О.В. ТЕРЕЩЕНКО

ПРИКЛАДНАЯ СТАТИСТИКА ДЛЯ СОЦИАЛЬНЫХ НАУК

Компьютерный практикум для студентов
гуманитарных специальностей

Минск, 2002

УДК [31:3](075.8)
ББК 60.6я73
Т35

Рецензент: к.с.н., доцент Е.А. Кечина

Рекомендовано Ученым советом
факультета философии и социальных наук
15 марта 2001 г., протокол № 5

Терещенко О.В.
Прикладная статистика социальных наук. Компьютерный практикум для студентов гуманит. спец. / О.В. Терещенко. – Мн: БГУ, 2002. – 93 с.
ISBN 985-445-628-5.

Пособие содержит программу учебного курса и методические материалы по компьютерному практикуму, в рамках которого осуществляется самостоятельная работа студентов, направленная на развитие и закрепление навыков компьютерного статистического анализа данных. Предназначено для студентов гуманитарных специальностей, изучающих статистику.

УДК [31:3](075.8)
ББК 60.6я73

ISBN 985-445-628-5

ВВЕДЕНИЕ

Настоящее учебно-методическое пособие предназначено для активизации самостоятельной работы студентов по изучению прикладной статистики и формированию навыков ее практического использования. Пособие включает план и программу учебного курса "Основы прикладной статистики" для студентов гуманитарных специальностей, описание компьютерного практикума, сопровождающего теоретический курс, краткий англо-русский словарь статистических и компьютерных терминов.

Практикум по компьютерному статистическому анализу является перспективной формой самостоятельной контролируемой работы студентов. Он сокращает характерный для гуманитарных специальностей разрыв между преподаванием прикладной статистики и информатики. Задания практикума органично встроены в программу курса, их выполнение приурочено к теоретическому рассмотрению соответствующих тем.

Каждое из 12 заданий включает требования к выполнению, формы отчетности, методические материалы, которые носят справочный характер и могут использоваться студентами при выполнении заданий практикума.

Методические материалы практикума рассчитаны на использование наиболее распространенного в мире программного средства для статистического анализа данных социальных исследований SPSS, версия 10.0.

УЧЕБНЫЙ ПЛАН КУРСА

№	Наименование разделов и тем	Количество часов				№ за-дания
		Лекц.	Практ.	Сам.	Всего	
	РАЗДЕЛ 1. СТАТИСТИЧЕСКИЕ ДАННЫЕ					
1.	Природа статистики. Особенности статистического подхода в социальном исследовании	2			2	
2.	Основные этапы эмпирического социального исследования.	4	2	2	8	1
3.	Измерение в количественном социальном исследовании.	4	6	2	12	2
4.	Данные социального исследования.	2	2		4	
5.	Подготовка данных к компьютерному статистическому анализу.		4	2	6	3
	РАЗДЕЛ 2. ОПИСАТЕЛЬНАЯ СТАТИСТИКА					
6.	Одномерные частотные распределения.	4	6	2	12	4
7.	Графическое представление одномерных распределений	4	4	2	10	5
8.	Меры центральной тенденции.	2	2		4	6
9.	Меры разброса данных.	2	2		4	
10.	Анализ формы распределения. Стандартизация переменных.	2	2	2	6	
	РАЗДЕЛ 3. ОСНОВЫ СТАТИСТИЧЕСКОГО ВЫВОДА					
11.	Теория вероятностей как методологическая основа статистического вывода.	2			2	
12.	Основные теоретические распределения.	2	4		6	
13.	Оценивание параметров генеральной совокупности. Репрезентативность выборки.	2	2	2	6	7
14.	Статистическая проверка гипотез.	4	4	2	10	8
	РАЗДЕЛ 4. АНАЛИЗ СТАТИСТИЧЕСКИХ СВЯЗЕЙ					
15.	Понятия статистической связи и независимости в социальных науках.	2			2	
16.	Исследование связей по таблице сопряженности.	4	4	2	10	9
17.	Меры связи, основанные на рангах.	2	2		4	
18.	Линейная статистическая модель парной связи.	4	4	2	10	10
19.	Нелинейные модели парной связи.	2	2	2	6	11
20.	Основы планирования эксперимента.	2			2	12
21.	Дисперсионный анализ.	4	4	2	10	

ПРОГРАММА УЧЕБНОГО КУРСА

Р а з д е л 1. СТАТИСТИЧЕСКИЕ ДАННЫЕ

Тема 1. Природа статистики. Особенности статистического подхода в социальных исследованиях

Статистика как наука, имеющая дело со сбором, обработкой, анализом и интерпретацией данных о массовых явлениях и процессах.

Статистические совокупности: генеральные, выборочные. Элементы статистических совокупностей (случаи). Виды генеральных совокупностей: конечные/бесконечные, конкретные/гипотетические, однородные/неоднородные.

Дизайн социального исследования: сплошной и выборочный подходы. Виды исследований генеральной совокупности: количественные/качественные. Количественные (статистические) методы исследования: сплошное обследование, выборочное обследование, факторный эксперимент. Качественные методы исследования: полевое исследование, фокус-группа, исследование случаев (case study).

Временной подход в социальных исследованиях: мониторинговое, лонгитюдное, когортное исследования.

Функции статистики: описание, обобщение, объяснение (прогнозирование).

Особенности статистического подхода в социальных науках. Абстрагирование от индивидуальности. Оценочный характер полученных результатов. Корректное использование специфических статистических методов сбора, обобщения и анализа данных.

Тема 2. Основные этапы эмпирического социального исследования

Основные этапы социального исследования. Выдвижение гипотез; роль теории в прикладном исследовании. Операционализация гипотез: выбор дизайна, обоснование выборки исследования; определение измеряемых показателей, разработка инструментария исследования. Полевой этап исследования: сбор данных. Обработка и анализ данных: ввод данных в компьютер, обработка и подготовка данных к

статистическому анализу; анализ данных; проверка гипотез исследования. Интерпретация полученных результатов, построение моделей исследуемых явлений и процессов.

Место статистики в социальном эмпирическом исследовании.

Практическое занятие: выдвижение и операционализация гипотез.

Практикум
Задание 1: операционализация гипотез, выбор дизайна исследования

Тема 3. Измерение в количественном социальном исследовании

Понятие переменной. Измерение как процедура присвоения символа наблюдаемому объекту. Цель измерения. Шкала измерения. Типы измерительных шкал: количественные / качественные.

Виды шкал. Шкала наименований (номинальная шкала); дихотомическая шкала как частный случай шкалы наименований. Шкала порядка; оценочная и ранговая шкалы как частные случаи шкалы порядка. Количественные шкалы интервалов и отношений. Свойства шкал; допустимые преобразования на шкалах разных видов.

Представления различных видов шкал в инструментарии исследования. Закрытые и открытые вопросы. Кодирование открытых вопросов. Требования к инструментарию.

Вторичные измерения: переменные-индикаторы и переменные-индексы. Основные методы построения индексов. Относительные и абсолютные показатели. Применение логарифмических шкал.

Особенности измерения в социальных исследованиях: принципиально качественный характер данных; реактивность. Нереактивные методики измерения.

Практическое занятие: измерительные шкалы.

Практикум
Задание 2: разработка инструментария исследования; проведение опроса

Тема 4. Данные социального исследования

Понятие информации. Статистический характер информации в количественном социальном исследовании. Данные социального исследования как формализованная и структурированная информация об объекте исследования.

Этапы формализации информации: определение генеральной совокупности (объекта исследования); построение выборки (выбор единиц наблюдения); операционализация понятий (выбор измеряемых показателей и определение способов их измерения); измерение показателей на единицах наблюдения. Соотношение этапов формализация информации с этапами социального исследования.

Критерии структурирования данных: объект / переменная / время измерения.

Структурирование данных в одномоментном исследовании: матрица данных "объект-признак" ("случай–переменная").

Понятие временно́го ряда. Виды временных рядов. Требования к временным рядам. Структурирование данных в лонгитюдном исследовании: куб данных "объект-признак-время".

Подготовки данных к вводу в компьютер. Проверка комплектности и полноты и правильности заполнения инструментария. Проверка принадлежности объектов к выборке исследования. Кодирование открытых вопросов.

Проблема пропущенных значений. Легальные и нелегальные пропущенные значения. Кодирование пропущенных значений.

Практическое занятие: Матрица данных.

Тема 5. Подготовка данных к компьютерному статистическому анализу.

Создание матрицы данных с помощью SPSS для WINDOWS. Описание переменных: имя, метка, формат, метки значений, пропущенные значения (системные, пользовательские). Ввод данных в матрицу.

Подготовка данных к статистическому анализу. Создание вторичных переменных: перекодировка, вычисление новых переменных.

Отбор случаев с заданными характеристиками (использование фильтров). Сортировка. Объединение данных из нескольких файлов.

Сохранение файла данных. Чтение файла данных.

Практическое занятие (в компьютерном классе): создание матрицы данных, ввод данных в компьютер, преобразование и подготовка данных к статистическому анализу.

Практикум
Задание 3: ввод и подготовка данных к статистическому анализу

ЛИТЕРАТУРА К РАЗДЕЛУ 1

1. *Афанасьев В.И.* Методические указания по курсу математической статистики с применением пакета SPSS. М., 1996.
2. *Батыгин Г.С.* Лекции по методологии социологических исследований. Учебник для Вузов. М., 1995.
3. *Бутенко И.А.* Анкетный опрос как искусство общения социолога с респондентом. Учеб. пособие. М., 1989.
4. *Кимбл Г.* Как правильно пользоваться статистикой. М., 1982.
5. *Паниотто В.И., Максименко В.С.* Количественные методы в социологических исследованиях. Киев, 1982.
6. *Паниотто В.И., Максименко В.С.* Зачем социологу математика? Киев, 1988.
7. *Пациорковский В.В., Петрова А.И., Пациорковская В.В.* Использование SPSS в социологии. Ч. 1. Ввод и контроль данных. М., 1998.
8. *Сатаров Г.А.* Математика в социологии: стереотипы, предрассудки, заблуждения // Социологические исследования. 1986. №3.
9. *Терещенко О.В.* Социолог и ЭВМ. Мн., 1990.
10. *Терещенко О.В.* Статистическая обработка и анализ социологической информации // Социология. / Под ред. А.Н. Елсукова. Мн., 2000.
11. *Терещенко О.*В. Первые шаги в SPSS для Windows. Мн., 2001.
12. *Толстова Ю.Н.* Измерение в социологии. М., 1998.
13. SPSS Base 7.5 для Windows. Руководство по применению. М., 1997.

Р а з д е л 2. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Тема 6. Одномерные частотные распределения

Дескриптивная статистика как средство описания выборочной совокупности.

Вариационный ряд. Абсолютная и относительная частота. Одномерное частотное распределение дискретной переменной (номинальной, порядковой, количественной).

Одномерное частотное распределение непрерывной переменной. Понятие группировки. Виды группировок: типологическая, аналитическая, процентильная. Проблема точных границ интервалов; подходы к ее решению. Открытые и закрытые интервалы.

Типологическая группировка: применение, особенности, правила построения.

Аналитическая группировка: применение, особенности, правила построения.

Понятие квантиля (процентиля). Дециль, квинтиль, квартиль, медиана. Процентильная группировка: применение, особенности, правила построения. Стандартные процентильные группировки.

Накопленная (кумулятивная) частота: восходящая, нисходящая. Распределение накопленных частот для дискретных и непрерывных переменных. Использование накопленных частот в построении процентильных группировок.

Практическое занятие: распределения абсолютных, относительных и накопленных частот; построение группировок.

Практикум
Задание 4: построение одномерных частотных распределений и группировок

Тема 7. Графическое представление одномерных распределений

Графики как способ визуализации одномерных распределений.

Основные виды графиков: диаграммы, гистограммы, полигоны распределения, кумуляты (графики накопленных частот), статистические карты, пиктограммы, графики временных рядов.

Общие требования к графикам. Использование площади фигур в качестве наиболее общего способа представления частот. Шкалы переменных и частот: масштаб, прерывание шкал.

Диаграммы для дискретных переменных (построение, особенности использования): круговая диаграмма; диаграммы полос и столбцов; ленточная диаграмма; пиктограмма.

Гистограмма и полигон распределения для количественных переменных. Плотность распределения. Правила построения гистограмм и полигонов. График интерквартильного диапазона: построение, применение. Статистические карты.

Графики накопленных частот (кумуляты). Правила построения кумулят для дискретных и непрерывных переменных.

Графическое представление временных рядов.

Практическое занятие: построение графиков.

Практикум
Задание 5: построение и редактирование графиков

Тема 8. Меры центральной тенденции

Характеристики положения распределения количественной переменной на действительной оси: минимальное и максимальное значения, квантили распределения.

Понятие центра распределения как разновидности нормы, вокруг которой колеблются значения всех наблюдений. "Среднее" как "типичное".

Мода (вероятностное среднее). Определение моды для дискретных и непрерывных распределений, для сгруппированных данных. Свойства моды. Применение моды в социальной практике.

Медиана (ранговое среднее). Вычисление медианы для сгруппированных данных. Определение медианы по распределению накопленных частот.

Среднее арифметическое. Вычисление среднего арифметического для сгруппированных данных и оценочных (квазиколичественных) шкал. Среднее арифметическое для дихотомических шкал.

Взвешенное среднее арифметическое. Определение весов. Перевзвешивание выборки.

Другие виды средних, используемые в статистическом анализе: среднее квадратическое, среднее геометрическое, среднее гармоническое.

Практическое занятие: вычисление и интерпретация квантилей, мер центральной тенденции.

Тема 9. Меры разброса данных

Вариационный размах как наиболее простой показатель разброса данных. Недостатки вариационного размаха в качестве меры разброса данных.

Среднее абсолютное отклонение. Его вычисление и использование.

Среднее квадратическое отклонение и дисперсия. Их преимущества перед другими мерами разброса данных. Вычисление среднего квадратического отклонения и дисперсии для сгруппированных данных и оценочных шкал. Среднее квадратическое отклонение и дисперсия для дихотомических шкал Среднее квадратическое отклонение от моды и медианы.

Коэффициент вариации; его использование в сравнительном анализе.

Моменты эмпирического распределения как его наиболее общие характеристики. Начальные и центральные моменты. Среднее арифметическое и дисперсия в качестве моментов распределения.

Практическое занятие: вычисление и интерпретация показателей разброса данных, моментов эмпирического распределения.

Тема 10. Анализ формы распределения. Стандартизация переменных

Форма эмпирического распределения, ее основные характеристики: модальность, протяженность, симметричность.

Виды распределений в зависимости от количества и расположения мод. Одномодальные распределения: "колокол", J-образное распределение. Бимодальные распределения: "двойной колокол", U-образное распределение. Полимодальное распределение. Природа би-

модальности и полимодальности: неоднородность генеральной совокупности. Анализ бимодальных и полимодальных распределений.

Симметричность одномодального распределения. Правая и левая асимметрия: коэффициент асимметрии.

Форма распределения " симметричный колокол", коэффициент эксцесса.

Анализ формы распределения: использование мер центральной тенденции, показателей разброса данных, графиков.

Стандартизация количественных переменных: z-оценка как безразмерная стандартизированная переменная. Распределение z-оценок, его свойства.

Практическое занятие: анализ формы распределения; вычисление и проверка свойств z-оценок.

Практикум
Задание 6: вычисление характеристик одномерного распределения

ЛИТЕРАТУРА К РАЗДЕЛУ 2

1. *Гласс Дж., Стенли Дж*. Статистические методы в педагогике и психологии. М., 1976.
2. *Кимбл Г*. Как правильно пользоваться статистикой. М., 1982.
3. *Мюллер Д., Шусслер К*. Статистические методы в социологии. Ч. 1. М., 1968.
4. *Ноэль Э*. Массовые опросы: введение в методику демоскопии. М., 1978.
5. *Паниотто В.И., Максименко В.С*. Количественные методы в социологических исследованиях. Киев, 1982.
6. Статистические методы анализа информации в социологических исследованиях. М., 1979.
7. SPSS Base 7.5 для Windows. Руководство по применению. М., 1997.

Р а з д е л 3. ОСНОВЫ СТАТИСТИЧЕСКОГО ВЫВОДА

Тема 11. Теория вероятностей как методологическая основа статистического вывода

Статистический вывод – область статистики, позволяющая делать выводы о неизвестных характеристиках и свойствах генеральной совокупности на основании результатов выборочного исследования. Задачи статистического вывода: оценивание неизвестных параметров

генеральной совокупности, статистическая проверка гипотез. Случайность отбора как предпосылка статистического вывода. Простая случайная выборка.

Случайная величина, ее значения: дискретные и непрерывные случайные величины. Выборочное пространство. Случайное событие. Вероятность случайного события. Свойство аддитивности вероятности. Теорема Байеса. Байесов подход в статистике.

Распределение случайной величины. Закон распределения дискретной случайной величины. Функция плотности распределения непрерывной случайной величины. Функция распределения дискретной и непрерывной случайной величины. Свойства закона распределения, функции плотности распределения, функции распределения.

Параметры генеральной совокупности и выборочные статистики. Ошибка выборки. Случайная и систематическая составляющие ошибки выборки, их источники.

Тема 12. Основные теоретические распределения

Понятие теоретического распределения. Теоретические распределения, наиболее часто используемые в анализе данных социальных исследований: Гаусса (нормальное), Стьюдента, Фишера, Хи-квадрат.

Нормальное распределение Гаусса $N(\mu,\sigma)$, его параметры, свойства. Стандартное нормальное распределение $Z(0,1)$. Таблица стандартного нормального распределения. Использование таблицы стандартного нормального распределения для работы с произвольными нормальными распределениями.

Распределение Стьюдента $t(df)$, его параметр, свойства. Таблица распределения Стьюдента.

Распределение Хи-квадрат $\chi^2(df)$, его параметр, свойства. Таблица распределения Хи-квадрат.

Распределение Фишера $F(df_1, df_2)$, его параметры, свойства. Таблицы распределения Фишера.

Практическое занятие: исчисление вероятностей для случайных величин, подчиняющихся важнейшим теоретическим распределениям.

Тема 13. Оценивание параметров генеральной совокупности. Репрезентативность выборки

Выборочное распределение статистики. Следствие из Центральной предельной теоремы.

Точечные оценки параметров генеральной совокупности. Свойства точечных оценок: несмещенность, эффективность, состоятельность. Точечное оценивание методом моментов.

