МАТЕМАТИЧЕСКАЯ ВЕРТИКАЛЬ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И СТАТИСТИКА. 7 КЛАСС (2 ч./нед.)

Урок 10. Диаграммы рассеивания (продолжение)

Материалы представляют собой примерный сценарий урока. Учитель может на свое усмотрение использовать сценарий целиком или частично наряду с собственными разработками и материалами учебника[1]. Авторы будут благодарны за замечания и предложения по структуре и содержанию сценариев.

Важно! Урок предполагает наличие у учащихся знаний о линейной функции, ее графике и угловом коэффициенте прямой. Если эти темы еще не изучены, то целесообразно перенести этот урок на более позднее время.
Оборудование: линейка или угольник, цветные карандаши или фломастеры, миллиметровая бумага, калькулятор.
Цель урока – формирование представления об оценке параметров линейных зависимостей с помощью диаграмм рассеивания. Учащиеся должны научиться проводить ось диаграммы и выражать линейную связь между величинами формулой.

Пример 1. На одном из популярных сайтов, посвященных продаже подержанных автомобилей, есть данные о ценах на автомобили Шкода «Октавия». В таблице приведены данные о 29 предложениях в октябре 2018 года.

	Год выпуска	Пробег (тыс. км)	Цена (тыс. руб).		Год выпуска	Пробег (тыс. км)	Цена (тыс. руб).		Год выпуска	Пробег (тыс. км)	Цена (тыс. руб).
1	2015	26	1170	11	2015	95	707	21	2011	150	427
2	2010	40	410	12	2012	100	519	22	2012	156	495
3	2016	50	828	13	2014	100	580	23	2007	160	315
4	2014	70	698	14	2012	101	440	24	2008	170	395
5	2014	52	750	15	2013	105	515	25	2008	176	350
6	2013	60	615	16	2010	110	470	26	2009	210	355
7	2015	66	610	17	2011	120	584	27	2009	210	290
8	2016	70	865	18	2013	130	550	28	2007	285	230
9	2016	75	684	19	2009	135	329	29	2008	280	130
10	2015	93	735	20	2011	141	479				

Можно рассмотреть разные величины и их соответствующие значения. Рассмотрим, например, пару «Возраст машины – цена». Ясно, что чем старше автомобиль, тем он дешевле (ведь это не раритетные коллекционные машины).

[1] Математика 7–9 класс. Теория вероятностей и статистика / Ю. Н. Тюрин, А. А. Макаров, И. Р. Высоцкий, И. В. Ященко. – 3-е изд., стереотипное. – М.: МЦНМО: ОАО «Московские учебники», 2011. – 256 с.: ил.

Цена каждый год снижается примерно на 10%[2]. Поэтому связь между ценой и возрастом нелинейная. Давайте рассмотрим связь между возрастом автомобиля на момент продажи и его пробегом. Очевидно, пробег автомобиля в целом пропорционален его сроку службы. Если машина прослужила год, то она могла наездить, предположим, 20 000 км. Еще год – еще примерно столько же, если образ жизни водителя не поменялся, и так далее. Логично считать, что между величинами «возраст авто» и «пробег авто» должна быть *положительная линейная связь*. Построим диаграмму рассеивания и на глаз проверим, правдоподобно ли это предположение.

