6	5	3	!	2
26	1	6	13	6	6	1	1
27	2	5	4	5	5	1	1
28	2	4	2	3	2	2	2
29	1	4	4	5	3	1	2
30	1	3	3	7	5	1	2

Задание для выполнения

1. Описать необходимые переменные и ввести данные.
2. Рассчитать частотное распределение для переменных «***пол***», «***использование Интернет для покупок»*** и «***использование Интернет для банковских операций»***.
3. Рассчитать среднее число часов в неделю, затраченное пользователями на использование Интернет, а также моду, медиану, стандартное отклонение и минимальное и максимальное значения для этого признака.
4. Интерпретировать полученные данные.
5. Сохранить данные и результаты расчетов в личной папке (должно быть два файла).

Требования к отчету:

Отчет должен содержать:

– ответы на контрольные вопросы;
– файлы с результатами расчетов.

Контрольные вопросы

1. Как открыть файл с данными в SPSS для Windows и как сохранить результаты расчетов в файле?
2. Что такое частотное распределение переменной и с помощью каких команд меню оно строится?
3. Как рассчитать для выбранной переменной моду, медиану, среднее, дисперсию, стандартное отклонение?
4. Назначение кнопок в главном диалоговом окне.
5. Как вернуться из окна просмотра результатов SPSS «Viewer» к окну редактора данных SPSS «Data Editor»? Указать 3 способа.
6. Как отказаться от показа частотных таблиц в окне просмотра результатов?
7. Как построить гистограмму SPSS для Windows?
8. Как отредактировать график?
9. Как сохранить построенный график в файле?
10. Назовите основные элементы окна редактора диаграмм.

Практическая работа 1.9
Обработка данных анкетного опроса в среде R

Цель работы: на примере опроса покупателей магазина демпинг научиться создавать таблицу с данными в среде R и строить графики частотного распределения

Ход работы

Определите таблицу данных, представляющую данные опроса покупателей магазина Демпинг (см. раб..1.6).

1. Определите числовой вектор для переменной «возраст»:

```
age <- c(50,31,21,18,17,34,32,25,22,21,19)
```

2. Определите текстовый (категориальный вектор) пол:

```
> sex <- c("male", "female", "male", "male", "female", "female", "female", "female", "female", "female")
```

Проверьте, будет ли это текстовый вектор?

```
> is.character(sex)
[1] TRUE
```

Проверьте, будет ли это вектор?

```
>  is.vector(sex)
[1] TRUE
```

Рассмотрите его структуру:

```
> str(sex)
```

```
 chr [1:11] "male" "female" "male" "male" "female" "female"
"female" "female" "female" ...
```

Проинформируйте R о том, что перед нами категориальный тип данных:

```
 > sex.f <- factor(sex)
> sex.f
 [1] male   female male   male   female female female female
female female female female
Levels: female male
```

3. Определите текстовый вектор для переменной «удовлетворенность покупками».

```
>udovl <- c( "sknet", "skda","sknet", "neud",
"ud","ud","ud","ud","skda","skda", "neud")
> udovl.f <-factor(udovl)
> udovl.f

 [1] sknet skda  sknet neud  ud    ud    ud    ud    skda  skda
neud
Levels: neud skda sknet ud
```

Укажите, что у нас упорядоченная категориальная переменная:

```
 >udovl.o <- ordered(udovl.f, levels=c("neud", "sknet", "skda",
"ud"))
> udovl.o
 [1] sknet skda  sknet neud  ud    ud    ud    ud    skda  skda
neud
Levels: neud < sknet < skda < ud
```

4. Рассчитайте частотное распределение покупателей по полу и удовлетворенности покупками.

```
> table(sex)
sex
female   male
     8      3
```

```
> table(udovl)
udovl
 neud  skda sknet    ud
    2     3     2     4
```

5. Постройте графики для переменных пол и удовлетворенность покупками.

```
> plot(sex.f)
```

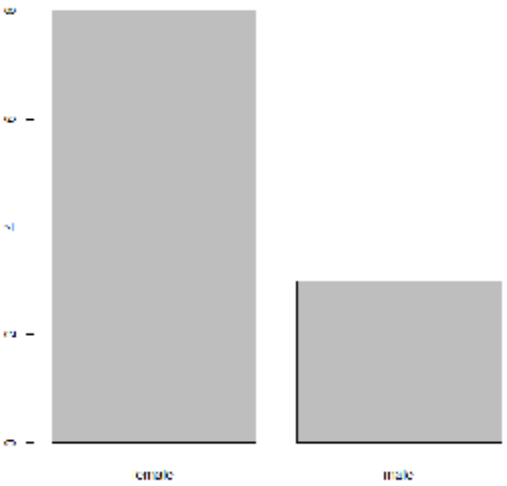


Рис. 1.9.1. Распределение покупателей по полу

```
> plot(udovl.o)
```

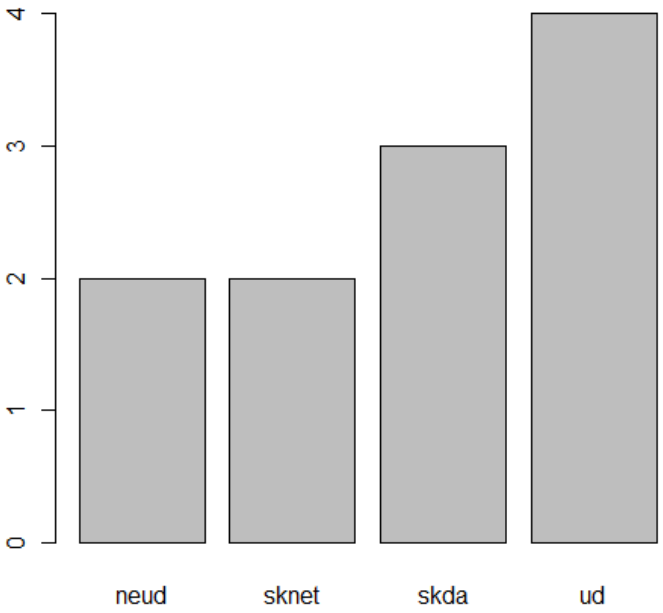


Рис. 1.9.2. Распределение покупателей по удовлетворенности покупками

6. Подсчитайте самостоятельно средний возраст покупателей, а также моду, медиану, стандартное отклонение, минимальный и максимальный возраст.

Заполните таблицу по результатам расчетов:

Мера	Значение
Среднее	
Мода	
Медиана Стандартное отклонение	
Минимальный	
Максимальный	

7. Объедините три созданных вектора в одну таблицу данных

```
>demping <- data.frame(sex1=sex.f, age1=age,
udovl1=udovl.o)
> demping
    sex1 age1 udovl1
1    male   50  sknet
2  female   31   skda
3    male   21  sknet
4    male   18   neud
5  female   17     ud
6  female   34     ud
7  female   32     ud
8  female   25     ud
9  female   22   skda
10 female   21   skda
11 female   19   neud
```

8. Просмотрите структуру созданной таблицы данных:

```
> str(demping)
'data.frame':   11 obs. of  3 variables:
 $ sex1  : Factor w/ 2 levels "female","male": 2 1 2 2 1 1 1 1 1
1 ...
```

```
 $ age1  : num  50 31 21 18 17 34 32 25 22 21 ...
 $ udovl1: Ord.factor w/ 4 levels "neud"<"sknet"<..: 2 3 2 1 4 4
4 4 3 3 ...
```

9. Выведите информацию только по женщинам-покупательницам.

```
 > demping[demping$sex=="female",]
    sex1   age1 udovl1
2  female   31   skda
5  female   17     ud
6  female   34     ud
7  female   32     ud
8  female   25     ud
9  female   22   skda
10 female   21   skda
11 female   19   neud
```

10. Сохраните рабочее пространство: File, Save Workplace

Контрольные вопросы

1. Как определяется числовой вектор в среде R?
2. Как определяется текстовый вектор?
3. Как задаются факторы (категориальные переменные)?
4. Каким образом задается упорядоченность в градациях фактора?
5. Как построить гистограмму частотного распределения?
6. Выведите информацию 1) по мужчинам-покупателям: 2) по покупателям, полностью удовлетворенным покупками; 3) по покупателям, полностью неудовлетворенным покупками.
7. Дайте содержательную интерпретацию результатам расчетов в данной работе.

2. ВЫЯВЛЕНИЕ РАЗЛИЧИЙ В РАСПРЕДЕЛЕНИИ ПРИЗНАКА С ПОМОЩЬЮ КРИТЕРИЯ ПИРСОНА χ^2

Множество задач маркетинговых, социологических, психологических, политологических, педагогических исследований предполагает те или иные сопоставления.

Одна из часто встречающихся задач в исследовании – сопоставление эмпирического распределения признака с каким-либо теоретическим законом распределения (равномерным, нормальным) или сравнение двух эмпирических распределений между собой, чтобы выявить сходство или различия в форме распределения.

Для сравнения эмпирического распределения признака с теоретическим (нормальным, равномерным и т.п.) и для сравнения двух, трех или более эмпирических распределений друг с другом может использоваться критерий Пирсона χ^2 (произносится – "хи-квадрат»), который рассчитывается по следующей формуле:

Критерий Пирсона χ^2

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{эмпj} - f_{теорj})^2}{f_{теорj}}$$

- где $f_{эмпj}$ - эмпирические частоты
$f_{теорj}$ - теоретические частоты
k - число разрядов признака

Замечания и ограничения по применению критерия Пирсона χ^2

- Объем выборки должен быть достаточно большим (>30).
- Теоретическая частота для каждой клетки таблицы должна быть больше 5.
- Выбранные разряды должны охватывать весь диапазон вариативности признака, а группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях.
- Разряды должны быть не перекрещивающимися, т.е. если наблюдение отнесено к одному разряду, то оно не может быть отнесено к другому.
- Сумма наблюдений по разрядам должна быть равна общему числу наблюдений.
- Если сравниваются два эмпирических распределения, то теоретические частоты рассчитываются как отношение произведения итогов по строке на итог по столбцу к общему числу наблюдений.

Практическая работа 2.1
Использование критерия Пирсона χ^2 для проверки согласованности распределений с помощью MS Excel и SPSS для Windows

Цель работы: научиться представлять концентированные данные для обработки в Excel и SPSS для Windows, строить для них таблицы сопряженности и вычислять критерий Пирсона χ^2.

Постановка задачи

Исследователей в области маркетинга интересует, зависит ли объем покупок модной одежды от семейного положения покупателей различного пола. В таблицах 2.1.1 и 2.1.2 приведены данные для мужчин и женщин соответственно.

Таблица 2.1.1

Покупка модной одежды мужчинами

Покупает	Семейное положение	
	Женат	Неженат
Много	32	12
Мало	70	19
Практически не покупает	5	5
Итоги	107	36

Таблица 2.1.2

Покупка модной одежды женщинами

Покупает	Семейное положение	
	Замужем	Незамужем
Много	24	31
Мало	75	20
Практически не покупает	7	5
Итоги	106	56

С помощью критерия хи-квадрат определить, есть ли связь между семейным положением и объемом покупок модной одежды для мужчин и женщин.

Решение с помошью электронных таблиц Excel

Ход работы

Алгоритм расчета χ^2 с помощью MS Excel

1. Занести в таблицу наименования разрядов признака (столбец A)
2. Занести соответствующие им эмпирические частоты для женатых и неженатых мужчин (столбцы B и C).
3. Подсчитать итоги по строкам и столбцам (столбец D).
4. Вычислить теоретические частоты для женатых и неженатых мужчин как отношение произведения итогов по строке и итогов по столбцу к общему числу объектов (столбцы E и F).
5. Для расчета $\chi^2_{эмп.}$ воспользоваться функциями ХИ2ТЕСТ и ХИ2ОБР. Функции ХИ2ТЕСТ передаются два параметра: интервал с эмпирическими частотами и интервал с теоретическими частотами. Функции ХИ2ОБР передаются два параметра: результат, полученный в ХИ2ТЕСТ, и число степеней свободы, определяющееся по формуле: $\nu=(k-1)*(l-1)$, где k – число разрядов признака или строк в таблице (для нашего примера $k=3$), l – число столбцов в таблице (для нашего примера 2).
6. По таблицам критических точек распределения χ^2 для данного числа степеней свободы определить $\chi^2_{критич.}$

Если $\chi^2_{эмп.}<\chi^2_{критич.}$, то расхождения между распределениями статистически недостоверны (распределения согласуются между собой), а, следовательно, женатые и неженатые мужчины не различаются по уровню покупок модной одежды.

Формулы для расчетов приведены в таблице 2.1.3, результаты расчетов в таблице 2.1.4.

Таблица 2.1.3

Таблица с формулами

	A	B	C	D	E	F
1	**Связь между покупкой модной одежды и семейным положением для мужчин**					
2	Покупает	Эмпирическая частота		Итоги	Теоретическая частота	
3		Семейное положение			Семейное положение	
4		Женат	Неженат		Женат	Неженат
5	Много	32	12	=СУММ(B5:C5)	=$D5*B$8/$D$8	=$D5*C$8/$D$8
6	Мало	70	19	=СУММ(B6:C6)	=$D6*B$8/$D$8	=$D6*C$8/$D$8
7	Практически не покупает	5	5	=СУММ(B7:C7)	=$D7*B$8/$D$8	=$D7*C$8/$D$8
8	Итоги	=СУММ(B5:B7)	=СУММ(C5:C7)	=СУММ(D5:D7)	=СУММ(E5:E7)	=СУММ(F5:F7)
9						
10						
11	хи2тест		=ХИ2ТЕСТ(B5:C7;E5:F7)			
12	***хи-квадрат***		=ХИ2ОБР(C11;2)			

Таблица 2.1.4

Таблица с результатами:

	A	B	C	D	E	F
1	**Связь между покупкой модной одежды и семейным положением для мужчин**					
2	Покупает	Эмпирическая частота		Итоги	Теоретическая частота	
3		Семейное положение			Семейное положение	
4		Женат	Неженат		Женат	Неженат
5	Много	32	12	44	32,9	11,1
6	Мало	70	19	89	66,6	22,4
7	Практически не покупает	5	5	10	7,5	2,5
8	Итоги	107	36	143	107	36
9						
10						
11	хи-тест		0,131			
12	***хи-квадрат***		4,066			

Аналогично решите задачу про покупку модной одежды женщинами.

Решение задачи в SPSS

Ход работы

1. Определить в редакторе 3 переменные:
   - Покупка с градациями (1 – много, 2 – мало, 3 – практически не покупает),
   - Семейное положение (1 – женат, 2 – неженат),
   - freq (частота) – указывает частоту каждого сочетания.
2. Ввести данные для заданных переменных согласно следующей таблицы 2.5.

Таблица 2.5

Покупка	Сем. положение	freq
Много	Женат	32
Мало	Женат	70
Практически не покупает	Женат	5
Много	Неженат	12
Мало	Неженат	19
Практически не покупает	Неженат	5

3. Взвесить данные, выбрав меню
Data (Данные)
Weight Cases (Взвесить случаи).

В диалоговом окне ***Weight Cases*** (рис. 2.1.1) выбрать опцию ***Weight Cases by*** и перенести переменную freq в поле ***Frequency Variable***. Закрыть диалоговое окно.

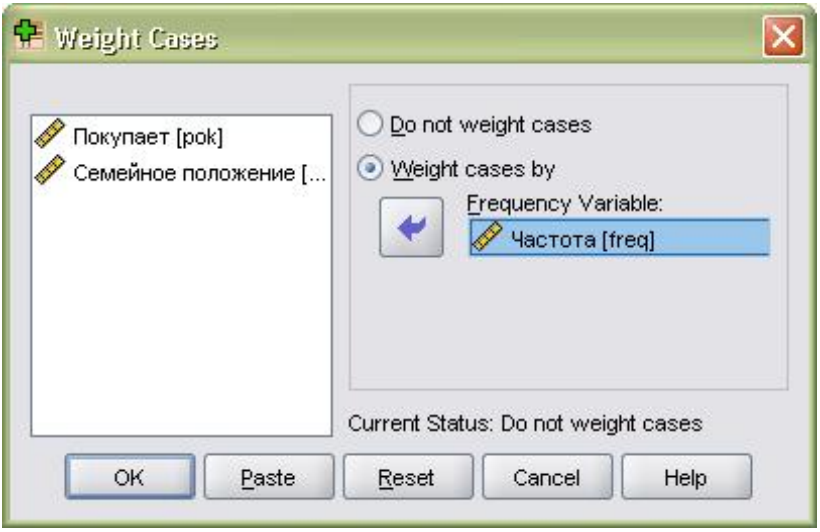


Рис. 2.1.1. Диалоговое окно «*Weight Cases*»

4. Построить таблицы сопряженности, выбрав меню ***Analyze*** (Анализ), ***Descriptive statistics*** (Описательная статистика), ***Crosstabs*** (Таблицы сопряженности) (рис. 2.1.2):

- Перенести переменную «**Покупка»** в список переменных строк.
- Переменную «**Семейное положение» –** в список переменных столбцов.

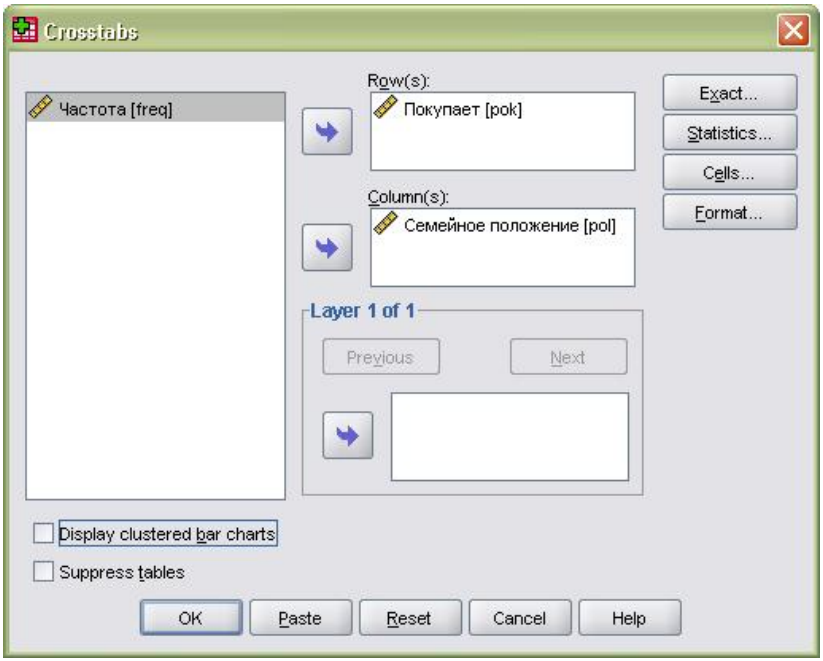


Рис. 2.1.2. Диалоговое окно «*Crosstabs*»

- Нажать кнопку ***Statistics***…
- В диалоговом окне ***Crosstabs: Statistics*** … (Частоты: Статистика) щелкнуть на опции ***Chi*-square** и затем на кнопке ***Continue*** **(Далее)** (рис. 2.1.3).

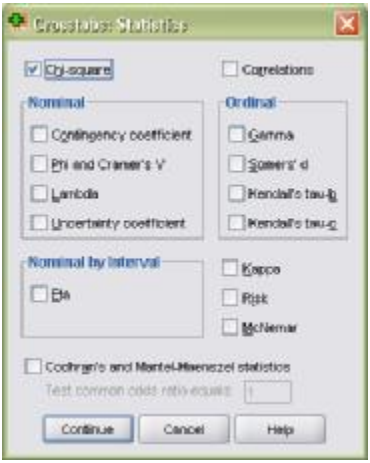


Рис. 2.1.3. Диалоговое окно «*Crosstabs: Statistics*»

o Щелкнуть на кнопке ***Cell*** (Ячейка). В группе ***Counts*** установить опцию наблюдаемые и ожидаемые частоты. В группе ***Percentage*** поставить галочку возле опции ***Column*** (Процент по колонке). Нажать **Continue OK** (рис. 2.1.4).

Рис. 2.1.4. Диалоговое окно «*Crosstabs: Cell Display*»

5. Рассмотреть и проанализировать результаты в окне вывода (табл. 2.1.6 – 2.1.7).

Таблица 2.1.6

Покупает * Семейное положение Crosstabulation

			Семейное положение		Total
			женат	неженат	
Покупает	Много	Count	32	12	44
		Expected Count	32,9	11,1	44,0
		% within Семейное положение	29,9%	33,3%	30,8%
	Мало	Count	70	19	89
		Expected Count	66,6	22,4	89,0
		% within Семейное положение	65,4%	52,8%	62,2%
	Практически не покупает	Count	5	5	10
		Expected Count	7,5	2,5	10,0
		% within Семейное положение	4,7%	13,9%	7,0%
Total		Count	107	36	143
		Expected Count	107,0	36,0	143,0
		% within Семейное положение	100,0%	100,0%	100,0%

Таблица 2.1.7

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,066[a]	2	,131
Likelihood Ratio	3,649	2	,161
Linear-by-Linear Association	,279	1	,597
N of Valid Cases	143		

a. 1 cells (16.7%) have xpected count less than 5.

По аналогии решить задачу о покупке модной одежды женщинами.

Задачи для самостоятельного решения

1. По данным таблицы 2.1.8 определите, существует ли зависимость между возрастом клиента и его затратами на летний отдых.

Таблица 2.1.8

Зависимость между возрастом клиента и затратами на летний отдых

Возраст	Уровень затрат на летний отдых			Всего
	Высокий	Средний	Низкий	
Молодые	25	40	31	
Среднего возраста	52	63	47	
Пожилые	12	23	58	
Всего				

2. По данным таблицы 2.9 проверить предположение о том, что частота обращений в телефонную службу доверия неравномерно распределяется по дням недели (с помощью критерия χ^2 сравнить заданное эмпирическое распределение с равномерным).

Таблица 2.1.9

Распределение обращений в телефонную службу доверия по дням недели

День недели	Число обращений
Понедельник	9
Вторник	6
Среда	7
Четверг	6
Пятница	9
Суббота	15
Воскресенье	18
Всего	70

Требования к отчету

Отчет должен содержать:

- ответы на контрольные вопросы;
- файл с результатами расчетов.

Контрольные вопросы.

1. Назначение критерия Пирсона χ^2.
2. Ограничения критерия Пирсона χ^2.
3. Что такое эмпирические частоты?
4. Что такое теоретические частоты?
5. Как вычисляется критерий Пирсона χ^2 в Excel?
6. Как представить данные для расчетов критерия Пирсона χ^2 с помощью SPSS?
7. Дайте содержательную интерпретацию полученным при решении задач результатам.

Практическая работа 2.2
Критерий Пирсона хи-квадрат для концентрированных данных в R

Ход работы

Попробуем с помощью R решить задачу о сравнении двух эмпирических распределений из предыдущей практической работы.

Создаем исходные данные (три вектора):

```
 >a<-c(1,2,3,1,2,3)
> b<-c(1,1,1,2,2,2)
> c<-c(32,70,5,12,19,5)
```

Здесь для вектора **a** цифрами закодировано: 1- много, 2 – мало, 3 – практически не покупает ; для вектора **b**: 1 – женат, 2 – неженат. Вектор **c** представляет столбец частот для взвешивания данных.

Объединяем их в общую таблицу данных с помощью команды **data.frame**:

```
>data2<-data.frame(pokupka=a, sempol=b, freq=c)
```

И проверяем, содержимое таблицы данных data2

```
> data2
  pokupka sempol freq
1       1       1   32
2       2       1   70
3       3       1    5
4       1       2   12
5       2       2   19
6       3       2    5
```

Поскольку данные концентрированные, восстанавливаем из них неконцентрированный массив, путем дублирования:

```
>data2.x <- data2[rep(1:nrow((data2),data2$freq),]
```

Команда xtab позволяет представить эти данные в виде кросс-таблицы:

```
>xtabs(frq ~ pokupka + sempol, data2)
```

```
          sempol
pokupka    1  2
      1   32 12
      2   70 19
      3    5  5
```

Рис. 2.2.1. Кросстаблица «покупка*семейное положение»

Для получения кросстаблицы в стиле SPSS можно использовать пакет **gmodels** и команду **CrossTable**:

```
> library(gmodels)
>CrossTable(data2.x$pokupka, data2.x$sempol,chisq = TRUE)
```

Вот какая таблица получится в результате:

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  143


                | data2.x$sempol
data2.x$pokupka |         1 |         2 | Row Total |
----------------|-----------|-----------|-----------|
              1 |        32 |        12 |        44 |
                |     0.026 |     0.077 |           |
                |     0.727 |     0.273 |     0.308 |
                |     0.299 |     0.333 |           |
                |     0.224 |     0.084 |           |
----------------|-----------|-----------|-----------|
              2 |        70 |        19 |        89 |
                |     0.174 |     0.518 |           |
                |     0.787 |     0.213 |     0.622 |
                |     0.654 |     0.528 |           |
                |     0.490 |     0.133 |           |
----------------|-----------|-----------|-----------|
              3 |         5 |         5 |        10 |
                |     0.824 |     2.448 |           |
                |     0.500 |     0.500 |     0.070 |
                |     0.047 |     0.139 |           |
                |     0.035 |     0.035 |           |
----------------|-----------|-----------|-----------|
   Column Total |       107 |        36 |       143 |
                |     0.748 |     0.252 |           |
----------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  4.066283     d.f. =  2     p =  0.1309236
```

Рис. 2.2.2. Результаты кросстабуляции и вычисления критерия Пирсона хи-квадрат для концентрированных данных

Таким образом, мы получили эмпирическое значение критерия Пирсона хи-квадрат, равное 4,06, что согласуется с расчетами, проведенными нами в SPSS и Excel в п.2.

Контрольые вопросы

1. Как самим создать таблицу данных для текущего примера?
2. Как восстановить из этих данных неконцентрированный массив?
3. Объясните, что делает следующая команда?

```
>xtabs(freq ~ pokupka + sempol, data2)
```

4. В каком пакете находится команда CrossTable?
5. По рис.2.5. объясните, какие данные выводятся в результате работы этой команды?
6. Решите аналогичную задачу о женщинах в среде R (см. п. 2).

Литература

1. https://stat.ethz.ch/pipermail/r-help/2008-November/180925.html
2. http://statmethods.net/stats/frequencies.html

3. КОРРЕЛЯЦИЯ

Виды зависимостей:		
причинно-следственные	прямые	функциональные
не причинно-следственные	обратные	корреляционные

Функциональная зависимость имеет место, если каждому значению одной переменной (X) соответствует вполне определенное значение другой переменной (Y).

Корреляционная зависимость имеет место, если определенному значению одной переменной соответствует несколько значений другой переменной**.**

Или, другими словами, **корреляционная зависимость** – это зависимость, при которой определенному значению одной величины соответствует целый комплекс значений другой, представляющий собой ряд распределения, причем при изменении данной величины меняется ряд распределения и его среднее.

Коэффициенты корреляции могут принимать значения от -1 до +1.

На рис. 3.1. представлено схематическое изображение силы и направления корреляции.

Существуют два подхода к оценке силы корреляции. Первый (см. табл. 3.1) ориентирован только на абсолютную величину коэффициента корреляции, а второй – на уровень значимости данного коэффициента корреляции при данном объеме выборки (см. таб. 3.2).

Таблица 3.1

Оценка силы корреляции по абсолютной величине коэффициента корреляции

Абс. величина коэффициента корреляции |r|	**Оценка силы корреляции**
$\|r\| \geq 0,70$	сильная (тесная)
$0,50 \leq \|r\| \leq 0,69$	средняя
$0,30 \leq \|r\| \leq 0,49$	умеренная
$0,20 \leq \|r\| \leq 0,29$	слабая
$\|r\| \leq 0,19$	очень слабая

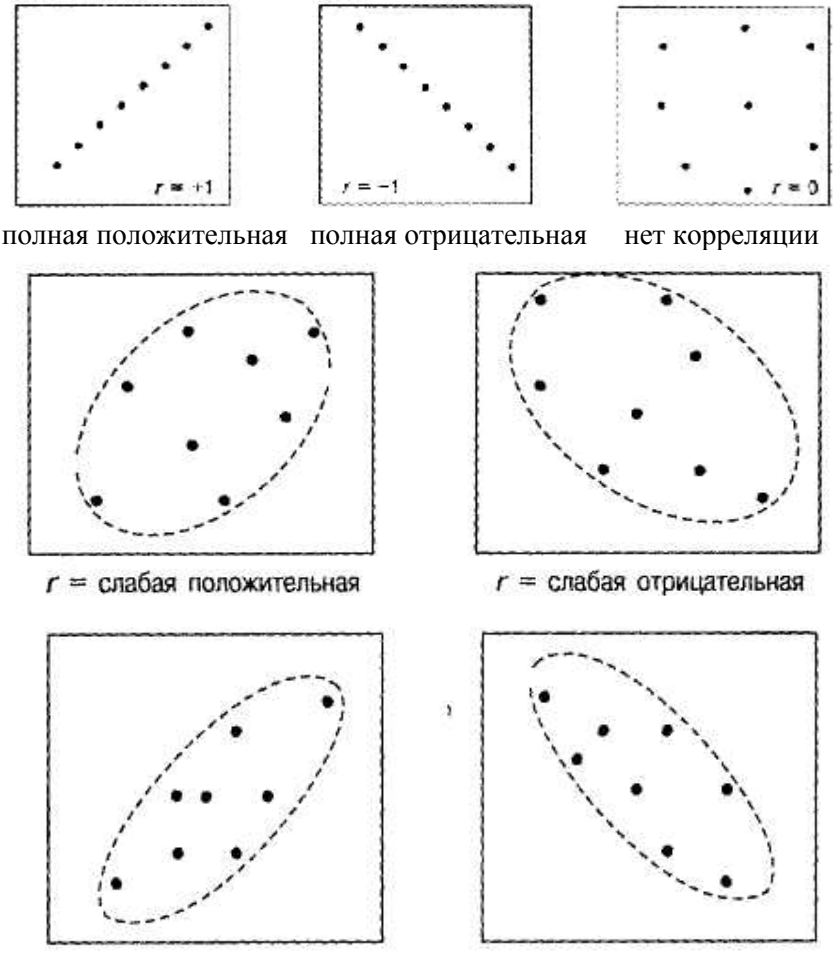


Рис. 3.1. Схематическое изображение силы и направления корреляции

В соответствии со вторым подходом (см. табл. 3.2), чем больше объем выборки, на которой изучалась корреляция между переменными, тем меньший по абсолютной величине коэффициент корреляции признается показателем достоверности корреляционной связи.

Таблица 3.2

Оценка силы корреляции по уровню статистической значимости

Уровень статистической значимости коэффициента корреляции	**Оценка силы корреляции**
$p \leq 0{,}01$	высокая значимая корреляция
$0{,}01 < p \leq 0{,}05$	значимая корреляция
$0{,}05 < p \leq 0{,}10$	тенденция достоверной связи
коэффициент корреляции не достигает уровня статистической значимости	незначимая корреляция

Наиболее целесообразным является сочетание двух подходов, когда при формулировании выводов о наличии или отсутствии корреляции между переменными, учитывается и абсолютная величина коэффициента корреляции, и уровень значимости этого коэффициента для данного объема выборки.

Причинность и корреляция

Наличие корреляции между переменными отнюдь не означает, что между ними существует причинно-следственная связь. Возможно, либо одна из переменных является частичной причиной другой, либо обе они являются следствием каких-то общих причин. Наличие значимой корреляции между переменными дает нам основание не отвергать гипотезу о причинно-следственной связи между переменными, но не может служить подтверждением такой гипотезы.

Изучая корреляционную зависимость между переменными, исследователь, интерпретируя значения коэффициентов корреляции, должен учитывать сущность, природу этих переменных.

В качестве мер корреляции используются различные коэффициенты. Выбор коэффициента зависит от типа данных, которые обрабатываются. Широко распространенными в маркетинговых, социологических, психолого-педагогических и других исследованиях являются коэффициенты ассоциации и контингенции, коэффициент Пирсона $\chi 2$, коэффициент ранговой корреляции Спирмена, коэффициент ранговой корреляции Кенделла, коэффициент Пирсона-Браве для непрерывных шкал.

Практическая работа 3.1
Вычисление коэффициента корреляции Пирсона-Браве для метрических шкал

Цель работы: научиться вычислять коэффициент корреляции Пирсона и строить диаграмму рассеивания, определять значимость полученного коэффициента

Краткие теоретические сведения

Для определения корреляционной зависимости между двумя переменными, заданными метрическими шкалами, служит ***коэффициент Пирсона-Браве.*** Он характеризует меру линейной связи между переменными.

Пусть X и Y две коррелирующих между собой переменных. Коэффициент Пирсона-Браве вычисляется по формуле:

$$r = \frac{\sum_{i=1}^{n} Z_i^x \mathbf{o} Z_i^y}{n-1}, \text{ где} \qquad (3.1.1)$$

$$Z_i^x = \frac{X_i - X_{cp}}{S_x}, \text{ где} \qquad (3.1.2)$$

S_x *- стандартное отклонение,*
X_i *- i-значение признака,*
X_{cp} *- среднее значение признака X*

$$Z_i^y = \frac{Y_i - Y_{cp}}{S_y}, \text{ где} \qquad (3.1.3)$$

S_y *- стандартное отклонение,*
Y_i *- i-значение,*
Y_{cp} *- среднее значение признака Y*

Постановка задачи

В таблице 3.1.1 приведены данные о количестве машин, ежедневно паркуемых на открытой стоянке и в гараже около крупного университета. Найдите и интерпретируйте корреляцию между этими двумя величинами.

Таблица 3.1.1

Количество автомобилей, которые парковались около университета

День недели	Открытая стоянка	Гараж
Понедельник	140	180
Вторник	120	200
Среда	130	190
Четверг	110	210
Пятница	160	160
Суббота	135	185

Решение с помощью Excel

Ход работы

1. Ввести данные.

2. Вычислить коэффициент Пирсона для этих данных вначале по формулам, а потом выполнить то же с помощью функции КОРРЕЛ (рис. 3.1.1.).

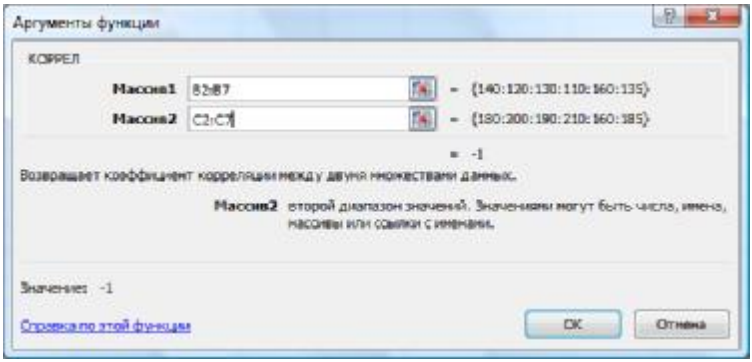


Рис. 3.1.1. Диалоговое окно «Аргументы функции КОРРЕЛ»

3. Заполнить в тетради таблицу 3.1.2 (внести расчетное и критическое значения коэффициента, определить его значимость по таблице критических значений).

Таблица 3.1.2

Объем выборки	Эмпирическое значение коэффициента	Критическое значение	Резюме (значим/ незначим)

4. Построить диаграмму рассеивания (выделить данные, вызвать мастер диаграмм, выбрать тип диаграммы – «точечная», подписать все заголовки) (см. рис. 3.1.2.).

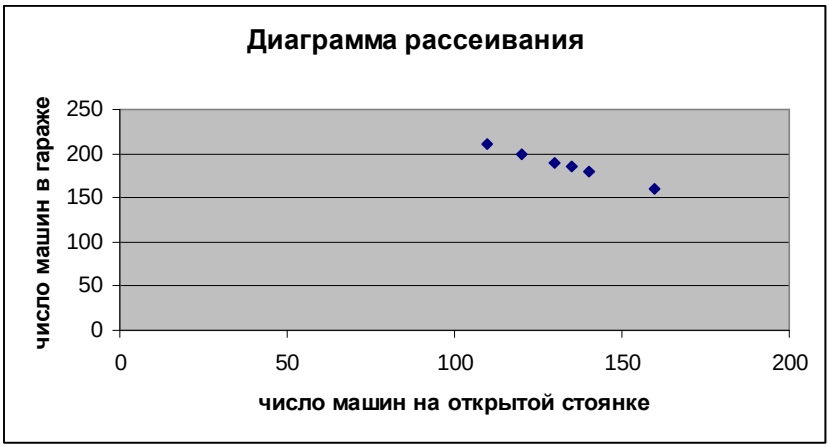


Рис. 3.1.2. Диаграмма рассеивания, построенная по данным таблицы 3.1.1.

Решение с помощью SPSS
Ход работы

1. Создать файл данных в SPSS: описать две переменные (Open – число машин на открытой стоянке, Garag – число машин в гараже).

2. Ввести или скопировать из файла Excel данные.

3. Вычислить коэффициент Пирсона для этих данных:

Меню: ***Analyze*** (Анализ), ***Correlate*** (Корреляция), ***Bivariate*** (Двумерная).

В окне ***Bivariate Correlation*** (рис. 3.1.3) задать необходимые параметры.

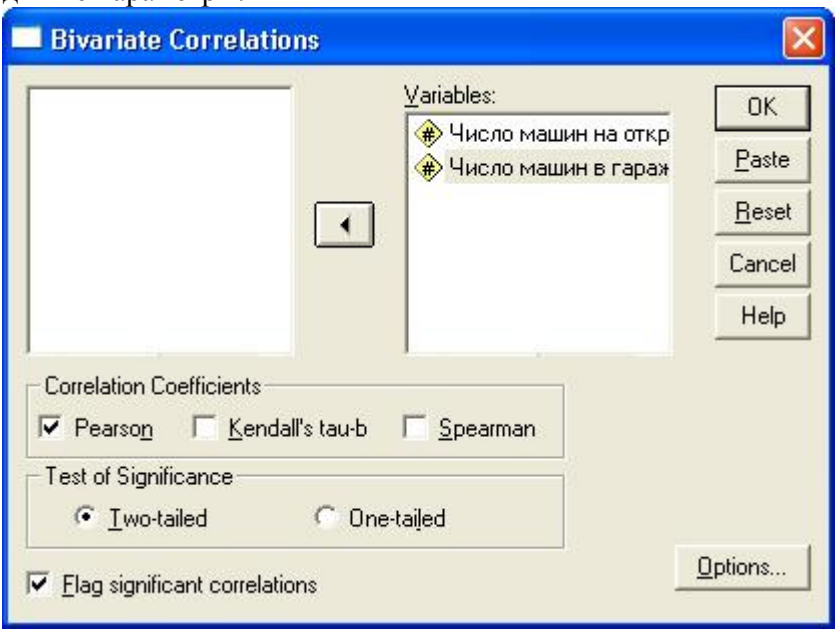


Рис. 3.1.3. Диалоговое окно «Bivariate Correlation»

4. Ознакомиться с полученными результатами в окне просмотра (рис. 3.1.4). Результаты включают: коэффициент корреляции (*Pearson Correlation*), уровень значимости (*Sig*), объем выборки (*N*). Записать их в тетрадь. Определить, значим ли коэффициент корреляции.

Correlations

		Число машин на открытой стоянке	Число машин в гараже
Число машин на открытой стоянке	Pearson Correlation	1	-1,000**
	Sig. (2-tailed)	.	.
	N	6	6
Число машин в гараже	Pearson Correlation	-1,000**	1
	Sig. (2-tailed)	.	.
	N	6	6

**. Correlation is significant at the 0.01 level (2-tailed).

Рис. 3.1.4. Окно просмотра для примера с парковкой

5. Построить диаграмму рассеивания:

Меню: ***Graphs*** (Графики), ***Scatter…***(Диаграмма рассеивания), ***Simple*** (Простая) (рис. 3.1.5).

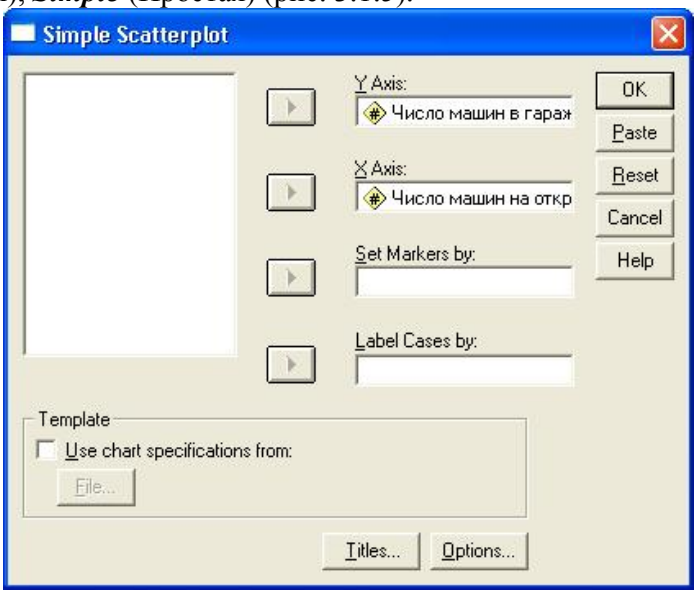


Рис. 3.1.5. Диалоговое окно «Simple Scatterplot» (простая точечная диаграмма)

Рассмотреть полученную диаграмму рассеивания (рис. 3.1.6).

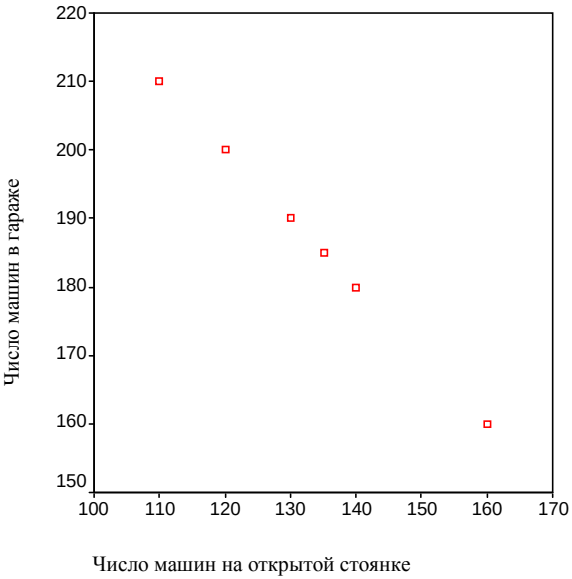


Рис. 3.1.6. Точечная диаграмма для примера с парковкой

6. Сохранить данные расчетов и исходные данные в файлах.

7. Сделать вывод о наличии или отсутствии значимой корреляции между двумя признаками.

Задания для самостоятельной работы

1. В таблице 3.1.3 содержатся данные о ежедневном объеме продаж в долларах лазерных принтеров, картриджей к принтерам и школьных принадлежностей. Найдите и интерпретируйте корреляцию между этими величинами. Решите задачу в среде Microsoft Excel и в среде SPSS.

Таблица 3.1.3

Ежедневный объем продаж ($)

День	Лазерные принтеры	Картриджи	Школьные принадлежности
1	17291	4379	3618
2	13734	2258	3514
3	18802	4206	3587
4	12171	2137	3007
5	16402	3775	3850
6	19904	4781	3675
7	14023	1991	3120
8	17513	2663	3345
9	17847	4451	2045
10	12718	1648	3292
11	12292	2342	3405
12	11846	2646	2799
13	17088	3216	2417
14	13523	2184	2405

Требования к отчету

Отчет должен содержать:

- ответы на контрольные вопросы;
- файл с результатами расчетов.

Контрольные вопросы

1. Дайте определение фукциональной зависимости.
2. Дайте определение корреляционной зависимости.
3. Какие значения могут принимать коэффициенты корреляции?
4. Как вычислить коэффициент корреляции Пирсона-Браве с помощью электронных таблиц?
5. Для каких шкал он используется?
6. Как построить диаграмму рассеивания?
7. Какие параметры задаются функции КОРРЕЛ?
8. Как рассчитать коэффициент корреляции Пирсона-Браве и построить диаграмму рассеивания в SPSS?
9. Как определить значимость коэффициента корреляции в SPSS?
10. Дайте содержательную интерпретацию результатам расчетов.

Практическая работа 3.2

Корреляция по Пирсону с помощью R

Ход работы

Запустите R. Решим в среде R две задачи из предыдущей лабораторной работы.

Задание 1.

1. Определить два вектора, представляющие число машин, припаркованных на открытой стоянке и в гараже около крупного университета:

