Семинар 2. SPSS
Описательные статистики

Арина Кузьмичева
ТГ – @fevrier_rin
Почта – adkuzmicheva@hse.ru

Зачем нам описательные статистики?

- Описание изучаемой группы объектов по некоторому признаку или признакам;
- Описание среднего уровня и разброса интересующего признака в некоторой совокупности;

Частотное распределение

1) **Какой вопрос был задан респондентам, сколько было опрошено и сколько на этот вопрос ответило.**

Опрошенные – это все респонденты, участвовавшие в опросе;

Ответившие – это те, кто ответил на вопрос, причем определенным образом («затрудняюсь ответить»?)

Не ответившие - это те:

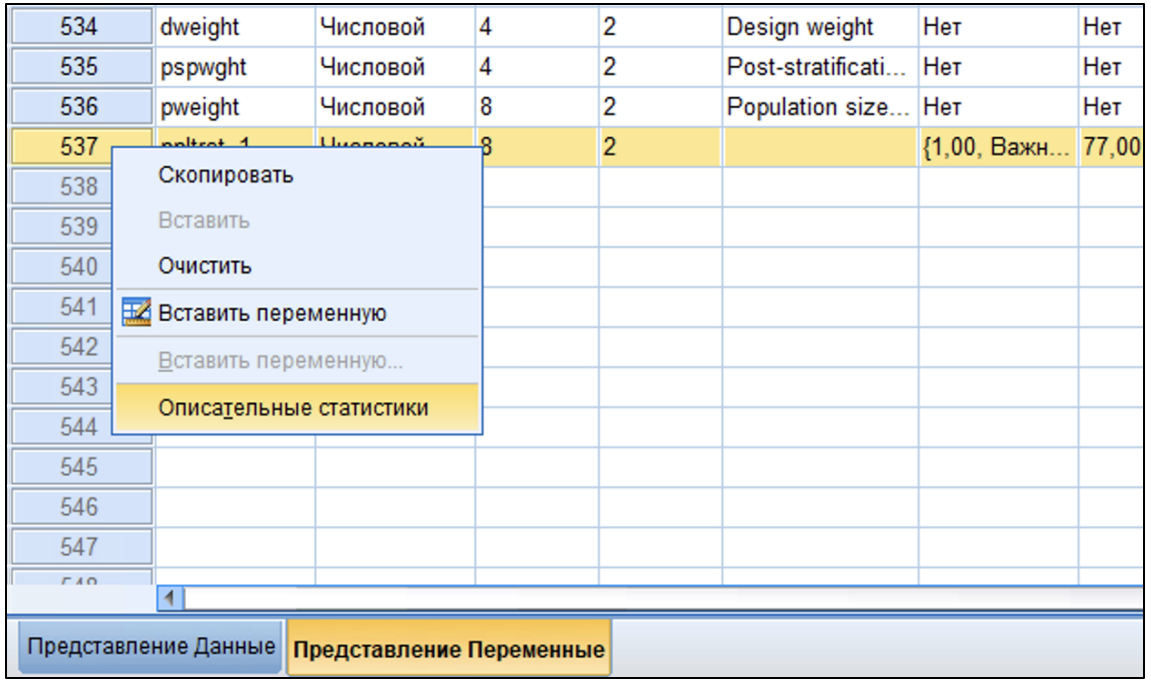
- Кто не стал отвечать на этот вопрос, пропустил его;
- Кому этот вопрос не задавался (например, вопрос о том, где респондент работает, может задаваться только тем респондентам, которые указали, что являются работающими);
- Кто выбрал в качестве ответа неподходящий для нас как исследователей вариант ответа. Чаще всего (но не всегда!), таким вариантом бывает «не знаю», «затрудняюсь ответить»

Частотное распределение

2) **Описываем особенности распределения**

Указываем, какие варианты выбирались чаще всего, а какие реже всего (приводим %);

Также указываем доли выбравших интересующие нас варианты



Перенос на генеральную совокупность

Как правило, когда мы имеем дело с **выборкой**, целью исследования является **получение выводов по всей генеральной совокупности**.

