Анализ одномерных распределений

Построение частотных распределений

- *Анализ частотных распределений* результатов количественного социологического исследования — это первый шаг при обработке собранной информации
- Выполняет функции получения общих представлений об изучаемых социальных группах
- Сжатие исходной информации, компактного ее представления для дальнейшего осмысления

Методы одномерного описательного анализа

- построение частотных распределений;
- графическое представление поведения анализируемой переменной
- получение статистических характеристик распределения анализируемой переменной

Частотное распределение

- **Частотное распределение** – это упорядоченный подсчет количества признаков по каждому значению какой-

Код	Значение	Частота (число случаев)	Процент (опрошенные)	Валидные процент (ответившие)	Накопленный процент
1	«Синие воротнички»	25	25%	25%	25
2	«Белые воротнички»	23	23%	23%	48
3	Специалисты	22	22%	22%	70
4	Фермеры	20	20%	20%	90
5	Безработные	10	10%	10%	100
	Всего:	**100**	**100%**	**100%**	

Валидные проценты

Статистика

Учет мнения граждан при принятии в

N	Валидные	600
	Пропущенные	0

Учет мнения граждан при принятии внешнеполитических решений

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	затрудняюсь ответить	77	12,8	12,8	12,8
	должен учитывать	422	70,3	70,3	83,2
	не должен учитывать	101	16,8	16,8	100,0
	Всего	600	100,0	100,0	

Статистика

Учет мнения граждан при принятии в

N	Валидные	523
	Пропущенные	77

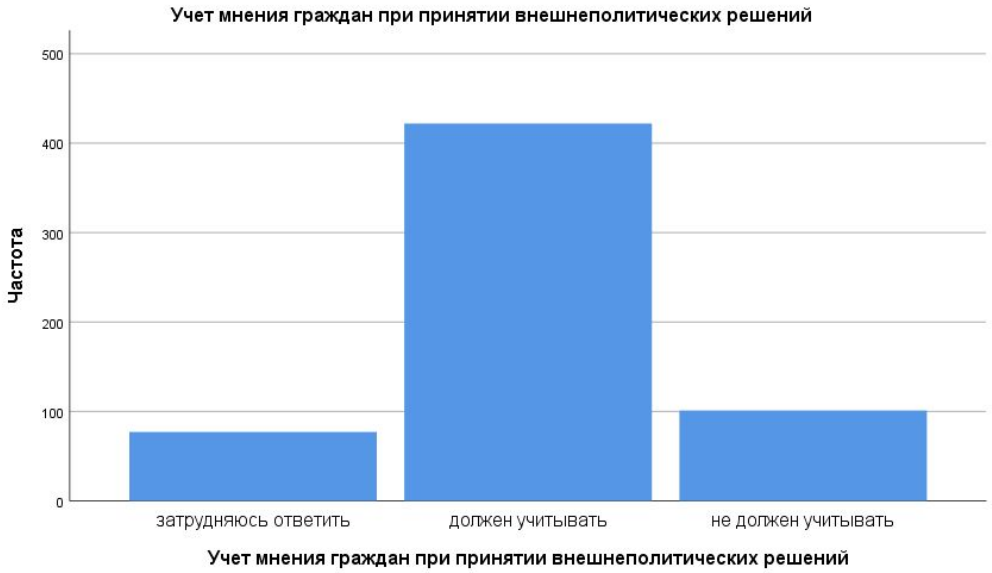
Учет мнения граждан при принятии внешнеполитических решений

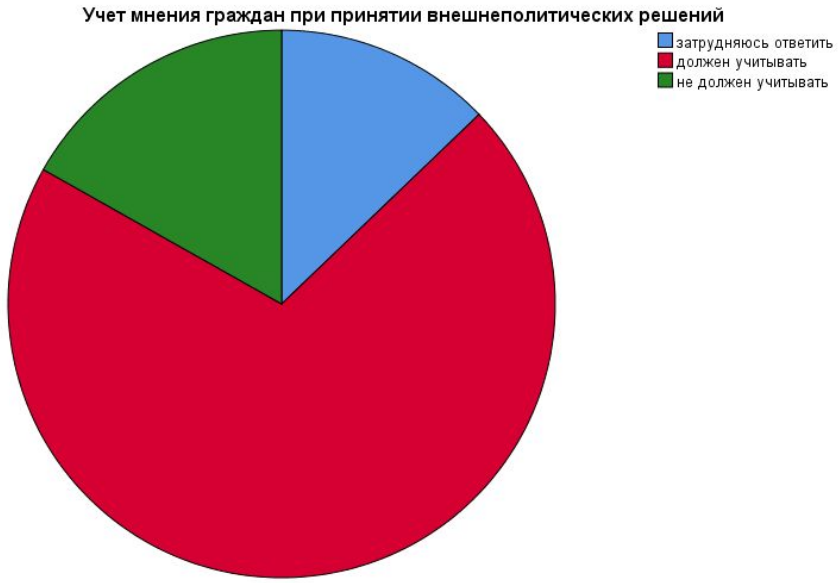
		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	должен учитывать	422	70,3	80,7	80,7
	не должен учитывать	101	16,8	19,3	100,0
	Всего	523	87,2	100,0	
Пропущенные	затрудняюсь ответить	77	12,8		
Всего		600	100,0		