```
> open <- c(140, 120,130,110,160,135)
> garag <- c(180, 200,190,210,160,185)
```

2. Подсчитать корреляцию по Пирсону:

```
>  cor.test(open,garag)

        Pearson's product-moment correlation

data:  open and garag
t = NaN, df = 4, p-value = NA
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 NaN NaN
sample estimates:
cor
 -1

Warning messages:
1: In sqrt(1 - r^2) : NaNs produced
2: In atanh(r) : NaNs produced
```

3. Нарисовать диаграмму рассеивания:

```
>plot(open, garag,  main="диаграмма рассеивания", xlab="число машин на открытой стоянке",  ylab="число машин в гараже")
```

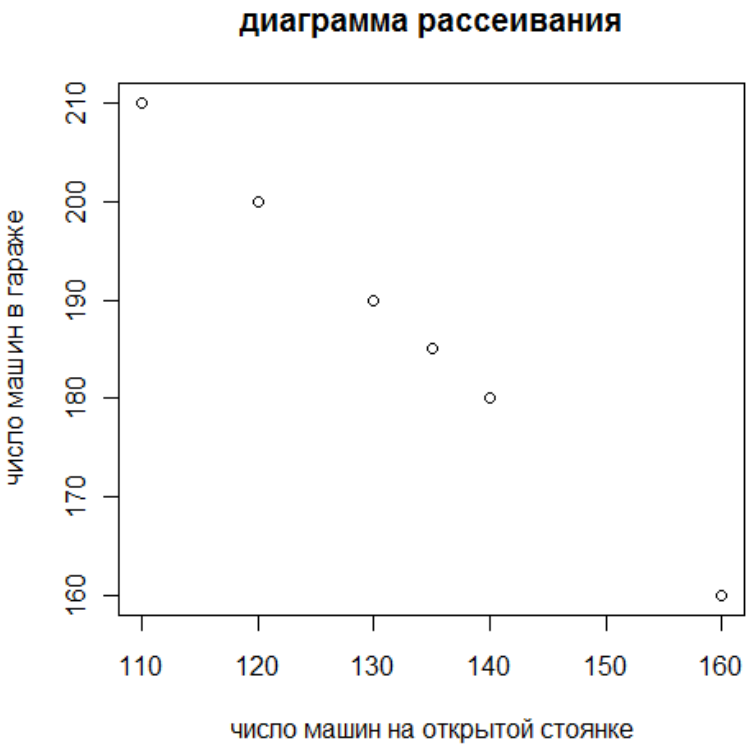


Рис. 3.2.1.Диаграмма рассеивания

Задание 2.

4. Введите данные в Excel (см. рис.3.2.2.) и сохраните под именем printer.csv

	A	B	C	D
1	День	Принтеры	Катриджи	Канцтовары
2	1	17291	4379	3618
3	2	13734	2258	3514
4	3	18802	4206	3587
5	4	12171	2137	3007
6	5	16402	3775	3850
7	6	19904	4781	3675
8	7	14023	1991	3120
9	8	17513	2663	3345
10	9	17847	4451	2045
11	10	12718	1648	3292
12	11	12292	2342	3405
13	12	11846	2646	2799
14	13	17088	3216	2417
15	14	13523	2184	2405
16				
17				

Лист1 | **Лист2** | Лист3

Рис.3.2.2. Исходные данные в Excel

5. Загрузите данные в таблицу данных data4.

```
>data4<-read.table("K:\\Rexample\\printer.csv", sep=";", dec=",", header=TRUE)
> data4
  День Принтеры Катриджи Канцтовары
1    1   17291     4379       3618
2    2   13734     2258       3514
```

3	3	18802	4206	3587
4	4	12171	2137	3007
5	5	16402	3775	3850
6	6	19904	4781	3675
7	7	14023	1991	3120
8	8	17513	2663	3345
9	9	17847	4451	2045
10	10	12718	1648	3292
11	11	12292	2342	3405
12	12	11846	2646	2799
13	13	17088	3216	2417
14	14	13523	2184	2405

6. Найдите коэффициент корреляции между продажами принтеров и катриджей:

```
>cor.test(data4$Принтеры,data4$Катриджи)

        Pearson's product-moment correlation

data:  data4$Принтеры and data4$Катриджи
t = 5.8127, df = 12, p-value = 8.306e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6035144 0.9545418
sample estimates:
      cor
0.8590239
```

В нашем случае получена значимая корреляция продаж принтеров и катриджей, коэффициент корреляции 0.8590239.

7. Проведите корреляционный анализ для всей матрицы данных:

```
>cor(data4)
                День       Принтеры   Катриджи  Канцтовары
День         1.0000000 -0.3182177 -0.3589594 -0.6655984
Принтеры    -0.3182177  1.0000000  0.8590239  0.1679915
Катриджи    -0.3589594  0.8590239  1.0000000  0.1426419
Канцтовары -0.6655984  0.1679915  0.1426419  1.0000000
```

Можно воспользоваться пакетом ellipse, чтобы визуализировать корреляционную матрицу. При этом коэффициенты корреляции рисуются в виде эллипсов. Чем ближе коэффициент корреляции к +1 или -1, тем более узким становится эллипс. Воспользуемся функцией plotcorr.

```
>install.packages(pkgs=c("ellipse"))
```

8. Постройте график, иллюстрирующий результаты корреляционного анализа (см. рис.)

```
> library(ellipse)
> plotcorr(cor(data4))
```

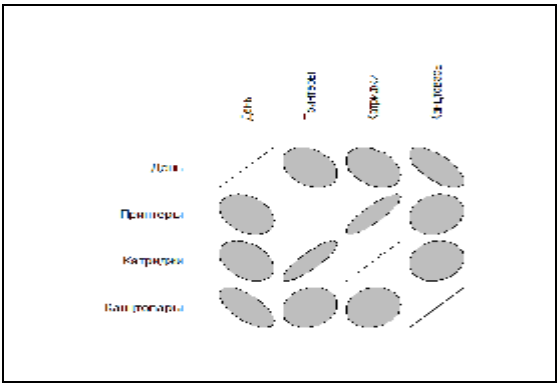


Рис.3.2.3. Результаты корреляционного анализа

Контрольные вопросы

1. Как подготовить данные для корреляционного анализа в R?
2. Как осуществляется корреляционный анализ для двух переменных?
3. Как строится корреляционная матрица нескольких переменных?
4. Как построить диаграммы рассеивания для пары переменных?
5. Как можно визуализировать данные корреляционного анализа?
6. Охарактеризуйте рисунок 1.
7. Заполните следующую таблицу по результатам расчетов:

Корреляция между	Коэффициент корреляции Пирсона	p-Значение	Вывод (корреляция значима/ незначима)
Принтерами и катриджами			
Катриджами и канцтоварами			
Канцтоварами и принтерами			

Практическая работа 3.3
Вычисление коэффициента ранговой корреляции Спирмена

Цель работы: научиться вычислять коэффициент ранговой корреляции Спирмена и определять его значимость.

Краткие теоретические сведения

Пусть N объектов могут быть упорядочены как по признаку X, так и по Y. Обозначим

R_i^x - ранг i-объекта по признаку X

R_i^y - ранг i-объекта по признаку Y

$d_i = R_i^x - R_i^y$ - мера несовпадения рангов

Для определения корреляционной зависимости между двумя переменными, заданными ранговыми шкалами служит коэффициент ранговой корреляции Спирмена, который вычисляется по формуле:

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} di^2}{N(N^2 - 1)} \qquad (3.3.1)$$

Постановка задачи.

Исследование, проведенное на британском рынке, выявило следующую значимость факторов успеха новых продуктов, производимых японскими и британскими фирмами (см. табл. 3.3.1, в которой ***n*** означает число опрошенных фирм).

Из таблицы 3.3.1 видно, что наиболее часто ключевым фактором успеха называли степень адаптации к требованиям покупателей. И только относительно небольшая доля фирм (1/4), как японских, так и британских, считает эффективный маркетинг важным критерием успеха.

Сравнение оценок, полученных от японских и британских фирм, обнаруживает, что первые из них придают большее значение наличию конкурентного преимущества, а вторые – большей адаптации к запросам потребителей.

Таблица 3.3.1

Факторы успеха новых продуктов, выявленные на британском рынке

Факторы успеха	Процент фирм	
	японских (n=116)	британских (n=86)
Хорошая адаптация к потребностям	69,8	75,6
Превосходство над конкурентами по качеству	79,3	59,3
Превосходство над конкурентами по дизайну	69,9	45,3
Превосходство над конкурентами по соотношению достоинство/цена	58,6	61,6
Превосходство над конкурентами по конструкции	55,2	48,8
Весьма конкурентная цена	41,4	27,9
Адаптация	39,7	34,9
Уникальность	36,2	29,1
Эффективный маркетинг	27,6	25,6
Глубокий анализ рынка	27,6	18,6
Большой объем рынка	20,7	16,3
Синергия производство/маркетинг	16,4	18,6
Уклонение от рынков с высокой конкуренцией и удовлетворенными покупателями	7,8	10,5
Уклонение от динамичных рынков с частой сменой товара	2,6	4,7

Определите, коррелируют ли между собой оценки японских и британских фирм.

Решение задачи с помощью Excel

Ход работы

1. Ввести данные в столбцы A, B, C.

2. В колонках D и E записать формулы для вычисления рангов.

3. В ячейке F3 записать формулу для вычисления разности рангов. Скопировать эту формулу в ячейки F4:F16.

4. В ячейку G3 записать формулу для возведения этой разности в квадрат. Скопировать ее в ячейки G4:G16.

5. В ячейке G17 просуммировать разности квадратов.

6. В ячейке B18 подсчитать число измерений (объектов) с помощью функции СЧЕТ (в нашем случае это число факторов успеха – 14), а в ячейке G18 – формулу для вычисления коэффициента ранговой корреляции Спирмена:

=1-6*G17/(B18*(B18*B18-1)) (см. табл. 3.3.2).

Таблица 3.3.2

Ввод формул для расчета коэффициента ранговой корреляции Спирмена

	A	B	C	D	E	F	G
1	**Факторы успеха**	**Процент фирм**		**Ранги**		**Разность рангов**	**Разность рангов в квадрате**
2		**японских**	**британских**	**японских**	**британских**		
3	Хорошая адаптированность к потребностям	69,8	75,6	=РАНГ(B3;$B$3:$B$16)	=РАНГ(C3;C$3:$C$16)	=D3-E3	=F3*F3
4	Превосходство над конкурентами по качеству	79,3	59,3	=РАНГ(B4;$B$3:$B$16)	=РАНГ(C4;C$3:$C$16)	=D4-E4	=F4*F4
5	Превосходство над конкурентами по дизайну	69,9	45,3	=РАНГ(B5;$B$3:$B$16)	=РАНГ(C5;C$3:$C$16)	=D5-E5	=F5*F5
6	Превосходство над конкурентами по соотношению достоинство/цена	58,6	61,6	=РАНГ(B6;$B$3:$B$16)	=РАНГ(C6;C$3:$C$16)	=D6-E6	=F6*F6
7	Превосходство над конкурентами по конструкции	55,2	48,8	=РАНГ(B7;$B$3:$B$16)	=РАНГ(C7;C$3:$C$16)	=D7-E7	=F7*F7
8	Весьма конкурентная цена	41,4	27,9	=РАНГ(B8;$B$3:$B$16)	=РАНГ(C8;C$3:$C$16)	=D8-E8	=F8*F8
9	Адаптированность к возможностям фирмы	39,7	34,9	=РАНГ(B9;$B$3:$B$16)	=РАНГ(C9;C$3:$C$16)	=D9-E9	=F9*F9
10	Уникальность	36,2	29,1	=РАНГ(B10;$B$3:$B$16)	=РАНГ(C10;C$3:$C$16)	=D10-E10	=F10*F10
11	Эффективный маркетинг	27,6	25,6	=РАНГ(B11;$B$3:$B$16)	=РАНГ(C11;C$3:$C$16)	=D11-E11	=F11*F11
12	Глубокий анализ рынка	27,6	18,6	=РАНГ(B12;$B$3:$B$16)	=РАНГ(C12;C$3:$C$16)	=D12-E12	=F12*F12
13	Большой объем рынка	20,7	16,3	=РАНГ(B13;$B$3:$B$16)	=РАНГ(C13;C$3:$C$16)	=D13-E13	=F13*F13
14	Синергия производство/маркетинг	16,4	18,6	=РАНГ(B14;$B$3:$B$16)	=РАНГ(C14;C$3:$C$16)	=D14-E14	=F14*F14
15	Уклонение от рынков с высокой конкуренцией и удовлетворенными покупателями	7,8	10,5	=РАНГ(B15;$B$3:$B$16)	=РАНГ(C15;C$3:$C$16)	=D15-E15	=F15*F15
16	Уклонение от динамичных рынков с частой сменой товара	2,6	4,7	=РАНГ(B16;$B$3:$B$16)	=РАНГ(C16;C$3:$C$16)	=D16-E16	=F16*F16
17							=СУММ(G3:G16)
18	Объектов	=СЧЁТ(B3:B17)			Коэф.Спирмена		=1-6*G17/(B18*(B18*B18-1))

В результате вычислений получится таблица с результатами (см. табл. 3.3.3).

Таблица 3.3.3

Результаты вычисления коэффициента Спирмена

	A	B	C	D	E	F	G
1	**Факторы успеха**	**Процент фирм**		**Ранги**		**Разность рангов**	**Разность рангов в квадрате**
2		**японских**	**британских**	**японских**	**британских**		
3	Хорошая адаптированность к потребностям	69,8	75,6	3	1	2	4
4	Превосходство над конкурентами по качеству	79,3	59,3	1	3	-2	4
5	Превосходство над конкурентами по дизайну	69,9	45,3	2	5	-3	9
6	Превосходство над конкурентами по соотношению достоинство/цена	58,6	61,6	4	2	2	4
7	Превосходство над конкурентами по конструкции	55,2	48,8	5	4	1	1
8	Весьма конкурентная цена	41,4	27,9	6	8	-2	4
9	Адаптированность к возможностям фирмы	39,7	34,9	7	6	1	1
10	Уникальность	36,2	29,1	8	7	1	1
11	Эффективный маркетинг	27,6	25,6	9	9	0	0
12	Глубокий анализ рынка	27,6	18,6	9	10	-1	1
13	Большой объем рынка	20,7	16,3	11	12	-1	1
14	Синергия производство/маркетинг	16,4	18,6	12	10	2	4
15	Уклонение от рынков с высокой конкуренцией и удовлетворенными покупателями	7,8	10,5	13	13	0	0
16	Уклонение от динамичных рынков с частой сменой товара	2,6	4,7	14	14	0	0
17							34
18	Объектов	14			Коэф.Спирмена		0,925274725

Решение задачи с помощью SPSS для Windows

Ход работы

1. Описать в редакторе данных две переменных с именами japan, brit (японские и британские фирмы).

Рис. 3.3.1. Описание переменных в SPSS

Ввести данные согласно таблице, представленной на рис. 3.2.2.

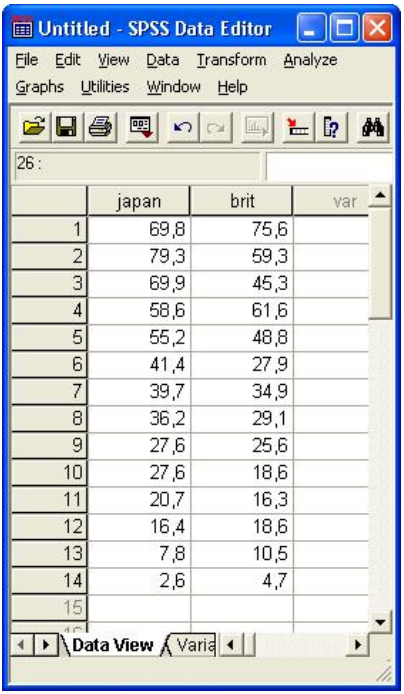


Рис. 3.3.2. Ввод данных в SPSS

2. Вычислить ранги для переменных japan, brit:

Меню: ***Transform*** (Преобразовать), ***Rank Cases***…(Ранги) (см. рис. 3.3.3)

В результате появятся две новые переменные:

From variable	New variable	Label
JAPAN	RJAPAN	RANK of JAPAN
BRIT	RBRIT	RANK of BRIT

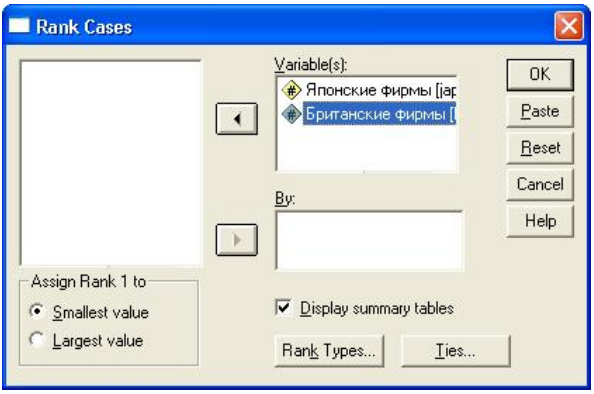


Рис. 3.3.3. Окно «Rank Cases»

3. Вычислить коэффициенты Спирмена и Кенделла для данных, представленных рангами:

Меню: ***Analyze*** (Анализ), ***Correlate*** (Корреляция), ***Bivariate*** (Двумерная).

В окне ***Bivariate Correlation*** задать необходимые параметры (см. рис. 3.3.4).

Примечание: можно вычислять коэффициенты ранговой корреляции без предварительного вычисления рангов, так как SPSS сам проделает необходимые расчеты.

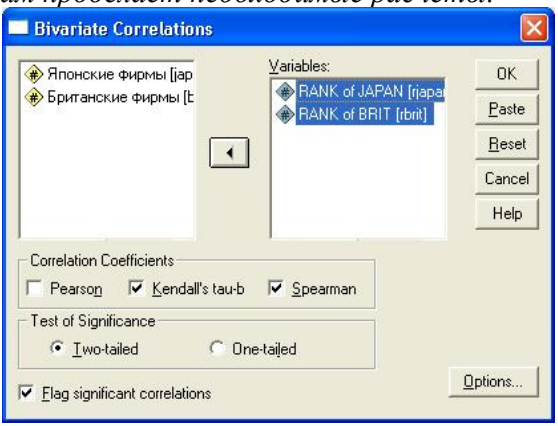


Рис. 3.3.4. Окно «Bivariate Correlation»

4. Ознакомиться с полученными результатами в окне просмотра (рис. 3.3.5).

			Японские фирмы	Британские фирмы
Kendall's tau_b	Японские фирмы	Correlation Coefficient	1,000	,811(**)
		Sig. (2-tailed)	.	,000
		N	14	14
	Британские фирмы	Correlation Coefficient	,811(**)	1,000
		Sig. (2-tailed)	,000	.
		N	14	14
Spearman's rho	Японские фирмы	Correlation Coefficient	1,000	,928(**)
		Sig. (2-tailed)	.	,000
		N	14	14
	Британские фирмы	Correlation Coefficient	,928(**)	1,000
		Sig. (2-tailed)	,000	.
		N	14	14

** Correlation is significant at the 0.01 level (2-tailed).

Рис. 3.3.5. Результаты вычисления коэффициентов корреляции

*Примечание: коэффициенты, обозначенные * –- значимы на уровне 0,05;*

*** – на уровне 0,01.*

Задание для самостоятельной работы

Руководство компании Marston Book Services (Великобритания), занимающейся предоставлением комплекса услуг по доставке потребителям печатной продукции (книги, плакаты, открытки и т.п.), разработало план усиления ориентации

деятельности компании на запросы потребителей. На одном из этапов разработки данного плана было проведено изучение мнений сотрудников компании и потребителей относительно важности отдельных показателей качества предоставляемых услуг. В таблице 3.2.4 приводятся рейтинги отдельных показателей качества.

Таблица 3.3.4

Рейтинг показателей качества услуг

Показатели качества услуг	Оценки сотрудников фирмы	Оценки потребителей
Быстрота обслуживания	1	7
Надежность	4	1
Неповрежденность упаковки заказов	3	8
Удовлетворение срочных заказов	8	3
Простота оформления заказов	7	6
Низкий уровень рекламаций	2	4
Предоставление информации по запросам	6	5
Выдерживание сроков выполнения заказов	5	2

Определите, согласуются ли оценки сотрудников компании и потребителей. Решите задачу в среде Excel и в среде SPSS для Windows.

Требования к отчету

Отчет о работе должен содержать:

– постановку задачи, файлы с исходными данными и результатами, формулы для вычисления, значения вычисленных коэффициентов;

– ответы на контрольные вопросы.

Контрольные вопросы

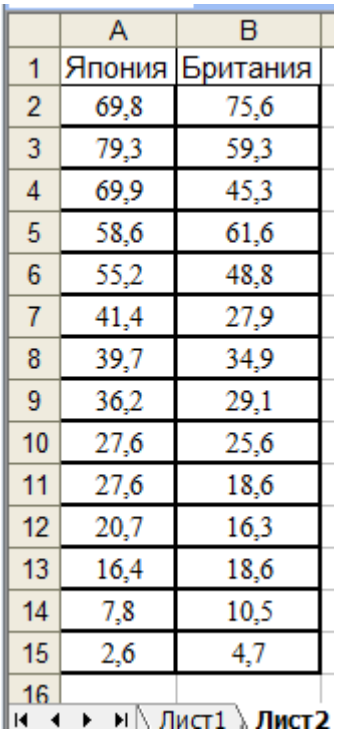
1. Какие коэффициенты ранговой корреляции вы знаете?
2. Для чего предназначен коэффициент корреляции Спирмена?
3. По какой формуле он вычисляется? Какие предельные значения принимает?
4. Как определить его значимость? Определите значимость коэффициента Спирмена в примере «японские-британские фирмы».
5. Как вычислить коэффициент Спирмена с помощью электронных таблиц Excel?
6. Назначение функции «СЧЕТ». Для чего она используется в данной лабораторной работе?
7. Назначение функции «РАНГ». Какие параметры передаются этой функции? Для чего используется в формуле для вычисления ранга абсолютная адресация?
8. Как преобразовать данные в ранги? Какие типы рангов есть в SPSS для Windows?
9. Как вычислить коэффициенты ранговой корреляции в SPSS?
10. Как определить значимость коэффициентов в SPSS?
11. Дайте содержательную интерпретацию полученным результатам.

Практическая работа 3.4

Вычисление коэффициентов ранговой корреляции с помощью R

Ход работы

1. Запустить R
2. Ввести данные в Excel



	A	B
1	Япония	Британия
2	69,8	75,6
3	79,3	59,3
4	69,9	45,3
5	58,6	61,6
6	55,2	48,8
7	41,4	27,9
8	39,7	34,9
9	36,2	29,1
10	27,6	25,6
11	27,6	18,6
12	20,7	16,3
13	16,4	18,6
14	7,8	10,5
15	2,6	4,7
16		

Лист1 Лист2

Рис.3.4. Исходные данные

3. Сохранить в формате .CSV (разделители запятые)

4. Импортировать в переменную data

```
>data<-read.table("K:\\Rexample\\lab321.csv", sep=";",
dec=",", header=TRUE)
> data
   Япония Британия
1    69.8     75.6
2    79.3     59.3
3    69.9     45.3
4    58.6     61.6
5    55.2     48.8
6    41.4     27.9
7    39.7     34.9
8    36.2     29.1
9    27.6     25.6
10   27.6     18.6
11   20.7     16.3
12   16.4     18.6
13    7.8     10.5
14    2.6      4.7
```

5. Подсчитать коэффициент корреляции Спирмена:

```
>   cor(data$Япония,data$Британия, method="spearman")
[1] 0.9284141
```

6. Вывести более полную статистику расчетов:

```
>cor.test(data$Япония,data$Британия, method="spearman")

        Spearman's rank correlation rho

data:  data$Япония and data$Британия
S = 32.5716, p-value = 1.663e-06
alternative hypothesis: true rho is not equal to 0
```

```
sample estimates:
      rho
0.9284141

Warning message:
In cor.test.default(data$Япония, data$Британия, method =
"spearman") :
  Cannot compute exact p-values with ties
```

7. Подсчитать коэффициент корреляции Кенделла:

```
> cor.test(data$Япония,data$Британия, method="kendall")

        Kendall's rank correlation tau

data:  data$Япония and data$Британия
z = 4.0083, p-value = 6.115e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8111111
Warning message:
In cor.test.default(data$Япония, data$Британия, method = "kend-
all") :
  Cannot compute exact p-value with ties
```

Контрольные вопросы

1.Как рассчитать коэффициент Спирмена в R?
2.Как рассчитать коэффициент Кенделла в R?
3.Как определить их значимость?

4. ЛИНЕЙНАЯ РЕГРЕССИЯ

Еще одним методом определения связи между варьирующими признаками является регрессионный анализ. Впервые термин «регрессия» был применен английским ученым-статистиком Ф.Гальтоном в 1886 г. в теории наследственности. Ф.Гальтон регрессией (regression to mediocrity – возврат к среднему состоянию) назвал явление, состоящее в том, что дети тех родителей, чей рост превышал среднее значение на ***a*** единиц, имели в среднем рост, превышающий среднее значение менее чем на ***a*** единиц.

Как мы уже выяснили, задача корреляционного анализа – установление корреляционной зависимости между переменными и определение величины этой зависимости в виде коэффициентов корреляции.

Задача регрессионного анализа – выражение корреляционной зависимости в виде функциональных отношений. Мы рассмотрим простейшую форму этой функциональной зависимости – линейную.

Имея функциональное отношение связи между переменными X и Y, можно оценивать Y по X, т.е. устанавливать причинно-следственные связи между переменными.

Переменная, которую мы хотим оценить, называется зависимой переменной Y, а переменная, используемая для ее оценки, – независимой переменной или фактором X.

Примеры задач на оценивание Y по X:

1. Предсказание отметок по английскому языку в университете (Y) по оценкам по этому предмету в школе (X).
2. Предсказание общей учебной успешности в школе (Y) по интеллекту (X).
3. Предсказание объема продаж (Y) по затратам на рекламу (X).
4. Прогнозирование текучести кадров (Y) в зависимости от уровня удовлетворенности работой (X) и т.п.

Чтобы вывести способ оценивания переменной Y на основе значений переменной X, мы должны знать, как связаны между собой X и Y. Рассмотрим алгоритм оценивания на примере.

Пусть мы хотим научиться предсказывать средний балл сессии у студентов (Y) в зависимости от того, сколько времени они тратят на просмотр телесериалов в период подготовки к экзаменам (X). Для этого нам необходимо:

1) узнать, сколько времени (часов в неделю) тратят на просмотр сериала n студентов накануне сессии;
2) измерить средний балл сессии у этих же студентов;
3) вывести уравнение, связывающее Y и X;
4) далее использовать это уравнение для предсказания среднего балла сессии (Y) у тех студентов, для которых нам известно, сколько времени в период подготовки к сессии они потратили на просмотр сериалов (X).

Данные измерений времени на просмотр сериалов и среднего балла сессии можно представить графически в виде диаграммы рассеивания (см. рис. 4.1), т.к. эти две переменные связаны между собой корреляционной зависимостью. Задача регрессионного анализа, как уже отмечалось, – представить корреляционную зависимость в виде функциональных отношений. В случае линейной регрессии (а мы рассматриваем именно этот случай), эта задача сводится к нахождению такой прямой, сумма квадратов отклонений от которой точек диаграммы рассеивания (сумма квадратов ошибки оценки) была бы минимальной. Эта прямая называется линией предсказания или линией тренда. Уравнение такой прямой – уравнение регрессии – в общем виде выглядит следующим образом:

$$y = b_1 * x + b_0 \qquad (4.1)$$

где b_1 и b_0 выбраны таким образом, чтобы сумма квадратов ошибок оценки была бы минимальной (критерий наименьших квадратов):

$$b_1 = r_{xy} * S_y / S_x \qquad (4.2)$$

$$b_0 = y_{ср} - b_1 * x_{ср} \qquad (4.3)$$

где r_{xy} – коэффициент корреляции Пирсона между X и Y;
S_x, S_y – стандартные отклонения по X и по Y.
$x_{ср}$, $y_{ср}$ – средние значения для X и Y.

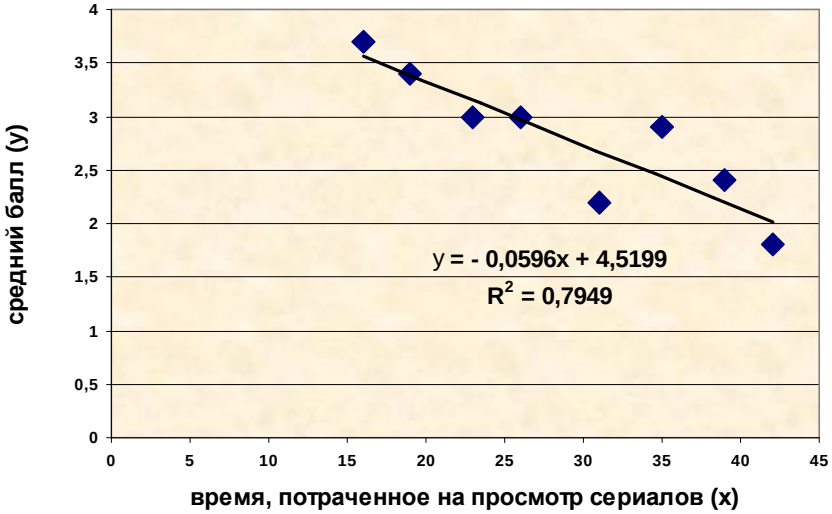


Рис. 4.1. Диаграмма рассеивания, линия предсказания, уравнение регрессии

Практическая работа 4.1
Вычисление коэффициентов корреляции и прогноз с помощью линейной регрессии в MS Excel

***Цель работы*:** научиться использовать электронные таблицы MS Excel для вычисления коэффициентов корреляции и прогнозирования с использованием линейной регрессии.

Постановка задачи

В таблице 4.1.1 приведены результаты тестирования 20-ти абитуриентов на вступительном экзамене в школу бизнеса (назовем его «тест 1» – это признак X) и результаты тестирования этих же лиц уже на выпускном курсе («тест 2», признак Y). Построить диаграмму рассеивания для этих данных, вычислить коэффициент корреляции Пирсона-Браве и вывести уравнение регрессии, связывающее признаки X и Y.

Таблица 4.1.1

Результаты тестирования в школе бизнеса

Тест 1 (X)	Тест 2 (У)
66	94
62	100
70	101
80	102
82	103
74	105
73	104
79	106
83	106
80	109
81	110
84	110
80	111
89	112
88	112
95	114
98	114
97	115
94	117
89	118

Ход работы

1. Ввести исходные данные согласно таблице 4.1.1 (см. рис. 4.1.1).

	A	B	C	D	E	F	G
1	**Результаты тестирования в школе бизнеса**						
2	**№ студента**	**тест 1**	**тест 2**				
3	1	66	94				
4	2	62	100				
5	3	70	101				
6	4	80	102				
7	5	82	103				
8	6	74	105				
9	7	73	104				
10	8	79	106				
11	9	83	106				
12	10	80	109				
13	11	81	110				
14	12	84	110				
15	13	80	111				
16	14	89	112				
17	15	88	112				
18	16	95	114				
19	17	98	114				
20	18	97	115				
21	19	94	117				
22	20	89	118				

Рис. 4.1.1. Ввод результатов тестирования

2. Вычислить коэффициент корреляции Пирсона-Браве:

а) меню ***Данные, Анализ данных, Корреляция (см. рис. 4.1.2);***

б) в диалоговом окне "***Корреляция***" (рис. 4.1.3) вывести и записать в тетрадь справку;

в) задать параметры: входной интервал (оба ряда данных), группировка по столбцам. Нажать ОК. Записать в тетрадь полученный коэффициент корреляции. Определить его значимость по таблице.

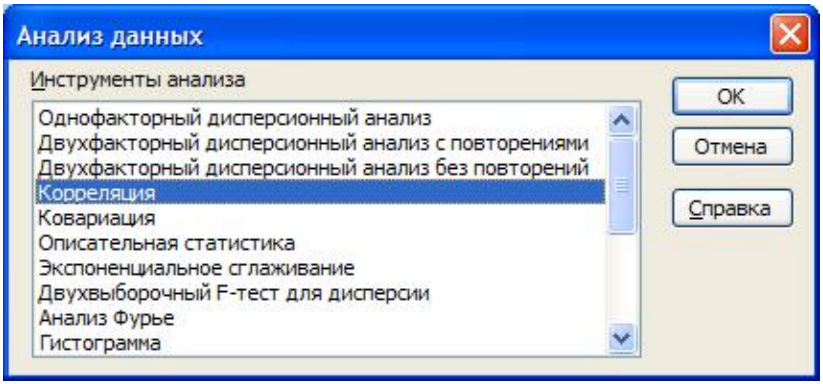


Рис. 4.1.2. Диалоговое окно «Анализ данных»

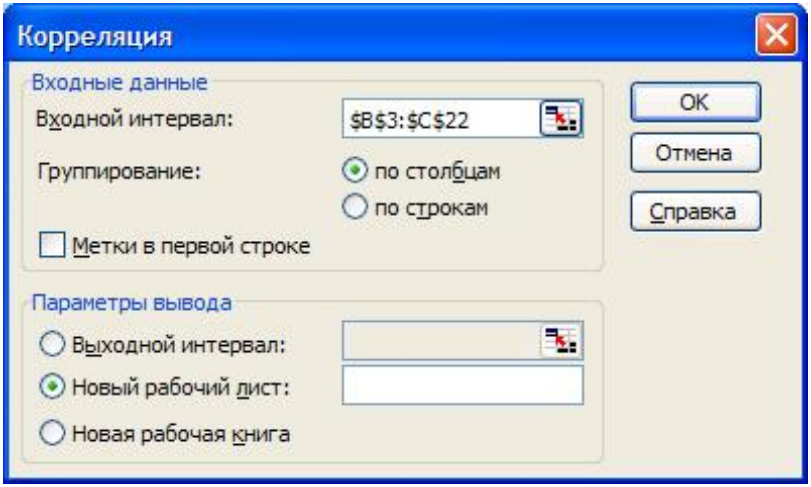


Рис. 4.1.3. Диалоговое окно «Кореляция»

3. Построить диаграмму рассеивания с помощью ***Мастера диаграмм*** (тип «точечная»). Сделать предположение о наличии или отсутствии корреляции между X и У по внешнему виду диаграммы.

4. Построить линию предсказания по методу наименьших квадратов (линейная регрессия):

а) щелкнуть правой кнопкой по точкам диаграммы;

б) в контекстном меню выбрать «***Добавить линию тренда»;***

в) в диалоговом окне "***Линия тренда***" на вкладке "***Тип***" (рис. 4.1.4) укажите – "**линейная**", на вкладке "***Параметры***" – поставьте флажок у опции "**Показывать уравнение на диаграмме**" и «**Поместить величину достоверности аппроксимации**» (см. рис. 4.1.5).

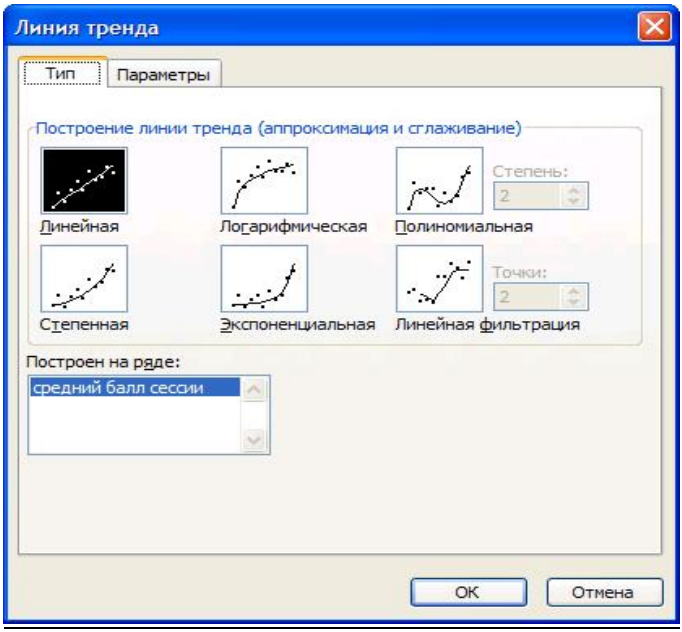


Рис. 4.1.4. Диалоговое окно «Линия тренда», вкладка «Тип»

В итоге должна получиться точечная диаграмма с линией тренда, уравнением регрессии и коэффициентом аппроксимации, как это показано на рис. 4.1.6.

Уравнение нужно записать в тетрадь. И, пользуясь MS Excel, предсказать по нему значения выпускного теста для лиц, у которых по результатам вступительного теста были следующие баллы:

а) 43;
б) 75;
в) 99.

5. Сохранить таблицу в личной папке.

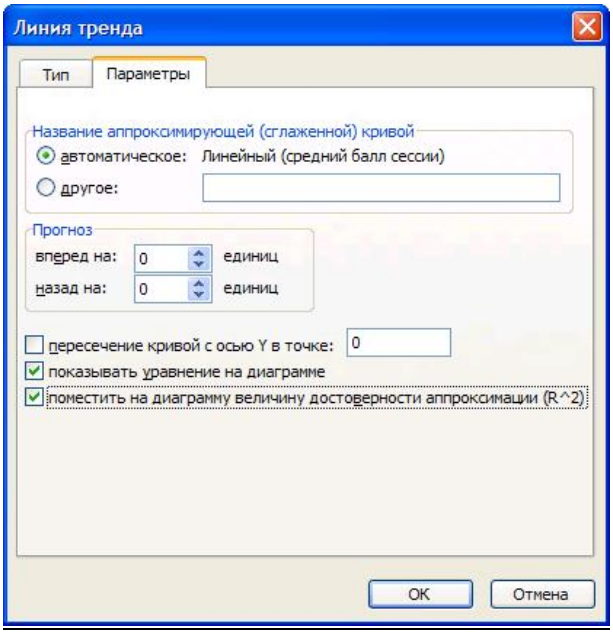


Рис. 4.1.5. Диалоговое окно «Линия тренда», вкладка «Параметры»

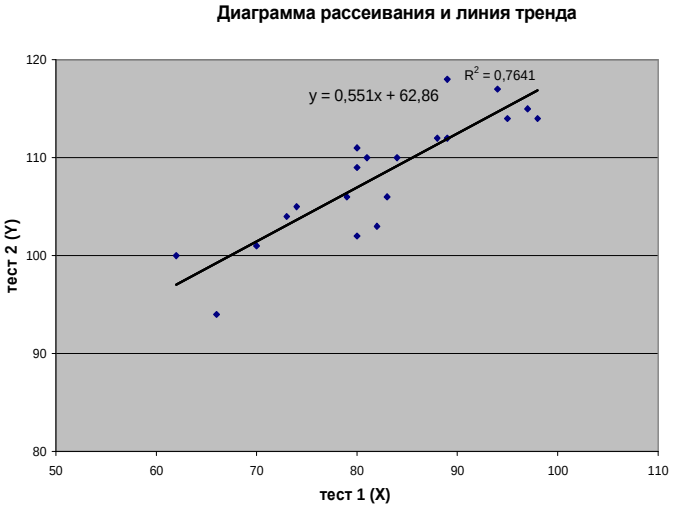


Рис. 4.1.6. Диаграмма рассеивания, линия тренда и уравнение регрессии

Задания для самостоятельного решения

1.В таблице приведены коэффициенты интеллекта IQ родителей и детей десяти семей. Определите, существует ли корреляционная зависимость между этими признаками.

IQ родителей	IQ детей
125	110
120	105
110	95
105	125
105	120
95	105
95	75
90	95
80	90
75	80

2.Постройте модель линейной регрессии между годом и численностью научных работников в Украине

Научные кадры и количество организаций Украины

Год	Количество организаций, которые ведут научные исследования и разработки	Численность научных сотрудников, чел.	Численность докторов наук в экономике Украины, чел.	Численность кандидатов наук в экономике Украины, чел.
1991	**1344**	**295010**	**8133**	**…**
1992	**1350**	**248455**	**8797**	**…**
1993	**1406**	**222127**	**9224**	**…**
1994	**1463**	**207436**	**9441**	**…**
1995	**1453**	**179799**	**9759**	**57610**
1996	**1435**	**160103**	**9974**	**58132**
1997	**1450**	**142532**	**10322**	**59332**
1998	**1518**	**134413**	**10446**	**59703**
1999	**1506**	**126045**	**10233**	**59547**
2000	**1490**	**120773**	**10339**	**58741**
2001	**1479**	**113341**	**10603**	**60647**
2002	**1477**	**107447**	**11008**	**62673**
2003	**1487**	**104841**	**11259**	**64372**
2004	**1505**	**106603**	**11573**	**65839**
2005	**1510**	**105512**	**12014**	**68291**
2006	**1452**	**100245**	**12488**	**71893**
2007	**1404**	**96820**	**12845**	**74191**
2008	**1378**	**94138**	**13423**	**77763**
2009	**1340**	**92403**	**13866**	**81169**
2010	**1303**	**89534**	**14418**	**84000**
2011	**1255**	**84969**	**14895**	**84979**
2012	**1208**			

Эти данные можно найти на сайте Госкомстата Украины: http://www.ukrstat.gov.ua/ . Постройте также экспоненциальную и квадратичную модель. Сравните их (рис.4.1.7)

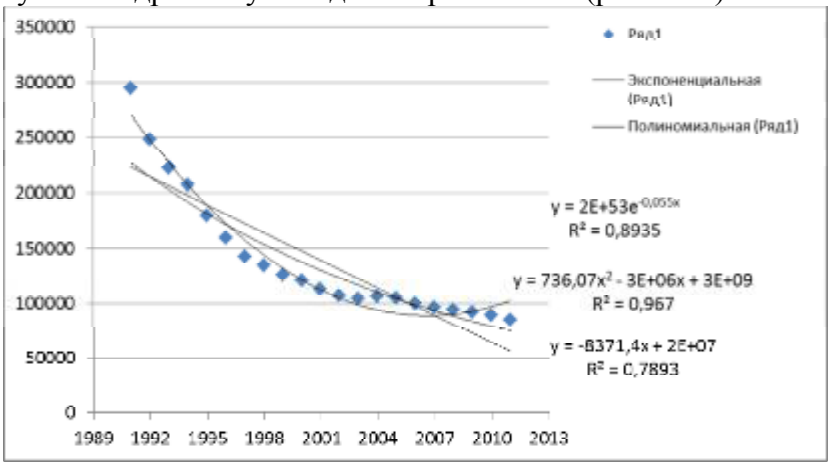


Рис.4.1.7.Линейная, экспоненциальная и квадратичная модель

Требования к отчету

Отчет о работе должен содержать:

– постановку задачи, исходные данные, коэффициент корреляции, уравнение линии предсказания, результаты предсказания;

– файл с результатами.

Контрольные вопросы

1. В чем состоит задача регрессионного анализа?
2. Что представляет собой диаграмма рассеивания?
3. Как получить уравнение линии предсказания?
4. Как осуществляются прогнозы с помощью уравнения линии предсказания (линии тренда)?
5. Как вычислить коэффициент корреляции Пирсона-Браве?
6. Что показывает величина достоверности аппроксимации?

Практическая работа 4.2.
Вычисление коэффициентов корреляции и прогноз с помощью линейной регрессии в R

Ход работы

Создадим для этих данных файл в Excel.

IQ родителей	IQ детей
125	110
120	105
110	95
105	125
105	120
95	105
95	75
90	95
80	90
75	80