Значение, которое мы получаем на выборке (выборочное значение), может быть подвержено смещениям из-за стат. погрешностей и ошибок выборки.

Чтобы учесть влияние разного рода ошибок и получить истинное значение, мы строим **доверительные интервалы** (ДИ) для доли каждого значения переменной (признака).

Доверительные интервалы – 1

Рассчитать выборочное значение оцениваемого параметра (P – выборочная оценка доли (внимание, опрошенные или ответившие))

Выбрать доверительную вероятность – ту вероятность, с которой генеральное значение параметра попадет в границы интервала, и делают поправку на эту вероятность. Обычно выбирают 90%, 95% и 99% доверительную вероятность. (t - поправка, отражающая доверительную вероятность – точка распределения Стьюдента).

Величина поправки:

90%: t =1,64

95%: t = 1,96

99%: t = 2,58

P.S. При 99% доверительной вероятности доверительный интервал шире
При 90% доверительной вероятности доверительный интервал уже

Доверительные интервалы – 2

Рассчитать стандартную ошибку S.E. (Δ)

S.E. =√ P*(1-P)/N

N – число опрошенных/ответивших (внимательно смотрите, что требует задание/исследовательская задача)

Подставить рассчитанные значения в формулу и получить границы доверительного интервала.

Нижняя граница доверительного интервала = P – t*S.E.

Верхняя граница доверительного интервала = P + t*S.E.

«С вероятностью 95% в генеральной совокупности доля выбравших вариант ответа 1 находится в интервале от… до….»

Для экономии времени используйте файл «Считалка для ДИ»

Доверительные интервалы. Пример

Найдите в массиве переменную prtvtdru (строка 49) и рассчитайте доверительный интервал для ЕР, КПРФ и ЛДПР при 95% доверительной вероятности от ответивших.

Доверительные интервалы. Пример

Найдите в массиве переменную prtvtdru (строка 49) и рассчитайте доверительный интервал для ЕР, КПРФ и ЛДПР при 95% доверительной вероятности от ответивших.

Party voted for in last national election, Russian Federation

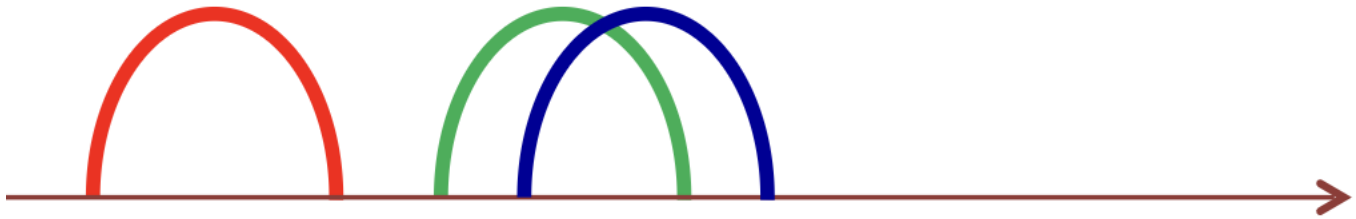
		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	United Russia (ER)	736	1,7	62,7	62,7
	Communist Party of the Russian Federation (KPRF)	139	,3	11,8	74,6
	LDPR	133	,3	11,3	94,2
	Всего	1173	2,6	100,0	

«С вероятностью 95% в генеральной совокупности доля выбравших ЕР находится в интервале от 59,9% до 65,5%»

ПРОЦЕНТ	ВЫБОРОЧНАЯ ДОЛЯ	S.E.	N (опр/отв)	t(95%)	нижняя гран	верхняя гран
62,7	0,627	0,01412014[illegible]6	1173	1,96	59,9%	65,5%
11,8	0,118	0,00941946[illegible]6	1173	1,96	10,0%	13,6%
11,3	0,113	0,00924383[illegible]2	1173	1,96	9,5%	13,1%

Пересечение ДИ

Если границы доверительных интервалов для долей разных значений пересекаются – это значит, что на выборке есть различия между долями, а вот **на генеральной совокупности доли следует считать равными**. Мы имеем право говорить о значимой разнице между долями только тогда, когда ДИ не пересекаются.



