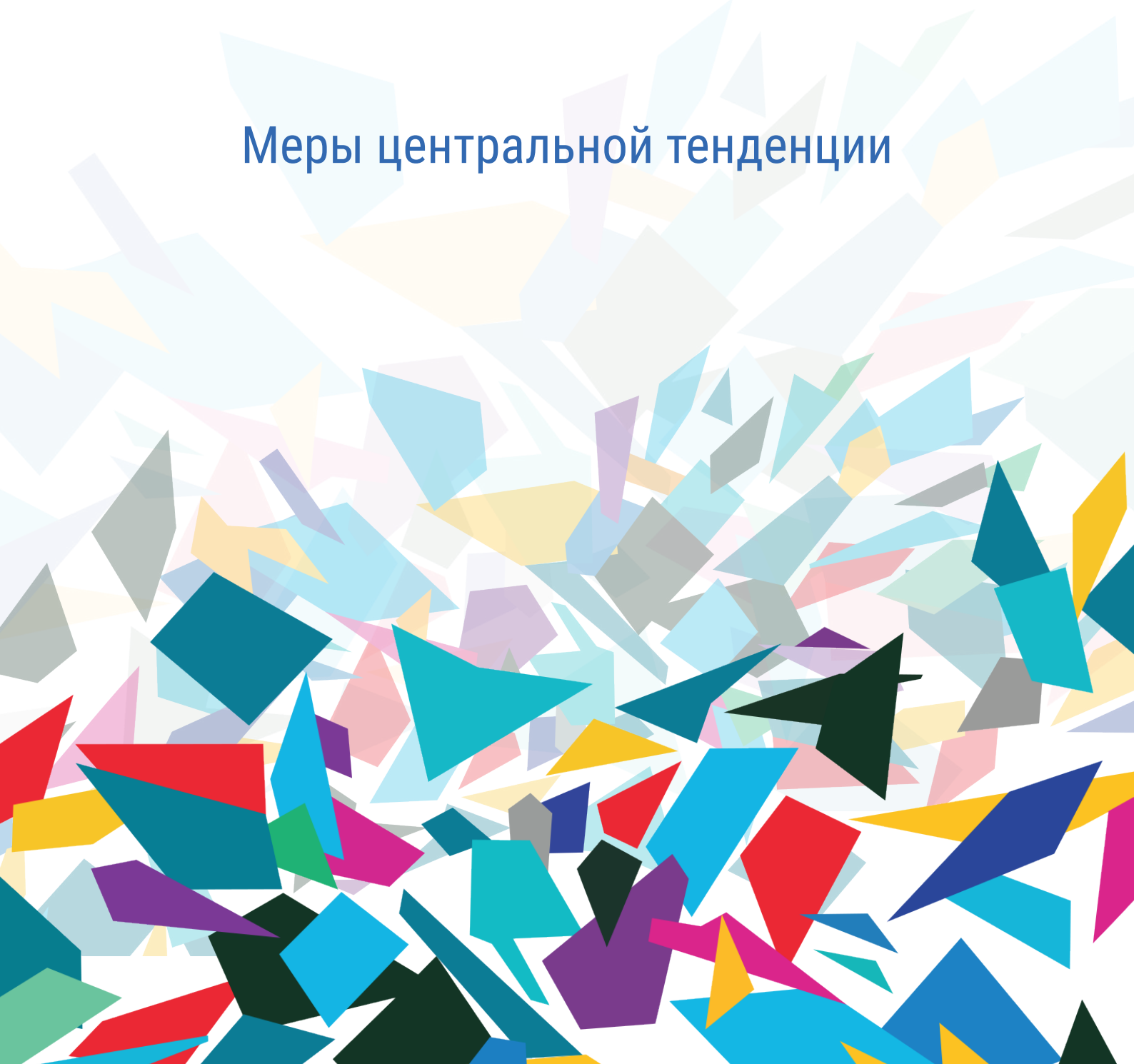
Лекция 4

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Меры центральной тенденции



Здравствуйте!

После изучения материала лекции вы сможете:

1. Объяснять значения мер центральной тенденции и интерпретировать информацию
2.Рассчитывть, объяснять, сравнивать и сопоставлять моду, медиану и среднее.
3. Объясните математические характеристики среднего.
4. Выбирать подходящую меру центральной тенденции в соответствии с уровнем измерения и перекосом.

