Введение в анализ данных
Лекция 6.3

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

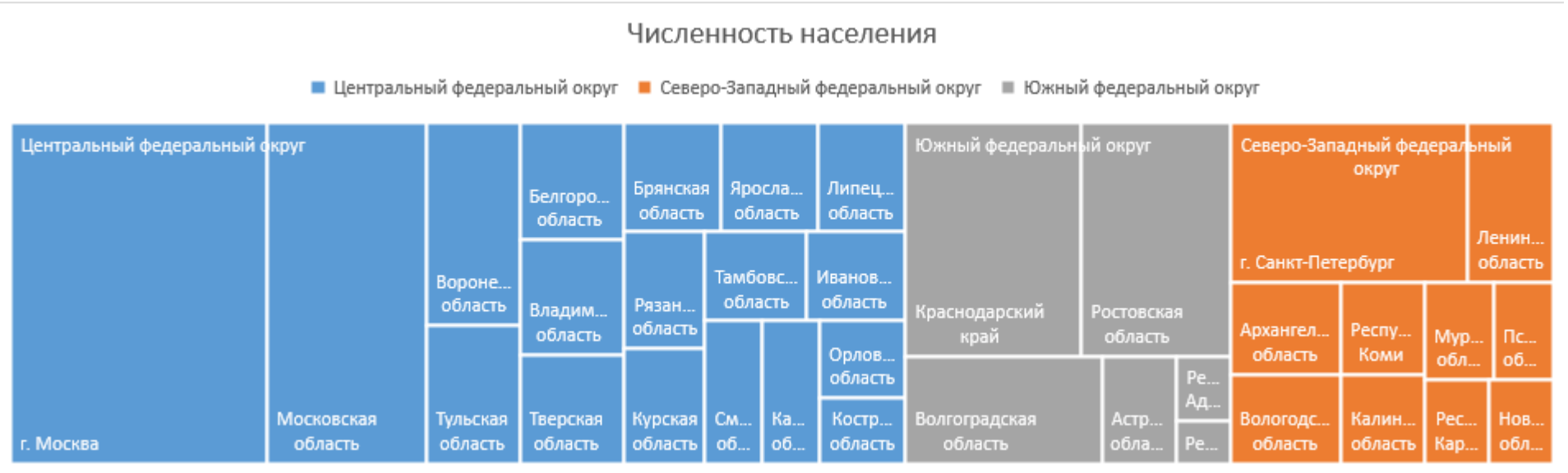
Визуализация данных

Иерархические диаграммы. Статистические диаграммы. Диаграммы с пользовательскими элементами управления.

Иерархические диаграммы

Иерархическая диаграмма (Treemap Chart), дерево

Древовидная диаграмма отображает иерархическое представление данных.
Ветви дерева представляются прямоугольники из родительских и дочерних веток, дочерние отображается как вложенные прямоугольники меньшего размера.
Древовидная диаграмма отображает категории, цвет и расположение, легко отображает данные больших объемов, которые будет сложно отобразить с помощью диаграмм других типов.
Диаграмма "дерево" удобна, если нужно сравнить пропорции в иерархии.



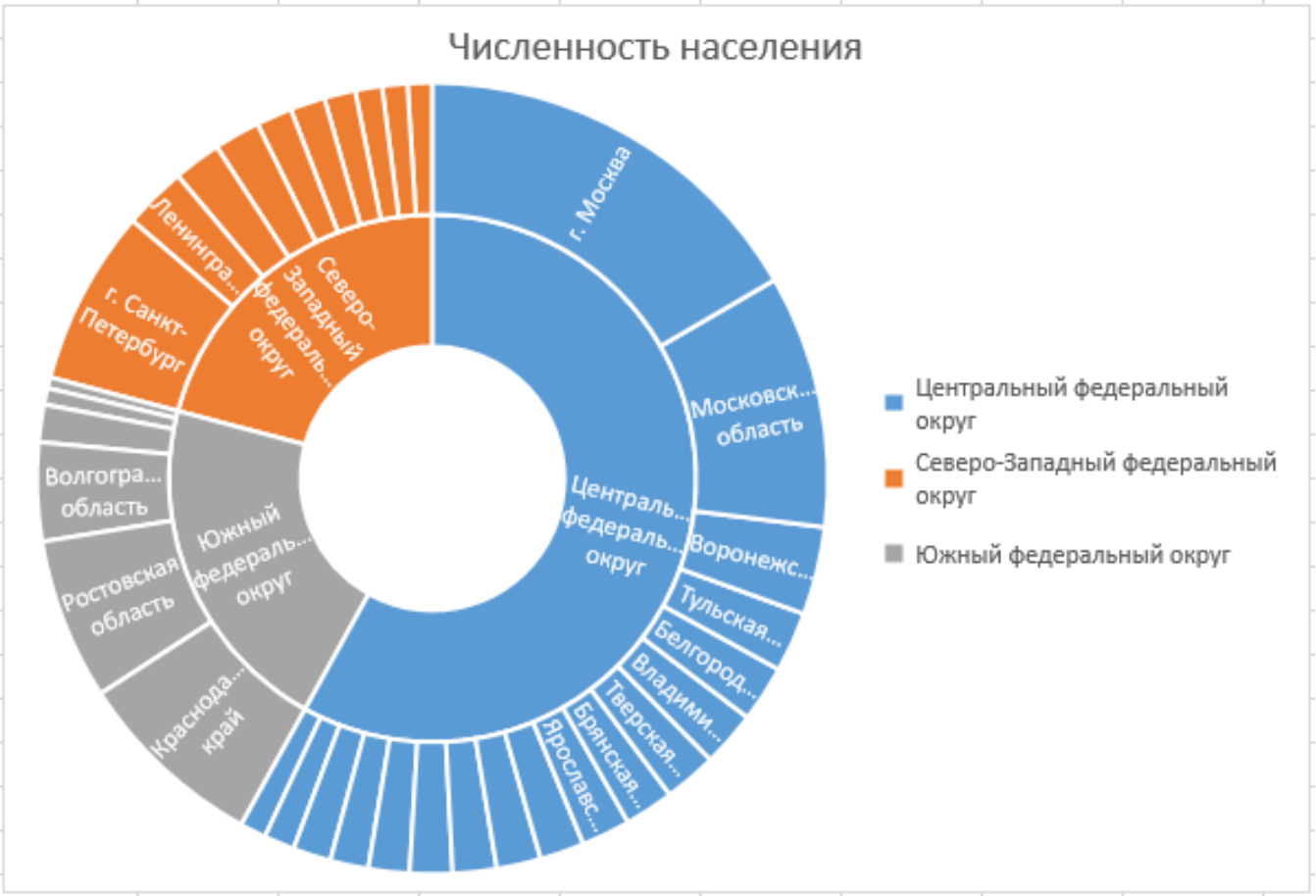
Древовидная диаграмма

	A	B	C
1	**Округ**	**Область**	**2005**
2	Центральный федеральный округ	Белгородская область	1512
3	Центральный федеральный округ	Брянская область	1327
4	Центральный федеральный округ	Владимирская область	1486
5	Центральный федеральный округ	Воронежская область	2361
6	Центральный федеральный округ	Ивановская область	1102
7	Центральный федеральный округ	Калужская область	1023
8	Центральный федеральный округ	Костромская область	700
9	Центральный федеральный округ	Курская область	1178
10	Центральный федеральный округ	Липецкая область	1194
11	Центральный федеральный округ	Московская область	6784
12	Центральный федеральный округ	Орловская область	822
13	Центральный федеральный округ	Рязанская область	1189
14	Центральный федеральный округ	Смоленская область	1025
15	Центральный федеральный округ	Тамбовская область	1139
16	Центральный федеральный округ	Тверская область	1415
17	Центральный федеральный округ	Тульская область	1615
18	Центральный федеральный округ	Ярославская область	1313
19	Центральный федеральный округ	г. Москва	10924
20	Северо-Западный федеральный округ	Республика Карелия	676
21	Северо-Западный федеральный округ	Республика Коми	963
22	Северо-Западный федеральный округ	Архангельская область	1282
23	Северо-Западный федеральный округ	Вологодская область	1235
24	Северо-Западный федеральный округ	Калининградская область	936
25	Северо-Западный федеральный округ	Ленинградская область	1685
26	Северо-Западный федеральный округ	Мурманская область	839

Для отображения вложенных категорий нужны соответствующим образом подготовленные данные. В приведенном примере данные группируются внутри округов, дочерние элементы соответствуют областям, размер элементов диаграммы соответствует численности населения

Солнечные лучи

	A	B	C
1	**Округ**	**Область**	**2005**
2	Центральный федеральный округ	Белгородская область	1512
3	Центральный федеральный округ	Брянская область	1327
4	Центральный федеральный округ	Владимирская область	1486
5	Центральный федеральный ок		
6	Центральный федеральный ок		
7	Центральный федеральный ок		
8	Центральный федеральный ок		
9	Центральный федеральный ок		
10	Центральный федеральный ок		
11	Центральный федеральный ок		
12	Центральный федеральный ок		
13	Центральный федеральный ок		
14	Центральный федеральный ок		
15	Центральный федеральный ок		
16	Центральный федеральный ок		
17	Центральный федеральный ок		
18	Центральный федеральный ок		
19	Центральный федеральный ок		
20	Северо-Западный федеральны		
21	Северо-Западный федеральны		
22	Северо-Западный федеральны		
23	Северо-Западный федеральны		
24	Северо-Западный федеральны		
25	Северо-Западный федеральны		
26	Северо-Западный федеральны		

