ТВИМС 4

Математическая статистика

Генеральная совокупность и выборка

Изучая социально-экономические процессы, мы рассматриваем как отдельные объекты, так и их совокупности. Например, человек, промышленная организация, торговая организация являются объектами следующих совокупностей: население региона, промышленные предприятия и торговые организации региона.

Каждый объект или элемент наблюдения обладает определенным набором качественных или количественных признаков, которые могут меняться при переходе от одного элемента к другому. Например, каждый студент факультета может характеризоваться такими количественными признаками как: возраст, рост, вес, также его могут характеризовать некоторые качественные признаки: цвет глаз, пол или семейное положение, также каждый студент характеризуется определенным уровнем рейтинга успеваемости. Таким образом, каждой совокупности объектов наблюдения соответствуют определенные совокупности значений того или иного признака.

При изучении какого-либо признака, присущего всем элементам совокупности, представляет интерес распределение его значений среди этих элементов. То есть, какое количество, или какая их доля, обладает некоторым конкретным значением рассматриваемого признака. Или для какого количества элементов значение рассматриваемого признака попадает в интересующий наблюдателя интервал значений.

Если же нас интересуют события, которые произойдут в будущем то, в этом случае мы можем говорить, лишь о вероятности того, что данное событие произойдет или не произойдет. Например, в середине лета, мы с большой вероятностью знаем, что на следующий день не выпадет снег. Но, жизненный опыт подсказывает, что природа иногда преподносит сюрпризы, а это значит что событие «утром выпал снег» возможно и летом, но оно маловероятно. Причина наших сомнений заключается в том, что мы не можем знать все факторы, которые влияют на то или иное событие. То есть, вероятность это не свойство природы, а результат неполноты наших знаний о ней.

На вопрос: «Какой процент студентов Университета в будущую сессию сдаст экзамен на «хорошо» и «отлично?» однозначный ответ дать нельзя. Доля студентов, успешно сдавших экзамен, в будущем может быть любой в интервале от 0% до 100%, но вероятность для каждого значения из этого интервала будут различной. Иначе говоря, она будет определенным образом распределена между всеми возможными значениями доли студентов. Мы можем сделать оценку распределения этих вероятностей, используя подобные распределения за предшествующие несколько лет. Таким образом, для оценок характеристик будущих событий используют конкретные данные по событиям, которые уже произошли.

Теория вероятностей изучает закономерности в случайных событиях, которые могут произойти в будущем, а также закономерности в генеральных совокупностях. ***Целью математической статистики является оценка параметров этих закономерностей на основе данных, получаемых в ходе наблюдений за выборочными совокупностями.***

Основными задачами математической статистики является разработка методов сбора, описания и анализа статистических данных.

Определение. Совокупность, объединяющая все множество элементов, называется генеральной совокупностью. Эта совокупность может быть конечной или бесконечной.

Определение. Выделенная по определенным правилам часть генеральной совокупности называется выборочной совокупностью или выборкой. Эта совокупность может быть только конечной.

Определение. Число объектов в генеральной или выборочной совокупности называется объемом совокупности.

Характеристики выборочной совокупности определяются непосредственно и точно. Для характеристик генеральной совокупности делаются оценки на основе выборочных данных.

Итак, выборка это своего рода модель генеральной совокупности. Для того, чтобы правильно отражать основные свойства генеральной совокупности, выборка должна удовлетворять двум требованиям репрезентативности:

1. отбор элементов в выборочную совокупность должен быть случайным;
2. выборка должна правильно отражать структурные соотношения генеральной совокупности.

Различают следующие способы формирования выборок.

1. Генеральная совокупность не делится на части:

- случайный бесповторный отбор, при котором отобранный объект в генеральную совокупность не возвращается;
- случайный повторный отбор, при котором объект перед следующим отбором возвращается в генеральную совокупность.

2. Генеральная совокупность делится на части:

- характерный отбор, при котором объекты случайным образом отбираются из групп, на которые определенным образом делится генеральная совокупность;
- серийный отбор, при котором случайным образом отбираются целые группы объектов;
- механический отбор, объекты выбираются через определённый интервал.

Предположим, имеется генеральная совокупность. Каждый ее элемент характеризуется некоторыми количественными признаками, значения которых заранее неизвестны и определяются в процессе наблюдения. Наблюдаемый количественный признак генеральной совокупности можно интерпретировать как случайную величину X, принимающую в процессе наблюдения (испытания) определенные значения.

Определение. Полученные в результате опыта значения случайной величины называются вариантами.

Определение. Множество вариант, расположенных в порядке возрастания, называется вариационным рядом.

Статистическое распределение выборки. Полигон и гистограмма

Пусть из генеральной совокупности извлечена выборка при этом варианта x_1 наблюдается n_1 раз, x_2 наблюдается n_2 раз, и так далее, x_k наблюдается n_k раз. В этом случае объём выборки будет $n = \sum_{i=1}^{k} n_i$.

Число n_i называются частотой варианты x_i, а $w_i = \frac{n_i}{n}$ - относительной частотой этой варианты. Заметим, что утверждение теоремы Пуассона, позволяет нам рассматривать относительную частоту w_i, как оценку вероятности p_i события $X=x_i$.

Определение. Статистическим распределением выборки называется перечень вариант и соответствующих им частот или относительных частот.

Выборочное распределение может быть представлено в табличном или графическом виде.

Определение. Таблицей частот называется таблица, состоящая из двух строк, в первой строке записывают значения вариант, а во второй – значения соответствующих частот или относительных частот.

Определение. Полигоном частот (относительных частот) называют ломаную линию, соединяющую точки с координатами (x_i, n_i) или (x_i, w_i).

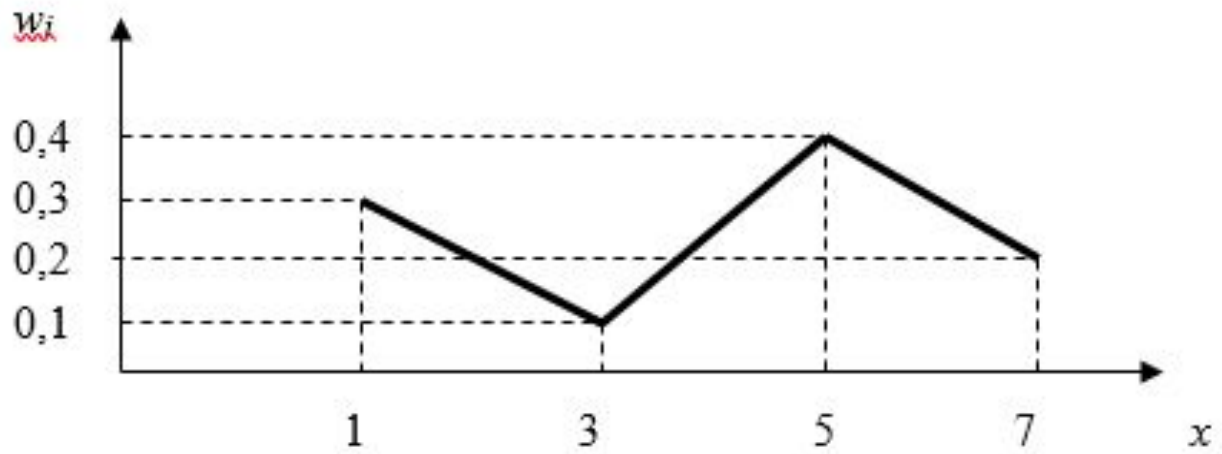
В качестве примера рассмотрим выборку объема $n = 10$:

1,1,1,3,5,5,5,5,7,7.

Статистическое распределение выборки представим в виде таблицы частот:

x_i	1	3	5	7
n_i	3	1	4	2
w_i	0,3	0,1	0,4	0,2

На рисунке изображен полигон относительных частот рассматриваемой выборки



В том случае, когда количество вариант велико, область, которой они принадлежат, разбивают на интервалы. Интервалы могут быть равной ширины или отличаться по ширине. Левый конец каждого интервала считают закрытым, а правый открытым. Таким образом, интервалы имеют вид $[x_i, x_{i+1})$. Частотой в этом случае считается число вариант n_i находящихся внутри i-го интервала. Интервальное распределение выборки может быть представлено в виде интервальной таблицы частот. Графически интервальный вариационный ряд представляют в виде гистограммы.