Одно из преимуществ частотных распределений и графиков состоит в том, что они суммируют общую форму распределения оценок таким образом, чтобы их можно было быстро понять. Однако вам почти всегда нужно сообщать более подробную информацию о ситуации. В частности, чрезвычайно полезны два дополнительных вида статистики: некоторое представление о типичном или среднем случае в распределении (например, «Средняя начальная зарплата социальных работников в рассматриваемом регионе штат 120 000 тенге») и некоторое представление о том, насколько разнообразны есть распределение («В рассматриваемом регионе начальная заработная плата для социальных работников колеблется от 95 000 до 137 000 тенге в месяц).

Три наиболее часто используемых показателя центральной тенденции – мода, медиана и среднее. Все, вероятно, вам знакомы. Они суммируют распределение баллов, описывая наиболее распространенный балл (режим), балл в среднем (медиана) или средний балл (среднее) этого распределения. Эти статистические данные являются мощными, потому что они могут сократить огромные массивы данных до единого легко понятного числа. Помните, что главной целью описательной статистики является обобщение или «сокращение» данных. Несмотря на то, что они имеют общую цель, три показателя центральной тенденции – это разные статистические данные, и они будут иметь одинаковое значение только при определенных условиях. Они различаются с точки зрения уровня измерения и, возможно, что более важно, с точки зрения того, как они определяют центральную тенденцию. Они не обязательно идентифицируют ту же оценку или случай, что и «типичный». Таким образом, ваш выбор подходящего показателя центральной.

Мода

Мода (типичность, максимальная частота) – наиболее часто встречающееся значение в совокупности наблюдений. Применяется, например, для определения размера одежды, обуви, калибра патронов, пользующихся популярностью у покупателей, анализа технических экспериментов, а также определение часто встречающегося значения среди данных, имеющих не числовую природу происхождения (например, цвета: синий, красный, желтый, синий, зеленый…).

Рисунок 4.1 Значения показателя.

Давайте найдем моду – максимально встречающееся значение в данной совокупности:

7	15	15	5	7	6	5	13	17	4	11	2	5
6	13	15	8	9	6	8	8	11	13	11	11	13
15	17	11	6	5	6	3	8	7	10	10	13	3
16	7	6	16	7	13	14	16	12	12	8	13	17
14	15	14	8	4	13	6	13	18	8	10	12	14
2	9	7	5	10	13	4	5	1	11	12	9	12
10	4	14	6	13	18	6	9	15	6	13	14	6
4	16	16	11	6	17	5	12	7	15	7	6	2
13	11	13	16	13	3	4	16	14	13	9	7	5
7	14	8	2	11	4	1	5	9	12	3	12	14
13	14	16	9	13	13	3	4	12	3	14	7	6
7	7	5	4	10	13	3	8	10	15	3	5	17
8	12	15	15	17	7	8	14	9	5	15	7	16
14	4	5	4	13	13	14	18	13	6	8	6	4
5	9	12	13									

Рассчитаем значение моды у нас получилось 13. Т.е. максимально часто встречающееся значение в данной совокупности является значение 13.