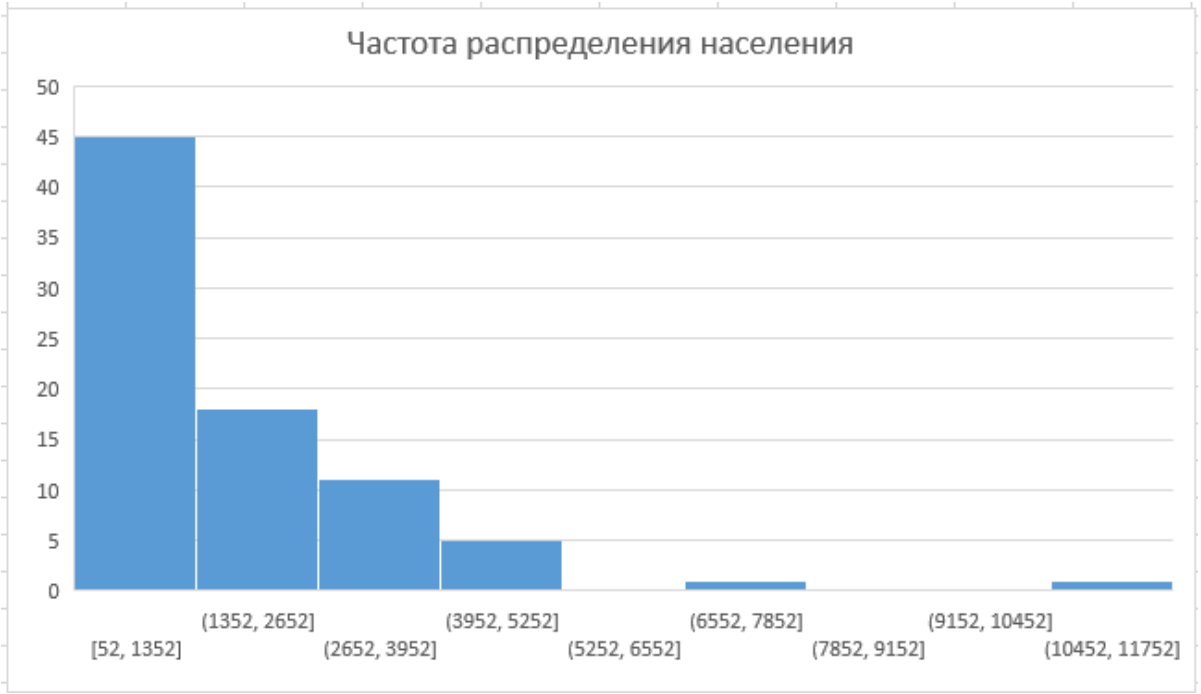


Статистические диаграммы

Гистограмма частот

В более ранних версиях Excel можно изобразить частоты с помощью диаграмм, но для этого предварительно необходимо данные сгруппировать.
То есть для каждой категории (интервала, группы, года и т.д.) должно быть свое значение.

B	C
Область	**2005**
Белгородская область	1512
Брянская область	1327
Владимирская область	1486
Воронежская область	2361
Ивановская область	1102
Калужская область	1023
Костромская область	700
Курская область	1178
Липецкая область	1194
Московская область	6784
Орловская область	822
Рязанская область	1189
Смоленская область	1025
Тамбовская область	1139
Тверская область	1415
Тульская область	1615
Ярославская область	1313



Гистограмма частот

Вариант "Автоматическая" (формула Скотта)

$$Интервал\ (h) = \frac{3.5 \times \sigma}{\sqrt[3]{n}}$$

σ = стандартное отклонение источника данных

n = количество значений в источнике данных

Формула Скотта минимизирует отклонение вариационного ряда на гистограмме по сравнению с набором данных, исходя из предположения о нормальном распределении данных.

Вариант "Выход за верхнюю границу интервала"

$$\bar{\chi} + 3 \times \sigma$$

$$\bar{\chi} = среднее\ источника\ данных$$

$$\sigma = стандартное\ отклонение\ источника\ данных$$

Вариант "Выход за нижнюю границу интервала"

$$\bar{\chi} - 3 \times \sigma$$

$$\bar{\chi} = среднее\ источника\ данных$$

$$\sigma = стандартное\ отклонение\ источника\ данных$$

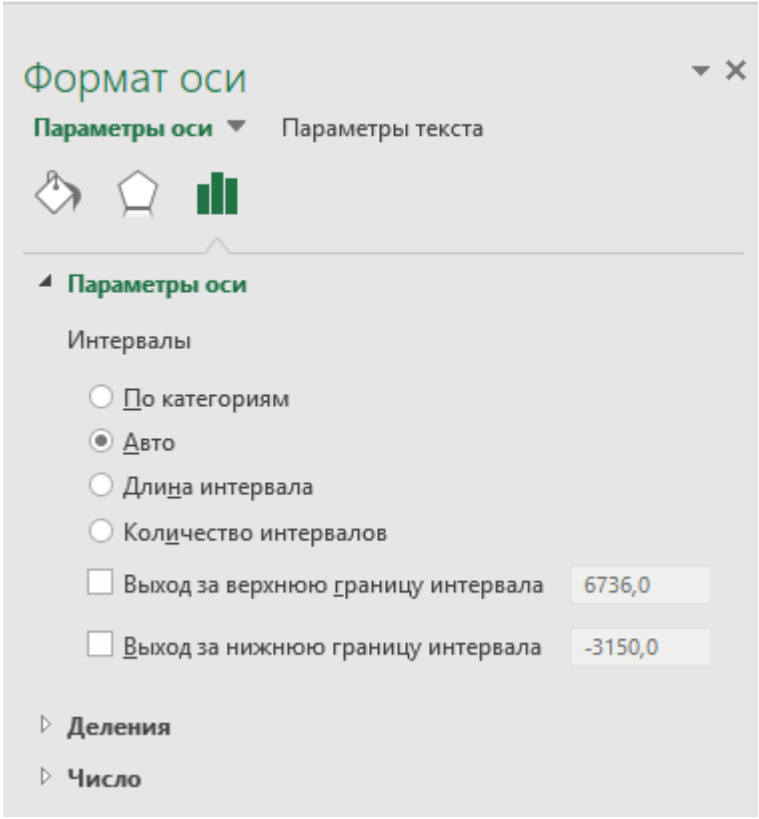
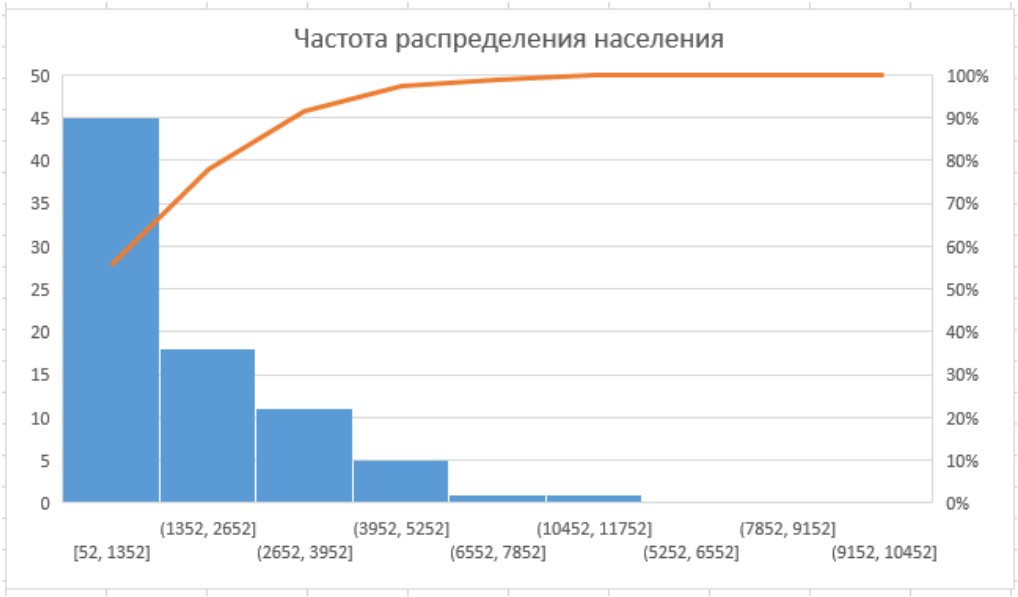


Диаграмма Парето

Аналог частотной диаграммы. В диаграмме Парето, как правило, используются не числовые диапазоны, а категории. Категории могут располагаться в произвольном порядке. Допускается несколько строк на одну и ту же категорию.
Диаграмма Парето является комбинированной: наряду с частотной диаграммой присутствует кумулятивная кривая накопленной доли категорий. Для этой кривой используется вторая ось ординат – справа.



Столбики – это отсортированные по убыванию значения отдельных элементов. График – соответствующие накопленные доли. Последнее значение равно 100%.
Чтобы построить подобную диаграмму в Excel 2013 и более ранних версиях, нужно выполнить следующие действия:
1) отсортировать данные по убыванию, чтобы значения, имеющие наибольший вклад, были в начале списка;
2) рассчитать столбец с накопленными долями;
3) использовать комбинированную диаграмму, чтобы столбиками показать отдельные элементы, а графиком накопленные доли.

Диаграмма размахов (ящик с усами)

- Диаграммы размахов (box plot) иллюстрируют распределение значений непрерывной переменной, отображая пять параметров:
  - минимум,
  - нижний квартиль (25-й процентиль),
  - медиану (50-й процентиль),
  - верхний квартиль (75-й процентиль)
  - максимум.
- На этой диаграмме также могут быть отображены вероятные выбросы (значения, выходящие за диапазон в ±1.5 межквартильного размаха, разности верхней и нижней квартилей).
- По умолчанию каждый «ус» продолжается до минимального или максимального значения, которое не выходит за пределы 1.5 межквартильного размаха. Выходящие за эти пределы значения отмечаются точками