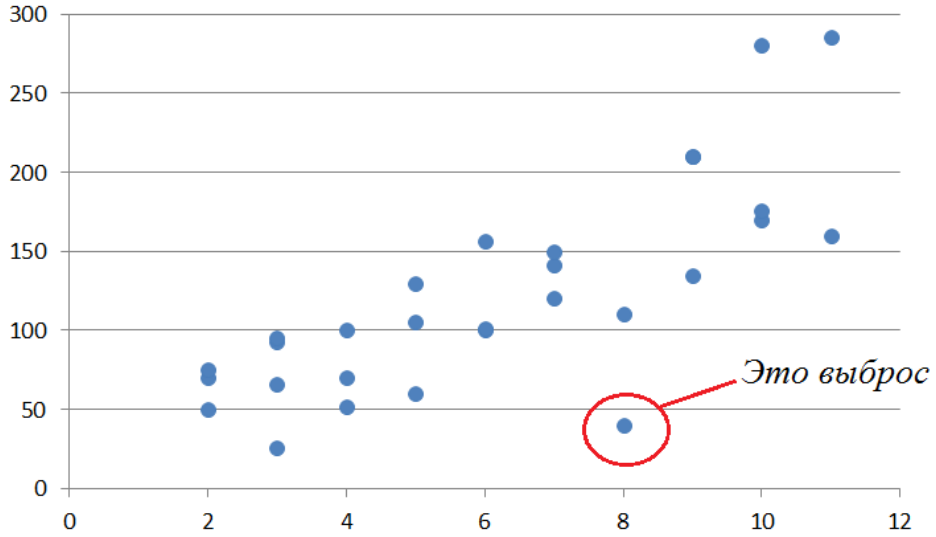


Рис. 1. Диаграмма «Возраст (г) – Пробег (тыс.км)

Желательный результат обсуждения. Сначала таблицу нужно сделать удобнее: добавим столбец «Возраст» вместо ненужного уже столбца «Цена». На построенной диаграмме рассеивания для пары «Возраст – Пробег» (рис.1) почти все точки образуют хорошо оформленное наклонное и вытянутое облако. Исключение составляет точка (8; 40). Эта выбивающаяся точка – *выброс*, то есть нетипичное и даже немного удивительное явление. Как мог получиться выброс? Возможны разные причины. Предложите школьникам подумать, как могло получиться, что у этой машины мог оказаться такой небольшой пробег за восемь лет. Возможны разные варианты. Например:

– Хозяин заболел и не смог ездить – машина стояла в гараже несколько лет, а теперь он или его родственники продают машину;

– Автомобиль ездил только летом – вывозил семью на дачу;

– Может быть, хозяин пытается «набить цену» и «скрутил пробег»? Такое тоже бывает...

– Может быть, человек внося данные на сайт, просто ошибся, и 40 тыс. км на самом деле 140 тыс. км?

[2] Данные Российского союза страховщиков.

Определить по диаграмме точную причину невозможно, да и не нужно. Важно, что эта точка выбивается из общей картины, и мы ее не будем учитывать, а рассмотрим облако без нее. *Выбросы нужно удалять – эта процедура чистки данных позволяет получить более точную картину явления.*
Важно! Не существует единого правила, какие точки считать выбросами, а какие нет. Есть общепринятые методы, но в данном случае мы просто видим, что точка далеко от облака.

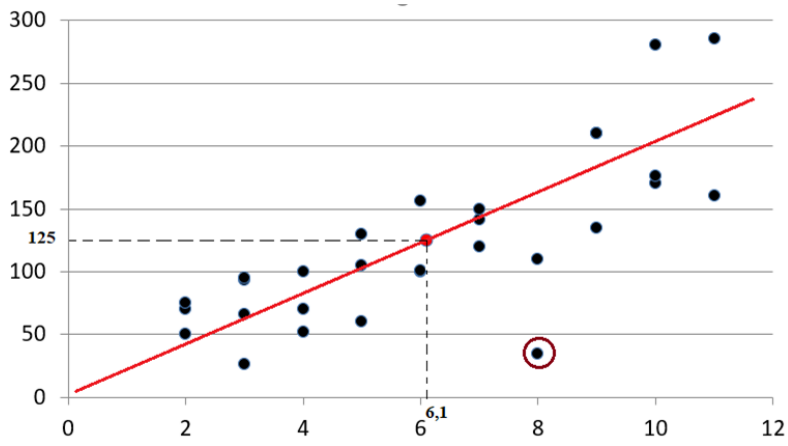
Как расположены точки в облаке? Они сгруппированы около некоторой прямой. Это подтверждает предположение о линейном характере связи. Но что это за прямая? Если мы ее проведем, то сможем написать ее уравнение. Тогда мы будем знать, как связь между пробегом и возрастом приближенно выражается формулой.

Провести прямую можно разными способами. Мы будем использовать самый простой – «глазомерный». Найдем прямую, которую можем назвать *осью диаграммы*[3]. Искомая прямая будет описывать *линейную модель связи* величин в нашей задаче.

Сначала найдем центральную точку. Например, можно найти средний возраст (6,1 года) и средний пробег (125 тыс. км). Напомните учащимся исключить выброс из расчётов. Поставим эту центральную точку.

