Генеральная совокупность и выборка

При изучении конкретной задачи всегда предполагается делать выводы относительно какой-то группы людей или объектов. Например, если вы хотите узнать, как женщины старше 18 в России относятся к новой рекламе какого-либо продукта, то этой группой будут абсолютно все женщины, проживающие в России, чей возраст превышает 18 лет. Или, если вам интересен прогресс в обучении у первоклассников, то такой группой будут абсолютно все дети, которые обучаются в школе первый год. Такая группа людей (или каких-то еще объектов) называется генеральной совокупностью.

Генеральная совокупность *— это совокупность всех объектов, которые представляют интерес в конкретном исследовании.*

Конечно, хотелось бы исследовать всю генеральную совокупность. Однако чаще всего это представляется невозможным, так как это слишком долго, дорого и ресурсозатратно. Если нам нужно опросить несколько миллионов людей, то нам нужно очень много сотрудников и времени на сбор данных: пока мы будем всех опрашивать, исследование может даже стать неактуальным. А иногда это просто нереализуемо. Например, если часть людей откажется участвовать в опросе и отвечать на наши вопросы. По этой причине исследуется не вся генеральная совокупность, а лишь ее часть — выборка. То есть, из всей генеральной совокупности мы отбираем только часть объектов (выборку), проводим на них исследование, а получившиеся результаты обобщаем на всю генеральную совокупность.

Выборка *— это группа объектов, отобранных из генеральной совокупности для исследования.*

Но здесь возникает новая проблема. Допустим, что мы хотим опросить всех женщин, чтобы узнать их мнение о новой рекламе косметики. Но в выборку возьмем лишь студенток. Разумеется, наши выводы тогда окажутся неверными, так как мнение женщин другого возраста (например, 30-40 лет) могут отличаться. Или, если мы отберем в выборку только домохозяек, то снова получим ошибочный результат, потому что не учтем взгляды работающих женщин.

Все эти примеры относятся к понятию репрезентативности. Если в нашей генеральной совокупности 40% женщин и 60% мужчин, то и в выборке должно быть такое же соотношение женщин и мужчин. Если в генеральной совокупности треть людей — студенты, то и в выборке мы должны учесть это соотношение. Для того чтобы было выполнено условие репрезентативности, все основные особенности исследуемой группы людей или объектов должны быть одни и те же у выборки и у генеральной совокупности. При соблюдении процедуры такую выборку можно, например, получить с помощью случайного выбора, но сделать его не всегда просто.

Репрезентативность *— соответствие характеристик выборки характеристикам генеральной совокупности.*

Но важно понимать, что даже для репрезентативной выборки будет наблюдаться ошибка выборки — несоответствие между характеристиками генеральной совокупности и характеристиками, полученными для данной выборки.

Предположим, что у нас есть 1000 школьников, которые сдали ЕГЭ. И из этой группы мы выбираем две выборки по 5 человек. Если мы посчитаем средний балл на всей совокупности и средние баллы для двух выборок, то эти три значения практически точно будут отличаться. То есть, характеристики для выборки почти всегда отличаются от характеристик совокупности и могут меняться, если сформировать другую выборку.

Иногда может происходить ***смещение выборки*** — явление, при котором статистические характеристики выборки сильно отличаются (смещены) относительно характеристик генеральной совокупности. Например, такое случается при нерепрезентативной выборке (допустим, психолог изучает особенности восприятия людей в целом, но в качестве респондентов использует только своих студентов), при отсутствии части данных (в социальных опросах люди могут избегать ответов на какие-то вопросы, про религию, деньги и другие личные темы) или вследствие явления социальной желательности (люди не склонны отвечать даже анонимно то, что не вызовет социальное одобрение: даже если все респонденты когда-либо совершали кражи в магазине, они с большой вероятностью не признаются в этом). Получается, что данные, которые мы получаем в таких ситуациях, смещаются относительно того, что происходит в реальности

Частотные распределения и диаграммы

Обычно после сбора данных у нас есть большое количество чисел, которые может быть необходимо представить в более организованной форме. Это поможет сразу отследить какие-то закономерности.

Один из самых простых вариантов представления большого количества чисел — это частотная таблица. Для того чтобы построить частотную таблицу, необходимо выстроить от меньшего к большему все возможные значения, которые встречаются в наших данных. А потом для каждого их них посчитать частоту — количество раз, сколько встречается это значение.