Сравнение плотности распределения и ящика с усами

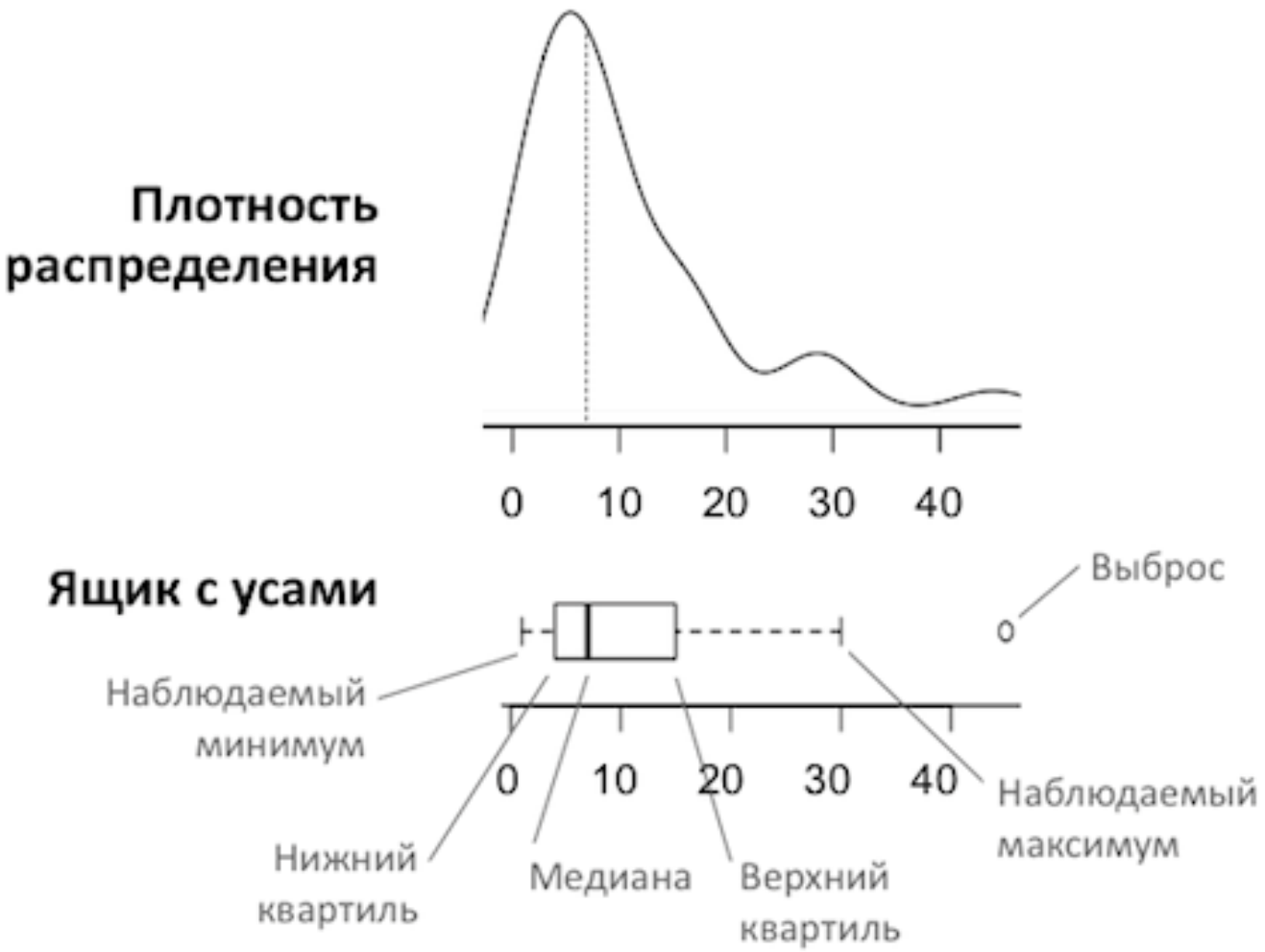


Диаграмма размахов (ящик с усами)

Дополнительно:
- среднее арифметическое
- выбросы

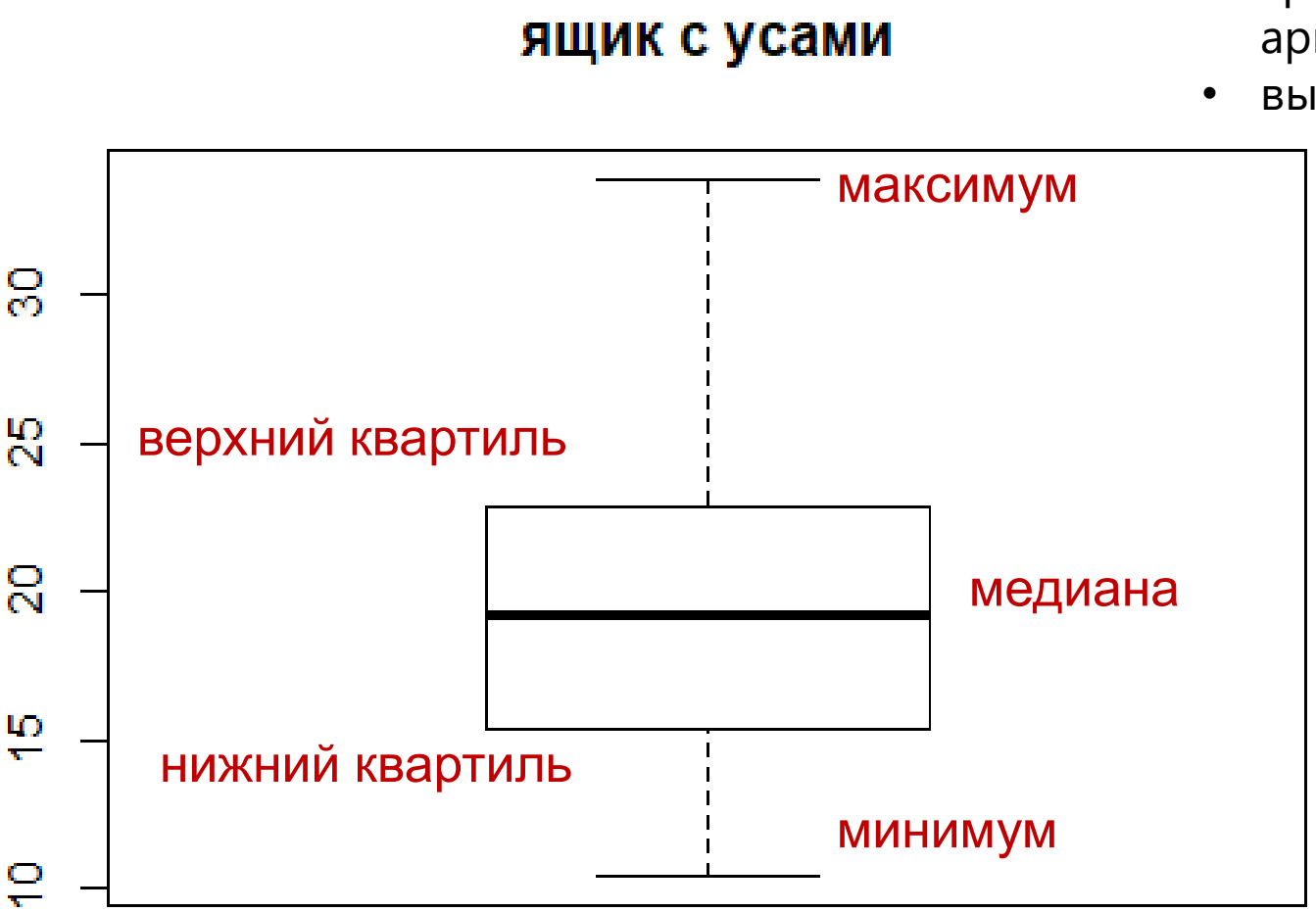


Диаграмма размахов (ящик с усами)

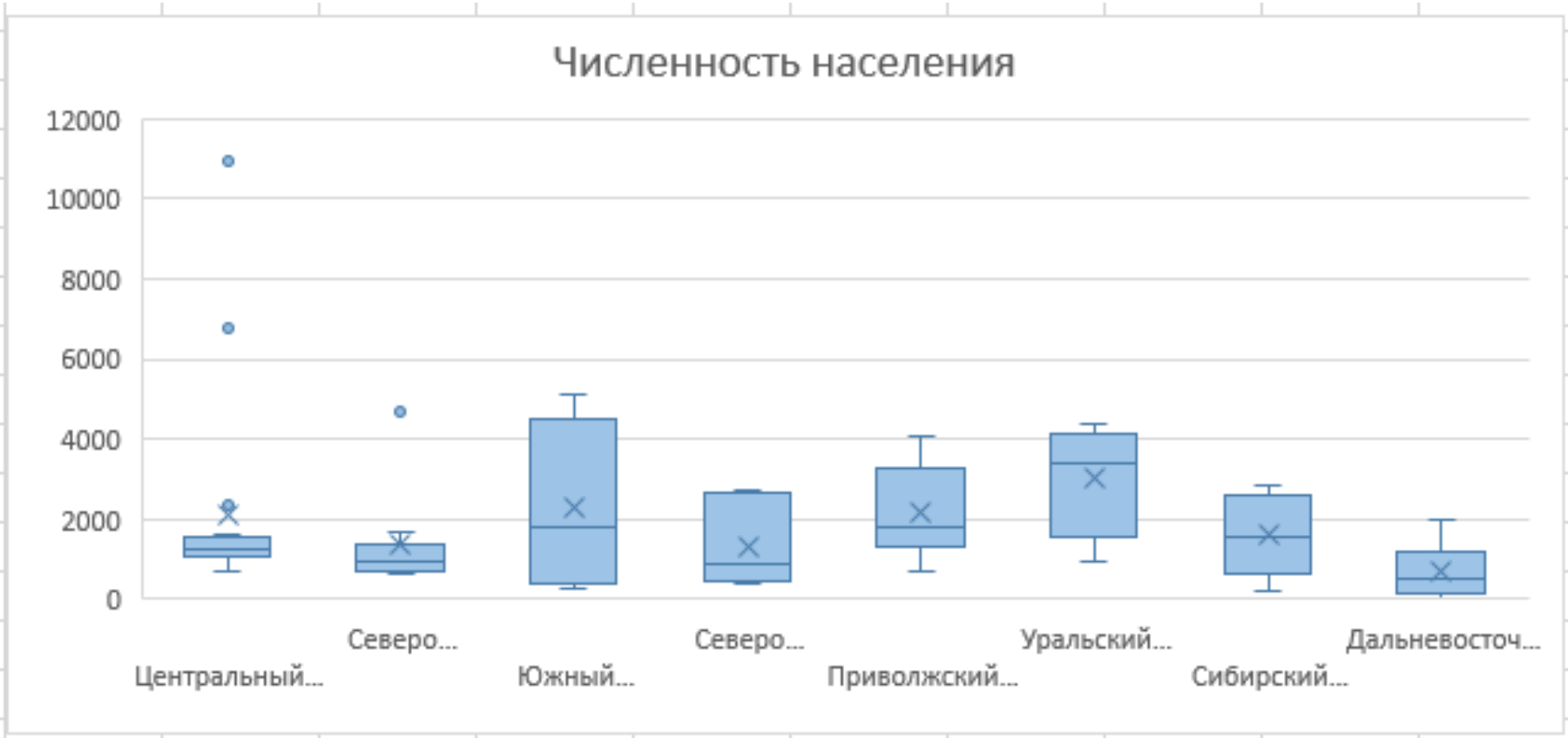
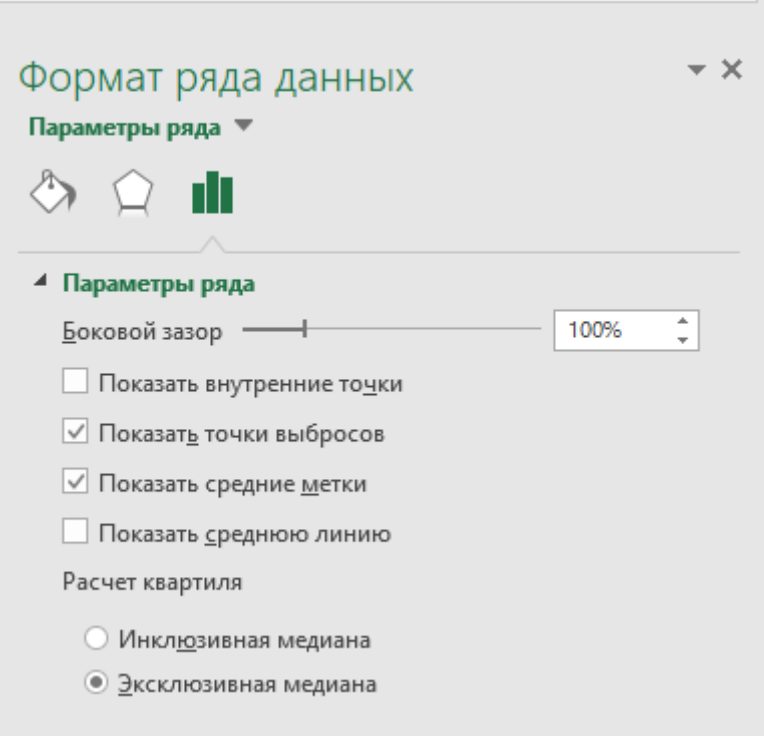