```
> data411<-read.table("K:\\Rexample\\lab411.csv", sep=";",
dec=",", header=TRUE)
```

Получим уравнение регрессии:

```
 fit <- lm(data411$IQ.детей ~ data411$IQ.родит)
> summary(fit)

Call:
lm(formula = data411$IQ.детей ~ data411$IQ.родит)

Residuals:
   Min      1Q  Median      3Q     Max
-22.074  -6.370  -1.888   6.370  22.074
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        41.4894   29.0066   1.430   0.1905
data411$IQ.родит    0.5851    0.2867   2.041   0.0756 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.9 on 8 degrees of freedom
Multiple R-squared: 0.3423,     Adjusted R-squared: 0.2601
F-statistic: 4.165 on 1 and 8 DF,  p-value: 0.07559
```

Коэффициенты в уравнении резрессии представлены в столбце Estimate. Подставив их в уравнение регрессии, получим выражение: IQ.детей = 41.4894 + 0,5851*IQ.родит.

Кроме значений коэффициентов R показывает нам величины ошибок, или стандартного отклонения, для каждого коэффициента. Для проверки гипотезы о значимости коэффициентов, используется t-критерий Стьюдента. R вычисляет как саму величину t так и степень значимости нашей гипотезы Pr(>|t|). Так, в нашем случае величина 0.1905 означает, что мы на 100*(1-0.1905) = 81% уверены в том, что свободный член в нашем выражении отличен от нуля.

Далее мы можем проверить, насколько точно наша модель описывает данные. Для этого используются коэффициенты R^2 . Чем ближе величина этих значений к 1, тем лучше. 1 это идеальный результат, означающий, что модель на 100% описывает данные.

И, наконец, последнее, что мы можем проверить, это то, насколько предсказываемая величина зависит от предикторов. Для этого выдвигается нулевая гипотеза, что предсказываемая величина вообще не зависит от предикторов. Для этой гипотезы определяется p-значение. В нашем случае, оно получилось равным 0.07559. Т.е. мы можем быть уверены приблизительно на 92%, что предсказываемая величина действитель-

но зависит от предикторов. Обычно, имеет смысл смотреть на этот параметр в первую очередь, ведь он определяет, насколько вообще наша модель адекватна.

Контрольные вопросы

1. Как подготовить данные для регрессионного анализа?
2. С помощью какой функции строится линейная регрессионная модель?
3. Чем отличается функция cor.test от функции cor?

Практическая работа 4.3
Вычисление коэффициентов корреляции и прогноз с помощью линейной регрессии в SPSS для Windows

Цель работы*:** научиться использовать ***SPSS для вычисления коэффициентов корреляции и прогнозирования с использованием линейной регрессии.

Постановка задачи

Руководство швейной фабрики хотело бы иметь информацию о том, сколько платьев каждого фасона цвета и размера будет продано. С одной стороны, нельзя допускать перепроизводства, т.к. в этом случае придется продавать товар по заниженным ценам, с другой стороны, нехватка приводит к необходимости выпускать больше изделий, чем было намечено к данному сроку. Если иметь прогноз через 5 недель после поступления первого наименования в продажу, можно обеспечить изготовление необходимого числа продукции к намеченному сроку. В таблице 4.3.1 приведены данные о суммарном объеме проданных изделий каждого фасона, цвета и размера и соответствующие цифры за пять недель, полученные в начале исследования. Нужно представить графически данные и получить оценки наименьших квадратов в модели $y=b_0+b_1x$.

Таблица 4.3.1.

Продано в первые пять недель	Продано всего
235	392
122	190
34	74
196	307
121	200
127	185
177	291
125	191
159	314
75	140
140	235
127	188
89	142
139	253
70	96
62	96

Решение с помощью SPSS

Ход работы

1. Описать в редакторе две переменные «**Продано всего**» и «**Продано в первые пять недель**» и ввести данные таблицы 4.3.1.

2. Выбрать в меню ***Analyze... (Анализ), Regression...(Регрессия), Linear... (Линейная).*** Появится диалоговое окно Linear Regression (Линейная регрессия) (рис. 4.3.1).

3. Перенести переменную «**Продано Всего**» в поле для зависимых переменных и присвоить переменной «**Продано в первые 5 недель**» статус независимой переменной. Начать расчёт нажатием **ОК**.

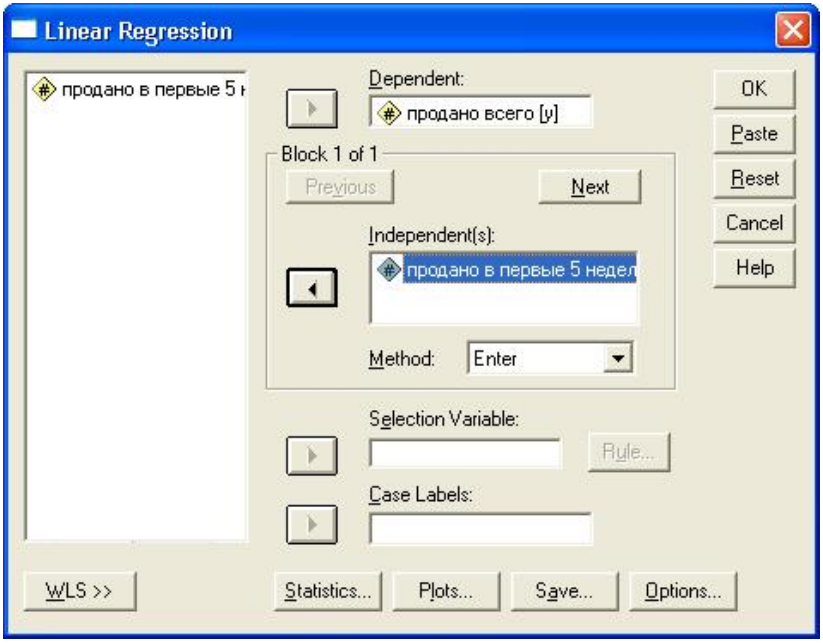


Рис. 4.3.1. Диалоговое окно «Linear Regression»

4. Вывод основных результатов выглядит так, как представлено на рис. 4.3.2 и 4.3.3.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,974[a]	,949	,946	20,666

a. Predictors: (Constant), продано в первые 5 недель

Рис. 4.3.2. Вывод результатов. Суммарная модель

Суммарная модель регрессии (рис. 4.3.2) включает: значение коэффициента корреляции Пирсона R=0,974; значение коэффициента детерминации R Square=0,949; значение уточненного коэффициента детерминации и стандартную ошибку оценки.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,827	13,821		-,132	,897
	продано в первые 5 недель	1,663	,103	,974	16,202	,000

a. Dependent Variable: продано всего

Рис. 4.3.3. Вывод результатов. Коэффициенты уравнения регрессии

В таблице «Коэффициенты» (см. рис. 4.3.3) окна вывода приводятся значения коэффициентов уравнения регрессии, их стандартной ошибки, стандартизованные коэффициенты, а также значение статистики Стьюдента t и ее значимость Sig. Коэффициент является значимым, если Sig<0,05.

Полученное уравнение регрессии: y=1,663x – 1,827.

5. Построить линию регрессии. Для этого необходимо проделать следующие шаги:

- Построить диаграмму рассеивания.
- Щёлкнуть дважды на этом графике, чтобы перенести его в редактор диаграмм.
- Выбрать в редакторе диаграмм меню ***Chart... (Диаграмма) Options... (Опции)***
- ***Откроется*** диалоговое окно Scatterplot Options (Опции для диаграммы рассеяния) (см. рис. 4.3.4).

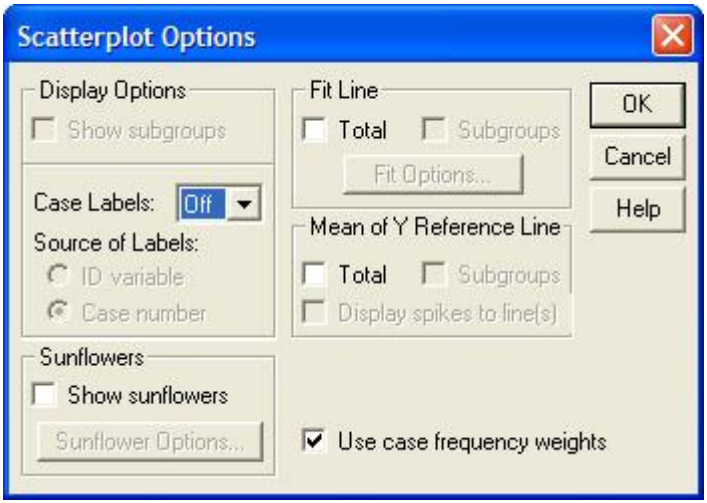


Рис. 4.3.4. Диалоговое окно «Scatterplot Options» (Опции для диаграммы рассеяния**)**

- В рубрике ***Fit Line*** (Приближенная кривая) поставить флажок напротив опции ***Total*** (Целиком для всего файла данных) и щёлкнуть на кнопке ***Fit Options*** (Опции для приближения). Откроется диалоговое окно ***Scatterplot Options: Fit Line*** (Опции для диаграммы рассеяния: приближенная кривая) (см. рис. 4.3.5).

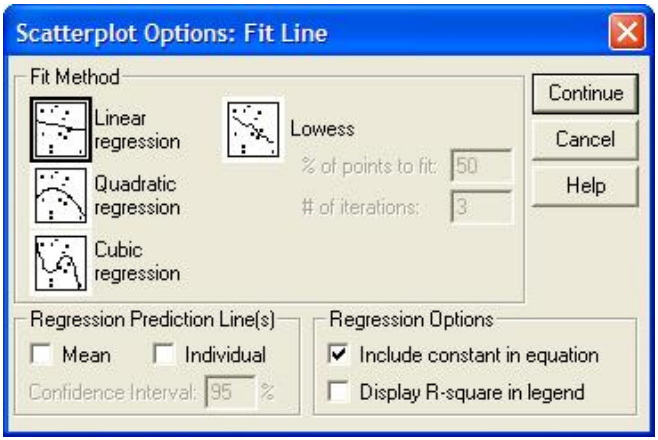


Рис. 4.3.5.Диалоговое окно «Scatterplot Options: Fit Line»

o Подтвердить предварительную установку **Linear Regression** (Линейная регрессия) щелчком **Continue** (Далее) и затем на **ОК**.

o Закрыть редактор диаграмм и щёлкнуть один раз где-нибудь вне графика. В окне просмотра появится диаграмма рассеивания с линией регрессии (см. рис. 4.3.6).

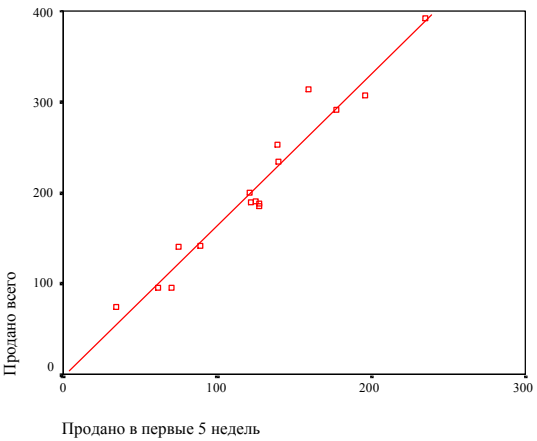


Рис. 4.3.6. Диаграмма рассеивания с линией регрессии

Требования к отчету

Отчет о работе должен содержать:

– постановку задачи, исходные данные, коэффициент корреляции, уравнение линии предсказания, результаты предсказания;

– файл с результатами.

Контрольные вопросы

1. Каким образом вы описали переменную «Продано в первые 5 недель» в редакторе данных?
2. Описать элементы диалогового окна «**Scatterplot Options**».
3. Описать элементы диалогового окна «**Scatterplot Options: Fit Line**».
4. Как получить уравнение линии предсказания в SPSS?
5. Дайте интерпретацию полученным результатам.

5. ПРОВЕРКА ГИПОТЕЗ

Гипотеза – это научно обоснованное предположение о структуре изучаемых объектов, о характере элементов и связей, образующих эти объекты, о механизмах их функционирования и развития.

В ходе исследования могут быть выдвинуты структурные либо объяснительные гипотезы.

Структурными называются гипотезы о структуре изучаемых объектов.

Пример структурной гипотезы: наиболее важными факторами выбора покупателя являются: цена, внешний вид, качество товара.

Объяснительными называются гипотезы о причинно-следственных связях в изучаемых объектах.

Примеры объяснительных гипотез: 1) самообразовательная деятельность будущих менеджеров способствует повышению уровня их профессиональной подготовленности; 3) увеличение затрат на рекламу не повлияет на доходы фирмы.

К научным гипотезам предъявляются два основных требования:

1. Гипотеза должна быть непротиворечивой (формулировка гипотезы должна подвергаться логическому анализу, устанавливающему ее непротиворечивость).
2. Гипотеза должна быть доступна проверке в ходе исследования (иначе нет смысла ее выдвигать, ведь проверить, истинна гипотеза или нет, будет невозможно).

Следует различать **научные** и **статистические** гипотезы.

Научная гипотеза – это, как уже отмечалось, разумное, обоснованное и развитое предположение о структуре изучаемого объекта или причинно-следственных связях в нем, о возможных путях решения той или иной проблемы.

Научная гипотеза формулируется как теорема и проверяется в ходе эксперимента – опыт отвечает на вопрос, является

гипотеза истинной или нет. Грамотная формулировка научной гипотезы – это подлинно творческий акт и важнейший этап исследования.

В экономике, психологии, педагогике, социологии, растениеводстве и многих других науках часто выдвигают гипотезы, которые можно и нужно проверять статистически, т.е. опираясь на результаты измерений в выборке.

Под **статистической гипотезой** понимают всякое высказывание о генеральной совокупности, проверяемое по выборке. При этом под **генеральной совокупностью** понимается тот объект, на который будут распространяться выводы по итогам исследования. Под **выборочной совокупностью** понимается сформированная в соответствии с определенными правилами (сформулированными в теории выборочного метода) часть генеральной совокупности, на которой непосредственно проводится исследование.

Статистическая гипотеза представляет собой утверждение относительно неизвестного параметра (неизвестных параметров) генеральной совокупности.

Статистическую гипотезу принято обозначать буквой ***H***.

Примеры статистических гипотез:

1. Утверждение $H: \mu=125$ – статистическая гипотеза, которая гласит, что неизвестное среднее конкретной совокупности=125.

Такое утверждение либо справедливо, либо ошибочно.

2. $H: \sigma^2_1 = \sigma^2_2$ – статистическая гипотеза, утверждающая, что дисперсии 1-й и 2-й совокупностей равны.

3. $H: \mu_1 = \mu_2 = \mu_3$ – статистическая гипотеза, состоящая в том, что средние значения трех совокупностей равны**.**

История проверки статистических гипотез ведет начало с XYIII века. Первый известный пример испытания статистической гипотезы – в работе Дж.Арбутнота (1667 – 1735 гг.), датированной 1710-м годом, «Доводы в пользу божественных

пророчеств, выведенные на основе постоянных и систематических наблюдений над рождением обоих полов».

Отметив, что записи на протяжении 82-х лет свидетельствуют о большем числе родившихся мальчиков, Арбутнот показал, что эти данные опровергают гипотезу о том, что рождения мальчиков и девочек равновероятны. Ибо, если вероятность рождения мальчика точно=1/2, то вероятность того, что за 82 года родилось больше мальчиков, чем девочек, была бы бесконечно мала. Арбутнот пришел в выводу, что большая доля рождения мужчин – результат вмешательства Провидения.

Грамотное формулирование статистических гипотез систематизирует предположения исследователя и представляет их в четком и лаконичном виде. Благодаря такой гипотезе «исследователь не теряет путеводной нити в процессе расчетов и ему легко понять после их окончания, что, собственно, он обнаружил».

Проверка гипотез является одним из направлений в теории **статистического вывода** (см. рис. 5.1).

Статистический вывод – это утверждение о параметрах генеральной совокупности на основании изучения выборочной совокупности.

Числовые характеристики, описывающие генеральную совокупность, называют параметрами и обозначаются буквами греческого алфавита. Те же характеристики для выборки называются статистиками и обозначаются буквами латинского алфавита.

Например:

μ - среднее в генеральной совокупности, x – среднее в выборке;

σ^2 - дисперсия в генеральной совокупности, S^2 – дисперсия в выборке.

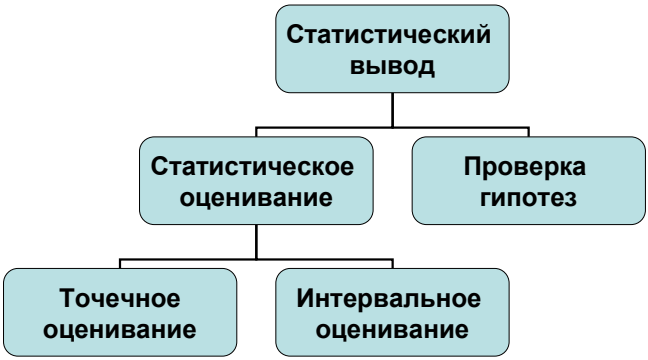


Рис. 5.1. Направления в статистическом выводе

Статистическое оценивание заключается в том, что исследователь по выборке ищет показатель, наиболее близкий к оцениваемому параметру (*точечное оценивание*), или интервал, в границах которого с большой вероятностью лежит этот параметр (*интервальное оценивание*).

Проверка гипотез состоит в том, что исследователь заранее формулирует некоторое утверждение о параметрах ГС (гипотезу), затем оценивает степень соответствия результатов, полученных в выборочном исследовании, сформулированной гипотезе и принимает решение об истинности или ложности гипотезы.

Статистические гипотезы можно разделить на: **гипотезы о законах распределения** (***например***: производительность труда рабочих, выполняющих одинаковую работу в одинаковых условиях, имеет нормальное распределение) и **гипотезы о параметрах распределения**, т.е. о средних, дисперсиях, коэффициентах корреляции, долях признаков и т.п. (***например***: 1) средние размеры деталей, производимых на однотипных параллельно рабо-

тающих станках, не различаются между собой; 2) средняя успеваемость студентов 1 курса специальности «Социология» не отличается от средней успеваемости студентов 1 курса специальности «Маркетинг») (см. рис. 5.2).

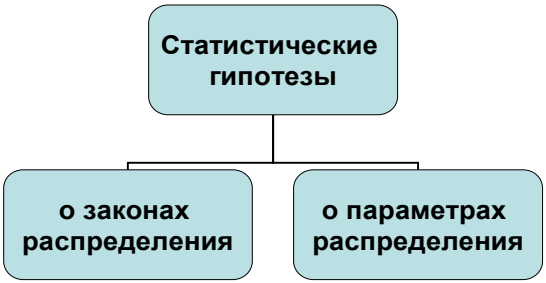


Рис. 5.2. Виды статистических гипотез в зависимости от их содержания

Кроме того, статистические гипотезы подразделяются на нулевые и альтернативные, направленные и ненаправленные (см. рис. 5.3).

Нулевая гипотеза – это гипотеза об отсутствии различий. Оно обозначается как H_0 и называется нулевой потому, что содержит в своей формулировке число нуль:

$$H_0 : X_1 - X_2 = 0$$

где X_1 , X_2 *– сопоставляемые значения признаков.*

Нулевая гипотеза опровергается, если различия окажутся значимыми.

Альтернативная гипотеза – это гипотеза о значимости различий. Она обозначается как H_1.

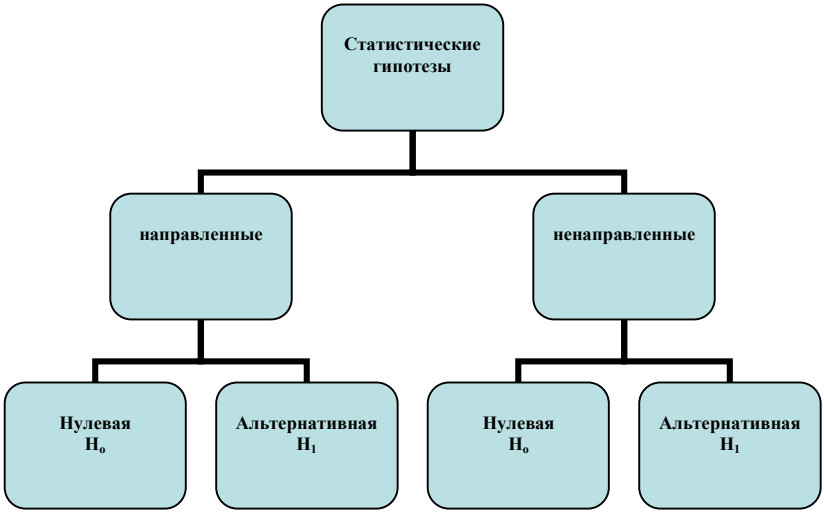


Рис. 5.3. Виды статистических гипотез в зависимости от их направленности и содержания

И нулевая, и альтернативная гипотезы могут быть направленными и ненаправленными.

Направленные гипотезы:

H_0: X_1 не превышает X_2

H_1: X_1 превышает X_2

Ненаправленные гипотезы:

H_0: X_1 не отличается от X_2

H_1: X_1 отличается от X_2

Если исследователь предполагает, что в группе, члены которой участвовали в специальном тренинге, индивидуальные значения испытуемых по какому-либо признаку, например, по инициативности, выше, а в группе, члены которой не участвова-

ли в тренинге, – ниже, то для проверки значимости этих различий необходимо сформулировать направленную гипотезу.

Т.е., если мы хотим проверить, действительно ли в группе А под влиянием определенных экспериментальных воздействий произошли более выраженные изменения, чем в группе В, мы должны сформулировать направленную статистическую гипотезу.

Если мы хотим проверить, различаются ли формы распределения изучаемого признака в группах А и В, нам необходимо сформулировать ненаправленную статистическую гипотезу.

Статистические гипотезы проверяются путем расчета **статистических критериев**.

Статистический критерий – это решающее правило, обеспечивающее надежное поведение, то есть принятие истинной и отклонение ложной гипотезы с высокой вероятностью.

Статистические критерии обозначают также метод расчета определенного числа и само это число.

По соотношению эмпирического и критического значений критерия мы можем судить о том, подтверждается или опровергается нулевая гипотеза. Например, если $\chi^2_{эмп} > \chi^2_{крит}$, H_0 отвергается.

В большинстве случаев для того, чтобы мы признали различия значимыми, необходимо, чтобы эмпирическое значение критерия превышало критическое, хотя есть критерии (например, критерий Манна-Уитни, критерий знаков, Т-критерий Уилкоксона), в которых мы должны придерживаться противоположного правила.

Алгоритм проверки статистических гипотез

1. Формулирование нулевой и альтернативной гипотезы.
2. Выбор подходящего статистического критерия, назовем его k.

3. Расчет по данным выборки эмпирического значения $k_{эмп}$.

4. На основании объема выборки, уровня значимости, числа степеней свободы определение критического значения критерия $k_{крит}$.

5. Сравнение эмпирического и критического значений критерия.

6. Если $k_{эмп} > k_{крит}$, то нулевая гипотеза отвергается (Исключения: критерий Манна-Уитни, критерий знаков, критерий Т-Уилкоксона).

Статистические критерии делятся на **параметрические** и **непараметрические**.

Параметрические статистические критерии включают в формулу расчета параметры распределения, т.е. средние и дисперсии (***t***-критерий Стьюдента, ***F***-критерий и др.).

Непараметрические статистические критерии не включают в формулу расчета параметры распределения и основаны на оперировании частотами или рангами (критерий Розенбаума, критерий Уилкоксона и др.).

Мощность статистического критерия – это его способность выявлять различия, если они есть.

Уровни статистической значимости

Уровень значимости – это вероятность того, что мы сочли случайные различия существенными, достоверными.

Так, ***например***, если мы утверждаем, что различия существенны (или достоверны) на 5%-ном уровне значимости (или при $p \leq 0{,}05$), это означает, что вероятность того, что они все-таки незначимы (недостоверны, несущественны) – не более 5%.

Низшим уровнем статистической значимости принято считать 5% (или 0,05); достаточным – 1% (0,01); высшим – 0,1% (0,001).

Ошибки при проверке гипотез

При проверке любой статистической гипотезы решение исследователя никогда не принимается со стопроцентной уверенностью – всегда имеется риск принятия неправильного решения.

Сущность проверки статистической гипотезы заключается в том, что эта проверка является средством контроля и оценки этого риска.

Ошибка, состоящая в том, что мы отклонили нулевую гипотезу, в то время как она верна, **называется ошибкой I рода.**

Вероятность такой ошибки обычно обозначается буквой α**;** тогда вероятность правильного решения – $(1 - \alpha)$. Таким образом, чем меньше вероятность ошибки (α), тем больше вероятность правильного решения.

Ошибка, состоящая в том, что мы приняли нулевую гипотезу, в то время, как она неверна, **называется ошибкой II рода.**

Вероятность такой ошибки обозначается буквой β.

($1 - \beta$) – это фактически мощность статистического критерия, его способность не допустить ошибку II рода.

Выбор статистического критерия для проверки гипотезы

Выбирая тот или иной статистический критерий для проверки статистических гипотез, нужно руководствоваться следующим:

– какова сама гипотеза, а соответственно, и какой тип статистической задачи предстоит решать;

– в каких шкалах осуществлялись измерения данных;

– каковы размеры выборки (выборок);

– если исследование опирается на более чем одну выборку, то можно ли применять выбранный критерий к неравным по объему выборкам;

– какова мощность критерия;

– какими возможностями располагает исследователь для расчета критерия (можно ли рассчитать критерий «вручную» или для его расчета необходимы определенные программные средства, имеются ли в распоряжении исследователя такие средства) и т.п.

Основные случаи проверки гипотез

Основные случаи проверки гипотез о параметрах генеральной совокупности:

– гипотезы о средних (рис. 5.4);
– гипотезы о дисперсиях (рис. 5.5);
– гипотезы о коэффициентах корреляции (рис. 5.6);
– гипотезы о долях признака (рис. 5.7);
– гипотезы о независимости признаков в корреляционной таблице.

Гипотезы о средних можно разделить на гипотезы о равенстве среднего определенному значению и гипотезы о значимости различия между средними двух совокупностей (см. рис. 5.4).

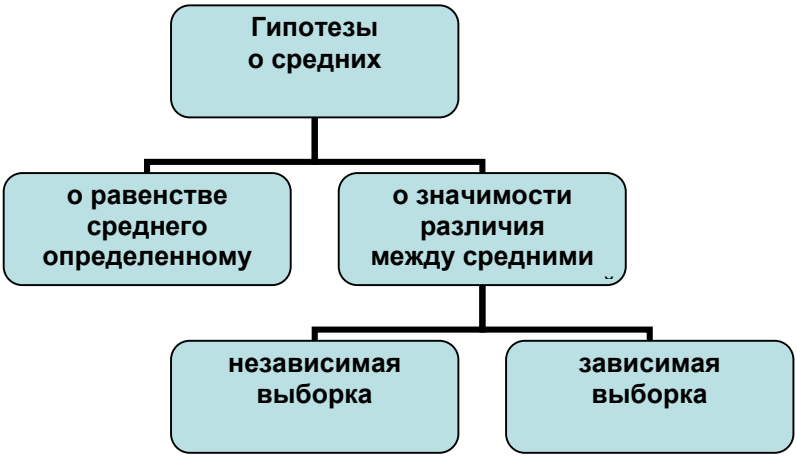
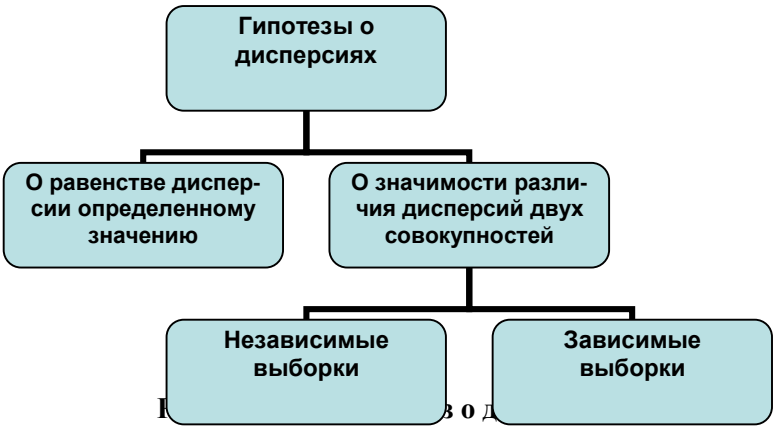


Рис. 5.4. Виды гипотез о средних значениях

На рис. 5.5 представлены виды гипотез о дисперсиях в генеральной совокупности.



На рис. 5.6 представлены виды гипотез о коэффициентах корреляции.

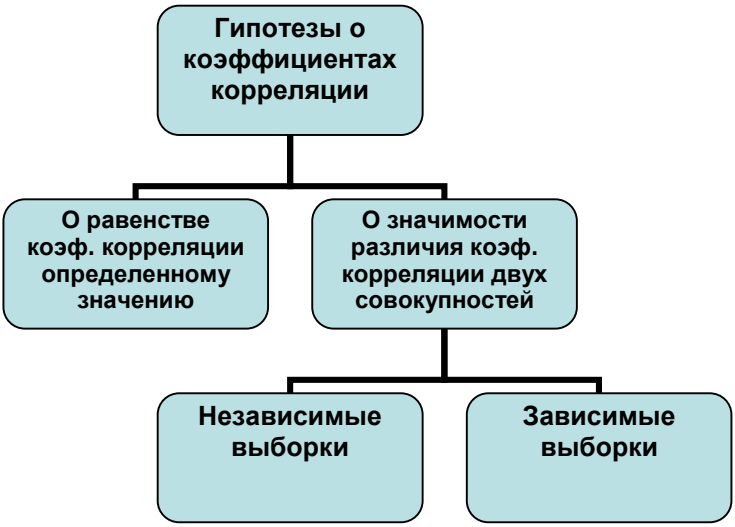


Рис. 5.6. Виды гипотез о коэффициентах корреляции

На рис. 5.7 представлены виды гипотез о долях признака.

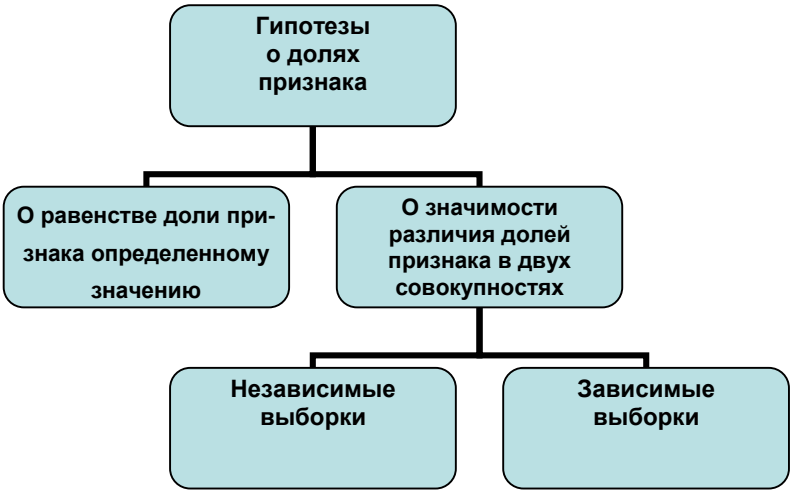


Рис. 5.7. Виды гипотез о долях признака

Сравнение средних

К наиболее часто применяемым методам статистического анализа относится сравнение средних значений двух выборок.

При сравнении средних значений выборок предполагается, что обе выборки подчинены нормальному распределению.

Если это не так, то используются **непараметрические тесты** (подробнее о них – в работах Дж.Полларда и Р.Руниона).

При сравнении средних значений выборок выделяют 4 тестовые ситуации:

Сравнение двух независимых выборок	→	***T***-тест Стьюдента для независимых выборок
Сравнение двух зависимых выборок	→	***T***-тест Стьюдента для зависимых выборок
Сравнение более двух независимых выборок	→	Однофакторный дисперсионный анализ
Сравнение более двух зависимых выборок	→	Однофакторный дисперсионный анализ с повторными измерениями

Практическая работа 5.1
Проверка гипотез о значимости различий средних в MS Excel

Цель работы: научиться использовать пакет статистического анализа MS Excel для проверки статистических гипотез.

Постановка задачи

Фирма наняла 40 человек для распространения своей продукции и организовала их обучение, разделив на две группы по 20 человек. В программу обучения первой группы входило знакомство с особенностями продукции, которую участникам группы предстояло распространять, в программу обучения второй группы дополнительно были введены основы психологии. В таблице 5.1.1 приведено число продаж, которые осуществили представители обеих групп в первый месяц работы. Свидетельствуют ли представленные в таблице 5.1.1. данные о различии в подготовленности к работе распространителей, обучавшихся по разным программам?

Таблица 5.1.1

Число продаж в зависимости от вида предварительного обучения

Число продаж	
Группа 1	**Группа 2**
16	9
17	7
5	5
7	6
7	7
8	8
8	4
6	6
6	6
8	8
8	8
15	12
8	4
10	4
10	10
13	5
4	4
4	3
8	8
9	3

Ход работы

1. Загрузить Excel. Ввести исходные данные согласно таблице 5.1.1.

2. Для каждого столбца данных вычислить среднее и дисперсию с помощью функций **СРЗНАЧ** и **ДИСП**. Посмотреть, равные ли в обеих группах получились средние и дисперсии.

3. Запустить пакет анализа: меню *«Данные», «Анализ данных», «Двухвыборочный t-тест с различными дисперсиями»* (см. рис. 5.1.1). Прочитать о нем справку и записать ее в тетрадь.

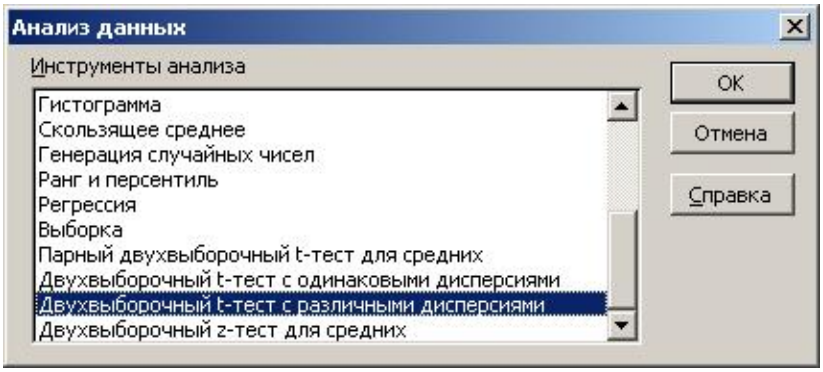


Рис. 5.1.1. Диалоговое окно «Анализ данных»

4. В окне «***Двухвыборочный t-тест…***» задать необходимые параметры: интервал переменной 1, интервал переменной 2, гипотетическая разность средних – 0, уровень значимости 0,05. Нажать кнопку *ОК* (рис. 5.1.2).

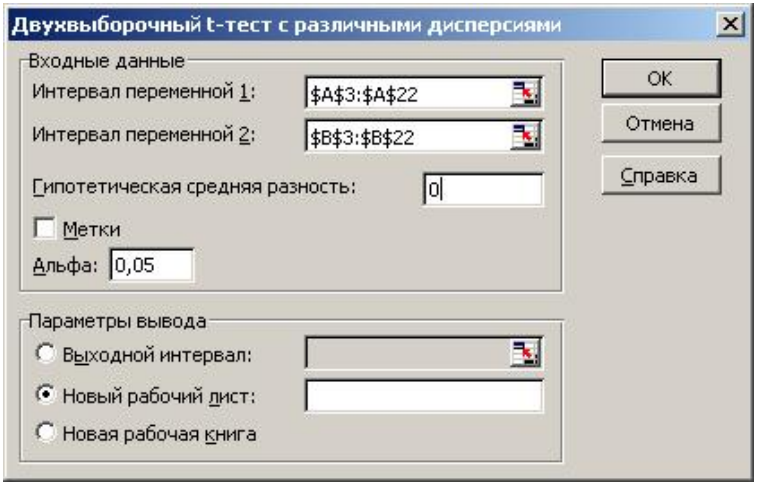


Рис. 5.1.2. Диалоговое окно «*Двухвыборочный t-тест…*»

5. Выписать таблицу с полученными результатами (см. рис. 5.1.3) и сделать вывод относительно значимости различий.

Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	8,85	6,35
Дисперсия	13,92368421	5,923684211
Наблюдения	20	20
Гипотетическая разность средних	0	
df	33	
t-статистика	2,509594424	
P(T<=t) одностороннее	0,008586329	
t критическое одностороннее	1,692360258	
P(T<=t) двухстороннее	0,017172658	
t критическое двухстороннее	2,034515287	

Рис. 5.1.3. Результаты расчета двухвыборочного t-теста с различными дисперсиями

6. Сохранить таблицу в личной папке.

Требования к отчету

Отчет о работе должен содержать:

– постановку задачи, исходные данные, описание двухвыборочного t-теста с различными дисперсиями;

– файл с результатами, выводы об истинности или ложности гипотезы.

Контрольные вопросы

1. Назовите виды статистических гипотез. Приведите примеры. Сформулируйте нулевую и альтернативную гипотезу для задачи.

2. Охарактеризуйте понятие «статистический критерий». В чем состоит различие между параметрическими и непараметрическими критериями?

3. Какие виды статистических критериев реализованы в Excel? Как они осуществляются?

Практическая работа 5.2
Проверка гипотезы о равенстве средних двух групп в R
Ход работы

1. Запустите R.

Определите два вектора, представляющие данные экспериментальной и контрольной группы с помощью функции scan.

Функция scan() принимает значения из файла или стандартного ввода (клавиатуры) и записывает их в переменные. Это значит, что после ее запуска, надо просто вводить значения переменной, при этом каждое значение вводится с новой строки. Ее удобно использовать для работы с небольшим числом данных, которые не записаны в файл, доступный для импорта в R.

Аргумент **what** задает тип значений, которые будут содержаться в переменной, а **n** — количество этих значений. Это не единственные доступные аргументы, об остальных можете узнать в справке по функции. Следует отметить, что число значений (n) можно не указывать, в этом случае, для прекращения ввода данных необходимо дважды нажать Enter.

2. Определите числовой вектор для первой группы:

```
> gr1<-scan(what=numeric(), n=20)
1: 16
2: 17
3: 5
4: 7
5: 7
6: 8
7: 8
8: 6
9: 6
10: 8
```

```
11: 8
12: 15
13: 8
14: 10
15: 10
16: 13
17: 4
18: 4
19: 8
20: 9
Read 20 items
```

3. Определить числовой вектор для данных второй группы:

```
gr2<-scan(what=numeric(), n=20)
1: 9
2: 7
3: 5
4: 6
5: 7
6: 8
7: 4
8: 6
9: 6
10: 8
11: 8
12: 12
13: 4
14: 4
15: 10
16: 5
17: 4
18: 3
19: 8
20: 3
Read 20 items
```

4. Проведите сравнение групп по критерию Стьюдента:

```
>t.test(gr1,gr2)

      Welch Two Sample t-test

data:  gr1 and gr2
t = 2.5096, df = 32.689, p-value = 0.01722

alternative hypothesis: true difference in means is not equal to
0

95 percent confidence interval:
 0.4725301 4.5274699

sample estimates:
mean of x mean of y
    8.85      6.35
```

Поясним результаты расчетов. Метод Уэлша не требует равенства дисперсий двух групп и использует другой способ вычисления степеней свободы. В среде R для функции t.test() по умолчанию используется метод Уэлша [http://voliadis.ru/r-two-samples].

После названия теста нам показано, какие группы были протестированы: группа1 и группа 2.

```
data:  gr1 and gr2
```

Далее показаны: t-статистика, число степеней свободы и точное значение p.

```
t = 2.5096,  df = 32.689,  p-value = 0.01722
```

Затем приведена формулировка альтернативной гипотезы:

```
alternative hypothesis: true difference in means is not equal to
0
```

После этого приведен доверительный 95% интервал. Для изменения доверительного интервала на, допустим, 99% доверительный интервал можно использовать аргумент conf.level=0.99.

```
95 percent confidence interval:
 0.4725301 4.5274699
```

И, наконец, приведены средние значения для каждой из групп:

```
sample estimates:
mean of x mean of y
    8.85      6.35
```

Полученные результаты расчетов свидетельствуют о значимых различиях в средних значениях между первой и второй группой испытуемых ($p<0,05$).

5. Объедините два созданных вектора в одну таблицу данных

```
> d <- data.frame(group1=gr1, group2=gr2)
> d
  group1 group2
1     16     9
2     17     7
3      5     5
```

4	7	6
5	7	7
6	8	8
7	8	4
8	6	6
9	6	6
10	8	8
11	8	8
12	15	12
13	8	4
14	10	4
15	10	10
16	13	5
17	4	4
18	4	3
19	8	8
20	9	3

6. Сохраните рабочее пространство: File, Save Workplace

Контрольные вопросы

1. Как вводятся данные с помощью функции scan?
2. Какие другие способы ввода данных вы знаете?
3. Каким образом производится сравнение средних значений групп по критерию Стьюдента?
4. Как интерпретируются его результаты?
5. Заполните следующую таблицу по результатам расчетов:

Мера	**Группа1**	**Группа2**
Среднее		
Наблюдения		
t-статистика		
Число степеней свободы		
p		
Уровень значимости		
Доверительный интервал		

Практическая работа 5.3
Проверка гипотез о значимости различий дисперсий в MS Excel

Цель работы: научиться использовать пакет статистического анализа MS Excel для проверки статистических гипотез.

Постановка задачи

В одном из исследований детям давались обычные арифметические задачи, а затем одной случайно выбранной половине учащихся сообщалось, что они не выдержали испытания, а остальным – обратное. Затем у каждого ребенка спрашивали, сколько секунд ему понадобилось бы для решения новой задачи. Экспериментатор вычислял разность между ожидаемым временем решения задачи (в сек.), которое назвал ребенок, и результатами ранее выполненного задания.

Проверяемая на уровне значимости 0,05 гипотеза состоит в том, что дисперсия совокупности детских оценок постоянна независимо от того, сообщалось ли детям о плохих результатах испытания или нет. Результаты эксперимента приведены в таблице 5.3.1. Группе 1 сообщалось о положительном результате, группе 2 – о неудаче. В каждой группе было по 12 детей.

Таблица 5.3.1

Результаты эксперимента

Разность между ожидаемым и реальным временем решения задачи, сек.	
Группа 1	Группа 2
8	40
9	10
10	30
11	20
7	38
6	22
5	30
9	30
12	45
15	26
9	25
6	25

Ход работы

1. Ввести исходные данные согласно таблице 5.3.1.
2. Для каждого столбца данных вычислить дисперсию с помощью функции **ДИСП**. Посмотреть, равные ли получились дисперсии в двух группах.
3. Запустить пакет анализа: меню ***«Данные», «Анализ данных», «Двухвыборочный F-тест для дисперсий*** (см. рис. 5.3.1).

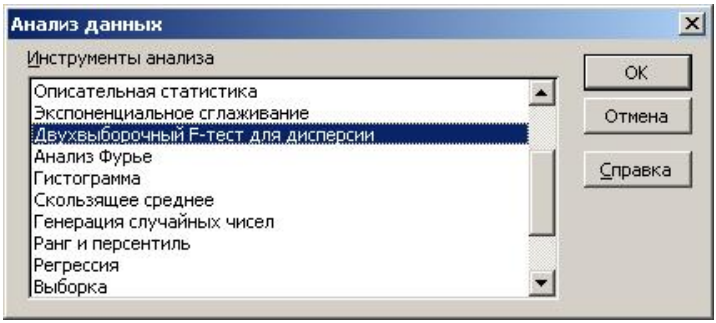


Рис. 5.3.1. Окно «Двухвыборочный F-тест для дисперсий»

4. Прочитать справку и записать ее в тетрадь.
5. В окне «***Двухвыборочный F-тест…***» задать необходимые параметры.
6. Выписать таблицу с полученными результатами (рис. 5.3.2).

Двухвыборочный F-тест для дисперсии		
	Группа 1	*Группа 2*
Среднее	8,916666667	28,41666667
Дисперсия	8,083333333	89,90151515
Наблюдения	12	12
Df	11	11
F	0,089913205	
F критическое одностороннее	0,354869911	

Рис. 5.3.2. Результаты расчета двухвыборочного *F*-теста с различными дисперсиями

7. Сделать вывод о подтверждении или опровержении гипотезы о равенстве дисперсий. (Указание: т.к. значение F-статистики 0,089 меньше нижнего критического значения 0,354, т.е. попадает в критическую область, то гипотеза о равенстве дисперсий двух групп данных отклоняется на уровне значимости 0.05)

8. Сохранить таблицу в личной папке.

Требования к отчету

Отчет о работе должен содержать:

– постановку задачи, исходные данные, описание двухвыборочного F-теста, результаты;

– выводы об истинности или ложности гипотезы;

– файл с результатами.

Контрольные вопросы

1. Приведите примеры известных вам параметрических и непараметрических статистических критериев.

2. Сформулируйте нулевую и альтернативную гипотезу для задачи из данной лабораторной работы.

3. Какие виды статистических критериев реализованы в Excel?

Практическая работа 5.4

Проверка гипотезы о равенстве дисперсий двух групп в *R*

1. Введите данные в электронных таблицах Excel (см. рис. 5.4.1)

	A	B	C
1	**bal**	**group**	
2	8	kg	
3	9	kg	
4	10	kg	
5	11	kg	
6	7	kg	
7	6	kg	
8	5	kg	
9	9	kg	
10	12	kg	
11	15	kg	
12	9	kg	
13	6	kg	
14	40	eg	
15	10	eg	
16	30	eg	
17	20	eg	
18	38	eg	
19	22	eg	
20	30	eg	
21	30	eg	
22	45	eg	
23	26	eg	
24	25	eg	
25	25	eg	
26			
27			

lab52 | Лист2 | Лист3

Рис.5.4.1.Исходные данные

2. Сохраните в формате .CSV (разделители запятые)

3. Импортируйте в переменную data

```
data<-read.table("K:\\Rexample\\lab52.csv", sep=";", dec=",",
                                        header=TRUE)
> data
   bal group
1    8    kg
2    9    kg
3   10    kg
4   11    kg
5    7    kg
6    6    kg
7    5    kg
8    9    kg
9   12    kg
10  15    kg
11   9    kg
12   6    kg
13  40    eg
14  10    eg
15  30    eg
16  20    eg
17  38    eg
18  22    eg
19  30    eg
20  30    eg
21  45    eg
22  26    eg
23  25    eg
24  25    eg
```

Функция read.table() импортирует таблицу из файла в переменную data. Тип переменной data.frame. Первый аргумент — это имя файла, если файл лежит в рабочей директо-

рии, то полный путь указывать необязательно. Параметр sep задает разделитель полей, dec — указывает разделитель в десятичных дробях. Этот аргумент следует задавать каждый раз, когда десятичный разделитель в исходном файле не является точкой, как это принято в среде R. И последним указан аргумент header. Если задать ему значение TRUE, то первая строка таблицы будет расцениваться как заголовок и будет использована для задания имен столбцам в переменной.

Для того, чтобы провести сравнение групп по «классическому» методу Стьюдента, необходимо проверить равенство дисперсий

```
> var.test(bal ~ group, data)

        F test to compare two variances

data:  bal by group

F = 11.1218, num df = 11, denom df = 11, p-value = 0.0003821

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:
  3.201727 38.633914

sample estimates:
ratio of variances
          11.12184
```

Функция `var.test()` проверяет гипотезу равенства дисперсий, а точнее их того, что их соотношение равно 1. В аргументах задано имя переменной data и т. н. формула модели — это специальный синтаксис для задания статистических

моделей, он используется во многих функциях, в том числе и в тех, которые не применяются для работы со статистическими моделями. Параметры `var` и `class` — это имена столбцов, содержащих значения зависимой и независимой переменных соответственно. Поскольку `p-value =` 0.0003821, считаем, что дисперсии значимо различаются.

Контрольные вопросы

1. Опишите параметры функции read.table().
2. Как проверяется равенство дисперсий двух выборок в R?
3. Объясните результаты расчетов, полученных в данной работе.

Практическая работа 5.5
Проверка гипотез о равенстве средних двух независимых выборок в SPSS

Цель работы: научиться производить проверку гипотез о равенстве средних двух независимых выборок с помощью Excel и SPSS и интерпретировать результаты

Постановка задачи

50 студентов университета были случайно распределены по двум группам. 25 студентов изучали курс анализа данных по традиционной вузовской методике, другие 25 - дистанционно. В конце эксперимента был проведен тест на усвоение знаний. Результаты его приведены в таблице 5.5.1. Свидетельствуют ли результаты теста о различии в методике преподавания темы?

Таблица 5.5.1

Тест в группе 1	Тест в группе 2
6	6
7	7
5	5
7	7
7	7
8	8
8	4
6	6
6	6
8	8
8	8
15	12
8	4
10	4
10	10
13	5
4	4
4	3
8	8
9	3
8	8
8,25	3
4	3
6	3
8	8

Нулевая гипотеза: Средние баллы по результатам теста в двух группах не отличаются.

Решение с помощью Excel

Ход работы

1. Загрузить Excel. Ввести исходные данные согласно таблице 5.5.1.

2. Для каждого столбца данных вычислить среднее и дисперсию помощью функций СРЗНАЧ и ДИСП. Посмотреть равные ли получились дисперсии.

3. Запустить пакет анализа: меню «*Данные», «Анализ данных», «Двухвыборочный тест с различными дисперсиями».* Прочитать в нем справку и записать ее в тетрадь.

4. В окне «Двухвыборочный тест…» задать необходимые параметры: интервал переменной 1 (данные первой группы), интервал переменной 2 (данные второй группы), гипотетическая разность средних – 0, уровень значимости 0,05. Нажать кнопку *ОК*.

5. Выписать таблицу с полученными результатами (см. рис.5.5.1)

Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	7,65	6
Дисперсия	6,5	5,916666667
Наблюдения	25	25
Гипотетическая разность средних	0	
Df	48	
t-статистика	2,341269661	
t критическое двухстороннее	2,01063358	

Рис. 5.5.1. Таблица с результатами

6. Сравнить эмпирическое значение t-критерия Стьюдента (t-статистику) с критическим двусторонним.

7. Сделать вывод о принятии или отвержении нулевой гипотезы (если эмпирическое значение больше критического, то нулевая гипотеза отвергается).

8. Сохранить таблицу в личной папке.

Решение с помошью SPSS

Ход работы

Проверка гипотезы с помощью SPSS для Windows включает следующие шаги:

1. Задание двух переменных: зависимой переменной **ball** (балл по тесту), независимой переменной-фактора **group** (группа) с градациями: «1 – первая», «2- вторая»).

2. Ввод исходных данных (представление данных в SPSS отличается от представления в электронных таблицах) и приводится на рис. 5.5.2.

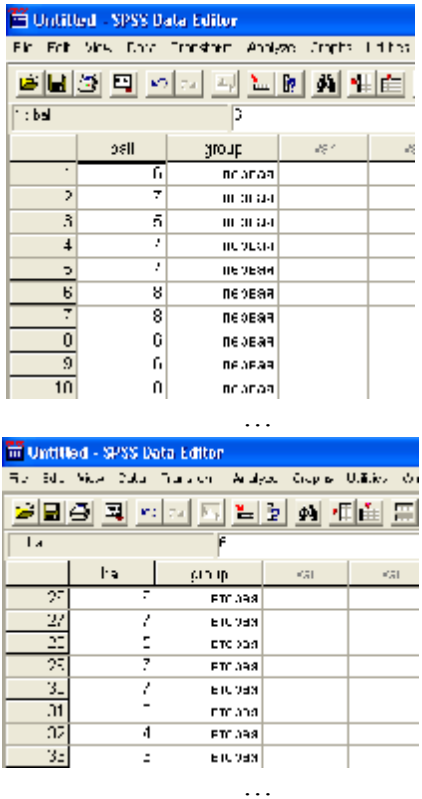


Рис.5.5.2. Представление исходных данных в SPSS для Windows

3. Выбор меню ***Analyze, Compare means, Independent-Samples T Test*** (для независимых выборок)

4. Перенос переменной **балл по тесту** в список тестируемых переменных, а переменную **группа** – в группирующую переменную (см. рис. 5.5.3)

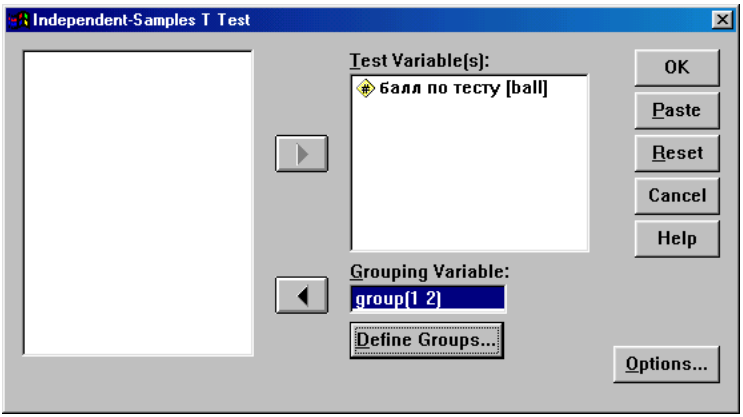


Рис. 5.5.3. Диалоговое окно Independent Samples T-Test

Щелчком на кнопке «**Определить группы**» задается кодировка, показанная на рис. 5.5.4.

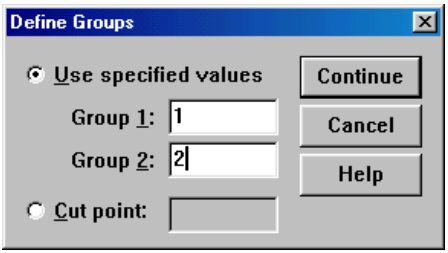


Рис. 5.5.4. Диалоговое окно «Define Groups»

5. Рассмотрение результатов в окне вывода (рис.5.5.5).

Group Statistics

	группа	N	Mean	Std. Deviation	Std. Error Mean
балл по тесту	первая	25	7,6500	2,5495	,5099
	вторая	25	6,0000	2,4324	,4865

Independent Samples Test

		ene's Test f Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. 2-tailed	Mean ifferenc	td. Erro ifferenc	5% Confidenc nterval of th Difference Lower	Upper
балл по т	Equal vari assumed	,261	,612	2,341	48	,023	1,6500	,7047	,2330	,0670
	Equal vari not assume			2,341	7,894	,023	1,6500	,7047	,2329	,0671

Рис. 5.5.5. Результаты сравнения средних двух независимых выборок

Для нашего примера значение критерия Стьюдента t*эмпирическое=2,341 и значимость p=0,023.* Нулевая гипотеза отклоняется, если значимость меньше 0,05.

Практическая работа 5.6
Проверка гипотез о равенстве средних двух зависимых выборок в SPSS

Цель работы: научиться производить проверку гипотез о равенстве средних двух зависимых выборок в Excel и SPSS и интерпретировать результаты.

Постановка задачи

Определите по данным следующей таблицы, значимо ли различаются средние процентных долей работающих женщин в 1968 и 1972 годах в 19 городах США.

Таблица 5.6.1.

Процент работающих женщин в 1968 и 1972 годах в 19 городах США

Город	1968	1972
N,Y,	0,42	0,45
L,A,	0,5	0,5
Chicago	0,52	0,52
Philadelphia	0,45	0,45
Detroit	0,43	0,46
San Francisco	0,55	0,55
Boston	0,45	0,6
Pitt,	0,34	0,49
St, Louis	0,45	0,35
Connecticut	0,54	0,55
Wash,, D,C,	0,42	0,52
Cinn,	0,51	0,53
Baltimore	0,49	0,57
Newark	0,54	0,53
Minn/St, Paul	0,5	0,59
Buffalo	0,58	0,64
Houston	0,49	0,5
Patterson	0,56	0,57
Dallas	0,63	0,64

Нулевая гипотеза: Средние процентных долей работающих женщин в 1968 и 1972 годах в 19 городах США значимо не различаются.

Альтернативная гипотеза: Средние процентных долей работающих женщин в 1968 и 1972 годах в 19 городах США значимо различаются.

Решение задачи с помощью электронных таблиц Excel

Ход работы

1. Загрузить Excel. Ввести исходные данные согласно таблице 5.6.1.

2. Для каждого столбца данных вычислить среднее и дисперсию помощью функций СРЗНАЧ и ДИСП.

3. Запустить пакет анализа: меню «*Данные», «Анализ данных», «Парный двухвыборочный t- тест для средних».* Прочитать в нем справку и записать ее в тетрадь.

4. В окне «*Парный двухвыборочный t- тест* …» задать необходимые параметры: интервал переменной 1, интервал переменной 2, гипотетическая разность средних – 0, уровень значимости 0.05. Нажать кнопку *ОК* (см. рис. 5.6.1).

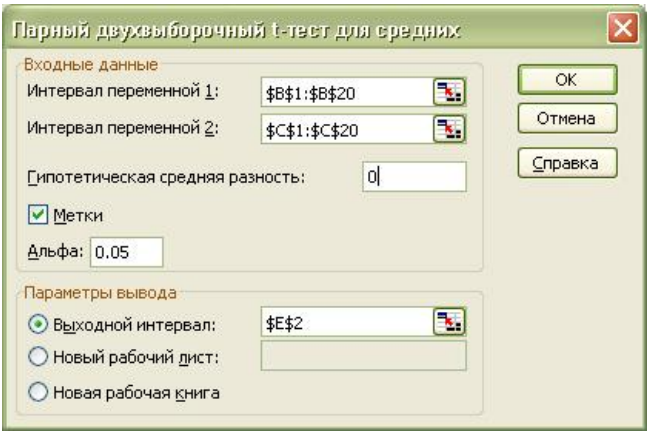


Рис. 5.6.1. Окно «Парный двухвыборочный t- тест для средних»

5. Выписать таблицу с полученными результатами (см. таблицу 5.6.2).

Таблица 5.6.2.

Результаты парного двухвыборочного t-теста для средних

Парный двухвыборочный t-тест для средних		
	1968	*1972*
Среднее	0.493158	0.526842105
Дисперсия	0.004623	0.005011696
Наблюдения	19	19
Корреляция Пирсона	0.630073	
Гипотетическая разность средних	0	
df	18	
t-статистика	-2.4577	
P(T<=t) одностороннее	0.012176	
t критическое одностороннее	1.734064	
P(T<=t) двухстороннее	0.024353	
t критическое двухстороннее	2.100922	

6. Сравнить эмпирическое значение t-статистики Стьюдента с критическим (двухсторонним). Сделать вывод о принятии или отвержении нулевой гипотезы.

7. Сохранить таблицу в личной папке.

Решение с помощью SPSS

Ход работы

1. Опишите три переменные, как показано на рис. 5.6.2, заполните данными таблицы 5.6.1 и сохраните в файле женщины.sav.

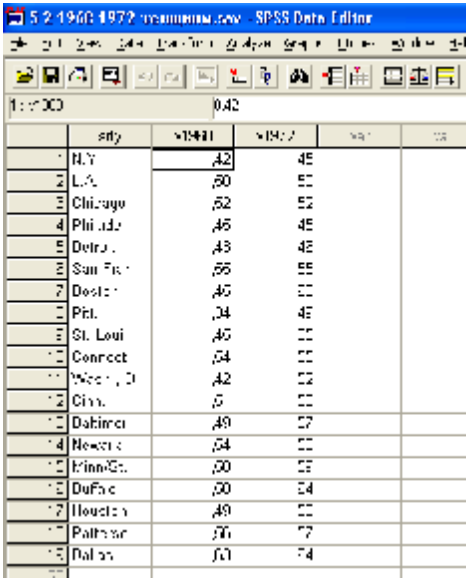


Рис.5.6.2. Содержание вкладки «Variable View» и «Data View» для задачи о процентной доле работающих женщин

2. Выберите в меню ***Analyze, Compare means, Paired Samples T-Test*** (тест для парных выборок).

3. Теперь в поле тестируемых переменных нужно выделить две необходимые переменные и эту пару перенести в поле для спаренных переменных. В нашем примере такими переменными являются v1968 и v1972 (рис. 5.6.3.)

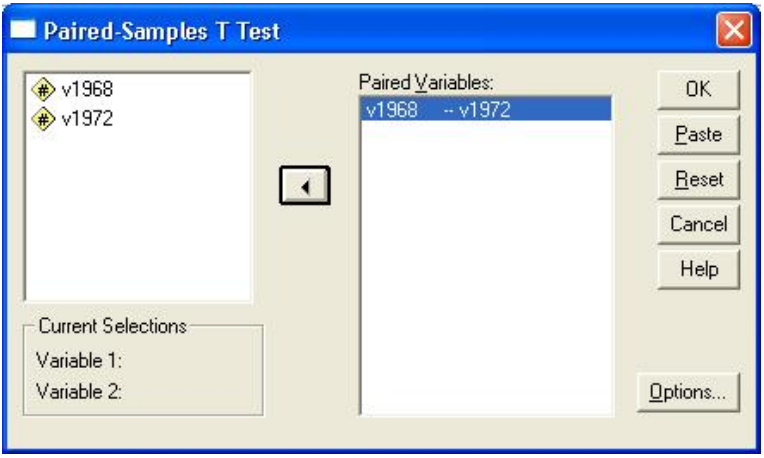


Рис. 5.6.3. Диалоговое окно «Paired-Samples T Test»

4. Запустите тест на исполнение нажатием клавиши ОК. В окне просмотра появятся результаты расчёта (см. рис. 5.6.4).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	V1968	,4932	19	,06799	,01560
	V1972	,5268	19	,07079	,01624

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
				95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1 V1968 - V1972	-,0337	,05974	,01371	-,0625	-,0049	-2,458	18	,024

Рис. 5.6.4. Результаты расчетов парного двухвыборочного теста для средних

Обратите внимание на эмпирическое значение статистики Стьюдента – 2,458 и ее значимость – 0,024. Нулевая гипотеза отвергается при значимости $p<0,05$.

Полученная в приведенном примере величина $p = 0,024$ свидетельствует о значимой разнице в средних.

Задания для самостоятельного решения

1.По результатам оценки размеров дебиторской задолженности, проведенной финансовым директором фирмы, были внесены изменения в ее кредитную политику. По истечении отчетного квартала было решено проанализировать, действительно ли изменения в кредитной политике фирмы оказали влияние на сокращение срока оплаты дебиторской задолженности. В таблице 5.6.3 представлены данные о продолжительности сбора дебиторской задолженности (в днях) при старой и новой кредитной политике фирмы.

Таблица 5.6.3.

Исходные данные

Тип кредитной политики	Продолжительность сбора дебиторской задолженности (дни)
Старая	39 42 28 35 38 30 30 29 36 34
Новая	28 31 29 30 24 37 27 22 33 31

2.Майерс [39] в 1976 г. провел интересный эксперимент по проверке эффективности индивидуального и группового тестирования. Испытуемые одной группы проверяли программу из 63 операторов, предназначенную для обработки строк, индивидуально, причем использовали набор тестов, составленных после изучения спецификации программы. Они не располагали листингом программы. Испытуемые второй группы располагали теми же средствами плюс листинг. Испытуемые третьей группы были разбиты на бригады из трех человек, и от них требовалось тестировать программы вручную методом проверки кода. Результаты эксперимента приведены в таблице 5.6.4.

Таблица 5.6.4.

Результаты эксперимента по тестированию программ

Характеристика	Индивидуальный просмотр + спецификация + терминал	Индивидуальный просмотр + спецификация + терминал +	Групповой просмотр + спецификации + листинг

		листинг	
Среднее число найденных ошибок	4,5	5,4	5,7
Дисперсия	4,8	5,5	3,0
Минимальное число ошибок	1	2	3
Максимальное число ошибок	7	9	9
Затраты на ошибку, чел.-мин	37	29	75

- Определите по критерию Стьюдента значимо ли различаются среднее число найденныхошибок в этих трех группах, если в каждой из групп было по 25 испытуемых
- Определите, значимо ли различаются дисперсии и средние затраты на ошибку трех групп?

Указание. При сравнении средних двух независимых выборок эмпирическое значение критерия Стьюдента вычисляется по формуле:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n1-1)\cdot S_1^2 + (n2-1)\cdot S_2^2}{n1+n2-2}\cdot\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

где

X1	Среднее первой выборки
X2	Среднее второй выборки
S1	Стандартное отклонение первой выборки
S2	Стандартное отклонение второй выборки
n1	Объем первой выборки
n2	Объем второй выборки

и сравнивается с критическим

$$t_{n1+n2-2}(1-(\alpha/2))$$

Если эмпирическое значение превышает критическое, то нулевая гипотеза о равенстве средних двух выборок отвергается на уровне значимости α.

Требования к отчету

Отчет о работе должен содержать: постановку задачи, исходные данные, файлы с результатами, выводы об истинности или ложности гипотезы.

Контрольные вопросы

1. Назовите виды статистических гипотез. Приведите примеры. Сформулируйте нулевую и альтернативную гипотезу для задачи.
2. Охарактеризуйте понятие «статистический критерий». В чем состоит различие между параметрическими и непараметрическими критериями?
3. Какие виды статистических критериев реализованы в Excel? Как они осуществляются?
4. Как осуществляется проверка гипотез о равенстве средних для двух независимых групп с помощью SPSS?
5. Объясните, что выводится в таблицах «Статистика групп» и «Тест для независимых выборок» в окне вывода?
6. Сравните с результатами, полученными в Excel.
7. Какие группы считаются зависимыми или парными?
8. Как осуществляется проверка гипотез о равенстве средних для двух зависимых групп с помощью SPSS? Каково эмпирическое значение критерия Стьюдента? Каковы границы доверительного интервала для разности средних? Попадает ли туда критическое значение?
9. Дайте содержательную интерпретацию результатам.

6. ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ – это один из методов анализа изменчивости признака под влиянием контролируемых факторов. Автором фундаментальной концепции дисперсионного анализа является Р.Фишер.

В англоязычной литературе термину «дисперсионный анализ» соответствует аббревиатура **ANOVA** (**An**alysis **o**f **Va**riance – анализ вариативности/дисперсии).

Для того, чтобы грамотно осуществить дисперсионный анализ, изучаемое явление нужно представить как систему переменных, среди которых следует выделить гипотетически (предположительно) независимую переменную (или контролируемый фактор) и гипотетически зависимую переменную (признак). Независимая переменная (контролируемый фактор) предположительно является причиной вариативности зависимой переменной, т.е. зависимая переменная предположительно изменяется под воздействием независимой переменной (контролируемого фактора). Такая модель – это довольно грубое упрощение реальности (т.к. на зависимую переменную чаще всего влияют и другие факторы, которые неизвестны исследователю или изучение влияния которых не входит в планы исследователя. В последнем случае влияние таких факторов стараются либо устранить либо уравновесить во всех выборках), но статистические методы, в частности, дисперсионный анализ, позволяют пренебречь этим упрощением и получить статистически достоверный результат анализа изменчивости зависимой переменной под влиянием контролируемого фактора.

Целью дисперсионного анализа является проверка значимости различия между средними с помощью сравнения (т.е. анализа) дисперсий. А именно, – разделение общей дисперсии на несколько источников, позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью.

Сущность дисперсионного анализа заключается в том, что из общей вариативности признака вычленяются три вида вариативности:

1) вариативность, обусловленная действием каждого из исследуемых факторов;

2) вариативность, обусловленная взаимодействием исследуемых факторов;

3) случайная вариативность, обусловленная действиями других факторов (неизвестных факторов или известных факторов, влияние которых не входит в задачи исследования).

Вариативность, обусловленная действием исследуемых факторов и их взаимодействием, соотносится со случайной вариативностью. Показателем этого соотношения является F-критерий Фишера.

$$F_{эмпA} = (\text{Вариативность, обусловленная фактором A}) / (\text{Случайная вариативность})$$

$$F_{эмпB} = (\text{Вариативность, обусловленная фактором B}) / (\text{Случайная вариативность})$$

$$F_{эмпAB} = (\text{Вариативность, обусловленная взаимодействием факторов A и B}) / (\text{Случайная вариативность})$$

Чем в большей степени вариативность признака обусловлена исследуемыми факторами или их взаимодействием, тем выше эмпирические значения F-критерия Фишера.

Нулевая гипотеза в дисперсионном анализе: средние величины исследуемого признака при всех градациях исследуемого фактора одинаковы. Например: средняя длительность пребывания посетителей в ресторане (исследуемый признак) одинакова в случае, если в зале звучит тихая, средняя или громкая музыка (громкость – предполагаемый фактор, который влияет на длительность пребывания посетителей в ресторане, а «тихая», «средняя» и «громкая» – градации этого фактора.

Альтернативная гипотеза: Средняя длительность пребывания посетителей в ресторане различается в зависимости от громкости звучащей в зале музыки.

При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой изменчивостью, должна быть близкой к оценке межгрупповой дисперсии.

Метод дисперсионного анализа становится незаменимым тогда, когда мы исследуем одновременное воздействие двух или более факторов, поскольку он позволяет выявить взаимодействие факторов в их влиянии на один и тот же признак.

Существует несколько видов дисперсионного анализа. Требуемый вариант выбирается с учетом числа факторов и имеющихся выборок из генеральной совокупности.

Однофакторный дисперсионный анализ служит для анализа дисперсии по данным двух или нескольких выборок. При анализе проверяется нулевая гипотеза о том, что каждый пример извлечен из одного и того же базового распределения вероятности. Альтернативная гипотеза: базовые распределения вероятности во всех выборках разные. Если имеется всего две выборки, применяют функцию ТТЕСТ. Для более, чем двух выборок не существует обобщения функции ТТЕСТ, и вместо этого можно воспользоваться моделью однофакторного дисперсионного анализа.

Двухфакторный дисперсионный анализ с повторениями применяется, если данные можно систематизировать по двум параметрам. Например, в исследовании, направленном на поиск лучших средств для похудения, добровольцам предлагались три разные диеты (A, B и C) и два варианта поведения – регулярные занятия фитнесом и отсутствие специализированных физических нагрузок. Таким образом, для каждой из 6 возможных пар условий {диета, физическая нагрузка} имеется набор наблюдений за снижением веса добровольцев. С помощью двухфакторного дисперсионного анализа с повторениями можно проверить следующие гипотезы:

1. Извлечены ли данные о снижении веса для различных диет из одной генеральной совокупности независимо от физической нагрузки.

2. Извлечены ли данные о снижении веса для различного характера физической нагрузки из одной генеральной совокупности независимо от типа диеты.

3. Извлечены ли 6 выборок, представляющих все пары значений { диета, физическая нагрузка }, используемые для оценки влияния различных типов диет (шаг 1) и типа физической нагрузки (шаг 2), из одной генеральной совокупности. Альтернативная гипотеза предполагает, что влияние конкретных пар { диета, физическая нагрузка } превышает влияние отдельно диеты и отдельно физической нагрузки.

Двухфакторный дисперсионный анализ без повторения полезен при классификации данных по двум измерениям, как и двухфакторный дисперсионный анализ с повторением. Однако при этом анализе предполагается только одно наблюдение для каждой пары. При этом анализе можно добавлять проверки в шаги 1 и 2 двухфакторного дисперсионного анализа с повторениями, но недостаточно данных для добавления проверок в шаг 3.

Дисперсионный анализ позволяет констатировать изменение признака, но при этом не указывает направления этих изменений. Чтобы получить наглядное представление о направлении изменений, необходимо специально графически представлять полученные данные по градации фактора.

Практическая работа 6.1
Однофакторный дисперсионный анализ

***Цель работы*:** научиться использовать пакет статистического анализа электронных таблиц Excel и возможности SPSS для Windows для однофакторного дисперсионного анализа.

Постановка задачи

Директор по маркетингу хочет установить, значимо ли различаются продажи в группах магазинах в зависимости от уровня используемой рекламы. Результаты продаж представлены в таблице 6.1.1.

Таблица 6.1.1

Результаты продаж в зависимости от уровня рекламы

	A	B	C	D
1	Номер недели	Уровень рекламы высокий	Уровень рекламы средний	Уровень рекламы низкий
2	1	8	7	4
3	2	7	8	5
4	3	9	5	3
5	4	5	4	6
6	5	6	6	2
7	6	8	7	4
8	Среднее			

Нулевая гипотеза:

Различия в продажах групп магазинов с разным уровнем рекламы, являются не более выраженными, чем случайные различия внутри каждой группы. (Другими словами: средние продаж в каждой из трех групп не различаются).

Решение с помощью электронных таблиц
Ход работы

1. Загрузить Excel.

2. Ввести данные. Рассчитать средние значения продаж для каждой группы магазинов.

3. Построить график для сравнения средних трех групп магазинов (см. рис. 6.1.1).

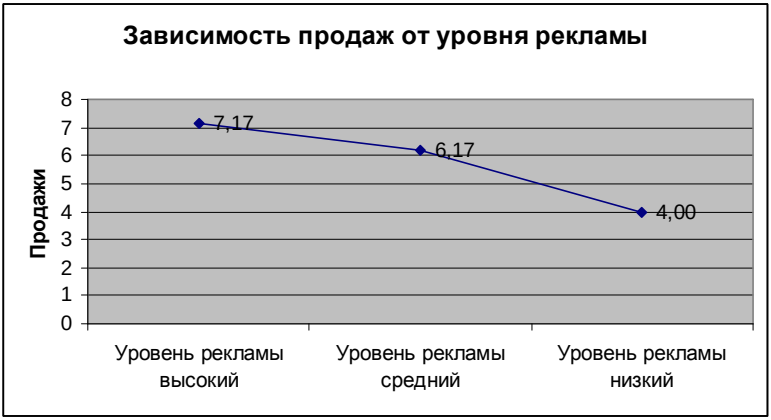


Рис. 6.1.1. График для сравнения средних

4. Выбрать меню ***Сервис***, ***Анализ данных***, ***Однофакторный ДА***.

5. В окне «***Однофакторный дисперсионный анализ***» (рис. 6.1.2) задать необходимые параметры.

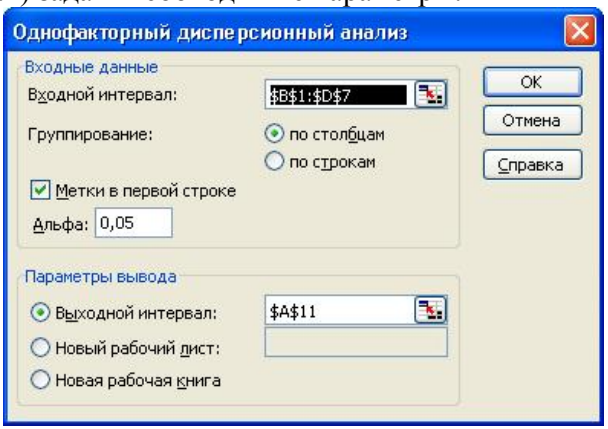


Рис. 6.1.2. Окно «Однофакторный дисперсионный анализ»

6. В окне результатов сравнить *F критическое и F эмпирическое* (см. рис. 6.1.3).

Если *F критическое* < *F эмпирического,* нулевая гипотеза отвергается.

Однофакторный дисперсионный анализ						
ИТОГИ						
Группы	***Счет***	***Сумма***	***Среднее***	***Дисперсия***		
Уровень рекламы высокий	**6**	**43**	**7,17**	**2,17**		
Уровень рекламы средний	**6**	**37**	**6,17**	**2,17**		
Уровень рекламы низкий	**6**	**24**	**4**	**2**		
Дисперсионный анализ						
Источник вариации	***SS***	***df***	***MS***	***F***	***P-Значение***	***F критическое***
Между группами	**31,44444444**	**2**	**15,72**	**7,447**	**0,005671839**	**3,682**
Внутри групп	**31,66666667**	**15**	**2,11**			
Итого	**63,11111111**	**17**				

Рис. 6.1.3. Окно результатов

Решение задачи в SPSS

1. Задать две переменные: Y – зависимая переменная – число продаж, F1 – независимая переменная – фактор (уровень рекламы) с градациями: 1 – низкий, 2 – средний, 3 – высокий.

2. Ввести исходные данные (представление данных в SPSS отличается от представления в электронных таблицах и приводится в таблице 6.1.2).

Таблица 6.1.2.

Представление исходных данных в SPSS для Windows

F1 (уровень рекламы)	Y (число продаж)
высокий	8
высокий	7
высокий	9
высокий	5
высокий	6
высокий	8
средний	7
средний	8
средний	5
средний	4
средний	6
средний	7
низкий	4
низкий	5
низкий	3
низкий	6
низкий	2
низкий	4

3. Выбрать меню ***Analyze, General Linear Model, Univariate***. Перенести переменную Y в список зависимых переменных, переменную F1 – в список независимых переменных (см. рис. 6.1.4).

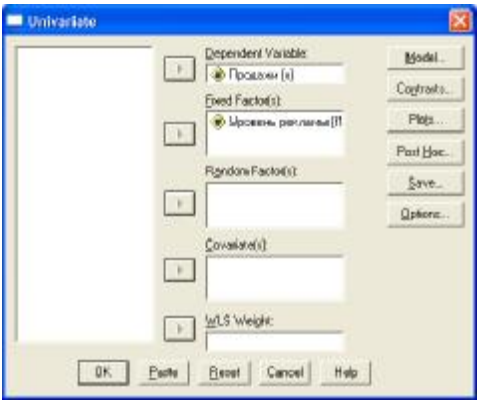


Рис. 6.1.4. Окно «Univariate»

4. Рассмотреть результаты в окне вывода (рис. 6.1.5).

Для нашего примера значение критерия Фишера *Fэмпирическое=7,447 и значимость p=0,006,* следовательно нулевая гипотеза отклоняется.

Tests of Between-Subjects Effects

Dependent Variable: Продажи

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	31,444[a]	2	15,722	7,447	,006
Intercept	600,889	1	600,889	284,632	,000
F1	31,444	2	15,722	7,447	,006
Error	31,667	15	2,111		
Total	664,000	18			
Corrected Total	63,111	17			

a. R Squared = ,498 (Adjusted R Squared = ,431)

Рис. 6.1.5. Окно вывода

5. Построить график для сравнения средних: ***Analyze, General Linear Model, Univariate,*** нажать кнопку ***Plot,*** появится диалоговое окно «***Profile Plots***» (см. рис. 6.1.6); перенести переменную *F1* в строку ***Horizontal Axis,*** нажать кнопку ***Add,*** затем ***Continue, OK***. На рис. 6.1.7. представлен график, который должен получиться в результате этих действий.

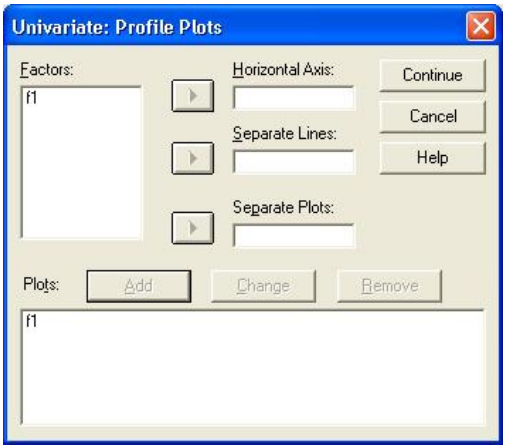


Рис. 6.1.6. Окно «Univariate: Profile Plots»

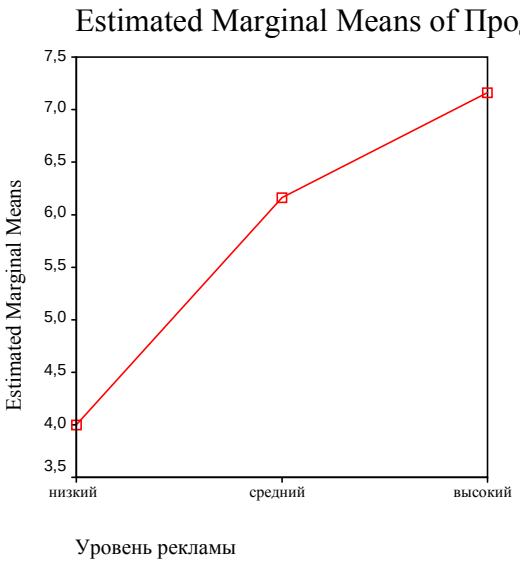


Рис. 6.1.7. График для сравнения средних

Задачи для самостоятельного решения.

1. Выполните проверку по F-критерию нулевой гипотезы на уровне 0,01 по представленным в табл. 6.1.3 данным, которые описывают весовые потери в килограммах испытуемыми, соблюдавшими различные диеты.

Таблица 6.1.3

Весовые потери испытуемых (кг), соблюдавших различные диеты

Диета 1	Диета 2	Диета 3	Диета 4
2,7	4,95	9,45	2,25
3,6	5,85	9	4,05
1,35	6,75	7,65	4,5
2,25		7,2	3,15
2,7			3,15

2. Для выяснения влияния денежного стимулирования на производительность труда программистов шести однородным группам из 5 человек были предложены задачи одинаковой трудности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличаются между собой величиной денежного вознаграждения за решаемую задачу (таблица 6.1.4).

Таблица 6.1.4.

Величина вознаграждения от меньшей к большей

Гр1	**Гр2**	**Гр3**	**Гр4**	**Гр5**	**Гр6**
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	2	16	19	20	24

Определите с помощью однофакторного дисперсионного анализа влияет ли величина вознаграждения на производительность труда программистов?

Требования к отчету

Отчет о работе должен содержать:

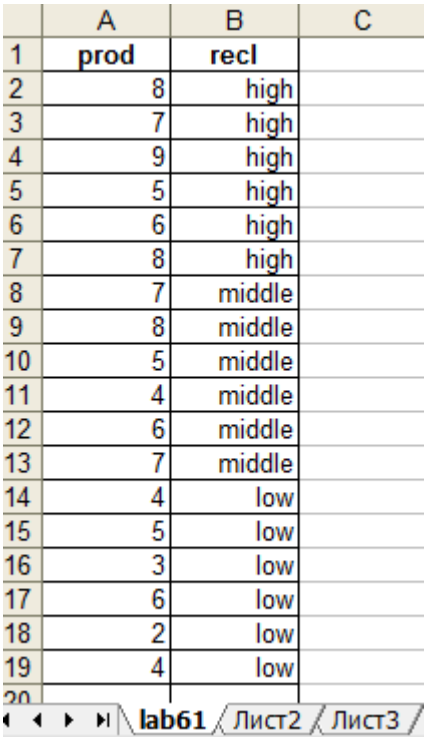
- постановку задачи, исходные данные, формулировку гипотезы, результаты, выводы об истинности или ложности гипотезы;
- файл с данными.

Контрольные вопросы

1. Назначение однофакторного дисперсионного анализа.
2. Как производится однофакторный дисперсионный анализ с помощью Excel?
3. Как производится однофакторный дисперсионный анализ с помощью SPSS?
4. Дайте содержательную интерпретацию полученным результатам.

Практическая работа *6.2*
Однофакторный дисперсионный анализ в R
Ход работы

1. Введите данные в электронных таблицах Excel, связывающие продажи с уровнем рекламы (см. рис.6.2.1)



	A	B	C
1	**prod**	**recl**	
2	8	high	
3	7	high	
4	9	high	
5	5	high	
6	6	high	
7	8	high	
8	7	middle	
9	8	middle	
10	5	middle	
11	4	middle	
12	6	middle	
13	7	middle	
14	4	low	
15	5	low	
16	3	low	
17	6	low	
18	2	low	
19	4	low	
20			

lab61 Лист2 Лист3

Рис. 6.2.1. Исходные данные

2. Сохраните в формате .CSV (разделители запятые) под именем lab61

3. Импортируйте в переменную data1 с помощью функции read.table