Доверительные интервалы

Чем больше размер выборки (N), тем точнее наши оценки (стандартная ошибка S.E. становится меньше) → тем уже доверительный интервал (границы близки друг к другу)

https://www.nejm.org/doi/full/10.1056/NEJMc2104974

https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00947-8/fulltext

Страна	Катар	Израиль
Эффективность	97,4 %	97,5 %
Число прививок	265 410	4 714 932
Число критических случаев	3	364
Доверительный интервал	92,2%–99,5%	97,1%–97,8%

Меры центральной/средней тенденции

Выбор одного числа, которое наилучшим образом описывает все значения признака по изучаемой группе.
Такое число называют центром, типическим значением для набора данных, мерой центральной тенденции, мерой среднего, мерой средней тенденции.

	Мода	Медиана	Среднее арифметическое
Номинальная	ДА	НЕТ	НЕТ
Порядковая	ДА	ДА	НЕТ
Интервальная	ДА	ДА	ДА
Дихотомическая	ДА	ДА	ДА

Мода

Наиболее часто встречающееся значение в выборке, наборе данных (имеет наибольшую частоту).

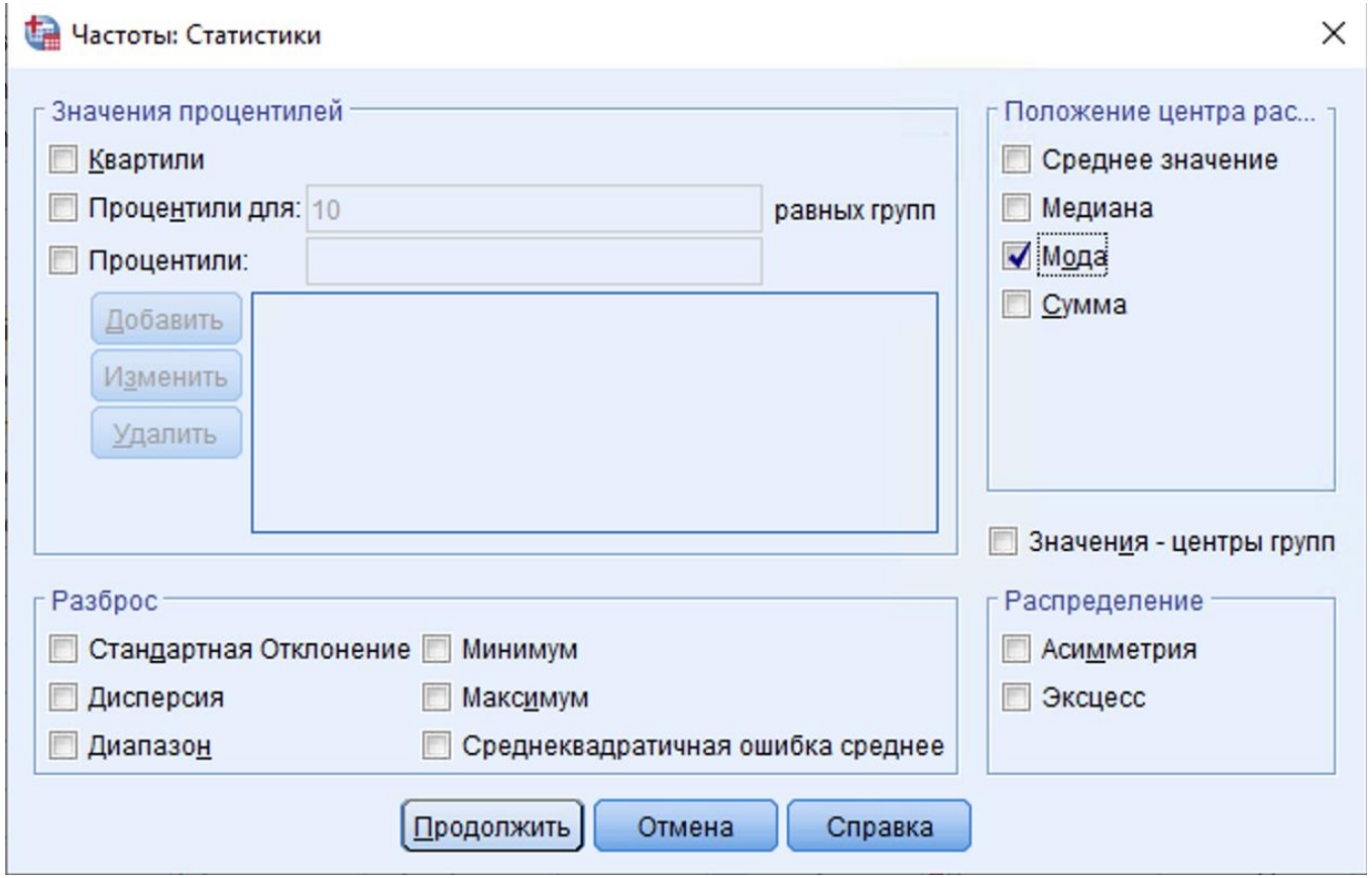
В распределении может быть:

- 1 модальное значение – унимодальное распределение
- 2 модальных значения – бимодальное распределение
- Или несколько значений – мультимодальное распределение

Может быть и так, что все значения признака встречаются одинаковое количество раз – тогда моды нет вообще и распределение равномерное

Мода

Анализ → Описательные статистики → Частоты → Выбираете нужные для анализа переменные → Раздел «Статистики» (отмечаете нужные параметры)



Статистика

How interested in politics

N	Валидные	44290
	Пропущенные	97
Мода		2

How interested in politics

		Частота	Проценты
Валидные	Very interested	5415	12,2
	Quite interested	15539	35,0
	Hardly interested	15248	34,4
	Not at all interested	8088	18,2
	Всего	44290	99,8
Пропущенные	Refusal	36	,1
	Don't know	57	,1
	No answer	4	,0
	Всего	97	,2
Всего		44387	100,0

Медиана

Значение, которое делит **вариационный ряд** пополам.
Половина значений – меньше медианы, а половина – больше.

Вариационный ряд - упорядоченные данные, расположенные в порядке возрастания значений признака, либо в порядке убывания.

Если число значений нечетно, медиана равна значению серединного элемента, если четно – среднему арифметическому 2-х серединных значений.

1. Зайдите в тот же раздел, где искали моду.
2. Или найдите то значение признака, для которого накопленный процент в разделе описательных статистик превысит 50%. Он и является медианным.