Диаграмма размахов (ящик с усами). Настройки



Инклюзивная медиана включает в «ящик» квартильные значения, а **эксклюзивная медиана** не включает.

При выборе **эксклюзивной медианы** верх и низ «ящика» соответствует средней между квартильным и следующим (от центра) значением. По умолчанию стоит **эксклюзивная**.

В то время как медиана разделяет упорядоченный массив пополам, квартили разбивают набор данных на четыре части. Первый квартиль – это число, разделяющее выборку на две части: 25% элементов меньше, а 75% — больше значения первого квартиля. Третий квартиль — это число, разделяющее выборку также на две части: 75% элементов меньше, а 25% — больше третьего квартиля.

Для расчета квартилей в Excel2007 и более ранних версиях использовалась функция КВАРТИЛЬ. Начиная с версии Excel2010 применяются две функции: КВАРТИЛЬ.ВКЛ и КВАРТИЛЬ.ИСКЛ

Диаграммы с пользовательскими элементами управления

Диаграмма с включением/выключением рядов данных

Чтобы не строить несколько диаграмм или постоянно не изменять исходные данные для просмотра данных по отдельности, можно построить диаграммы с возможностью управления рядами, т.е. при необходимости ряды включать или выключать на диаграмме.

Построение такого рода диаграмм возможно с использованием дополнительной таблицы, в которую данные копируются или не копируются из исходной таблицы, если ряд данных включен или выключен.

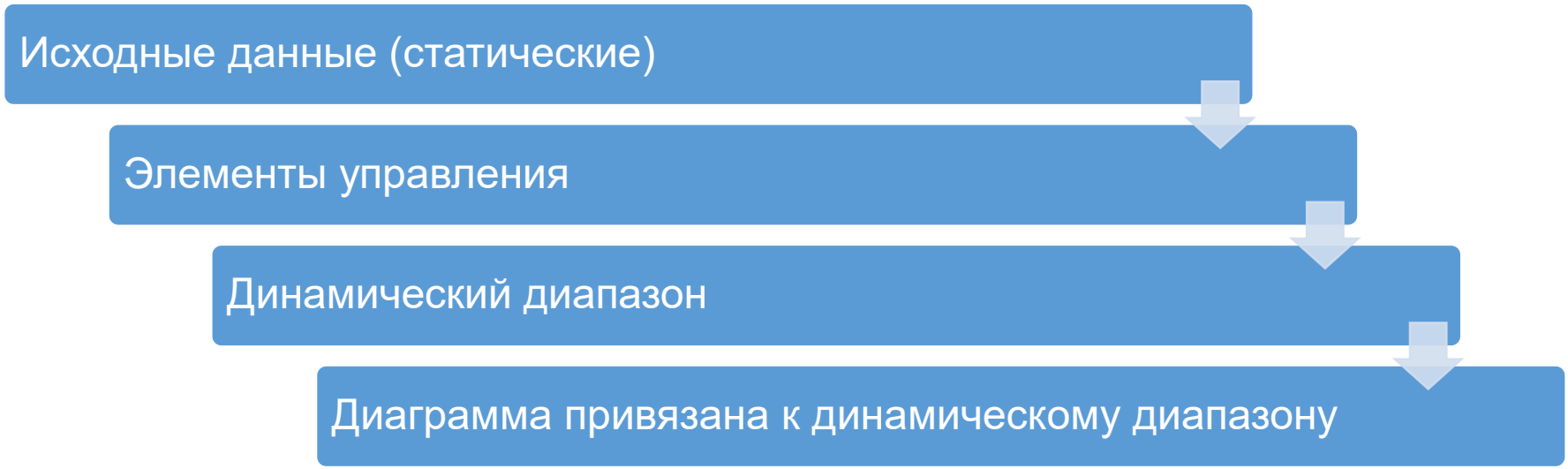
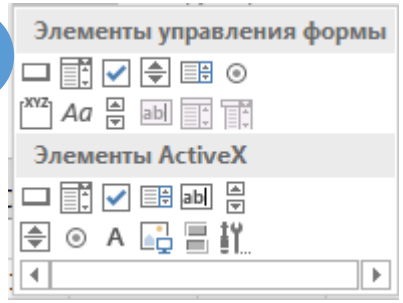


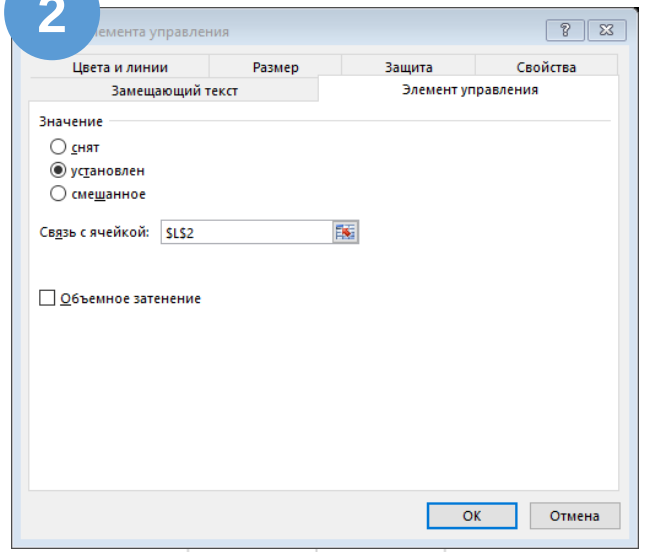
Диаграмма с включением/выключением рядов данных

B	C	D	E	F	G	H	I	J
Область	**2005**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Белгородская область	1512	1532	1536	1541	1544	1548	1550	1553
Брянская область	1327	1275	1264	1254	1242	1233	1226	1221
Владимирская область	1486	1441	1432	1422	1413	1406	1397	1390
Воронежская область	2361	2335	2332	2330	2329	2331	2333	2335
Ивановская область	1102	1060	1054	1049	1043	1037	1030	1023
Калужская область	1023	1009	1008	1006	1005	1011	1010	1014
Костромская область	700	666	662	659	656	654	651	648
Курская область	1178	1126	1122	1119	1119	1117	1120	1123
Липецкая область	1194	1172	1166	1162	1160	1158	1156	1156
Московская область	6784	7106	7199	7048	7134	7231	7319	7423
Орловская область	822	786	781	776	770	765	760	755
Рязанская область	1189	1152	1148	1144	1141	1135	1130	1127
Смоленская область	1025	983	981	975	968	965	959	953
Тамбовская область	1139	1090	1082	1076	1069	1062	1050	1040
Тверская область	1415	1350	1342	1334	1325	1315	1305	1297
Тульская область	1615	1550	1545	1532	1522	1514	1506	1499
Ярославская область	1313	1271	1271	1272	1272	1272	1272	1271
г. Москва	10924	11541	11613	11980	12108	12197	12330	12381
Республика Карелия	676	643	640	637	634	633	630	627
Республика Коми	963	899	890	880	872	864	857	850
Архангельская область	1282	1225	1213	1202	1192	1183	1174	1166

1



2



Управляющими элементами для отображения рядов данных будут флажки из панели меню **Разработчик**, элементы управления формы. Значение флажка привязывается к ячейке.