Графическое представление поведения анализируемой переменной

Учет мнения граждан при принятии внешнеполитических решений

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	затрудняюсь ответить	77	12,8	12,8	12,8
	должен учитывать	422	70,3	70,3	83,2
	не должен учитывать	101	16,8	16,8	100,0
	Всего	600	100,0	100,0	





Меры центральной тенденции и меры разброса

- *Меры центральной тенденции –* статистики, описывающие, где находятся наиболее типичные значения (мода, медиана, среднее арифметическое)
- *Меры разброса -* статистики, описывающие вариабельность значений признака (дисперсия, стандартное отклонение, размах, квартильный размах).

Среднее арифметическое и дисперсия

- статистические расчеты должны соответствовать уровню измерений данных

Номинальный уровень измерений	Порядковый уровень измерений	Интервальный уровень измерений
Мода	Медиана	Среднее арифметическое
Коэффициент вариации	Квартильный размах	Стандартное отклонение

Измерения для номинальных переменных

- *Мода* – это наиболее часто встречающееся значение признака в серии зарегистрированных наблюдений

«Синие воротнички»	25
«Белые воротнички»	23
Специалисты	22
Фермеры	20
Безработные	10

Мода = «синие воротнички»

- *Коэффициент вариации*

$$v = \frac{\Sigma f_{немодальное}}{N}$$

где $\Sigma f_{немодальное}$ – сумма всех случаев, *не* входящих в модальную категорию; N – общее число случаев

$\vartheta = 23+22+20+10 / 100 = 75/100 = 0{,}75$

Значение коэффициента вариации колеблется между 0 (когда все случаи принимают одно и то же значение) и $1–1/N$ (когда каждый случай имеет свое значение)

Чем меньше коэффициент вариации, тем типичнее, или значимее (верно отражает картину), мода

Упражнение: определите моду и коэффициент вариации

Голосование за партию

		Частота	Процент	Валидный процент	Накопленный процент
Валидные	"Единая Россия"	77	38,5	38,5	38,5
	КПРФ	13	6,5	6,5	45,0
	ЛДПР	24	12,0	12,0	57,0
	"Партия Роста"	2	1,0	1,0	58,0
	"Патриоты России"	1	,5	,5	58,5
	"Спаведливая Россия"	8	4,0	4,0	62,5
	"Яблоко"	3	1,5	1,5	64,0
	Другая партия	3	1,5	1,5	65,5
	не пошел(ла) бы на выборы	29	14,5	14,5	80,0
	Не проголосовал(ла) бы ни за одну партию	23	11,5	11,5	91,5
	затрудняюсь ответить	17	8,5	8,5	100,0
	Итого	200	100,0	100,0	

Измерения для порядковых переменных

- ***Медиана*** - это такое значение признака, которое разделяет ранжированный ряд распределения на две равные части - со значениями признака меньше медианы и со значениями признака больше медианы

Снижение поддержки населением проводимой политик

		Частота	Процент	Валидный процент	Накопленный процент
Валидные	не согласен	43	22,2	22,2	22,2
	скорее не согласен	46	23,7	23,7	45,9
	не могу точно сказать	47	24,2	24,2	70,1
	скорее согласен	44	22,7	22,7	92,8
	согласен	14	7,2	7,2	100,0
	Итого	194	100,0	100,0	

$$№_{Me} = \frac{N+1}{2}$$

- ***Межквартильный размах*** – показывает, насколько плотно различные значения группируются вокруг медианы, или насколько типична или репрезентативна медиана для распредел

Статистики

Снижение поддержки населением проводимой политик

N	Валидные	194
	Пропущенные	0
Медиана		3,00
Процентили	20	1,00
	40	2,00
	60	3,00
	80	4,00