Статистика

How interested in politics

N	Валидные	44290
	Пропущенные	97
Медиана		3,00

How interested in politics

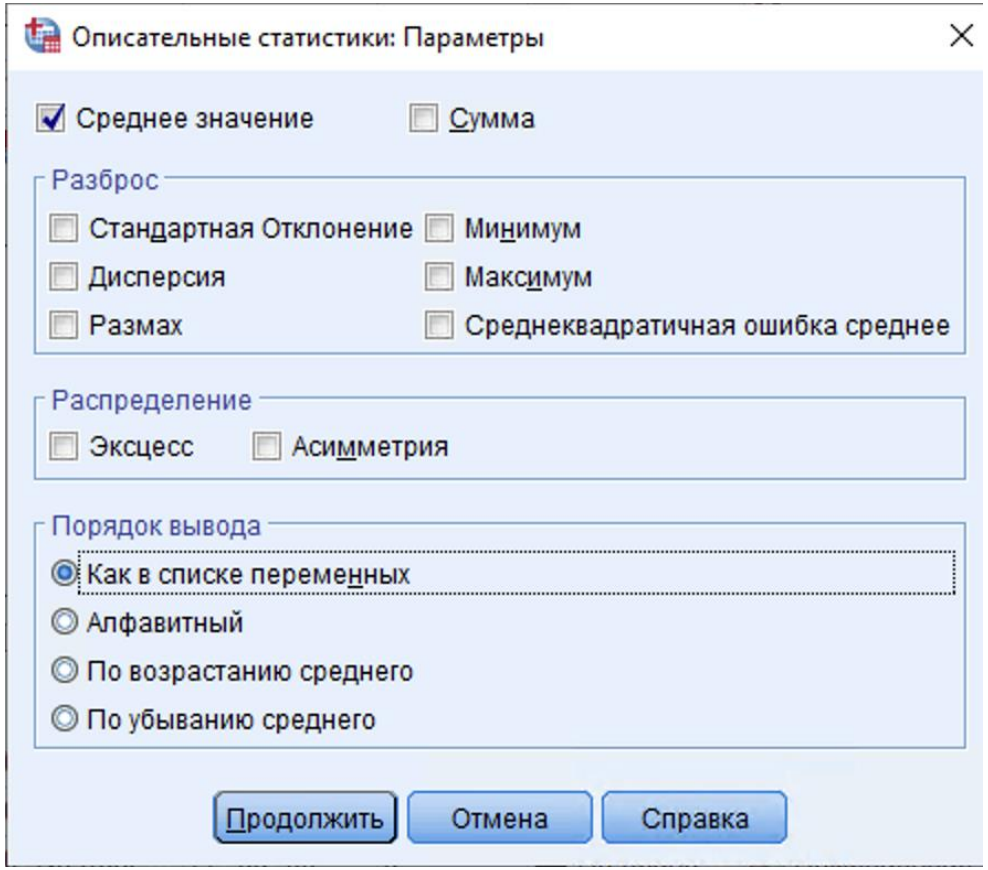
		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	Very interested	5415	12,2	12,2	12,2
	Quite interested	15539	35,0	35,1	47,3
	Hardly interested	15248	34,4	34,4	81,7
	Not at all interested	8088	18,2	18,3	100,0
	Всего	44290	99,8	100,0	
Пропущенные	Refusal	36	,1		
	Don't know	57	,1		
	No answer	4	,0		
	Всего	97	,2		
Всего		44387	100,0		

Среднее арифметическое

Сумма всех значений, деленная на объем (число опрошенных или ответивших!)

1) Анализ → Описательные статистики → Частоты → Выбираете нужные для анализа переменные → Раздел «Статистики» (отмечаете нужные параметры)

2) Анализ → Описательные статистики → Описательные статистики → Выбираете нужные для анализа переменные → Раздел «Параметры» (отмечаете нужные параметры)



Описательные статистики

	N	Среднее
Internet use, how much time on typical day, in minutes	30113	197,63
N валидных (по списку)	30113	

ДИ для среднего

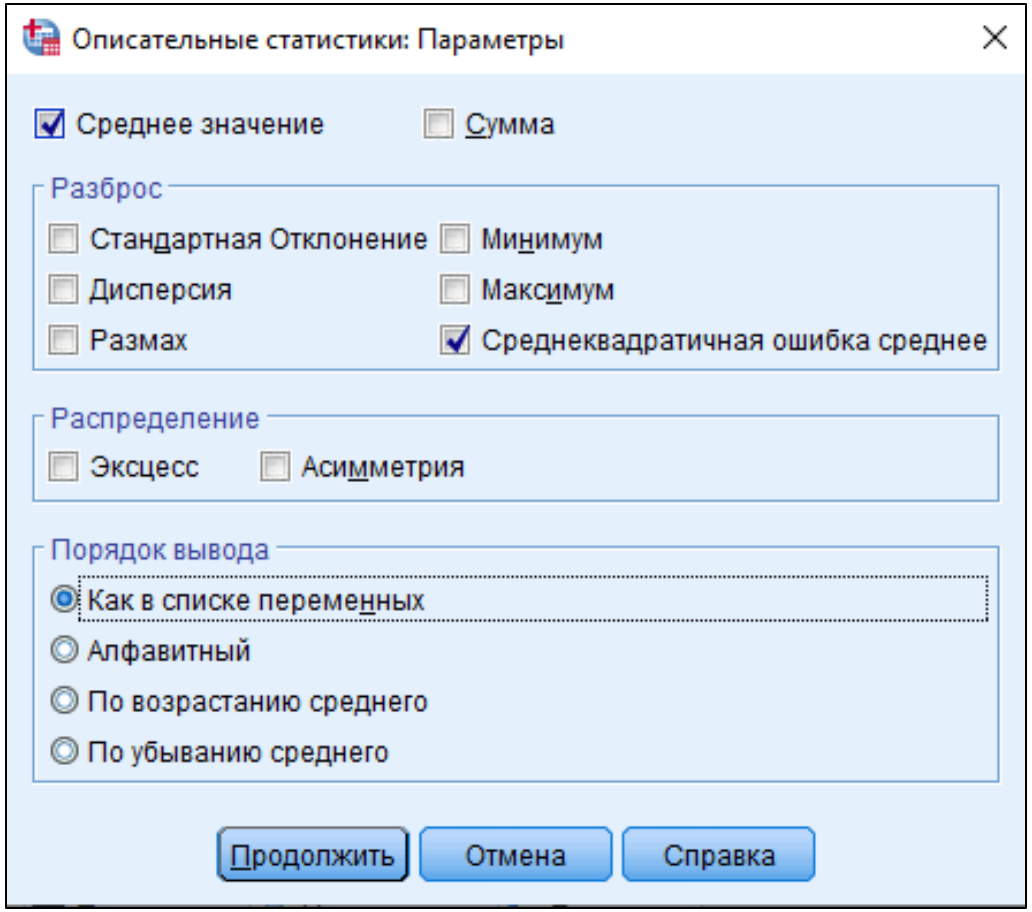
Среднее арифметическое нужно переносить на ГС.