Интервальное оценивание параметров генеральной совокупности. Стандартная ошибка выборки. Доверительная вероятность, ее стандартные значения: 0.99, 0.95, 0.9. Интервал, соответствующий доверительной вероятности. Построение доверительных интервалов для математического ожидания, доли, дисперсии.

Доверительный интервал для случайной ошибки выборки. Показатели репрезентативности выборки: предельно допустимая ошибка и доверительная вероятность. Расчет минимального объема репрезентативной выборки. Поправка на конечный объем генеральной совокупности.

Конкретный и субъективный характер репрезентативности. Влияние дизайна выборочного исследования на вычисление доверительных интервалов и оценивание случайных ошибок выборки.

Представление данных о репрезентативности выборки в научных работах и прессе.

Практическое занятие: интервальное оценивание параметров генеральной совокупности и случайных ошибок выборки; вычисление объема простой случайной репрезентативной выборки.

Практикум
Задание 7: оценка параметров генеральной совокупности и ошибок выборки

Тема 14. Статистическая проверка гипотез

Понятие и структура статистической гипотезы. Понятие и назначение нулевой гипотезы H_0. Понятие и назначение альтернативных гипотез H_1. Односторонние и двусторонние альтернативные гипотезы.

Верные и ошибочные решения при проверке статистической гипотез. Ошибка первого рода, ее вероятность (уровень значимости). Ошибка второго рода, ее вероятность (мощность критерия). Критерий нулевой гипотезы. Статистика критерия, ее распределение, свойства.

Процедура проверки статистической гипотезы. Критическое значение и критическая область. Односторонние и двусторонние критические области. Использование статистических таблиц при проверке гипотез. Проверка статистических гипотез с помощью компьютера: p -значение.

Связь математического аппарата проверки статистических гипотез с процедурами построения доверительных интервалов.

Проверка гипотез о равенстве математических ожиданий (t- критерий); вероятностей положительного ответа (Z-критерий); дисперсий (F-критерий, критерий хи-квадрат); распределений вероятностей (критерий хи-квадрат).

Надежность результатов статистического вывода, границы доверия к ним.

Практическое занятие: проверка статистических гипотез.

Практикум
Задание 8: проверка статистических гипотез

ЛИТЕРАТУРА К РАЗДЕЛУ 3

1. *Афифи А., Эйзен С.* Статистический анализ: подход с использованием ЭВМ. М., 1982.
2. *Гласс Дж., Стенли Дж.* Статистические методы в педагогике и психологии. М., 1976.
3. *Закс Л.* Статистическое оценивание. М., 1976.
4. *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи. М., 1973.
5. *Кимбл Г.* Как правильно пользоваться статистикой. М., 1982.
6. *Кокрен У.* Методы выборочного исследования. М., 1976.
7. *Мюллер Д.,Шусслер К.* Статистические методы в социологии. Ч. 2. М.,1 968
8. *Паниотто В.И.* Качество социологической информации. Киев, 1986.
9. *Паниотто В.И., Максименко В.С.* Количественные методы в социологических исследованиях. Киев,1982.
10. Статистические методы анализа информации в социологических исследованиях. М., 1979.

Р а з д е л 4. АНАЛИЗ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

Тема 15. Понятия статистической связи и независимости в социальных науках

Понятия статистической связи и статистической независимости. Природа статистической связи. Парная и множественная связь между переменными.

Ненаправленная связь (для номинальных переменных). Прямая и обратная связь (для порядковых и количественных переменных).

Понятие меры связи; требования к мерам связи.

Корреляционная и причинная связь. Определение причинной связи. Формальные критерии причинности (каузальности). Зависимые и независимые переменные. Анализ причинных связей в выборочных и экспериментальных исследованиях.

Тема 16. Исследование связей по таблице сопряженности

Таблица сопряженности как средство представления совместного распределения двух переменных. Элементы таблицы сопряженности, правила ее заполнения. Маргинальные частоты.

Определение статистической связи и независимости по таблице сопряженности. Частотная модель связи: критерий статистической независимости строк и столбцов таблицы Хи-квадрат. Эмпирические и теоретические частоты. Проверка гипотезы о наличии связи между строками и столбцами таблицы сопряженности. Измерение силы (тесноты) связи. Меры связи, основанные на критерии хи-квадрат: коэффициенты контингенции (φ), Чупрова (T), Крамера (C). Преимущества и недостатки коэффициентов; их интерпретация.

Таблицы сопряженности размерности 2*2. Понятия абсолютной (двусторонней) и полной (односторонней), прямой и обратной связи для таблиц 2*2. Меры связи для таблиц 2*2: коэффициенты Юла Q и контингенции Φ. Использование коэффициентов Q и Φ при анализе и интерпретации связи двух дихотомических переменных.

Теоретико-информационный подход к исследованию причинной связи. Общая и дополнительная информация о распределении признака. Энтропия. Уменьшение энтропии, связанное с получением допол-

нительной информации. Теоретико-информационная модель причинной связи. Теоретико-информационные меры связи: λ_a, λ_b, λ-симметричное Гудмена; τ_a, τ_b, τ-симметричное Гудмена и Краскалла. Проверка гипотезы о статистической значимости теоретико-информационных мер связи. Интерпретация значений теоретико-информационных мер связи.

Практическое занятие: построение таблиц сопряженности, вычисление мер связи для таблиц сопряженности, проверка гипотез о наличии связи между строками и столбцами таблицы.

Практикум
Задание 9: построение и анализ таблиц сопряженности

Тема 17. Меры связи, основанные на рангах

Понятие ранговой корреляции, особенности ее применения в прикладных задачах. Прямая и обратная ранговая связь. Коэффициенты ранговой корреляции Спирмена (r_s) и Кендалла (τ), интерпретация их значений.

Понятие связных рангов. Коэффициенты Спирмена и Кендалла для связных рангов. Коэффициенты Спирмена и Кендалла для упорядоченных таблиц сопряженности. Прямая и обратная связь в упорядоченных таблицах. Интерпретация коэффициентов ранговой корреляции для таблиц сопряженности.

Множественный коэффициент ранговой корреляции, его вычисление и интерпретация.

Проверка гипотез о статистической значимости коэффициентов ранговой корреляции.

Практическое занятие: вычисление коэффициентов ранговой корреляции, проверка гипотез об их статистической значимости.

Тема 18. Линейная статистическая модель парной связи

Линейная статистическая модель парной связи $y = bx + b_0$.

Диаграмма рассеяния как способ графического представления совместного распределения двух количественных переменных. Анализ

линейности, направления и тесноты связи двух переменных по диаграмме рассеяния.

Ковариация как мера совместного рассеяния двух количественных признаков, ее свойства. Коэффициент линейной корреляции Пирсона, его свойства. Связь коэффициента линейной корреляции с коэффициентом Ф для дихотомических переменных и ранговым коэффициентом корреляции Спирмана. Проверка гипотезы о статистической значимости коэффициента линейной корреляции. Симметричность коэффициента линейной корреляции.

Регрессионный анализ как метод исследования причинной линейной связи. Зависимая и независимая переменные в уравнении линейной регрессии. Построение уравнения парной линейной регрессии методом наименьших квадратов. Вычисление и интерпретация параметров уравнения регрессии. Проверка гипотез о статистической значимости параметров уравнения линейной регрессии.

Анализ остатков регрессии. Модель с разделением дисперсии: дисперсия общая, объясненная, остаточная. Коэффициент детерминации, его интерпретация в качестве доли объясненной дисперсии.

Практическое занятие: вычисление ковариации, коэффициента линейной корреляции, коэффициента детерминации; построение уравнения парной линейной регрессии; проверка гипотез о статистической значимости параметров уравнения регрессии и коэффициента корреляции.

Практикум
Задание 10: анализ линейной статистической связи

Тема 19. Нелинейные модели парной связи

Нелинейные модели парной связи, их вид на диаграмме рассеяния.

Понятие кусочно-линейной математической функции. Аппроксимация нелинейной функции кусочно-линейной. Разделение дисперсии зависимой переменной: общая, внутригрупповая и межгрупповая (объясненная) дисперсия. Корреляционное отношение как показатель тесноты нелинейной связи между зависимой и независимой переменными. Преимущества и недостатки корреляционного отношения.

Нелинейные модели с использованием логарифмических, экспоненциальных, степенных и тригонометрических функций.

Практическое занятие: вычисление корреляционного отношения; построение и интерпретация уравнения регрессия с использованием логарифмических и экспоненциальных функций.

Практикум
Задание 11: вычисление корреляционного отношения, построение и анализ уравнения парной нелинейной регрессии

Тема 20. Основы планирования эксперимента

Метод эксперимента в социальных исследованиях. Зависимая и независимые переменные (факторы); требования к ним. Прямые эффекты факторов и эффекты взаимодействия.

Классический трехстадийный эксперимент: формирование групп; проведение эксперимента; оценивание результатов. Характеристики эксперимента: число факторов (однофакторный, двухфакторный и т.п.); характер факторов (выделяемые, контролируемые); число групп; схема (внутригрупповая, межгрупповая). Виды экспериментов: лабораторный, естественный (квазиэксперимент), полевой.

Представление результатов экспериментального исследования: план эксперимента.

Тема 21. Дисперсионный анализ

Модель дисперсионного анализа: общая, объясняемая (межгрупповая) и остаточная (внутригрупповая) дисперсия. Разложение общей суммы квадратов на меж- и внутригрупповую суммы квадратов. Числа степеней свободы для сумм квадратов.

Однофакторный дисперсионный анализ; его гипотеза. Проверка нулевой гипотезы об отсутствии различий между средними значениями зависимой переменной в группах, образованных различными градациями переменной-фактора: t-критерий; F-критерий.

Двухфакторный дисперсионный анализ. Главные эффекты факторов и взаимодействие факторов как источники межгрупповой дисперсии. Гипотезы двухфакторного дисперсионного анализа. Проверка гипотез: F-критерий.

Многофакторный дисперсионный анализ; его гипотезы. Проверка гипотез многофакторного дисперсионного анализа.

Применение методов множественного сравнения в дисперсионном анализе. Т-метод Тьюки. S-метод Шеффе.

Практическое занятие: однофакторный и двухфакторный дисперсионный анализ.

Практикум
Задание 12: обработка данных научного эксперимента: дисперсионный анализ

ЛИТЕРАТУРА К РАЗДЕЛУ 4

1. *Адлер Ю.П., Грановский Ю.В., Маркова Е.В*. Теория эксперимента: прошлое, настоящее, будущее. М., 1982.
2. *Аптон Г*. Анализ таблиц сопряженности. М., 1982.
3. *Вихалемм П*. Эксперимент в социологическом исследовании // Методы сбора информации в социологическом исследовании. Кн.2. М., 1990.
4. *Гласс Дж., Стенли Дж*. Статистические методы в педагогике и психологии. М., 1976.
5. *Кендалл М.Дж., Стьюарт А*. Статистические выводы и связи. М., 1973.
6. *Кимбл Г*. Как правильно пользоваться статистикой. М., 1982.
7. *Кэмбелл Д*. Модели экспериментов в психологии и прикладных исследованиях. М., 1980.
8. *Мюллер Д., Шусслер К*. Статистические методы в социологии. Ч. 2, 3. М., 1968
9. *Паниотто В.И., Максименко В.С*. Количественные методы в социологических исследованиях. Киев,1982.
10. Справочник по прикладной статистике. В 2 т. / Под ред. Э. Ллойда, У. Лидермана, Ю.Н. Тюрина. М., 1989, 1990.
11. Статистические методы анализа информации в социологических исследованиях. М., 1979.
12. *Финни Д.* Введение в теорию планирования экспериментов. М., 1970.

КОМПЬЮТЕРНЫЙ ПРАКТИКУМ

ЗАДАНИЕ 1. ОПЕРАЦИОНАЛИЗАЦИЯ ГИПОТЕЗ, ВЫБОР ДИЗАЙНА ИССЛЕДОВАНИЯ

1. Разработайте и операционализируйте гипотезу для самостоятельного исследования.
2. Определите генеральную совокупность и дизайн исследования.

Необходимая подготовка:
- Темы 1, 2 программы курса.

Форма отчетности: письменная работа (2–3 стр.)

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 1

Дизайн исследования – общая стратегия проведения исследования, формирования выборки и обоснования ее репрезентативности.

Выбор дизайна исследования определяется
1) *стратегическими факторами:*
- исследовательской парадигмой,
- целями исследования;

2) *тактическими факторами:*
- характеристиками генеральной совокупности (ГС),
- имеющимися ресурсами (материальными, человеческими, временными).

Основные *парадигмы эмпирического исследования*:
- количественная (в целом, соответствует позитивистскому и структурно-функциональному подходу);
- качественная (в целом, соответствует интерпретативному подходу).

Количественные исследования претендуют на получение "объективных" данных о социальных явлениях и процессах.

Основные *цели* количественных исследований:

- описание ГС, проверка гипотез о параметрах ГС (выборочное исследование);
- проверка гипотез о причинно-следственных связях между социальными явлениями (эксперимент, выборочное исследование);
- исследование структуры группы и внутригрупповых социальных ролей.

Основные *методы количественных исследований*:

- полное или выборочное обследование ГС;
- факторный контролируемый эксперимент.

Основные *методы сбора данных*:

- анкетирование;
- стандартизированное интервью;
- контент-анализ документов и текстов.

Качественные исследования претендуют на понимание социальных явлений и процессов через призму человеческих взаимоотношений и судеб.

Основные *цели* качественных исследований:

- исследование образа жизни, социальной идентичности, "картины мира", языка и т.п. отдельных социальных групп;
- исследование процессов социализации и формирования идентичности, "картины мира" и языка человека.

Основные *методы качественных исследований*:

- полевое исследование (этнометодология);
- фокус-группа;
- исследование отдельных случаев (case-study).

Основные *методы сбора данных в качественных исследованиях*:

- наблюдение;
- нестандартизированное и полустандартизированное интервью;
- "качественный" анализ документов и текстов.

Цели исследования в зависимости от протяженности во времени:

1) *статические* или единовременные (cross-sectional) – "срез" социальной реальности в определенный момент времени;

2) *динамические* – исследование социальных процессов во времени:

- ретроспективные исследования;

- мониторинговые (cross-sectional) исследования;
- лонгитюдные исследования.

Основные виды количественных исследований

Дескриптивное обследование генеральной совокупности (survey).
Цель: получение "объективной" информации о ГС.
Примеры: национальные переписи населения, официальные отчеты, данные текущего статистического учета.
Методы анализа данных: описательная статистика (классификация объектов, получение частотных распределений, определения центральных тенденций), точное измерение связей между анализируемыми показателями).
Преимущества: качество и надежность собранной информации, абсолютно надежная проверка гипотез исследования.
Недостатки: высокая стоимость и трудоемкость.

Дескриптивное выборочное обследование (sample survey) имеет дело не со всей генеральной совокупностью, но только с некоторой ее частью, называемой выборкой.
Цель: получение "объективной" информации о ГС.
Примеры: опросы общественного мнения, электоральные исследования и т.п.
Методы анализа данных: описательная выборочная статистика (классификация объектов, получение частотных распределений, определения центральных тенденций), статистический вывод (оценивание ошибок выборки, параметров ГС, проверка гипотез относительно параметров ГС и связей между измеряемыми показателями).
Преимущества: возможность обследовать ГС, располагая ограниченными ресурсами.
Недостатки: наличие случайных и систематических ошибок выборки, необходимость применять специальные процедуры обобщения результатов исследования на ГС.

Контролируемый факторный эксперимент – сравнительный анализ специально сформированных и подвергнутых различным воздействиям однородных групп.

Цель: строгое исследование причинных связей между непрерывной зависимой переменной и набором влияющих на нее факторов.
Примеры: психология, педагогика, медицина, сравнительные исследования.
Методы: планирование эксперимента, статистический анализ экспериментальных данных (дисперсионный анализ).
Преимущества: экспериментальный дизайн позволяет устранить влияние на зависимую переменную третьих (не изучаемых в эксперименте) факторов, и представить исследуемую связь в "очищенном" виде.
Недостатки: ограниченность результатов, эффекты смешения воздействий.

Выборочный метод

Репрезентативность выборки – способность представлять изучаемые явления, процессы, закономерности адекватно тому, как они протекают в генеральной совокупности; соответствие результатов выборочного исследования положению дел в генеральной совокупности.
Общие проблемы выборочного метода:
- определение ГС;
- однородность ГС;
- доступность элементов ГС (готовность участвовать в исследовании).

Методы обоснования репрезентативности:
- статистический (вероятностные или случайные методы отбора);
- внестатистический (целевые методы отбора).

В количественных исследованиях основным методом обобщения результатов выборочного исследования на ГС является специальный раздел статистики – ***статистический вывод***, который включает:
- оценивание ошибок выборки;
- оценивание параметров ГС;
- проверку гипотез о параметрах ГС.

Статистический вывод основывается на теоретических выводах математической статистики, которая, в свою очередь, базируется на приложениях теории вероятностей, предполагающей случайный отбор изучаемых объектов из Г.С.

Случайный отбор предполагает, что все элементы из генеральной совокупности имеют одинаковую вероятность попасть в выборку.

Виды случайного отбора:

- простой;
- стратифицированный (районированный);
- кластерный (гнездовой).

Отбор является *простым случайным*, если он производится из полного списка элементов однородной ГС с применением специальных процедур (лотереи, таблиц и компьютерных датчиков случайных чисел и т.п.). Часто вместо случайных процедур отбора используют *рандомизирующие (квази-случайные) процедуры* – систематический отбор, маршрутную выборку и т.п.

Стратифицированный (районированный) случайный отбор применяется для неоднородных ГС, а также в отсутствие полных списков элементов ГС. ГС разделяется на относительно однородные части, из каждой части производится простой случайный отбор. *Примеры:* национальный опрос со стратификацией ГС по областям; маркетинговое исследование со стратификацией магазинов по размерам.

Кластерный (гнездовой) случайный отбор применяется к ГС, естественным образом разделенным на относительно небольшие группы – кластеры. Составляется полный список кластеров; из которого производится простой или стратифицированный случайный отбор. Выбранные кластеры обследуются полностью. *Примеры*: семьи, академические группы, отдельные населенные пункты.

Если кластеры слишком велики для сплошного обследования, их можно обследовать выборочно. В этом случае отбор является *многоступенчатым*.

Целевой (неслучайный) отбор применяется, когда случайный отбор невозможен по организационным или финансовым причинам.

Виды целевого (неслучайного) отбора:

- квотный отбор;
- метод доступной выборки;
- метод основного массива;
- метод "снежного кома".