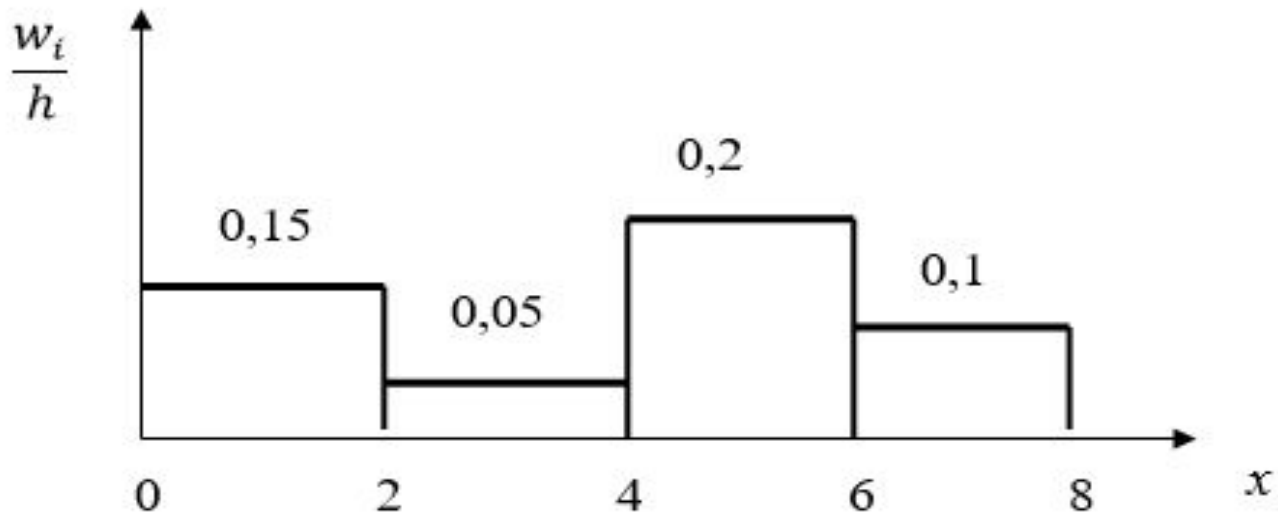
Определение. Гистограммой частот (относительных частот) называют ступенчатую фигуру, состоящую из прямоугольников, основания которых равны $h_i = x_{i+1} - x_i$, а высоты $\frac{n_i}{h}$ или $\frac{w_i}{h}$.

В качестве примера разобьем рассмотренную ранее выборку на интервалы равной ширины h=2.

Представим интервальное распределение выборки в виде таблицы:

$[x_{i+1} - x_i)$	0-2	2-4	4-6	6-8
n_i	3	1	4	2
w_i	0,3	0,1	0,4	0,2

Гистограмма относительных частот для этой выборки показана на рисунке



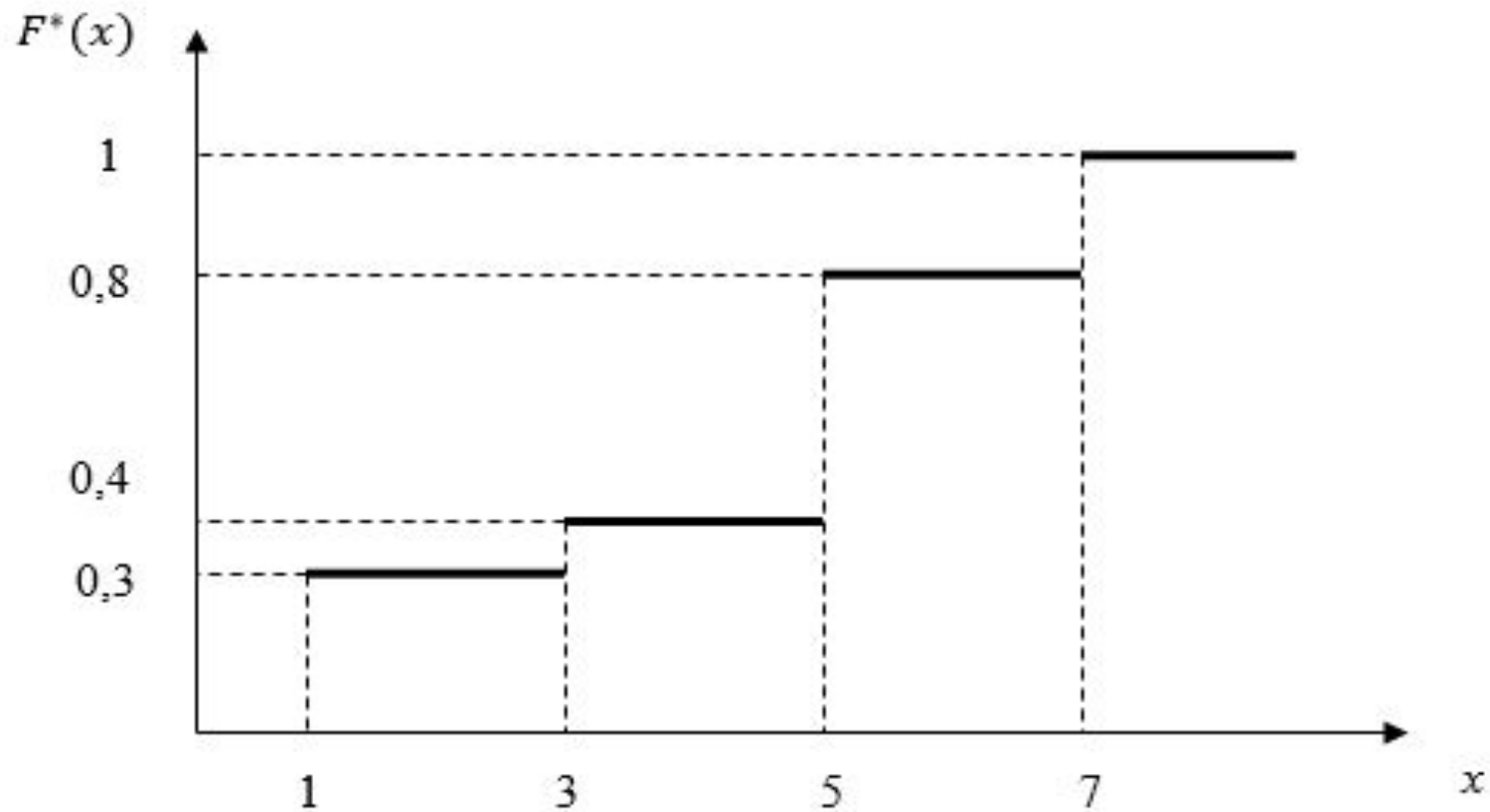
Определение. Эмпирической функцией распределения случайной величины называют функцию $F^*(x)$ относительной частоты события $X < x$.

$$F^*(x) = F_x^* = \frac{n_x}{n}$$

Графическое представление этой функции называется кумулятой.

Составим функцию распределения для выборки, представленной ранее:

$$F_x^* = \begin{cases} 0; & \text{если } x \le 1 \\ 0{,}3; & \text{если } 1 < x \le 3 \\ 0{,}4; & \text{если } 3 < x \le 5 \\ 0{,}8; & \text{если } 5 < x \le 7 \\ 1 & \text{если } 7 < x \end{cases}$$



ПРИМЕР. Анализируется выборка из 100 малых предприятий региона. Цель обследования – измерение коэффициента соотношения заемных и собственных средств (x_i) на каждом i-м предприятии. Результаты представлены в табл. 1.

Требуется построить гистограмму и график накопленных частот.

Решение. 1. Построим сгруппированный ряд наблюдений (табл. 2).

2. Определим в выборке $x_{\min} = 5{,}05$ и $x_{\max} = 5{,}85$.