```
>data1<-read.table("K:\\Rexample\\lab61.csv", sep=";", dec=",",
header=TRUE)
> data1
   prod   recl
1     8   high
2     7   high
3     9   high
4     5   high
5     6   high
6     8   high
7     7 middle
8     8 middle
9     5 middle
10    4 middle
11    6 middle
12    7 middle
13    4    low
14    5    low
15    3    low
16    6    low
17    2    low
18    4    low
```

4. Проведите однофакторный дисперсионный анализ:

```
> anova(lm(prod~recl, data1))
Analysis of Variance Table

Response: prod
           Df       Sum         Sq Mean Sq   F value      Pr(>F)
recl        2       31.444      15.7222        7.4474
0.005672 **
Residuals 15        31.667       2.1111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Функция anova() в качестве параметра принимает объект, являющийся линейной регрессионной моделью. Функция lm() как раз и отвечает за подгонку линейной модели. Первый параметр функции — это формула модели: слева от тильды - имя столбца с данными (в нашем случае prod), справа — фактор или группирующая переменная (уровень рекламы recl). Второй параметр — таблица данных для анализа (data1).

В нашем случае p=0.005672, поэтому нулевая гипотеза отвергается на уровне 0,01 (таким образом, продажи различаются в зависимости от уровня рекламы).

Постройте график, иллюстрирующий результаты дисперсионного анализа (см. рис.)

Для этого вначале, упорядочим значения переменной уровень рекламы:

```
>data1$recl.o<- ordered(data1$recl, levels=c("low", "middle", "high"))
> data1$recl.o
 [1] high   high   high   high   high   high   middle middle middle middle middle middle low    low    low
[16] low    low    low
Levels: low < middle < high
```

А теперь начертим диаграмму (рис.6.2.2):

```
>plot(data1$recl.o,data1$prod, xlab="уровень рекламы", ylab="продажи", main="зависимость продаж от уровня рекламы")
```

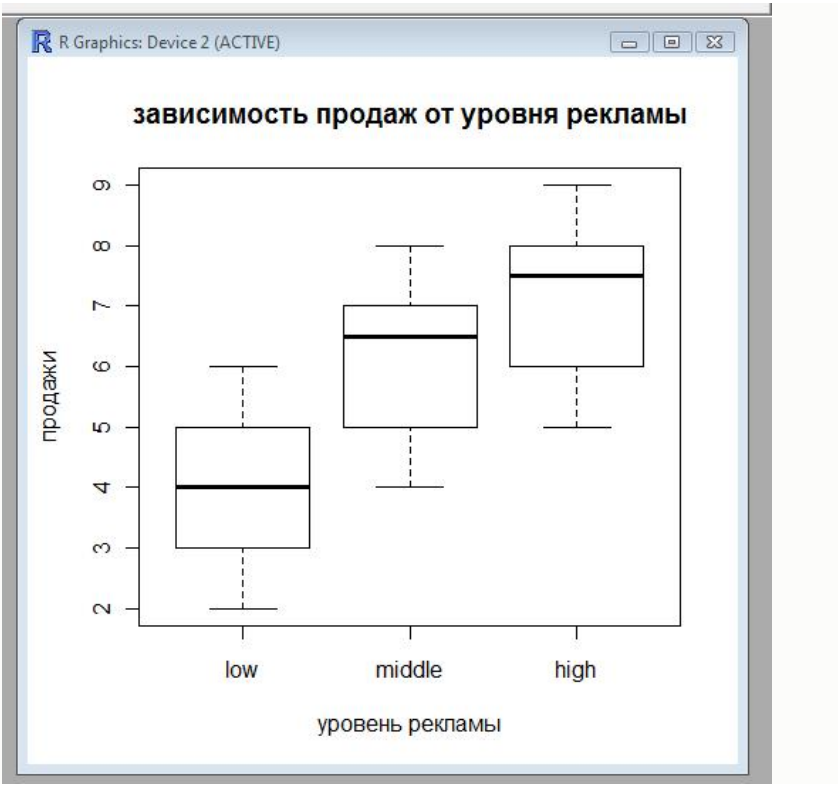


Рис. 6.2.2.Зависимость продаж от уровня рекламы

Контрольные вопросы

1. Как готовятся данные для дисперсионного анализа?
2. Как осуществляется дисперсионный анализ?
3. Как построить график для сравнения групп?
4. Заполните следующую таблицу по результатам расчетов:

Мера	Значение
F-статистика	
p	

Практическая работа 6.3
Двухфакторный дисперсионный анализ

***Цель работы*:** научиться использовать пакет статистического анализа электронных таблиц Excel и возможности SPSS для Windows для двухфакторного дисперсионного анализа.

Постановка задачи

Маркетологи использовали три уровня рекламы товаров в магазине: высокий, средний и низкий. У купонной распродажи было два уровня. Купон на 20-долларовую скидку либо давали потенциальным покупателям (уровень в этом случае обозначали номером 1), либо не давали (этот уровень обозначали номером 2 в табл. 6.3.1 и в табл. 6.3.2).

Результаты экспериментов с рекламой и купоном объединили в таблицу размером 3х2 с шестью ячейками. Тридцать магазинов были выбраны случайным образом, и для каждой комбинации условий эксперимента (уровней рекламы и уровней купонной распродажи) случайным образом взяли по пять магазинов, как показано в табл. 6.3.2. Эксперимент продолжался два месяца. Определили объем продаж в каждом магазине, нормализовали его, приняв во внимание посторонние факторы (размер магазина, товарооборот и т.д.) и пересчитали по десятибалльной шкале. В дополнение была получена качественная оценка относительного числа постоянных покупателей для каждого магазина, также с использованием десятибалльной шкалы. Полученные данные приведены в табл. 6.3.1. [17, с. 613].

Нужно проверить, влияют ли уровень рекламы и уровень купонной распродажи на число продаж.

Таблица 6.3.1

Данные о продажах в зависимости от уровня рекламы и уровня купонной распродажи

Номер магазина	Уровень купонной распродажи	Внутри-магазинная реклама	Продажи	Постоянные покупатели
1	1	1	10	9
2	1	1	9	10
3	1	1	10	8
4	1	1	8	4
5	1	1	9	6
6	1	2	8	8
7	1	2	8	4
8	1	2	7	10
9	1	2	9	6
10	1	2	6	9
11	1	3	5	8
12	1	3	7	9
13	1	3	6	6
14	1	3	4	10
15	1	3	5	4
16	2	1	8	10
17	2	1	9	6
18	2	1	7	8
19	2	1	7	4
20	2	1	6	9
21	2	2	4	6
22	2	2	5	8
23	2	2	5	10
24	2	2	6	4
25	2	2	4	9
26	2	3	2	4
27	2	3	3	6
28	2	3	2	10
29	2	3	1	9
30	2	3	2	8

Решение задачи с помощью SPSS для Windows

1. В редакторе данных описать переменные: Y – зависимая переменная – число продаж; F1 – независимая переменная – фактор 1 – уровень купонной распродажи с градациями: 1 – купон давали, 2 – купон не давали; F2 – независимая переменная 2 (фактор 2) – уровень рекламы с градациями: 1 – низкий, 2 – средний, 3 – высокий.
2. Ввести исходные данные.
3. Перенести переменные в окне «***Univariate***» (см. рис. 6.3.1)

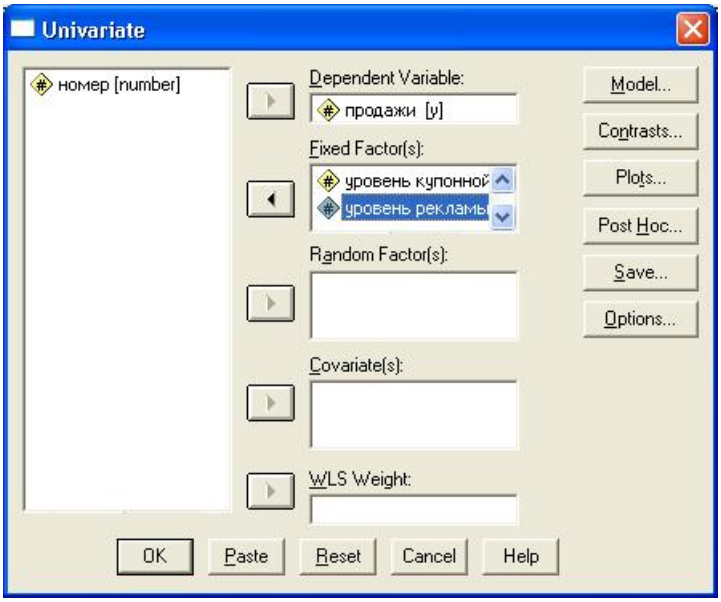


Рис. 6.3.1. Окно «Univariate»

4. Для построения графика нажать кнопку Plot, перенести переменную f1 в поле **Horizontal Axis**, f2 в поле **Separate Lines.** Щёлкнуть на выключателе **Add (см. рис. 6.3.2),** нажать **Continue, OK.**
5. Рассмотреть итоги в окне вывода (рис. 6.3.3 – 6.3.4).

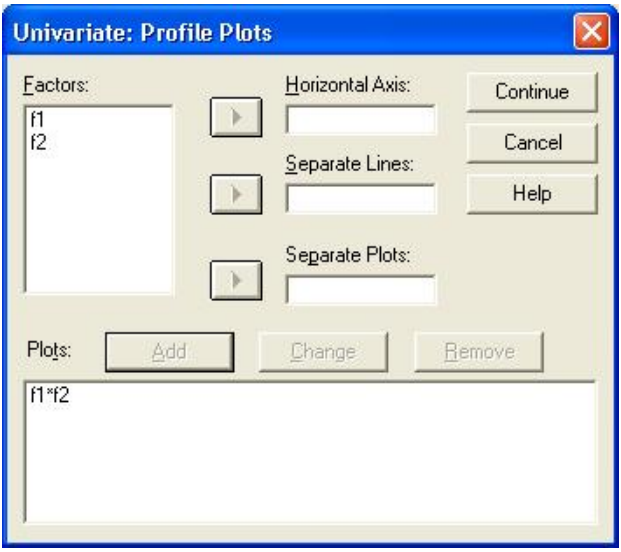


Рис. 6.3.2. Диалоговое окно «Univariate: Profile Plots»

Tests of Between-Subjects Effects

Dependent Variable: продажи

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	162.667(a)	5	32.533	33.655	.000
Intercept	1104.133	1	1104.133	1142.207	.000
KUPON	53.333	1	53.333	55.172	.000
RECLAMA	106.067	2	53.033	54.862	.000
KUPON * RECLAMA	3.267	2	1.633	1.690	.206
Error	23.200	24	.967		
Total	1290.000	30			
Corrected Total	185.867	29			

a R Squared = .875 (Adjusted R Squared = .849)

Рис. 6.3.3. Результаты выполнения дисперсионного анализа в окне вывода

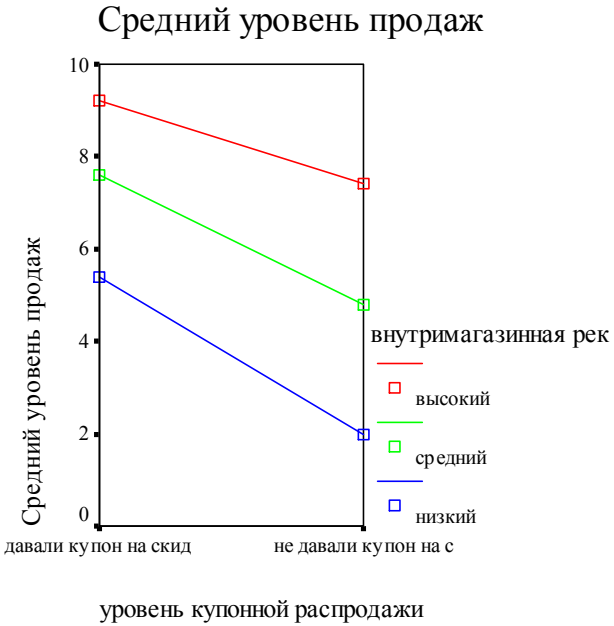


Рис. 6.3.3. График для средних

Анализ полученных результатов показывает, что на уровень продаж влияет уровень купонной рекламы и уровень внутримагазинной рекламы, а их взаимодействие оказывается не значимым.

Решение задачи двухфакторного дисперсионного анализа с помощью электронных таблиц

Постановка задачи

В одном из исследований изучалась такая личностная черта, как доминантность взрослых мужчин и женщин. Авторы предполагали, что доминантность должна быть выше у людей, которые были первенцами в своих семьях. Оказалось, что влияние каждого из двух исследуемых факторов – пола и порядка рождения незначимо, а их взаимодействие – значимо. У мужчин доминантность с увеличением порядка рождения

снижается, а у женщин – наоборот повышается. По данным рис. 6.3.4. проверьте с помощью двухфакторного дисперсионного анализа (ДА) приведенное выше утверждение.

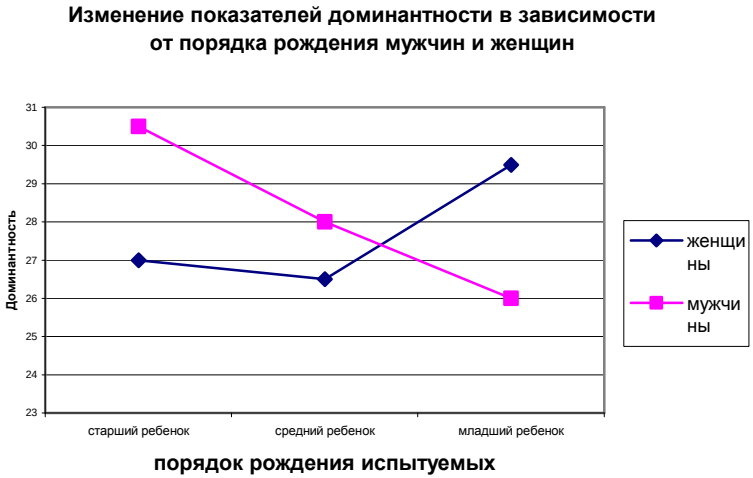
	B	C	D
1	Показатели доминантности у мужчин и женщин		
2	в зависимости от порядка рождения		
3		A1 - женщины	A2 - мужчины
4	B1 - старший ребенок	26	30
5		28	31
6	B2 - средний ребенок	27	27
7		26	29
8	B3 - младший ребенок	29	25
9		30	27

Рис. 6.3.4. Показатели доминантности у мужчин и женщин в зависимости от порядка рождения

Ход работы

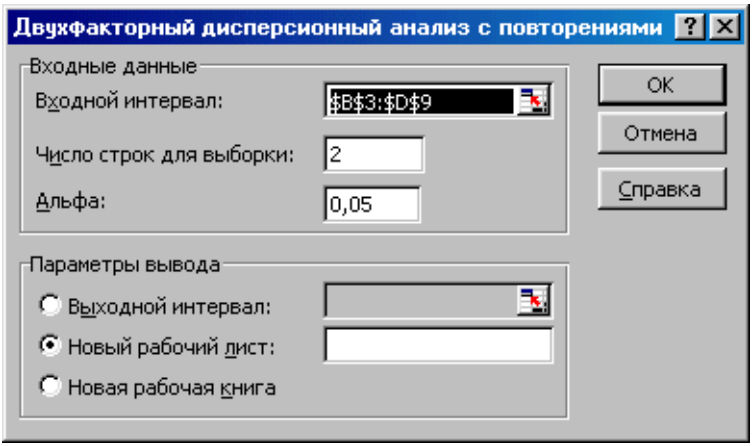
1. Ввести данные. Рассчитать средние значения показателей доминантности для шести групп.

2. Построить диаграмму для сравнения средних значений показателя доминантности шести групп испытуемых.



3. Выбрать меню *Данные, Анализ данных, Двуфакторный ДА*.

4. В окне дисперсионного анализа задать необходимые параметры.



5. Рассмотреть результаты.

6. В окне результатов сравнить три значения *F критического* и три значения *F эмпирического.* Определить, влияет ли на доминантность ребенка фактор пола, фактор порядка рождения и их взаимодействие.

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Выборка	4,67	2,00	2,33	1,87	0,23	5,14
Столбцы	0,75	1,00	0,75	0,60	0,47	5,99
Взаимодействие	26,00	2,00	13,00	10,40	0,01	5,14
Внутри	7,50	6,00	1,25			
Итого	38,92	11,00				

Если *F критическое < F эмпирического, нулевая гипотеза отвергается.*

Задание для самостоятельной работы

1. В одном из экспериментов по психологии программирования изучалось влияние комментариев в листинге на легкость модификации программы. Студенты начального курса были разделены на две группы по 6 человек в каждой. В качестве теста было предложено две формы 26-строчной программы на Фортране, оперирирующей информацией об оценках студентов. Первая форма содержала единственный блок комментариев высокого уровня в начале программы, сообщающий о ее назначении. Вторая форма программы не содержала комментариев высокого уровня, но имела 19 однострочных комментариев низкого уровня. Изучив программу и те, и другие студенты должны были провести три ее модификации, на которые давалось 30 минут. Три задания на модификацию оценивались по 10-балльной шкале. Результаты эксперимента приведены в таблице.

Оценки производительности в эксперименте с комментированием программы

	Оценка модификации		
Комментарии	1-я мод.	2-я мод.	3-я мод.
Высокоуровневые	6	8	3
	7	10	4
	5	8	4
	7	7	5
	7	9	5
	7	9	4
Низкоуровневые	6	6	1
	6	7	2
	7	8	3
	5	6	2
	6	6	2
	5	5	1

Влияет ли на результаты теста (производительность труда студентов) уровень комментированности программы, вид ее модификации или их взаимодействие?

Требования к отчету

Отчет о работе должен содержать:

- постановку задачи, исходные данные, формулировку гипотезы, результаты, выводы об истинности или ложности гипотез; файлы с данными.

Контрольные вопросы

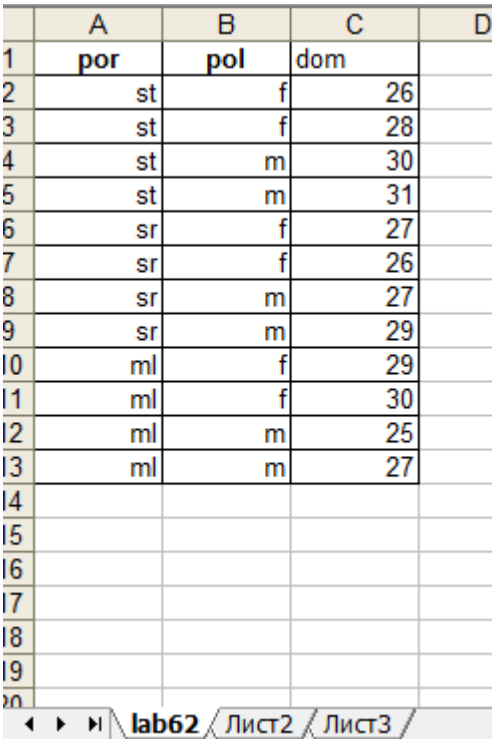
1. Какова цель двухфакторного дисперсионного анализа?
2. Как производится двухфакторный дисперсионный анализ с помощью Excel?
3. Как интерпретируются его результаты?
4. Как задаются данные для дисперсионного анализа с помощью SPSS? Как интерпретируются результаты?
5. Приведите свой пример задачи для двухфакторного дисперсионного анализа и решите ее.

Практическая работа 6.4

Двуфакторный дисперсионный анализ в R

Ход работы

1. Введите данные, связывающие доминантность ребенка в семье, с его полом и порядком рождения в Excel:



	A	B	C	D
1	**por**	**pol**	dom	
2	st	f	26	
3	st	f	28	
4	st	m	30	
5	st	m	31	
6	sr	f	27	
7	sr	f	26	
8	sr	m	27	
9	sr	m	29	
10	ml	f	29	
11	ml	f	30	
12	ml	m	25	
13	ml	m	27	
14				
15				
16				
17				
18				
19				
20				

lab62 | Лист2 | Лист3

Рис.6.4.1. Исходные данные

2. Загрузите данные в таблицу данных data3

```
 >data3<-read.table("K:\\Rexample\\lab6.2.csv",  sep=";",  dec=",",
header=TRUE)
 > data3
```

```
   por pol dom
1   st   f  26
2   st   f  28
3   st   m  30
4   st   m  31
5   sr   f   27
6   sr   f   26
7   sr   m  27
8   sr   m  29
9   ml   f  29
10  ml   f  30
11  ml   m  25
12  ml   m  27
```