Строим доверительный интервал:
$SE = \sqrt{(S2/N)}$,
где N – число опрошенных или ответивших, а S2 – дисперсия признака

Нижняя граница ДИ: xcp-t*SE
Верхняя граница ДИ: xcp+t*SE

С вероятностью 95% среднее значение лежит в интервале от... до...

SE в SPSS высчитывается автоматически (вам нужно только подставить значения в формулу...)



ДИ для среднего. Пример

Найдите в массиве переменную nwspol (строка 7). Рассчитайте среднее и доверительный интервал для среднего при 90% доверительной вероятности.

Описательные статистики						
	N	Диапазон	Среднее		Среднекв.отклонение	Дисперсия
	Статистика	Статистика	Статистика	Стандартная ошибка	Статистика	Статистика
News about politics and current affairs, watching, reading or listening	43863	1428	85,43	,653	136,799	18713,934
N валидных (по списку)	43863					

Среднее для переменной nwspol – 85,43 (в среднем опрошенные во всех странах тратят на потребление новостей 85,43 мин в день)
Нижняя граница ДИ: 85,43 – 1,64*0,653 = 84,36
Верхняя граница ДИ: 85,43 + 1,64*0,653 = 86,5

С вероятностью 90% среднее значение лежит в интервале от 84,36 до 86,5 мин.

Среднее для дихотомической шкалы

Если два значения признака кодируются 0 и 1, то среднее указывает долю (относительную частоту) единиц в выборке.

Пример: 1, 0, 0, 0, 1, 1, 1, 1, 1, 0
Среднее равно 0,6. То есть 60% значений выборки принимают значение, равное единице. Далее мы смотрим, что закодировано 1, и говорим, что доля этих лиц у нас составляет 60%.

С осторожностью рассчитывайте среднее для дихотомической шкалы в SPSS!
1) Переменные часто закодированы как 1 и 2 (а не 0 и 1)

2) В переменных часто указываются варианты 9, 99 и подобные (которые не должны быть учтены при расчете среднего)

Меры разброса

Меры разброса/меры однородности дают информацию о единодушии/однородности наших объектов.

Если все респонденты выбрали один и тот же вариант ответа – выборка однородна с точки зрения этого признака, если все выбрали разные ответы – выборка неоднородна.

