

ТЕМА 2

МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ, МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ И РАЗБРОСА

1. Расчет описательных статистик в SPSS

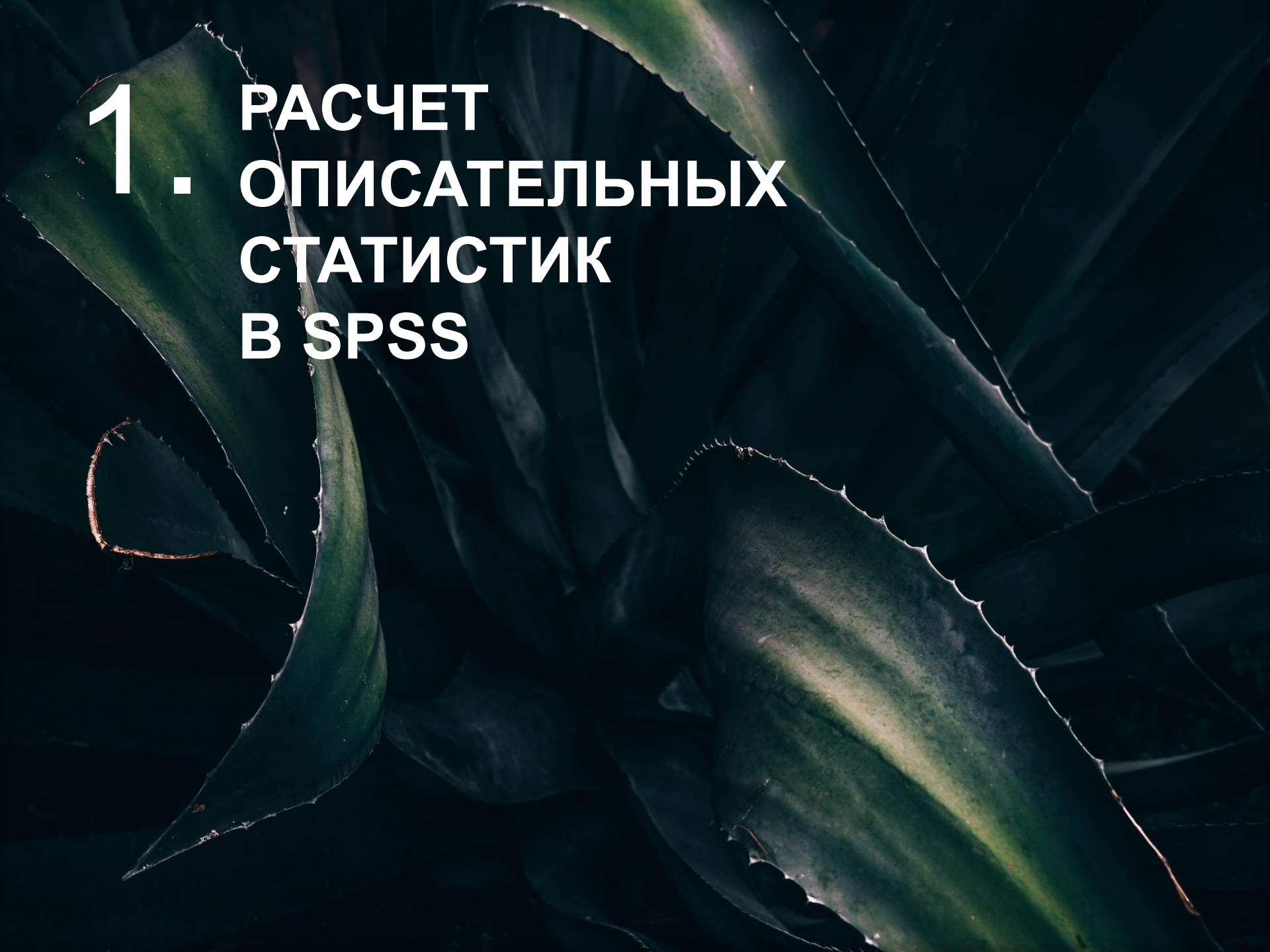
1.1. Описательные статистики для дискретных данных (для номинальных и порядковых шкал)

1.2. Описательные статистики для непрерывных данных (для интервальных шкал / шкал отношений)

2. Анализ множественных ответов

3. Формы представления статистических данных

4. Создание и редактирование графиков и диаграмм

The background of the slide is a close-up photograph of several aloe vera leaves. The leaves are a deep green color with prominent, lighter green longitudinal veins. The edges of the leaves are serrated with small, dark spines. The lighting is dramatic, with strong highlights and deep shadows, creating a textured and organic feel.

1. РАСЧЕТ ОПИСАТЕЛЬНЫХ СТАТИСТИК В SPSS

1. Расчет описательных статистик в SPSS

Описательные статистики (Descriptive Statistics) - это основные статистические параметры, которыми можно описать имеющееся распределение данных, если оно носит характер близкий к нормальному распределению.

Из лекции по Теме 1 мы узнали, что все данные условно делятся на два больших класса:

Дискретные

(отдельные значения признака, общее число которых конечно)



Номинальные

Пол респондента (GENDER):

- 1 – Мужской
- 2 – Женский

Порядковые

Возрастная группа:

- 1 – «18-24 года»
- 2 – «25-34 года»
- 3 – «35-44 года»

Непрерывные

(могут принимать любое значение в некотором интервале)



Интервальные (числовые)

Доход работника (INCOME):

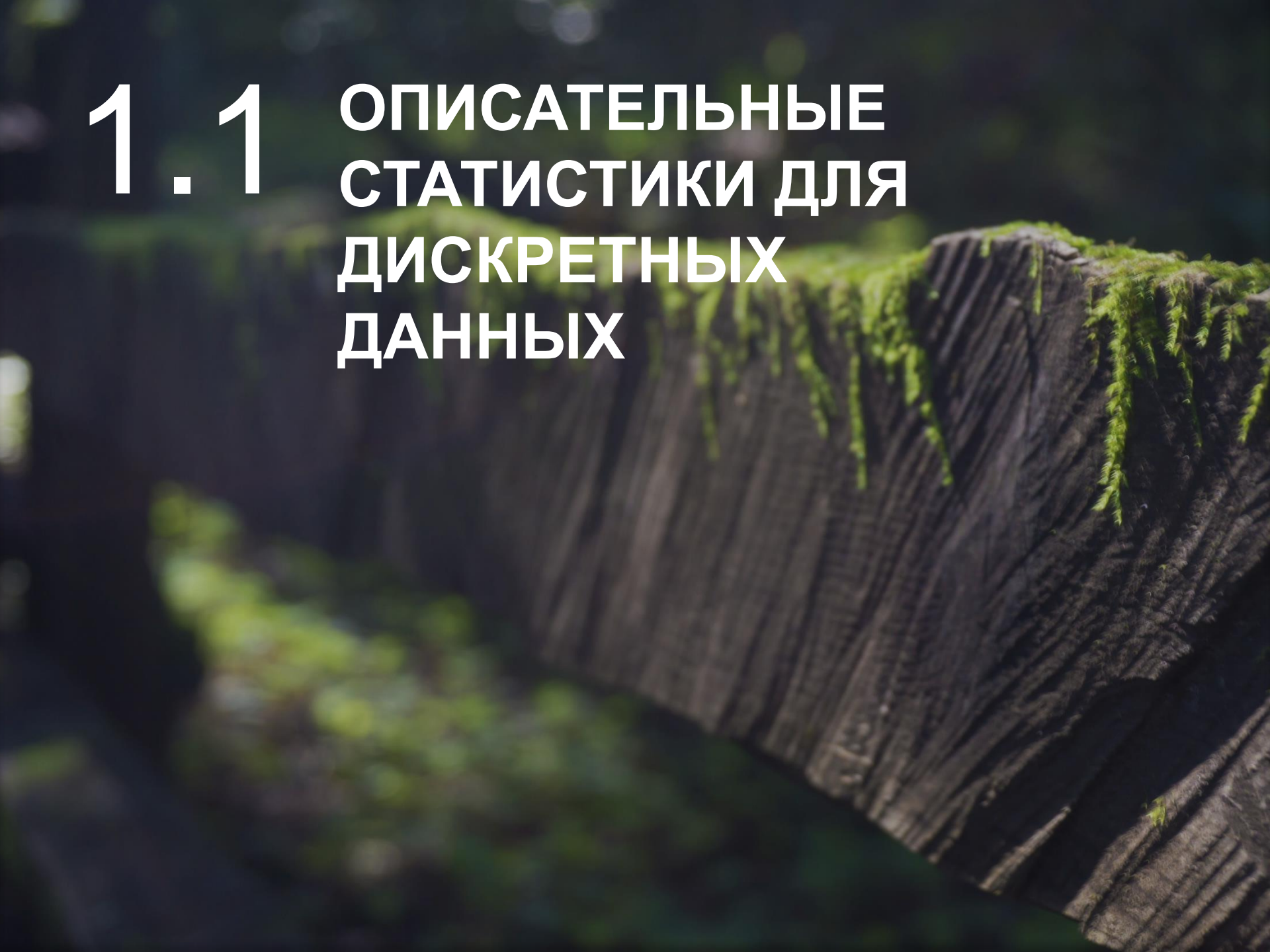
100\$.....100 000\$

Возраст, лет:

18 лет 78 лет

Рассмотрим, как вычисляются описательные статистики для двух типов данных.