3. Проведите двуфакторный дисперсионный анализ:

```
>summary(aov(dom~por*pol, data=data3))
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
por          2  4.6667   2.3333   1.8667  0.23424
pol          1  0.7500   0.7500   0.6000  0.46799
por:pol      2 26.0000  13.0000  10.4000  0.01122 *
Residuals    6  7.5000   1.2500
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Функция aov() в качестве первого параметра функции требует формулу модели: слева от тильды - имя столбца с данными (в нашем случае dom, справа — взаимодействующие факторы (пол и порядок рождения). Второй параметр — таблица данных для анализа (data3).

В нашем случае на доминантность влияет только взаимодействие порядка рождения и пола p=0,01122, поэтому нулевая гипотеза отвергается на уровне 0,01 (таким образом продажи различаются в зависимости от уровня рекламы).

Постройте график, иллюстрирующий результаты дисперсионного анализа (см. рис.6.4.2).

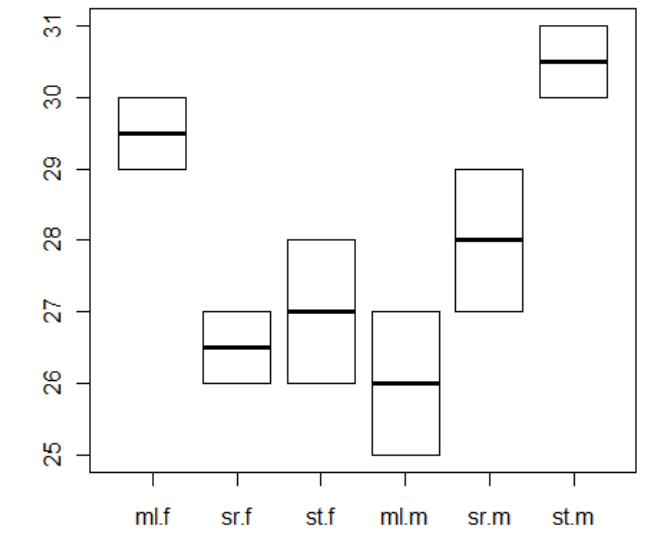


Рис. 6.4.2. Результаты двуфакторного дисперсионного анализа

Контрольные вопросы

1. Как подготовить данные для двуфакторного дисперсионного анализа?
2. Как осуществляется двуфакторный дисперсионный анализ?
3. Как построить график для сравнения групп?
4. Заполните следующую таблицу по результатам расчетов:

Формулировка нулевой гипотезы	**F-статистика**	**p-Значение**	**Вывод**
1- я:			
2- я:			
3- я:			

7. ДИСКРИМИНАНТНЫЙ АНАЛИЗ

С помощью дикриминантного анализа на основании некоторых признаков (независимых переменных) объект может быть причислен к одной из двух или нескольких групп (число групп определяется числом категорий зависимой переменной). В двумерной дискриминантном анализе объекты относятся к одной из двух групп, например, купившие или не купившие данных продукт. А независимыми переменными в этом случае выступают возраст, доход покупателей, и др. показатели.

Дискриминантный анализ используется для анализа данных в том случае, когда зависимая переменная категориальная, а предикторы (независимые переменные) — интервальные. В результате дискриминантного анализа строится так называемая дискриминантная функция

$$d = b_1x_1+b_2x_2+... + b_nx_n+a ,$$

где x_1 и x_n — значения переменных, соответствующих рассматриваемым случаям, b_1-b_n и a — коэффициенты, которые и предстоит оценить с помощью дискриминантного анализа. Коэффициенты подбираются так, чтобы по значениям дискриминантной функции можно было с максимальной четкостью провести разделение по группам.

Процедура **дискриминантного** анализа состоит из пяти шагов. Первый шаг — формулирование проблемы, требует **определения** целей, зависимой и независимых переменных. Выборку делят на две части. Анализируемую выборку используют для вычисления дискриминантной функции; **проверочную** — для проверки достоверности модели. Второй **шаг** — определение функции, включает выведение такой линейной комбинации предикторов **(дискриминантных** функций), чтобы группы максимально возможно различались между собой значениями предикторов

Определение статистической значимости представляет собой третий шаг. Она включает проверку нулевой гипотезы о том, что в совокупности средние всех дискриминантных функ-

ций во всех группах равны между собой. Если нулевую гипотезу отклоняют, то имеет смысл интерпретировать результаты.

Четвертый шаг — интерпретация дискриминантных весов или **коэффициентов** аналогична такой же стадии во множественном регрессионном анализе.

Пятый шаг — проверка достоверности. Она включает разработку классификационной матрицы. Дискриминантные веса, определенные с помощью анализируемой выборки, умножают на значения независимых переменных в проверочной выборке, чтобы получить дискриминантные показатели для случаев в этой выборке. Затем случаи распределяют по группам, исходя из дискриминантных показателей и соответствующего правила принятия решения. Определяют процент верно классифицированных случаев и сравнивают его с процентом случаев, которое можно ожидать на основе классификации методом случайного выбора.

Для оценки коэффициентов существует два известных подхода. Прямой метод включает оценку дискриминантной функции при одновременном введении всех предикторов. Альтернативный ему пошаговый метод включает последовательное введение предсказанных переменных, исходя из их способности дискриминировать группы.

Практическая работа 7.1
Дискриминантный анализ в SPSS

Цель работы: научиться использовать возможности SPSS для дискриминантного анализа данных и интерпретировать его результаты

Постановка задачи

Маркетологи изучали переменные, влияющие на семейный отдых. Семьям, которые отдыхали на курорте в последние два года, присвоен код 1; тем же, которые не посетили курорт за указанный период времени, присвоен код 2.

Обе выборки (как анализируемая, так и проверочная) сбалансированы с точки зрения посещаемости курорта. Анализируемая выборка содержит 15 семей каждой категории, а проверочная — по 6 семей каждой категории.

Кроме того, получены данные о доходе, отношении к путешествию, значении, придаваемом семейному отдыху, размеру семьи и возрасту главы семьи. Все данные приведены в таблице 7.1.1.

Таблица 7.1.1.

Факторы, влияющие на семейный отдых

Номер	Посещение курорта	Ежегодный доход	Отношение к путешествию	Значение, придаваемое семейному отдыху	Размер семьи	Возраст главы семьи
1	1	50,2	5	8	3	43
2	1	70,3	5	7	4	61
3	1	52,9	7	5	6	52
4	1	46,5	7	5	5	36
5	1	52,7	6	6	6	55
6	1	75	8	7	5	68
7.1	1	46,2	5	3	3	62

8	1	57	2	4	6	51
9	1	64,1	7	5	4	57
10	1	68,1	7	6	5	45
11	1	73,4	6	7	5	44
12	1	7,9	5	8	4	64
13	1	56,2	1	8	6	54
14	1	49,3	4	2	3	56
15	1	62	5	6	2	58
16	2	32,1	5	4	3	58
17	2	36,2	4	3	2	55
18	2	43,2	2	5	2	57
19	2	50,4	5	2	4	37
20	2	44,1	6	6	3	42
21	2	38,3	6	6	2	45
22	2	55	1	2	3	57
23	2	45,1	3	5	3	51
24	2	35	6	4	5	64
25	2	37,3	2	7	4	54
26	2	41,8	5	3	3	56
27	2	57.1	8	3	2	36
28	2	33,4	6	8	2	50
29	2	37,5	3	2	3	48
30	2	41,3	3	3	2	42

Решение задачи с помощью SPSS
Ход работы

1. Описать переменные и ввести данные согласно таблицы 7.1.1.

2. Выбрать Меню ***Analyze, Classify, Discriminant Analisis.***

3. В открывшемся окне перенести переменную ***отдых*** *(посещение курорта)* в окно группирующей переменной, а остальные переменные в список независимых переменных или предикторов.

4. После щелчка по выключателю ***Define Range***... (Определить промежуток) ввести минимальное и максимальное значения переменной ***отдых***: 1 и 2 (рис.7.1.1).

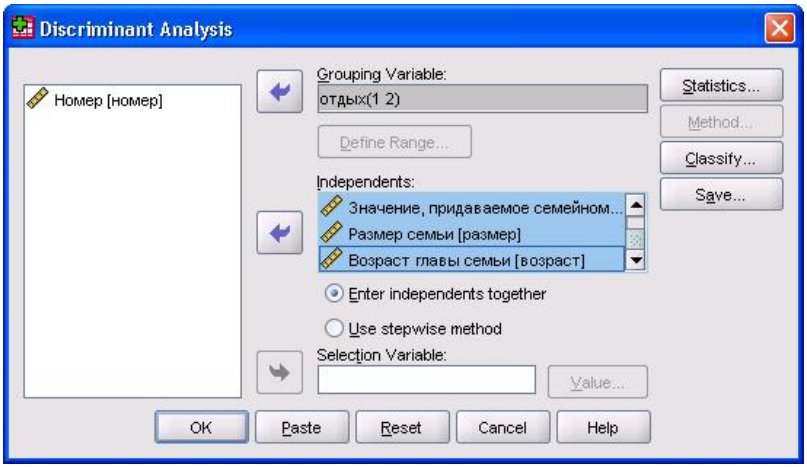


Рис. 7.1.1. Диалоговое окно «Discriminant Analisis»

5. Для начала оставить установленный по умолчанию метод: ***Enter independents together*** (Одновременный учет всех независимых переменных), при котором в анализе одновременно будут участвовать все независимые переменные.

6. Нажать кнопку ***Statistics*** и активировать опции: ***Means*** (Средние значения), ***Univariate ANOVAs*** (Одномерные тесты ANOVA), ***Unstandardized Function Coefficients*** (Нестандартизированные коэффициенты функции) и ***Within-group Correlation Matrice*** (Корреляционная матрица внутри группы) (рис.7.1.2).

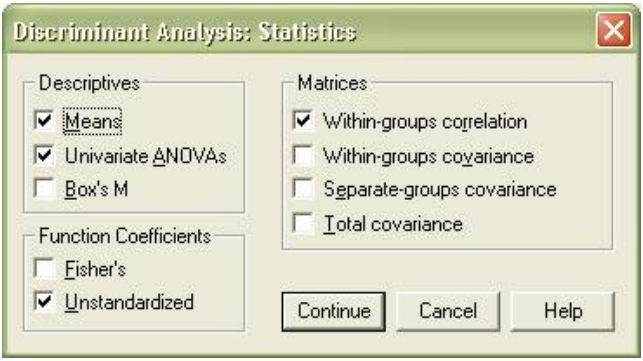


Рис. 7.1.2. Диалоговое окно «Discriminant Analisis Statistics»

7. Нажать кнопку Classify и сделать дополнительно запрос на вывод диаграмм по отдельным группам (Separate-groups Plots), результатов для отдельных наблюдений (Casewise results) и сводной таблицы (Summary table) (рис.7.1.3.)

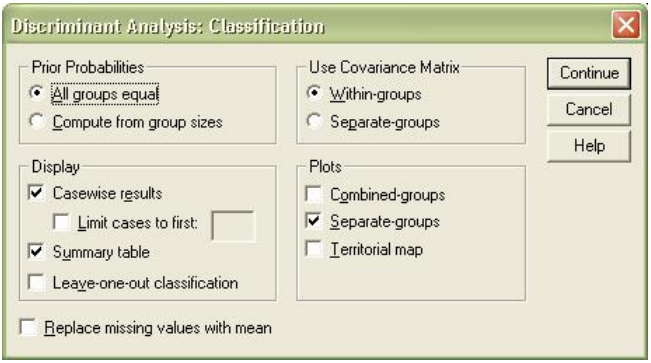


Рис. 7.1.3. Диалоговое окно «Discriminant Analisis: Classification»

8. При помощи выключателя ***Save***... сохранить значения дискриминантной функции в дополнительной переменной (Discriminant Scores).

9. Рассмотреть и проанализировать результаты в окне вывода.

В таблице «Статистика групп» (рис.7.1.4) приводятся средние значения и стандартные отклонения для всех признаков в каждой группе отдельно и суммарные показатели.

Group Statistics

Посещение курорта		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
1	Ежегодный доход	59.7200	10.16951	15	15.000
	Отношение к путешествию	5.3333	1.91485	15	15.000
	Значение, придаваемое семейному отдыху	5.8000	1.82052	15	15.000
	Размер семьи	4.4667	1.30201	15	15.000
	Возраст главы семьи	53.7333	8.77062	15	15.000
2	Ежегодный доход	41.8467	7.51588	15	15.000
	Отношение к путешествию	4.3333	1.95180	15	15.000
	Значение, придаваемое семейному отдыху	4.2000	1.89737	15	15.000
	Размер семьи	2.8667	.91548	15	15.000
	Возраст главы семьи	50.1333	8.27101	15	15.000
Total	Ежегодный доход	50.7833	12.64178	30	30.000
	Отношение к путешествию	4.8333	1.96668	30	30.000
	Значение, придаваемое семейному отдыху	5.0000	2.00000	30	30.000
	Размер семьи	3.6667	1.37297	30	30.000
	Возраст главы семьи	51.9333	8.57395	30	30.000

Рис. 7.1.4. Статистика групп

В следующей таблице (рис. 7.1.5) приводятся результаты теста о том, насколько значимо различаются между собой переменные в обеих группах. Для этого приводятся значения тестовой величины Лямбда Уилкса ("Wilks-Lambda") и применяется простой дисперсионный анализ. Для всех переменных (кроме возраста и отношения к путешествию) получается значимое различие между обеими группами.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Ежегодный доход	.483	29.966	1	28	.000
Отношение к путешествию	.933	2.006	1	28	.168
Значение, придаваемое семейному отдыху	.834	5.554	1	28	.026
Размер семьи	.649	15.158	1	28	.001
Возраст главы семьи	.954	1.338	1	28	.257

Рис. 7.1.5. Тест на равенство групповых средних

Затем вычисляется корреляционная матрица (рис. 7.1.6).

Pooled Within-Groups Matrices

		Ежегодный доход	Отношение к путешествию	Значение, придаваемое семейному отдыху	Размер семьи	Возраст главы семьи
Correlation	Ежегодный доход	1.000	.135	.112	.017	.010
	Отношение к путешествию	.135	1.000	.089	-.044	-.213
	Значение, придаваемое семейному отдыху	.112	.089	1.000	.048	.045
	Размер семьи	.017	-.044	.048	1.000	-.007
	Возраст главы семьи	.010	-.213	.045	-.007	1.000

Рис. 7.1.6. Корреляционная матрица

На следующем шаге вычисляются и анализируются коэффициенты дискриминантной функции**.**

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.751[a]	100.0	100.0	.798

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.364	25.803	5	.000

Рис. 7.1.7. Собственные значения и лямбда Уилкоксона

При помощи Лямбда Уилкса (рис. 7.1.7) производится тест на то, значимо ли в обеих группах отличаются друг от друга средние значения дискриминантной функции; значение $p < 0,05$ указывает на очень значимое различие.

Значение, выводимое под именем "Eigenvalue" (Собственное значение), соответствует отношению суммы квадратов между группами к сумме квадратов внутри групп. Большие собственные значения указывают на удачно подобранные дискриминантные функции.

Следующая таблица дает представление о том, как сильно отдельные переменные, применяемые в дискриминантной функции, коррелируют со стандартизированными значениями этой дискриминантной функции. При этом корреляционные коэффициенты были рассчитаны в обеих группах по отдельности и затем усреднены.

Далее приводятся сами коэффициенты дискриминантной функции (рис. 7.1.8).

Canonical Discriminant Function Coefficients

	Function
	1
Ежегодный доход	.081
Отношение к путешествию	.077
Значение, придаваемое семейному отдыху	.112
Размер семьи	.481
Возраст главы семьи	.022
(Constant)	-7.944

Unstandardized coefficients

Рис.7.1.8. Нестандартизованные коэффициенты

В завершении приводится классификационная таблица (рис. 7.1.9) с указанием достигнутой точности прогнозирова-

ния. Значение этой точности равно 90%, что является хорошим результатом.

Classification Results[a]

			Predicted Group Membership		
		Посещение курорта	1	2	Total
Original	Count	1	12	3	15
		2	0	15	15
	%	1	80.0	20.0	100.0
		2	.0	100.0	100.0

a. 90.0% of original grouped cases correctly classified.

Рис.7.1.9. Результаты классификации

Требования к отчету

- файлы с исходными данными и результатами расчетов
- ответы на контрольне вопросы

Контрольные вопросы

1. Назначение дискриминантного анализа.
2. Как он осуществляется в SPSS?
3. Назовите этапы выполнения дискриминантного анализа.
4. Как разделить общую выборку для целей анализа и проверки?
5. Что такое коэффициент λ Уилкса? Для каких целей его используют?
6. Что такое классификационная матрица?
7. Как определяют статистическую значимость дискриминантного анализа?
8. Чем отличается пошаговый дискриминантный метод от прямого?
9. Дайте содержательную интерпретацию полученным результатам.

Практическая работа 7.2
Дискриминантный анализ с помощью R

Прочитайте данные для дискриминантного анализа из файла Excel

	A	B	C	D	E	F
1	G	x1	x2	x3	x4	x5
2	1	50,2	5	8	3	43
3	1	70,3	5	7	4	61
4	1	52,9	7	5	6	52
5	1	46,5	7	5	5	36
6	1	52,7	6	6	6	55
7	1	75	8	7	5	68
8	1	46,2	5	3	3	62
9	1	57	2	4	6	51
10	1	64,1	7	5	4	57
11	1	68,1	7	6	5	45
12	1	73,4	6	7	5	44
13	1	71,9	5	8	4	64
14	1	56,2	1	8	6	54
15	1	49,3	4	2	3	56
16	1	62	5	6	2	58
17	2	32,1	5	4	3	58
18	2	36,2	4	3	2	55
19	2	43,2	2	5	2	57
20	2	50,4	5	2	4	37
21	2	44,1	6	6	3	42
22	2	38,3	6	6	2	45
23	2	55	1	2	3	57
24	2	45,1	3	5	3	51
25	2	35	6	4	5	64
26	2	37,3	2	7	4	54
27	2	41,8	5	3	3	56
28	2	57	8	3	2	36
29	2	33,4	6	8	2	50
30	2	37,5	3	2	3	48
31	2	41,3	3	3	2	42
32						
33						

lab7 Лист2 Лист3

```
> data7<-read.table("K:\\Rexample\\lab7.csv", sep=";", dec=",",
header=TRUE)
```

Загрузите пакет для дискриминантного анализа

```
>library(MASS)
```

Выполните линейный дикриминантный анализ (G – группа, к которой относится семья; x1, x2, x3, x4, x5 - предикторы, на основании которых мы причисляем объект к той или иной группе).

```
> fit <- lda(G ~ x1 + x2 + x3 + x4 + x5, da-
ta=data7,na.action="na.omit", CV=TRUE)
```

Выведите информацию о проделанном анализе

```
> fit
```

```
$class
 [1] 2 1 1 2 1 1 2 1 1 1 1 1 1 2 1 2 2 2 2 2 2 1 2 1 2 2 1 2 2 2
Levels: 1 2

$posterior
              1            2
1  0.113258268 8.867417e-01
2  0.996016716 3.983284e-03
3  0.966736499 3.326350e-02
4  0.216380790 7.836192e-01
5  0.975200087 2.479991e-02
6  0.999981302 1.869757e-05
7  0.032339307 9.676607e-01
8  0.932150071 6.784993e-02
9  0.975051346 2.494865e-02
10 0.995991340 4.008660e-03
11 0.999088805 9.111950e-04
12 0.998607048 1.392952e-03
13 0.973751874 2.624813e-02
14 0.026965196 9.730348e-01
15 0.501536398 4.984636e-01
16 0.011722597 9.882774e-01
17 0.003733323 9.962667e-01
18 0.026041504 9.739585e-01
19 0.434079763 5.659202e-01
20 0.128832186 8.711678e-01
21 0.013306218 9.866938e-01
22 0.556359394 4.436406e-01
23 0.095818165 9.041818e-01
24 0.737445243 2.625548e-01
25 0.180494502 8.195055e-01
26 0.058080298 9.419197e-01
27 0.539887579 4.601124e-01
28 0.012271705 9.877283e-01
29 0.007514953 9.924850e-01
30 0.004324549 9.956755e-01
```

Вначале выведены для всех 30-ти семей предсказанные группы, а затем вероятность принадлежности к первой и второй группе.

Выведите теперь данные о корректно классифицированных случаях:

```
> ct <- table(data7$G, fit$class)
```

```
> ct

   1  2
1 11  4
2  3 12
```

Мы видим, что мы неправильно классифицировали 4+3=7 случаев. Для четырех семей мы предсказали, что они будут относиться ко второй группе, в то время как они относились к первой; а также три семьи мы отнесли к первой группе, в то время как они относились ко второй.

В целом, процент верно классифицированных случаев можно получить так:

```
> sum(diag(prop.table(ct)))
[1] 0.7666667
```

Для получения коэффициентов дискриминантной функции:

```
>  fit <- lda(G ~ x1 + x2 + x3 + x4 + x5, data=data7)
> fit
Call:
lda(G ~ x1 + x2 + x3 + x4 + x5, data = data7)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
        x1       x2  x3       x4       x5
1 59.72000 5.333333 5.8 4.466667 53.73333
2 41.84667 4.333333 4.2 2.866667 50.13333

Coefficients of linear discriminants:
           LD1
x1 -0.08134025
x2 -0.07673785
x3 -0.11176810
x4 -0.48101971
x5 -0.02156085
```

Таким образом, наша модель верно классифицирует 76% случаев, что является достаточно хорошим показателем.Чтобы узнать, какая информация генерируется при выполнении дискриминантного анализа можно также ввести команду:

```
 > names(fit)
 [1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
"call"
 [9] "terms"   "xlevels"
```

А затем последовательно вывести всю эту информацию.

```
 fit$counts
 1  2
15 15
```

Мы узнали, что в нашей задаче используются 2 группы, по 15 объектов в каждой.

```
> fit$means
       x1       x2  x3       x4       x5
1 59.72000 5.333333 5.8 4.466667 53.73333
2 41.84667 4.333333 4.2 2.866667 50.13333
```

Мы вывели информацию о средних значениях всех пяти предикторов для двух групп.

```
> fit$scaling
          LD1
x1 -0.08134025
x2 -0.07673785
x3 -0.11176810
x4 -0.48101971
x5 -0.02156085
```

Получили коэффициенты дискриминантного уравнения (без константы) .

Контрольные вопросы

1. Как задаются данные для выполнения дискриминантного анализа в R?
2. Как загрузить пакет для дискриминантного анализа и как он называется?
3. Объясните следующую строку кода R

```
fit <- lda(G ~ x1 + x2 + x3 + x4 + x5, data=data7,na.action="na.omit", CV=TRUE)
```

4. Как получить коэффициенты дискриминантного уравнения?
5. Как получить информацию о числе корректно классифицированных случаев?
6. Сравните информацию, выводимую R, с информацией, выводимой SPSS.

Литература

1. http://www.statmethods.net/advstats/discriminant.html
2. http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

8. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ (таксономия, автоклассификация, распознавание образов) можно назвать методом конденсации информации. Задачей кластерного анализа является такое представление многомерного массива информации в сжатом виде, чтобы потеря информации не была чрезмерной. Он позволяет объединить множество объектов в небольшое число однородных групп.

Пример. Пусть мы имеем матрицу данных, которая включает характеристики 22-х людей (объектов) по двум количественным признакам: стаж и зарплата. Откладывая значения признаков по осям координат, мы можем изобразить все объекты на плоскости в виде точек: X – стаж, Y – зарплата.

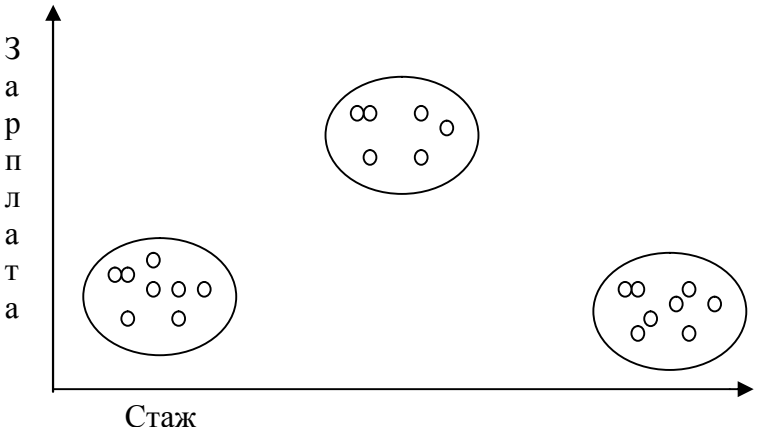


Рис. 8.1. Графическое представление кластеров

Как видно из рисунка 8.1, все объекты можно разбить на три группы так, что внутри одной группы находятся объекты с близкими значениями признаков. Первая группа – это люди с небольшим стажем работы и низкой зарплатой, вторая – со средним стажем и средней зарплатой и третья группа – люди с большим стажем и низкой зарплатой.

Множество близких между собой точек называется **кластером** или **таксоном** и при интерпретации результатов рассматривается как некоторый тип.

Одной из разновидностей кластерного анализа является **иерархический** кластерный анализ. Перед началом иерархической кластеризации все объекты считаются отдельными кластерами, которые в ходе алгоритма объединяются. На первом шаге выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате число кластеров становится N-1. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Результат работы алгоритма определяют способы вычисления расстояния между объектами и определения близости между кластерами. Наиболее употребительной мерой измерения расстояния является **евклидово расстояние** между объектами.

Процесс агрегирования данных может быть графически представлен с помощью дендрограммы (другие названия: "**дендограмма-дерево** "**диаграмма-дерево**").

Дендограмма – это графическое изображение процедуры и результатов процесса последовательной кластеризации, который осуществляется в терминах матрицы расстояний или сходства. Пример дендограммы приведен на рисунке 8.2.

В дендрограмме-дереве объекты располагаются вертикально слева (номера объектов на рис.8.2 обозначены обычным шрифтом), а результаты кластеризации — справа. Значения расстояний или сходства, отвечающие построению новых кластеров, изображаются в виде горизонтального ряда цифр, расположенного поверх дендрограммы (на рис. 8.2 – цифры, набранные курсивом).

Решение о числе кластеров принимают по теоретическим и практическим соображениям. В иерархической кластеризации важным критерием принятия решения о числе кластеров являются расстояния, при которых происходит объединение кластеров.

Относительные размеры кластеров должны быть такими, чтобы имело смысл сохранить данный кластер, а не объединить его с другими. Кластеры интерпретируют с точки зрения **кластерных центроидов** (средних значений переменных).

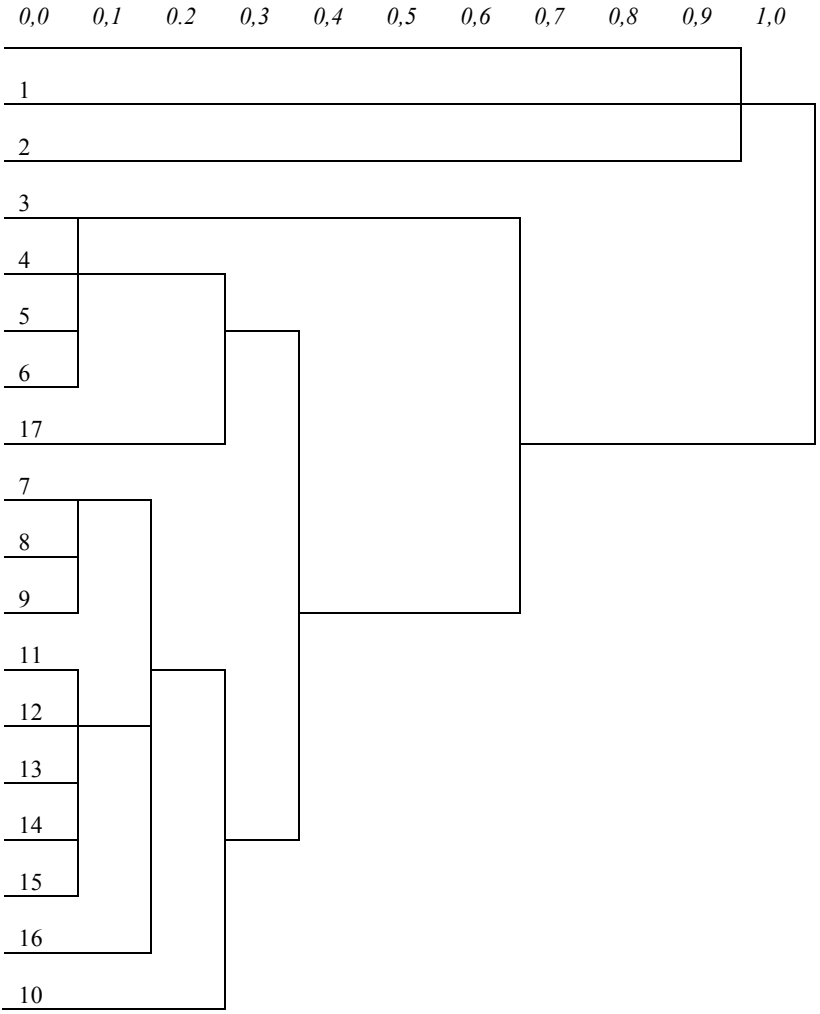


Рис. 8.2. Пример дендрограммы-дерева

Результатом работы алгоритмов таксономии обычно является разбиение множества объектов на группы (кластеры) в пространстве признаков, заданных исследователем, а также расчет некоторых обобщенных характеристик каждого кластера (центр, средние, меры вариации).

Существуют методы, позволяющие проводить классификацию в пространстве признаков, измеренных с помощью разного типа шкал: номинальных, порядковых, ранговых.

Практическая работа 8.1
Иерархический кластерный анализ в SPSS

Цель работы: научиться производить иерархический кластерный анализ и описывать полученные кластеры.

Постановка задачи

18.1 претендентов прошли 10 различных тестов в отделе кадров предприятия [8.1]. Максимальная оценка, которую можно было получить по тесту, – 10 баллов. Список тестов был следующим:

1. Память на числа.
2. Математические задачи.
3. Находчивость при прямом диалоге.
4. Тест на составление алгоритмов.
5. Уверенность во время выступления.
6. Командный дух.
7. Находчивость.
8. Сотрудничество.
9. Признание в коллективе.
10. Сила убеждения.

Распределите претендентов в кластеры, исходя из результатов тестирования.

Ход работы

1. Опишите переменные и введите данные в соответствии со следующей таблицей (для переменной с фамилией задайте тип ***String***).

Таблица 8.1.1.

Результаты тестирования претендентов

№	Претендент	T1	T2	T3	T4	T5	T6	T7	T8.1	T9	T10
1	Volker R	10	10	9	10	10	10	9	10	10	9
2	Sigrid K	10	10	4	10	5	5	4	5	4	3
3	Elmar M	5	4	10	5	10	4	10	5	3	10
4	Peter B	10	10	9	10	10	10	9	10	10	9
5	Otto R	4	3	5	4	3	10	4	10	10	5
6	Elke M	10	10	4	10	5	4	3	4	5	5
7	Sarah K	4	4	5	5	4	10	5	10	10	6
8	Peter T	4	5	3	4	5	10	4	10	10	4
9	Gudrun M	4	5	10	4	10	5	10	4	3	10
10	Siglinde P	10	10	4	10	5	4	4	5	4	4
11	Werner W	4	5	10	5	10	4	10	4	5	10
12	Achim Z	10	10	9	10	10	9	9	10	10	10
13	Dieter K	6	5	4	3	5	10	5	10	10	5
14	Boris P	4	5	10	4	10	5	10	3	4	10
15	Silke W	10	10	9	10	10	9	10	9	10	10
16	Clara T	6	5	3	4	4	10	4	10	10	5
17	Manfred K	10	10	5	10	4	5	4	3	4	5
18	Richard M	4	5	10	4	10	4	10	4	4	10

2. Выберите меню ***Analyze, Classify, Hierarchical Cluster*** (Иерархический кластерный анализ) (рис. 8.1.3)

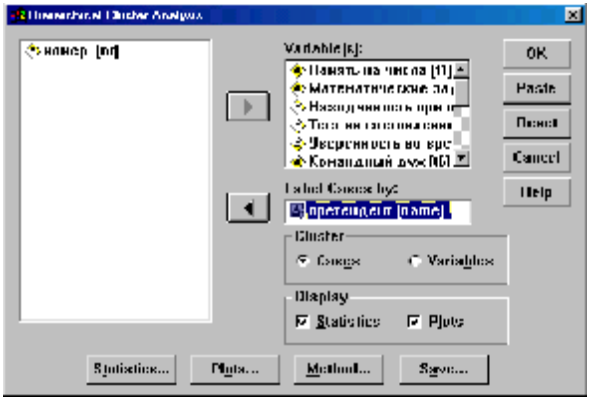


Рис. 8.1.3. Диалоговое окно «Hierarchical Cluster Analysis»

3. Поместите переменные ***t1*** – ***t10*** в поле тестируемых переменных, а переменную имя претендента (***name***) – используйте для обозначения случаев.

Обзорная таблица порядка слияния должна выглядеть следующим образом (см. рис. 8.1.4).

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	,000	0	0	6
2	14	18	2,000	0	0	4
3	12	15	2,000	0	0	6
4	9	14	2,000	0	2	8
5	2	10	2,000	0	0	13
6	1	12	3,000	1	3	15
7	13	16	4,000	0	0	12
8	9	11	4,000	4	0	11
9	5	7	5,000	0	0	14
10	6	17	6,000	0	0	13
11	3	9	6,500	0	8	15
12	8	13	7,000	0	7	14
13	2	6	7,500	5	10	16
14	5	8	12,833	9	12	16
15	1	3	194,000	6	11	17
16	2	5	198,500	13	14	17
17	1	2	219,407	15	16	0

Рис. 8.1.4. Таблица порядка слияния кластеров

Поскольку значительный скачок в показателе ***Coefficient*** наблюдается после 14-го шага, то для данных, состоящих из 18.1-ти случаев, оптимальным является решение, состоящее из 18.1-14=4 кластера.

4. Вновь активируйте окно **«Hierarchical Cluster Analysis» (**Иерархический кластерный анализ) и, щелкнув на кнопке ***Statistics,*** укажите в разделе принадлежность к кластеру ***Single solution*** и ***4 clusters*** (рис.8.1.5).

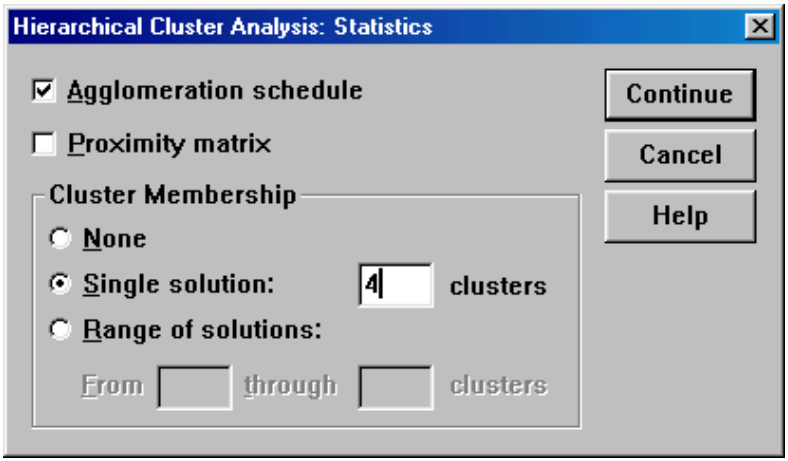


Рис. 8.1.5. Диалоговое окно «Hierarchical Cluster Analysis: Statistics»

5. Пройдите выключатель ***Save.*** Теперь для каждого случая будет выводиться и информация о принадлежности к кластеру.

6. Получите таблицу о принадлежности к кластерам (см. рис. 8.1.6).

Разобраться в значении кластеров помогут кластерные профили. Они представляют собой средние значения переменных, которые включены в анализ, распределенные по кластерной принадлежности.

7. Выберите в меню ***Analyze, Compare Means, Means.*** Переменным t1 – t10 присвойте статус зависимых, а переменной «принадлежность к кластеру» – независимой и начните расчет (см. рис. 8.1.7.). Результаты расчета приведены на рис. 8.1.8.

Cluster Membership

Case	4 Clusters
1:Volker R	1
2:Sigrid K	2
3:Elmar M	3
4:Peter B	1
5:Otto R	4
6:Elke M	2
7:Sarah K	4
8:Peter T	4
9:Gudrun M	3
10:Siglinde P	2
11:Werner W	3
12:Achim Z	1
13:Dieter K	4
14:Boris P	3
15:Silke W	1
16:Clara T	4
17:Manfred K	2
18:Richard M	3

Рис. 8.1.6. Принадлежность к кластерам

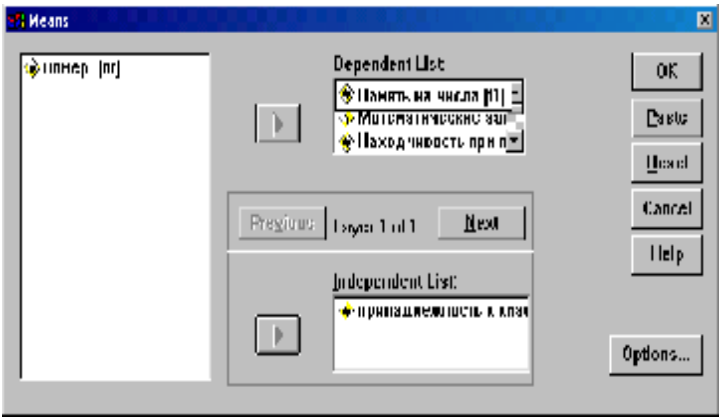


Рис. 8.1.7. Вычисление кластерных центров

8. Дайте содержательное описание кластерам (рис. 8.1.8).

Report

Mean

	принадлежность к кластеру				
	1	2	3	4	Total
Память на числа	10,00	10,00	4,20	4,80	6,94
Математические задачи	10,00	10,00	4,80	4,40	7,00
Находчивость при прямом диалоге	9,00	4,25	10,00	4,00	6,83
Тест на составление алгоритмов	10,00	10,00	4,40	4,00	6,78
Уверенность во время выступления	10,00	4,75	10,00	4,20	7,22
Командный дух	9,50	4,50	4,40	10,00	7,11
Находчивость	9,25	3,75	10,00	4,40	6,89
Сотрудничество	9,75	4,25	4,00	10,00	7,00
Признание в коллективе	10,00	4,25	3,80	10,00	7,00
Сила убеждения	9,50	4,25	10,00	5,00	7,22

Рис. 8.1.8. Средние значения всех переменных по кластерам

Задача для самостоятельного решения

Создайте файл пиво.sav, который содержит некоторые данные о 17-ти сортах пива (табл. 8.1.2).

Таблица 8.1.2.

Информация о 17-ти сортах пива

Марка пива	Страна-производитель	Расходы	Калории	% алкоголя
Budweiser	1	0,43	144	4,7
Lowenbrau	1	0,48	157	4,9
Michelob	1	0,50	162	5,0
Kronenbourg	3	0,73	170	5,2
Heineken	4	0,77	152	5,0
Schmidts	1	0,30	147	4,7
Pabst	1	0,38	152	4,9
Miller	1	0,43	99	4,3
Budweiser	1	0,44	113	3,7

Coors	1	0,46	102	4,1
Dos	5	0,70	145	4,5
Becks	2	0,76	150	4,7
Rolling	1	0,36	144	4,7
Pabst	1	0,38	68	2,3
Tuborg	1	0,43	155	5,0
Olympia	1	0,46	72	2,9
Schlitz	1	0,47	97	4,2

- переменная «производитель» указывает на страну-производителя пива, где США закодированы с помощью единицы;
- расходы приведены в долларах США для ёмкости, равной 12-ти унциям для жидкости (примерно одна треть литра);
- калорийность указана для одинакового количества пива;
- содержание алкоголя приводится в процентах.

Распределите 17 сортов пива в кластеры, исходя из двух переменных: «калории» и «расходы». (Указание: используйте иерархический кластерный анализ, значения переменных – стандартизуйте – вкладка ***Method***, ***Z-оценки***). Как определить оптимальное число кластеров? Охарактеризуйте полученные кластеры.

Требования к отчету

Отчет должен содержать:

- файлы с данными и результатами расчетов;
- ответы на контрольные вопросы.

Контрольные вопросы

1. Назначение кластерного анализа. В чем отличие кластерного анализа от факторного анализа.
2. Назовите примеры использования кластерного анализа в маркетинге, социологии, психологии.
3. Какие методы кластерного анализа вы знаете? Как он осуществляется в SPSS?
4. Как определить оптимальное число кластеров при иерархическом анализе?
5. Что является наиболее распространенной мерой сходства в кластерном анализе?
6. Что показывает дендрограмма и как ее построить?

Практическая работа 8.2
Кластерный анализ в R
Ход работы

1. Представьте данные о тестировании претендентов из предыдущей работы в файле Excel

№	Претендент	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
1	Volker R	10	10	9	10	10	10	9	10	10	9
2	Sigrid K	10	10	4	10	5	5	4	5	4	3
3	Elmar M	5	4	10	5	10	4	10	5	3	10
4	Peter B	10	10	9	10	10	10	9	10	10	9
5	Otto R	4	3	5	4	3	10	4	10	10	5
6	Elke M	10	10	4	10	5	4	3	4	5	5
7	Sarah K	4	4	5	5	4	10	5	10	10	6
8	Peter T	4	5	3	4	5	10	4	10	10	4
9	Gudrun M	4	5	10	4	10	5	10	4	3	10
10	Siglinde P	10	10	4	10	5	4	4	5	4	4
11	Werner W	4	5	10	5	10	4	10	4	5	10
12	Achim Z	10	10	9	10	10	9	9	10	10	10
13	Dieter K	6	5	4	3	5	10	5	10	10	5
14	Boris P	4	5	10	4	10	5	10	3	4	10
15	Silke W	10	10	9	10	10	9	10	9	10	10
16	Clara T	6	5	3	4	4	10	4	10	10	5
17	Manfred K	10	10	5	10	4	5	4	3	4	5
18	Richard M	4	5	10	4	10	4	10	4	4	10

2. Загрузите данные в таблицу данных data8:

```
>data8<-read.table("K:\\Rexample\\lab8.csv", sep=";", dec=",", header=TRUE)
>data8
  number      name t1 t2 t3 t4 t5 t6 t7 t8 t9 t10
1      1  Volker R 10 10  9 10 10 10  9 10 10   9
2      2  Sigrid K 10 10  4 10  5  5  4  5  4   3
3      3   Elmar M  5  4 10  5 10  4 10  5  3  10
4      4   Peter B 10 10  9 10 10 10  9 10 10   9
5      5    Otto R  4  3  5  4  3 10  4 10 10   5
6      6    Elke M 10 10  4 10  5  4  3  4  5   5
7      7   Sarah K  4  4  5  5  4 10  5 10 10   6
8      8   Peter T  4  5  3  4  5 10  4 10 10   4
```

```
9      9   Gudrun M  4  5 10  4 10  5 10  4  3  10
10    10 Siglinde P 10 10  4 10  5  4  4  5  4   4
11    11   Werner W  4  5 10  5 10  4 10  4  5  10
12    12    Achim Z 10 10  9 10 10  9  9 10 10  10
13    13   Dieter K  6  5  4  3  5 10  5 10 10   5
14    14    Boris P  4  5 10  4 10  5 10  3  4  10
15    15    Silke W 10 10  9 10 10  9 10  9 10  10
16    16    Clara T  6  5  3  4  4 10  4 10 10   5
17    17  Manfred K 10 10  5 10  4  5  4  3  4   5
18    18  Richard M  4  5 10  4 10  4 10  4  4  10
```

3. Проведите иерархический кластерный анализ и постройте дендрограмму:

```
>data8.dist <- daisy(data8[,2:11])
> data8.h <- hclust(data8.dist, method="ward")
>plot(data8.h,labels=abbreviate(data8[,1],1, method="both.sides"))
```

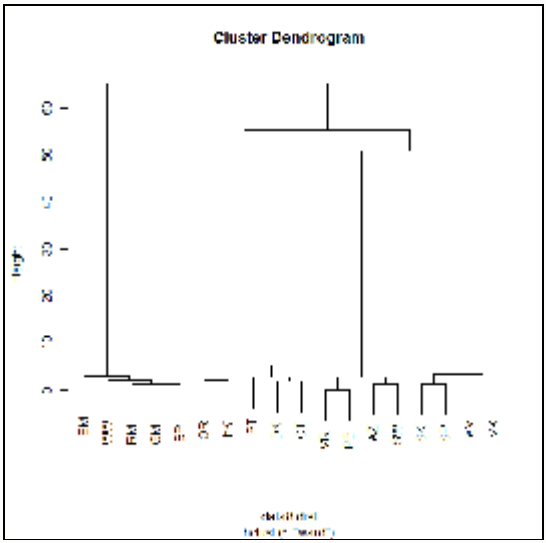
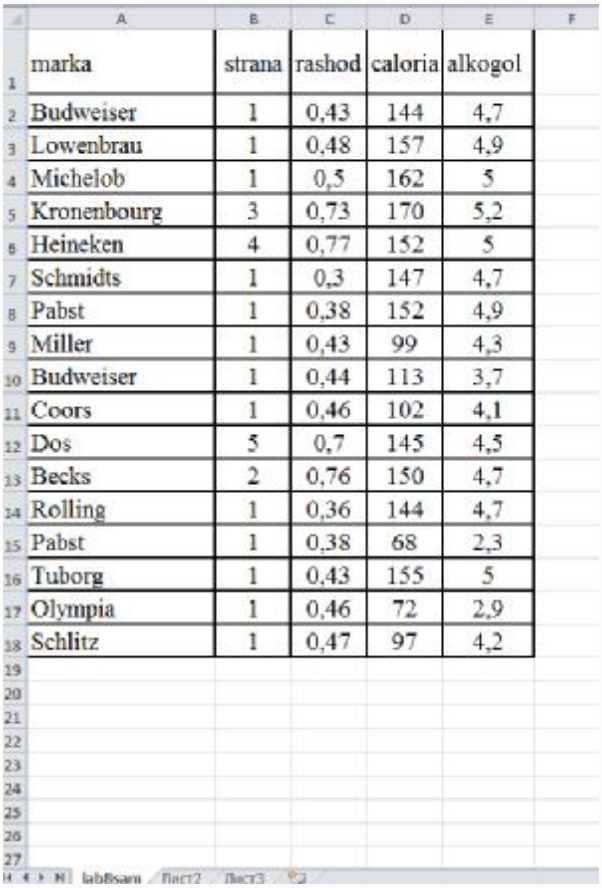


Рис. 8.2.1.Дендрограмма

4. Изучите справку по функции hclust.
5. Для примера с пивом (см. пред. работу) подготовьте файл с данными:



	A	B	C	D	E	F
1	marka	strana	rashod	caloria	alkogol	
2	Budweiser	1	0,43	144	4,7	
3	Lowenbrau	1	0,48	157	4,9	
4	Michelob	1	0,5	162	5	
5	Kronenbourg	3	0,73	170	5,2	
6	Heineken	4	0,77	152	5	
7	Schmidts	1	0,3	147	4,7	
8	Pabst	1	0,38	152	4,9	
9	Miller	1	0,43	99	4,3	
10	Budweiser	1	0,44	113	3,7	
11	Coors	1	0,46	102	4,1	
12	Dos	5	0,7	145	4,5	
13	Becks	2	0,76	150	4,7	
14	Rolling	1	0,36	144	4,7	
15	Pabst	1	0,38	68	2,3	
16	Tuborg	1	0,43	155	5	
17	Olympia	1	0,46	72	2,9	
18	Schlitz	1	0,47	97	4,2	
19						
20						
21						
22						
23						
24						
25						
26						
27						

Введите данные в R и просмотрите их:

```
>data8sam<-read.table("K:\\Rexample\\lab8sam.csv", sep=";",
dec=",", header=TRUE)
> data8sam
```

	marka	strana	rashod	caloria	alkogol
1	Budweiser	1	0.43	144	4.7
2	Lowenbrau	1	0.48	157	4.9
3	Michelob	1	0.50	162	5.0
4	Kronenbourg	3	0.73	170	5.2
5	Heineken	4	0.77	152	5.0
6	Schmidts	1	0.30	147	4.7
7	Pabst	1	0.38	152	4.9
8	Miller	1	0.43	99	4.3
9	Budweiser	1	0.44	113	3.7
10	Coors	1	0.46	102	4.1
11	Dos	5	0.70	145	4.5
12	Becks	2	0.76	150	4.7
13	Rolling	1	0.36	144	4.7
14	Pabst	1	0.38	68	2.3
15	Tuborg	1	0.43	155	5.0
16	Olympia	1	0.46	72	2.9
17	Schlitz	1	0.47	97	4.2

6. Стандартизуйте данные с третьей до пятой колонок, так как они представлены различными шкалами, в то время как в предыдущем примере все тесты измерялись в шкале от 0 до 10.