3. Разобьем весь диапазон $[x_{\min}, x_{\max}]$ на k равных интервалов: $k \approx 1 + \log_2 100 = 7{,}62$; $k = 8$, отсюда длина интервала

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{5{,}85 - 5{,}05}{8} = 0{,}1.$$

Таблица 1

Коэффициенты соотношений
заемных и собственных средств предприятий

5,56	5,45	5,48	5,45	5,39	5,37	5,46	5,59	5,61	5,31
5,46	5,61	5,11	5,41	5.31	5,57	5,33	5,11	5,54	5,43
5,34	5,53	5,46	5,41	5,48	5,39	5,11	5,42	5,48	5,49
5,36	5,40	5,45	5,49	5,68	5,51	5,50	5,68	5,21	5,38
5,58	5,47	5,46	5,19	5,60	5,63	5,48	5,27	5,22	5,37
5,33	5,49	5,50	5,54	5,40	5.58	5,42	5,29	5,05	5,79
5,79	5,65	5,70	5,71	5,85	5,44	5,47	5,48	5,47	5,55
5,67	5,71	5,73	5,05	5,35	5,72	5,49	5,61	5,57	5,69
5,54	5,39	5,32	5,21	5,73	5,59	5,38	5,25	5,26	5,81
5,27	5,64	5,20	5,23	5,33	5,37	5,24	5,55	5,60	5,51

Таблица 2

Сгруппированный ряд наблюдений

Номер интервала	**Интервалы**	**Середины интервалов x_i**	w_i	w_i^C	$f_n(x)$
1	5,05–5,15	5,1	0,05	0,05	0,5
2	5,15–5,25	5,2	0,08	0,13	0,8
3	5,25–5,35	5,3	0,12	0,25	1,2
4	5,35–5,45	5,4	0,20	0,45	2,0
5	5,45–5,55	5,5	0,26	0,71	2,6
6	5,55–5,65	5,6	0,15	0,86	1,5
7	5,65–5,75	5,7	0,10	0,96	1,0
8	5,75–5,85	5,8	0,04	1,00	0,4

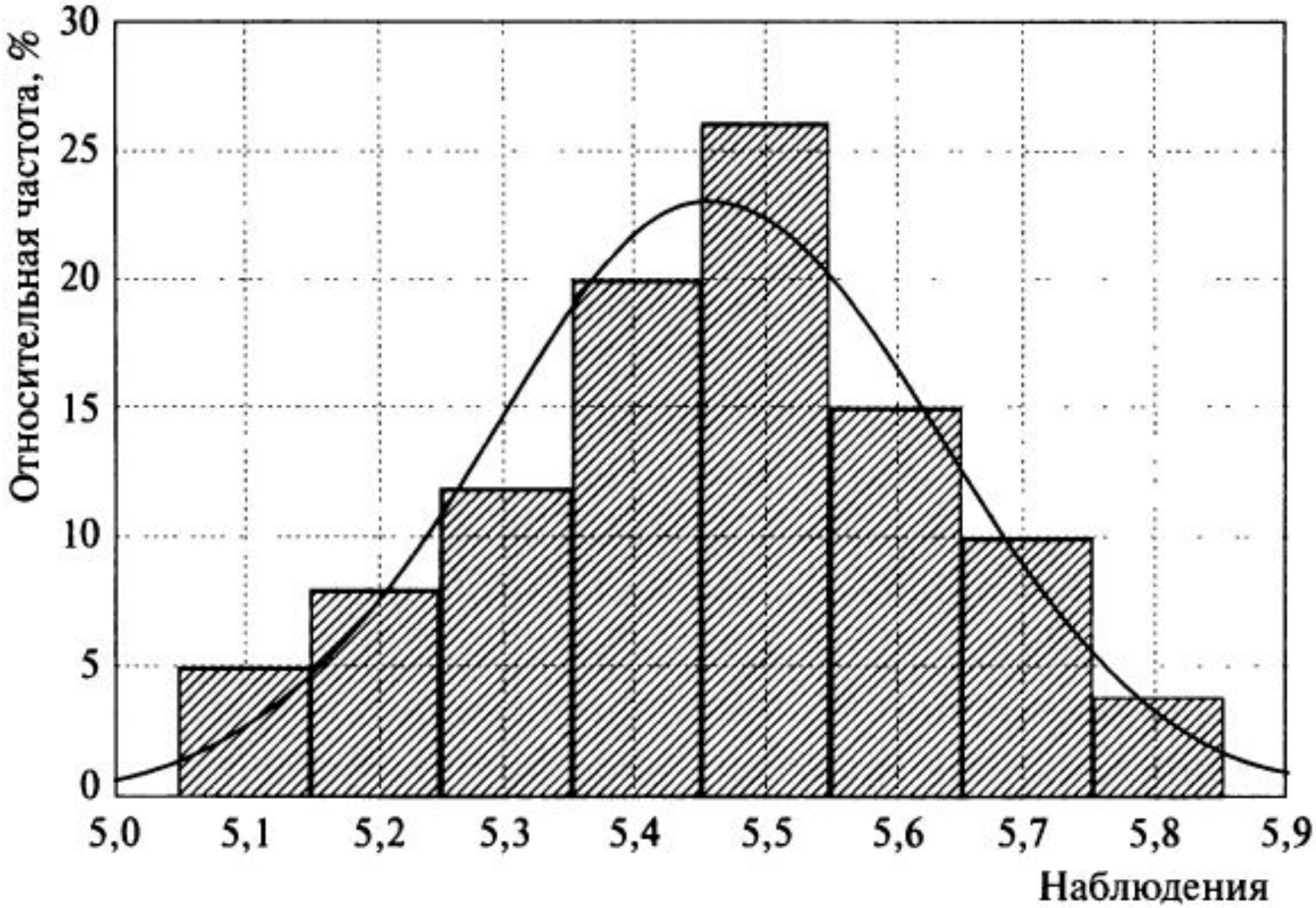


Рис. 12. Гистограмма

характеристики положения

Определение. Выборочной средней называется среднее арифметическое значений выборки.

$$\overline{X_в} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Здесь и в дальнейшем чертой сверху будем обозначать средние величины.

Значения выборочной средней для фиксированной выборки рассчитываются по формуле:

$$\overline{x_в} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \sum_{i=1}^{n} x_i w_i .$$

Определение. Выборочной медианой называется число равное среднему элементу выборки при нечетном n и среднему арифметическому двух средних элементов выборки при четном n

$$Me = X_{\frac{n+1}{2}} \ (n - \text{нечетное}), \ Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \ (n - \text{четное}).$$

Заметим, что при четном n значение медианы не является вариантой, хотя и может совпадать со значениями двух средних элементов, если они равны.

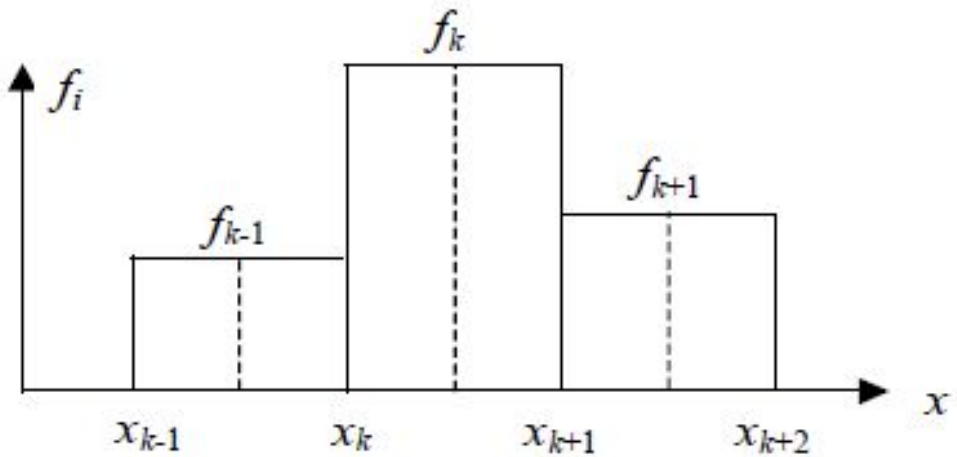
Определение. Модой вариационного ряда называется варианта с наибольшей частотой.