Но если построить график, то получается такая картина

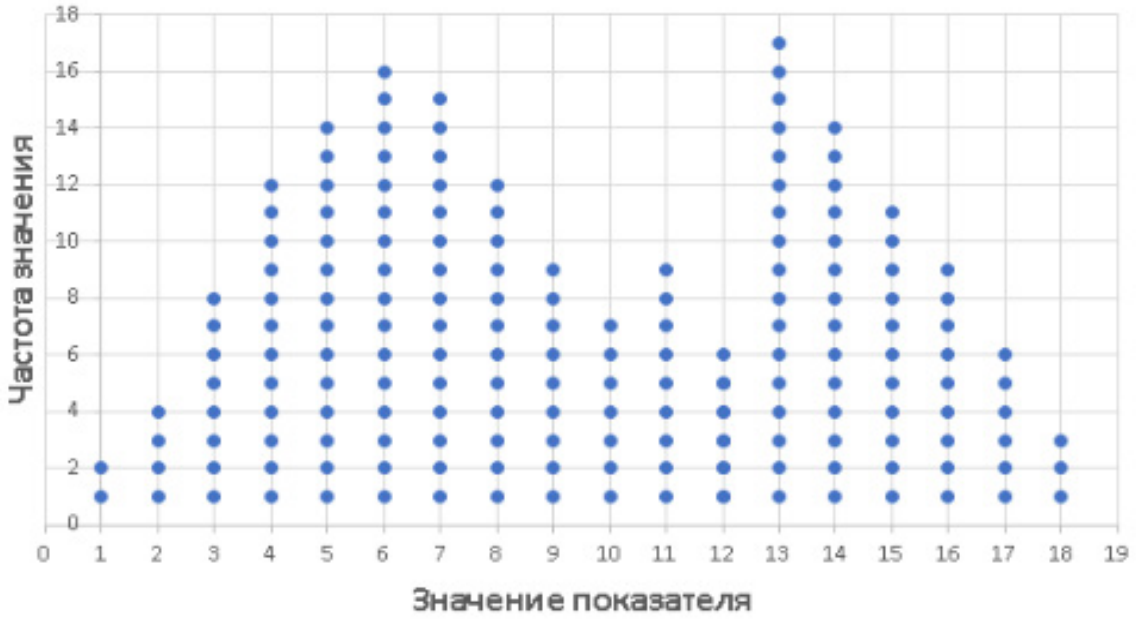


Рисунок 4.2 Значение показателя 2

Видим, что на анализируемый показатель влияет 2 значения: это значения показателей 6, который встречается 16 раз и 13, встречающийся 17 раз. Например, такая ситуация может возникнуть при выборе кандидата в президенты: первая вершина — отданные голоса городского населения, вторая — сельского. Такой эффект называется мультимодальностью и, как правило, указывает что набор данных не подчиняется нормальному распределению.

Среднее арифметическое. Среднее арифметическое – сумма всех чисел, деленное на их количество, зависимое от разброса наблюдений.

$$\text{Среднее арифметическое} = \frac{\text{сумма чисел}}{\text{количество слагаемых}}$$

Например, среднее арифметическое чисел 3, 7, 11 будет: (3+7+11) /3 = 7.

Недостатком данной меры является чувствительность к различным отклонениям и неоднородностям в выборке, другими словами, оно подвержено существенным искажениям со стороны «отщепенцев» (значений) резко отклоняющихся от центра распределения. Для распределений с большим коэффициентом асимметрии может не соответствовать понятию среднего.

В приведенном примере аномальные значения («отщепенцы») будут наращивать среднее значение: если считать среднее арифметическое число проблем с качеством на 1 принтер, то получим 9,1. Впечатляющая цифра! Медиана проблем равняется 1.
Чтобы уяснить эту концепцию представьте 3-х мужчин, находящихся в одной социальной сфере.

Рисунок 4.3 Пример искажения среднего.

Пример искажения среднего. Предположим, что у каждого из мужчин годовой доход составляет 1 440 000 тенге. Но тут, к ним подсаживается бизнесмен, с годовым доходом 545 000 000 тенге, то мы ошибочно будем полагать что он составляет 137 330 000. Что на самом деле не соответствует действительности.

Три характеристики среднего значения.

Среднее значение является наиболее часто используемым показателем центральной тенденции, и мы рассмотрим его математические и статистические характеристики более подробно.
1. Среднее значение баллов по всем счетам. Во-первых, среднее значение является отличной мерой центральной тенденции, поскольку оно действует как точка опоры, которая «уравновешивает» все оценки: среднее значение - это точка, вокруг которой сводятся все оценки. Мы можем выразить это свойство символически:

$$\sum \left(X_i - \overline{X} \right) = 0$$

Это выражение говорит, что, если мы вычтем среднее из каждой оценки (Xi) в распределении, а затем сложим различия, результат всегда будет равен 0.
Для иллюстрации рассмотрим результаты тестов, представленные в таблице 3.6. Среднее из этих пяти баллов составляет 390/5 или 78. Разница между каждым баллом и средним значением указана в правом столбце, а сумма этих различий равна 0. Сумма отрицательных различий (-19) точно равен сумме положительных разностей (+19), как всегда будет. Таким образом, среднее значение «балансирует» баллы и находится в центре распределения.