Квотный отбор предполагает, что известно распределение ГС по основным социально-демографическим показателям. Численные кво-

ты распределяются пропорционально численности социальных групп в ГС; интервьюер может произвольно выбирать респондентов в пределах квот. Допускается в исследованиях установок, ценностных ориентаций, предпочтений, мотиваций; не допускается в исследованиях социального неравенства, структуры, мобильности.

Метод доступной выборки применяется, если невозможно получить список ГС. Пример: интерактивный опрос аудитории телеканала. Как правило, не может претендовать на репрезентативность.

Метод основного массива – обследуется большая часть относительно небольшой ГС.

Метод "снежного кома" – список членов небольшой "специфичной" ГС составляется со слов ее отдельных представителей. Процедура продолжается, пока фамилии в списке не начнут повторяться.

Метод факторного контролируемого эксперимента

Факторный эксперимент в социальных науках – метод исследования причинных связей между условиями, в которых находятся люди, и их поведением, реакциями и т.п. Заключается в том, что формируются специально подобранные группы участников, которые затем подвергаются различным воздействиям. Процедуры формирования групп планируются так, чтобы "отсечь" все посторонние факторы и исследовать связь между факторами, включенными в эксперимент, и зависимой переменной "в чистом виде".

Исследуемые реакции – *зависимые переменные*, условия эксперимента и правила формирования групп – *независимые переменные*.

Цель эксперимента – определить, влияет ли фактор, положенный в основу формирования групп или условий проведения эксперимента, на значения зависимой переменной.

Если в эксперименте несколько факторов, необходимо получить ответы на вопросы:

1) имеется ли эффект от воздействия каждого отдельно взятого фактора (главные эффекты факторов);
2) зависит ли величина воздействия фактора от значений других факторов (эффект взаимодействия).

Классический трехстадийный план эксперимента:

1. Формирование одинаковых по составу и объему групп (которые часто называют опытной и контрольной).

2. Проведение эксперимента - помещение групп в разные условия или применение к ним разных методов воздействия.
3. Оценивание результатов эксперимента.

Классификация экспериментов:

- по числу групп (бивалентные / поливалентные);
- по числу факторов (однофакторные, двухфакторные и т.п.);
- по характеру факторов (выделяемые / регулируемые);
- по схеме эксперимента (внутригрупповые / межгрупповые).

ЗАДАНИЕ 2. РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ИССЛЕДОВАНИЯ

1. Составьте анкету для проведения исследования (не менее 6 вопросов, должны быть представлены все уровни измерения).
2. Проведите опрос (не менее 30 человек).

Необходимая подготовка:

- Тема 3 программы курса.
- Задание 1 практикума.

Форма отчетности: анкета; заполненный инструментарий опроса.

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 2

Измерением называется процедура присвоения наблюдаемым объектам определенных символов в соответствии с некоторым правилом. Символы могут быть просто метками, представляющими классы или категории объектов в генеральной совокупности, или числами, характеризующими степень выраженности у объекта измеряемых свойств. Символы-метки могут быть представлены цифрами, но при этом не обязательно несут в себе характерную "числовую" информацию.

Алгоритм (правило) присвоения символа объекту называется *измерительной шкалой*. Как всякая модель, измерительные шкалы должны правильно отражать изучаемые характеристики объекта и, следовательно, иметь те же свойства, что и измеряемые показатели.

Основные виды измерительных шкал

Шкала наименований (номинальная шкала) используется только для обозначения принадлежности объекта к одному из нескольких непересекающихся классов. Приписываемые объектам символы, которые могут быть цифрами, буквами, словами или некоторыми специальными символами, представляют собой метки соответствующих классов. Характерной особенностью номинальной шкалы является принципиальная невозможность упорядочить классы по измеряемому признаку – к ним нельзя применять суждения типа "больше – меньше", "лучше – хуже", и т.п. Примерами номинальных шкал являются пол и национальность, специальность по образованию, марка сигарет, предпочитаемый цвет одежды и т.п.

Частным случаем шкалы наименований является *дихотомическая шкала*, с помощью которой фиксируют наличие / отсутствие у объекта определенного качества или соответствие / несоответствие объекта некоторому требованию. По установившейся традиции, при измерении дихотомических показателей применяют следующие стандартные значения: 0 – если объект не обладает требуемым свойством, 1 – если обладает.

Шкала порядка позволяет не только разбивать объекты на классы, но и упорядочивать классы по возрастанию (убыванию) изучаемого признака. Однако при этом порядковые шкалы не дают ответа на вопрос, *на сколько* или *во сколько раз* это свойство выражено сильнее у объектов из одного класса, чем у объектов из другого класса. Примерами шкал порядка могут служить уровень образования, военные и академические звания, тип поселения (большой город, средний город, малый город, село). Можно сказать, что выпускник университета имеет более высокий образовательный уровень, чем выпускник средней школы, но разница в уровне образования не поддается непосредственному измерению.

Частным случаем шкалы порядка являются *оценочные шкалы*, при использовании которых объект получает (или сам выставляет) оценки, исходя из определенного числа баллов. К ним относятся, например, школьные оценки. Строго говоря, подобные шкалы являются частным случаем шкалы порядка, так как нельзя определить, на

сколько знания "отличника" больше, чем знания "троечника", но в силу некоторых теоретических соображений с ними часто обращаются, как со шкалами более высокого уровня – шкалами интервалов. В частности, вычисляют средний балл по аттестату зрелости.

Другим частным случаем шкалы порядка является *ранговая шкала*, которую обычно применяют, когда признак заведомо не поддается объективному измерению (например, красота или степень неприязни), или когда порядок объектов более важен, чем точная величина различий между ними (например, места, занятые в спортивных соревнованиях). В таких случаях измерение заключается в ранжировании по определенному критерию некого списка объектов, качеств, мотивов, и т.п.

В силу того, что символы, присваиваемые объектам в соответствии с порядковыми и номинальными шкалами, не обладают "числовыми" свойствами, даже если записываются с помощью цифр, эти два типа шкал получили общее название *качественных*, в отличие от количественных шкал интервалов и отношений.

Количественные ***шкалы интервалов и отношений*** имеют общее свойство, отличающее их от качественных шкал: они предполагают не только определенный порядок между объектами или их классами, но и наличие некоторой *единицы измерения*, позволяющей определять, на сколько значение признака у одного объекта больше или меньше, чем у другого.

Основное различие между этими двумя типами шкал состоит в том, что *шкала отношений* имеет абсолютный нуль, не зависящий от произвола наблюдателя и соответствующий полному отсутствию измеряемого признака, а на *шкале интервалов* нуль устанавливается произвольно или в соответствии с некоторыми условными договоренностями.

Примерами шкал интервалов являются календарное время, температурные шкалы Цельсия и Фаренгейта. Шкала оценок с заданным количеством баллов часто рассматривается как квази-интервальная, в предположении, что минимальное и максимальное положения на шкале соответствуют некоторым крайним оценкам или позициям, и интервалы между баллами шкалы имеют одинаковую длину.

К шкалам отношений относится абсолютное большинство измерительных шкал, применяемых в науке, технике и быту: рост и вес, возраст, расстояние, сила тока, время (длительность промежутка между двумя событиями), температура по Кельвину (абсолютный нуль). Шкала отношений является единственной шкалой, на которой определено *отношение отношения*, то есть разрешены арифметические действия умножения и деления и, следовательно, возможен ответ на вопрос, *во сколько раз* одно значение больше или меньше другого.

Количественные шкалы делятся на дискретные и непрерывные. *Дискретные* шкалы измеряются посредством счета: число детей в семье, количество решенных задач, и т.п. Они могут принимать только целые неотрицательные значения *Непрерывные* шкалы предполагают, что измеряемое свойство изменяется непрерывно, и при наличии соответствующих приборов и средств, могло бы быть измерено с любой необходимой степенью точности. Результаты измерения непрерывных показателей довольно часто выражаются целыми числами (например, шкала IQ для измерения интеллекта), но это связано не с природой самих показателей, а с характером измерительных процедур.

Все номинальные и порядковые шкалы, значения которых выражаются целыми числами, считаются дискретными.

Измерительный инструментарий

В социальных науках большинство показателей не поддаются непосредственному измерению с помощью традиционных технических средств. Вместо них применяются всевозможные анкеты, тесты, стандартизированные интервью и т.п., получившие общее название измерительного *инструментария*.

С точки зрения дизайна анкеты, полезно различать открытые и закрытые, альтернативные и неальтернативные вопросы. Большинство вопросов, для которых предполагается статистическая обработка, бывают за*крытыми*. В них респонденту предлагается отметить один или несколько вариантов ответа из предложенного списка. Если из списка можно выбрать только один ответ, вопрос является *альтернативным*, если несколько – *неальтернативным*. *Открытые* вопросы предполагают, что респондент должен сформулировать свой ответ самостоятельно и записать его в отведенное для этого место.

Вопросы с использованием номинальных шкал

Номинальные шкалы обычно предъявляются респондентам в виде альтернативных закрытых вопросов. Для этого составляется полный список возможных ответов на вопрос; ответы нумеруются в произвольном порядке.

Укажите, пожалуйста, Ваш пол:
1. мужской
2. женский

Если вопрос является неальтернативным (можно выбрать несколько вариантов ответа), он не может быть представлен единственной шкалой, а порождает *блок дихотомических шкал*, каждая из которых фиксирует, выбран данный пункт респондентом или не выбран.

Какие качества, на Ваш взгляд, свойственны белорусским предпринимателям?
1. трудолюбие
2. неразборчивость в средствах достижения цели
3. рационализм
4. склонность к жульничеству
5. деловая хватка
6. авантюризм
7. высокий профессионализм
8. высокий уровень общей культуры
9. нежелание честно трудиться
10. непрофессионализм, некомпетентность
11. инициативность, настойчивость
12. жажда наживы
13. честность, порядочность
14. низкий уровень общей культуры
15. способность отстаивать свои интересы

Если вопрос, предполагающий альтернативный дихотомический ответ, задается отдельно (вне блока), ответы нумеруются цифрами 1 и 2.

Нравится ли Вам работать в библиотеке?
1. да
2. нет

Перекодировка дихотомических переменных к классическому виду (с использованием нулей и единиц) производится не при заполнении анкеты респондентом, а позже, при подготовке данных к анализу.

Вопросы с использованием порядковых шкал

Порядковые шкалы представляются в анкете закрытыми альтернативными вопросами, варианты ответов на которые пронумерованы в порядке возрастания или убывания измеряемого свойства.

Ваше образование:

1. начальное
2. базовое среднее
3. среднее общее
4. среднее специальное
5. высшее

Оценочные порядковые шкалы могу быть представлены несколькими способами, в зависимости объекта оценивания и подготовленности аудитории

Как Вы оцениваете сегодняшнее материальное положение Вашей семьи по сравнению с тем, что было 3–4 месяца назад?

1. стало значительно лучше
2. стало немного лучше
3. осталось таким же
4. стало немного хуже
5. стало значительно хуже

Какое значение имеет для Вас национальный состав коллектива, в котором Вы работаете?

не имеет никакого значения	1	2	3	4	5	имеет очень большое значение

Оценочные шкалы чаще, чем другие, предъявляют в блоках:

Оцените, пожалуйста, по 5-балльной шкале некоторые качества передачи Белорусского телевидения "Экономикст":

1. актуальность, злободневность сюжетов ____
2. объективность оценок ____
3. разнообразие тематики ____
4. глубина анализа ____
5. доходчивость, ясность изложения ____
6. привлекательность подачи материала ____

Какие из информационно-аналитических программ Вы смотрите по телевидению и как часто?

	стараюсь не пропускать	смотрю от случая к случаю	фактически не смотрю
Новости (БТ)	1	2	3
Новости (ОРТ)	1	2	3
Новости (РТР)	1	2	3
Сегодня (НТВ	1	2	3

Измерение по *шкале рангов* предполагает, что задан список объектов (ценностей, мотивов, и т.п.), который необходимо упорядочить по определенному критерию. В инструментарии рядом с названиями объектов следует предусмотреть место для того, чтобы респондент мог указать их ранги.

Проранжируйте, пожалуйста, перечисленные ценности в порядке их значимости для Вас:

____ любовь
____ работа
____ семья
____ дружба
____ дети

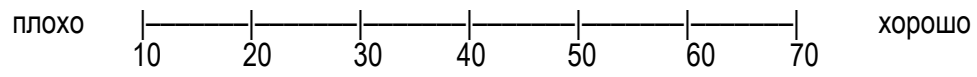
Вопросы с использованием количественных шкал

Количественные переменные, как дискретные, так и непрерывные, обычно представляются открытыми вопросами:

Сколько чашек кофе в день Вы обычно выпиваете? ____ чашек.

Укажите, пожалуйста, свой рост: ____ см.

Количественную шкалу можно создать самостоятельно. Существует два основных способа создания шкал в анкете. В обоих используется прямая линия, на которой респондент отмечает свою позицию. В первом случае градации шкалы размещаются на равном расстоянии друг от друга. Расстояние между ними Вы можете установить в произвольных единицах.



Приведенная выше шкала измеряет, насколько хорошо респондент относится к оцениваемому объекту, по шкале от 10 до 70 баллов. Чтобы получить более точную оценку можно использовать линию, не разделенную на интервалы. В этом случае позицию респондента придется измерять обычной линейкой:

плохо |————————————————————————————————| хорошо

ЗАДАНИЕ 3. ВВОД И ПОДГОТОВКА ДАННЫХ К СТАТИСТИЧЕСКОМУ АНАЛИЗУ

1. Создайте матрицу данных с помощью SPSS. Определите переменные своего исследования: задайте для каждой из них имя, тип, формат, метку переменной, метки значений (где это требуется), коды пропущенных значений, уровень измерения.
2. Введите данные своего исследования.
3. При необходимости сгруппируйте количественные шкалы в интервалы (создайте для этого новые переменные и задайте для них все параметры).

***Необходимая подготовка*:**

- Темы 4,5 программы курса.
- Задания 1, 2 практикума.
- Практическое занятие в компьютерном классе с преподавателем прикладной статистики или информатики.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных в своей папке

Форма отчетности: файл данных с расширением **.sav* (на дискете).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 3

Основные окна и файлы SPSS

Окна	***Файлы***
Окно редактора данных SPSS Data Editor (две закладки: Data View / Variable View)	файл данных (*.SAV)
Окно результатов SPSS Viewer	файл результатов (*.SPO)
Окно команд SPSS Syntax Editor	файл команд (*.SPS)
Окно редактора графиков SPSS Chart Editor	файл графиков (*.SCT)

Этапы подготовки данных к статистическому анализу

1. *Определение переменных:*
   - имя переменной;
   - тип переменной;
   - формат переменной;
   - метка переменной;
   - метки значений переменной;
   - пропущенные значения (системные, пользовательские);
   - уровень измерения.
2. *Ввод данных.*
3. *Проверка качества ввода и чистка данных:*
   - одномерные распределения;
   - таблицы сопряженности;
   - чистка данных.
4. *Преобразования матрицы данных:*
   - добавление переменной;
   - удаление переменной;
   - добавление случая;
   - удаление случая;
   - сортировка данных;
   - отбор случаев по заданным критериям (с помощью фильтров).
5. *Перекодировка, группировка, вычисление переменных:*
   - перекодировка дискретных значений;
   - группировка (построение интервалов);

- перцентильная группировка;
- вычисление переменных: арифметические операции, стандартные функции;
- вычисление переменных в соответствии с заданными условиями;
- вычисление переменных: подсчет одинаковых значений в нескольких переменных;
- стандартизация переменных (z-оценки).

6. *Обмен данными и результатами с другими приложениями Windows:*
   - чтение файлов данных SPSS и других форматов;
   - сохранение файлов данных SPSS и других форматов;
   - копирование результатов в другие приложения Windows.

Краткое руководство по предварительной подготовке данных в SPSS

Действие	Реализация
Определение переменной (*окно редактора данных, закладка* ***Variable View***)	
Имя переменной	*Столбец* ***Name***. Длина имени – до 8 символов. Запрещены слова: **all, and, or, not, with, by, to, eq, ne, lt, le, gt, ge**. Запрещены символы: **пробел**, **!**, **?**, **'**, *****. Имя не может заканчиваться точкой. Каждое имя должно быть уникальным. Имена нечувствительны к регистру.
Тип переменной	*Столбец* ***Type***. Основные типы: **Numeric** (числовой) **String** (строковый)
Формат переменной	*Столбцы:* ***Width*** (общее число позиций) ***Decimals*** (число позиций после десятичной точки) *Рекомендуется начинать установку формата со столбца* ***Decimals***.
Метка переменной	*Столбец* ***Label*** (любой поясняющий текст). Может быть на русском языке.

Действие	**реализация**
Метки значений переменной	*Столбец* ***Values*** (могут задаваться для каждого значения, особенно для номинальных и порядковых переменных). Могут быть на русском языке. Ввести значение переменной в окошко **Value** | ввести его метку в окошко **Value Label** | кнопка **Add** (метка появится в окне) | последовательно ввести все метки, убедиться, что они перечислены в окне | **OK**
Пропущенные значения (пользовательские)	*Столбец* ***Missing***. Выбрать вид пропущенных значений: *дискретные пропущенные значения*: ввести до трех значений; *интервал пропущенных значений*: ввести границы интервала и, при необходимости, шаг.
Пропущенные значения (системные)	Пустая клетка в окне редактора (*закладка* ***Data View***) или перекодировка (см. "Перекодировка", "Группировка").
Уровень измерения	*Столбец* ***Measure*****: Scale** (количественный) **Ordinal** (порядковый) **Nominal** (номинальный)
Ввод данных	
Ввод данных	*Окно редактора данных (закладка* ***Data View****).* Переход между клетками – по клавише **Enter** или по клавишам управления курсором.
Проверка качества ввода и чистка данных	
Построение одномерных распределений	**Analyze** | **descriptive statistics** | **frequencies…** | имена нужных переменных переместить в окно **Variable(s)** | **OK**.
Построение таблиц сопряженности	**Analyze** | **descriptive statistics** | **crosstabs…** | поместить имя переменной, образующей строки таблицы, в окно **Row(s)** | имя переменной, образующей столбцы, – в окно **Column(s)** | **OK**.
Чистка данных	*Окно редактора данных (закладка* ***Data View****)* или перекодировка (см. "Перекодировка", "Группировка").