1.1 ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ ДИСКРЕТНЫХ ДАННЫХ



1.1. Описательные статистики для дискретных данных

Для анализа дискретных данных проводят **частотный анализ**.

- **Частота (Frequency)** – количество наблюдений, в которых признак принимает определенное значение или находится в определенном интервале.
- **Распределение частот (Frequency Distribution)** показывает частоты во взаимосвязи с результатами наблюдений.

Пример: Узнать распределение частот по переменной Уровень образования (EDUCAT) (job.sav)

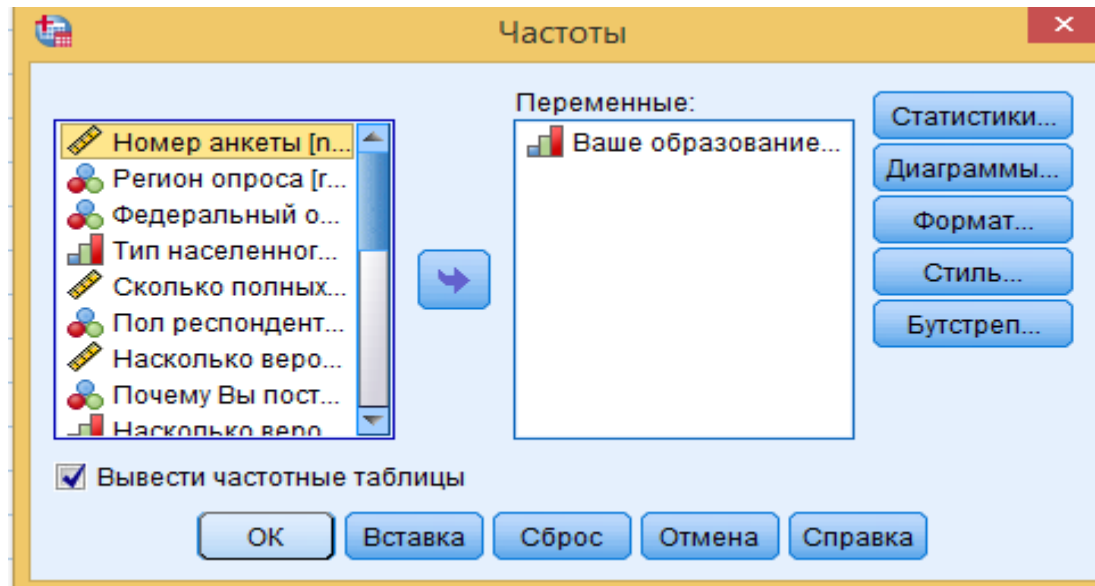
Респондент (num_ank)	Уровень образования (EDUCAT)
1	Среднее
2	Среднее специальное
3	Высшее
4	Среднее
5	Высшее
...	...
500	Среднее специальное



Уровень образования	Доля респондентов
Неполное среднее	1%
Среднее	15%
Среднее специальное	49%
Неоконченное высшее	5%
Высшее	30%

1.1. Описательные статистики для дискретных данных

1. Сначала загрузите файл job.sav, выбрав команды меню File (Файл) / Open... (Открыть...).
2. Выберите в меню команды Analyze (Анализ) / Descriptive Statistics (Описательные статистики) / Frequencies (Частоты). Появится диалоговое окно Частоты (Frequencies).
3. Кнопкой со стрелкой перенесите переменную EDUCAT в список выходных переменных и подтвердите операцию кнопкой ОК.



1.1. Описательные статистики для дискретных данных

1. Каждая строка частотной таблицы описывает одно возможное значение.
2. Первый столбец содержит метки отдельных значений (уровень образования).
3. Во втором столбце под заголовком «Частота» приведена частота каждого из вариантов уровня образования.

Пример: 3 респондента имеют «Неполное среднее образование или ниже», а большее число респондентов 246 закончили «Средне специальные учебные заведения».

Ваше образование?					
		Частота	Проценты	Процент допустимых	Накопленный процент
Допустимо	Неполное среднее образование или ниже	3	,6	,6	,6
	Среднее образование (школа или ПТУ)	76	15,2	15,2	15,8
	Среднее специальное образование (техникум)	246	49,2	49,2	65,0
	Незаконченное высшее (с 3-го курса ВУЗа)	25	5,0	5,0	70,0
	Высшее образование	150	30,0	30,0	100,0
	Всего	500	100,0	100,0	

1.1. Описательные статистики для дискретных данных

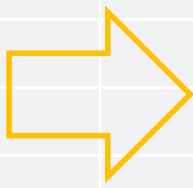
4. В третьем столбце показана процентная частота каждого варианта образования.
5. В четвертом столбце дано допустимое процентное значение (исключены потерянные данные).
6. Последний столбец «Накопленный процент» содержит сумму процентных частот.

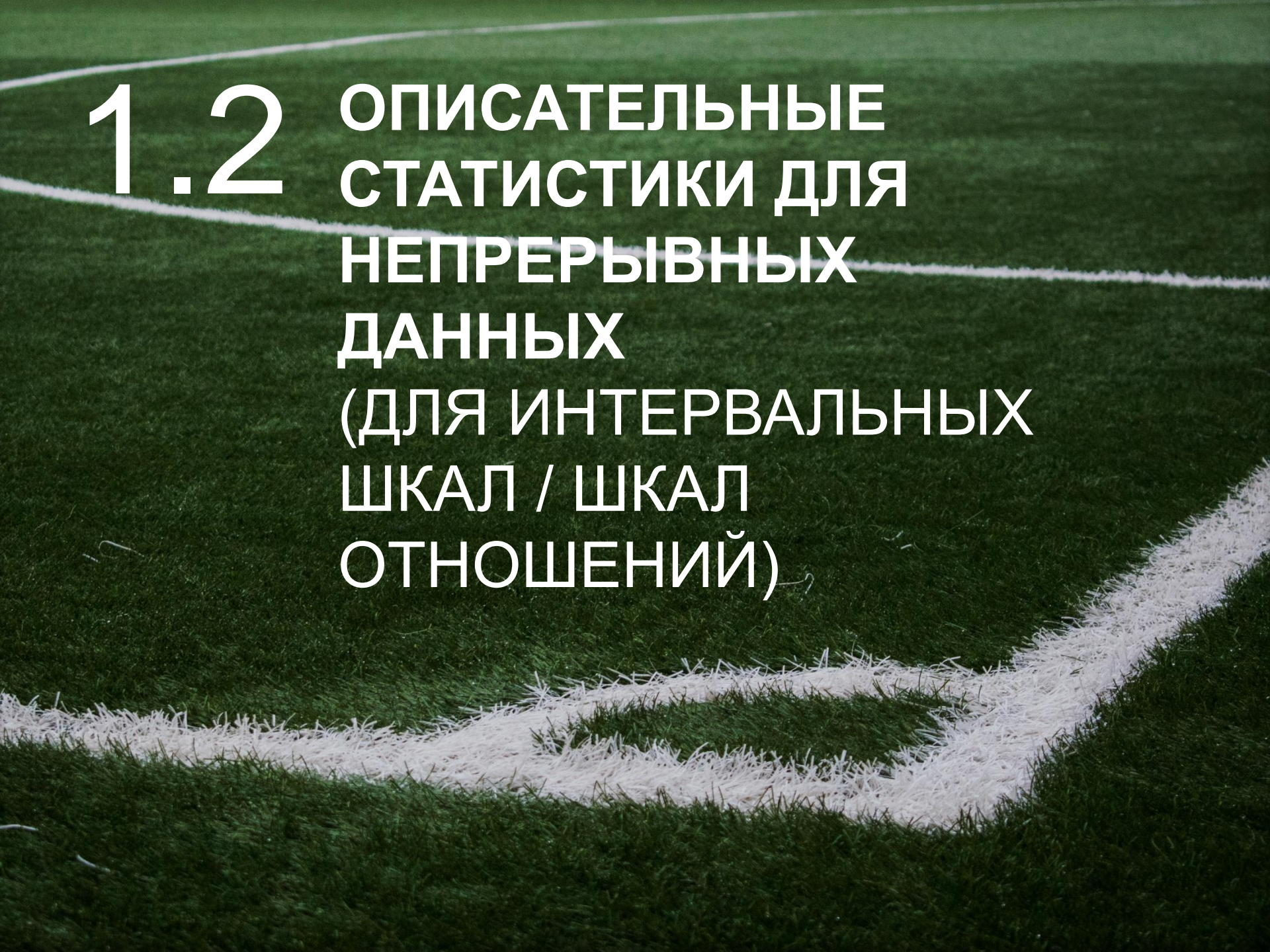
Ваше образование?					
		Частота	Проценты	Процент допустимых	Накопленный процент
Допустимо	Неполное среднее образование или ниже	3	,6	,6	,6
	Среднее образование (школа или ПТУ)	76	15,2	15,2	15,8
	Среднее специальное образование (техникум)	246	49,2	49,2	65,0
	Незаконченное высшее (с 3-го курса ВУЗа)	25	5,0	5,0	70,0
	Высшее образование	150	30,0	30,0	100,0
	Всего	500	100,0	100,0	

1.1. Описательные статистики для дискретных данных

- Непрерывные данные можно перекодировать в дискретные, объединяя их в интервалы.
- Например, часто в анкете просят указать возраст респондента числом (получают непрерывную переменную), а затем перекодируют ее в порядковую шкалу, объединяя респондентов в возрастные группы.