Замечание))

Мода M_0 – наиболее часто встречающееся значение в вариационном ряду. Для интервальной таблицы мода определяется в виде

$$M_0 = x_k + \frac{f_k - f_{k-1}}{2f_k - (f_{k-1} + f_{k+1})} \cdot h,$$

где входящие в формулу величины определяются из фрагмента гистограммы, представляющей собой интервал с наибольшей частотой и два соседних с ним интервала



Пример. Найти характеристики положения для выборки:

0,0,1,1,1,3,3,4,7,7,7,7,7,8,9,9

Решение.

Для удобства представим вариационный ряд в виде таблицы частот

x_i	0	1	3	4	7	8	9
n_i	2	3	2	1	5	1	2
w_i	0,1250	0,1875	0,1250	0,0625	0,3125	0,0625	0,1250

Находим выборочную среднюю:

$$\overline{x_в} = 0 \cdot 0{,}125 + 1 \cdot 0{,}1875 + 3 \cdot 0{,}125 + 4 \cdot 0{,}0625 + 7 \cdot 0{,}3125 + 8 \cdot 0{,}0625 + 9 \cdot 0{,}125 = 2{,}6875$$

Находим значения медианы и моды: $m_е$=(4+7)/2=5,5; $m_о$=7.

характеристики рассеивания

Определение. Размахом выборки называется разность между наибольшей и наименьшей вариантой:

$$R_в = X_{max} - X_{min}.$$

Определение. Выборочной дисперсией называется среднее арифметическое квадратов отклонений вариант от выборочной средней:

$$D_в = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_в)^2}{n}.$$

Определение. Выборочным средним квадратическим отклонением или стандартным отклонением называется корень квадратный из выборочной дисперсии.

$$d_в = \frac{\sum_{i-1}^{k}(x_i - \overline{x_в})^2 \cdot n_i}{n} = \sum_{i-1}^{n}(x_i - \overline{x_в})^2 w_i \ .$$

$$d_в = \sum (x_i - \overline{x_в})^2 w_i = \sum (x_i^2 w_i - 2x_i \overline{x}_в w_i + (\overline{x_в})^2 w_i) =$$

$$\sum x_i^2 w_i - 2\overline{x}_в \sum x_i w_i + (\overline{x}_в)^2 \sum w_i = \overline{x^2} - \left(\overline{x_в}\right)^2$$

$$\sigma_в = \sqrt{d_в}.$$

Задача??? (найти выборочное ско)

x_i	1	3	7	9	12
m_i	2	10	4	24	10

характеристики асимметрии и эксцесса

Определение. Начальным эмпирическим моментом порядка m выборки называется величина, значения которой рассчитываются по формуле

$$\nu_m^* = \frac{\sum_{i=1}^{k} n_i \cdot x_i^m}{n}.$$

Определение. Центральным эмпирическим моментом выборки порядка m называется величина, значения которой рассчитываются по формуле

$$\mu_m^* = \frac{\sum_{i=1}^{k} n_i \cdot \left(x_i - \overline{x_в}\right)^m}{n}.$$

Параметрами, характеризующими геометрическую форму выборочного распределения (полигона частот или гистограммы) служат выборочный коэффициент асимметрии и выборочный коэффициент эксцесса. Выборочный коэффициент асимметрии равен отношению центрального эмпирического момента третьего порядка к кубу выборочного среднего квадратического отклонения:

$$a_s = \frac{\mu_3^*}{\sigma_в^3}$$

Коэффициент асимметрии характеризует степень несимметричности выборочного распределения относительно среднего значения. Для симметричного распределения a_s=0, для положительной (правосторонней) асимметрии a_s>0. Для отрицательной (левосторонней) a_s <0.

Выборочный коэффициент эксцесса равен отношению центрального эмпирического момента четвертого порядка к выборочному среднему квадратическому отклонению в четвертой степени:

$$e_k = \frac{\mu_4^*}{\sigma_в^4} .$$

Коэффициент эксцесса характеризует степень *островершинности* выборочного распределения по сравнению с нормальным распределением. Для нормального распределения $e_k = 0$, для *островершинного* распределения $e_k > 0$, для *плосковершинного* $e_k < 0$.

Коэффициент вариации — ЭТО

отношение стандартного отклонения σ к среднему значению $\bar{x}$, выраженное в процентах: $V = \frac{\sigma}{\bar{x}} \cdot 100\%$

Коэффициент вариации и среднеквадратическое отклонение могут использоваться как меры риска, например, при финансовых операциях.

Коэффициент вариации может быть использован при сравнении стандартных отклонений, которые вычислены по данным, имеющим различные средние.

Пример. Предположим, что цены на ценные бумаги широко колеблются. Инвестор, который покупает акции по низкой цене, а продает по высокой, имеет хороший доход. Однако если цены на акции падают ниже стоимости, по которой инвестор купил, то он теряет доход.

Чтобы оценить меру риска, инвестор может использовать коэффициент вариации и среднеквадратическое отклонение.

Какую информацию о степени риска может дать коэффициент вариации по сравнению со среднеквадратическим отклонением?

Допустим, за пять недель цены:

на акции 1 представлялись в виде \$57, 68, 64, 71, 62;

на акции 2 представлялись в виде \$12, 17, 8, 15, 13.

Средняя цена на акции 1 $\overline{X}$ = \$64.40 и S = \$4.84.

Средняя цена на акции 2 $\overline{X}$ = \$13.00 и S = \$3.03.

Со среднеквадратическим отклонением как мерой риска акции 1 более рискованные. Однако среднее арифметическое акций 1 почти в 5 раз больше среднего арифметического акций 2. Коэффициент вариации, используемый в данном случае, дает следующие результаты:

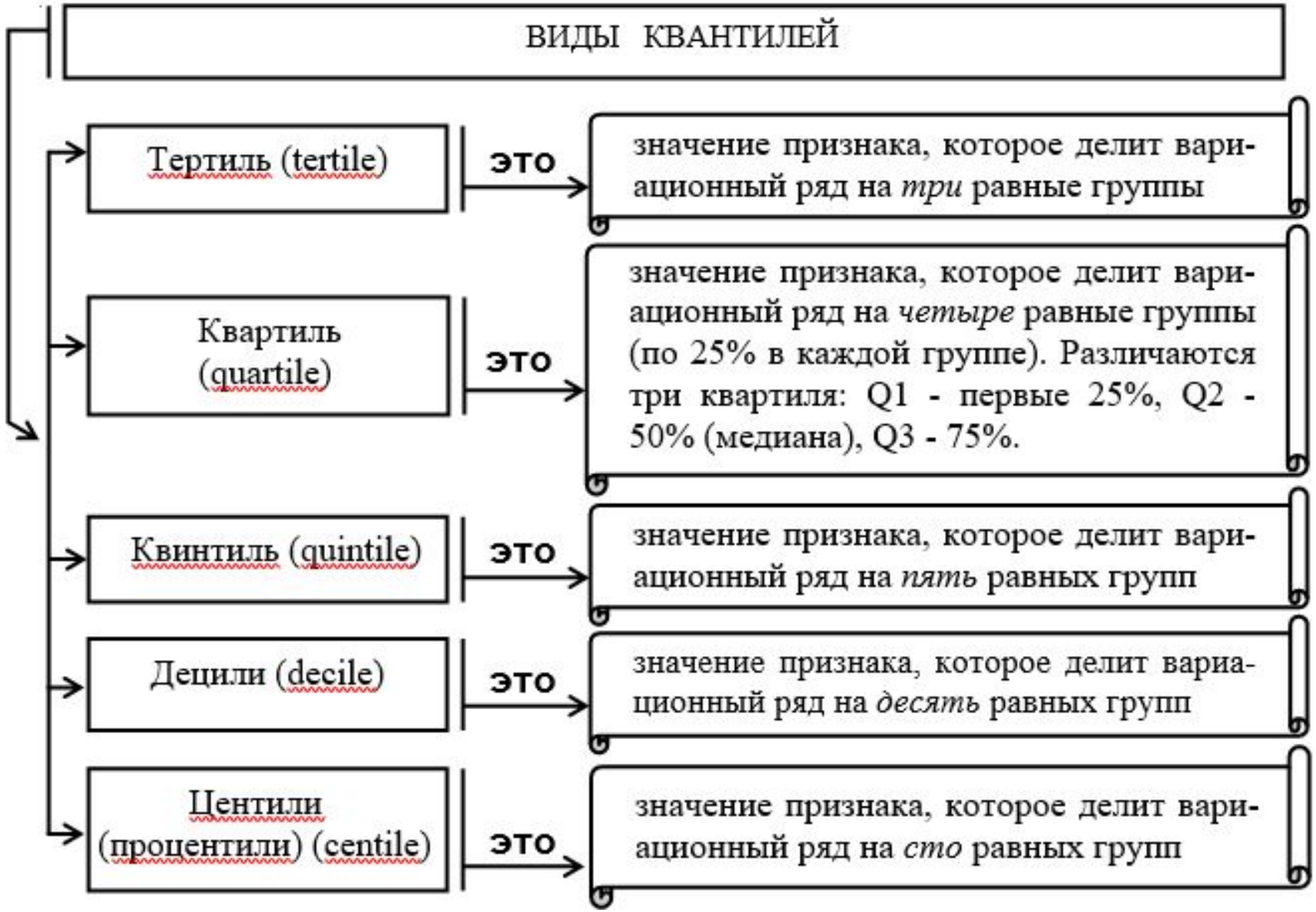
$$V_1 = \frac{4.84}{64.40} \cdot 100 = 7.52\ \%;$$