Через эту точку на глазок (но по линейке) проводим прямую так, чтобы она прошла примерно посередине облака. У нас получилась прямая, которая проходит очень близко к началу координат. И это разумно: ведь новая машина имеет нулевой пробег[4].

Чем плотнее точки группируются вокруг оси диаграммы, тем сильнее связь и тем лучше построенная прямая описывает эту связь.



[3] Иногда используется термин «линия тренда». Этот термин оправдан тогда, когда речь идет о процессе во времени: изменение какой-либо величины во времени. В нашем случае ось диаграммы с натяжкой можно назвать трендом. А вот в следующем примере «Протяженность линии метро – количество станций на ней» зависимости от времени нет. Здесь упоминание тренда неуместно. Поэтому мы везде используем нейтральное название «ось диаграммы рассеивания».

[4] Это одно из возможных рассуждений – простое и не самое лучшее. Другое рассуждение: «Машина, возраст которой 0 полных лет, не старше года. Будем считать, что ей полгода. Если за год машина проходит k км, то за полгода она проходит в два раза меньше: $k/2$ км. Это пробег машины возраста 0 лет. Поэтому $b = 0,5k$. Получаем уравнение $y = kx + 0,5k$. Теперь достаточно одной центральной точки $(6,1;125)$, чтобы определить k: $125 = k(6,1 + 0,5)$, откуда $k \approx 18,94$. Получается уравнение $y = 18,94x + 9,47$. Разные предположения дают разные методы и приводят к разным, но близким моделям.

Уравнение прямой составить несложно. Напишем уравнение прямой

$$y = kx + b.$$

Теперь возьмем на построенной прямой две точки. Например, в нашем случае удобно взять начало координат и центр (см. таблицу справа), и подставим эти числа в уравнение. Получаем систему двух уравнений с неизвестными k и b

x	0	6,1
y	0	125

$$\begin{cases} 0 = k \cdot 0 + b, \\ 125 = k \cdot 6{,}1 + b, \end{cases}$$

откуда $b = 0$, $k = \frac{125}{6{,}1} \approx 20{,}49$. Получаем уравнение оси: $y = 20{,}49x$. Это значит, что в машины рассмотренной совокупности пробегают в среднем по 20,49 тыс. км в год.

С помощью полученной формулы можно ответить на вопрос о приблизительном (типичном, ожидаемом) пробеге легкового автомобиля известного возраста. Например, машина, которой 5,5 лет, согласно нашей линейной модели, имеет пробег $20{,}49 \cdot 5{,}5 \approx 112{,}7$ (тыс. км).

Предложите учащимся самостоятельно провести ось диаграммы, построенной в задаче о московском метрополитене (см. урок 9) и написать ее уравнение.

Пример 2. Найдите координаты центра облака данных и проведите ось диаграммы. Найдите уравнение оси. При помощи полученной формулы связи между протяженностью линии метро и числом станций на ней в московском метрополитене, оцените ожидаемое количество станций на линии длиной 30 км.

	длина, км	кол-во станций
Сокольническая	32,5	22
Замоскворецкая	42,8	23
Арбатско-Покровская	45,1	22
Филевская	14,9	13
Кольцевая	19,3	12
Калужско-Рижская	37,6	24
Таганско-Краснопресненская	42,2	23
Калининская	16,5	8
Солнцевская	24,9	12
Серпуховско-Тимирязевская	41,2	25
Люблинско-Дмитровская	38,3	23
Большая кольцевая	10,5	5
Каховская	3,3	3
Бутовская	10	7
Монорельс	4,7	6
МЦК	54	31

Желательный результат. Выбросов на первый взгляд нет (см. рис.2). Координаты центра облака (27,4; 16,2). Через нее учащиеся на глаз строят прямую. В этом случае из некоторых соображений разумно взять точку $(0; 1)$ (почему?) и провести прямую через нее и через центр. Не исключено, что эта прямая окажется подходящей осью.

Оси у каждого ученика получатся разные, поэтому и уравнения будут немного отличаться. Теперь предложите найти, какое число станций должно

было бы быть на линии протяженностью 30 км (такой линии на самом деле нет).

В нашем примере получилось $y = 0{,}53x + 1{,}7$. Это значит, что (согласно нашей модели) в московском метро на каждый километр линий приходится примерно 0,53 станции или 53 станции на 100 км.

На линии протяженностью 30 км окажется примерно 17,6 станций.