2. Среднее значение минимизирует изменение оценок. Вторая характеристика среднего называется принципом «наименьших квадратов», который выражается в утверждении

$$\sum \left(X_i - \overline{X}\right)^2$$ = minimum или: среднее значение - это точка в распределении, вокруг которой сводится к минимуму изменение оценок (как указано в квадрате различий). Если различия между оценками и средними возводятся в квадрат, а затем добавляются, итоговая сумма будет меньше суммы квадратов различий между оценками и любой другой точкой в распределении.

Чтобы проиллюстрировать этот принцип, рассмотрим таблицу 4.2. Столбец 1 таблицы показывает те же пять баллов, которые показаны в Таблице 4.1, а столбец 2 отображает различия между баллами и средним значением. В столбце 3 различия между оценками и средними возводятся в квадрат, а сумма этих различий составляет 388.

Если бы мы выполняли те же математические операции с любым числом, отличным от среднего, скажем, 77, то результатом была бы сумма, превышающая 388. Эта точка проиллюстрирована в столбце 4 таблицы 4.7, который показывает, что сумма квадратов разностей вокруг 77 - 393, значение больше 388.

Этот принцип наименьших квадратов подчеркивает тот факт, что среднее значение ближе ко всем оценкам, чем другие показатели центральной тенденции. Кроме того, эта характеристика среднего важна для статистических методов корреляции и регрессии, тем, которые мы поднимаем в конце этого текста.

3. На среднее значение влияют все оценки, и оно может вводить в заблуждение, если распределение имеет «выбросы». Последняя важная характеристика среднего значения заключается в том, что все оценки в распределении включаются в его. Напротив, режим отражает только наиболее распространенную оценку, а медиана имеет дело только с оценкой среднего случая.

С одной стороны, эта характеристика является преимуществом, поскольку среднее использует всю доступную информацию. С другой стороны, когда у распределения есть внешние или некоторые чрезвычайно высокие или низкие оценки, среднее значение может ввести в заблуждение: оно может не представлять центральную или типичную оценку. Распределения с выбросами имеют перекос. Если есть какие-то чрезвычайно высокие оценки, это называется положительной пилой, и распределение с несколькими очень низкими баллами есть отрицательный перекос.

Следует помнить, что среднее значение будет вытягиваться в направлении отдаленных оценок относительно медианы. При положительной асимметрии среднее значение будет больше, чем медиана, и с отрицательной асимметрией произойдет противоположное.

Почему это проблематично? Поскольку медиана использует только средние значения, она всегда будет отражать центр распределения. Среднее значение, поскольку оно использует все случаи (включая выбросы), может быть намного выше или ниже, чем основная часть баллов, и давать ложное представление о центральности.

Таблица 4.1 Пример, что все очки уравновешиваются вокруг среднего

X_i	$X_i - \bar{X}$
65	65-78 = -13
73	73-18 = -5
77	77-78 = -1
85	85-78 = 7
90	90-78 = 12
$\sum X_i = 390$	$\sum(X_i - \bar{X}) = 0$
$\bar{X} = 392/5 = 78$	

Таблица 4.2. Пример, что средним является минимизированное изменение

1	2	3	4
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - 77)^2$
65	65-78 = -13	$(-13)^2 = 169$	$65 - 77 = (-12)^2 = 144$
73	73-18 = -5	$(-5)^2 = 25$	$73 - 77 = (-4)^2 = 16$
77	77-78 = -1	$(-1)^2 = 1$	$77 - 77 = (0)^2 = 0$
85	85-78 = 7	$(7)^2 = 49$	$85 - 77 = (8)^2 = 64$
90	90-78 = 12	$(12)^2 = 144$	$90 - 77 = (13)^2 = 169$
$\sum X_i = 390$	$\sum(X_i - \bar{X}) = 0$	$\sum(X_i - \bar{X})^2 = 388$	$\sum(X_i - 77)^2 = 393$

Таблица 4.3. Пример, что среднее затронут каждым счетом

1	2	3	4	5	6
Очки	Меры из Центральных Тенденция	Очки	Меры из Центральных Тенденция	Очки	Меры из Центральных Тенденция
15		15		0	
20	Средний = 25	20	Средний = 718	20	Средние = 22
25	Медиана = 25	25	Медиана = 25	25	Медиана = 25
30		30		30	
35		3500		35	

Медиана.