3

ИСТИНА	2005
ИСТИНА	2010
ИСТИНА	2011
ИСТИНА	2012
ИСТИНА	2013
ИСТИНА	2014
ИСТИНА	2015
ИСТИНА	2016

Диаграмма с включением/выключением рядов данных

=ЕСЛИ($L$2;Лист1!C2;#Н/Д)

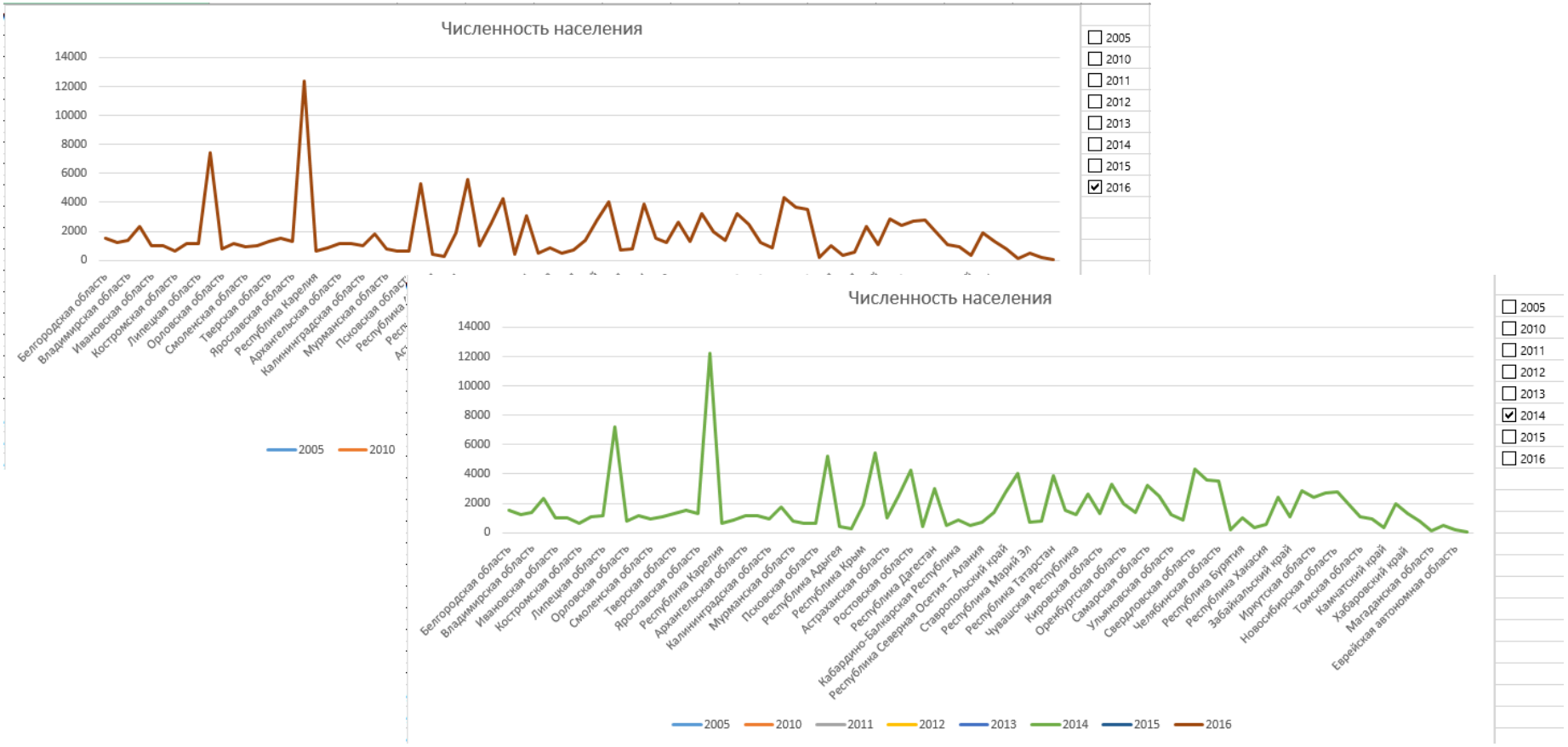
B	C	D	E	F	G	H	I	J	K	L	M
Область	**2005**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**			
Белгородская область	1512	1532	1536	1541	1544	1548	1550	1553		ИСТИНА	☑ 2005
Брянская область	1327	1275	1264	1254	1242	1233	1226	1221		ИСТИНА	☑ 2010
Владимирская область	1486	1441	1432	1422	1413	1406	1397	1390		ИСТИНА	☑ 2011
Воронежская область	2361	2335	2332	2330	2329	2331	2333	2335		ИСТИНА	☑ 2012
Ивановская область	1102	1060	1054	1049	1043	1037	1030	1023		ИСТИНА	☑ 2013
Калужская область	1023	1009	1008	1006	1005	1011	1010	1014		ИСТИНА	☑ 2014
Костромская область	700	666	662	659	656	654	651	648		ИСТИНА	☑ 2015
Курская область	1178	1126	1122	1119	1119	1117	1120	1123		ИСТИНА	☑ 2016
Липецкая область	1194	1172	1166	1162	1160	1158	1156	1156			
Московская область	6784	7106	7199	7048	7134	7231	7319	7423			

Формируется динамическая таблица.

Наименования категорий жестко привязаны к статическим данным.
Заголовки рядов тоже жестко привязаны к статическим данным.

Значения рядов данных в зависимости от состояния флажка, соответственно значения связанной ячейки принимают значение или #Н/Д

Диаграмма с включением/выключением рядов данных



Значения связанных ячеек лучше не отображать. Этого можно достичь пользовательским числовым форматом ";;;"

Диаграмма с выбором значений

Задача: просматривать данные рядов по отдельности.
Построение таких диаграмм возможно с использованием элементов управления: счетчик, список или полоса прокрутки. Управляя элементами, в дополнительную таблицу копируются значения нужного ряда, на основании которых происходит построение диаграммы.

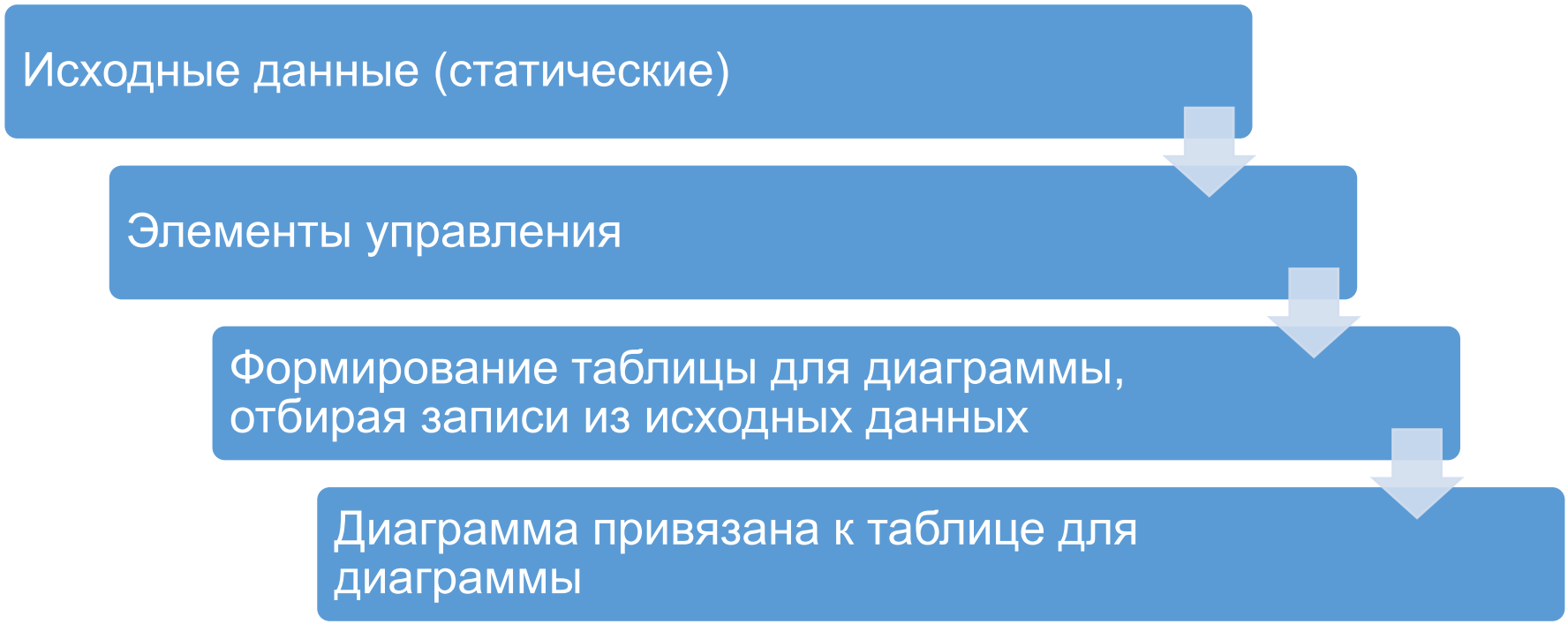


Диаграмма с выбором значений

L3 | =ДВССЫЛ(АДРЕС(СТРОКА(A2);ПОИСКПОЗ($L$2;$B$1:$I$1;0)+1;1;0);ЛОЖЬ)

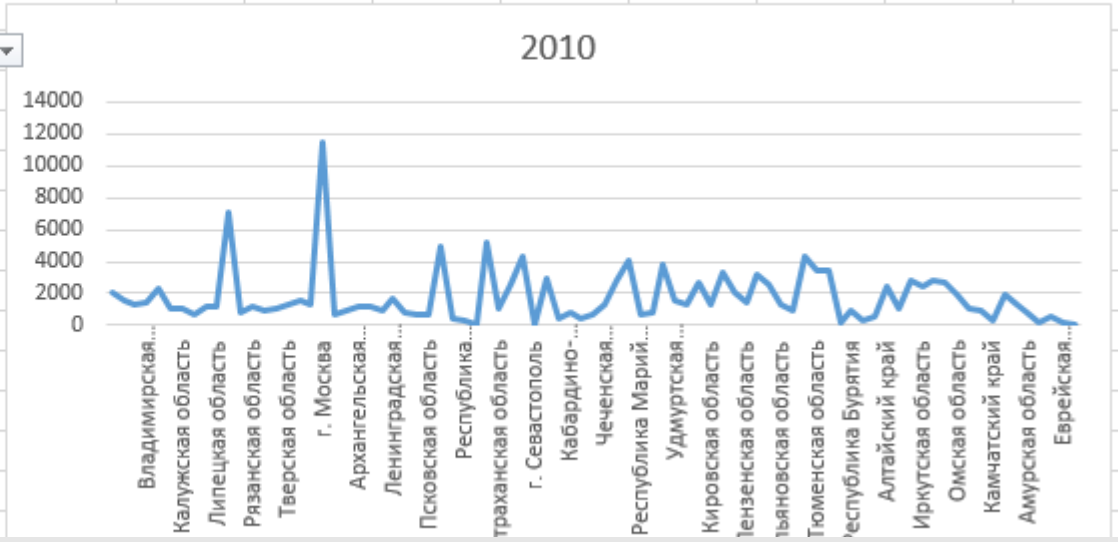
	A	B	C	D	E	F	G	H	I	J	K	L
1	**Область**	**2005**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**			**Год**
2	Белгород	1512	1532	1536	1541	1544	1548	1550	1553		**Область**	2005
3	Брянская	1327	1275	1264	1254	1242	1233	1226	1221		Белгородская область	1512
4	Владимир	1486	1441	1432	1422	1413	1406	1397	1390		Брянская область	1327
5	Воронеж	2361	2335	2332	2330	2329	2331	2333	2335		Владимирская область	1486
6	Ивановск	1102	1060	1054	1049	1043	1037	1030	1023		Воронежская область	2361
7	Калужска	1023	1009	1008	1006	1005	1011	1010	1014		Ивановская область	1102
8	Костромс	700	666	662	659	656	654	651	648		Калужская область	1023
9	Курская о	1178	1126	1122	1119	1119	1117	1120	1123		Костромская область	700
10	Липецкая	1194	1172	1166	1162	1160	1158	1156	1156		Курская область	1178
11	Московск	6784	7106	7199	7048	7134	7231	7319	7423		Липецкая область	1194
12	Орловска	822	786	781	776	770	765	760	755		Московская область	6784
13	Рязанская	1189	1152	1148	1144	1141	1135	1130	1127		Орловская область	822
14	Смоленск	1025	983	981	975	968	965	959	953		Рязанская область	1189