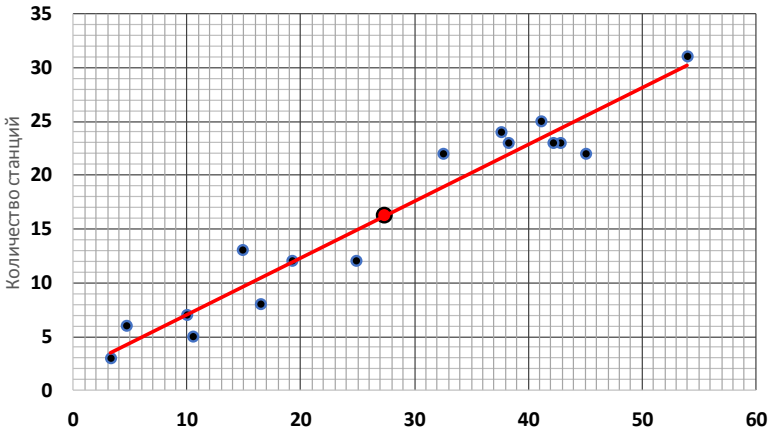


Рис. 2. Диаграмма «Протяженность линии метро (км) – число станций»

Важно! Об округлении: в жизни количество станций, безусловно, целое число. Многие школьники захотят округлить полученный результат до целого. Округление возможно, но не обязательно: метод может дать дробное число.

На прошлом уроке учащиеся построили диаграмму рассеивания «Дальность прыжка – время забега». Можно ли считать, что если человек вдвое дальше прыгает, то он вдвое быстрее бегает? Или, что есть вообще какая-то линейная связь между дальностью прыжка и временем забега на 60 м? Обсудите это. Если есть мнение, что диаграмма являет подобие линейной зависимости, предложите построить ось этой диаграммы.

Прыжок, см	180	194	190	215	210	170	175	202	205	195	205	200	190	186
Бег, с	10,8	10,2	10,6	9,6	10,2	11,0	11,6	10,4	10,0	11,0	9,8	10,6	10,8	10,7

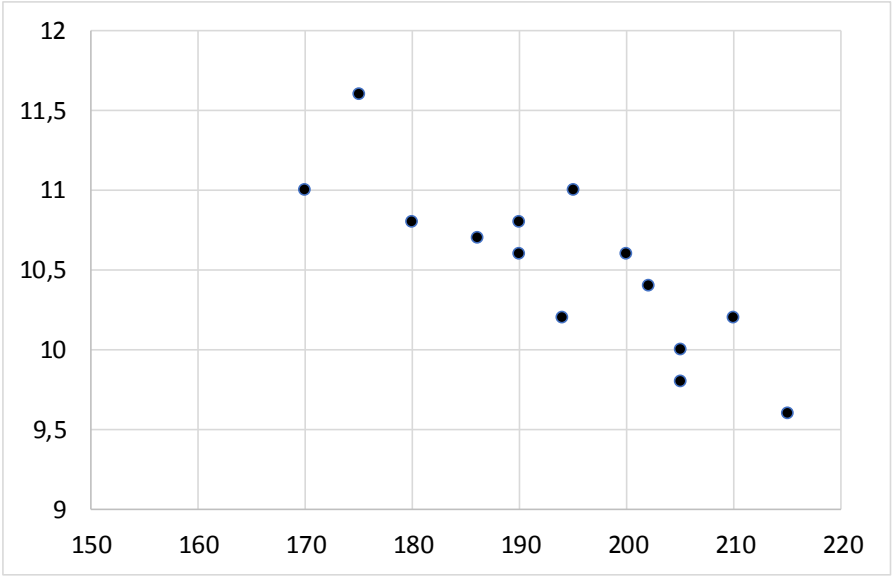


Рис. 3. Таблица и диаграмма «Дальность прыжка (см) – время забега на 60 м (с)»

Желательный результат обсуждения. Чем быстрее человек бегает, тем дальше он прыгнет. Но многое еще зависит от силы и направления толчка при

прыжке. Взаимосвязь между данными сложная. Однако форма диаграммы позволяет надеяться, что связь *похожа на линейную.* Можно провести ось и написать ее уравнение, но нужно помнить, что эта линейная формула опишет модель не точно и вряд ли позволит нам делать хорошие оценки значений.