Респондент (num_ank)	Возраст, лет	Возрастная группа
1	20	1 = «от 18 до 24 лет»
2	19	1 = «от 18 до 24 лет»
3	31	2 = «от 25 до 34 лет»
4	18	1 = «от 18 до 24 лет»
5	37	3 = «от 35 до 44 лет»
...	
500	62	4 = «от 45 лет и старше»





1.2 ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ НЕПРЕРЫВНЫХ ДАННЫХ (ДЛЯ ИНТЕРВАЛЬНЫХ ШКАЛ / ШКАЛ ОТНОШЕНИЙ)

1.2. Описательные статистики для непрерывных данных

ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ ДЛЯ НЕПРЕРЫВНЫХ ДАННЫХ УСЛОВНО
МОЖНО РАЗБИТЬ НА НЕСКОЛЬКО ГРУПП:



1. МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

- Среднее
- Мода
- Медиана



2. МЕРЫ ИЗМЕНЧИВОСТИ

- Дисперсия
- Средне-
квадратическое
(стандартное)
отклонение
- Стандартная ошибка
- Размах



3. МЕРЫ ОТКЛОНЕНИЯ ФОРМЫ РАСПРЕДЕЛЕНИЯ

- Асимметрия
- Эксцесс

Все это различные вычисляемые статистические показатели, характеризующие распределение значений переменной.

1.2. Описательные статистики для непрерывных данных

1. Меры центральной тенденции

Измерение центральной тенденции (Measure of central tendency) состоит в выборе одного числа, которое наилучшим образом описывает все значения признака из набора данных.

Такое число называют **центром**, типическим значением для набора данных.

- **Среднее значение (Mean)** – сумма всех значений, отнесенная к общему числу наблюдений (очень чувствительна к выбросам).
- **Мода (Mode)** – наиболее часто встречающееся значение переменной.
- **Медиана (Median)** – среднее по порядку значение.

1.2. Описательные статистики для непрерывных данных

2. Меры изменчивости (Measure of Variation)

Выделяют две величины, характеризующие изменчивость (разброс), значений распределения относительно среднего:

Дисперсия (Variance) - равна сумме квадратов отклонений каждого значения от среднего, деленной на N , где N - число значений в распределении.

Стандартное отклонение (Standard Deviation) - равно квадратному корню из дисперсии.

Дополнительными мерами изменчивости являются 4 простые характеристики, отражающие границы распределения и его размах:

- Размах
- Квартильный размах
- Коэффициент вариации
- Стандартная ошибка среднего

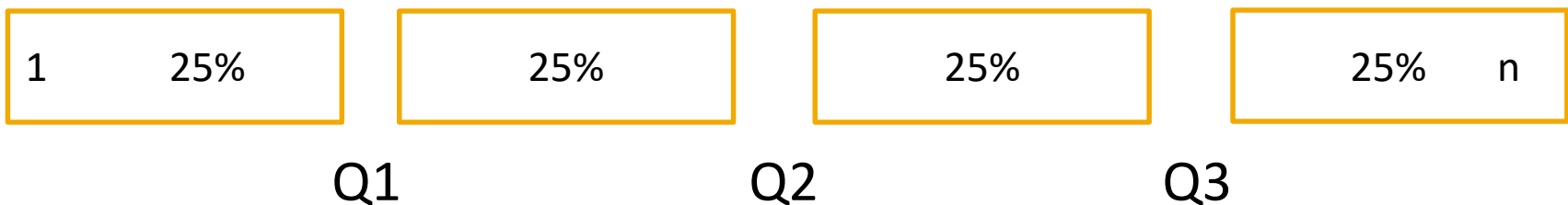
1.2. Описательные статистики для непрерывных данных

Размах – разность между наибольшим значением набора данных и наименьшим.

$$R = x_{\max} - x_{\min}$$

Пример: Для набора данных 29, 8, 4, 12, 10, 26, 6, 19 размах равен $R = 29 - 4 = 25$.

Квартили - значения, которые делят вариационный ряд на четыре равные части.



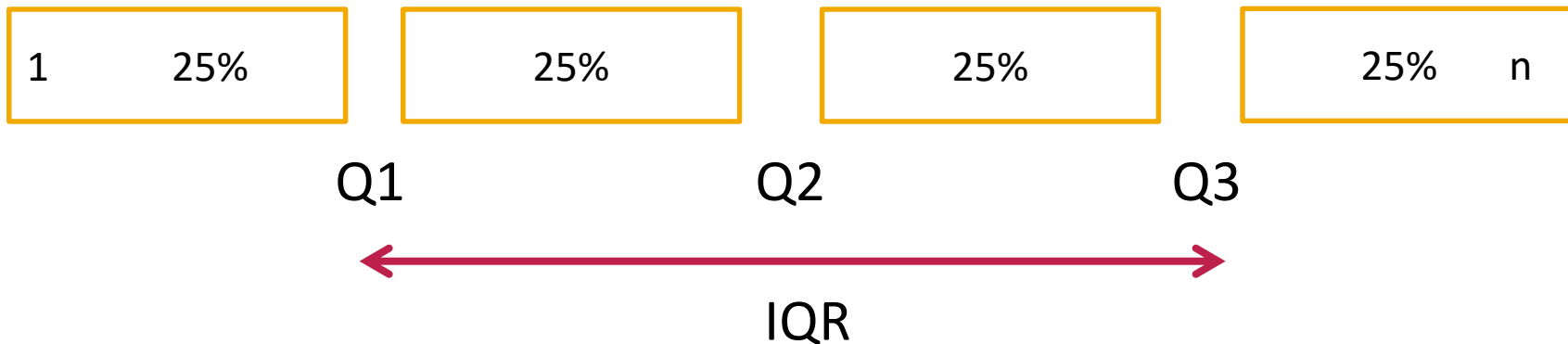
Ниже первого квартиля расположено 25% всех данных. Между первым и вторым квартилем также расположено 25% данных. Второй квартиль совпадает с медианой.

1.2. Описательные статистики для непрерывных данных

Размах квартилей – это разница между третьим и первым квартилем.

$$\text{IQR} = Q3 - Q1$$

Между Q1 и Q3 расположены 50% всех данных.



1.2. Описательные статистики для непрерывных данных

- **Коэффициент вариации** – отношение стандартного отклонения к среднему арифметическому, выраженное в %. Это относительная мера разброса значений признака.
- **Стандартная ошибка среднего (S.E. Mean)** – определяется как стандартное отклонение, деленное на квадратный корень из объема выборки. Является характеристикой точности или стабильности величины, для которой она вычисляется.

Для более детального анализа также можно вычислить две дополнительные характеристики меры изменчивости

Минимум (Minimum) –
равен наименьшему из
значений распределения.

Пример:
Для распределения [3 5 7 5 6 8 9]
минимум равен 3.

Максимум (Maximum) –
равен наибольшему из значений
распределения.

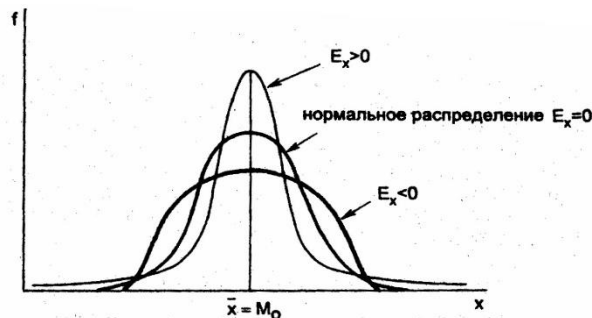
Пример:
Для распределения [3 5 7 5 6 8 9]
максимум равен 9.