```
> scale(data8sam[3:5])
          rashod    caloria    alkogol
 [1,] -0.468979784  0.4004072  0.3817709
 [2,] -0.128267975  0.8044711  0.6362848
 [3,]  0.008016748  0.9598803  0.7635417
 [4,]  1.575291070  1.2085350  1.0180556
 [5,]  1.847860518  0.6490619  0.7635417
 [6,] -1.354830488  0.4936527  0.3817709
 [7,] -0.809691593  0.6490619  0.6362848
 [8,] -0.468979784 -0.9982755 -0.1272570
 [9,] -0.400837422 -0.5631298 -0.8907987
```

```
[10,] -0.264552699 -0.9050300 -0.3817709
[11,]  1.370863985  0.4314890  0.1272570
[12,]  1.779718156  0.5868982  0.3817709
[13,] -0.945976317  0.4004072  0.3817709
[14,] -0.809691593 -1.9618125 -2.6723960
[15,] -0.468979784  0.7423074  0.7635417
[16,] -0.264552699 -1.8374851 -1.9088543
[17,] -0.196410337 -1.0604392 -0.2545139
attr(,"scaled:center")
    rashod     caloria     alkogol
 0.4988235 131.1176471   4.4000000
attr(,"scaled:scale")
   rashod    caloria    alkogol
0.1467516 32.1731300  0.7858117
```

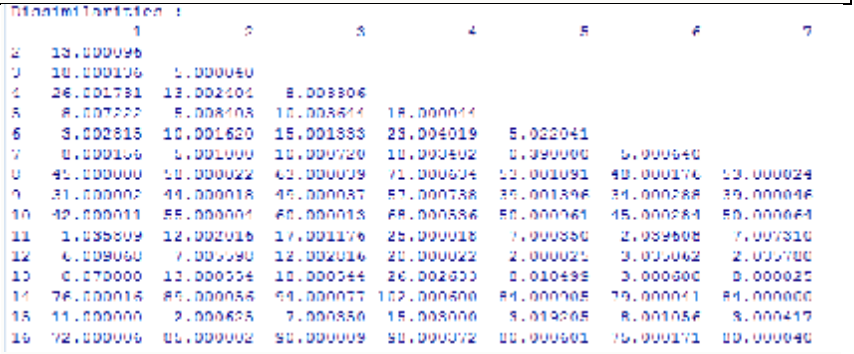
7. Загрузите пакет для кластерного анализа: меню ***Package, Load Package,cluster***

И постройте матрицу расстояний

> data8sam.dist <- daisy(data8sam[,3:4])

Посмотрите на получившуюся матрицу расстояний:

```
> data8sam.dist
```



Обведите красной линией границы 2-х кластеров

```
> rect.hclust(data8sam.h, k=2, border="red")
```

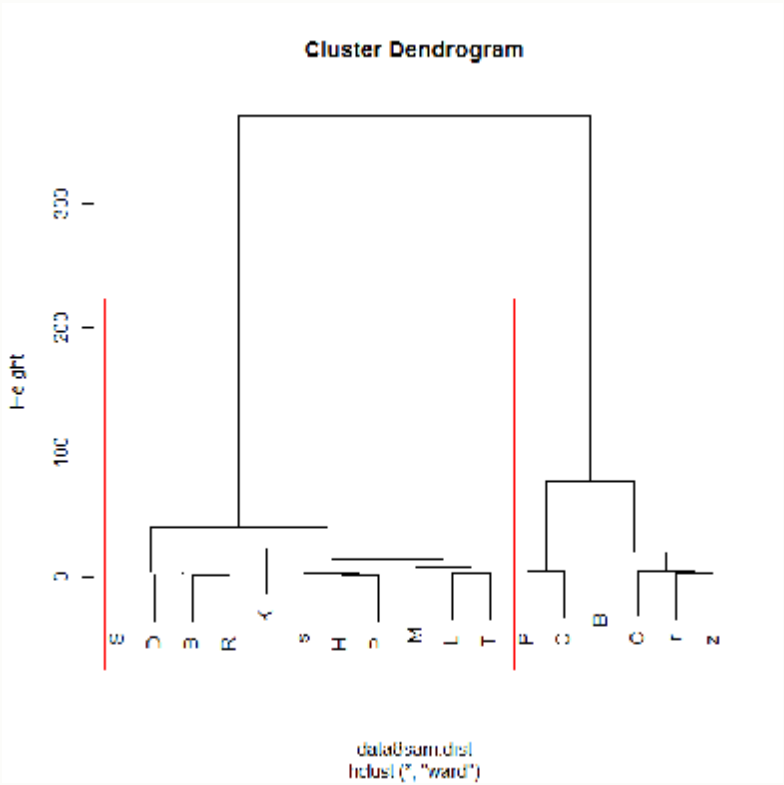


Рис.8.2.2. Два кластера, выделенные красными прямоугольниками на дендрограмме

Обведите красной линией границы 3-х кластеров:

```
>rect.hclust(data8sam.h, k=3, border="red")
```

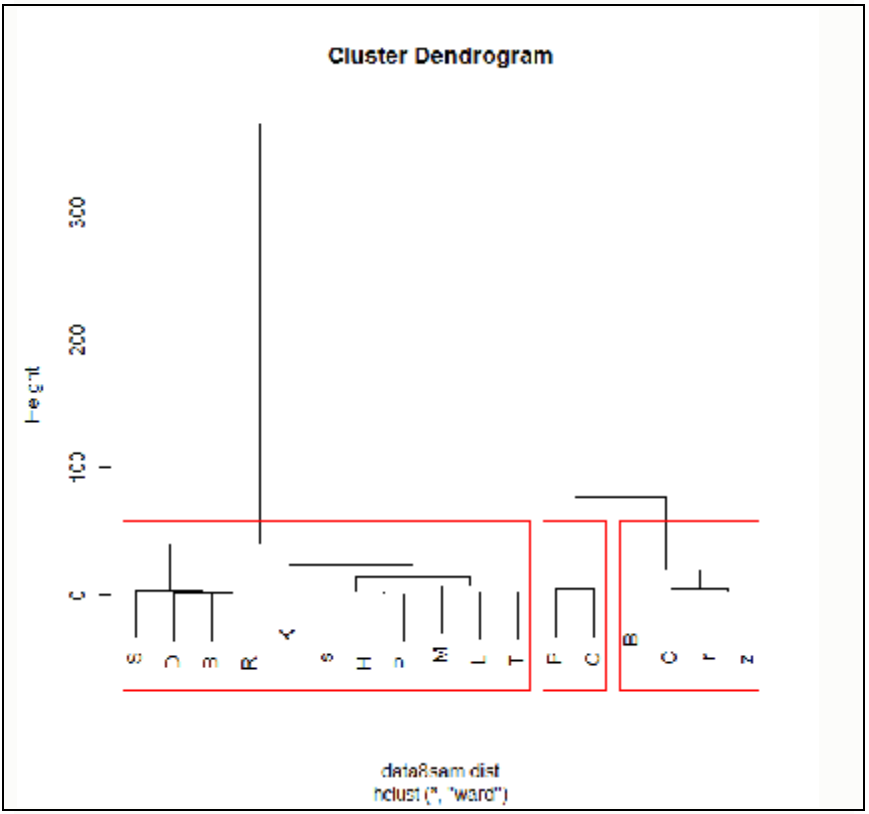


Рис.8.2.3. Три кластера, выделенные красными прямоугольниками на дендрограмме

8. Сохраните рабочее пространство: File, Save Workplace.

9. ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ так же, как и кластерный, является средством конденсации информации. Но, в отличие от кластерного анализа, задача которого – объединить множество объектов в небольшое число однородных групп, задача факторного анализа – найти максимально взаимосвязанные группы признаков. Эти группы представляют собой новые, комплексные переменные, называемые **факторами**. Число факторов существенно меньше, чем число измеряемых признаков.

Основное назначение **факторного анализа** – упорядочение кажущейся хаотичности изучаемого явления и нахождение такой простой структуры, которая достаточно точно отражала и воспроизводила бы реально существующие зависимости, т.е. отражала бы сущность этого явления.

Исходным для факторного анализа является положение о наличии взаимосвязи между признаками (переменными), которые подлежат анализу. В качестве количественной меры связи между парами переменных используется коэффициент корреляции.

Результатом факторного анализа является так называемая **факторная матрица** (другие названия – **матрица факторных нагрузок** или **матрица факторного отображения)**.

Общий вид факторной матрицы (без числовых значений) представлен в табл. 9.1.

Таблица 9.1

Общий вид матрицы факторных нагрузок (факторное отображение)*

Признаки	***Факторы***					
	F_1	F_2	F_3	F_4	F_5	F_6
P_1		*				
P_2		*	*			
P_3	*		*			*
P_5	*					
P_6	*				*	
…	*					
P_m	*					

– знаком «» помечены ячейки, в которых помещены нагрузки, существенно отличающиеся от нуля.*

Каждый фактор характеризуется столбцом, а каждая переменная – строкой матрицы факторного отображения.

На рис. 9.1 представлена классификация факторов, которые получаются в результате факторного анализа.

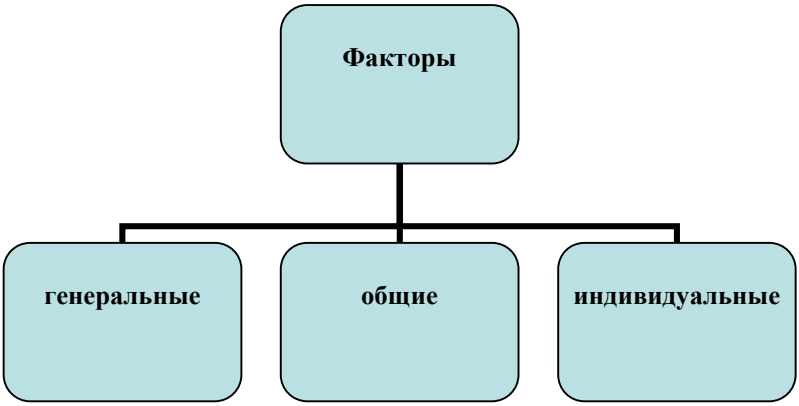


Рис. 9.1. Виды факторов

Фактор называется **генеральным**, если он имеет от всех переменных нагрузки, значительно отличные от нуля.

Фактор называется **общим**, если хотя бы две его нагрузки отличаются от нуля.

Индивидуальным называется фактор, который имеет ненулевую нагрузку только от одной переменной.

Переменные также можно различать по количеству высоких нагрузок. Количество высоких нагрузок переменной на факторы называется ее **сложностью**. Так, ***например***, если переменная значительно нагружает три фактора, ее сложность – 3.

Проблема факторов не имеет однозначного решения, но из бесконечного множества возможных решений нужно выбрать только одно. Такой выбор осуществляется путем враще-

ния конфигурации векторов (переменных) вокруг точки их начала. При геометрической интерпретации факторами являются координатные оси, на которые натянуто пространство общих факторов – пространство наименьшей размерности, в котором можно представить переменные в виде векторов.

Вращение факторов, как и все предшествующие ему процедуры факторного анализа, осуществляется в компьютере автоматически. Но, получив факторную матрицу после вращения, исследователь должен сам решить – какие факторы нужно оставить для дальнейшего анализа и интерпретации. Как правило, для дальнейшего анализа и интерпретации оставляют те факторы, на которые приходится более 5% полной дисперсии и абсолютные значения дисперсии которых не меньше единицы.

Прежде, чем делать окончательные выводы по результатам факторного анализа, необходимо обратить внимание на такие моменты:

– появляется ли тот или иной фактор при повторных измерениях одной или нескольких анализируемых переменных?

– не отражает ли фактор известную или тривиальную связь между переменными? (Если отражена тривиальная связь, фактор неинтересен для исследователя);

– не является ли данный фактор результатом неоднородности данных?

– отражают ли выявленные факторы сущность изучаемого явления?

– появляются ли выявленные факторы в случае повторной выборки из той же самой генеральной совокупности?

Практическая работа 9.1
Факторный анализ в SPSS

Цель работы: научиться проводить факторный анализ в SPSS для Windows и интерпретировать его результаты

Постановка задачи

В таблице 9.1.2 содержатся ответы тридцати респондентов на шесть вопросов, касающихся выбора зубной пасты. Респондентам предлагалось оценить степень своего согласия с приведенными ниже утверждениями по 7-ми бальной шкале (1 – «абсолютно не согласен», 7 – «полностью согласен»):

1. Важно приобретать зубную пасту, которая предотвращает развитие кариеса.
2. Мне нравится зубная паста, которая придает зубам белизну.
3. Зубная паста должна укреплять десны.
4. Я предпочитаю зубную пасту, которая освежает дыхание.
5. Предотвращение порчи зубов не является важным преимуществом данной зубной пасты.
6. Наиболее важной причиной покупки определенного вида зубной пасты является ее способность придать зубу привлекательный вид.

Таблица 9.1.2

Ответы 30-ти респондентов на вопросы о зубной пасте

№ респондента	Номер вопроса					
	v1	v2	v3	v4	v5	v6
1	7	3	6	4	2	4
2	1	3	2	4	5	4
3	6	2	7	4	1	3
4	4	5	4	6	2	5
5	1	2	2	3	6	2
6	6	3	6	4	2	4
7	5	3	6	3	4	3
8	6	4	7	4	1	4
9	3	4	2	3	6	3
10	2	6	2	6	7	6
11	6	4	7	3	2	3
12	2	3	1	4	5	4
13	7	2	6	4	1	3

14	4	6	4	5	3	6
15	1	3	2	2	6	4
16	6	4	6	3	3	4
17	5	3	6	3	3	4
18	7	3	7	4	1	4
19	2	4	3	3	6	3
20	3	5	3	6	4	6
21	1	3	2	3	5	3
22	5	4	5	4	2	4
23	2	2	1	5	4	4
24	4	6	4	6	4	7
25	6	5	4	2	1	4
26	3	5	4	6	4	7
27	4	4	7	2	2	5
28	3	6	2	6	4	3
29	4	7	3	7	2	7
30	2	3	2	4	7	2

Ход работы

1. Определите переменные в SPSS для Windows и введите данные согласно таблице 9.2.

2. Выберите меню ***Analyze, Data reduction, Factor.*** Перенесите переменные v1-v6 в поле выбранных переменных (см. рис. 9.1.2).

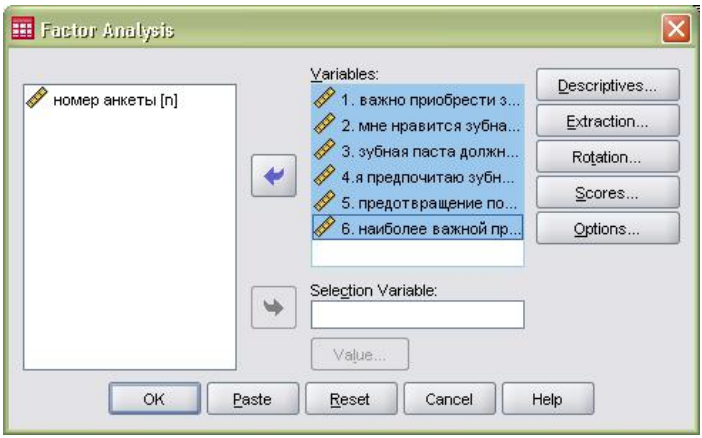


Рис. 9.1.2. Окно «Factor Analysis». Результаты переноса шести переменных в поле выбранных переменных («Variables»)

3. Нажмите кнопку ***Descriptives.*** Откроется диалоговое окно ***Factor Analysis: Descriptives*** (Факторный анализ: Описательные статистики) (см. рис. 9.1.3):

o Оставьте установленную по умолчанию опцию вывода ***Initial solution*** (Первичного решения).

o Выберите опции **Coefficients** (для вывода корреляционной матрицы) и **KMO and Bartlett's test of sphericity** (для проверки пригодности факторного анализа для собранных данных).

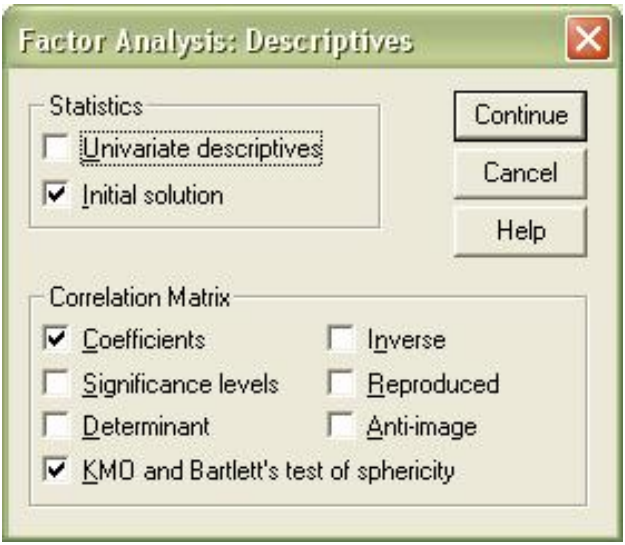


Рис. 9.1.3. Окно «*Factor Analysis: Descriptives*»

4. Щелкните на кнопке ***Extraction.*** Появится диалоговое окно ***Factor Analysis: Extraction*** (см. рис. 9.1.4)

o Оставьте установку ***Principal components*** (Анализ главных компонентов).

o Здесь количество факторов сознательно ограничим двумя. Щёлкните поэтому на опции ***Number of factors*** (количество факторов) и введите число 2.

o Щелчком на соответствующей опции деактивируйте вывод неповернутых значений факторов ***Unrotated factor solution***.

o Активируйте опцию ***Scree plot*** (Точечная диаграмма). Точечная диаграмма графически представляет собственные значения факторов, упорядоченные по величине.

Диалоговое окно ***Factor Analysis: Extraction*** (Факторный анализ: Извлечение) должно теперь выглядеть так, как представлено на рисунке 9.1.4.

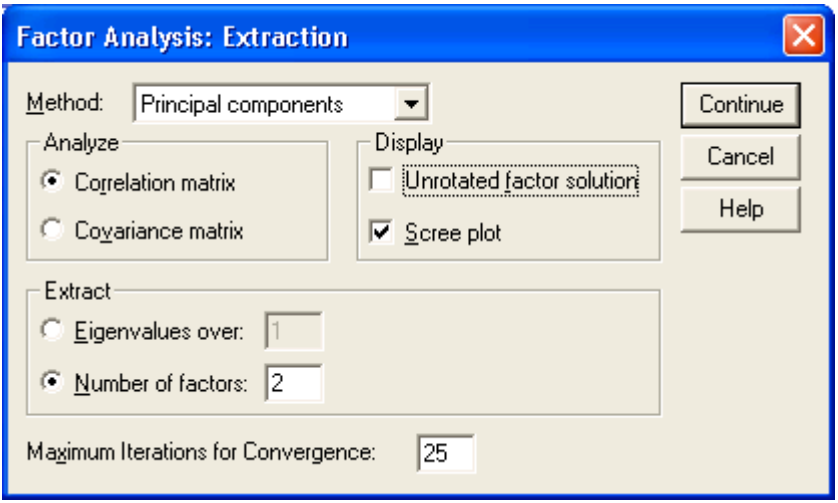


Рис. 9.1.4 Окно «*Factor Analysis: Extraction*»

5. Щелкните на кнопке ***Rotation*** **(**Вращение) и выберите метод варимакс (***Varimax***) в диалоговом окне «***Factor Analysis: Rotation***» (см. рис. 9.1.5).

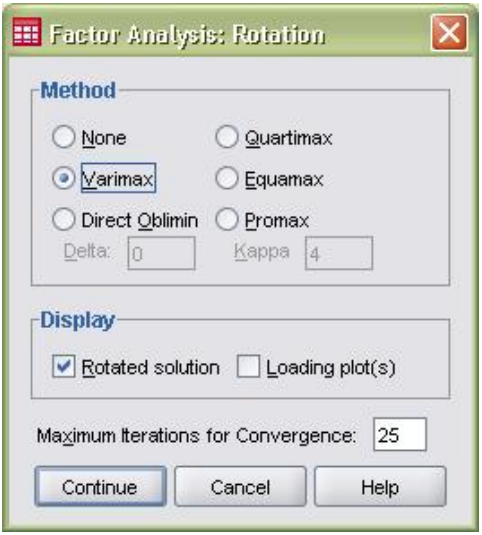


Рис. 9.1.5. Окно «*Factor Analysis: Rotation*»

6. Щелкнуть кнопку ***Scores*** и активировать ***Save as variables*** (Сохранить как переменные), чтобы рассчитанные значения факторов сохранить в виде дополнительных переменных (см. рис. 9.1.6).

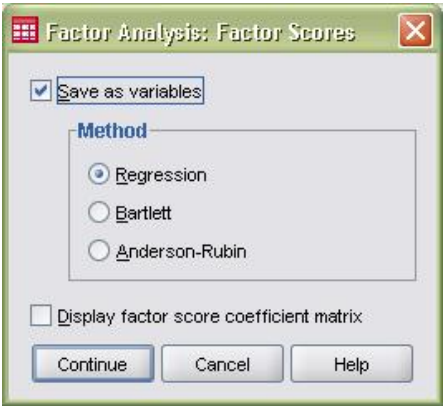


Рис. 9.1.6. Окно «*Factor Analysis: Factor Scores* »

7. Щелкнуть кнопку ***Options***..., чтобы организовать вывод коэффициентов, отсортированных по размеру (см. рис. 9.1.7).

- Активировать опцию ***Sorted by size*** (отсортированные по размеру).
- Запретить вывод малых факторных нагрузок. Для этого активировать опцию ***Suppress absolute values less than:*** (не выводить абсолютные значения меньшие, чем:), и ввести предельное значение 0,40**.**

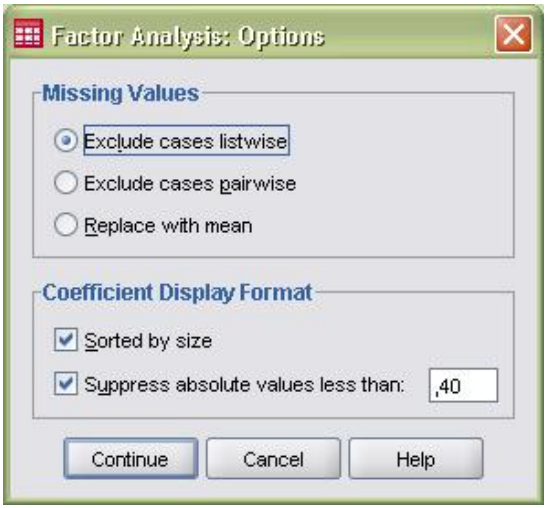


Рис. 9.1.7. Окно «*Factor Analysis: Options* »

8. Рассмотрите результаты в окне вывода. Они содержат:

- корреляционную матрицу (рис. 9.1.8);
- значение критерия Бартлетта (см. рис. 9.1.9);
- % дисперсии, объясняемой моделью (рис. 9.1.11 и 9.1.12);
- диаграмму «каменная осыпь» (рис. 9.1.10);
- повернутую матрицу факторных нагрузок (рис. 9.1.13).

Корреляционная матрица

Correlation Matrix

		1. важно приобрести зубную пасту,которая предотвращает развитие кариеса;	2. мне нравится зубная паста,которая придает зубам белизну;	3. зубная паста должна укреплять десна;	4.я предпочитаю зубную пасту,которая освежает дыхание;	5. предотвращение порчи зубов не является важным преимуществом данной зубной пасты;	6. наиболее важной причиной покупки данной зубной пасты является способность зуб-
Correlation	1. важно приобрести зубную пасту,которая предотвращает развит кариеса;	1.000	-.041	.873	-.086	-.858	.004
	2. мне нравится зубная паста,которая придает зубам белизну;	-.041	1.000	-.143	.590	-.007	.713
	3. зубная паста должна укреплять десна;	.873	-.143	1.000	-.248	-.778	-.018
	4.я предпочитаю зубну пасту,которая освежае дыхание;	-.086	.590	-.248	1.000	-.007	.640
	5. предотвращение порчи зубов не являетс важным преимущество данной зубной пасты;	-.858	-.007	-.778	-.007	1.000	-.136
	6. наиболее важной причиной покупки данной зубной пасты является способность зуб-	.004	.713	-.018	.640	-.136	1.000

Рис. 9.1.8. Фрагмент окна вывода. Корреляционная матрица

Как видно из рис. 9.1.8, относительно высокое значение коэффициентов корреляции наблюдается между v1(предотвращение кариеса), v3 (укрепление десен) и v5 (предотвращение порчи зубов). Можно ожидать, что эти переменные коррелируют с одним и тем же набором факторов.

Аналогично, относительно высокие корреляции наблюдаются между v*2* (отбеливание зубов), v4 (свежее дыхание) и v6 привлекательность внешнего вида зубов). Также можно ожидать, что эти переменные коррелируют с одними и теми же факторами.

Критерий сферичности Бартлетта

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.655
Bartlett's Test of Sphericity	Approx. Chi-Square	115.879
	df	15
	Sig.	.000

Рис. 9.1.9. Фрагмент окна вывода.
Критерий сферичности Бартлетта

Как видно из рис. 9.1.9, приближенное значение статистики χ^2 (Approx. Chi-Sguare) равно 115,879 с 15-ю степенями свободы, она является значимой при уровне <0,05.

Значение **статистики КМО** (Kaiser-Meyer-Olkin Measure of Sampling Adeguacy), равное 0,665, также большое (>0,5). Следовательно, факторный анализ можно рассматривать как приемлемый метод для анализа корреляционной матрицы, представленной на рис. 9.1.8.

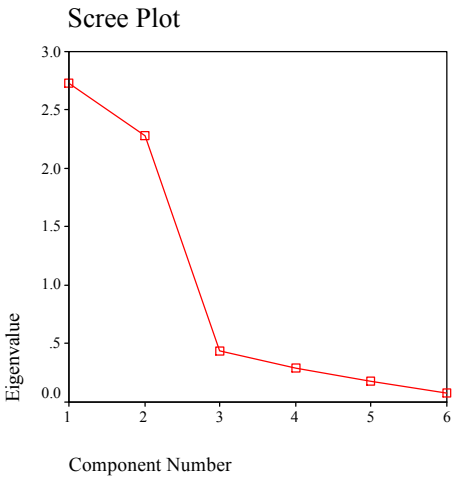


Рис. 9.1.10. Фрагмент окна вывода.
Диаграмма «Каменная осыпь»

«Каменистая осыпь» (рис. 9.1.10) представляет собой график зависимости собственных значений факторов от их номеров в порядке выделения.

Обычно график имеет четкий разрыв между крутой частью кривой, где факторам свойственны большие собственные значения, и плавной хвостовой частью (*осыпью)*. Опыт показывает, что точка, с которой начинается осыпь, указывает на действительное число факторов. В нашем случае осыпь начинается правее третьего фактора.

На рис. 9.1.11 представлен фрагмент окна вывода, в котором показаны так называемые **общности** (communalities). Общность – доля дисперсии отдельной переменной, которую она делит с другими переменными**.**

Communalities

	Initial	Extraction
1. важно приобрести зубную пасту,которая предотвращает развитие кариеса;	1.000	.926
2. мне нравится зубная паста,которая придает зубам белизну;	1.000	.769
3. зубная паста должна укреплять десна;	1.000	.894
4.я предпочитаю зубную пасту,которая освежает дыхание;	1.000	.727
5. предотвращение порчи зубов не является важным преимуществом данной зубной пасты;	1.000	.877
6. наиболее важной причиной покупки данной зубной пасты является способность зуб-	1.000	.817

Extraction Method: Principal Component Analysis.

Рис. 9.1.11. Фрагмент окна вывода. Общности

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.725	45.424	45.424	2.725	45.424	45.424
2	2.284	38.072	83.496	2.284	38.072	83.496
3	.434	7.241	90.737			
4	.289	4.821	95.557			
5	.183	3.057	98.614			
6	.083	1.386	100.000			

Extraction Method: Principal Component Analysis.

Рис. 9.1.12. Фрагмент окна вывода.
Суммарный процент объясняемой дисперсии

На рис. 9.1.12 представлен фрагмент окна вывода с суммарным процентом объясняемой факторной моделью дисперсии. Как видно из представленной на рис. 9.12 таблицы, имеются два собственных значения, больших единицы. Это указывает на необходимость извлечения двух факторов.

Следующий показатель показывает процент полной дисперсии, приписываемый каждому фактору. В нашем случае два фактора объясняют 83% дисперсии (что является более чем удовлетворительным), а три – более 90% (см. колонку Cumulative % на рис. 9.1.12).

На рис. 9.1.13 представлен фрагмент окна вывода с повернутой матрицей факторных нагрузок. Как видно из представленной на рисунке 9.1.13 таблицы, первая, третья и пятая переменные имеют высокие нагрузки по первому фактору. Вторая, четвертая и шестая – по второму. Факторы интерпретируют исходя из понимания тех переменных, которые его нагружают. Поэтому первый фактор можно интерпретировать как «здоровье зубов», а второй – «внешний вид зубов».

Таким образом, можно сделать вывод, что при выборе зубной пасты покупатели руководствуются двумя основными факторами: как паста влияет на здоровье зубов и как она влияет на их внешний вид.

Повернутая матрица факторных нагрузок

Rotated Component Matrix[a]

	Component	
	1	2
1. важно приобрести зубную пасту, которая предотвращает развитие кариеса;	,962	
3. зубная паста должна укреплять десна;	,935	
5. предотвращение порчи зубов не является важным преимуществом данной зубной пасты;	-,932	
6. наиболее важной причиной покупки данной зубной пасты является способность зуб-		,900
2. мне нравится зубная паста, которая придает зубам белизну;		,876
4.я предпочитаю зубную пасту, которая освежает дыхание;		,846

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Рис. 9.1.13. Фрагмент окна вывода.
Повернутая матрица факторных нагрузок

Задача для самостоятельного решения

В исследовании взаимосвязи между поведением в семье и поведением при покупке получены данные по семибалльной шкале (1 — «не согласен», 7 — «согласен») по следующим заявлениям о стиле жизни:

- Я лучше спокойно провел бы вечер дома, чем пошел на вечеринку.
- Я всегда проверяю цены, даже на товар с маленькой ценой
- Магазины меня интересуют больше, чем кинофильмы.
- Я не покупаю товары, которые рекламируют на бил-

лбордах.

- Я – домосед.
- Я экономлю, используя купоны для покупки товаров.
- Компании зря тратят большие деньги на рекламу.

Данные, полученные на выборке из 25-ти респондентов, приведены в таблице 9.1.3.

Таблица 9.1.3

Ответы респондентов на вопросы о поведении в семье и поведении при покупке

№	v1	v2	v3	V4	v5	v6	v7
1	6	2	1	6	5	3	5
2	5	7	5	6	6	6	4
3	5	3	4	5	6	6	7
4	3	2	2	5	1	3	2
5	4	2	3	2	2	1	3
6	2	6	2	4	3	7	5
7	1	3	3	6	2	5	7
8	3	5	1	4	2	5	6
9	7	3	5	3	5	2	4
10	6	3	3	4	4	6	6
11	6	6	2	6	4	4	7
12	3	2	2	7	6	1	6
13	5	7	6	2	2	6	1
14	6	3	5	5	7	2	3
15	3	2	4	3	2	6	5
16	2	7	5	1	4	5	2
17	3	2	2	7	2	4	6
18	6	4	5	4	7	3	3
19	7	2	6	2	5	2	1
20	5	6	6	3	4	5	3
21	2	3	3	2	1	2	6
22	3	4	2	1	4	3	6
23	2	6	3	2	1	5	3
24	6	5	7	4	5	7	2
25	7	6	5	4	6	5	3

Проведите факторный анализ данных и проинтерпретируйте его результаты.

Требования к отчету

Отчет по работе должен содержать:

– файлы с данными;

– файлы с результатами;

– ответы на контрольные вопросы.

Контрольные вопросы

1. Назначение факторного анализа.
2. Что такое матрица факторного отображения?
3. Как факторный анализ осуществляется в SPSS?
4. Какие кнопки расположены в главном окне факторного анализа и для чего они служат?
5. Какими способами определяется число факторов в модели факторного анализа?
6. Что показывает график «каменистая осыпь»?
7. Сколько факторов мы извлекли в примере с зубной пастой? Почему? Дайте им содержательную интерпретацию.

Практическая работа 9.2
Факторный анализ в R

Ход работы

1. Прочитайте данные из файла Excel в переменную data9.

```
> data9<-read.table("K:\\Rexample\\lab9.csv", sep=";", dec=",",
header=TRUE)
```

2. Вызовите функцию для анализа методом главных компонент (princomp) и выведите суммарные статистики (summary), включающие стандартное отклонение, процент, объясняемой дисперсии, накопленный процент объясняемой дисперсии.

```
> fit <- princomp(data9, cor=TRUE)
> summary(fit)

Importance of components:
                          Comp.1       Comp.2       Comp.3
Comp.4       Comp.5     Comp.6
Standard deviation       1.6508862 1.5113935 0.65914145
0.53780594 0.42824693 0.28838562
Proportion of Variance 0.4542375 0.3807184 0.07241124
0.04820587 0.03056591 0.01386104
Cumulative Proportion  0.4542375 0.8349559 0.90736718
0.95557305 0.98613896 1.00000000
```

Как видно из приведенных данных, процент объясняемой дисперсии для первого фактора 45,4%, для второго – 38,1%, вместе (накопленный процент) эти два фактора объясняют 83,5% наблюдаемой дисперсии.

Для выведения матрицы факторных весов используется команда loadings(fit):

```
> loadings(fit)
```

```
> loadings(fit)

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
v1  0.562 -0.171        -0.260  0.176  0.742
v2 -0.178 -0.547 -0.535 -0.602        -0.136
v3  0.567        -0.177  0.145  0.590 -0.520
v4 -0.211 -0.515  0.765 -0.137  0.255 -0.147
v5 -0.524  0.238 -0.182         0.744  0.286
v6 -0.115 -0.585 -0.242  0.728         0.237

               Comp.1 Comp.2 Comp.3 Comp.4 C
SS loadings     1.000  1.000  1.000  1.000
Proportion Var  0.167  0.167  0.167  0.167
Cumulative Var  0.167  0.333  0.500  0.667
```

Рис.9.2.1. Результаты выполнения команды loading

Для построения графика каменная осыпь используйте команду plot.

```
>plot(fit,type="lines")
```

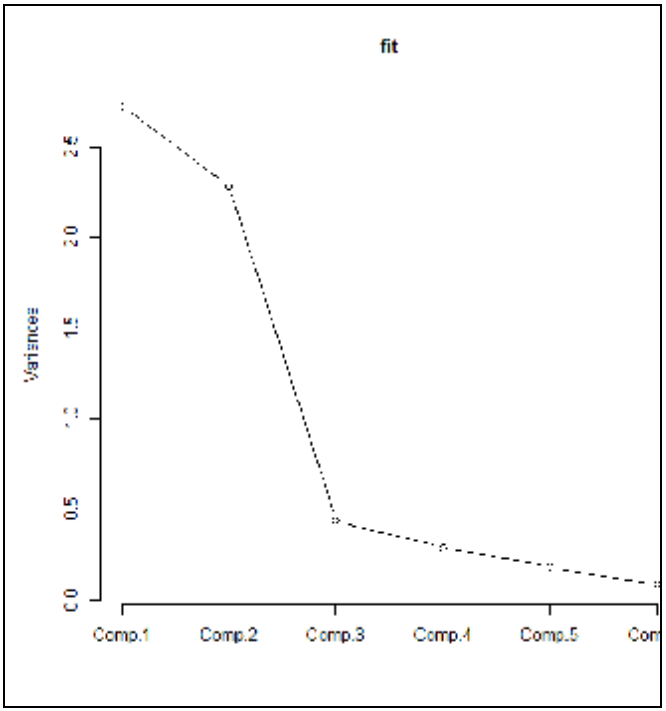


Рис.9.2.2. Диаграмма «Каменная осыпь»

Как видно из рисунка 9.2.2 каменная осыпь начинается с третьего фактора, так что оптимальным для нашего случая будет выбор двух факторов.

Для извлечения и вращения факторов можно использовать функцию **factanal ( ).** В самом общем случае при ее вызове указываются такие параметры, как данные, число факторов, которые необходимо извлечь, и метод вращения. Метод вращения может быть выбран один из следуюших: "none", "varimax", "quatimax", "promax", "oblimin", "simplimax", или "cluster" .