Например, если мы хотим построить распределение оценок, которые получили школьники за текущую четверть по пятибалльной шкале, то у нас будет пять вариантов значений: 1,2,3,4,5. И для каждого значения мы можем посчитать, сколько раз такая оценка была выставлена. По полученному распределению можно сразу оценить основную тенденцию: каких оценок большего всего, часто ли ученики получают неудовлетворительные оценки, какова доля четверок и т.д.

Давайте рассмотрим наш пример с оценками не в теории, а на практике. Пусть у нас есть 25 школьников, которые написали контрольную по программированию и получили разные оценки от 2 до 5 3,2,3,3,5,4,3,2,4,2,3,4,5,2,4,3,4,5,5,4,3,3,2,3,2

Оформим эти вычисления в таблицу и получим непосредственно частотное распределение:

Значение	Частота
2	5
3	10
4	6
5	4

Смотря на такую таблицу, мы сразу можем сделать ряд выводов: например, о том, что контрольная была явно сложная, так как очень мало отличных оценок, достаточно много неудовлетворительных и больше всего оценок «удовлетворительно».

С помощью такого распределения можно так же быстро сделать дополнительные вычисления: для начала мы можем посчитать сумму всех оценок, но не складывая все оценки по очереди, а складывая произведения значения и его частоты: $2 \cdot 5 + 3 \cdot 10 + 4 \cdot 6 + 5 \cdot 4 = 10 + 30 + 24 + 20 = 84$

И так же просто можно посчитать среднее значение, если поделить полученную сумму на количество: $84/25 = 3.36$

Иногда в распределениях считают не абсолютные частоты, а пропорции или проценты. Так проще показать, какую долю занимают те или иные значения. В нашем примере мы видим, что двойку получило 5 человек из 25. Значит, мы можем рассчитать, что это $5/25 = 0.2$ от общего количества,если считать долю. Или $0.2 \cdot 100 = 20\%$, если считать в процентах.

Добавим к нашим значениям в таблицу представление в виде доли и в виде процента:

Значение	Частота	Доля	Процент
2	5	0.2	20%
3	10	0.4	40%
4	6	0.24	24%
5	4	0.16	16%

Если различных значений не очень много, то такое представление достаточно удобно. Но если представлено много разных чисел, то тогда таблица получится очень объемной, и делать по ней выводы будет достаточно сложно. Для таких случаев можно строить распределение частот с группировкой: считать частоты не для конкретного значения, а для группы значений. Это особенно актуально для количественных переменных.

Например, теперь у нас есть данные, которые содержат не оценки за контрольную, а баллы по ЕГЭ по информатике для этих же учеников:

45,55,20,63,65,34,55,88,73,78,75,20,98,63,93,60,14,88,90,75,95,73,34,27,100

Если мы будем выписывать частоту для каждого значения, таблица получится громоздкой, поэтому сгруппируем значения в пять групп одинакового размера: 1-

20 баллов, 21-40 баллов, 41-60 баллов, 61-80 баллов, 81-100 баллов. Получаем следующее распределение:

Значения	Частота
0-20	3
1-40	3
41-60	4
61-80	8
81-100	7

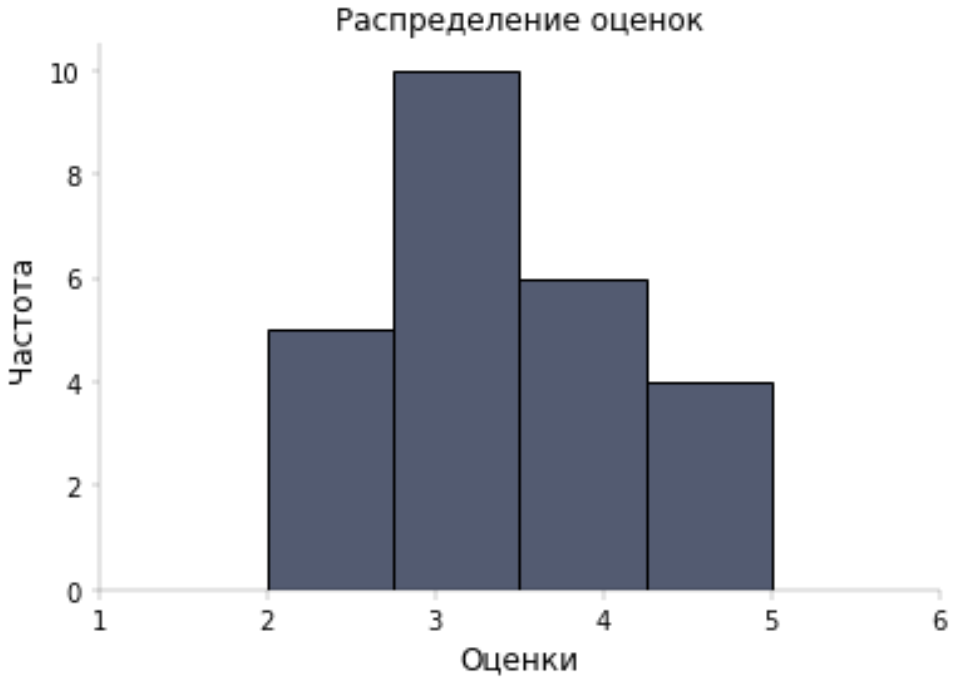
Разумеется, так анализировать распределение стало удобнее. Однако стоит отметить, что в таком виде оно потеряло часть информативности. Например, мы знаем, что восемь человек получили баллы в интервале от 61 до 80, но мы совершенно не понимаем, какие именно это баллы: выше 70 или ниже? Может быть, они все близки к нижней границе, а может быть — к верхней. Такие уточнения могут быть нам полезны для заключения выводов, однако после группировки мы теряем эту информацию. Поэтому всегда стоит понимать — действительно ли в данном конкретном случае частотное распределение с группировкой будет более подходящим.

Визуализация частотного распределения

Частотное распределение можно представить не только в виде таблицы, но и визуализировать его. Для этого есть два типа диаграмм: ***гистограмма*** *и* ***полигон***.

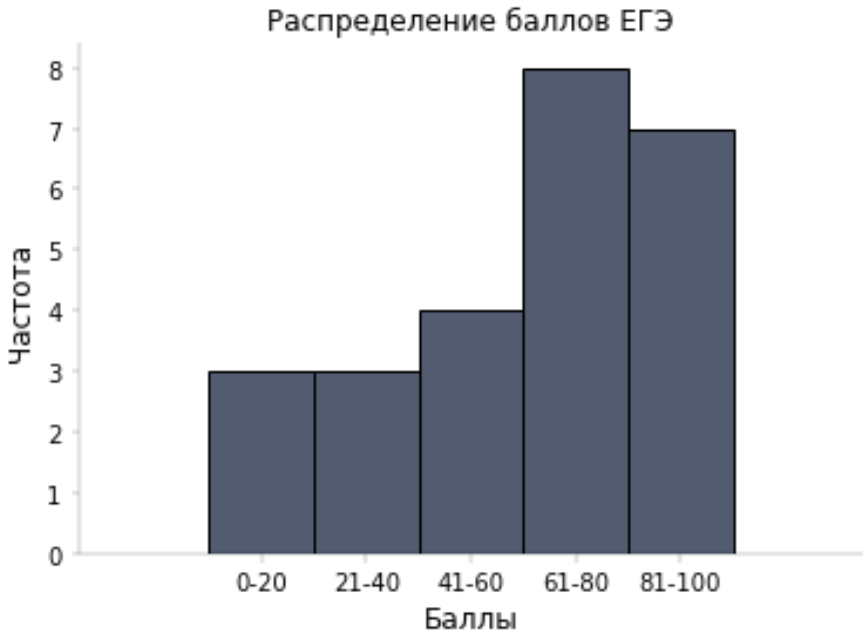
Для построения гистограммы по оси абсцисс (оси x) откладываются все возможные значения, а по оси ординат (оси y) — частоты. Если мы построим гистограмму для примера с оценками, то она будет выглядеть следующим образом:

Значение	Частота
2	5
3	10
4	6
5	4



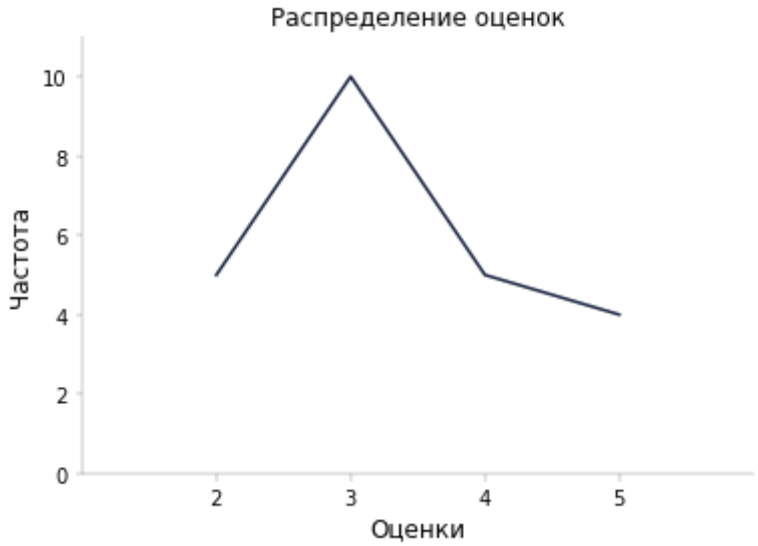
Ровно так же, как построена данная гистограмма, мы можем построить и гистограмму по сгруппированным частотам для примера с баллами ЕГЭ, который мы рассматривали ранее:

Баллы	**Частота**
0-20	3
1-40	3
41-60	4
61-80	8
81-100	7

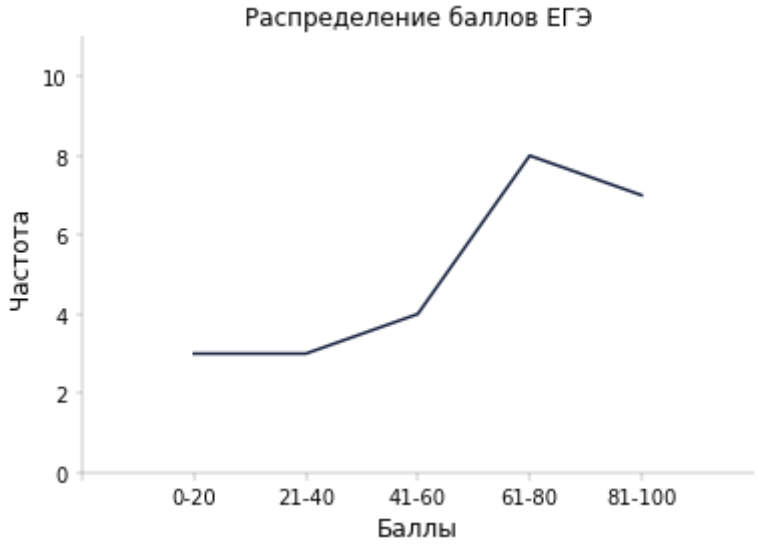


Как вы можете видеть, гистограмма достаточно информативна и легко интерпретируется. Однако важно помнить, что ее можно построить только для количественных данных.

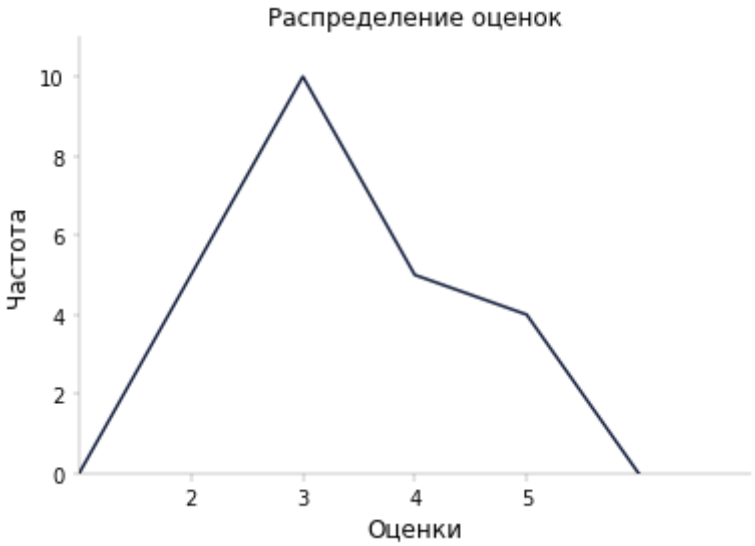
Гистограмма — не единственный вариант визуализации частотного распределения. Также мы можем построить полигон распределения. Для этого нам необходимо отметить точки на области построения, которые соответствуют высоте столбца (как если бы мы строили гистограмму). После того как все точки отмечены, их соединяют, и получается полигон:



Разумеется, для сгруппированных данных мы тоже можем построить полигон распределения:



Есть два варианта отображения полигона: его концы могут быть в крайних точках (как на диаграммах выше) либо соединяться с осью абсцисс в точках, соседних к крайним:



В разных источниках можно встретить как первый вариант отображения полигона, так и второй, так что оба являются допустимыми.