$$V_2 = \frac{3.03}{13} \cdot 100 = 23.31\ \%.$$

Для акций 2 коэффициент вариации почти в три раза больше, чем коэффициент вариации для акций 1.

Используя коэффициент вариации в данном случае, можно сделать заключение, что покупать акции 2 более рискованно.

При анализе результатов измерения значений признака иногда необходимо сгруппировать результаты в равные группы при помощи «точек деления» – *квантилей.*



Процентили широко используются в различного рода отчетах. Для того чтобы определить P-й процентиль, необходимо выполнить следующее:

- представить результаты наблюдений в виде вариационного ряда;
- вычислить номер P-го процентиля в вариационном ряду

$$i = \frac{P}{100}(n),$$

где P – значения процентиля;

n – объем выборки;

i – номер процентиля в ряду наблюдений;

- определить значение P-го процентиля:

а) если i – целое, то P-й процентиль является средней величиной i-го и $(i+1)$-го наблюдений в вариационном ряду;

б) если i не является целым числом, то номер P-го процентиля определяется как целая часть от значения $(i+1)$.

Пример. Определить 30-й процентиль для следующего ряда наблюдений: 14 12 19 23 5 13 28 17.

Решение. Вариационный ряд

5 12 13 14 17 19 23 28;

$$i = \frac{30}{100} \cdot 8 = 2.4 .$$

Так как i не является целым числом, то номер 30-го процентиля в данном вариационном ряду определяется как целая часть от значения 2.4+1=3.4, т. е. 3. Следовательно 30-м процентилем является значение $x_3 = 13$.

Задача????

Пример. Определить Q_1, Q_2, Q_3 для следующей выборки:

109 121 122 129 106 116 125 114.

Этика.

Представляя письменный отчет, результаты необходимо излагать честно, нейтрально и объективно. Неэтичным является использование среднего арифметического для явно асимметричных данных, для данных с резко выделяющимися наблюдениями. При использовании числовых характеристик средней тенденции необходимо указывать соответствующие характеристики рассеивания: для среднего арифметического такой характеристикой является стандартное отклонение, для медианы соответствующей характеристикой рассеяния является интерквартильный размах, для моды – размах.

Статистические оценки

О параметрах генеральной совокупности мы знаем то, что они объективно существуют, но определить их непосредственно невозможно в силу того, что генеральная совокупность или бесконечна или чрезмерно велика. Поэтому может стоять вопрос только об оценке этих характеристик.

Ранее было установлено, что для выборки, извлеченной из генеральной совокупности, при соблюдении условий репрезентативности, можно определить характеристики, которые являются аналогами характеристик генеральной совокупности.

Определение. Приближенные значения параметров распределения, найденные по выборке, называются оценкой параметра.

Обозначим оцениваемый параметр случайной величины (генеральной совокупности) как Θ, а его оценку, полученную с помощью выборки, $\tilde{\Theta}$.

Оценка $\tilde{\Theta}$ является случайной величиной, поскольку любая выборка является случайной. Оценки, полученные для разных выборок, будут отличаться друг от друга. Поэтому будем считать $\tilde{\Theta}$ функцией, зависящей от выборки: $\tilde{\Theta} = \tilde{\Theta}(X_в)$.

Определение. Статистическая оценка называется состоятельной, если она стремиться по вероятности к оцениваемому параметру:

$$P_{n\to\infty}(\tilde{\Theta} = \Theta) = 1.$$

Это равенство означает, что событие $\tilde{\Theta} = \Theta$ становится достоверным при неограниченном возрастании объема выборки.

Определение. Статистическая оценка называется эффективной, если она имеет наименьшую дисперсию при одних и тех же объёмах выборки.

Рассмотрим оценку $\tilde{M}_X$ математического ожидания M_X случайной величины X. В качестве такой оценки выберем $\bar{X}_в$. Найдем математическое ожидание случайной величины $\bar{X}_в$.

Сначала сделаем важное утверждение: учитывая то, что все случайные величины X_i извлекаются из одной и той же генеральной совокупности X, а значит, имеют одно и то же распределение что и X, можно записать:

$$M(X_i) = M_X.$$

Теперь найдем $M(\bar{X}_в)$:

$$M(\bar{X}_в) = M\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}\left(M(X_1) + M(X_2) + \cdots + M(X_n)\right) = \frac{nM_X}{n} = M_X.$$

Мы установили, что в рассматриваемом случае математическое ожидание выбранной нами оценки (случайной величины) равно самому оцениваемому параметру. Оценки, обладающие таким свойством, занимают особое место в математической статистике, они называются несмещенными.

Определение. Статистическая оценка $\tilde{\theta}$ называется несмещенной, если её математическое ожидание равно оцениваемому параметру

$$M(\tilde{\theta}) = \theta.$$

Если это требование не выполнено, то оценка называется смещенной.

Таким образом, выборочная средняя является несмещенной оценкой математического ожидания.

Проведем анализ смещенности выборочной дисперсии $D_в$, если ее выбрать в качестве оценки генеральной дисперсии D_X.

$$M(D_в) = M(\bar{X}^2 - (\bar{X})^2) = M\left(\frac{\sum x_i^2}{n} - \frac{(\sum x_i)^2}{n^2}\right) = M\left(\frac{\sum x_i^2}{n}\right) - M\left(\frac{(\sum x_i)^2}{n^2}\right).$$

Преобразуем каждое из двух полученных слагаемых:

$$M\left(\frac{\sum X_i^2}{n}\right) = \frac{1}{n} M\left(\sum X_i^2\right) = \frac{1}{n}\sum M(X_i^2) = \frac{n}{n} M(X^2) = M(X^2).$$

Здесь было использовано равенство $M(X_i^2) = M(X^2)$, (смотри ранее).

Рассмотрим второе слагаемое. С помощью формулы квадрата суммы n слагаемых получаем

$$M\left(\frac{(\sum X_i)^2}{n^2}\right) = \frac{1}{n^2}\sum M(X_i^2) + \frac{2}{n^2}\sum M(X_1X_2 + X_1X_3 + \cdots + X_2X_3 + \cdots + X_{n-1}X_n),$$

учитывая также то, что X_i и X_j независимые случайные величины запишем $M(X_iX_j) = M(X_i)M(X_j) = M_x^2$, $M(X_i^2) = M(X^2)$ и окончательно получим:

$$M\left(\frac{(\sum X_i)^2}{n^2}\right) = \frac{1}{n}M(X^2) + \frac{2}{n^2}\frac{(n-1)n}{2}M_x^2 = \frac{1}{n}M(X^2) + \frac{n-1}{n}M_x^2.$$

Подставим полученные результаты

$$M(D_в) = M(X^2) - \frac{1}{n}M(X^2) - \frac{n-1}{n}M_x^2.$$

После преобразования получим

$$M(D_в) = \frac{n-1}{n}(M(X^2) - M_x^2) = \frac{n-1}{n}D_x.$$

Таким образом, можно сделать вывод, что выборочная дисперсия является *смещенной* оценкой генеральной дисперсии.