Действие	реализация
Преобразование матрицы данных	
Добавление переменной	*Окно редактора данных (закладка **Variable View**):* выделить строку с переменной, перед которой должна быть вставка | вызвать контекстное меню (правой кнопкой мыши) | **Insert Variables.** *Окно редактора данных (закладка **Data View**):* выделить столбец с переменной, перед которой должна быть вставка | вызвать контекстное меню (правой кнопкой мыши) | **Insert Variables.**
Удаление переменной	*Окно редактора данных (закладка **Variable View**):* выделить строку с удаляемой переменной | вызвать контекстное меню (правой кнопкой мыши) | **Clear.** *Окно редактора данных (закладка **Data View**):* выделить столбец с удаляемой переменной | вызвать контекстное меню (правой кнопкой мыши) | **Cut** или **Clear.**
Добавление случая	*Окно редактора данных (закладка **Data View**):* выделить строку, перед которой должна быть вставка | вызвать контекстное меню (правой кнопкой мыши) | **Insert Cases.**
Удаление случая	*Окно редактора данных (закладка **Data View**):* выделить удаляемую строку | вызвать контекстное меню (правой кнопкой мыши) | C**ut** или **Clear.**
Сортировка случаев	*Окно редактора данных (закладка **Data View**):* выделить столбец с переменной, по которой производится сортировка | вызвать контекстное меню (правой кнопкой мыши) | **Sort ascending** (по возрастанию) или **Sort descending** (по убыванию). *Окно редактора данных*: **Data** | **Sort Cases…** | перенести в окно **Sort by** имена переменных, по которым осуществляется сортировка | **Sort Order**: выбрать **Ascending** или **Descending** | **OK**.
Отбор случаев по заданным критериям с помощью фильтров.	*Окно редактора данных*: **Data** | **Select Cases…** | в разделе **Select** выбрать **If condition is satisfied** | кнопка **If…** | в открывшемся окне задать критерии отбора, используя логические операторы | **Continue** | **Unselected Cases Are** выбрать **Filtered** | **OK**. Для снятия фильтра в разделе **Select** выбрать **All cases.**

Действие	реализация
Перекодировка, группировка, вычисление переменных	
Перекодировка дискретных значений (в ту же переменную)	*Окно редактора данных*: **Transform** | **Recode** | **Into Same Variable(s)…** | переместить имя переменной в окно **Variables** | кнопка **Old and New Values** | в разделе **Old Value** выбрать **Value** | ввести старое значение | в разделе **New Value** выбрать **Value** | ввести новое значение | кнопка **Add** (правило перекодировки появится в окне) | задать последовательно все правила перекодировки, проверить их наличие в окне | **Continue** | **Change** | **OK.** В качестве **Old Value** может быть выбрано пропущенное значение (**System-missing** или **System- or user-missing**). В качестве **New Value** может быть выбран системный пропуск (**System-missing**).
Перекодировка дискретных значений (в другую переменную)	*Окно редактора данных*: **Transform** | **Recode** | **Into Different Variable(s)…** | перенести имя перекодируемой переменной в окно **Input Variable → Output Variable** | ввести имя новой переменной в отрывшееся окно **Output Variable Name** | кнопка **Old and New Values** | в разделе **Old Value** выбрать **Value** | ввести старое значение | в разделе **New Value** выбрать **Value** | ввести новое значение | кнопка **Add** (правило перекодировки появится в окне) | задать последовательно все правила перекодировки, проверить их наличие в окне | **Continue** | **Change** | **OK.** В качестве **Old Value** может быть выбрано пропущенное значение (**System-missing** или **System- or user-missing**). В качестве **New Value** может быть выбран системный пропуск (**System-missing**).

Действие	реализация
Группировка (построение интервалов) *Рекомендуется выполнять только в другую переменную.*	*Окно редактора данных*: **Transform** | **Recode** | **Into Different Variable(s)…** | перенести имя перекодируемой переменной в окно **Input Variable → Output Variable** | ввести имя новой переменной в отрывшееся окно **Output Variable Name** | кнопка **Old and New Values** | в разделе **Old Value** выбрать один из трех вариантов интервалов **Range** (с двумя границами, с открытой верхней границей или с открытой нижней границей) | задать границы интервала | в разделе **New Value** выбрать **Value** | ввести новое значение | кнопка **Add** (правило перекодировки появится в окне) | задать последовательно все интервалы, проверить наличие перекодировок в окне | **Continue** | **Change** | **OK.** В качестве **New Value** может быть выбран системный пропуск (**System-missing**).
Процентильная группировка	**Analyze** | **descriptive statistics** | **frequencies…** | имена группируемых переменных переместить в окно **Variable(s)** | кнопка **Statistics…** | в разделе **Percentile Values** выбрать один из трех вариантов группировки | **Continue** | **OK** | полученные границы использовать для построения интервалов ("Группировка").
Вычисление переменных с использованием арифметических операций и стандартных математических функций	*Окно редактора данных*: **Transform** | **Compute…** | ввести имя вычисляемой переменной в окно **Target Variable** | в окне **Numeric Expression** задать формулу для вычисления, с использованием арифметических выражений и/или стандартных математических функций | **OK**.
Вычисление переменных в соответствии с некоторыми условиями.	*Окно редактора данных*: **Transform** | **Compute…** | ввести имя вычисляемой переменной в окно **Target Variable** | в окне **Numeric Expression** задать число или формулу для вычисления | кнопка **If…** | выбрать **Include if case satisfies condition** | в окне задать условие | **Continue** | **OK** | На вопрос **Change existing variable?** ответить **OK.** Повторять столько раз, сколько необходимо задать условий.

Действие	реализация
Вычисление переменных: подсчет одинаковых значений в нескольких переменных.	*Окно редактора данных*: **Transform** | **Count…** | в окне **Target Variable** задать имя новой переменной | в окно **Variables** поместить имена переменных, для которых производится подсчет одинаковых значений | кнопка **Define Values** | в разделе **Values** выбрать **Value** | ввести значение подсчитываемых оценок | кнопка **Add** | **Continue** | **OK**
Стандартизация переменной (вычисление *z* -оценок)	*Окно редактора данных*: **Transform** | **Rank Cases…** | поместить имена переменных в окно **Variable(s)** | кнопка **Rank Types…** | кнопка **More >>** | флажок **Normal Scores…** | **Continue** | **OK** В результате в матрице данных появится новая переменная с тем же именем, к которому впереди добавлена буква *n*.
Обмен данными и результатами с другими приложениями Windows	
Чтение файлов данных SPSS и других форматов	**File** | **Open** | **Data…** | в открывшемся окне перейти в нужную папку | выбрать нужный тип файла (*.sav, *.dbf, *.xls и т.п.) | выбрать нужный файл | кнопка **открыть.**
Сохранение файлов данных SPSS и других форматов	*Окно редактора данных*: **File** | **Save As…** | в открывшемся окне перейти в нужную папку | выбрать нужный тип файла (*.sav, *.dbf, *.xls, *.dat и т.п.) | ввести имя файла | кнопка **сохранить.**
Копирование результатов, полученных в SPSS, в другие приложения Windows (Word, Excel, PowerPoint и т.п.)	*Окно результатов:* выделить нужную таблицу или рисунок | открыть контекстное меню (правая кнопка мыши) | **Copy objects** | перейти в другое приложение | поместить курсор в место, куда следует вставить объект | вызвать контекстное меню | **вставить.**

ЗАДАНИЕ 4. ПОСТРОЕНИЕ ОДНОМЕРНЫХ РАСПРЕДЕЛЕНИЙ

1. Загрузите файл с данными своего исследования.
2. Постройте одномерные распределения для всех переменных.
3. Проверьте по распределениям качество ввода данных (отсутствие значений, выходящих за допустимые пределы); корректность груп-

пировок; правильность использования пропущенных значений – системных и пользовательских.
4. При необходимости исправьте найденные ошибки и повторите вычисления.
5. Откорректируйте файл результатов, оставив в нем только правильно выполненные задания.
6. Сохраните результатов под именем *lesson4.spo.*

Необходимая подготовка:
- Тема 6 программы курса.
- Задания 1–3 практикума.
- Наличие файла с данными проведенного исследования.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности: файл результатов *lesson4.spo* (на дискете).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 4

Основные понятия

Абсолютная частота (frequency f_i) значения переменной – количество объектов, обладающих данным значением.

Относительная частота ($\frac{f_i}{n} \times 100\%$, где n – объем выборки)– процент или доля объектов, обладающих данным значением.

Одномерное частотное распределение (frequency distribution) – таблица, содержащая значения переменной и их частот.

Валидный процент (valid percent) – без учета пропущенных значений, от числа опрошенных

Накопленная частота (cumulative frequency), используется только для порядковых и количественных переменных. Вычисляется по формуле $F_i = \sum_{j=1}^{i} f_i$ – количество (процент, доля) объектов, имеющих значения, не превосходящие текущее значение.

Частоты отдельных значений используются только для дискретных переменных.
Непрерывные переменные должны быть сгруппированы в интервалы[1]; понятие частоты относится к интервалу.
При необходимости, дискретные шкалы также могут быть сгруппированы.

Виды группировок:

- *типологическая* (интервалы произвольной длины, интерпретируются содержательно);
- *аналитическая* (интервалы одинаковой длины, содержательная интерпретация не требуется);
- *процентильная* (интервалы с одинаковыми частотами, длина – какая получится, интерпретация не требуется).

Построение одномерных распределений

Analyze | **descriptive statistics** | **frequencies…** | поместить имена нужных переменных в окно **Variable(s)** | **OK**

Пример:

отметка по статистике

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	3.8	4.0	4.0
	неуд.	3	11.5	12.0	16.0
	удовл.	6	23.1	24.0	40.0
	хорошо	9	34.6	36.0	76.0
	отлично	4	15.4	16.0	92.0
	6	2	7.7	8.0	100.0
	Total	25	96.2	100.0	
Missing	не явился	1	3.8		
Total		26	100.0		

ошибки ввода данных

[1] Построение группировок см. в Методических материалах к заданию 3.

ЗАДАНИЕ 5. ПОСТРОЕНИЕ И РЕДАКТИРОВАНИЕ ГРАФИКОВ

1. Загрузите файл с данными своего исследования.
2. Постройте для каждой переменной подходящий график.
3. Проверьте корректность построения графиков.
4. При необходимости отредактируйте графики.
5. Отредактируйте файл результатов, оставив в нем только правильно выполненные задания.
6. Сохраните файл результатов под именем *lesson5.spo*.

Необходимая подготовка:
- Тема 7 программы курса.
- Задания 1-3 практикума.
- Наличие файла с данными проведенного исследования.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности: файл результатов *lesson5.spo* (на дискете).

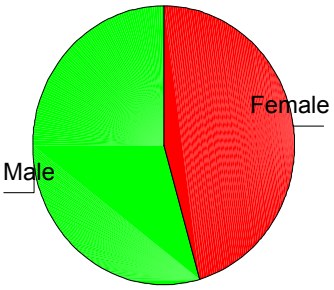
МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 5

Любые одномерные распределения могут быть представлены графически. Одномерные графики – средство визуализации распределений.

Основные требования к графикам и правила построения

1. График должен отражать уровень измерения переменной (для количественных и порядковых шкал направление оси должно соответствовать возрастанию значений; дискретность переменной подчеркивается промежутками между столбцами, непрерывность – отсутствием промежутков).
2. Частоты на графике, в большинстве случаев, изображаются площадями фигур.
3. Для оси частот, а также для осей порядковых и количественных переменных обязательно соблюдение масштаба.

Графики для дискретных переменных



Круговая диаграмма[2]

(только для номинальных шкал!).

Graphs | **Pie…** | выбрать **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** | поместить имя переменной в окно **Define Slices by** | **OK'**

Диаграмма столбцов

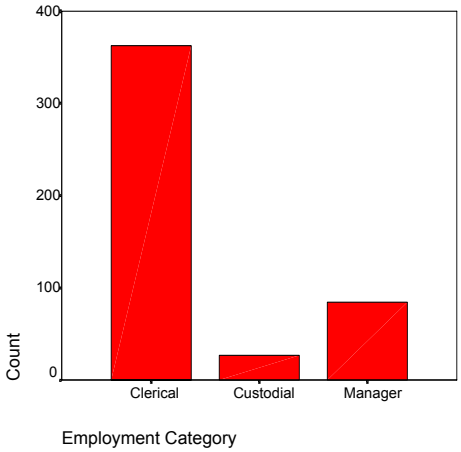
(для номинальных, порядковых, дискретных количественных шкал).

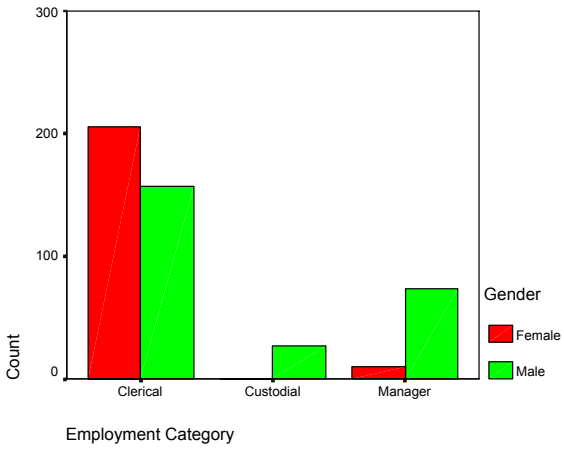
Graphs | **Bar…** | **Simple** | **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** для частот; **Cum. n of cases** или **Cum. % of cases** для накопленных частот | поместить имя переменной в окно **Category Axes** | **OK**

Кластеризованная диаграмма столбцов

(для сравнительного анализа распределений переменной в двух или нескольких группах).

Graphs | **Bar…** | **Clustered** | **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** для частот; **Cum. n of cases** или **Cum. % of cases** для накопленных частот | поместить имя группирующей переменной в окно **Define clusters by** | поместить имя переменной в окно **Category Axes** | **OK**



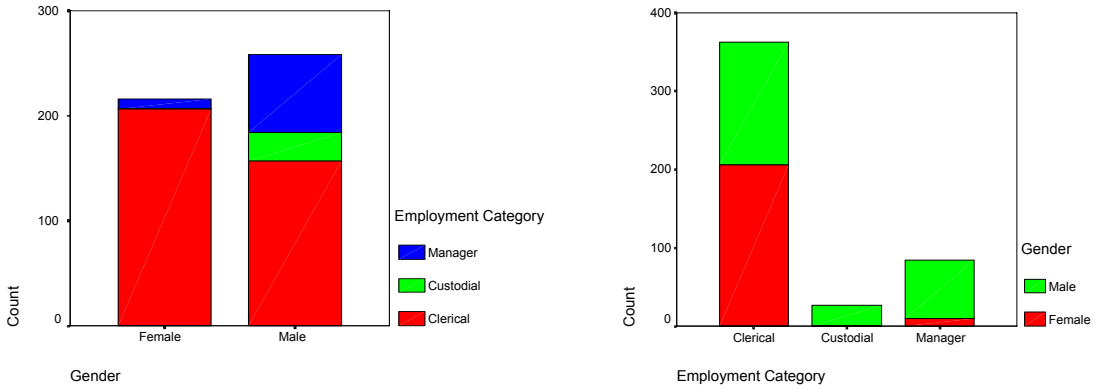


[2] Здесь и далее примеры на основе данных из файла *employee data.sav* из базы данных SPSS.

Ленточная диаграмма

(для исследований структуры группы объектов).

Graphs | **Bar…** | **Stacked** | **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** для частот; **Cum. n of cases** или **Cum. % of cases** для накопленных частот | поместить имя переменной в окно **Category Axes** | поместить имя структурирующей переменной в окно **Define stacks by** | **OK**



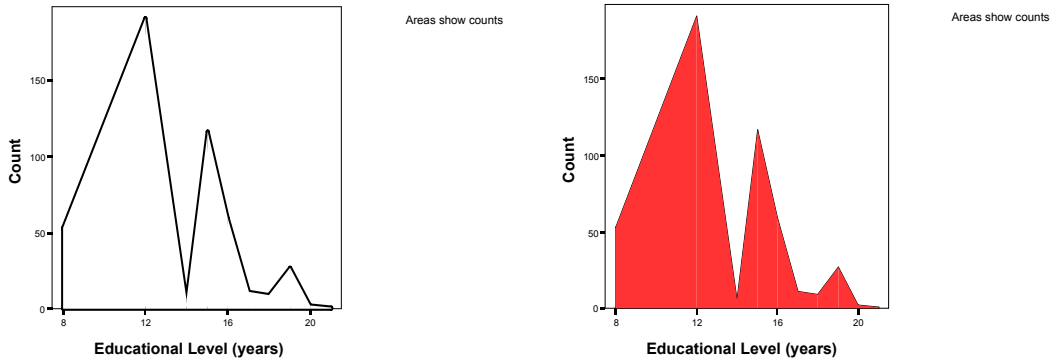
Два способа представления структуры выборки по полу и должности.

Полигон

(для порядковых и дискретных количественных шкал).

Graphs | **Line…** | **Simple** | **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** для частот; **Cum. n of cases** или **Cum. % of cases** для накопленных частот | поместить имя переменной в окно **Category Axes** | **OK**

Graphs | **Area…** | **Simple** | **Summaries for group of cases** | **Define** | выбрать **N of cases** или **% of cases** для частот; **Cum. n of cases** или **Cum. % of cases** для накопленных частот | поместить имя переменной в окно **Category Axes** | **OK**



Если некоторые из допустимых значений порядковой или количественной переменной не встречаются в выборке (имеют частоту, равную нулю), диаграммы и полигоны строятся с нарушениями масштаба по шкале переменной. В некоторых случаях эту ошибку удается исправить, построив график с помощью процедуры **Graphs | Interactive…**. Для этого необходимо убрать флажок **Exclude empty categories** в разделе **Options**.