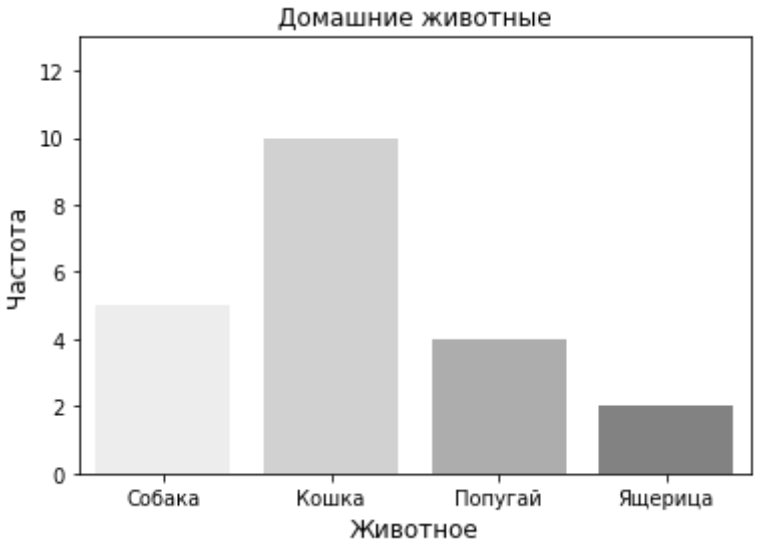
Таким образом, частотное распределение можно представлять как в формате таблицы, так и в графическом: гистограммы или полигона. Однако важно отметить, что полигон является скорее вспомогательным типом визуализации, который используется редко, обычно для визуального представления распределений наилучшим вариантом является гистограмма.

Гистограмма отлично подходит для визуализации количественных данных, но нам важно уметь в аналогичном виде представлять распределение частот для порядковых и номинальных данных. Аналогом гистограммы для качественных данных является **столбчатая диаграмма**. Для ее построения нужны ровно те же действия, что и для гистограммы, но визуально она немного отличается: между столбцами для разных категорий есть расстояние.

Предположим, что мы решили узнать, какие домашние животные есть у группы людей. Мы узнали, что собаки есть у пятерых людей, кошки — у десяти, четверо держат попугаев, и еще у двоих дома живут ящерицы:

Животное	**Частота**
Собака	5
Кошка	10
Попугай	4
Ящерица	2

Теперь по данной таблице частот мы можем построить столбчатую диаграмму. Для каждого животного просто нарисуем столбец, высота которого будет соответствовать частоте.



Если мы визуализируем распределение порядкового признака, то категории обычно располагаются в упорядоченном виде: по возрастанию или убыванию. Для номинального признака значения никак нельзя упорядочить, поэтому столбцы могут располагаться в любом порядке, но чаще для простоты интерпретации их располагают от самого высокого до самого низкого.

Меры вариативности

Зачем нужны меры вариативности?

Меры вариативности, или как их еще называют — меры разброса, помогают нам понять, насколько наши данные разнообразны. Например, вы купили упаковку клубники в супермаркете — все ягоды ярко-красные, примерно одного размера и одинаковые на вкус (безвкусные или сладкие — здесь как повезет). Здесь мы не наблюдаем вариативность ни в одной из характеристик клубники. А клубника, собранная у бабушки на даче, может быть совсем другой — размер ягод от мелких до крупных, цвет от зеленого до темно-бордового, а вкус — от кислой до очень сладкой. Здесь все наши три переменные (размер, цвет и вкус) принимают разнообразные значения, а значит и вариативность присутствует. Меры вариативности как раз и помогают нам определить, насколько наши данные похожи (или нет) друг на друга. Между прочим, в обоих примерах средний вес ягод может быть одинаковым — вес мелких ягоды от бабушки будут компенсироваться весом более крупных. А значит, одной только меры центральной тенденции недостаточно, чтобы знать, что же происходит в наших данных.