	Коэффициент качественной вариации	Коэффициент вариации	Квартильный размах	Дисперсия/ стандартное отклонение
Номинальная	ДА	НЕТ	НЕТ	НЕТ
Порядковая	НЕТ	НЕТ	ДА	НЕТ
Интервальная	НЕТ	ДА	ДА	ДА
Дихотомическая	НЕТ	ДА	ДА	ДА

Коэффициент вариации / Коэффициент качественной вариации

Коэффициент вариации для интервальных переменных

$$CV = \frac{\text{Среднеквадрат.Отклонение}}{\text{Среднее арифметическое}} * 100$$

Подходит для интервальных шкал
Принимает значение от 0 до 100%, интерпретируется как «на сколько процентов данные отклоняются от среднего»
Но нужно быть аккуратными с интерпретацией

Коэффициент качественной вариации для номинальных переменных

$$K = (1 - \Sigma pi^2) / (1 - 1/k)$$

где:

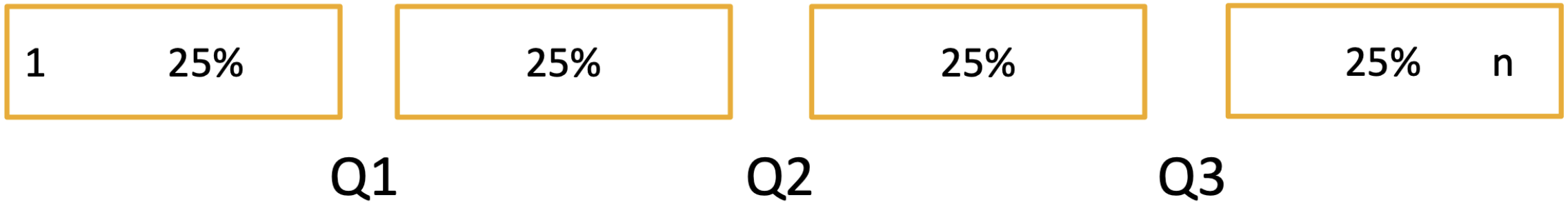
- pi - доля объектов в i-й категории
- k - число категорий

Подходит для номинальных шкал
Принимает значение от 0 до 1, где:
0 – максимально однородные данные
1 – максимально разнородные данные
Если получилось отрицательное значение, или значение, превышающее 1, значит, в расчете есть ошибка

В SPSS, к сожалению, нужно считать вручную (отдельной удобной команды нет)

Квартили

Точки, которые делят вариационный ряд на 4 равно наполненных интервала.



Квартильный размах

Используется для всех шкал кроме номинальной!
Квартильный размах – это разность между третьим и первым квартилем.

$$QR = Q3 - Q1$$

Значения квартилей находятся по накопленной частоте **(раздел описательные статистики):**
Q1 – это то значение признака, для которого накопленная частота превышает 25%
Q3 – это то значение, для которого накопленная частота превышает 75%

Минимальный возможный размах всегда = 0 (если первый и третий квартиль попали в одну и ту же категорию). Чем ближе полученный результат к 0, тем меньше QR;
Максимально возможный квартильный размах получается тогда, когда первый и третий квартиль попадают в крайние категории;

Интерпретация величины QR зависит от балльности шкалы. Чем ближе полученный QR к максимальному для этой шкалы, тем больше разброс, тем менее однородны данные (например, QR = 2 будет максимальным для 3-балльной шкалы, но слабым для 7-балльной шкалы).