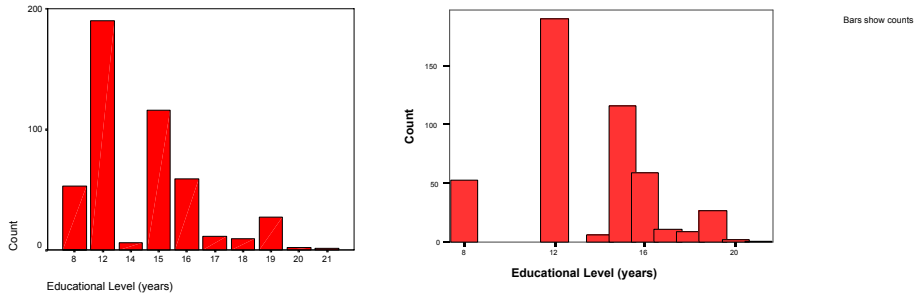


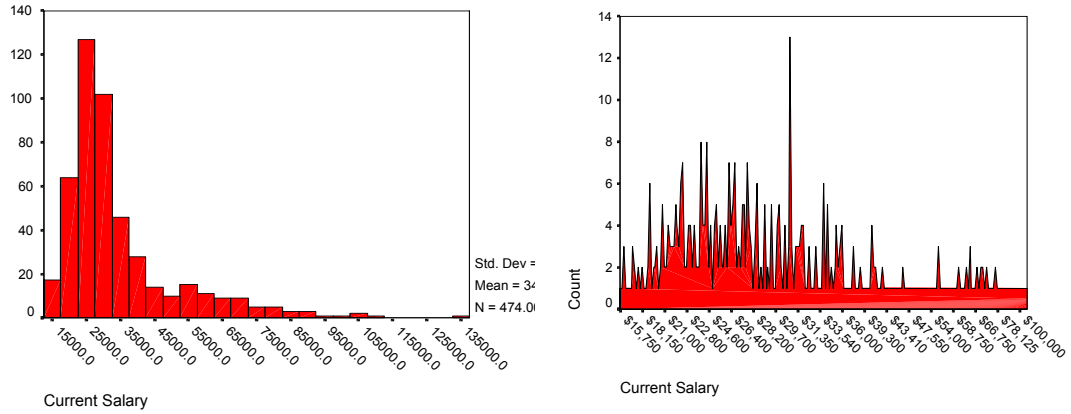
График слева построен с нарушением масштаба, график справа – правильно.

Графики для непрерывных переменных

Гистограмма

(только для непрерывных переменных;
SPSS автоматически осуществляет аналитическую группировку).

Graphs | Histogram… | Поместить имя переменной в окно **Variable | OK**



На графике слева – гистограмма, справа – полигон.

Редактирование графиков

При необходимости график можно отредактировать – изменить цвета, названия осей, масштаб и т.п. Для редактирования графика необходимо открыть окно **SPSS Chart Editor**, щелкнув дважды мышью по графику в окне результатов.

ЗАДАНИЕ 6. ВЫЧИСЛЕНИЕ ХАРАКТЕРИСТИК ОДНОМЕРНОГО РАСПРЕДЕЛЕНИЯ

1.1. Вычислите для всех переменных своего исследования адекватные описательные статистики с помощью процедур **Frequencies** или **Descriptives**.
1.2. Отредактируйте и сохраните файл результатов lesson6.spo.
1.3. Подготовьте (с использованием редактора MS Word) краткий отчет об основных результатах проведенного исследования; проиллюстрируйте его таблицами и графиками из файлов lesson4.spo, lesson5.spo, lesson6.spo.

2.1. Загрузите файл *employee data.sav* из базы данных SPSS.
2.2. Постройте гистограмму распределения переменной *salary* (*зарплата*).
2.3. Вычислите для переменной *salary* адекватные описательные статистики.
2.4. С помощью процедуры **Compute** вычислите z-оценки для переменной *salary* (назовите новую переменную z*salary*): $z = \left(x - \bar{x}\right)/s$.
2.5. Постройте гистограмму распределения переменной *zsalary.*
2.6. Вычислите для переменной *zsalary* адекватные описательные статистики.
2.7. Подготовьте краткий отчет о результатах сравнения распределений переменных *salary* и *zsalary*.

Необходимая подготовка:
- Темы 8-10 программы курса.
- Задания 1-5 практикума.
- Наличие файла с данными проведенного исследования; файлов *lesson4.spo, lesson5.spo.*

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson6.spo* (на дискете);
- отчет о проведенном исследовании (3–5 стр.);
- отчет о результатах сравнения распределений переменных *salary* и *zsalary* (1–2 стр.).

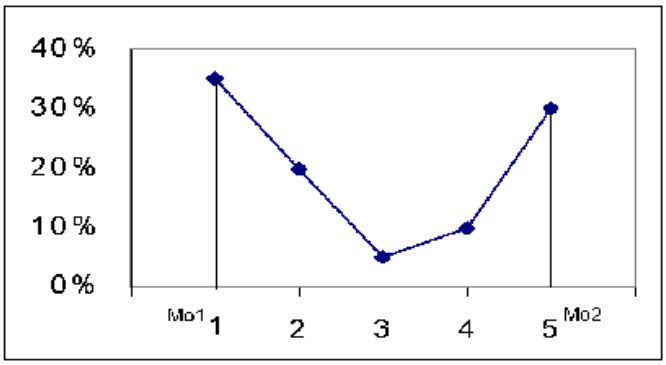
МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 6

Показатели центра распределения

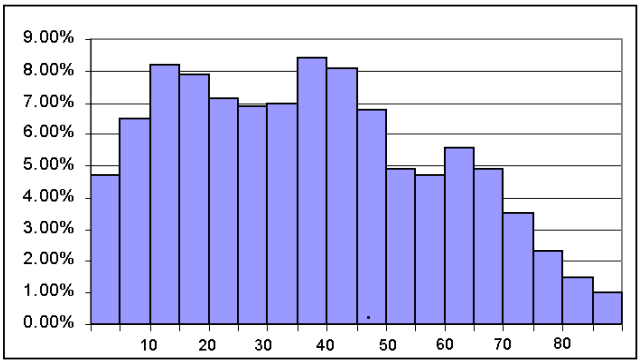
Для дискретной переменной ***модой*** (Mo) называется наиболее часто встречающееся значение признака, т.е. значение, обладающее максимальной частотой. Например, модальным полом на филологическом факультете университета является женский, на радиофизическом факультете – мужской.

Для непрерывной переменной говорят о *модальном интервале*, которому соответствует максимальная частота (для аналитической группировки).

Распределение может иметь более одной моды. Распределение с двумя модами называется *бимодальным*, с тремя и более модами – *полимодальным*. Определить би- или полимодальность проще всего при помощи графика – на нем в этом случае видны несколько ых "пиков".



Распределение степени согласия студентов с требованием разрешить в университете работу политических молодежных организаций имеет две моды: $Mo_1 = 1$ (частота 35%) и $Mo_2 = 5$ (частота 30%).



Распределение населения РБ по возрасту имеет три моды: в возрасте 10-15 лет, 35-40 лет и 60-65 лет

Мода является единственной универсальной мерой центральной тенденции; ее можно использовать для переменных всех типов – номинальных, порядковых и количественных, дискретных и непрерывных.

Для номинальных переменных мода является единственным корректным показателем центральной тенденции.

Медиана (Me) – граница 50%-ного интервала, значение переменной, которое делит вариационный ряд пополам – половина всех объектов из выборки имеют значения переменной, не превосходящие медиану, вторая половина – значения, которые больше, чем медиана.

Медиана определяется только для порядковых и количественных переменных; к номинальным переменным ее применять нельзя.

Среднее арифметическое ($\bar{x}$) – наиболее часто используемый показатель центра распределения *для количественных переменных*. Показывает, каким было бы значение переменной, если бы у всех объектов из выборки оно было одинаковым.

Среднее арифметическое представляет собой сумму всех значений переменной, разделенную на объем выборки:

$$\bar{x} = \sum x_i / n ,$$

где x_i – значение переменной x для объекта i;

n – объем выборки.

Среднее арифметическое может также вычисляться для дихотомических переменных. В этом случае $\bar{x} = p$, где p – доля положительных ответов.

Показатели степени разброса данных

Диапазон разброса значений переменной является простейшей мерой степени разброса данных: $d = x_{\max} - x_{\min}$.

Среднее квадратическое отклонение (СКО) показывает, насколько индивидуальные значения переменной в среднем отклоняются от среднего арифметического:

$$s = \sqrt{\sum \left(x_i - \bar{x}\right)^2 / (n-1)},$$

где x_i – значение переменной для объекта с номером i;
$\bar{x}$ – среднее арифметическое;
n – объем выборки.

В некоторых задачах вместо среднего квадратического отклонения используется **дисперсия**[3], которая представляет собой квадрат СКО:

$$s^2 = \sum \left(x_i - \bar{x}\right)^2 / (n-1).$$

Дисперсия и СКО могут вычисляться для количественных и дихотомических переменных. Для дихотомической переменной $s^2 = p(1-p)$.

Показатели формы распределения

Коэффициент асимметрии предназначен для проверки симметричности одномодального распределения:

$$\beta_1 = \sum \left(x_i - \bar{x}\right)^3 / ns^3 .$$

Распределение симметрично, если $\beta_1 = 0$; имеет положительную (левую) асимметрию, если $\beta_1 > 0$; имеет отрицательную (правую) асимметрию, если $\beta_1 < 1$.

Коэффициент эксцесса характеризует форму одномодального симметричного распределения:

[3] От англ. *dispersion* – рассеяние. Тем не менее по-английски дисперсия – *variance* (изменчивость).

$$\beta_2 = \sum \left(x_i - \bar{x}\right)^4 / n\, s^4 .$$

Распределение имеет форму нормального распределения, если $\beta_2 = 3$; является островершинным, если $\beta_2 > 3$; является плосковершинным, если $\beta_2 < 3$.

Вычисление характеристик распределения

Функция **Descriptives**:

Analyze | **Descriptive Statistics** | **Descriptives** | поместить имена переменных в окно **Variables** | клавиша **Options** | установить необходимые опции | **Continue** | **OK**

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Educational Level (yea	474	13	8	21	13.49	2.88	8.322
Valid N (listwise)	474						

Функция **Frequencies**:

Analyze | **Descriptive Statistics** | **Frequencies** | поместить имена переменных в окно **Variables** | клавиша **Statistics** | установить необходимые опции | **Continue** | **OK**

Statistics

Educational Level (years)

N	Valid	474
	Missing	2
Mean		13.49
Median		12.00
Mode		12
Std. Deviation		2.88
Variance		8.32
Skewness		-.114
Std. Error of Skewness		.112
Kurtosis		-.265
Std. Error of Kurtosis		.224
Range		13
Minimum		8
Maximum		21

Вывод распределения частот можно подавить, сняв флажок **Display frequency table**.

Опции параметра **Options** функции **Descriptives** и параметра **Statistics** функции **Frequencies**:

Mean	Среднее арифметическое
Median	Медиана
Mode	Мода
Minimum	Минимум
Maximum	Максимум
Range	Диапазон
Std. Deviation	СКО
Variance	Дисперсия
Skewness	Асимметрия
Kurtosis	Эксцесс

ЗАДАНИЕ 7. ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ И ОШИБОК ВЫБОРКИ

1. Оцените ошибку выборки для каждой переменной ($\alpha = 0.05; 0.01$).
2. Оцените с помощью доверительного интервала математические ожидания для количественных и вероятности положительных ответов для дихотомических переменных ($\alpha = 0.05; 0.01$).
3. Отредактируйте и сохраните файл результатов *lesson7.spo*.
4. Подготовьте (с использованием редактора MS Word) краткий отчет о параметрах ГС и репрезентативности выборки; приведите в нем необходимые данные из файла *lesson7.spo*.

Необходимая подготовка:

- Темы 12–13 программы курса.
- Задания 1–6 практикума.
- Наличие файла с данными проведенного исследования.

Методическое указание:

после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:

- файл результатов *lesson7.spo* (на дискете);
- отчет о репрезентативности проведенного исследования и параметрах ГС (2–3 стр.).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 7

***Статистикой выборки* (*выборочной статистикой*)** называется любая количественная характеристика выборки.

Параметром генеральной совокупности называется любая количественная характеристика ГС.

Цель выборочного метода – по значению выборочной статистики оценить значение параметра ГС.

Параметры ГС	***Выборочные статистики***
Вероятность p	Частота (относительная в долях) f
Функция распределения F	Накопленная частота F
Математическое ожидание μ	Среднее арифметическое $\bar{x}$
Дисперсия σ^2	Дисперсия s^2
СКО σ	СКО s

Следствие из Центральной предельной теоремы: если из бесконечной генеральной совокупности с математическим ожиданием μ дисперсией σ^2 извлечь бесконечное число случайных выборок одного и того же объема n, то выборочное среднее арифметическое будет иметь нормальное распределение с математическим ожиданием μ и дисперсией σ^2/n.

Этот теоретический результат позволяет оценить ошибку выборки и значения параметров ГС.

Ошибка выборки

Ошибкой выборки называется разносить между значениями выборочной статистики и параметра ГС: $\Delta = \bar{x} - \mu$ (для дихотомической переменной $\Delta = p_{выб} - p_{ген}$).

Ошибка выборки измеряется в тех же единицах, что и переменная.

Если точное значение параметра неизвестно, то нельзя вычислить и точное значение ошибки выборки. В этом случае ее необходимо оценить статистически с помощью доверительного интервала.

Доверительным интервалом для ошибки выборки называется интервал, в который она попадает с заданной *доверительной вероятностью* $1-\alpha$**:** $|\Delta| \le Z_{1-\alpha/2} \times SE$ или $-Z_{1-\alpha/2} \times SE \le \Delta \le +Z_{1-\alpha/2} \times SE$,

где $SE = \sqrt{s^2/n}$ – стандартная ошибка среднего арифметического (Standard Error), вычисляется по выборке,

$Z_{1-\alpha/2}$ – доверительный коэффициент, зависящий от выбранного уровня доверительной вероятности:

- при $1-\alpha = 0.9$ $Z_{0.95} = 1.64$;
- при $1-\alpha = 0.95$ $Z_{0.975} = 1.96$;
- при $1-\alpha = 0.99$ $Z_{0.995} = 2.58$.

Вычисление стандартной ошибки выборки и доверительного интервала для ошибки выборки

Функция **Descriptives**:

Analyze | **Descriptive statistics** | **Descriptives** | поместить имена переменных в окно **Variables** | кнопка **Options** | установить опции **Mean** и **S.E. mean** | **Continue** | **OK**

Descriptive Statistics

	N	Mean	
	Statistic	Statistic	Std. Error
Current Salary	476	$34,540.44	$798.23
Valid N (listwise)	476		

Стандартная ошибка для средней зарплаты (Current Salary)
SE = 798 *долларов.*

Ошибка выборки по отношению к средней зарплате ($34540):

- *с вероятностью 90% не превысит $1294 (*$1.64 \times \$798$*);*
- *с вероятностью 95% не превысит $1564 (*$1.96 \times \$798$*);*
- *с вероятностью 99% не превысит $2059 (*$2.58 \times \$798$*).*

Функция **Frequencies**:

Analyze | Descriptive statistics | **Frequencies** | поместить имена переменных в окно **Variables** | кнопка **Statistics** | установить опции **Mean** и **S.E. mean** | **Continue** | **OK**

Statistics

Minority Classification

N	Valid	474
	Missing	2
Mean		.22
Std. Error of Mean		1.90E-02

Стандартная ошибка для дихотомической переменной 'Принадлежность к национальному меньшинству' (Minority Classification)

$$SE = 1.90E - 02 = 0.019$$

Ошибка выборки по отношению к доле национального меньшинства (0.22):

- *с вероятностью 90% не превысит 3.1% (*$1.64 \times 0.019 \approx 0.031$*);*
- *с вероятностью 95% не превысит 3.7% (*$1.96 \times 0.019 \approx 0.037$ *);*
- *с вероятностью 99% не превысит 4.9% (*$2.58 \times 0.019 \approx 0.049$*).*

Если ошибка выборки вычисляется для частоты значения номинальной переменной, необходимо перекодировать исходную переменную в дихотомическую и вычислить для нее стандартную ошибку.

Оценивание параметров генеральной совокупности

Точечной оценкой параметра ГС называется функция от выборки, позволяющая приблизительно оценить значение параметра. Этому критерию удовлетворяют, в частности, выборочные статистики, многие из которых с успехом используются в качестве точечных оценок параметров ГС: $\mu \approx \bar{x}$, $\sigma \approx s$, $p_{ген} \approx p_{выб}$ и т.п.

Точность точечных оценок является достаточно неопределенной. Поэтому чаще применяется интервальное оценивание параметров ГС (в первую очередь, математического ожидания μ и вероятности p).

Вычисление доверительных интервалов для математического ожидания и вероятности

Математическое ожидание μ с заданной доверительной вероятностью $1-\alpha$ находится в интервале $\bar{x} \pm Z_{1-\alpha/2} \times SE$ или, что то же самое, в интервале $(\bar{x} - Z_{1-\alpha/2} \times SE; \bar{x} + Z_{1-\alpha/2} \times SE)$.

Средняя зарплата по ГС:

- *с вероятностью 90% находится в интервале \$34540± \$1294 (\$33246÷\$35834);*
- *с вероятностью 95% находится в интервале \$34540± \$1564 (\$32978÷\$36104);*
- *с вероятностью 99% находится в интервале \$34540± \$2059 (\$32481÷\$ 36599).*

Вероятность положительного ответа $p_{ген}$ (для дихотомической переменной) с заданной доверительной вероятностью $1-\alpha$ находится в интервале $p_{выб} \pm Z_{1-\alpha/2} \times SE$ или, что то же самое, в интервале $(p_{выб} - Z_{1-\alpha/2} \times SE; p_{выб} + Z_{1-\alpha/2} \times SE)$, где $p_{выб}$ – доля положительных ответов, вычисленная по выборке.

Численность национального меньшинства в ГС:

- *с вероятностью 90% находится в пределах 22%±3.1% (18.9%; 25.1%);*
- *с вероятностью 95% находится в пределах 22%±3.7% (18.3%; 25.7%);*
- *с вероятностью 99% находится в пределах 22%±4.9% (17.1%; 26.9%).*

Если доверительный интервал должен быть вычислен для вероятности значения номинальной или порядковой переменной, переменную следует предварительно дихотомизировать, перекодировав нужное значение в 1, все остальные – в 0.

ЗАДАНИЕ 8. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

1. Обсудите с преподавателем статистики набор гипотез, которые можно проверить по данным Вашего исследования.
2. Проверьте выбранные гипотезы.
3. Отредактируйте и сохраните файл результатов *lesson8.spo*.
4. Подготовьте (с использованием редактора MS Word) отчет о результатах проверки гипотез; приведите в нем необходимые данные из файла *lesson8.spo*.

Необходимая подготовка:
- Тема 14 программы курса.
- Задания 1-6 практикума.
- Наличие файла с данными проведенного исследования.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson8.spo* (на дискете);
- отчет о результатах проверки гипотез (3–5 стр.).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 8

Статистической гипотезой называется утверждение относительно параметров генеральной совокупности, сформулированное по определенным правилам. Любая содержательная гипотеза может быть сформулирована в виде статистической.