1.2. Описательные статистики для непрерывных данных

3. Меры отклонения формы распределения

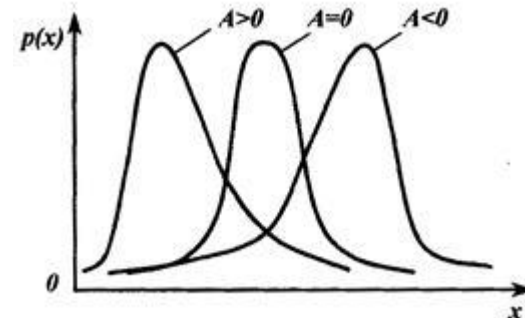
Экссесс E (Kurtosis) - является мерой «сглаженности» («остро-» или «плосковершинности») распределения.

- $E_x = 0$, значит распределение близко к нормальному
- $E_x < 0$ указывает на «плосковершинное» распределение
- $E_x > 0$ указывает на «островершинное» распределение
- Значения эксцесса, превышающие 5,0, говорят о том, что по краям распределения находится больше значений, чем вокруг среднего



Асимметрия A (Skewness) - показывает, в какую сторону относительно среднего сдвинуто большинство значений распределения.

- $A_x = 0$, значит распределение симметрично относительно среднего значения
- $A_x > 0$ указывает на сдвиг распределения в сторону меньших значений
- $A_x < 0$ указывает на сдвиг распределения в сторону больших значений
- $-1 < A_x < +1$ принимается за нормальное распределение



1.2. Описательные статистики для непрерывных данных

Для того, чтобы рассчитать описательные статистики используя статистический пакет SPSS, необходимо сделать следующие шаги:

1. Внести значения переменной, описательные статистики которой нам необходимо вычислить. Данные вносятся в таблицу «Редактор Данных».
2. Выбрать «**Анализ**» → «**Описательные статистики**» → «**Частоты**».
3. В появившемся окне «**Частоты**» перенести нужные переменные из левой части в правую.

1.2. Описательные статистики для непрерывных данных

4. Нажать кнопку «**Statistics**», и в появившемся окне выбрать **необходимые переменные** (отметить галочкой):

- Среднее арифметическое (Mean)
- Мода (Mode)
- Медиана (Median)
- Дисперсия (Variance)
- Стандартное отклонение (Std. deviation)
- Асимметрия (Skewness)
- Эксцесс (Kurtosis)

5. Нажать кнопку «**Продолжить**».

6. Нажать кнопку «**ОК**».

7. Интерпретировать результаты.

Частоты: Статистики

Значения процентилей

- ☐ Квартили
- ☐ Процентили для: 10 равных групп
- ☐ Процентили:
- Добавить
- Изменить
- Удалить

Положение центра распределения

- ☒ Среднее значение
- ☒ Медиана
- ☒ Мода
- ☐ Сумма
- ☐ Значения - центры групп

Разброс

- ☐ Стандартная Отклонение
- ☐ Дисперсия
- ☐ Диапазон
- ☐ Минимум
- ☐ Максимум
- ☐ Среднеквадратичная ошибка среднее

Распределение

- ☐ Асимметрия
- ☐ Эксцесс

Продолжить Отмена Справка

1.2. Описательные статистики для непрерывных данных

Пример: Вычислить среднее арифметическое, моду и медиану, дисперсию, стандартное отклонение, асимметрию и эксцесс по переменной AGE (возраст) из массива данных job.sav.

Статистика

Сколько полных лет Вам исполнилось?

N	Допустимо	500
	Пропущенные	0
Среднее значение		40,53
Медиана		39,00
Мода		35
Стандартная отклонения		12,151
Дисперсия		147,657
Асимметрия		,243
Стандартная Ошибка асимметрии		,109
Эксцесс		-,970
Стандартная ошибка эксцесса		,218

1.2. Описательные статистики для непрерывных данных

Интерпретация результатов

- Средний возраст 40 лет
- Медианное значение равно 39 годам
- Мода составила 35 лет, значит данное значение встречается наиболее часто
- Дисперсия равна 147,657
- Стандартное отклонение составило 12,151 лет
- Асимметрия является положительной, однако находится в пределах от -1 до 1, поэтому распределение по данной переменной можно считать нормальным
- Эксцесс отрицательный, но по модулю не превышает 1, поэтому и по этому параметру можно говорить о нормальности распределения

The background of the slide is a dense, close-up photograph of green leaves. The leaves are elongated and pointed, with prominent veins. They are arranged in a way that creates a textured, layered effect, with some leaves in the foreground being more in focus than others in the background. The overall color is a rich, vibrant green.

2. АНАЛИЗ МНОЖЕСТВЕННЫХ ОТВЕТОВ

2. Анализ множественных ответов

При анализе и кодировании множественных ответов (вопросы, на которые можно дать несколько ответов одновременно) используются два метода:

Метод множественной дихотомии

(для каждой из возможностей ответа определяется отдельная переменная)

Метод множественных категорий

(должно быть известно максимальное количество возможных ответов)

2. Анализ множественных ответов

Метод множественной дихотомии

Пример: В массиве данных job.sav был вопрос: «Какие из перечисленных проблем в наибольшей степени волнуют Вас на текущем месте работы?» Варианты ответов:

1. Задержки выплат заработной платы
2. Сокращение персонала/угроза увольнения и т.д.

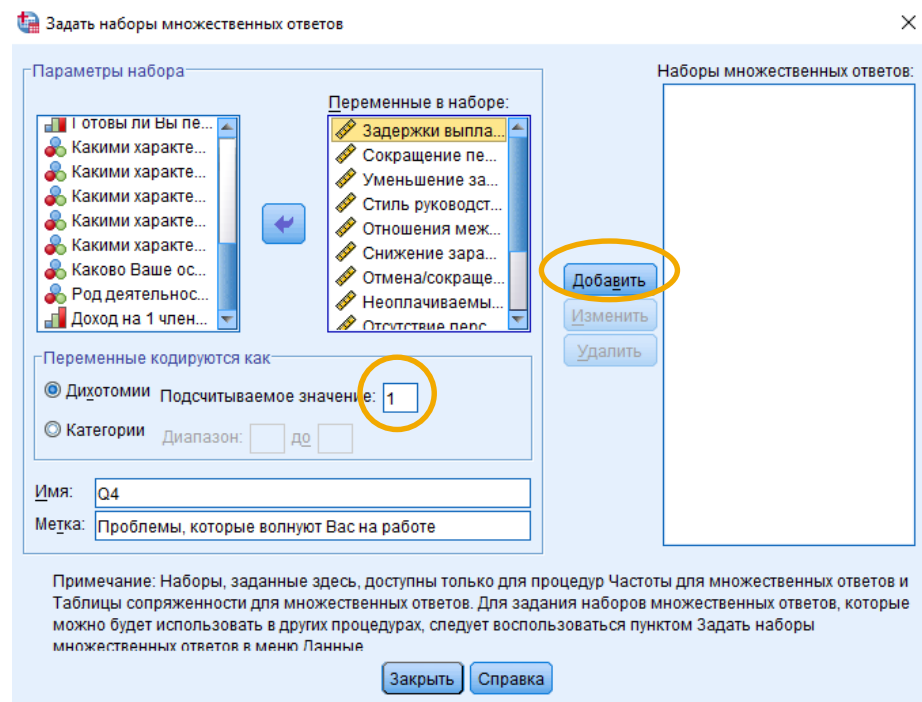
Кодирование переменных:

Если респондент ответит «задержки выплат з/п», то эта переменная примет значение «1», иначе «0», если респондент ответит «сокращение персонала/угроза увольнения», то эта переменная – «1», иначе «0» и т.д.

Определение наборов в SPSS:

«Анализ» → «Множественные ответы» → «Определить наборы»

- Выбрать переменные qq4_1-qq4_99
- Задать дихотомическую кодировку
- Выбрать подсчитываемое значение=1



2. Анализ множественных ответов

Метод множественных категорий

Пример: В массиве данных job.sav был вопрос: «Какими характеристиками должен обладать, с Вашей точки зрения, идеальный работодатель?» Варианты ответов:

1. Возможность карьерного роста
2. Высокая заработная плата, её связь с личным вкладом в работу и т.д.