Учитывая полученный результат, поставим задачу построить такую оценку генеральной дисперсии, которая удовлетворяла бы условию несмещенности. Для этого рассмотрим случайную величину

$$S^2 = \frac{n}{n-1} D_{\text{в}}.$$

Легко видеть, что для этой величины условие выполняется:

$$M(S^2) = D_x.$$

Следовательно, S^2 можно считать несмещенной оценкой генеральной дисперсии. Эта величина называется исправленной выборочной дисперсией. Значение исправленной дисперсии для конкрктной выборки рассчитывается по формуле

Оценка $s^2 = \frac{n}{n-1} D_B$ для генеральной совокупности не является эффективной, но для **нормального распределения** величины X оценка является **асимптотически эффективной**, т.е., при $n \to \infty$ отношение дисперсии этой оценки к минимально возможной дисперсии оценки стремится к единице.

Методы нахождения оценок

При выборе оценок характеристик случайных величин важно знать их точность. В некоторых случаях требуется высокая точность, а иногда достаточно иметь грубую оценку. Например, планируя перелет с пересадкой нам важно знать как можно точнее планируемое время прилета к месту стыковки авиарейсов. В другой ситуации, например, находясь дома и ожидая курьера с заказанным нами товаром, высокая точность времени его прибытия для нас не важна. В обоих случаях случайной величиной является время прибытия, а интересующей нас характеристикой случайной величины – среднее время в пути.

Оценки бывают двух видов. В первом случае ставится задача получить конкретное числовое значение параметра. В другом случае определяется интервал, в который с заданной вероятностью попадает интересующий нас параметр.

Определение. Оценка, определяемая одним параметром, называется точечной оценкой.

Наиболее простым и удобным методом построения точечных оценок является метод моментов, предложенный К. Пирсоном. Суть метода заключается в том, что устанавливается приближенное равенство между выборочными и теоретическими начальными и центральными моментами:

$$\nu_k^* = \nu_k ,$$

$$\mu_k^* = \mu_k .$$

В общем случае система уравнений для моментов может не иметь решения в элементарных функциях (тогда можно искать решение приближенными методами) или вообще оказаться неразрешимой (несовместной).

Оценки, полученные методом моментов, часто оказываются смещенными. К достоинствам метода моментов следует отнести его простую вычислительную реализацию, а также то, что оценки являются функциями выборочных моментов.

Иногда оценки, получаемые с помощью метода моментов, принимаются в качестве первого приближения, по которому другими методами можно построить оценки более высокого качества.

Оценка одного параметра. Пусть известен вид плотности распределения $f(x,\theta)$, зависящей от одного параметра θ, но не известно значение этого параметра. Для нахождения оценки этого параметра достаточно составить одно уравнение, например, для начальных моментов первого порядка: $\overline{x}_B = M(X)$. Так как математическое ожидание признака генеральной совокупности $M(X) = \int_{-\infty}^{\infty} x \cdot f(x,\theta)dx = \varphi(\theta)$ зависит от неизвестного параметра θ, а выборочное среднее $\overline{x}_B = \frac{x_1 + x_2 + ... + x_n}{n}$ зависит от реализации выборки $x_1, x_2, ..., x_n$, то после решения уравнения $\varphi(\theta) = \frac{x_1 + x_2 + ... + x_n}{n}$ мы получаем $\theta^* = \psi(x_1, x_2, ..., x_n)$, оценку неизвестного параметра как функцию значений конкретной выборки.

Пример. Найти оценку параметра λ для экспоненциального распределения.

Решение. Плотность экспоненциального распределения имеет вид

$$f(x,\lambda)=\begin{cases} 0, & x<0 \\ \lambda\cdot e^{-\lambda x}, & x\geq 0 \end{cases}$$

По выборке найдём начальный выборочный момент $\nu_1^* = \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_в$.

Для экспоненциального распределения

$$\nu_1 = \int_{-\infty}^{\infty} x\cdot f(x,\lambda)dx = \frac{1}{\lambda}.$$

Используя метод моментов, запишем:

$$\frac{1}{\lambda} = \overline{x_в}. \qquad \lambda^* \approx \lambda = \frac{1}{\overline{x_в}}.$$

Оценка двух параметров. Пусть задан вид плотности распределения, зависящей от двух неизвестных параметров, $f(x,\theta_1,\theta_2)$. Для их оценки можно, например, приравнять начальные моменты первого порядка и центральные моменты второго порядка, т.е.,

$$\alpha_k^*[X]=\overline{x^k}=\frac{1}{n}\sum_{i=1}^{n}(x_i)^k,\ \mu_2^*[X]=\mu_2[X],\ \text{или}\ M(X)=\overline{x}_B,\ D(X)=D_B.$$

Аналогично случаю одного параметра, теоретические моменты есть функции параметров θ_1,θ_2, а выборочные моменты зависят от реализации выборки. Решая полученную систему относительно неизвестных параметров, получаем их точечные оценки:

$$\theta_1^*=\psi_1(x_1,x_2,...,x_n),\qquad \theta_2^*=\psi_2(x_1,x_2,...,x_n).$$

Задача????

Случайная величина X распределена равномерно.

$$f(x) = \begin{cases} \dfrac{1}{b-a}, \text{если } x \in (a,b) \\ 0, \text{если } x \notin (a,b) \end{cases}$$

Получена выборка

x_i	2	3	4	5	6
n_i	4	6	5	12	8

Найти оценку параметров a и b.

метод максимального правдоподобия

Метод максимального правдоподобия, предложенный Р. Фишером, может быть применен для точечной оценки параметров распределения как дискретных, так и непрерывных случайных величин.

Дискретные случайные величины. Пусть X – дискретная случайная величина, для которой в результате опыта получена выборка значений $x_1, x_2, ..., x_n$.

Вид закона распределения известен; закон содержит неизвестный параметр θ, для которого требуется найти точечную оценку на основании данных выборки.

Обозначим вероятность $P(X = x_i) = p(x_i; \theta)$, $i = 1, 2, \ldots, n$.

Функцией правдоподобия дискретной случайной величины *X* называют функцию аргумента θ и данных выборки $x_1, x_2, ..., x_n$:

$$L(x_1, x_2, ..., x_n; \theta) = p(x_1; \theta) \cdot p(x_2; \theta) \cdot ... \cdot p(x_n; \theta).$$

В качестве точечной оценки параметра θ принимается значение $\theta^* = \theta^*(x_1, x_2, ..., x_n)$, при котором функция правдоподобия достигает наибольшего значения. Полученную оценку называют **оценкой максимального правдоподобия.**

Так как функции L и $\ln L$ достигают максимума при одном и том же значении θ, при практических вычислениях чаще используют вторую функцию, называемую **логарифмической функцией правдоподобия**:

$$\ln L(x_1, x_2, ..., x_n; \theta) = \ln p(x_1; \theta) + \ln p(x_2; \theta) + ... + \ln p(x_n; \theta).$$

После решения уравнения $\frac{d\ln L}{d\theta}=0$, называемого **уравнением правдоподобия**, которое дает критические точки функции правдоподобия, необходимо отобрать точки максимума, пользуясь условием $\frac{d^2\ln L}{d\theta^2}<0$.

Оценки, полученные методом максимального правдоподобия:
1) состоятельны (но могут оказаться смещенными);
2) распределены асимптотически нормально (при $n\to\infty$ закон их распределения приближается к нормальному);
3) среди всех асимптотически нормальных оценок оценки, полученные методом максимального правдоподобия, имеют наименьшую дисперсию, т.е., если для параметра θ существует эффективная оценка $\theta^*_{эфф}$, то она совпадает с полученной методом максимального правдоподобия.
Метод особенно полезен при малых объемах выборок; к недостаткам метода можно отнести громоздкие вычисления.

Случайная величина X распределена по закону Пуассона:

$$P_m(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!},$$

где m – длина серии испытаний, x_i – число появлений события в i-й серии испытаний, n– объем выборки (число проведенных серий испытаний), λ – неизвестный параметр распределения. Методом максимального правдоподобия по выборке $x_1, x_2, ..., x_n$ оценить значение этого параметра.