Обсудите еще одну задачу прошлого урока: про рост и вес. Можно ли считать, что вес и рост связаны линейно? Имеет ли смысл искать линейную зависимость в этом случае?

Рост, см	167	169	179	178	177	175	171	181	174	175	180	174	172	178	171
Вес, кг	62	67	70	72	70	69	63	80	73	66	75	70	67	74	66

Желательный результат обсуждения. Вспомним формулу объема параллелепипеда $V = abc$. Конечно, человек не параллелепипед, но между ростом и объемом тела (следовательно, весом) зависимость довольно сложная – третьей степени (не зря же объем измеряется в кубических метрах). Линейной связи нет. Построить ось и формулу можно, но для очень высоких и очень низких людей эта формула работать, скорее всего, не будет.

Выводы. Диаграммы рассеивания позволяют визуально выявить положительную или отрицательную связь. Но выявить характер связи более определенно – линейная она или более сложная – часто не удается. В двух первых примерах мы понимали, что линейная связь есть: во сколько раз больше возраст машины, во столько же раз больше пробег. Во сколько раз длиннее линия метро, во столько же раз на ней больше станций. Конечно, все изменчиво, разброс данных велик, но облака точек в этих примерах вытянуты вдоль прямых, и мы понимаем, что так должно быть, исходя из здравого смысла. В таких случаях ось диаграммы позволяет лучше понять характер связи и даже *количественно оценить ее параметры,* написав уравнение.

Важно! Характер линейной связи – положительная она или отрицательная – определяет знак углового коэффициента k в уравнении оси диаграммы.

Часто связь двух величин не описывается линейной функцией, хотя может казаться, что большая часть облака (или даже все облако) вытянута вдоль прямой. В этом случае прямолинейная ось диаграммы описывает истинную связь приближенно только на том участке, где облако прямолинейно.

Рекомендуемое домашнее задание: см. приложение.

Дополнительные материалы для подготовки к уроку. Учитель на свое усмотрение может свободно использовать таблицы[5], размещенные на сайте «Вероятность в школе» http://ptlab.mccme.ru/node/350. Для изучения темы «Диаграммы рассеивания» хорошо подходят географические таблицы с данными о крупных реках, озерах и т.п.

[5] Все данные взяты из открытых источников и представлены в формате .xls или .xlsx. Таблицы можно использовать как целиком, так и делая из них необходимые выдержки.

Приложение

Домашнее задание. Таблица содержит данные об атмосферном давлении в полдень 31 октября 2018 года в 34 городах России. В третьей колонке дана средняя высота центра города над уровнем моря.

1. Постройте диаграмму рассеивания по данным таблицы на миллиметровой бумаге.
2. Имеется ли связь между данными? Если да, охарактеризуйте ее (положительная она или отрицательная).
3. Что вызывает изменчивость атмосферного давления? Приведите один-два фактора, объясняющих, почему если связь между высотой и давлением существует, то она не очень сильная.
4. Как по вашему мнению, есть ли основания считать, что эта связь хотя бы приближенно описывается линейной функцией?
5. Найдите координаты центра облака данных и проведите ось диаграммы.
6. Напишите уравнение оси диаграммы.
7. Согласно полученному уравнению, какое атмосферное давление можно было бы ждать 31 октября в месте, которое находится на высоте 400 метров над уровнем моря?

Город	**Атм. давление (мм.р.ст.)**	**Высота города над ур. моря (м)**	**Город**	**Атм. давление (мм.р.ст.)**	**Высота города над ур. моря (м)**
Абакан	738	245	Невинномысск	742	320
Барнаул	749	180	Новосибирск	752	150
Владикавказ	712	692	Омск	760	90
Грозный	756	130	Пенза	770	200
Екатеринбург	753	270	Пермь	764	149
Ессентуки	701	640	Санкт-Петербург	770	3
Иркутск	721	440	Ставрополь	728	620
Казань	774	60	Тверь	764	146
Калининград	762	5	Тюмень	767	102
Кемерово	749	140	Улан-Удэ	716	500
Кизляр	778	-5	Уфа	771	150
Кисловодск	692	825	Хабаровск	753	72
Красноярск	740	287	Ханты-Мансийск	766	50
Магас	720	600	Челябинск	756	225
Москва	765	187	Черкесск	723	530
Нальчик	733	512	Чита	708	650