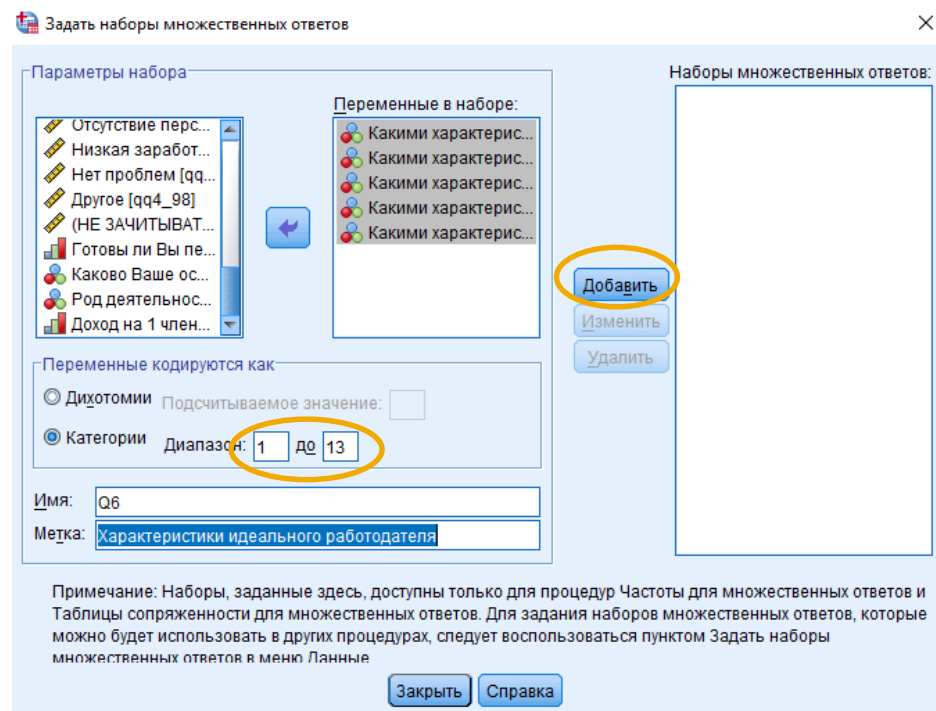
Кодирование переменных:

Максимальное количество возможных ответов равно 5. Каждая из пяти переменных кодируется одинаковыми категориями. Не зависимо от количества данных ответов область этих переменных заполняется слева направо.

Определение наборов в SPSS:

«Анализ» → «Множественные ответы» → «Определить наборы»

- Выбрать переменные Q6_1 – Q6_5
- Задать категориальную кодировку
- Выбрать диапазон значений: 1:13



3. ФОРМЫ ПРЕДСТАВЛЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ



3. Формы представления статистических данных

Визуализация данных



Цель визуализации — помочь интерпретировать данные без специальных знаний, используя способность зрительной системы видеть закономерности, выявлять тенденции и обнаруживать отклонения.



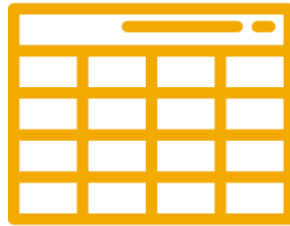
Визуализация данных является наиболее эффективным вариантом для глубокого и точного анализа и превращения чисел в наглядные представления.

3. Формы представления статистических данных

Существуют три основных способа
представления статистических данных:



Включение данных
в текстовый блок



Табличная форма



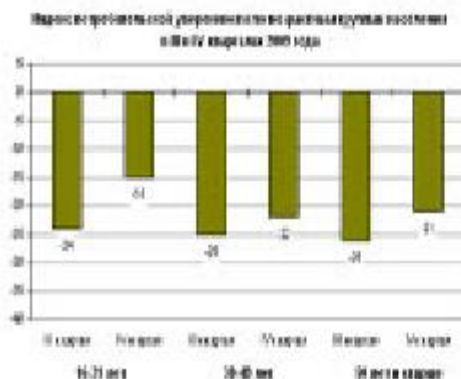
Графическая форма

3. Формы представления статистических данных

Графическое представление данных может осуществляться как в SPSS, так и в других средах (например, в Excel), на основе полученных в SPSS расчетов.

Динамика численности экономически активного населения

		Экономически активное население	В том числе		Уровень безработицы (%)	Уровень занятости (%)
			Занятые, чел.	Безработные, чел.		
Период	январь	74568930	67737181	6831749	9.2	60.8
	февраль	74445260	68029677	6415583	8.6	61.1
	март	74645510	68227702	6417807	8.6	61.2



3. Формы представления статистических данных

Графическая визуализация частотных распределений в SPSS

Столбчатая диаграмма (Bar chart) (вертикальная или горизонтальная) – используется при малом количестве дискретных категорий номинальных и порядковых переменных.

Гистограмма (Histogram) – используется для непрерывных интервальных переменных. В отличие от bar charts, в гистограмме «столбики» располагаются один за одним, без пространства. По гистограмме можно судить о нормальности распределения.

Линейный график (Line chart) – используется для непрерывных интервальных переменных (временные ряды, децильные группы и т.д.).

Неправильно использовать для номинальных переменных!

Круговая диаграмма (Pie chart) – используется для номинальных и порядковых переменных с относительно небольшим числом дискретных категорий (до 5).

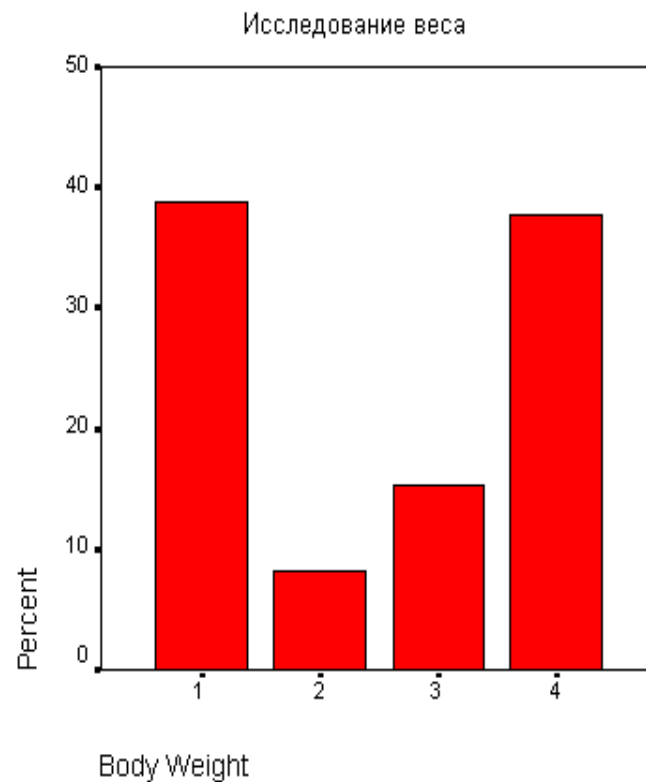
3. Формы представления статистических данных

Столбчатая диаграмма

Столбчатая диаграмма – отображает итоги для значений переменной.

Как правило, применяется:

- Для отображения частот переменных, относящихся к номинальной или порядковой шкале
- Для отображения средних значений, сумм или других показателей последовательных переменных (т.е. переменных, принадлежащих к интервальной шкале или к шкале отношений)



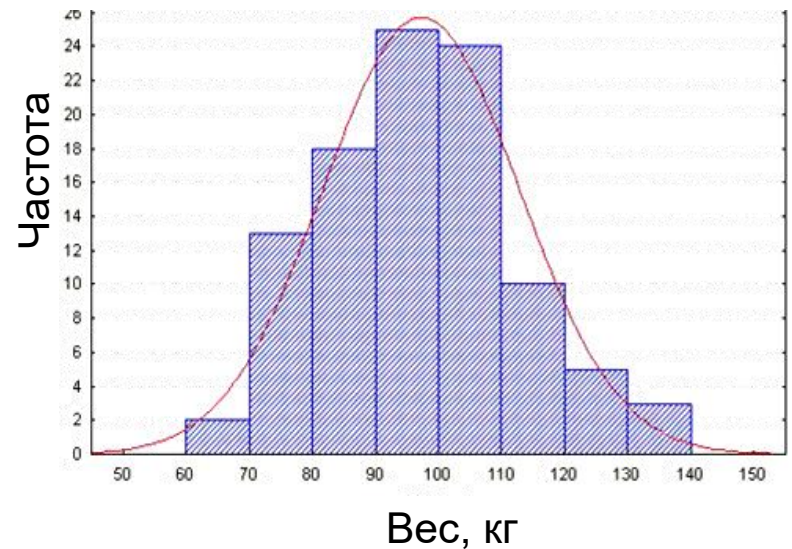
3. Формы представления статистических данных

Гистограмма

Гистограмма – графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает **ЧАСТОТУ**.

Интервалы:

- должны быть одного размера
- не должны иметь общих точек



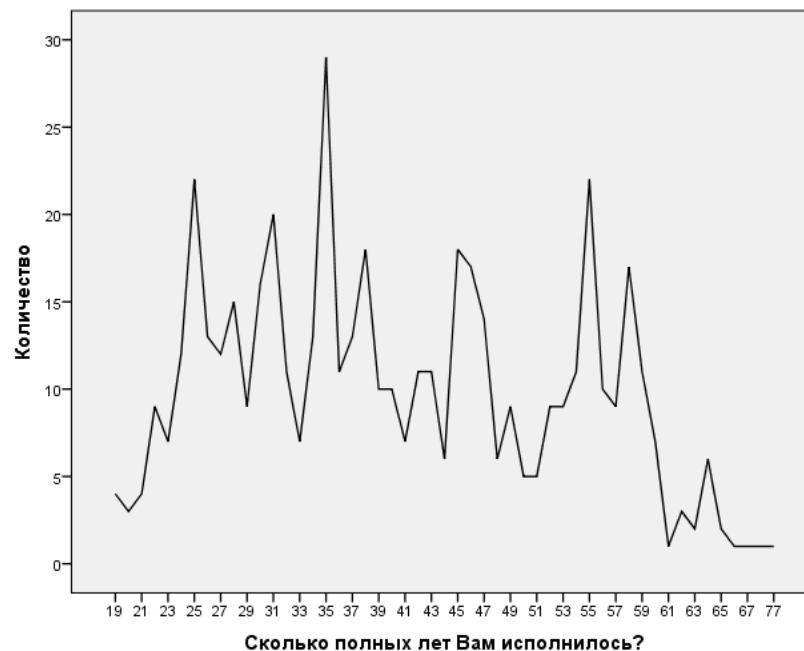
3. Формы представления статистических данных

Линейная диаграмма

Линейная диаграмма – отображает то же самое, что и столбчатая диаграмма, только в виде ломанной линии.