Межквартильный размах = $Q_3 - Q_1$

где q_3 – третий квартиль (значение, ниже которого находится 3/4, или 75% всех признаков) = $\frac{3(n+1)}{4}$

q_1 – первый квартиль (значение, ниже которого находится 1/4 или

Упражнение: определите медиану и межквартильный размах

1. Владельцем квартиры какой стоимости вы являетесь

Стоимость жиль	Частота
Свыше 5 000 000 рублей	3
4 999 999 – 4 000 000 рублей	4
3 999 999 - 3 000 000 рублей	10
2 999 999 – 2 000 000 рублей	15
1 999 999 – 1 000 000 рублей	25
Менее 1000 000 рублей	1

квартира

		Частота	Проценты	Валидный процент	Накопленный процент
Валидные	менее 1000000	1	1,7	1,7	1,7
	1000000-1999999	25	43,1	43,1	44,8
	2000000-29999999	15	25,9	25,9	70,7
	3000000-3999999	10	17,2	17,2	87,9
	4000000-4999999	4	6,9	6,9	94,8
	более 5000000	3	5,2	5,2	100,0
	Всего	58	100,0	100,0	

Статистика

квартира

N	Валидные	58
	Пропущенные	0
Медиана		3.00
Процентили	25	2.00
	50	3.00
	75	4.00

Измерение для интервальных шкал

- *Среднее арифметическое* - величина, полученная путем сложения всех членов числового ряда и деления суммы на число членов

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_{ср})^2}{n-1}}$$

- *Стандартное отклонение* – степень отклонения данных наблюдений от среднего значения.
- *Небольшое* стандартное отклонение указывает на то, что данные группируются вокруг среднего значения
- *Высокое* стандартное отношение указывает, что данные располагаются далеко от него

где X_i – значение каждого отдельного случая;

$\overline{X}$ – среднее геометрическое;

N – количество случаев;

$\sum_{i=1}^{N}$ – знак суммы всех отдельных случаев от 1 до N.

Измерение для интервальных шкал

Возраст респондента

	Валидные																														
	18	19	20	22	26	27	32	34	35	36	38	39	41	44	45	46	47	49	50	51	52	55	56	57	61	62	64	66	70	78	Всего
Частота	1	2	3	1	2	2	3	1	1	2	1	2	1	1	1	1	1	1	5	3	1	2	2	2	2	2	2	1	1	1	51
Проценты	2,0	3,9	5,9	2,0	3,9	3,9	5,9	2,0	2,0	3,9	2,0	3,9	2,0	2,0	2,0	2,0	2,0	2,0	9,8	5,9	2,0	3,9	3,9	3,9	3,9	3,9	3,9	2,0	2,0	2,0	100,0
Валидный процент	2,0	3,9	5,9	2,0	3,9	3,9	5,9	2,0	2,0	3,9	2,0	3,9	2,0	2,0	2,0	2,0	2,0	2,0	9,8	5,9	2,0	3,9	3,9	3,9	3,9	3,9	3,9	2,0	2,0	2,0	100,0
Накопленный процент	2,0	5,9	11,8	13,7	17,6	21,6	27,5	29,4	31,4	35,3	37,3	41,2	43,1	45,1	47,1	49,0	51,0	52,9	62,7	68,6	70,6	74,5	78,4	82,4	86,3	90,2	94,1	96,1	98,0	100,0	

Статистика

Возраст респондента

N	Валидные	51
	Пропущенные	0
Медиана		47,00
Мода		50

Статистика

Возраст респондента

N	Валидные	51
	Пропущенные	0
Среднее		44,08
Стандартная отклонения		15,388

Упражнение: определите среднее арифметическое и стандартное отклонение

Респондент	Доход (мес.)
Респондент 1	25 000
Респондент 2	100 000
Респондент 3	10 000
Респондент 4	75 000
Респондент 5	70 000
Респондент 6	20 000
Респондент 7	19000
Респондент 8	22000
Респондент 9	30000
Респондент 10	35000

χ^2 - критерий Пирсона

Назначения критерия

- для сопоставления *эмпирического* распределения признака с *теоретическим* - равномерным, нормальным или каким-то иным;
- для сопоставления *двух, трех или более эмпирических* распределений одного и того же признака
- Критерий χ^2 отвечает на вопрос о том, с одинаковой ли частотой встречаются разные значения признака в эмпирическом и теоретическом распределениях или в двух и более эмпирических распределениях

χ^2 - критерий Пирсона

Ограничения критерия

- Объем выборки должен быть достаточно большим: $n>30$
- Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5: $f>5$
- Выбранные разряды должны «вычерпывать» все распределение, то есть охватывать весь диапазон вариативности признаков. При этом группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях
- Необходимо вносить «поправку на непрерывность» при сопоставлении распределений признаков, которые принимают всего 2 значения
- Разряды должны быть неперекрещивающимися: если наблюдение отнесено к одному разряду, то оно уже не может быть отнесено ни к какому другому разряду

Статистические гипотезы

- Статистические гипотезы делятся на *нулевые* и *альтернативные*

Нулевая гипотеза - это гипотеза об отсутствии различий

- обозначается как H_0
- Называется нулевой потому, что содержит число 0: $X_1 - X_2 = 0$, где X_1, X_2 – сопоставляемые значения признаков
- Нулевая гипотеза - это то, что мы хотим опровергнуть, если перед нами стоит задача доказать значимость различий

Альтернативная гипотеза - это гипотеза о значимости различий.