=ДВССЫЛ(АДРЕС(СТРОКА(A2);ПОИСКПОЗ($L$2;$B$1:$I$1;0)+1;1;0);ЛОЖЬ)

- тест ссылки – в ссылку
- номер строки для ссылки
- номер столбца по значению года
- смещение
- стиль ссылки R1C1

Диаграмма с выбором значений

Область	Год 2010
Белгородская область	1532
Брянская область	1275
Владимирская область	1441
Воронежская область	2335
Ивановская область	1060
Калужская область	1009
Костромская область	666
Курская область	1126
Липецкая область	1172
Московская область	7106
Орловская область	786
Рязанская область	1152



Область	Год 2015
Белгородская область	1550
Брянская область	1226
Владимирская область	1397
Воронежская область	2333
Ивановская область	1030
Калужская область	1010
Костромская область	651
Курская область	1120
Липецкая область	1156
Московская область	7319
Орловская область	760
Рязанская область	1130
Смоленская область	959

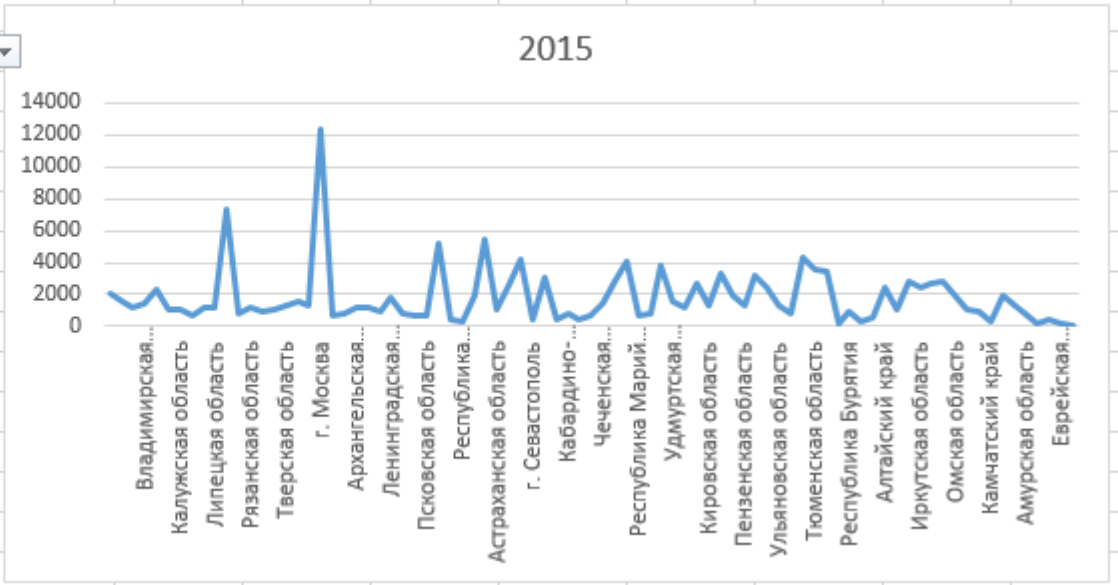


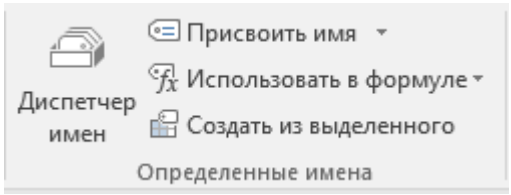
Диаграмма с зумом и прокруткой

Задача: сделать масштабируемой ось или обе оси координат.

Решение:

- Создается именованный диапазон с использованием функции СМЕЩ()
- СМЕЩ(ссылка;смещ_по_строкам;смещ_по_столбцам;[высота];[ширина])
  - Ссылка, от которой вычисляется смещение.
  - Смещ_по_строкам. Количество строк, которые требуется отсчитать вверх или вниз, чтобы левая верхняя ячейка результата ссылалась на нужную ячейку.
  - Смещ_по_столбцам. Количество столбцов, которые требуется отсчитать влево или вправо, чтобы левая верхняя ячейка результата ссылалась на нужную ячейку.
  - Высота (число строк) возвращаемой ссылки.
  - Ширина (число столбцов) возвращаемой ссылки.
- В данном случае создается управление параметрами Высота и/или Ширина для того, чтобы изменять границы диапазона

Диаграмма с зумом и прокруткой



Создаем именованный диапазон для оси X.
В данном случае четвертому аргументу **Высота** назначена ссылка на ячейку, значением которой будет управлять элемент формы.

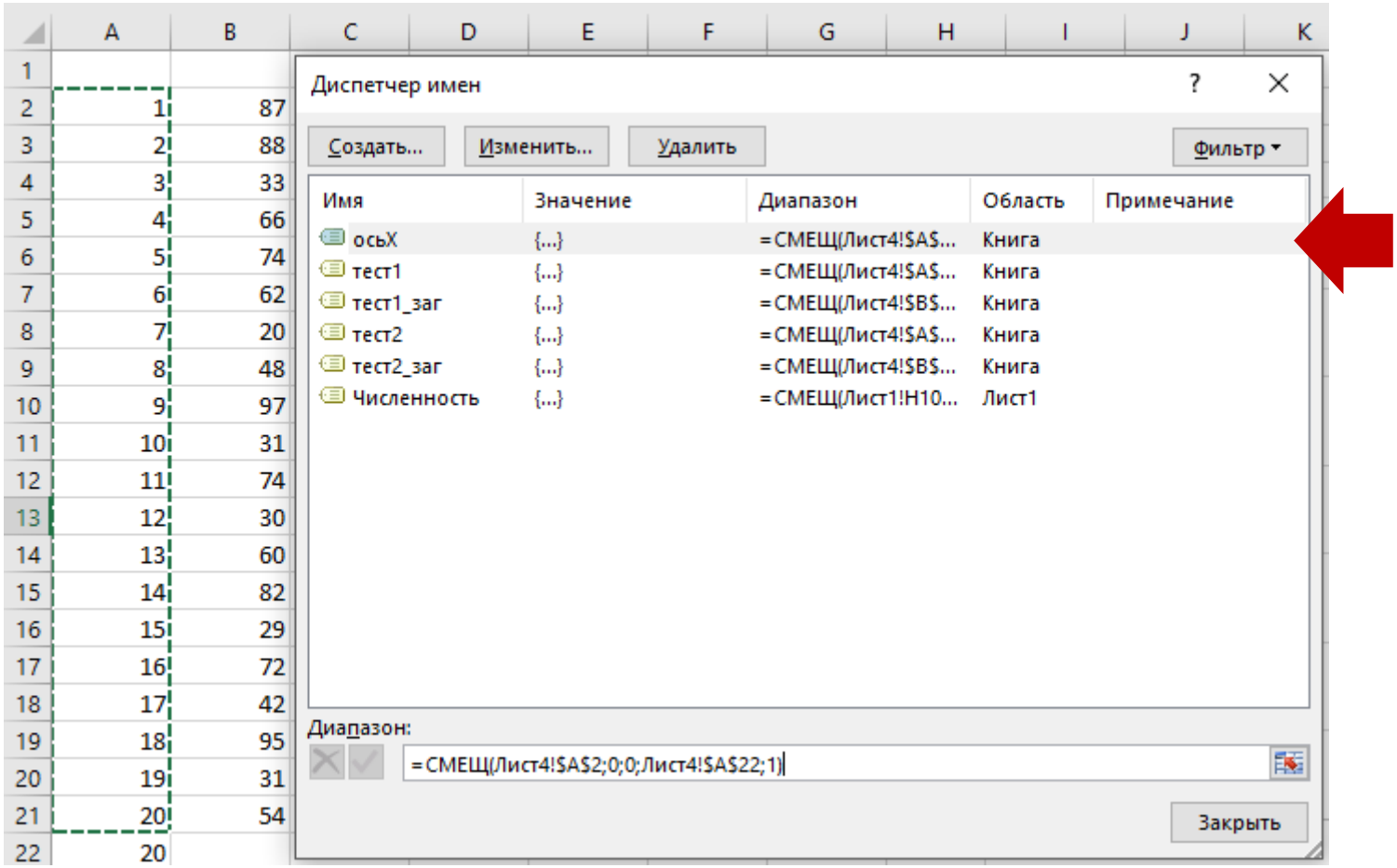
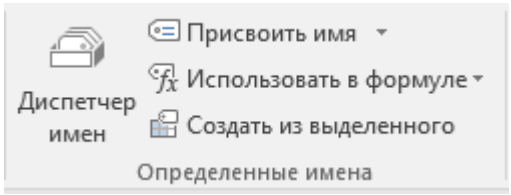


Диаграмма с зумом и прокруткой



Создаем именованный диапазон для оси Y.
В данном случае четвертому аргументу **Высота** назначена ссылка на ячейку, значением которой будет управлять элемент формы.