Нельзя использовать для **номинальных** переменных.

Диаграмма



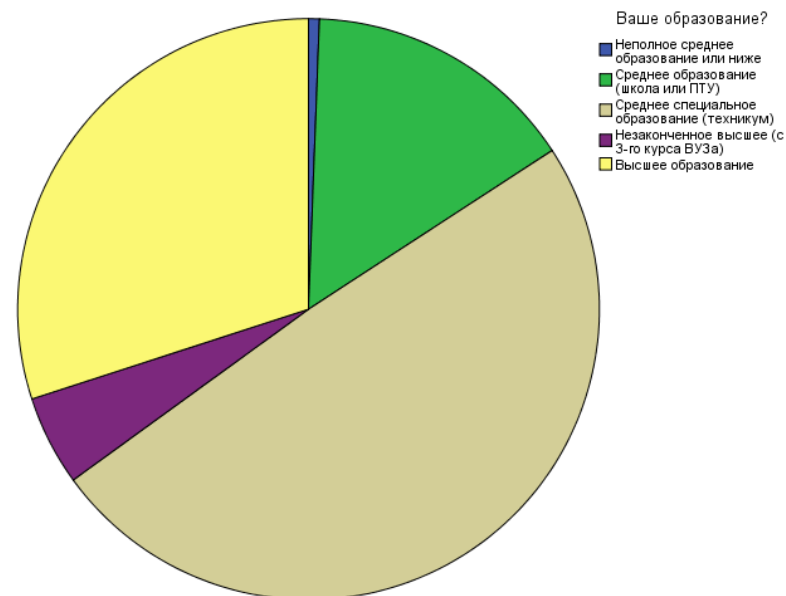
3. Формы представления статистических данных

Круговая диаграмма

Круговая диаграмма – применяется для иллюстрации распределения наблюдений в различных категориях.

Номинальные и порядковые переменные (до 5 дискретных категорий).

Диаграмма

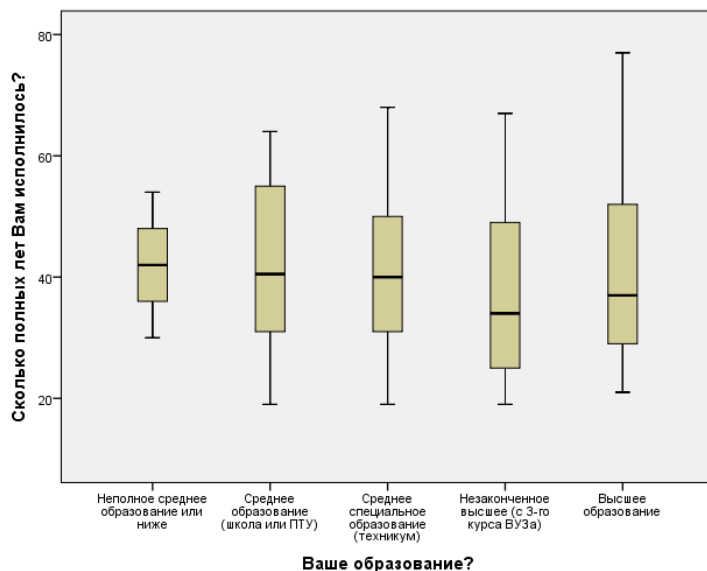


3. Формы представления статистических данных

Дополнительные графики

Существует множество других графиков, однако популярными являются еще два вида диаграмм:

Сколько полных лет Вам исполнилось?



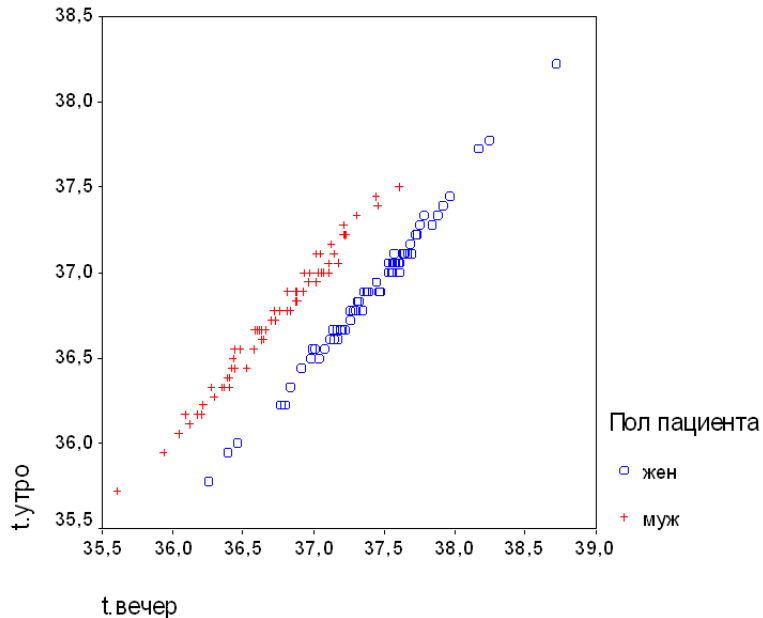
1. Коробчатая диаграмма (Boxplot) — применяется для отображения различия между группами с учетом распределения данных.

Представляет:

- медиану
- первый и третий квартили
- минимальное и максимальное значения
- аномальные и экстремальные значения

3. Формы представления статистических данных

Дополнительные графики



2. Диаграмма рассеяния (Scatterplot) – отображает характер взаимосвязи между переменными.

Используются интервальные переменные. Диаграмма представляется в форме скопления точек.

3. Формы представления статистических данных

Формат представления графиков

Правила: аккуратность, информативность.

Должны быть указаны:

- название и номер
- название переменных (рядов) (легенда) и категорий – при использовании второй оси значений, обязательно указать в легенде
- названия осей
- подписи данных – если это не загромождает график
- линии сетки – желательно, но чтобы не было «частотола»
- % от каких переменных (респондентов, ответивших , ответов и т.д. – особенно при множественном выборе)
- общее число респондентов – N
- источник данных

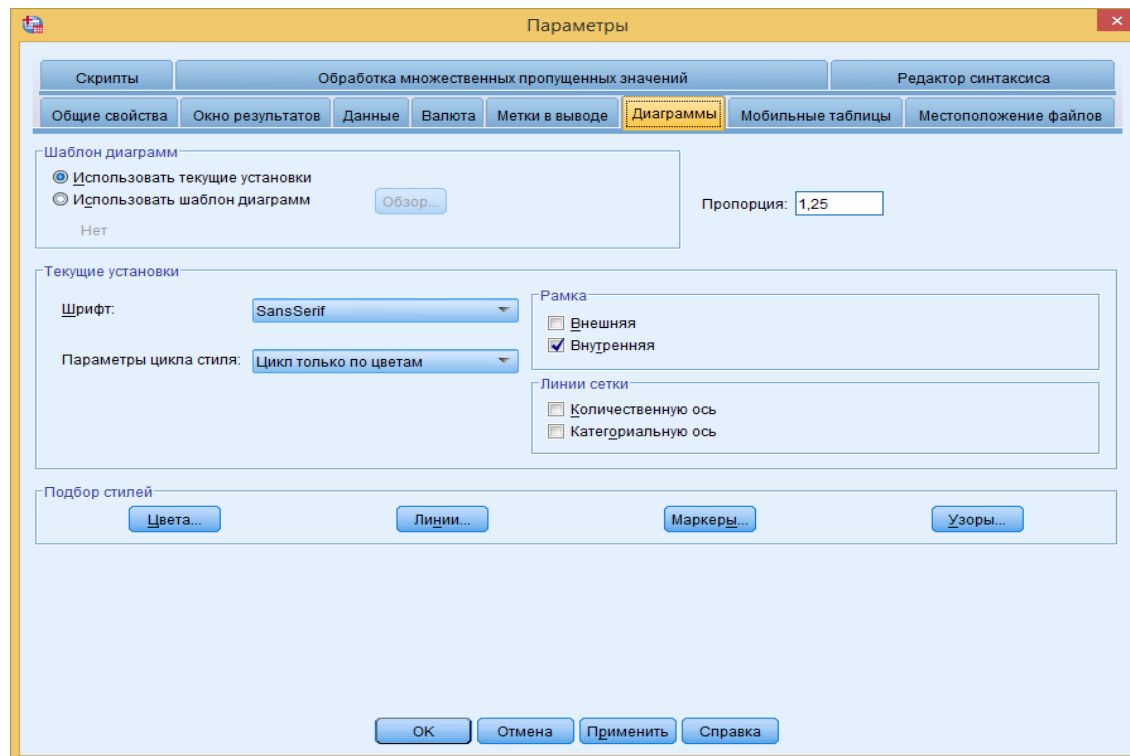
4. СОЗДАНИЕ И РЕДАКТИРОВАНИЕ ГРАФИКОВ И ДИАГРАММ В SPSS