Медиана (середина) – уровень показателя, который делит набор данных на 2 равные половины (50/50). Она не присваивает наблюдениям весовые коэффициенты исходя из того, на сколько они отдалены от средней точки, а лишь оценивает их в зависимости от расположения.

Развивая мысль можно также делить медиану на четверти — квартили:
0,25 квантиль — первый (нижний) квартиль;
0,5 квантиль — медиана — второй квартиль;
0,75 квантиль — третий (верхний) квартиль.

Еще один вариант разделить на децили, каждый из которых включает в себя 10% наблюдений. Например, если ваш расход топлива бензинового двигателя автомобиля в верхнем дециле общего распределения расходов топлива всех бензиновых двигателей, то это означает, ваш двигатель сжигает топлива больше, чем 90% остальных двигателей.

Разбив распределение на сотые доли получим процентили — 1% распределения: первый процентиль представляет нижний 1% данного распределения, а 99-й — его верхний 1%.

Рассмотрим набор нормально распределенных случайных чисел.

Рисунок 4.4 Нормально распределенных случайных чисел.

В данном примере видим идеальную ситуацию, когда медиана, среднее арифметическое и мода совпадают. Но, если рассмотреть ассиметричное распределение, которое может возникать при проведении технических замеров, например, скорости, может сложиться такая ситуация.

Рисунок 4.5 Медиана и Средняя арифметическая

Как видим из графика у нас присутствуют аномальные значения («отщепенцы»): 23, 28, 30, влияющие на среднее арифметическое, но никак не затрагивающие медиану.
Медиана — альтернатива среднему арифметическому, устойчивая к аномальным отклонениям («отщепенцам»).

Выбор меры центральной тенденции

Вы должны учитывать два основных критерия при выборе меры тенденции в сантиметрах. Сначала убедитесь, что вы знаете уровень измерения рассматриваемой переменной. Как правило, это говорит вам, следует ли сообщать о моде, медиане или значении.
В таблице 4.4. показана взаимосвязь между уровнем измерения и показателями центральной тенденции. Жирным шрифтом «ДА» обозначен наиболее подходящий показатель центральной тенденции для каждого уровня измерения, а не жирным шрифтом «Да» указаны уровни измерения, для которых измерение также разрешено. Запись «Нет» в таблице означает, что статистика не может.

Таблица 4.4. Измерения и меры центральной тенденции

Мера центральной тенденции	Уровень измерения		
	Номинал	Порядковый	Отношение интервала
Среднее значение	Да	Да	Да
Медиана	Нет	Да	Да
Меры	Нет	Да (?)	Да

Вывод:

При выборе меры центральной тенденции нужно учитывать ее устойчивость к значениям, резко отклоняющихся от центра применяемых в каждом конкретном случае. Нужно определить какое влияние оказывают «отщепенцы»: искажают его или наоборот играют важную роль.
Окончательный выбор меры центральной тенденции всегда лежит за исследователем.

При этом необходимо учесть

1. Три показателя центральной тенденции, представленные в этой главе, имеют общую цель. Каждый сообщает некоторую информацию о наиболее типичной или представительной ценности в распределении. Надлежащее использование этих статистических данных позволяет исследователю сообщать важную информацию обо всем распределении баллов в одном, легко понимаемом числе.

2. Режим сообщает о наиболее распространенной оценке и наиболее целесообразно используется с переменными номинального уровня.

3. Медиана (М сообщает оценку, которая является точным центром распределения). Она наиболее целесообразно используется с переменными, измеренными на порядковом уровне, и с переменными отношения интервалов, когда распределение искажено.

4. Среднее значение, наиболее часто используемое из трех показателей, сообщает о наиболее типичном балле. Он наиболее целесообразно использовать с переменными, измеренными на уровне отношения интервалов (кроме случаев, когда распределение сильно искажено).

5. Среднее имеет ряд важных математических характеристик. Во-первых, это распределение баллов, вокруг которого все остальные баллы аннулируются. Во-вторых, среднее - это точка минимального отклонения (это принцип «наименьших квадратов»). Наконец, среднее значение зависит от количества очков в распределении и поэтому тянется в направлении крайних значений.