Меры вариативности (разброса) *представляют собой количественные меры, выражающие различия между значениями в распределении. Более формально — это степень индивидуальных отклонений значений от центральной тенденции.*

Зная меру вариативности можно представить форму нашего распределения. Например, если мы попросим написать математический тест только победителей профильных олимпиад или детей, которые посещают математические кружки, — то их баллы не будут так сильно отличаться друг от друга и располагаться они будут единым кластером. Если этот же тест будут писать все ученики одной параллели среднестатистической общеобразовательной школы — то, пожалуй, оценки будут и более вариативны, и располагаться они будут на большем интервале значений.

Вариативность также помогает нам понять, насколько типичной для распределения является случайно взятая оценка. Если вариативность маленькая — то значения больше похожи друг на друга. А вот в распределениях с большой вариативностью утверждать, что случайное наблюдение будет несильно отличается от центральной тенденции, пожалуй, не стоит.

Как и в случае с мерами центральной тенденции, существуют несколько мер разброса. Об основных из них мы и поговорим в этой главе — **размах**, **интерквартильный размах, дисперсия** и **среднеквадратичное отклонение.**

Размах

Самая простая мера вариативности — **размах**. По сути размах описывает величину интервала, на котором располагаются значения переменной. Из величины размаха мы можем сказать насколько наши значения сгруппированы в центре или удалены друг от друга.

Размах *— это разница между самым большим и самым маленьким значением переменной.*

Размах не имеет специального обозначения и рассчитывается одинаково и для выборки, и для генеральной совокупности. При необходимости мы будем обозначать размах в формулах словом $range$.

$$range = X_{max} - X_{min}$$

Вы можете встретить и другие определения размаха, которые используют пределы целых чисел для непрерывных переменных и корректировку для дискретных. Но в нашем курсе мы будем пользоваться самым простым определением, представленным выше.

Рассчитаем размах для переменной $X = [3,5,6,7,4,6,5,4,6]$:

$$range = X_{max} - X_{min} = 7 - 3 = 4$$

Интерпретируем размах следующим образом: все значения нашей переменной находятся в диапазоне 4-х единиц измерения.

Размах найти легко, но, к сожалению, это не самая информативная мера разброса, которая зависит от экстремальных значений интервала переменной. Если в нашей переменной будет хотя бы одно нетипичное значение — размах, как и среднее, не будет отражать реальное положение дел. При очень большом размахе возможных значений, данные на самом деле могут быть сгруппированы в одной части распределения. Предположим, что у нашей переменной из прошлого примера появилось еще одно наблюдение — 100.

$$X=[3,5,6,7,4,6,5,4,6,100]$$

$$range=100-3=97$$

Действительно, видим, что размах теперь равен 97, хотя большинство значений все еще сгруппировано в диапазоне от 3 до 7.

Так что на эту метрику мы будем обращать внимание, но другие меры вариативности будут для нас более полезными.

Интерквартильный размах

Для того чтобы разобраться, что такое **интерквартильный размах**, нам сначала нужно понять, что такое **квартили**. На самом деле, мы с ними уже немного знакомы — медиана является вторым квартилем распределения — такой величиной, которая делит распределение пополам (50% и 50%).

Квартили *— это значения, которые делят распределение на четверти.*

- **Первый квартиль** (еще его называют **нижним**) — отделяет первые 25% значений от следующих 75%.
- **Второй квартиль** (он же **медиана**) — делит выборку пополам (50% и 50%).
- **Третий квартиль** (еще его называют **верхним**) — отделяет первые 75% выборки от следующих 25%.
- **Четвертый квартиль** по сути уже не делит выборку — ниже него располагается 100% значений.

Возможно, вы еще встречали такое понятие как **персентиль** или **процентиль** — это такое значение переменной ниже которого находится определенный процент наблюдений в наших данных. Например, некоторые зарубежные вузы используют в качестве проходного балла не сам балл за экзамен, а персентиль. Так, если указано, что нужно попасть в 95-й персентиль – это значит, что ваш балл за тест должен быть лучше, чем у 95% сдававших в этот год.