Статистическая гипотеза состоит из двух утверждений. Первое утверждение, постулирующее отсутствие различий между частями ГС или связи между переменными, называется *нулевой гипотезой* и обозначается H_0. Во втором утверждении, которое называется *альтернативной гипотезой* и обозначается H_1, формулируется наличие определенных различий между структурными частями ГС или связи между переменными.

Статистическая гипотеза формулируется в терминах параметров ГС с использованием отношений равенства (H_0) и неравенства – не равно, больше, меньше (H_1).

Примеры:

1. Гипотеза: зарплата женщин, в среднем, ниже, чем зарплата мужчин.
$H_0 : \mu_{ж} = \mu_{м}$ *– средняя зарплата женщин и мужчин одинакова.*
$H_1 : \mu_{ж} < \mu_{м}$ *– средняя зарплата женщин ниже, чем средняя зарплата мужчин.*

2. Гипотеза: на должность менеджера наниматели предпочитают брать мужчин.
$H_0 : p_м = 0.5$ *– среди менеджеров мужчины составляют 50%.*
$H_1 : p_м > 0.5$ *– среди менеджеров мужчины составляют более 50%.*

Альтернативные гипотезы, в соответствии с используемым отношением неравенства, делят на *двусторонние* ($\neq$) и *односторонние* (<, >); односторонние альтернативные гипотезы, в свою очередь, можно разделить на *левосторонние* (<) и *правосторонние* (>).

Проверка статистической гипотезы состоит в том, чтобы по данным выборочного исследования сделать вывод, следует ли в отношении ГС принять нулевую гипотезу или отклонить ее в пользу альтернативной. При этом нулевая гипотеза считается справедливой до тех пор, пока не будет найдено убедительное подтверждение того, что она не верна.

Решение о принятии или отклонении нулевой гипотезы принимается в соответствии с *критерием*, который строится на основе специально подобранной для каждой нулевой гипотезы численной функции ω. Функция ω вычисляется по выборке и называется *статистикой критерия*.

Поскольку решение (принять или отклонить H_0) принимается на основе выборки, оно может быть как правильным, так и ошибочным. При этом возможны два типа ошибок. Ошибка, заключающаяся в том, чтобы по данным выборки отклонить нулевую гипотезу, которая на самом деле верна, называется *ошибкой первого рода*; ее вероятность обозначается буквой α. Ошибка, состоящая в том, чтобы принять нулевую гипотезу, которая на самом деле не верна, называется *ошибкой второго рода*; ее вероятность обозначается буквой β.

Решение, принятое по выборке:	**Генеральная совокупность**	
	H_0 ***верна***	H_0 ***не верна***
Принять H_0	верное решение, вероятность $1-\alpha$	ошибка II рода, вероятность β
Отклонить H_0	ошибка I рода, вероятность α	верное решение, вероятность $1-\beta$

Проблема заключается в том, что с уменьшением вероятности ошибки I рода увеличивается вероятность ошибки II рода и наоборот.

Эту проблему учитывают при разработке критериев, благодаря чему исследователь, не являющийся профессиональным статистиком, может на практике руководствоваться простым правилом:

При проверке статистической гипотезы фиксируют некоторое малое значение α ($\alpha = 0.10,\ 0.05,\ 0.01$) и надеются, что β также будет мало.

Фиксированное значение α называется ***уровнем значимости.***

Процедура проверки гипотезы заключается в том, чтобы сравнить выбранный уровень значимости α с вычисленной SPSS величиной, которая называется *p –значением* и обозначается в распечатке *Sig.* (значимость) или *Sig. (2-tailed)* (двусторонняя значимость).

Для двусторонней альтернативной гипотезы:
- принимается H_1, если $Sig.(2-tailed) < \alpha$;
- принимается H_0, если $Sig.(2-tailed) > \alpha$.

Для правосторонней альтернативной гипотезы:
- принимается H_1, если $Sig.(2-tailed) < 2\alpha$ и вычисленное значение критерия ***больше*** 0;
- принимается H_0, во всех остальных случаях.

Для левосторонней альтернативной гипотезы:
- принимается H_1, если $Sig.(2-tailed) < 2\alpha$ и вычисленное значение критерия ***меньше*** 0;
- принимается H_0 во всех остальных случаях.

Проверка статистических гипотез

Сравнение средних арифметических и дисперсий в двух выборках

Analyze | **Compare means** | **Independent-Samples T Test** | поместить имя переменной, для которой сравниваются средние, в окно **Test variables** | поместить имя группирующей переменной в окно **Grouping variables** | щелкнуть по имени группирующей переменной | в открывшемся окне задать значения, определяющие группы | **Continue** | **OK**

Сравнение относительных частот (долей) в двух выборках

Используйте ту же процедуру, предварительно перекодировав переменную, для которой сравниваются относительные частоты, в дихотомическую.

Пример.
Проверить гипотезу о том, что зарплата женщин в среднем ниже, чем зарплата мужчин.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	Female	216	$26,031.92	$7,558.02	$514.26
	Male	258	$41,441.78	$19,499.21	$1,213.97

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Current Salary	Equal variances assumed	95.875	.000	-10.286	474	.000	-$14920.8	$1450.7	-$17771	-$12070
	Equal variances not assumed			-10.864	387	.000	-$14920.8	$1373.4	-$17621	-$12221

1. *По выборке средняя зарплата женщин $26032, мужчин – $41442.*
2. *Выберем для проверки гипотезы о равенстве дисперсий (F -критерий Фишера)* $\alpha = 0.05$. *Для дисперсий* $Sig. = 0.000 < 0.05$ *;следовательно, принимается гипотеза о том, что дисперсии зарплаты мужчин и женщин не равны. Следовательно, для проверки гипотезы о средних выбирается двухвыборочный t -критерий, не предполагающий равенства дисперсий (Equal variance assumed).*
3. *Выберем для проверки гипотезы о равенстве средних* $\alpha = 0.01$. *Для средних* $Sig.(2-tailed) = 0.000 < 2\alpha = 0.02$. *Кроме того,* $t = -10.945 < 0$, *что соответствует левосторонней альтернативной гипотезе. Следовательно, принимается альтернативная гипотеза, согласно которой средняя зарплата женщин ниже, чем средняя зарплата мужчин.*

Сравнение среднего арифметического с числом

Analyze | **Compare means** | **One-Sample T Test** | поместить имя переменной в окно **Test variables** | поместить число, с которым производится сравнение, в окно **Test Value** | **Continue** | **OK**

Сравнение относительной частоты с числом

Используйте ту же процедуру, предварительно перекодировав переменную в дихотомическую.

Пример:
проверить гипотезу о том, что должности менеджеров чаще занимают мужчины, чем женщины.

1. *Перекодируем переменную 'Пол' в дихотомическую (1 - мужской, 0 - женский).*
2. *Отберем для анализа группу менеджеров.*
3. *Доля мужчин среди менеджеров по выборке – 88% (mean).*

One-Sample Statistics

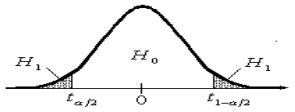
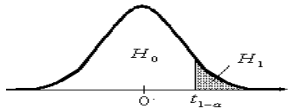
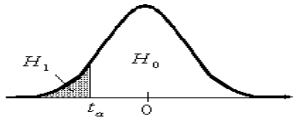
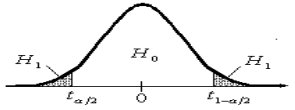
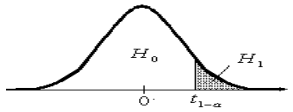
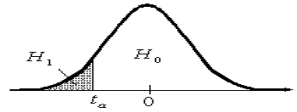
	N	Mean	Std. Deviation	Std. Error Mean
SEX	84	.8810	.3258	3.555E-02

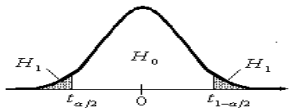
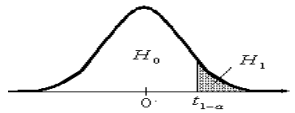
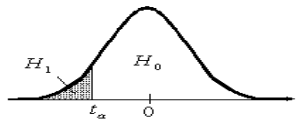
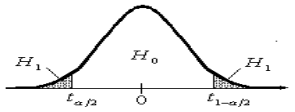
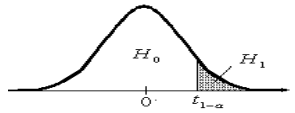
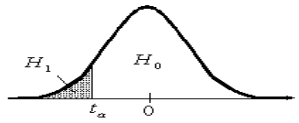
One-Sample Test

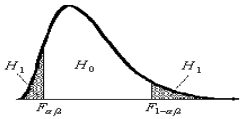
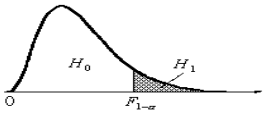
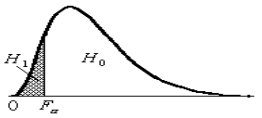
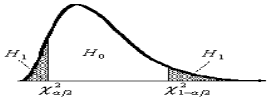
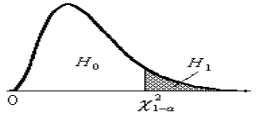
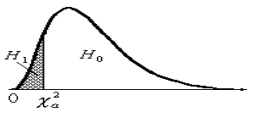
	Test Value = 0.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
SEX	10.717	83	.000	.3810	.3103	.4517

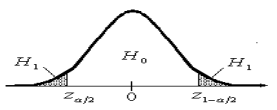
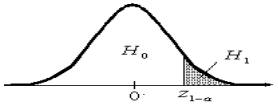
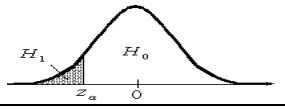
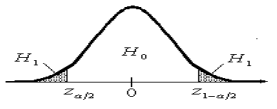
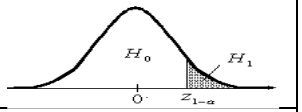
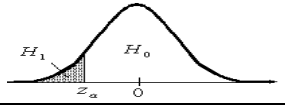
4. *Выберем значение* α*, например,* $\alpha = 0.01$.
5. $Sig.(2-tailed) = 0.000 < 2\alpha = 0.02$ *Кроме того* $t = +10.717 > 0$*, что соответствует правосторонней альтернативной гипотезе, следовательно, принимается гипотеза* H_1 *(при уровне значимости* $\alpha = 0.01$*).*

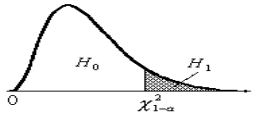
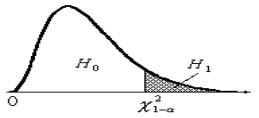
Некоторые статистические критерии для проверки гипотез

Гипотеза	Предположения	Статистика критерия	Критическая область	Примечания
Сравнение выборочного среднего арифметического с числом				
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	Генеральная совокупность имеет нормальное распределение с дисперсией σ^2. Дисперсия σ^2 неизвестна; используется оценка $\sigma^2 \approx s^2$	$t_н = \dfrac{\bar{x} - \mu_0}{\sqrt{s^2/n}}$	$\lvert t_н \rvert > t_{1-\alpha/2}$ 	t-распределение Стьюдента $df = n - 1$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$			$t_н > t_{1-\alpha}$ 	
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$			$t_н < t_\alpha$ 	
Сравнение средних арифметических двух независимых выборок с равными дисперсиями (одновыборочный t-критерий)				
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	Выборки извлечены из генеральных совокупностей с дисперсиями σ_1^2 и σ_2^2. Дисперсии σ_1^2 и σ_2^2. неизвестны, но приблизительно равны $\sigma_1^2 \approx \sigma_2^2$	$t_н = \dfrac{\bar{x}_1 - \bar{x}_2}{s_0}\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$, где $s_0 = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$	$\lvert t_н \rvert > t_{1-\alpha/2}$ 	t-распределение Стьюдента $df = n_1 + n_2 - 2$
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$			$t_н > t_{1-\alpha}$ 	
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$			$t_н < t_\alpha$ 	

Гипотеза	Предположения	Статистика критерия	Критическая область	Примечания
Сравнение средних арифметических двух независимых выборок с неравными дисперсиями (двухвыборочный t-критерий)				
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	Выборки извлечены из генеральных совокупностей с дисперсиями σ_1^2 и σ_2^2. Дисперсии σ_1^2 и σ_2^2 неизвестны и неравны $\sigma_1^2 \neq \sigma_2^2$	$t_н = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	$\lvert t_н \rvert > t_{1-\alpha/2}$ 	t-распределение Стьюдента $df = \dfrac{\left(s_1^2/n_1 + s_2^2/n_2\right.}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(s_2^2/n\right.}{n_2+}$
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$			$t_н > t_{1-\alpha}$ 	
$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$			$t_н < t_\alpha$ 	
Сравнение средних арифметических двух зависимых (связанных) выборок				
$H_0 : \mu_d = 0$ $H_1 : \mu_d \neq 0$	μ_d – математическое ожидание разностей $d_i = x_i - y_i$, вычисленных для каждой пары i.	$t_н = \dfrac{\bar{d}}{\sqrt{s_d^2/n}}$, где $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$ $s_d^2 = \dfrac{\sum \left(d_i - \bar{d}\right)^2}{n-1}$	$\lvert t_н \rvert > t_{1-\alpha/2}$ 	t-распределений Стьюдента $df = n - 1$
$H_0 : \mu_d = 0$ $H_1 : \mu_d > 0$			$t_н > t_{1-\alpha}$ 	
$H_0 : \mu_d = 0$ $H_1 : \mu_d < 0$			$t_н < t_\alpha$ 	

Гипотеза	Предположения	Статистика критерия	Критическая область	Примечания
Сравнение дисперсий двух независимых выборок				
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	выборки независимы, извлечены из генеральных совокупностей с нормальным распределением.	$F_н = \frac{s_1^2}{s_2^2}$	$F_н > F_{1-\alpha/2}$ $F_н < F_{\alpha/2}$ 	F-распределение Фишера $df_1 = n_1 - 1$ $df_2 = n_2 - 1$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$			$F_н > F_{1-\alpha}$ 	
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$			$F_н < F_{\alpha}$ 	
Сравнение дисперсий двух зависимых (связанных) выборок				
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	выборки зависимы	$\chi_н^2 = (n-1)\frac{s_1^2}{s_2^2}$	$\chi_н^2 > \chi_{1-\alpha/2}^2$ $\chi_н^2 < \chi_{\alpha/2}^2$ 	χ^2-распределение $df = n - 1$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$			$\chi_н^2 > \chi_{1-\alpha}^2$ 	
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$			$\chi_н^2 < \chi_{\alpha}^2$ 	

Гипотеза	Предположения	Статистика критерия	Критическая область	Примечания
Сравнение доли положительных ответов с числом				
$H_0 : p = p_0$ $H_1 : p \neq p_0$	дихотомическая переменная x $p = \Pr\{x = 1\}$	$z_н = \dfrac{p - p_0}{\sqrt{p_0(1-p_0)/n}}$	$\lvert z_н \rvert > z_{1-\alpha/2}$ 	Стандартное нормальное распределение $Z(0,1)$
$H_0 : p = p_0$ $H_1 : p > p_0$			$z_н > z_{1-\alpha}$ 	
$H_0 : p = p_0$ $H_1 : p < p_0$			$z_н < z_\alpha$ 	
Сравнение доли положительных ответов в двух выборках				
$H_0 : p_1 = p_2$ $H_1 : p_1 \neq p_2$	дихотомическая переменная x $p = \Pr\{x = 1\}$	$z_н = \dfrac{p_1 - p_2}{\sqrt{\dfrac{n_1 + n_2}{n_1 n_2} p_0(1-p_0)}}$ где $p_0 = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$	$\lvert z_н \rvert > z_{1-\alpha/2}$ 	Стандартное нормальное распределение $Z(0,1)$
$H_0 : p_1 = p_2$ $H_1 : p_1 > p_2$			$z_н > z_{1-\alpha}$ 	
$H_0 : p_1 = p_2$ $H_1 : p_1 < p_2$			$z_н < z_\alpha$ 	

Гипотеза	Предположения	Статистика критерия	Критическая область	Примечания
Сравнение эмпирического распределения с теоретическим				
$H_0 : p_i = p_i^0$, $\forall i = 1,k$ $H_1 : p_i \neq p_i^0$, для некоторых $i = \overline{1,k}$	Переменная x имеет k значений $\sum_{i=1}^{k} p_i^0 = 1$	$\chi^2_н = n\sum_{i=1}^{k} \frac{\left(p_i - p_i^0\right)^2}{p_i^0}$	$\chi^2_н > \chi^2_{1-\alpha}$ 	Распределение χ^2 $df = k - 1$
Сравнение распределений двух выборок				
$H_0 : p_i^{(1)} = p_i^{(2)}$ $\forall i = 1,k$ $H_1 : p_i^{(1)} \neq p_i^{(2)}$ для некоторых $i = \overline{1,k}$	Переменная x имеет k значений $\sum_{i=1}^{k} p_i^{(1)} = 1$ $\sum_{i=1}^{k} p_i^{(2)} = 1$	$\chi^2_н = n_1 n_2 \sum_{i=1}^{k} \frac{\left(p_i^{(1)} - p_i^{(2)}\right)^2}{n_1 p_i^{(1)} + n_2 p_i^{(2)}}$	$\chi^2_н > \chi^2_{1-\alpha}$ 	Распределение χ^2 $df = k - 1$

ЗАДАНИЕ 9. ПОСТРОЕНИЕ И АНАЛИЗ ТАБЛИЦ СОПРЯЖЕННОСТИ

1. Откройте файл *GSS93.sav* из базы данных SPSS.
2. Постройте таблицу сопряженности для двух номинальных переменных; измерьте связь между ними.
3. Постройте таблицу сопряженности для двух порядковых переменных; измерьте связь между ними.
4. Постройте таблицу сопряженности для двух количественных переменных; измерьте связь между ними (***Методическое указание****: при необходимости сгруппируйте переменные в интервалы*).
5. Отредактируйте и сохраните файл результатов *lesson9.spo.*
6. Подготовьте (с использованием редактора MS Word) отчет о результатах; приведите в нем необходимые данные из файла *lesson9.spo*.