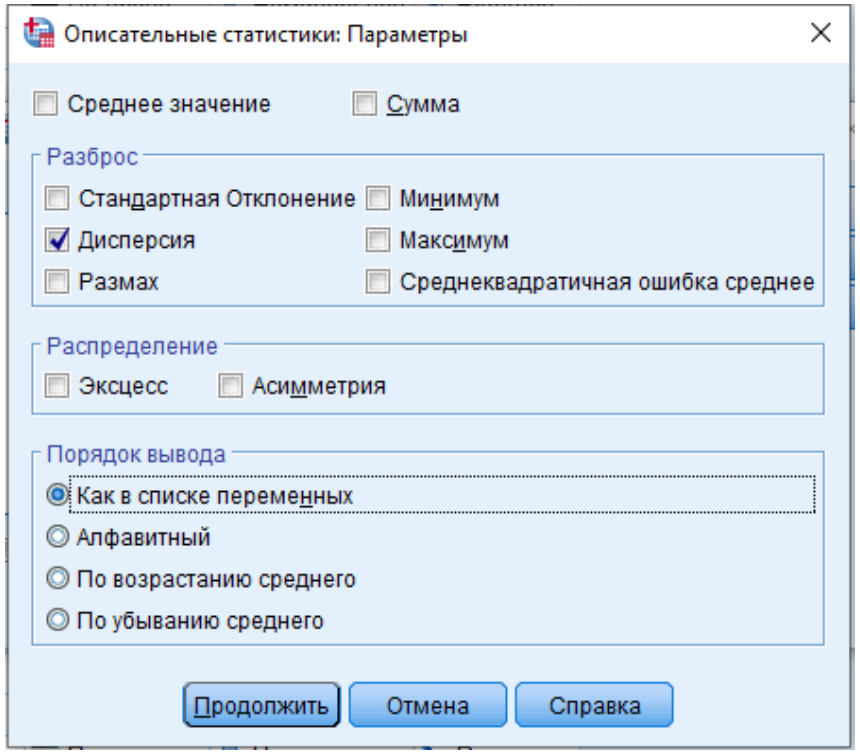
Дисперсия

Используется только для интервальной и дихотомической шкалы.

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

$$S^2 = p * (1 - p)$$

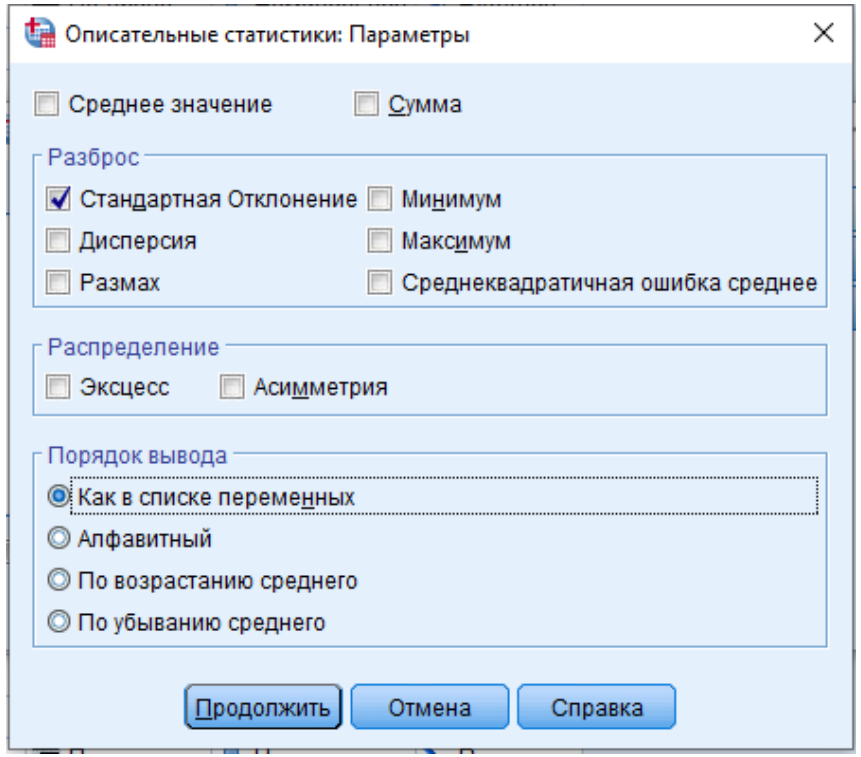
Дисперсия, равная 0, означает отсутствие разброса и полную однородность данных. Чем больше дисперсия, тем более разнородные данные. Максимально возможная дисперсия зависит от диапазона значений шкалы.
Для дихотомической шкалы максимально возможная дисперсия = 0,25



Стандартное отклонение

$$\sqrt{S^2}$$

Нагляднее дисперсии, имеет размерность.
Показывает разброс данных от их среднего значения.
Чем ниже стандартное отклонение, тем более сконцентрированы данные вокруг среднего; чем выше — тем сильнее разброс и больше вариативность значений.



Что может быть на проверочной

1. Опишите особенности распределения переменной X с переносом на генеральную совокупность.
2. Рассчитайте все возможные для данного типа шкалы меры среднего и проинтерпретируйте их.
3. Рассчитайте все возможные для данного типа шкалы меры разброса и проинтерпретируйте их.