- Она обозначается как H1
- Альтернативная гипотеза - это то, что мы хотим доказать, поэтому иногда ее называют экспериментальной гипотезой

χ^2 - критерий Пирсона

Проверяемые статистические гипотезы

- H_0 – эмпирическое распределение признака не отличается от теоретического равномерного распределения
- H_1 – эмпирическое распределение признака отличается от теоретического равномерного распределения

Расчет χ^2 - критерия Пирсона

1. Занести в таблицу наименования разрядов и соответствующие им эмпирические частоты (первый столбец)

2. Рядом с каждой эмпирической частотой записать теоретическую частоту (второй столбец)

3. Подсчитать разности между эмпирической и теоретической частотой по каждому разряду (строке) и записать их в третий столбец

4. Определить число степеней свободы по формуле: ***v=k-1,*** где *k* - количество разрядов признака

Если V=1, внести поправку на «непрерывность» (величины, указанные в столбце 3, уменьшить на 0,5)

5. Возвести в квадрат полученные разности и занести их в четвертый столбец

6. Разделить полученные квадраты разностей на теоретическую частоту и записать результаты в пятый столбец

7. Просуммировать значения пятого столбца. Полученную сумму обозначить как χ^2 эмпирическое

8. Определить по таблице критических знаний критические значения для данного числа степеней свободы V.

- Если χ^2 эмпирическое меньше критического значения, расхождения между распределениями статистически недостоверны.
- Если χ^2 эмпирическое равно критическому значению или превышает его, расхождения между распределениями статистически достоверны.

Расчет χ^2 - критерия Пирсона

Вид спорта	Эмпирическая частота	Теоретическая частота	fэмп – f теор.	(fэмп – f теор.)2	(fэмп – f теор.)2/fтеор.
Хоккей	15	9	6	36	4
Теннис	10	9	4	16	1,77
Футбол	9	9	0	0	0
Бокс	2	9	-7	49	5,4
					∑= 11,17

$\boldsymbol{V = 4 - 1 = 3}$; χ^2крит. = 5,991 (p= 0,05); χ^2крит = 11,345 (p = 0,001)

χ^2эмп > χ^2крит

Принимается гипотеза H1

Уровни статистической значимости

- Уровень значимости - это вероятность того, что мы сочли различия существенными, а они на самом деле случайны
- Когда мы указываем, что различия достоверны на 5%-ом уровне значимости, или при $p<0,05$, то мы имеем виду, что вероятность того, что они все-таки недостоверны, составляет 0,05.
- Когда мы указываем, что различия достоверны на 1%-ом уровне значимости, или при $p<0,01$, то мы имеем в виду, что вероятность того, что они все-таки недостоверны, составляет 0,01.

Правило отклонения H_0 и принятия H_1

- Если эмпирическое значение критерия равняется критическому значению, соответствующему $p > 0,05$ или превышает его, то H_0 отклоняется, но мы еще не можем определенно принять H_1
- Если эмпирическое значение критерия равняется критическому значению, соответствующему $p<0,01$ или превышает его, то H_0 отклоняется и принимается H_1.
- *Исключения:* критерий знаков G, критерий Т Вилкоксона и критерий U Манна-Уитни. Для них устанавливаются обратные соотношения.

Упражнение: определите, есть ли различия в распределении политических предпочтений россиян

Голосование за партию

		Частота	Процент	Валидный процент	Накопленный процент
Валидные	"Единая Россия"	77	38,5	38,5	38,5
	КПРФ	13	6,5	6,5	45,0
	ЛДПР	24	12,0	12,0	57,0
	"Партия Роста"	2	1,0	1,0	58,0
	"Патриоты России"	1	,5	,5	58,5
	"Спаведливая Россия"	8	4,0	4,0	62,5
	"Яблоко"	3	1,5	1,5	64,0
	Другая партия	3	1,5	1,5	65,5
	не пошел(ла) бы на выборы	29	14,5	14,5	80,0
	Не проголосовал(ла) бы ни за одну партию	23	11,5	11,5	91,5
	затрудняюсь ответить	17	8,5	8,5	100,0
	Итого	200	100,0	100,0	