Необходимая подготовка:
- Темы 14-18 программы курса.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson9.spo* (на дискете);
- отчет об исследовании связей по таблицам сопряженности (3–4 стр.).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 9

Статистической связью между двумя переменными называется такое их соотношение, когда изменение *значения* одной переменной приводит к изменению *распределения* второй переменной. Отсутствие связи между двумя переменными называется ***статистической независимостью***.

Таблица сопряженности – наиболее универсальное средство исследования статистических связей – представляет совместное распределение двух переменных. *Строки таблицы* образуются значе-

ниями одной переменной. *Столбцы таблицы* образуются значениями второй переменной. В клетке таблицы (на пересечении строки и столбца) указывается *частота совместного появления* соответствующих значений. Суммы частот по строке или по столбцу называются *маргинальными частотами. Распределения маргинальных частот* представляют собой одномерные распределения переменных.

Gender * Employment Category Crosstabulation

Count

		Employment Category			
		Clerical	Custodial	Manager	Total
Gender	Female	206		10	216
	Male	157	27	74	258
Total		363	27	84	474

В таблице сопряженности могут быть представлены как абсолютные, так и относительные частоты (по отдельности или одновременно).

Gender * Employment Category Crosstabulation

% within Gender

		Employment Category			
		Clerical	Custodial	Manager	Total
Gender	Female	95.4%		4.6%	100.0%
	Male	60.9%	10.5%	28.7%	100.0%
Total		76.6%	5.7%	17.7%	100.0%

Gender * Employment Category Crosstabulation

% within Employment Category

		Employment Category			
		Clerical	Custodial	Manager	Total
Gender	Female	56.7%		11.9%	45.6%
	Male	43.3%	100.0%	88.1%	54.4%
Total		100.0%	100.0%	100.0%	100.0%

Gender * Employment Category Crosstabulation

% of Total

		Employment Category			
		Clerical	Custodial	Manager	Total
Gender	Female	43.5%		2.1%	45.6%
	Male	33.1%	5.7%	15.6%	54.4%
Total		76.6%	5.7%	17.7%	100.0%

Таблица сопряженности может быть построена для дискретных переменных (номинальных, порядковых, количественных), а также для непрерывных переменных, сгруппированных в интервалы.

Следует избегать ситуаций, когда частоты в клетках таблицы слишком малы (за исключением тех случаев, когда отдельные категории объектов отсутствуют в принципе – например, женщины-охранники). Для повышения частот в клетках значения переменных рекомендуется группировать. В некоторых случаях, для проверки специфических гипотез о связи между переменными, из таблицы могут быть удалены целые строки или столбцы.

Gender * Employment Category Crosstabulation

			Employment Category		
			Clerical	Manager	Total
Gender	Female	Count	206	10	216
		% within Gender	95.4%	4.6%	100.0%
	Male	Count	157	74	231
		% within Gender	68.0%	32.0%	100.0%
Total		Count	363	84	447
		% within Gender	81.2%	18.8%	100.0%

При исследовании связи между полом и занимаемой должностью можно опустить должность охранника (custodial), на которой женщины не работают.

Построение таблицы сопряженности

Analyze | **Descriptive Statistics** | **Crosstabs** | поместить имена переменных, образующих строки и столбцы, в окна **Row(s)** и **Column(s)** | кнопка **Cells** | задать вид вычисляемых частот | **Continue** | кнопка **Statistics** | задать вычисляемые статистики и коэффициенты | **Continue** | **OK**

Опции параметра **Cells** функции **Crosstabs:**

Counts observed	Абсолютные частоты
Percentages Row	Проценты по строке
Percentages Column	Проценты по столбцу
Percentages Total	Проценты от объема выборки

Проверка гипотезы о значимости статистической связи между строками и столбцами таблицы сопряженности

$H_0 : f_{ij} = e_{ij}$

$H_1 : f_{ij} \neq e_{ij}$

осуществляется по критерию Хи-квадрат (опция **Chi-square** параметра **Statistics**).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.277[a]	2	.000
Likelihood Ratio	95.463	2	.000
N of Valid Cases	474		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.30.

Поскольку в листинге SPSS для критерия Хи-квадрат приводится двустороннее p-значение (*Asymp. Sig. 2-sided*), а при проверке гипотезы о статистической значимости связи используется односторонний критерий, напечатанное p-значение следует сравнивать с 2α: если $Sig. < 2\alpha$, принимается альтернативная гипотеза, согласно которой связь является статистически значимой при выбранном α (α может принимать значения 0.1, 0.05, 0.01 – см. тему 14, занятие практикума 8).

Теснота (степень, сила) связи между двумя переменными измеряется с помощью специальных коэффициентов (мер связи). Выбор мер связи зависит от типа используемых переменных.

Выбор мер связи

Тип переменных	**Меры связи**	**Опции параметра Statistics**
Обе переменные количественные	Коэффициент линейной корреляции Пирсона r	Correlations
Обе переменные порядковые или количественные	Коэффициент ранговой корреляции Смирмана r_s	Correlations
Обе переменные дихотомические	Коэффициент Ф	Phi and Cramer's V
Во всех остальных случаях	Коэффициент Крамера V	Phi and Cramer's V

Коэффициенты Пирсона и Спирмана измеряют направленные (прямые или обратные) связи между количественными и порядковыми переменными:

- связь является *прямой (положительной),* если значения двух переменных одновременно возрастают и убывают;
- связь является *обратной (отрицательной),* если при возрастании значений одной переменных значения второй переменной убывают, или наоборот.

EDUC_GR * QUALIF Crosstabulation

Count

		QUALIF			
		custodials	clericals	mangerials	Total
EDUC_GI	8.00	13	40		53
	12.00	13	182	1	196
	16.00	1	135	39	175
	17.00		5	15	20
	20.00		1	29	30
Total		27	363	84	474

Прямая связь между порядковыми переменными "образование" и "должность".

Коэффициент Ф измеряет направленные (прямые или обратные) связи между дихотомическими переменными:

- *прямая (положительная)* связь проявляется в том, что признаки чаще появляются или не появляются вместе, чем врозь;
- *обратная (отрицательная)* связь проявляется в том, что признаки чаще появляются врозь, чем вместе.

FEMAIL * MANAGER Crosstabulation

Count

		MANAGER		
		.00	1.00	Total
FEMAIL	.00	184	74	258
	1.00	206	10	216
Total		390	84	474

Обратная связь между дихотомическими переменными "пол" и "должность менеджера".

Коэффициент Крамера измеряет ненаправленные связи для таблиц сопряженности, в которых *хотя бы одна из переменных является номинальной*.

Свойства коэффициентов связи:

- при отсутствии статистической связи значение коэффициента равно нулю;
- коэффициенты ненаправленной связи принимают значения из интервала [0,1] – чем теснее связь, тем ближе значение коэффициента к 1;
- коэффициенты направленной связи принимают значения из интервала [-1,+1], при этом прямой связи соответствуют положительные значения коэффициентов, обратной – отрицательные; чем ближе значение коэффициента к +1 или к –1, тем теснее связь.

Вычисленные значения коэффициентов могут выводиться в окно результатов как вместе с таблицей сопряженности, так и без нее. В последнем случае вывод таблицы следует подавить (опция **Suppress tables**).

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.593	.027	16.001	.000[c]
Ordinal by Ordinal	Spearman Correlation	.600	.027	16.276	.000[c]
N of Valid Cases		474			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Ранговая корреляция между образованием и должностью

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.314	.000
	Cramer's V	.314	.000
N of Valid Cases		474	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Корреляция между женским полом и должностью менеджера

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.409	.000
	Cramer's V	.409	.000
N of Valid Cases		474	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Статистическая связь между полом и должностью

Проверка гипотезы о статистической значимости мер связи

Гипотеза о статистической значимости коэффициента связи

$H_0 : r = 0$ (связь отсутствует)

$H_1 : r \neq 0$ (связь имеет место)

может быть проверена при помощи p-значения (*Sig.*):

- если $Sig. < \alpha$, принимается H_1,
- если $Sig. > \alpha$, принимается H_0,

где α может принимать значения 0.1, 0.05, 0.01.

ЗАДАНИЕ 10. АНАЛИЗ ЛИНЕЙНЫХ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

1. Загрузите файл *World95.sav*.
2. Отберите подгруппу европейских стран (переменная *region*, группы стран 1 и 2).
3. Постройте гистограмму рассеяния для переменных *gdp_cap* (национальный доход на душу населения) и *babymort* (младенческая смертность); определить по ней наличие и характер связи между переменными.
4. Исследуйте корреляционную связь между переменными.
5. Постройте уравнение регрессии с зависимой переменной *babymort* и независимой переменной *gdp_cap*.
6. Проинтерпретируйте уравнение регрессии.
7. Отредактируйте и сохраните файл результатов *lesson10.spo.*
8. Подготовьте отчет о результатах; приведите в нем необходимые данные из файла *lesson10.spo.*

Необходимая подготовка:
- Темы 14, 15, 18 программы курса.

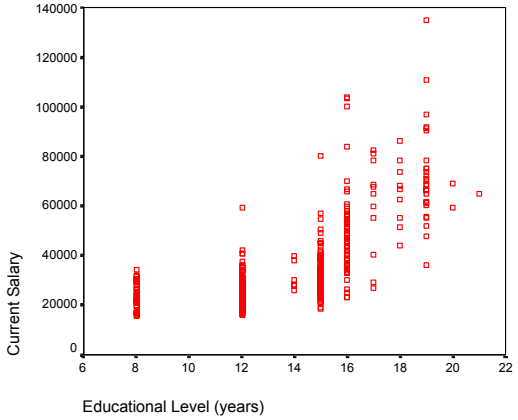
Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson10.spo* (на дискете);
- отчет об исследовании линейной связи (1–2 стр.).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 10

Для исследования связи между двумя количественными переменными наиболее часто используют линейную модель $y = bx + b_0$. Если между двумя переменными существует линейная связь, то при увеличении значения переменной x значение переменной y пропорционально увеличивается (*прямая, положительная связь*) или уменьшается (*обратная, отрицательная связь*).

Определить, существует ли связь между переменными и является ли она линейной, прямой или обратной, проще всего по диаграмме рассеяния.



Связь между образованием и зарплатой

Построение диаграммы рассеяния

Graphs | **Scatter…** | **Simple** | **Define** | поместить имена двух переменных в окошки **X Axis** и **Y Axis** | **OK**

Линейная связь является *полной*, если все точки диаграммы рассеяния лежат на прямой $y = bx + b_0$; *сильной* или *тесной*, если облако точек достаточно прилегает к прямой достаточно близко; *слабой*, если облако точек по отношению к прямой $y = bx + b_0$ широко разбросано.

Теснота (сила) линейной связи измеряется с помощью *коэффициента линейной корреляции Пирсона:*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Коэффициент Пирсона обладает следующими свойствами:

$-1 \le r \le +1$;

$r = 0$, если между переменными нет связи, или если связь не является линейной;

$r > 0$, если линейная связь является прямой (положительной);

$r < 0$, если линейная связь является обратной (отрицательной);

при этом чем ближе значение r к +1 или к –1, тем теснее связь;

$r = \pm 1$, если связь является полной.

Для сгруппированных переменных коэффициент Пирсона вычисляется по таблице сопряженности.

Вычисление коэффициента Пирсона

Analyze | **Correlate** | **Bivariate…** | поместить имена переменных в окно **Variables** | отметить флажками **Correlation coefficient: Pearson** | **Test of significance: Two-tailed** | **Flag significant correlations** | **OK**

Correlations

		Educational Level (years)	Current Salary
Educational Level (years)	Pearson Correlation	1.000	.661**
	Sig. (2-tailed)	.	.000
	N	474	474
Current Salary	Pearson Correlation	.661**	1.000
	Sig. (2-tailed)	.000	.
	N	474	474

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

Матрица корреляций для переменных "Образование" и "Зарплата"

Команда **Correlate** позволяет вычислить матрицу корреляций одновременно для нескольких переменных; для этого достаточно поместить их имена в окно **Variables**. Результаты представляются в виде ***матрицы корреляций***, в клетках которой указываются значения коэффициентов корреляции для соответствующих переменных. Матрица корреляций симметрична относительно главной диагонали, состоящей из единиц.

Для коэффициента корреляции r проверяется гипотеза о статистической значимости. Выборочное значение коэффициента является ***статистически значимым***, если по нему можно заключить, что значение коэффициента для генеральной совокупности будет отличаться от нуля:

$H_0 : r_{ген} = 0$

$H_1 : r_{ген} \neq 0$

Для проверки гипотезы выбранный уровень значимости ($\alpha = 0.1;0.05;0.01$) необходимо сравнить с напечатанным в клетке таблицы p-значением ***Sig. (two-tailed)***:

если $Sig < \alpha$, принимается H_1 (коэффициент корреляции статистически значим; между переменными существует линейная связь);

если $Sig > \alpha$, принимается H_0, наличие линейной статистической связи не подтверждается.

Термином *корреляция* обычно определяют связь между двумя переменными, не имеющую причинной компоненты. Если одна переменная (*независимая*) измеряет причину, а вторая (*зависимая*) – следствие, связь между переменными называется причинной и может быть описана посредством *уравнения линейной регрессии* $y = bx + x_0$. Теснота причинной линейной связи измеряется с помощью коэффициента линейной корреляции Пирсона. С помощью уравнения регрессии можно предсказать, каким будет среднее значение зависимой переменной y при определенном значении независимой переменной x.

Коэффициент b называется ***коэффициентом регрессии*** и вычисляется по формуле:

$$b = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Коэффициент регрессии показывает, насколько, в среднем, увеличится или уменьшится значение зависимой переменной y при увеличении значения независимой переменной x на 1. Знак коэффициента регрессии совпадает со знаком коэффициента корреляции; равенство значения коэффициента нулю свидетельствует об отсутствии линейной связи между переменными.

Коэффициент b_0 называется ***свободным членом уравнения регрессии*** и вычисляется по формуле $b_0 = \bar{y} - b\bar{x}$; во большинстве задач он не интерпретируется.

Вычисление уравнения регрессии

Analyze | **Regression** | **Linear…** | поместить имя зависимой переменной в окно **Dependent** | поместить имя независимой переменной в окно **Independent(s)** | **OK**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-18331.2	2821.912		-6.496	.000
	Educational Level (years)	3909.907	204.547	.661	19.115	.000

a. Dependent Variable: Current Salary

Уравнение регрессии: зарплата $= 3910 \times$ *образование* $- 3910$ *:*
(при повышении уровня образования на 1 год
годовая зарплата в среднем увеличивается на $3910)

Для параметров регрессии b и b_0 проверяются гипотезы о статистической значимости по тому же алгоритму, что и для коэффициента корреляции:

$H_0 : b_{ген} = 0$ $\qquad$ $H_0 : b_{0,ген} = 0$

$H_1 : b_{ген} \neq 0$ $\qquad$ $H_1 : b_{0,ген} \neq 0$

Качество уравнения парной регрессии, его объясняющая способность измеряется ***коэффициентом детерминации*** r^2. Коэффициент детерминации показывает, какая доля дисперсии (изменчивости) зависимой переменной y объясняется влиянием независимой переменной x.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661[a]	.436	.435	$12,833.54

a. Predictors: (Constant), Educational Level (years)

ЗАДАНИЕ 11. АНАЛИЗ НЕЛИНЕЙНЫХ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

1. Загрузите файл *World95.sav* из базы данных SPSS.
2. Отберите подгруппу европейских стран (переменная *region*, группы стран 1 и 2).
3. Постройте гистограмму рассеяния для переменных *gdp_cap* (национальный доход на душу населения) и *babymort* (младенческая

смертность); определить по ней наличие и характер связи между переменными.
4. Вычислите новую переменную – логарифм национального дохода *ln_gdp=ln(gdp_cap)*.
5. Постройте гистограмму рассеяния для переменных *ln_gdp* и *babymort*; сравните ее с предыдущей диаграммой.
6. Постройте уравнение регрессии с зависимой переменной *babymort* и независимой переменной *ln_gdp*.
7. Проинтерпретируйте уравнение регрессии.
8. Отредактируйте и сохраните файл результатов *lesson11.spo.*
9. Подготовьте отчет о результатах; приведите в нем необходимые данные из файла *lesson11.spo*.

Необходимая подготовка:
- Темы 14, 15, 18, 19 программы курса.
- Задание 10 практикума.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson11.spo* (на дискете);
- отчет об исследовании нелинейной связи (1–2 стр.).

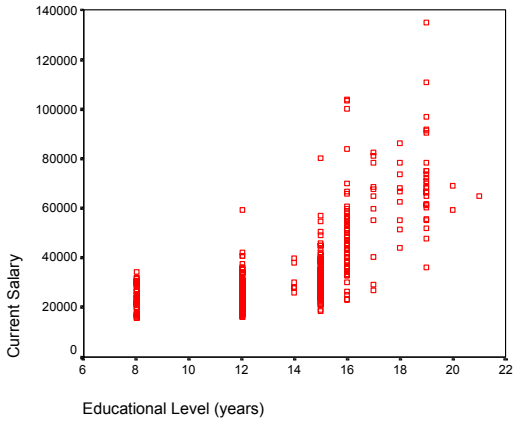
МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 11

Если между двумя количественными переменными наблюдается связь, которая не может быть достаточно хорошо описана уравнением линейной регрессии, можно попытаться построить для нее уравнение нелинейной регрессии. Наиболее часто в социальных науках используются логарифмические и экспоненциальные регрессионные модели.

Логарифмическая модель применяется, если облако точек на диаграмме рассеяния напоминает логарифмическую кривую. Она имеет вид $y = b\ln(x) + b_0$, где $\ln(x)$ – независимая переменная.

Эспоненциальная модель применяется, если облако точек на диаграмме рассеяния напоминает экспоненту. Она имеет вид $y = e^{bx+bo}$ или $\ln(y) = bx + b_0$, где $\ln(y)$ – зависимая переменная.

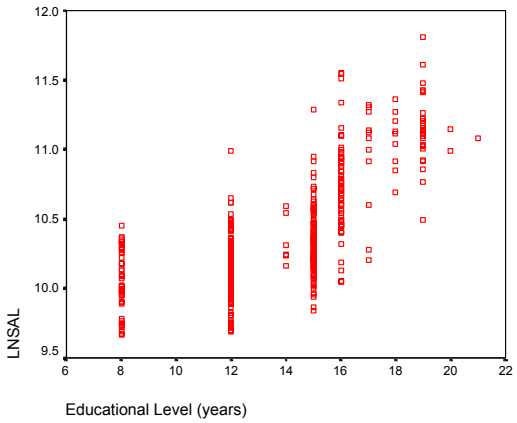
Пример (продолжение примера из задания 10):



Связь между образованием и зарплатой

Диаграмма рассеяния имеет форму экспоненты, поэтому для описания связи между образованием и заработной платой можно использовать экспоненциальную модель.