Вернемся к **интерквартильному размаху**. Для того чтобы его вычислить — нужно найти 1-й и 3-й квартили распределения — значения, которые делят его на 25% и 75% и 75% и

25% соответственно. Как уже, наверное, понятно — это мера, которая часто используется в паре с медианой — ведь они обе вычисляются на основе порядковых номеров значений переменной.

Интерквартильный размах *— интервал значений признака, содержащий центральные 50% наблюдений распределения, то есть интервал между первым и третьим квартилем.*

Интерквартильный размах тоже не имеет специального обозначения и рассчитывается одинаково для выборки и генеральной совокупности. Мы будем обозначать его как IQR (inter-quartile range). Квартили будем обозначать заглавной буквой Q и индексом — порядковым номером квартиля.

$$IQR = Q_3 - Q_1$$

Давайте рассчитаем интерквартильный размах для переменной $X = [3,5,6,7,4,6,5,4,6,100]$. Мы уже работали с этим примером и видели, что простой размах получается для нее не очень релевантным. Сначала упорядочим значения, чтобы найти квартили:

$$3,4,4,5,5,6,6,7,100$$

Проще всего найти квартили, поделив выборку сначала пополам, а потом найти серединки (медианы) каждой половинки. Если 5 — медиана нашей выборки, то 50% и 50% выборки у нас такие:

$$3,4,4,5 \text{и} 6,6,7,100$$

Дальше действуем по той же логике, что и с медианой — находим среднее арифметическое центральных элементов:

$$Q_1 = 4 + \frac{4}{2} = 4$$

$$Q_3 = 6 + \frac{7}{2} = 6.5$$

$$IQR = 6.5 - 4 = 2.5$$

Интерпретируем интерквартильный размах: 50% наблюдений в нашей примере (между первым и вторым квартилем) находятся в диапазоне двух единиц измерения. Если мы посмотрим на наше оригинальное распределение, то убедимся, что это правда:

$$3,4,4,5,5,6,6,7,100$$

А теперь вспомним, что размах для этой же переменной был равен 97. Так как при вычислении интерквартильного размаха мы как раз отсекли наше аномально большое значение 100, то эта мера описывает наше распределение лучше. Таким образом можно сделать вывод, что, как и медиана, интерквартильный размах менее чувствителен к наличию аномальных значений в переменной по сравнению с другими мерами вариативности.

Дисперсия и среднеквадратичное отклонение

Если интерквартильный размах — мера, которая используется в паре с медианой, то **среднеквадратичное отклонение** — лучший друг среднего арифметического. Отклонение, пожалуй, самая частотная мера вариативности — оно берет среднее арифметическое как точку отсчета и оценивает насколько данные сгруппированы вокруг него или наоборот удалены.

Но с начала давайте определим, что такое отклонение от среднего.

Отклонение от среднего *— это разница между значением переменной и ее средним арифметическим.*

Математически отклонение от среднего можно выразить для генеральной совокупности так:

$$X-\mu$$

И так для выборки:

$$X-\mathrm{M}$$

Например, если у нас есть генеральная совокупность, где $\mu=50$, и конкретное значение переменной $X=53$, то отклонение от среднего для этого значения будет следующим:

$$X-\mu=53-50=3$$

Отклонение может быть и отрицательным, если значение меньше среднего (например, для $X=45$ из того же распределения отклонение будет -5).

Так как мы хотим, чтобы наша мера разброса описывала наше распределение целиком, то было бы здорово найти такую метрику, которая агрегирует все наши отклонения от среднего. Давайте поработаем с такой генеральной совокупностью:

$$N=4, X=[8,1,3,0], \mu=3$$

Найдем для каждого значения его отклонение от среднего:

X	$X-\mu$
8	$8-3=5$
1	$1-3=-2$
3	$3-3=0$
0	$0-3=-3$

Пожалуй, следующим логичным шагом было бы найти среднее отклонений от среднего распределения. Но на самом деле из свойств среднего арифметического мы уже знаем, что сумма этих дистанций будет равна 0:

$$\Sigma(X-\mu)/N=5+(-2)+0+(-3)/4=0/4=0$$

Что нам может здесь помочь? Один из вариантов — возвести отклонения в квадрат. Так мы избавимся от знаков минус, а затем сможем найти среднее квадратичное отклонение. Такая мера вариативности называется — **дисперсия**.