Решение:

Составим логарифмическую функцию правдоподобия:

$$ln\,L(x_1,x_2,...,x_n;\lambda)=ln\left(\frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\cdot\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}\cdot...\cdot\frac{\lambda^{x_n}e^{-\lambda}}{x_n!}\right)=$$

$$=ln\left(\frac{\lambda^{\sum x_i}}{x_1!\cdot x_2!\cdot...\cdot x_n!}e^{-n\lambda}\right)=\left(\sum x_i\right)ln\,\lambda-n\lambda-ln\left(x_1!\cdot x_2!\cdot...\cdot x_n!\right).$$

Уравнение правдоподобия: $\frac{d\,ln\,L}{d\lambda}=\frac{\left(\sum x_i\right)}{\lambda}-n=0$,

его решение (критическая точка): $\lambda=\frac{\left(\sum x_i\right)}{n}=\overline{x_B}$.

Проверим выполнение достаточных условий экстремума: $\frac{d^2\,ln\,L}{d\lambda^2}=-\frac{\left(\sum x_i\right)}{\lambda^2}<0$, т.е., $\lambda=\overline{x_B}$ – точка максимума, и в качестве оценки максимального правдоподобия параметра λ распределения Пуассона нужно взять выборочное среднее $\lambda^*=\overline{x_B}$.

Непрерывные случайные величины. Пусть X – непрерывная случайная величина, для которой в результате опыта получена выборка значений $x_1, x_2, ..., x_n$. Вид плотности распределения известен; плотность содержит неизвестный параметр θ, $f(x) = f(x, \theta)$, для которого требуется найти точечную оценку на основании данных выборки.

Функцией правдоподобия непрерывной случайной величины X называют функцию аргумента θ и данных выборки $x_1, x_2, ..., x_n$:

$$L(x_1, x_2, ..., x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot ... \cdot f(x_n; \theta).$$

Дальнейшие вычисления аналогичны случаю дискретной случайной величины.

Случайная величина X распределена по показательному закону с плотностью распределения $f(x)=\lambda e^{-\lambda x}$ $(x \geq 0)$ и неизвестным параметром λ. Методом максимального правдоподобия по выборке $x_1, x_2, ..., x_n$ оценить значение этого параметра.

Решение. Составим логарифмическую функцию правдоподобия:

$$ln\,L(x_1,x_2,...,x_n;\lambda)=ln\left(\lambda e^{-\lambda x_1}\cdot\lambda e^{-\lambda x_2}\cdot...\cdot\lambda e^{-\lambda x_n}\right)=$$

$$=ln\left(\lambda^n e^{-\lambda\sum x_i}\right)=n\,ln\,\lambda-\lambda\left(\sum x_i\right).$$

Уравнение правдоподобия: $\frac{d\,ln\,L}{d\lambda}=\frac{n}{\lambda}-\left(\sum x_i\right)=0$,

его решение (критическая точка): $\lambda=\frac{n}{\left(\sum x_i\right)}=\left(\frac{\left(\sum x_i\right)}{n}\right)^{-1}=\frac{1}{\overline{x}_B}$.

Проверим выполнение достаточных условий экстремума: $\frac{d^2\,ln\,L}{d\lambda^2}=-\frac{n}{\lambda^2}<0$,

т.е., $\lambda=\frac{1}{\overline{x}_B}$ – точка максимума, и в качестве оценки максимального правдоподобия параметра λ показательного распределения нужно взять величину, обратную выборочному среднему, $\lambda^*=\frac{1}{\overline{x}_B}$, что совпадает с оценкой, полученной методом моментов.

Если плотность распределения зависит от двух неизвестных параметров, $f(x) = f(x;\theta_1,\theta_2)$, то и функция правдоподобия является функцией двух переменных, θ_1,θ_2. Для нахождения критических точек нужно решить систему

$$\begin{cases} \dfrac{\partial \ln L}{\partial \theta_1} = 0, \\ \dfrac{\partial \ln L}{\partial \theta_2} = 0, \end{cases}$$

Для того, чтобы в критической точке достигался максимум, достаточно, чтобы выполнялись условия $AC - B^2 > 0$, $A < 0$ (или $C < 0$), где A, B, C – значения вторых производных в критической точке $P(\theta_1, \theta_2)$:

$$A = \left.\frac{\partial^2 \ln L}{\partial x^2}\right|_P, \; B = \left.\frac{\partial^2 \ln L}{\partial x \partial y}\right|_P, \; C = \left.\frac{\partial^2 \ln L}{\partial y^2}\right|_P.$$

Случайная величина X распределена по нормальному закону с плотностью распределения $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$ и неизвестными параметрами a и σ. Методом максимального правдоподобия по выборке $x_1, x_2, ..., x_n$ оценить значение этих параметров.

Решение.

Составим логарифмическую функцию правдоподобия:

$$ln\,L\left(x_1,x_2,...,x_n;a,\sigma\right)=ln\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_1-a)^2}{2\sigma^2}}\cdot\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_2-a)^2}{2\sigma^2}}\cdot...\cdot\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_n-a)^2}{2\sigma^2}}\right)=$$

$$=ln\left(\frac{1}{\sigma^n\left(\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum(x_i-a)^2}\right)=-n\,ln\,\sigma-\frac{n}{2}ln\left(\sqrt{2\pi}\right)-\frac{1}{2\sigma^2}\sum\left(x_i-a\right)^2.$$

Уравнения правдоподобия:

$$\frac{\partial\,ln\,L}{\partial a}=\frac{\sum x_i-na}{\sigma^2}=0,\qquad\frac{\partial\,ln\,L}{\partial\sigma}=-\frac{n}{\sigma}+\frac{\sum\left(x_i-a\right)^2}{\sigma^3}=0,$$

их решение (критическая точка): $a=\frac{\left(\sum x_i\right)}{n}=\overline{x_B}$, $\sigma^2=\frac{\sum\left(x_i-a\right)^2}{n}=D_B$.

Проверка выполнения достаточных условий экстремума показывает, что при этих значениях действительно достигается максимум функции правдоподобия.

Распределения, связанные с нормальным.

Рассмотрим некоторые распределения, функционально связанные с нормальным, которые далее будут широко использоваться в задачах математической статистики.

Распределение χ^2 (Пирсона)

Рассмотрим совокупность независимых случайных величин $X_1, X_2, ..., X_n$, у которых математическое ожидание равно нулю, а среднее квадратическое отклонение – единице. Сумма квадратов этих величин распределена по закону, называемому «хи – квадрат с n степенями свободы»:

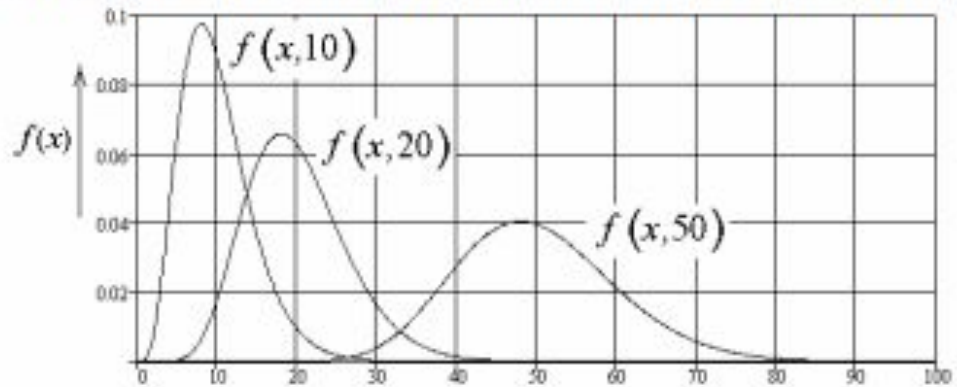
$$\chi^2 = \sum_{i=1}^{n} X_i^2 .$$

Если эти величины связаны одним линейным соотношением, например, $\sum_{i=1}^{n} X_i = n\overline{X}$, число степеней свободы уменьшается, $k = n - 1$.