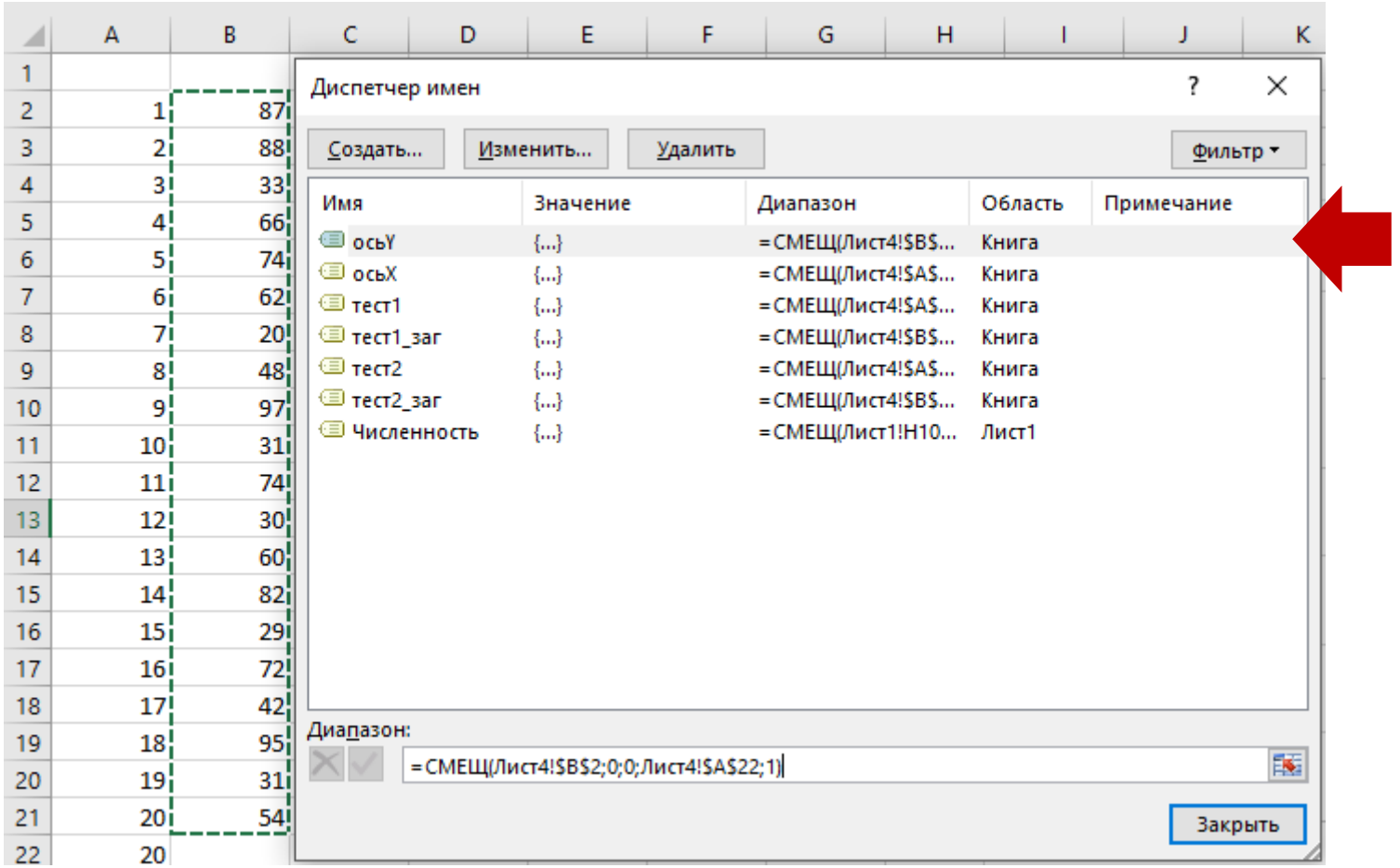
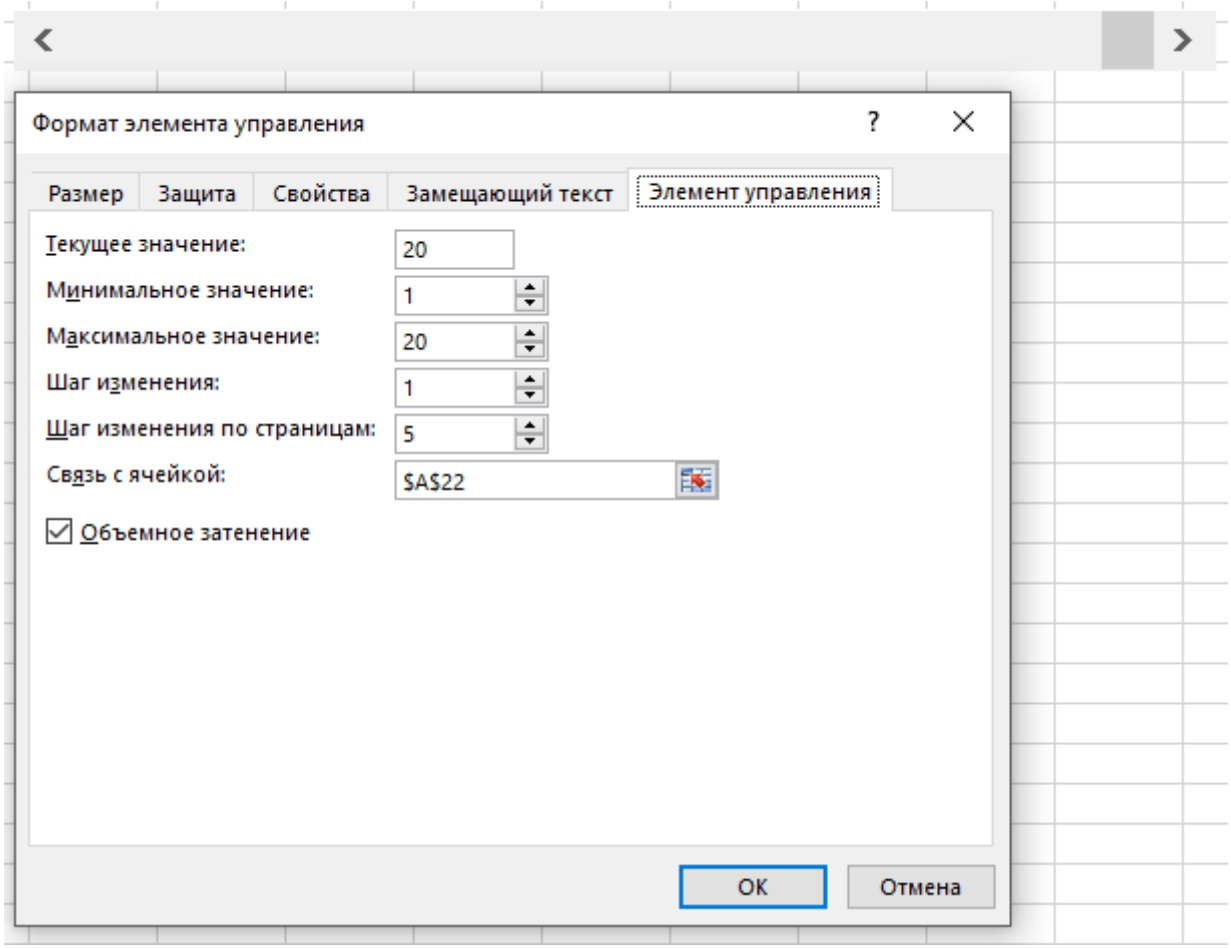