4. Создание и редактирование графиков и диаграмм

Для изменения общих настроек диаграмм необходимы следующие шаги:

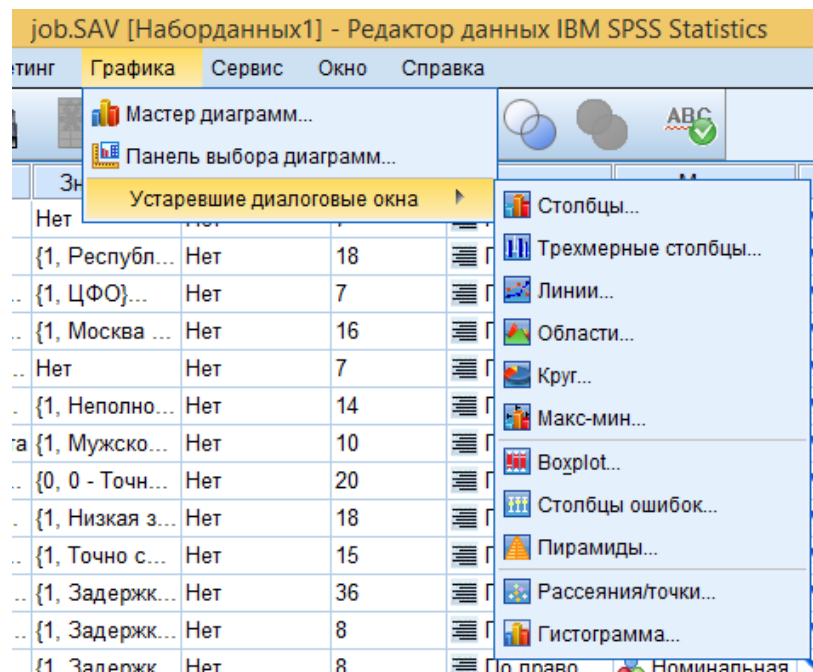
1. Меню **«Редактор данных»** → **«Правка»** → **«Параметры»**.
2. Перейти на вкладку **«Диаграммы»**. Можно задать самостоятельно основные свойства диаграмм, которые в дальнейшем будут применяться программой по умолчанию ко всем графикам.



4. Создание и редактирование графиков и диаграмм

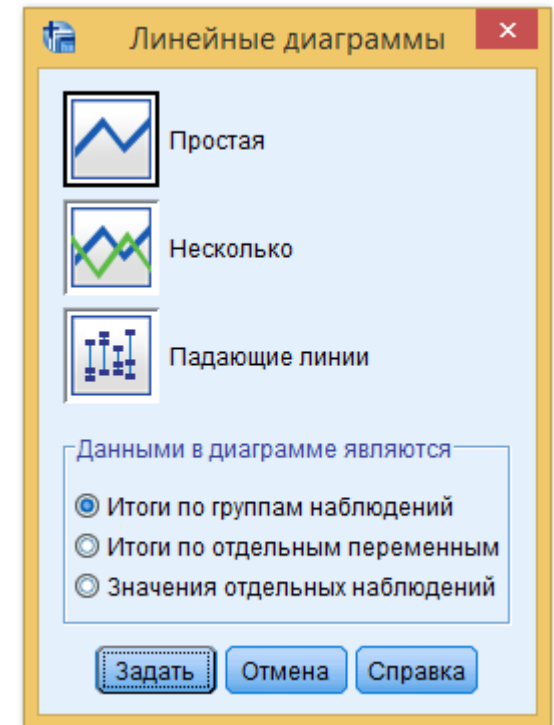
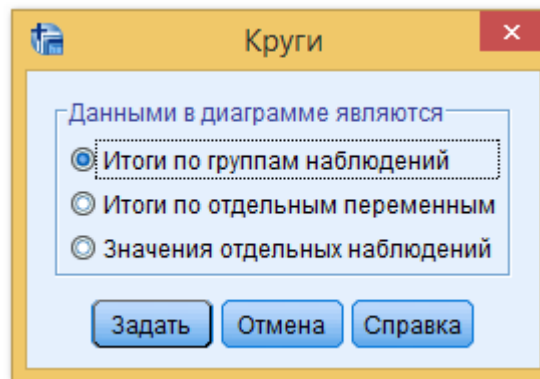
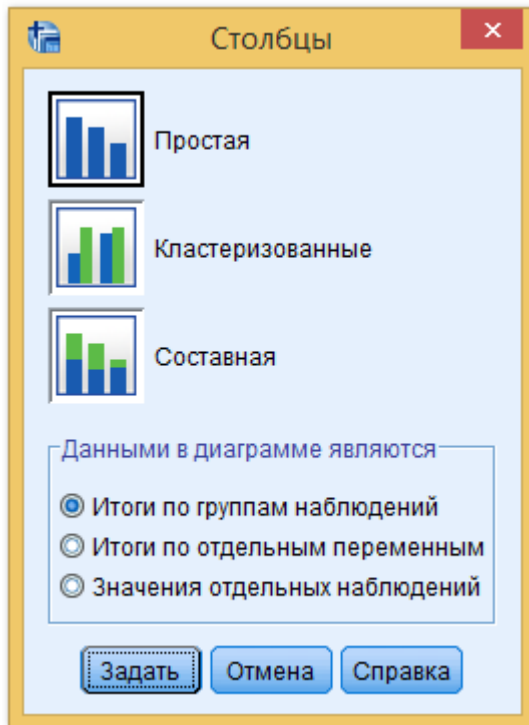
Чтобы построить график необходимо выполнить следующие шаги:

1. Сначала загрузите файл job.sav, выбрав команды меню File (Файл) / Open... (Открыть...).
2. Выберите в меню команды Graphs (Графики) / Legacy Dialogs (Устаревшие диалоговые окна). Далее в зависимости от типа графика выбираем:
 - Bar... (Столбцы...)
 - Line... (Линии...)
 - Histogram... (Гистограмма...)
 - Pie... (Круг...)



4. Создание и редактирование графиков и диаграмм

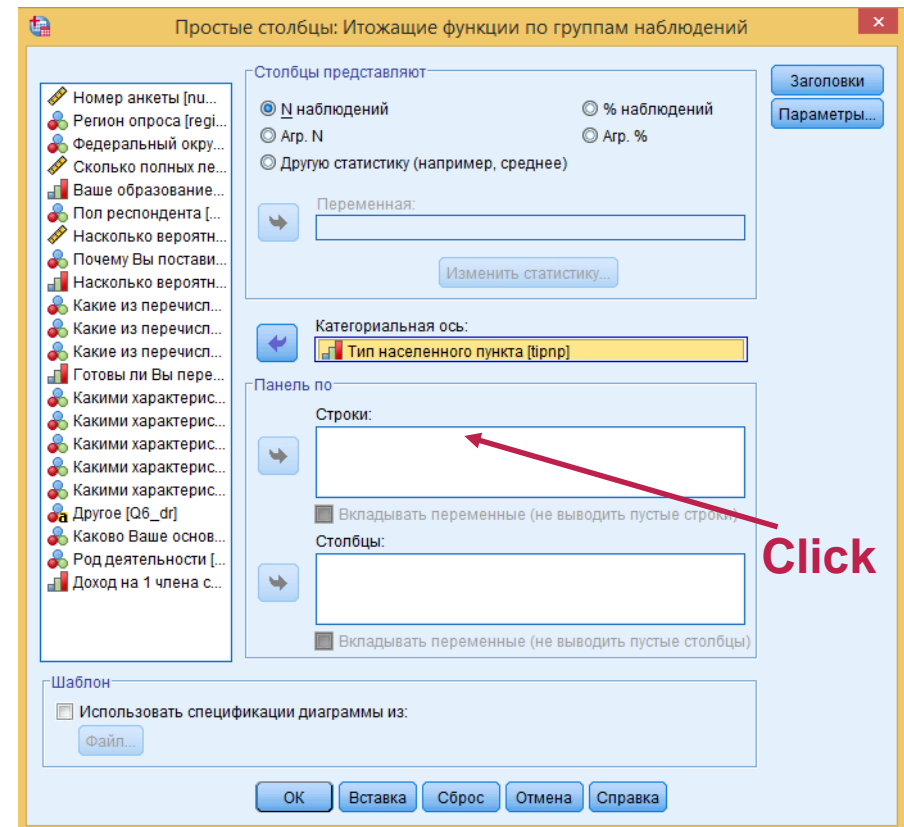
3. В зависимости от вида диаграммы можно выбрать его тип (характерно для столбчатых и линейных диаграмм).