Вычислим новую зависимую переменную – логарифм заработной платы: $\ln sal = \ln(salary)$. Вид диаграммы рассеяния изменится: облако точек станет более "линейным".



Связь между образованием и логарифмом зарплаты.

Построим уравнение *линейной* регрессии для логарифма зарплаты (зависимая переменная) и образования (независимая переменная).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.062	.063		144.445	.000
	Educational Level (years)	9.596E-02	.005	.697	21.102	.000

a. Dependent Variable: LNSAL

Уравнение регрессии имеет вид $\ln(y) = 0.096x + 9.06$ или $y = e^{0.096x+9.06}$. Качество построенной модели измеряется с помощью коэффициента детерминации R^2.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.697[a]	.485	.484	.2853

a. Predictors: (Constant), Educational Level (years)

Сравнив полученную величину $R^2 = 0.485$ с величиной $R^2 = 0.436$ для линейной модели (см. методические материалы к заданию 10), мы убеждаемся, что объясняющая способность экспоненциальной модели на 5% выше, чем объясняющая способность линейной модели.

ЗАДАНИЕ 12. ОБРАБОТКА ДАННЫХ НАУЧНОГО ЭКСПЕРИМЕНТА (ДИСПЕРСИОННЫЙ АНАЛИЗ)

1. Из каталога рабочих файлов SPSS загрузите файл *employee.sav*.
2. Проверьте гипотезу о зависимости зарплаты от пола.
3. Проверьте гипотезы о зависимости зарплаты от пола (фактор А) и принадлежности к национальному меньшинству (фактор В), а также от их взаимодействия; поясните, что в данном случае означает наличие или отсутствие взаимодействия факторов.
4. Сохраните результаты в файле *lesson12.spo.*
5. Подготовьте отчет в Word, в который включите:
   - постановку задачи;
   - план эксперимента;
   - гипотезы (в содержательной и статистической формулировках);

- таблицу дисперсионного анализа;
- результаты проверки гипотез, их интерпретацию.

Необходимая подготовка:
- Темы 14, 20, 21 программы курса.

Методическое указание:
после выполнения каждого пункта задания сохраняйте файл данных и файл результатов в своей папке.

Форма отчетности:
- файл результатов *lesson12.spo* (на дискете);
- отчет о результатах факторного анализа (3–4 стр.).

МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ К ЗАДАНИЮ 12

Дисперсионный анализ предназначен для исследования причинных связей, в первую очередь – для обработки данных научного факторного эксперимента.

Зависимая переменная всегда бывает количественной. Независимые переменные (***факторы***) могут быть номинальными, порядковыми, количественными (сгруппированными в интервалы).

Суть дисперсионного анализа заключается в проверке гипотезы о влиянии независимых переменных на зависимую в том смысле, что группы объектов, образованные значениями факторов, отличаются друг на от друга средними значениями зависимой переменной.

Для проверки гипотезы используется модель *разделения дисперсии зависимой переменной*, согласно которой дисперсия, как показатель степени разброса и неоднородности данных, включает, как минимум, две составляющие, одна из которых порождается вариабельностью зависимой переменной внутри группы (*внутригрупповая дисперсия*), а вторая – различиями между группами (*межгрупповая дисперсия*). Межгрупповая дисперсия, зависящая от степени неоднородности групп, рассматривается как показатель степени влияния груп-

пообразующих факторов на зависимую переменную и, соответственно, является основным предметом дисперсионного анализа.

Вместо дисперсии $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 \Big/ (n-1)$ в дисперсионном анализе используется только ее числитель – *общая сумма квадратов* $MSS_{общ} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Однофакторный дисперсионный анализ
(одна независимая переменная).

Согласно модели однофакторного дисперсионного анализа, общую сумму квадратов можно разделить на две составляющие: внутригрупповую $(MSS_{внр})$ и межгрупповую $(MSS_{мгр})$ суммы квадратов: $MSS_{общ} = MSS_{вгр} + MSS_{мгр}$. Внутригрупповая сумма квадратов является мерой рассеяния зависимой переменной внутри групп, выделенных соответственно значениям фактора; межгрупповая интерпретируется как часть общей суммы квадратов, обусловленная различиям между группами, т.е. влиянием фактора.

Гипотеза однофакторного анализа.
Пусть фактор имеет k значений (образует k групп).
$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ (во всех группах средние значения зависимой переменной равны);
$H_1 : \mu_i \neq \mu$ (хотя бы для некоторых групп средние значения не равны).

Для проверки гипотезы используется p-значение $(Sig.)$ F-критерия Фишера:

- если $Sig > \alpha$, принимается гипотеза H_0 об отсутствии влияния фактора на зависимую переменную;
- если $Sig < \alpha$, принимается гипотеза H_1 о том, что такое влияние существует.

Значение α, как обычно, выбирается из чисел 0.1, 0.05, 0.01.

Пример.

Гипотеза: средняя продолжительность рабочей недели зависит от уровня образования.

Зависимая переменная – продолжительность рабочей недели *Number of Hours*; независимая переменная (фактор) – уровень образования *Degree* (5 ступеней).

Descriptive Statistics

Dependent Variable: Number of Hours Worked Last Week

RS Highest Degree	Mean	Std. Deviation	N
Less than HS	43.69	8.72	52
High school	45.77	10.58	387
Junior college	45.87	11.66	54
Bachelor	46.38	12.89	162
Graduate	50.27	11.44	86
Total	46.29	11.27	741

Выделен план эксперимента

Первые два столбца таблицы (названия групп и соответствующие им средние значения зависимой переменной) образуют *план эксперимента*.

Основные результаты однофакторного дисперсионного анализа принято представлять в виде специальной таблицы, в которой указывается межгрупповая сумма квадратов (*Model*), в однофакторном анализе она порождается единственной независимой переменной – (*DEGREE*); внутригрупповая сумма квадратов (*Error*); общая сумма квадратов (*Total*).

Tests of Between-Subjects Effects

Dependent Variable: Number of Hours Worked Last Week

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	1589531.720[a]	5	317906.344	2539.158	.000
DEGREE	1589531.720	5	317906.344	2539.158	.000
Error	92148.280	736	125.201		
Total	1681680.000	741			

a. R Squared = .945 (Adjusted R Squared = .945)

Для независимой переменной *DEGREE* $Sig < 0.01$, следовательно, принимается гипотеза H_1, согласно которой продолжительность рабочей недели зависит от уровня образования, при уровне значимости $\alpha = 0.01$.

Недостатком дисперсионного анализа является то, что он не дает представления о характере связи между зависимой и независимой переменными. О нем можно судить только по плану эксперимента.

Двухфакторный дисперсионный анализ
(две независимые переменные).

В двухфакторном анализе рассматриваются три вида группировок зависимой переменной: по первому фактору (который принято называть *фактором A*), по второму фактору (*фактор B*), и перекрестная группировка по двум факторам.

Согласно модели двухфакторного дисперсионного анализа межгрупповая сумма квадратов $(MSS_{мгр})$ делится на сумму квадратов, порожденную влиянием первого фактора (MSS_A), сумму квадратов, порожденную влиянием второго фактора (MSS_B) и сумму квадратов, порожденную эффектом взаимодействия двух факторов (MSS_{AB}): $MSS_{мгр} = MSS_A + MSS_B + MSS_{AB}$.

Таким образом, модель дисперсионного двухфакторного анализа имеет вид: $MSS_{общ} = MSS_{вгр} + MSS_A + MSS_B + MSS_{AB}$.

Гипотезы двухфакторного дисперсионного анализа.
Для двухфакторного дисперсионного анализа проверяются гипотезы трех видов.

Гипотеза о влиянии на зависимую переменную фактора A (об эффекте фактора А):

$H_0 : \mu_1. = \mu_2. = ... = \mu_k. = \mu$ (во всех k группах, образованных фактором А, средние значения зависимой переменной равны);

$H_1 : \mu_i. \neq \mu$ (хотя бы для некоторых групп средние значения не равны).

Гипотеза о влиянии на зависимую переменную фактора B (об эффекте фактора B):

$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = ... = \mu_{\cdot l} = \mu$ (во всех l группах, образованных фактором B, средние значения зависимой переменной равны);

$H_1 : \mu_{\cdot j} \neq \mu$ (хотя бы для некоторых групп средние значения не равны).

Гипотеза об эффекте взаимодействия факторов A и B:

$H_0 : \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu = 0$ для любых $i = \overline{1,k}$; $j = \overline{1,l}$ (эффект взаимодействия отсутствует);

$H_1 : \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu \neq 0$ хотя бы для некоторых i и j (эффект взаимодействия имеет место).

Каждая гипотеза проверяется отдельно, аналогично тому, как проверяется гипотеза однофакторного дисперсионного анализа.

Пример.

Гипотезы: на продолжительность рабочей недели влияют образование, пол, а также имеет место эффект их взаимодействия.

Зависимая переменная – продолжительность рабочей недели (*Number of Hours*); зависимые переменные – пол (*Sex*), образование (*Degree*).

Descriptive Statistics

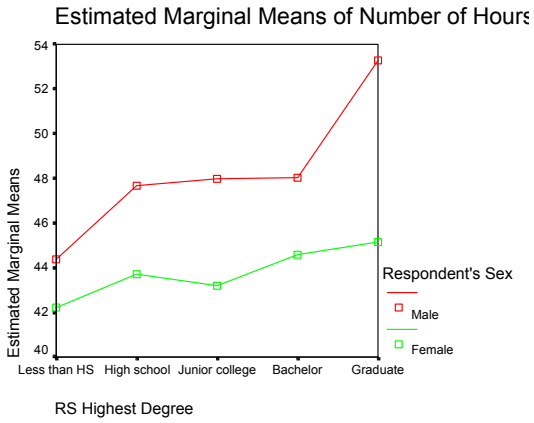
Dependent Variable: Number of Hours Worked Last Week

RS Highest Degree	Respondent's Sex	Mean	Std. Deviation	N
Less than HS	Male	44.36	7.59	36
	Female	42.19	11.00	16
	Total	43.69	8.72	52
High school	Male	47.69	10.90	201
	Female	43.70	9.83	186
	Total	45.77	10.58	387
Junior college	Male	48.00	11.07	30
	Female	43.21	12.06	24
	Total	45.87	11.66	54
Bachelor	Male	48.05	12.15	84
	Female	44.59	13.50	78
	Total	46.38	12.89	162
Graduate	Male	53.30	12.05	54
	Female	45.16	8.22	32
	Total	50.27	11.44	86
Total	Male	48.24	11.26	405
	Female	43.94	10.83	336
	Total	46.29	11.27	741

Следует заметить, что SPSS довольно нетрадиционно представляет план двухфакторного эксперимента. В классическом варианте он выглядел бы так:

ПОЛ	УРОВЕНЬ ОБРАЗОВАНИЯ					
	Less than HS	High School	Junior college	Bachelor	Graduate	Total
Male	44.36	47.69	48.00	48.05	53.30	48.24
Female	42.19	43.70	43.21	44.59	45.16	43.94
Total	43.69	45.77	45.87	46.38	50.27	46.29

Во многих случаях план эксперимента можно представить графически, что облегчает понимание исследуемых зависимостей. На следующем рисунке представлены изменения средней продолжительности рабочей недели в зависимости от уровня образования, отдельно для мужчин и женщин.



Результаты двухфакторного дисперсионного анализа принято представлять в виде таблицы, в которой указывается межгрупповая сумма квадратов (*Model*); ее составляющие – сумма квадратов, связанная с влиянием фактора A (*DEGREE*), сумма квадратов, связанная с эффектом фактора B (*SEX*) и сумма квадратов, порожденная эффектом взаимодействия (*DEGREE*SEX*); внутригрупповая сумма квадратов (*Error*); общая сумма квадратов (*Total*).

Tests of Between-Subjects Effects

Dependent Variable: Number of Hours Worked Last Week

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Model	1593245.1[a]	10	159324.512	1316.972	.000
DEGREE	1589531.7	5	317906.344	2627.804	.000
SEX	3326.066	1	3326.066	27.493	.000
DEGREE * SEX	387.337	4	96.834	.800	.525
Error	88434.876	731	120.978		
Total	1681680.0	741			

a. R Squared = .947 (Adjusted R Squared = .947)

Для факторов A и B принимаются альтернативные гипотезы о наличии их эффектов на зависимую переменную, при уровне значимости $\alpha = 0.01$.

Для эффекта взаимодействия принимается нулевая гипотеза (эффект взаимодействия отсутствует, о чем свидетельствует, в том числе, и одинаковая для обоих полов тенденция повышения продолжительности рабочей недели с ростом образования).

Выполнение дисперсионного анализа в SPSS

Analyze | **General Linear Model** | **Univariate** | поместить имя зависимой переменной в окно **Dependent Variable** | поместить имена независимых переменных в окно **Fixed Factor(s)** | в окне **Model** отметить пункт **Full factorial** | в окне **Sum of squares** указать **Type I** | отменить назначение **Include Intercept in Model** | в окне **Options** можно указать **Descriptive Statistics** (для получения плана эксперимента) | в окне **Plots** можно задать вид графика.

АНГЛО-РУССКИЙ СЛОВАРЬ КОМПЬЮТЕРНЫХ И СТАТИСТИЧЕСКИХ ТЕРМИНОВ

A

Active window	Активное окно
Analysis	Анализ
data analysis	анализ данных
univariate analysis	одномерный анализ
Association measuring	Измерение связей

B

Bar chart	Диаграмма столбцов

C

Case	Случай, реализация
Categorical data	Категориальные данные
numeric coding	цифровое кодирование
Cell	Ячейка (таблицы)
Charts	Графики
bar chart	диаграмма столбцов
changing style	изменение стиля
cluster chart	кластерная диаграмма
creating a chart	построение графика
editing	редактирование
line chart	полигон распределения
Chi-square Test	Критерий Хи-квадрат
Clustered Bar	Кластерная диаграмма
Column format	Формат столбца
Contingency coefficient	Коэффициент контингенции
Contingency table	Таблица сопряженности
Copying	Копирование
between Windows' applications	между приложениями Windows
Correlation	Корреляция
Cramer's V	Коэффициент Крамера
Creating variables	Создание переменных
column format	формат столбца
labels	метки
name	имя
type	тип
Crosstable	Таблица сопряженности

D

3-D effect	3-мерный эффект
Data	Данные
categorical	категориальные
categorising data	категоризация данных
coding	кодировка
import/export	импорт/экспорт
level	уровень
multiple response	множественный выбор (отклик)
nominal	номинальные
numeric	количественные
ordinal	порядковые
recoding	перекодировка
Data base	База данных
Data editor	Редактор данных
Data matrix	Матрица данных
Data sheet	Таблица данных
display labels	вывод меток
editing data	редактирование данных
editing structure	редактирование матрицы данных
entering data	ввод данных
grid lines	разделительные линии
opening files	открытие файла
ordering data	упорядочение данных
printing	печать
Descriptive statistics	Описательная статистика
Designated window	Назначенное окно

E

Editing data	Редактирование данных
Editor window	Окно редактора
Entering data	Ввод данных
Exporting files	Экспорт файлов

F

File	Файл
data file	файл данных
opening files	открытие файла
saving files	сохранение файла
saving as text	сохранение в формате текста
Frequency distribution	Распределение частот

G

Grid lines	Разделительные линии

H

Help system	Справочная система
topics	полная справочная система
tutorial	обучающая программа
statistics coach	статистическое руководство
Hypothesis testing	Проверка гипотез

I

Importing files	Импорт файлов
Inferential statistics	Статистический вывод

L

Line chart	Полигон
Listing file	Файл листинга

M

Mean	Среднее арифметическое
Measuring associations	Измерение связей
Menu	Меню
Multiple response data	Данные с множественным выбором

N

Nominal data	Номинальные данные
Numeric data	Количественные данные

O

One-Sample T Test	Одновыборочный t-критерий
One-Way ANOVA	Однофакторный дисперсионный анализ
Ordinal data	Порядковые данные
Output window	Окно результатов
designated window	назначенное окно
changing designated window	изменение назначенного окна
opening window	открытие окна
saving window	сохранение окна

P

Pearson's R	Коэффициент корреляции Пирсона
Population	Генеральная совокупность
Printing	Печать
results of analysis	результатов анализа
charts	графиков
data sheet	матрицы данных

Q

Questionnaire	Анкета
creating scales	создание шкал
open question	открытый вопрос
producing ranked data	получение ранжированных данных
questions	вопросы

R

Recoding data	Перекодировка данных
Reserved words	Зарезервированные слова

S

Sample	Выборка
Saving files	Сохранение файлов
Spearman correlation	Корреляция Спирмена
Spreadsheet	Электронная таблица
SPSS	SPSS
menu	меню
toolbar	панель инструментов
windows	окна
Standard deviation	Стандартное (среднее квадратическое отклонение)
Statistics	Статистика
associaion measuremetn	измерение связи
Chi-Square test	критерий Хи-квадрат
Cramer's V	коэффициент Крамера
crosstable	таблица сопряженности
descriptive	описательная
frequency	частота
inferential	статистический вывод
Pearson's correlation	корреляция Пирсона
Spearman's rank correlation	ранговая корреляция Спирмана
Student t-test	t-критерий Стьюдента

T

Text files	Текстовый файл
Toolbar	Панель инструментов
Tutorial	Обучающая программа

V

Variable	Переменная
deleting	удаление
labels	метки
length	длина
name	имя
type	тип
value	значение
setting variables	задание переменных
Variance	Дисперсия

W

Window	Окно
active	активное
chart	графиков
chart editor	редактора графиков
designated	назначенное
editor	редактора
output	вывода результатов
syntax	команд

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

1. *Афанасьев В.И.* Методические указания по курсу математической статистики с применением пакета SPSS. М., 1996.
2. *Пациорковский В.В., Петрова А.И., Пациорковская В.В.* Использование SPSS в социологии. М., 1998.
3. SPSS для WINDOWS. Руководство пользователя. Книга 1. Базовая система версии 6.1. Интерфейс. Разведочный анализ данных. М., 1995.
4. SPSS Base 7.5 для WINDOWS. Руководство пользователя. М., 1997.
5. SPSS Base 7.5 для WINDOWS. Руководство по применению. М., 1997.

СОДЕРЖАНИЕ