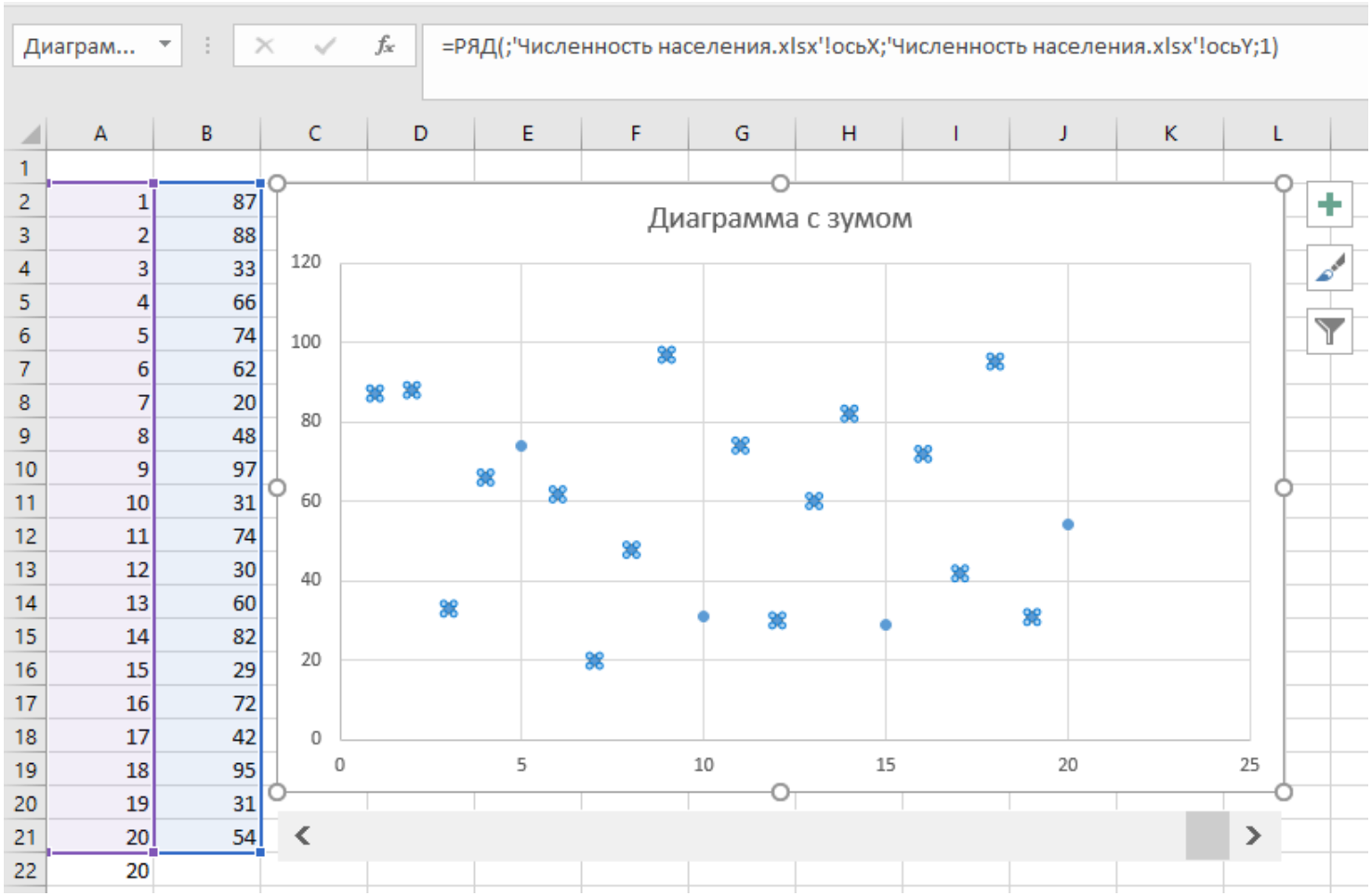
Диаграмма с зумом и прокруткой



Элементу Полоса прокрутки назначаются максимальные и минимальные значения, которые вычисляются из исходных данных.

В данном случае Полоса прокрутки будет управлять высотой отбираемого диапазона значений.

Диаграмма с зумом и прокруткой



	A	B
2	1	87
3	2	88
4	3	33
5	4	66
6	5	74
7	6	62
8	7	20
9	8	48
10	9	97
11	10	31
12	11	74
13	12	30
14	13	60
15	14	82
16	15	29
17	16	72
18	17	42
19	18	95
20	19	31
21	20	54
22	20	

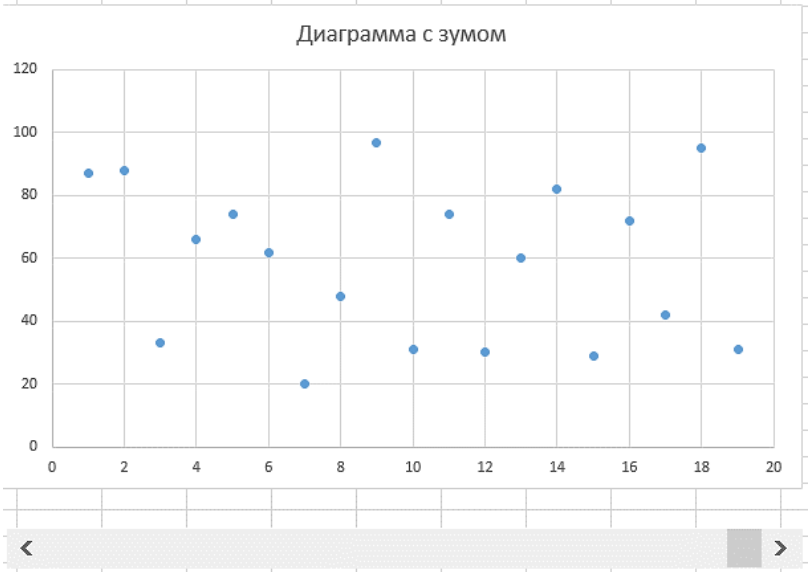
Строим по диапазону ячеек точечную диаграмму, выделяем ряд данных и в строке данных для ряда меняем диапазоны ячеек на ссылки на именованные диапазоны

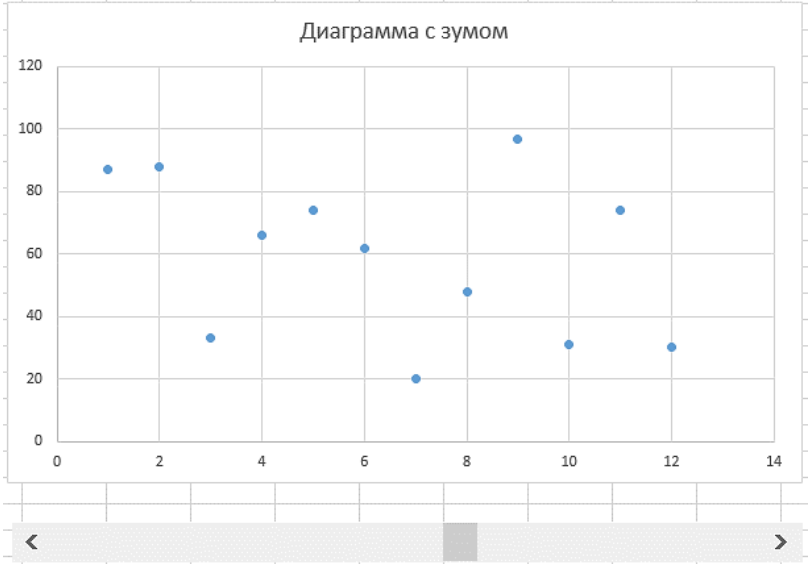
Диаграмма с зумом и прокруткой

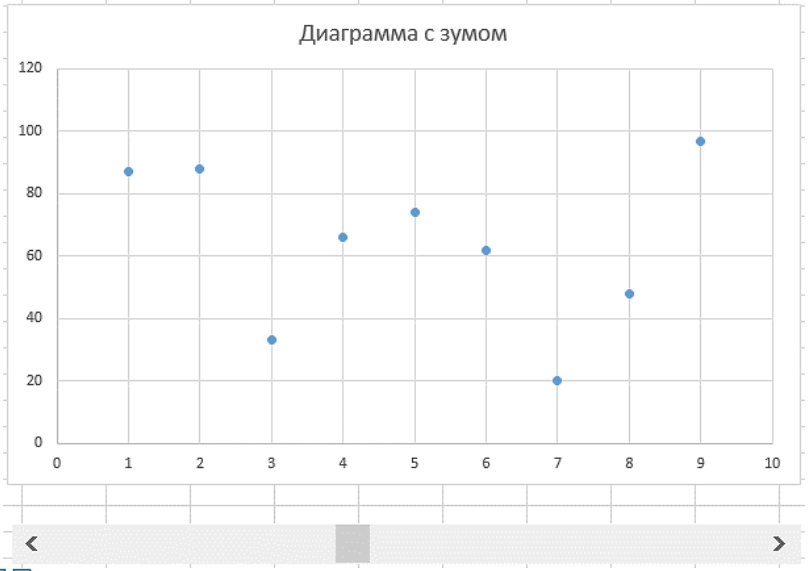
=РЯД(имя;Подписи_категорий(ось_X);Значения(ось_Y);Порядок)

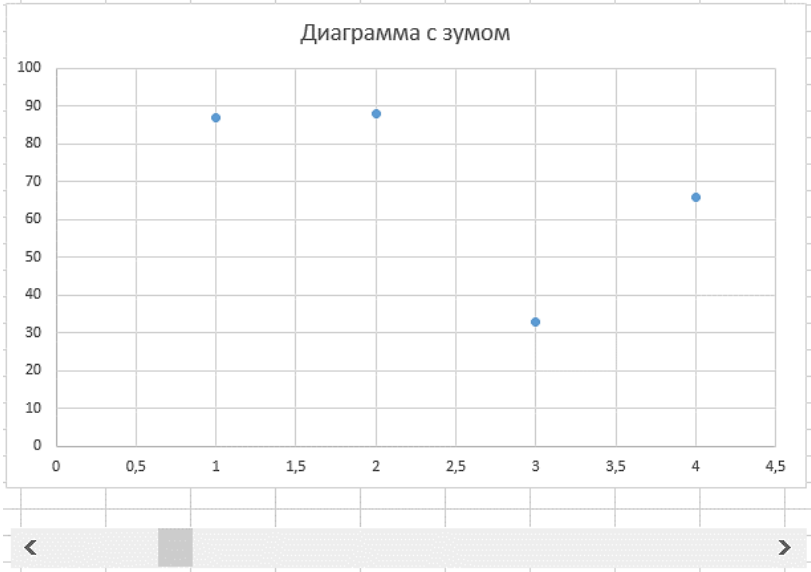
АРГУМЕНТ	ОБЯЗАТЕЛЬНЫЙ/ НЕ ОБЯЗАТЕЛЬНЫЙ	ОПРЕДЕЛЕНИЕ
Имя	Не обязательный	Имя ряда данных, которое отображается в легенде
Подписи_ категорий	Не обязательный	Подписи, которые появляются на оси категорий (если не указано, Excel использует последовательные целые числа в качестве меток)
Значения	Обязательный	Значения, используемые для построения диаграммы
Порядок	Обязательный	Порядок ряда данных

Диаграмма с зумом и прокруткой

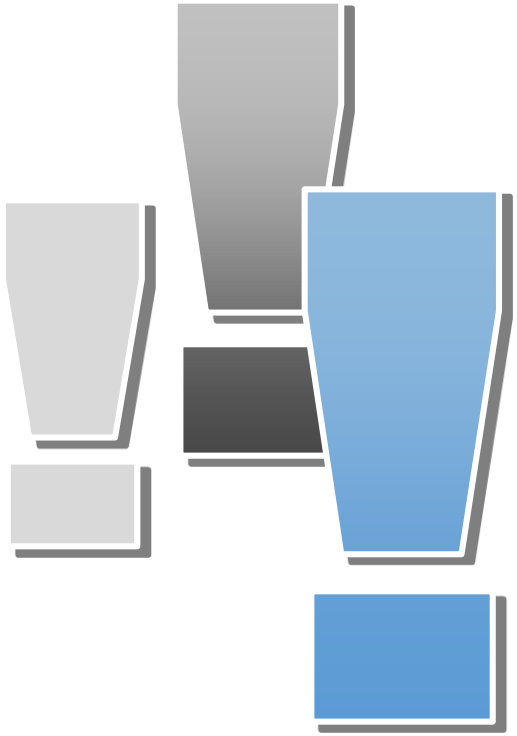








Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57