4. Также можно указать, какие данные использовать, установив переключатель.

4. Создание и редактирование графиков и диаграмм

5. Выберите нужную переменную и переместить ее в строку:
- Категориальная ось (для столбчатой и линейной диаграмм)
 - Задать сектора значениями (для круговой диаграммы)
 - Переменная (для гистограмм)
6. При необходимости заполните остальные строчки, если необходимо посмотреть распределение по группам.

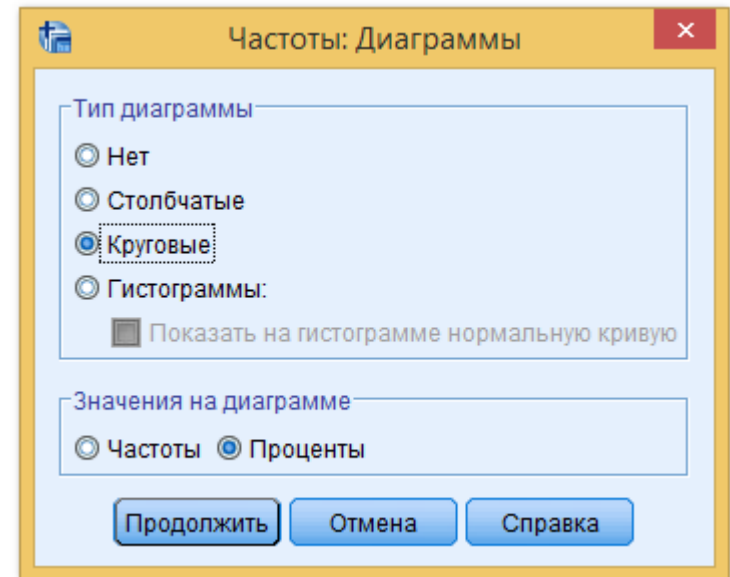


4. Создание и редактирование графиков и диаграмм

Гистограмма в частотах

Можно строить диаграммы не выходя из раздела «Описательные статистики».

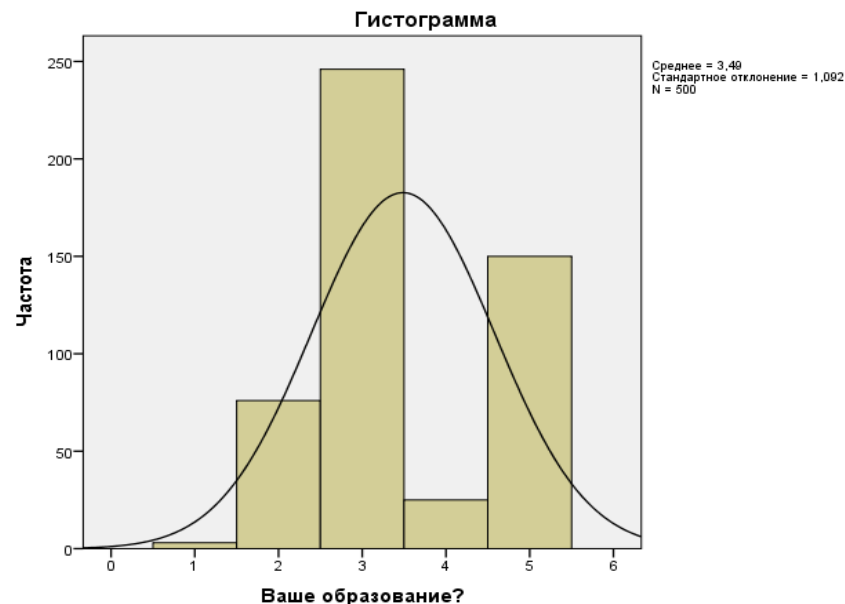
- «Анализ» → «Описательные статистики» → «Частоты» → указать переменные.
- «Диаграммы» → выбрать нужный вид :
 - «Столбчатые»
 - «Круговые»
 - «Гистограммы»
- «Значения на диаграмме» → выбрать «Частоты» или «Проценты».
- «Продолжить» → «Ок» → «Частоты» → опция «Вывести частотные таблицы».



4. Создание и редактирование графиков и диаграмм

Гистограмма в частотах

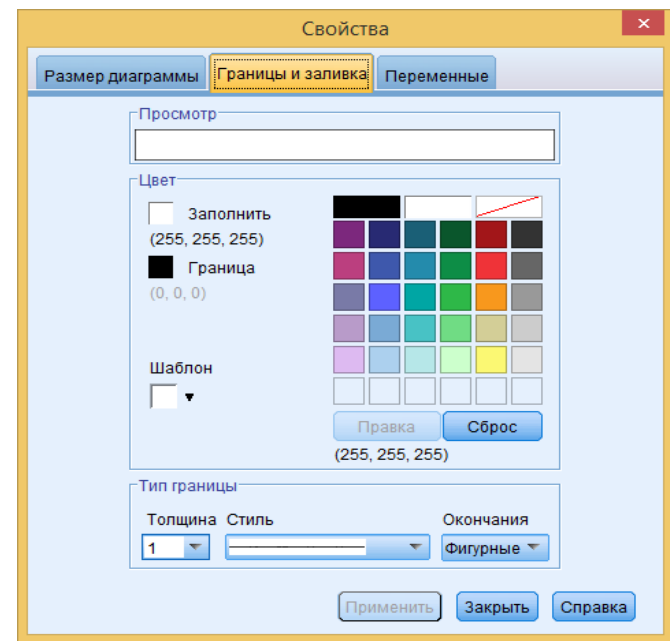
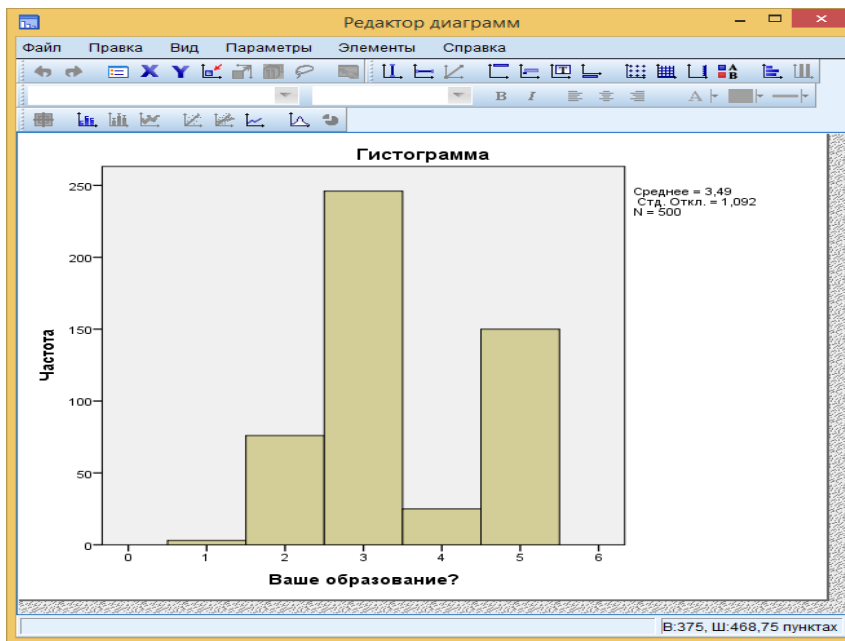
- «Анализ» → «Описательные статистики» → «Частоты» → указать переменные.
- «Диаграммы» → «Гистограммы» → установить флажок «Показывать на гистограмме нормальную кривую».
- «Продолжить» → «Ок» → «Частоты» → опция «Вывести частотные таблицы».



4. Создание и редактирование графиков и диаграмм

Редактирование диаграмм

- Для редактирования диаграммы необходимо в окне вывода дважды щелкнуть мышью по **выбранному графическому объекту**.
- На экране появится окно «**Редактор диаграмм**», содержащее строку меню с набором команд и панель инструментов.
- Дважды щелкнуть мышью по диаграмме → «**Свойства элемента**».



Литература по Теме 2

- 1. Бююль А., Цеффель П. SPSS: искусство обработки информации. – М., 2005**
 - Глава 3. Подготовка данных
 - Глава 5. Основы статистики
 - Глава 6. Частотный анализ
 - Глава 12. Анализ множественных ответов

- 2. Наследов А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. – СПб., 2013**
 - Глава 5. Диаграммы



Для свободного использования в образовательных целях
Copyright 2017 © Академия НАФИ. Москва
Все права защищены
www.nafi.ru