```
> fitAfterRotation <- factanal(data9,factors = 2, rotation = "varimax")
```

Напечатать повернутую матрицу факторных весов (для двух факторов), с подавлением печати факторных нагрузок, меньших, чем 0.3 и с упорядочением по убыванию:

```
> print(fitAfterRotation$loadings, cutoff = .30, sort = TRUE)
```

```
Loadings:
   Factor1 Factor2
v1  0.968
v3  0.896
v5 -0.887
v2           0.808
v4           0.747
v6           0.870

                 Factor1 Factor2
SS loadings        2.538   1.994
Proportion Var     0.423   0.332
Cumulative Var     0.423   0.755
```

Рис.9.2.3. Повернутая матрица факторных нагрузок (функция factanal)

Таким образом, мы видим, что первый фактор нагружают первый третий и пятый признаки, второй фактор – второй, четвертый и шестой. В соответствии со смыслом признаков-переменных, которые нагружают фактор, мы интерпрети-

руем первый фактор как «здоровье зубов», а второй фактор – как «красота зубов» (см. п.9).

Несколько другие, но похожие значения, дает использование функции *principal()*. Для ее работы необходимы два пакета *psych* и *GPArotation*. Вызываем ее следующей командой:

```
> fit <- principal(data9, nfactors=2, rotate="varimax")
```

А затем выводим полученную повернутую матрицу факторных нагрузок.

```
> fit
```

Результаты работы, приведенные на рис. ниже, совпадают в результатами, полученными нами ранее при выполнении факторного анализа методом главных компонент в SPSS (см. п.9).

```
> fit
Principal Components Analysis
Call: principal(r = data9, nfactors =
Standardized loadings (pattern matri
     RC1   RC2   h2    u2
v1  0.96 -0.02 0.93 0.074
v2 -0.04  0.88 0.77 0.231
v3  0.94 -0.14 0.89 0.106
v4 -0.11  0.85 0.73 0.273
v5 -0.93 -0.09 0.88 0.123
v6  0.07  0.90 0.82 0.183

                       RC1  RC2
SS loadings           2.69 2.32
Proportion Var        0.45 0.39
Cumulative Var        0.45 0.83
Proportion Explained  0.54 0.46
Cumulative Proportion 0.54 1.00
```

Рис. 9.2.4. Повернутая матрица факторных нагрузок (функция principal)

Контрольные вопросы

1. Как задаются данные для факторного анализа в R.
2. Как получить таблицу с процентом объясняемой дисперсии для каждого фактора?
3. Как определяется оптимальное число факторов? Как начертить график Scree plot?
4. Какие функции и из каких пакетов мы использовали? (заполните таблицу)

Функция	**Пакет**

5. Заполните таблицу с результатами расчетов для примера с зубной пастой

Название	**Объяснение или значение**
V1	
V2	
V3	
V4	
V5	
V6	
F1	
F2	
N (число объектов)	
% дисперсии, объясняемый 1 фактором	
% дисперсии, объясняемый 2 фактором	
Scree plot	

Литература

1. http://www.statmethods.net/advstats/factor.html
2. http://127.0.0.1:11083/library/stats/html/factanal.html
3. http://www.stat.cmu.edu/~cshalizi/350/2008/lectures/14/lecture-14.pdf

10. ИСПОЛЬЗОВАНИЕ MS EXCEL ДЛЯ РЕШЕНИЯ ЗАДАЧ МЕТОДОМ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Линейное программирование – метод, который широко применяют для принятия решений относительно наиболее эффективного использования ресурсов (денег, времени, техники, рабочей силы, сырья, производственных и складских помещений и т.п.).

Характерные признаки задач оптимизации, которые целесообразно решать методом линейного программирования:

1) наличие целевой функции – максимизировать, минимизировать или установить равным определенному числу результат (прибыль, цену, те или иные затраты и т.п.);

2) наличие определенных ограничений в возможности получить желаемый результат, сформулированный как целевая функция (например, фирма не в состоянии отремонтировать более тридцати автомобилей за неделю, в гостинице можна разместить не более двухсот гостей и т.п.);

3) наличие возможных альтернативных действий для выбора (например, если фирма производит четыре вида продукции, менеджеры используют линейное програмирование для того, чтобы определить, как наиболее рационально распределить ограниченные производственные ресурсы между этими видами продукции, чтобы получить максимальную прибыль);

4) целевая функция и ограничения должны быть описаны линейными уравнениями или неравенствами.

Алгоритм решения задач линейного программирования:

1) глубоко осознать сущность возникшей проблемы;
2) определить целевую функцию и ограничения;

3) определить переменные, анализ которых необходимо осуществить для принятия решения;

4) используя эти переменные, описать математически целевую функцию и ограничения;

5) применить MS Excel для решения задачи.

Для решения задач методом линейного программирования в MS Excel существует инструмент «**Поиск решения**» в меню «**Сервис**», с помощью которого можно решать задачи с использованием до 200 переменных, каждая из которых имеет по два ограничения и использовать дополнительно до 100 ограничений.

Процедура поиска решения позволяет найти оптимальное значение формулы, содержащейся в ячейке, которая называется целевой. Эта процедура работает с группой ячеек, прямо или косвенно связанных с формулой в целевой ячейке. Чтобы получить по формуле, содержащейся в целевой ячейке, заданный результат, процедура изменяет значения во влияющих ячейках.

Практическая работа 10
Использование MS Excel для решения задач методом линейного программирования

Цель работы – научиться использовать электронные таблицы MS Excel для решения задач методом линейного программирования.

Постановка задачи

Небольшая кондитерская фирма стремится получить максимальную прибыль, выпуская два вида продукции – торты и пирожные. Один торт дает фирме 6 гривень прибыли, одно пирожное – 1 гривню 50 копеек. На изготовление одного торта необходимо 4 часа, причем мощности фирмы таковы, что параллельно могут выпекаться 8 тортов. За это же время можно приготовить 50 пирожных. Таким образом, в среднем на изготовление одного торта тратится 0,5 человеко-часов (4/8), на изготовление одного пирожного – 0,08 человеко-часов (4/50). Кадровый ресурс фирмы – 48 человеко-часов в сутки. Спрос на продукцию фирмы – не более 50-ти тортов и не более 420-ти пирожных в сутки.

Целевая функция для такого примера имеет следующий вид:

$$6*X_1+1,5*X_2=max, \qquad (10.1)$$

где X_1 *– количество тортов,*

X_2 *– количество пирожных*

Ограничения:

$$0,5*X_1+0,08*X_2 \le 48 \qquad (10.2)$$

$$X_1 \le 50 \qquad (10.3)$$

$$X_2 \le 420 \qquad (10.4)$$

Ход работы

1. Запустить MS Excel.
2. Ввести данные в соответствии с рис. 10.1. Обратите внимание, что в ячейки B5 и C5 вводятся коэффициенты уравнения целевой функции; в ячейки B8 и C8 – коэффициенты неравенства, выражающего первое ограничение (неравенство *10.2*); в ячейки B6 и C6 – коэффициенты неравенства, выра-

жающего второе ограничение (неравенство *10.3*); в ячейки B7 и C7 – коэффициенты неравенства, выражающего третье ограничение (неравенство *10.4*).

	A	B	C	D
1	Расчет путей максимальной прибыли для кондитерской фирмы			
2				
3	Факторы, которые учитываются	Вид изделия		
4		торт (X1)	Пирожное (X2)	
5	прибыль	6	1,5	
6	спрос	1	0	
7		0	1	
8	ресурсы (кадры и время)	0,5	0,08	
9				
10	Результат			

Рис. 10.1. Ввод в MS Excel данных задачи

3. Ячейки B10 и C10 – «влияющие ячейки», в них будет получен искомый результат – X_1 – количество тортов (в ячейке B10) и X_2 – количество пирожных (ячейка C10), выпуск которых принесет наибольшую прибыль в заданных условиях. Перед началом решения задачи в каждую из этих ячеек необходимо ввести число «1» (см. рис. 10.2).

4. Ячейку D5 сделать целевой. В нее необходимо ввести выражение целевой функции (10.1). Процедура поиска решения заключается в том, ищется максимальное, минимальное или определенное значение формулы в целевой ячейке. Чтобы по формуле, содержащейся в целевой ячейке, получить заданный результат, процедура изменяет значения во влияющих ячейках (см. рис. 10.2).

	A	B	C	D	E	F
1	Расчет путей максимальной прибыли для кондитерской фирмы					
2						
3	Факторы, которые учитываются	Вид изделия				
4		торт (X1)	Пирожное (X2)			
5	прибыль	6	1,5	=B5*$B$10+C5*$C$10		
6	спрос	1	0	=B6*$B$10+C6*$C$10	<=	50
7		0	1	=B7*$B$10+C7*$C$10	<=	420
8	ресурсы (кадры и время)	0,5	0,08	=B8*$B$10+C8*$C$10	<=	48
9						
10	Результат	1	1			

Рис. 10.2. Ввод в MS Excel формул для решения задачи

5. В ячейки E6:E8 поставить знаки, соответствующие ограничениям в условиях задачи (в нашем случае все три условия звучали, как «не более», но вообще можно использовать «=», «≤», «≥»).

6. В ячейки F6:F8 ввести те значения, которые представлены в правых частях неравенств *10.2 – 10.4* (см. рис. 10.2). Теперь наша электронная таблица готова к применению процедуры «Поиск решения».

7. Зайти в меню «Сервис» и выбрать опцию «Поиск решения» (если она не представлена в меню «Сервис», войти в «Надстройки» и установить метку возле пункта «Поиск решения»). Откроется диалоговое окно «Поиск решения», представленное на рис. 10.3.

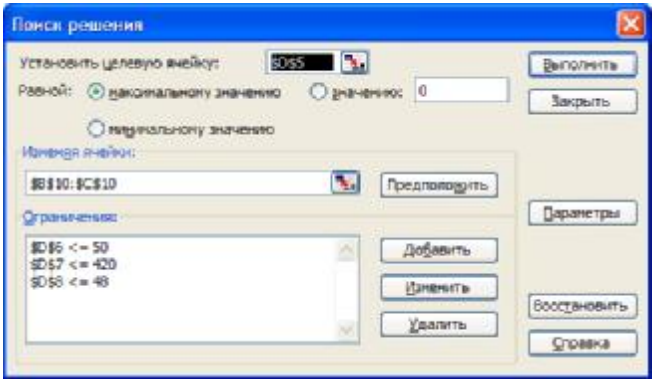


Рис. 10.3. Диалоговое окно «Поиск решения»

8. Установить параметры:

а) адрес целевой ячейки (для нашей задачи – это ячейка D5);

б) то, чему должно быть равно полученное в целевой ячейке значение – максимуму, минимуму или определенному значению (в нашем – случае – мы хотим сделать прибыль максимальной, т.е. максимизировать целевую функцию);

в) адреса влияющих ячеек – опция «изменяя ячейки»;

г) ограничения (после нажатия кнопки «Добавить» появляется диалоговое окно «Добавление ограничения», представленное на рис. 10.4; после введения каждого ограничения нажимают кнопку «добавить»; после того, как введены все ограничения, нажимают кнопку «ОК»).

9. После нажатия кнопки «ОК» снова откроется диалоговое окно «Поиск решения», представленное на рис. 10.3. Далее необходимо запустить процедуру «Поиск решения», нажав кнопку «Выполнить».

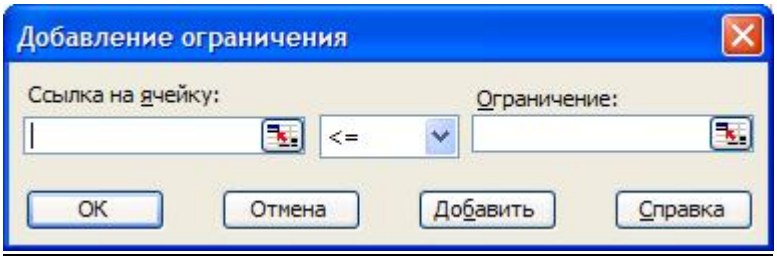


Рис. 10.4. Диалоговое окно «Добавление ограничения»

10. Результаты расчетов представлены в ячейках B10 и C10 (соответственно, количество тортов и пирожных, которое целесообразно выпускать в сутки) и D5 (максимальная прибыль в гривнях) (рис. 10.5).

	A	B	C	D	E	F
1	Расчет путей максимальной прибыли для кондитерской фирмы					
2						
3	Факторы, которые учитываются	Вид изделия				
4		торт (X1)	Пирожное (X2)			
5	прибыль	6	1,5	**803**		
6	спрос	1	0	29	<=	50
7		0	1	420	<=	420
8	ресурсы (кадры и время)	0,5	0,08	48	<=	48
9						
10	Результат	**29**	**420**			

Рис. 10.5. Результаты расчетов

Задачи для самостоятельного решения

10.1. Руководитель частного вуза формирует учебный план двухступенчатой магистров. Каждый обучающийся на первой ступени (бакалаврат) должен прослушать не менее 35-ти дисциплин и на второй ступени обучения (магистратура) не менее 15-ти дисциплин. Кроме того, в контракте каждому обучающемуся гарантировано, что в сумме он имеет право прослушать не более 60-ти дисциплин. Каждый курс первой ступени обучения обходится вузу в 50000 гривень (зарплата преподавателей и обслуживающего персонала, методическая литература, оборудование, аренда помещений и т.п.), каждый курс второй ступени обучения – в 75000 гривень. Сколько дисциплин нужно включить в учебный план на каждой из ступеней обучения, чтобы минимизировать затраты вуза? Решить задачу, используя инструмент «Поиск решения» MS Excel.

Подсказка:

пусть X_1 – искомое число дисциплин на первой ступени обучения; X_2 – искомое число дисциплин на второй ступени обучения

целевая функция: $50000 * X_1 + 75000 * X_2 = min\$$

ограничения:

$X_1 \geq 35$; $X_2 \geq 15$; $X_1 + X_2 \leq 60$

10.2. Решить задачу 10.1, изменив условия следующим образом: стоимость одной дисциплины на первой ступени обучения увеличилась на 1000 гривень, стоимость одной дисциплины на второй ступени обучения не изменилась, минимальное количество дисциплин, которые изучаются на первой ступени обучения, стало равным сорока, а на второй – не изменилось.

Требования к отчету

Отчет должен содержать:

- ответы на контрольные вопросы;
- файл с результатами расчетов.

Контрольные вопросы

1. Каковы характерные признаки задач оптимизации, которые целесообразно решать методом линейного программирования?
2. Что такое целевая функция?
3. Что такое ограничения?
4. С помощью какого инструмента MS Excel решаются задачи оптимизации?
5. Опишите назначения полей диалогового окна «Поиск решения».
6. Опишите назначения полей диалогового окна «добавление ограничений».
7. Дайте содержательную интерпретацию результатов решения задач 10.1 и 10.2.

11. ИСПОЛЬЗОВАНИЕ MS EXCEL ДЛЯ ПОСТРОЕНИЯ «ТАБЛИЦ ПРИНЯТИЯ РЕШЕНИЙ»

Альтернатива – вариант действий или стратегия, которую выбирает принимающий решение.

Исход – возможное последствие каждой альтернативы.

Основные стадии принятия решений:

1. Четко сформулировать задачу.
2. Определить возможные альтернативы, не упустив ни одной.
3. Определить возможные исходы, не упустив ни одного.
4. Представить в виде таблицы перечень всех комбинаций всех возможных альтернатив и исходов.
5. Выбрать одну из моделей математической теории принятия решений.
6. Применить выбранную модель и получить решение.

Ситуации принятия решения:

1. ***Принятие решения в условиях определенности*** – принимающему решение известны исходы (следствия) каждой альтернативы.
2. ***Принятие решения в условиях риска*** – когда существуют несколько возможных исходов для каждой альтернативы и принимающему решение известны вероятности наступления каждого исхода. В этом случае, как правило, применяют один из 2-х равнозначных критериев:
   a) максимизировать ожидаемую прибыль (выгоду);
   b) минимизировать ожидаемые потери.
3. ***Принятие решения в условиях неопределенности*** – когда существуют несколько возможных исходов для каждой альтернативы и принимающему решение неизвестны вероятности возможных исходов.

Практическая работа 11
Использование MS Excel для построения «таблиц принятия решений»

Цель работы: научиться использовать MS Excel для построения «таблиц принятия решений», выбора наилучшего решения на основе произведенных расчетов.

Постановка задачи

Сергей Коваленко, студент Луганского национального университета имени Тараса Шевченко, может добираться на занятия тремя способами:

1) пройти от своего дома примерно 400 метров до остановки маршрутки №122, доехать на ней до спортивного корпуса и пройти еще 200 метров до учебного корпуса;

2) пройти от своего дома примерно 500 метров до остановки маршрутки №129 и доехать до ней прямо до учебного корпуса;

3) сесть на остановке возле своего дома в маршрутку №170 и доехать в ней до остановки, которая находится напротив учебного корпуса. Для того, чтобы дойти до этого корпуса, Сергею нужно перейти дорогу шириной 45 метров.

Длина каждого из маршрутов разная. Кроме того, в разное время и в разные дни бывают разные ситуации на дорогах. Известно, что при условии благоприятной ситуации на самых насыщенных транспортом улицах нашего города – Советской и Оборонной – самым быстрым является первый вариант. Когда на этих улицах пробки, предпочтительнее второй или третий варианты. Сергей решил провести исследование и найти лучший вариант пути. В течение трех месяцев он добирался в университет разными способами и записывал время, которое уходило на дорогу. В таблице 11.1 приведено среднее время в минутах, которое Сергей тратил на все варианты пути в условиях разной загруженности основных дорог, по которым проходят маршруты.

Таблица 11.1

Фрагмент 1 «таблицы принятия решения».
Исходные данные

Варианты маршрута	Среднее время на маршрут (минуты)		
	дороги практически свободны	средняя загруженность дорог транспортом	«пробки» на основных дорогах
№122	15	30	45
№129	20	25	35
№170	30	30	30

Сергей зафиксировал, что за 90 дней его исследований пробки на дорогах были в течение 10-ти дней, средняя загруженность транспортом на дорогах была в течение 60-ти дней. Будем считать, что в течение этих 90 дней ситуация на дорогах была типичной для нашего города. Требуется:

а) составить «таблицу принятия решения»;

б) определить, какой вариант пути стоит выбрать Сергею?

Ход работы

1. Запустить Excel. Еще раз внимательно прочитать условие задачи и продумать, как будет выглядеть итоговая «таблица принятия решения», как рациональнее разместить ее фрагменты.

2. Ввести фрагмент 1 «таблицы принятия решения» (таб. 11.1).

3. Заполнить фрагмент 2 «таблицы принятия решения» (табл. 11.2), произвести расчет вероятностей, подставляя в формулы расчета адреса ячеек, в которых содержатся нужные значения.

4. Заполнить фрагмент 3 «таблицы принятия решения» (табл. 11.3), произвести расчет ожидаемых значений, подставляя в формулы расчета адреса ячеек, в которых содержатся

нужные числа. Помните, что ожидаемое значение = среднее значение*вероятность.

5. Определить лучший вариант пути как вариант с минимальным суммарным ожидаемым значением времени.

6. Заполнить фрагмент 4 «таблицы принятия решения» (табл. 11.4) и проанализировать его.

Таблица 11.2

Фрагмент 2 «таблицы принятия решения». Расчет вероятностей

	дороги практически свободны	**средняя загруженность дорог транспортом**	**«пробки» на основных дорогах**	**Всего**
Число дней	20 (90 – 10 – 60)	60 (дано по условию задачи)	10 (дано по условию задачи)	90 (дано)
Вероятность (расчет)	20/90	60/90	10/90	90/90
Вероятность (итог) – в таблицу не вводить!	*0,22*	*0,67*	*0,11*	*1,00*

Таблица 11.3

Фрагмент 3 «таблицы принятия решения». Расчет ожидаемых значений

Варианты маршрута	**Среднее время на маршрут (минуты)**			**Суммарное ожидаемое время (мин.)**
	дороги практически свободны	**средняя загруженность дорог транспортом**	**«пробки» на основных дорогах**	
№122	15*0,22	30*0,67	45*0,11	
№129	20*0,22	25*0,67	35*0,11	
№170	30*0,22	30*0,67	30*0,11	

Лучшим считается путь, суммарное ожидаемое значение времени для которого является минимальным.

Таблица 11.4

Фрагмент 4 «таблицы принятия решения».
Лучшие варианты пути для разных условий на дорогах

Ситуация на дорогах	Лучший вариант пути	Время в пути (мин.)	Вероятность
дороги практически свободны			
средняя загруженность дорог транспортом			
«пробки» на основных дорогах			

Задача для самостоятельного решения

Приобретая новый компьютер, клиент стоит перед выбором:

а) заключить контракт на полное бесплатное сервисное обслуживание сроком на 1 год, заплатив при этом 600 гривень дополнительно к цене компьютера;

б) заключить контракт на частичное сервисное обслуживание сроком на 1 год, заплатив при этом 300 гривень дополнительно к цене компьютера. В том случае, если в течение года после покупки компьютера ему потребуется крупный ремонт, покупателю придется заплатить за этот ремонт 1500 гривень;

в) не заключать контракт на сервисное обслуживание. В этом случае покупатель не платит ни копейки дополнительно к цене компьютера, но в случае, если потребуется крупный ремонт, он должен будет заплатить за этот ремонт 3000 гривень.

Вероятность того, что крупный ремонт компьютеру все-таки потребуется, – 20% (0,2). Как поступить покупателю? Что является альтернативами в этом примере и что является возможными исходами?

Создайте таблицу принятия решения, ориентируясь на таблицу 11.5, произведите необходимые расчеты и сделайте вывод. Помните, что ожидаемое значение «стоимости» каждой альтернативы = (стоимость первого исхода*вероятность первого исхода) + (стоимость второго исхода*вероятность второго исхода) + …(стоимость последнего возможного исхода*вероятность последнего возможного исхода).

Таблица 11.5

Расчеты для принятия решения при покупке компьютера

Вариант поведения покупателя	**Себестоимость возможных исходов (грн.)**		**Ожидаемое значение «стоимости» альтернативы**
	Обслуживание потребуется	**Обслуживание не потребуется**	
Не заключать соглашения о сервисном обслуживании	3000	0	3000*0,2+0*0,2
Заключить соглашение о частичном сервисном обслуживании	1500	300	300+240
Заключить соглашение о полном сервисном обслуживании	600	600	
Вероятность исхода	.20	.80	–

Контрольные вопросы

1. Что такое альтернатива в теории принятия решений?
2. Что такое исход в теории принятия решений?
3. Почему важно при принятии решения учесть все возможные альтернативы и все возможные исходы для каждой альтернативы?
4. Что представляет собой «таблица принятия решения»?
5. Обоснуйте выбор критериев, по которым выбиралось наилучшее решение в задачах.

Таблицы критических значений
Критические значения критерия Пирсона χ^2

df	P			df	P			df	P		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
1	3,842	6,635	10,829	**31**	44,993	52,203	61,118	**61**	80,232	89,591	100,887
2	5,992	9,211	13,817	**32**	46,202	53,498	62,508	**62**	81,381	90,802	102,165
3	7,815	11,346	16,269	**33**	47.408	54.789	63.891	**63**	82,529	92,010	103,442
4	9,488	13,278	18,470	**34**	48.610	56.074	65.269	**64**	83,675	93,217	104,717
5	11,071	15,088	20,519	**35**	49.810	57.356	66.641	**65**	84,821	94,422	105,988
6	12,593	16,814	22,462	**36**	51.007	58.634	68.008	**66**	85,965	95,626	107,257
7	14,068	18,478	24,327	**37**	52.201	59.907	69.370	**67**	87,108	96,828	108,525
8	15,509	20,093	26,130	**38**	53.393	61.177	70.728	**68**	88,250	98,028	109,793
9	16,921	21,669	27,883	**39**	54.582	62.444	72.080	**69**	89,391	99,227	111,055
10	18,309	23,213	29,594	**40**	55.768	63.707	73.428	**70**	90,531	100,425	112,317
11	19,677	24,729	31,271	**41**	56.953	64.967	74.772	**71**	91,670	101,621	113,577
12	21,028	26,221	32,917	**42**	58.135	66.224	76.111	**72**	92,808	102,816	114,834
13	22,365	27,693	34,536	**43**	59.314	67.477	77.447	**73**	93,945	104,010	116,092
14	23,688	29,146	36,132	**44**	60.492	68.728	78.779	**74**	95,081	105,202	117,347
15	24,999	30,583	37,706	**45**	61.668	69.976	80.107	**75**	96,217	106,393	118,599
16	26,299	32,006	39,262	**46**	62,841	71,221	81,431	**76**	97,351	107,582	119,850
17	27,591	33,415	40,801	**47**	64,013	72,463	82,752	**78**	99,617	109,958	122,347
18	28,873	34,812	42,323	**48**	65,183	73,703	84,069	**79**	100.749	111.144	123.595
19	30,147	36,198	43,832	**49**	66,351	74,940	85,384	**80**	101.879	112.329	124.839
20	31,415	37,574	45,327	**50**	67,518	76,175	86,694	**90**	113.145	124.116	137.208
21	32,675	38,940	46,810	**51**	68,683	77,408	88,003	**100**	124.342	135.807	149.449
22	33,929	40,298	48,281	**52**	69,846	78,638	89,308	**110**	135.480	147.414	161.582
23	35,177	41,647	49,742	**53**	71,008	79,866	90,609	**120**	146.567	158.950	173.618

24	36,420	42,989	51,194	**54**	72,168	81,092	91,909	**130**	157.610	170.423	185.573
25	37,658	44,324	52,635	**55**	73,326	82,316	93,205	**140**	138.613	181.841	197.450
26	38,891	45,652	54,068	**56**	74,484	83,538	94,499	**150**	179.581	193.207	209.265
27	40,119	46,973	55,493	**57**	75,639	84,758	95,790	**200**	233.994	249.445	267.539
28	41,343	48,289	56,910	**58**	76,794	85,976	97,078	**250**	287.882	304.939	324.831
29	42,564	49,599	58,320	**59**	77,947	87,192	98,365	**300**	341.395	359.906	381.424
30	43,780	50,904	59,722	**60**	79,099	88,406	99,649	**350**	394.626	414.474	437.487

Критические значения коэффициента корреляции r Пирсона

(для проверки ненаправленных альтернатив, n — объем выборки)

n	**P**			**n**	**P**			**n**	**P**		
	0,05	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**		**0,05**	**0,01**	**0,001**
5	0,878	0,959	0,991	**33**	0,344	0,442	0,547	**61**	0,252	0,327	0,411
6	0,811	0,917	0,974	**34**	0,339	0,436	0,539	**62**	0,250	0,325	0,408
7	0,754	0,875	0,951	**35**	0,334	0,430	0,532	**63**	0,248	0,322	0,405
8	0,707	0,834	0,925	**36**	0,329	0,424	0,525	**64**	0,246	0,320	0,402
9	0,666	0,798	0,898	**37**	0,325	0,418	0,519	**65**	0,244	0,317	0,399
10	0,632	0,765	0,872	**38**	0,320	0,413	0,513	**66**	0,242	0,315	0,396
11	0,602	0,735	0,847	**39**	0,316	0,408	0,507	**67**	0,240	0,313	0,393
12	0,576	0,708	0,823	**40**	0,312	0,403	0,501	**68**	0,239	0,310	0,390
13	0,553	0,684	0,801	**4!**	0,308	0,398	0,495	**69**	0,237	0,308	0,388
14	0,532	0,661	0,780	**42**	0,304	0,393	0,490	**70**	0,235	0,306	0,385
15	0,514	0,641	0,760	**43**	0,301	0,389	0,484	**80**	0,220	0,286	0,361
16	0,497	0,623	0,742	**44**	0,297	0,384	0,479	**90**	0,207	0,270	0,341

Критические значения коэффициента корреляции r Пирсона (продолжение)

17	0,482	0,606	0,725	**45**	0,294	0,380	0,474	**100**	0,197	0,256	0,324
18	0,468	0,590	0,708	**46**	0,291	0,376	0,469	**110**	0,187	0,245	0,310
19	0,456	0,575	0,693	**47**	0,288	0,372	0,465	**120**	0,179	0,234	0,297
20	0,444	0,561	0,679	**48**	0,285	0,368	0,460	**130**	0,172	0,225	0,285
21	0,433	0,549	0,665	**49**	0,282	0,365	0,456	**140**	0,166	0,217	0,275
22	0,423	0,537	0,652	**50**	0,279	0,361	0,451	**150**	0,160	0,210	0,266
23	0,413	0,526	0,640	**51**	0,276	0,358	0,447	**200**	0,139	0,182	0,231
24	0,404	0,515	0,629	**52**	0,273	0,354	0,443	**250**	0,124	0,163	0,207
25	0,396	0,505	0,618	**53**	0,271	0,351	0,439	**300**	0,113	0,149	0,189
26	0,388	0,496	0,607	**54**	0,268	0,348	0,435	**350**	0,105	0,138	0,175
27	0,381	0,487	0,597	**55**	0,266	0,345	0,432	**400**	0,098	0,129	0,164
28	0,374	0,479	0,588	**56**	0,263	0,341	0,428	**450**	0,092	0,121	0,155
29	0,367	0,471	0,579	**57**	0,261	0,339	0,424	**500**	0,088	0,115	0,147
30	0,361	0,463	0,570	**58**	0,259	0,336	0,421	**600**	0,080	0,105	0,134

Критические значения коэффициента ранговой корреляции Спирмена

(по В.Ю. Урбаху, 1964)

Связь достоверна, если $r_{s\ эмп} \geq r_{s\ 0.05}$, и тем более достоверна, если $r_{s\ эмп} \geq r_{s\ 0.01}$.

n	*p*	
	0.05	**0.01**
5	0.94	-
6	0.85	-
7	0.78	0.94
8	0.72	0.88
9	0.68	0.83
10	0.64	0.79
11	0.61	0.76
12	0.58	0.73
13	0.56	0.70
14	0.54	0.68
15	0.52	0.66
16	0.50	0.64

n	*p*	
	0.05	**0.01**
17	0.48	0.62
18	0.47	0.60
19	0.46	0.58
20	0.45	0.57
21	0.44	0.56
22	0.43	0.54
23	0.42	0.53
24	0.41	0.52
25	0.39	0.51
26	0.39	0.50
27	0.38	0.49
28	0.38	0.48

n	*p*	
	0.05	**0.01**
29	0.37	0.48
30	0.36	0.47
31	0.36	0.46
32	0.36	0.45
33	0.34	0.45
34	0.34	0.44
35	0.33	0.43
36	0.33	0.43
37	0.33	0.43
38	0.32	0.41
39	0.32	0.41
40	0.31	0.40

Источник таблиц: *Сидоренко Е. В.* Методы математической обработки в психологии - СПб.: ООО "Речь", 2000, .[30]

Литература

1. Адаменко Е.В. Математические методы в педагогике и психологии. – Луганск: Альма-матер, 2008. – 96 с.
2. Адаменко Е.В., Панченко Л.Ф. Кластерный анализ в педагогических исследованиях // Вісник Луганського держ. пед. ун-ту ім. Тараса Шевченка. – 2002. – №8(52). – С. 6–10.
3. Адаменко Е.В., Панченко Л.Ф. Обучение студентов университета использованию табличного процессора Excel для обработки результатов исследования // Наук.-метод. зб. статей за матеріалами семінару „Комп'ютерні та інноваційні технології у навчальному процесі (жовтень 2000)". – Алчевськ, 2000. – С. 56–60.
4. Адаменко Е.В., Панченко Л.Ф., Кондратенко П.В. Оценка формы распределения данных экспериментального исследования с помощью Excel // Наука на порозі нового тисячоліття: Матеріали наук. конф. – Луганськ: Альма-матер, 2001. – С. 4–5.
5. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. – М.:Финансы и статистика, 1989. – 607 с.
6. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985. – 487 с.
7. Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров. – М.: Мир, 2003. – 688 с.
8. Бююль А., Цефель П. SPSS: искусство обработки информации. – СПб.: ДиаСофтЮП, 2001. – 608с.
9. Винстон У.А. Microsoft Excel: анализ данных и построение бизнес моделей. – М.: Русская редакция, 2005. – 576 с.
10. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. – М.: Прогресс, 1976. – 496 с.

11. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 2003. – 352 с.

12. Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128 с.

13. Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных. – М.: Финансы и статистика, 1988. – 254 с.

14. Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики. – М.: ИНФРАМ, 2000. – 416 с.

15. Иберла К. Факторный анализ. – М.: Статистика, 1980. – 398 с.

16. Кендэлл М. Ранговые корреляции. – М.: Статистика, 1975. –212 с.

17. Малхотра Н.К. Маркетинговые исследования. Практическое руководство, 3-е издание. – М.: Издательский дом «Вильямс», 2002. – 960 с.

18. Мидлтон М.Р. Анализ статистических данных с использованием Microsoft Excel. – М.: Бином, 2005. – 296 с.

19. Наследов А.Д. SPSS: компьютерный анализ данных в психологии и социальных науках. – СПб.: Питер, 2005. – 416 с.

20. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. – К.: Наукова думка, 1982. – 272 с.

21. Паніотто В.І., Максименко В.С., Харченко Н.М. Статистичний аналіз соціологічних даних. – К.: КМ Академія, 2004. – 270 с.

22. Панченко Л. Ф. Використання вільного програмного забезпечення для навчання студентів аналізу даних // Вісн. Луган. нац. ун-ту імені Тараса Шевченка : Педагогічні науки. – 2010. – № 17(204). – С. 67–77.

23. Панченко Л.Ф. Математические методы в психологии. – Луганск, Альма-матер, 2005. – 60 с.

24. Панченко Л. Ф., Е. В. Адаменко. Компьютерный анализ данных : учеб. пособие для студентов высш. учеб. заведений. – Луганск : Изд-во ГУ „ДЗ ЛНУ імені Тараса Шевченка”, 2010. – 188 с.

25. Панченко Л.Ф., Адаменко Е.В. Выявление различий в распределении признака с помощью критерия Пирсона хи-квадрат // 2001 год – итоги науки. Материалы научной конференции. – Луганск: Альма Матер, 2002. – С. 46–48.

26. Петрунин Ю.Ю. Информационные технологии анализа данных. – М.: КДУ, 2008. – 292 с.

27. Плис А.И., Сливина Н.А. Практикум по прикладной статистике в среде SPSS. – М.: Финансы и статистика, 2004. – 284 с.

28. Поллард Дж. Справочник по вычислительным методам статистики. – М.: Финансы и статистика, 1982. – 344 с.

29. Саймон Д. Анализ данных в Excel. – М.: Вильямс, 2004. – 528 с.

30. Сидоренко Е.В. Методы математической обработки в психологии. – СПб.: Речь, 2000. – 350 с.

31. Статистическое обеспечение маркетинга продукта. – М.: МЭСИ, 2000. – 150 с.

32. Таганов Д.Н. SPSS: статистический анализ в маркетинговых исследованиях. – СПб.: Питер, 2004. – 192 с.

34. Технологии анализа данных / Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. – СПб.: BHV-Петербург, 2007. – 384 с.

35. Тюрин Ю.Н. Анализ данных на компьютере. – М.: Инфра-М, 2003. – 544 с.

36. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 216 с.

37. Чекотовский Э.В. Графический анализ статистических данных в Microsoft Excel 2000. – К.: Диалектика, 2002. – 464 с.

38. Харман Г. Современный факторный анализ. – М.: Статистика, 1972. – 489 с.

39. Шнейдерман Б. Психология программирования: Человеческие факторы в вычислительных и информационных системах. – М.: Радио и связь, 1984. – 304 с.

40. Эддоус М., Стэнсфилд Р. Методы принятия решений. – М.: ЮНИТИ, 1997. – 590 с.

41. StatSoft, Inc. (2001). Электронный учебник по статистике. Москва, StatSoft. – http://www.statsoft.ru/home/textbook/default.htm.

Дополнительная литература

42. Шипунов А.Б. и др. Наглядная статистика. Используем R! – М.: ДМК ПРЕСС, 2012. – 298 с.

43. R–project [Електронний ресурс]. – Режим доступу: http://www.r-project.org/about.html.

44. Quick-R [Електронний ресурс]. – Режим доступу: lhttp://www.statmethods.net/

45. Using R for Multivariate Analysis [Електронний ресурс]. – Режим доступу: http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

46. Рекомендации по преподаванию программной инженерии и информатики в университетах = Software Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering; Computing Curricula 2001: Computer Science: пер.с англ. – М.: ИНТУИТ.РУ «Интернет-Университет Информационных Технологий », 2007.

Панченко Л. Ф. Практикум з аналізу даних. – Навчальний посібник.

У навчальному посібнику розглядаються теоретичні та практичні питання комп'ютерного аналізу даних з використанням електронних таблиць MS Excel, статистичного пакету SPSS, середовища R. Посібник включає такі розділи аналізу як описова статистика, основи кореляційного й регресійного аналізу, перевірка гіпотез та дисперсійний аналіз, методи багатовимірного аналізу (дискримінантний, факторний і кластерний).

Навчальний посібник призначений для студентів та магістрантів спеціальностей „Інформатика", а також усіх інших спеціальностей, які пов'язані з аналізом даних на комп'ютері.

Ключові слова: аналіз даних, комп'ютерний аналіз даних, описова статистика, кореляція, регресія, дисперсійний аналіз, перевірка гіпотез, методи багатовимірного аналізу, Microsoft Excel, SPSS, R.

Панченко Л. Ф. Практикум по анализу данных. – Учебное пособие.

В учебном пособии рассматриваются теоретические и практические вопросы компьютерного анализа данных с использованием электронных таблиц MS Excel, статистического пакета SPSS, свободно распростарняемой среды R. Пособие включает такие разделы анализа как описательная статистика, основы корреляционного и регрессионного анализа, проверка гипотез и дисперсионный анализ, методы многомерного анализа (дискриминантный, кластерный и факторный анализ).

Учебное пособие предназначено для студентов и магистрантов специальностей „Информатика", а также всех других специальностей, которые связаны с анализом данных на компьютере.

Ключевые слова: анализ данных, компьютерный анализ данных, описательная статистика, корреляция, регрессия, про-

верка гипотез, дисперсионный анализ, дискриминантный анализ, кластерный анализ, факторный анализ, Microsoft Excel, SPSS, R.

Panchenko L. F. Data analysis practicum. – Textbook.

The textbook discusses the theoretical and practical aspects of computer data analysis using spreadsheets Microsoft Excel, the statistical package SPSS, environment R. The manual includes such sections as descriptive statistics, the basics of correlation and regression analysis, hypothesis testing and analysis of variance, multivariate analysis (discriminant analysis, cluster analysis and factor analysis).

The textbook is addressed for "Computer Science" students as well as for students of other specialties that are associated with the computer data analysis.

***Key words**:* data analysis, computer data analysis, descriptive statistics, correlation, regression, ANOVA, hypothesis testing, discriminant analysis, cluster analysis, factor analysis, Microsoft Excel, SPSS, R.

Навчальне видання

ПАНЧЕНКО Любов Феліксівна

ПРАКТИКУМ З АНАЛІЗУ ДАНИХ

Навчальний посібник
для студентів вищих навчальних закладів

Російською мовою

У навчальному посібнику розглядаються теоретичні та практичні питання комп'ютерного аналізу даних з використанням електронних таблиць MS Excel, статистичного пакету SPSS, середовища R, які складаються з таких розділів аналізу як описова статистика, основи кореляційного й регресійного аналізу, перевірки гіпотез, однофакторного й двохфакторного дисперсійного аналізу, методів багатовимірного аналізу.

Навчальний посібник призначений для студентів та магістрантів спеціальностей „Інформатика", а також усіх інших спеціальностей, які пов'язані з аналізом даних на комп'ютері.

За редакцією автора
Комп'ютерний макет – Л. Ф. Панченко

Здано до склад. 07.05.2013 р. Підп. до друку 5.06.2013 р.
Формат 60x84 1/16. Папір офсет. Гарнітура Times New Roman.
Друк ризографічний. Ум. друк. арк. 15,64. Наклад 300 прим. Зам. № 142.

Видавець і виготовлювач
Видавництво Державного закладу
„Луганський національний університет імені Тараса Шевченка"
вул. Оборонна, 2, м. Луганськ, 91011. т/ф: (0642) 58-03-20.
e-mail: alma-mater@list.ru
Свідоцтво суб'єкта видавничої справи ДК № 3459 від 09.04.2009 р.