Дисперсия *— среднее квадратов отклонений от среднего арифметического распределения.*

X	$X-\mu$	$(X-\mu)^2$
8	$8-3=5$	25
1	$1-3=-2$	4
3	$3-3=0$	0
0	$0-3=-3$	9

Дисперсия для нашей переменной будет равна:

$$\Sigma(X-\mu)^2$$

$$N=25+4+0+94=384=9.5$$

Таким образом мы знаем, что в среднем квадрат отклонения от среднего находится на расстоянии 9.5 единиц измерения от среднего

арифметического распределения. Полезная ли это мера? Пожалуй. Но интерпретировать ее не просто — гораздо полезнее знать, что между Санкт-Петербургом и Москвой 721 километр по трассе, чем квадрат этого расстояния — 519841 километр.

Мы возводили отклонения в квадрат, чтобы перехитрить арифметику — нам нужно было избавиться от отрицательных чисел, чтобы посчитать среднее отклонений. Теперь было бы неплохо вернуться от квадратичных к оригинальным единицам измерений. Сделать это на самом деле несложно — мы просто извлечем квадратный корень из нашей дисперсии. Это и будет **среднеквадратичное** (оно же **среднеквадратическое** или **стандартное**) отклонение.

Среднеквадратичное отклонение *(стандартное отклонение) — квадратный корень дисперсии. Мера, которая определяет среднее отклонение от среднего арифметического распределения.*

$$\sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{9.5} \approx 3.08$$

Таким образом мы можем сказать, что в нашей генеральной совокупности

$$N=4, X=[8,1,3,0], \mu=3$$

данные в среднем отличаются от μ на 3.08.

Давайте теперь запишем формулы для дисперсии и среднеквадратичного отклонения. Так как есть нюанс расчета среднеквадратичного отклонения для выборки, то будем использовать разные буквы — σ (сигма) для генеральной совокупности и s для выборки. Дисперсия — среднеквадратичное отклонение в квадрате — обозначается σ^2 и s^2 соответственно.

Формула дисперсии генеральной совокупности:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Формула среднеквадратичного отклонения генеральной совокупности:

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Как мы уже сказали выше, с выборкой есть нюанс.

Когда мы извлекаем значения из генеральной совокупности, даже если наша выборка сделана с соблюдением всех правил, то у более редких значений (наиболее отличных от среднего нашего распределения) шанс в эту выборку попасть все равно очень маленький. А значит наша выборка скорее всего недооценивает вариативность генеральной совокупности, из которой ее извлекли. Но математики выяснили, что эта величина довольно предсказуемая, а значит ее можно скорректировать.

Корректировка заключается в том, что в нашей формуле в знаменателе мы будем использовать не n (количество наблюдений в выборке), а $n-1$. А раз мы уменьшаем значение в знаменателе на 1, то результат деления станет немного больше — мы как будто искусственно немного увеличиваем значение нашей выборочной дисперсии и среднеквадратичного отклонения.

Формула дисперсии выборки:

$$s^2 = \frac{\Sigma(X-M)^2}{n-1}$$

Формула среднеквадратичного отклонения выборки:

$$s = \sqrt{\frac{\Sigma(X-M)^2}{n-1}}$$

Какую меру вариативности выбрать?

Как вы могли уже заметить, некоторые меры вариативности высчитываются на основе мер центральной тенденции (среднеквадратичное отклонение и среднее) или по схожему принципу (интерквартильный размах и медиана). **Поэтому среди этих двух логично выбирать метрику в пару той мере центральной тенденции, которую вы используете.**

Размах полезно определить для любой количественной переменной в пару к другой мере, чтобы лучше представлять диапазон, в котором существуют ваши данные.

Обратите внимание, что **среднеквадратичное отклонение нельзя вычислить для категориальных переменных — номинальных и порядковых.**

Размах и **интерквартильный размах** можно использовать с некоторыми **порядковыми** переменными, выраженными числами (например, ответы на вопрос про степень удовлетворенности сервисом по шкале от -5 до $+5$).

А вот для **номинальных переменных** единственной мерой вариативности будет только такая неформальная метрика как определение количества уникальных значений. Например, бывают переменные, в которых практически все значения уникальные (ФИО студента) — сразу понятно, что роль такой характеристики в анализе будет очень ограниченной.