Плотность этого распределения

$$f(x) = \begin{cases} 0, & x \le 0, \\ \dfrac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{\frac{x}{2}-1}, & x > 0, \end{cases}$$

Распределение Пирсона зависит от одного параметра – числа степеней свободы. С увеличением числа степеней свободы распределение приближается к нормальному, что можно наблюдать на приведенных ниже графиках.



Плотность распределения χ^2 Пирсона

t – распределение Стьюдента

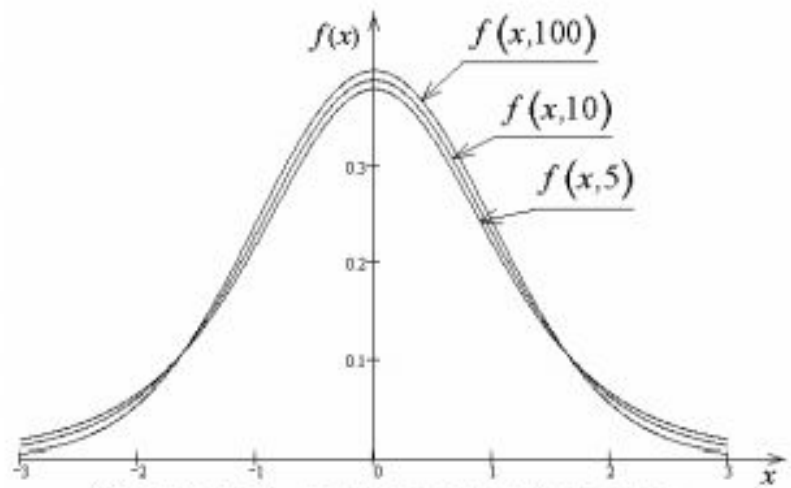
Пусть Z – стандартная нормальная величина (т.е. $M(Z)=0$, $\sigma(Z)=1$, а V – независимая от Z случайная величина, распределенная по закону χ^2 с k степенями свободы. Величина

$$T=\frac{Z}{\sqrt{\frac{\chi^2}{k}}}$$

распределена по закону, называемому t – распределением Стьюдента с k степенями свободы. Очевидно, распределение Стьюдента зависит от одного параметра – числа степеней свободы k. Плотность t – распределения Стьюдента выражается формулой

$$f(x,k)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}.$$

С увеличением числа степеней свободы распределение быстро приближается к нормальному, что можно наблюдать на приведенных ниже графиках.



Плотность t – распределения Стьюдента

Если независимые случайные величины U и V распределены по закону χ^2 с k_1 и k_2 степенями свободы соответственно, то величина

$$F = \frac{\left(U/k_1\right)}{\left(V/k_2\right)}$$

распределена по закону, называемому распределением Фишера – Снедекора со степенями свободы k_1 и k_2.

Плотность этого распределения

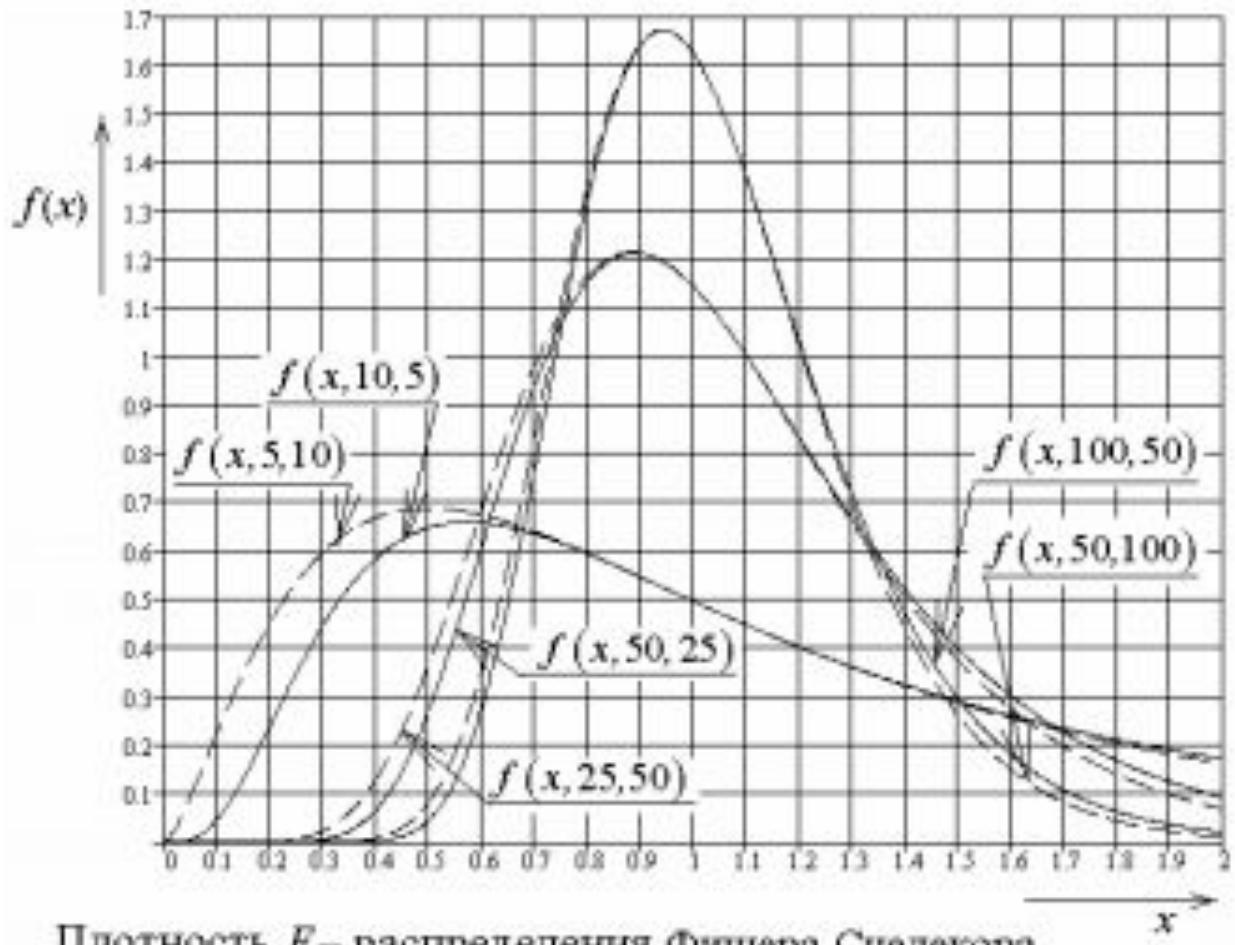
$$f(x) = \begin{cases} 0, & x \le 0, \\ C\dfrac{x^{\frac{k_1}{2}-1}}{\left(k_2 + k_1 x\right)^{\frac{k_1+k_2}{2}}} e^{-x/2}, & x > 0, \end{cases}$$

где

$$C = \frac{\Gamma\left(\dfrac{k_1+k_2}{2}\right)(k_1)^{\frac{k_1}{2}}(k_2)^{\frac{k_2}{2}}}{\Gamma\left(\dfrac{k_1}{2}\right)\Gamma\left(\dfrac{k_2}{2}\right)}.$$

***F* – распределение Фишера – Снедекора**

С увеличением числа степеней свободы распределение приближается к нормальному, что можно наблюдать на приведенных графиках.



Плотность F – распределения Фишера-Снедекора

Замечание…

Под числом степеней свободы понимается количество значений варьирования наблюдений, которые могут принимать «произвольные» значения, не изменяя общего уровня, около которого эти значения варьируют.

Пример. Необходимо вычислить среднее арифметическое по следующим наблюдениям:

0,24 0,02 0,01 0,22 0,04 0,14 0,18;

$$\overline{X} = \frac{0,95}{7} = 0,136.$$

Если перед нами поставлена задача произвольно отобрать еще одну совокупность таких же 7-ми значений, не изменяя при этом вычисленного уровня (т. е. $\overline{X} = 0.136$), то свободно варьирующих значений было бы не 7, а 6. Седьмое значение должно быть таким, чтобы общая сумма всех значений оказалась бы неизменной, т. е. $\sum x_i = 0.95$. Число степеней свободы в данном случае равно $7 - 1 = 6$, или объем выборки n минус 1.