Тема 3.1. ВЫБОРОЧНЫЙ МЕТОД

План лекции:

1. Выборка
2. Статистическое распределение выборки
3. Графическое изображение вариационных рядов
4. Эмпирическая функция распределения

Список литературы:

1. Вентцель, Е.С. Теория вероятностей [Текст] / Е.С. Вентцель. – М.: Высшая школа, 2006. – 575 с.
2. Гмурман, В.Е. Теория вероятностей и математическая статистика [Текст] / В.Е. Гмурман. - М.: Высшая школа, 2007. - 480 с.
3. Кремер, Н.Ш. Теория вероятностей и математическая статистика [Текст] / Н.Ш. Кремер - М: ЮНИТИ, 2002. – 543 с.

п.1 Выборка

Математическая статистика – это наука, занимающаяся разработкой методов сбора, регистрации и обработки результатов наблюдений (измерений) с целью познания закономерностей случайных массовых явлений.

Результаты измерений (наблюдений) называют *статистическими данными*.

Одним из основных способов сбора статистических данных является *выборочный метод*.

Во многих практических задачах, связанных с повторяющимися испытаниями, нельзя провести все возможные испытания, а можно проделать лишь доступную, выборочную их часть, а затем сделать обоснованный вывод. Например, условимся считать некоторый ноутбук стандартным, если продолжительность его работы составляет 5000 часов, в противном случае он считается нестандартным. Исследовать каждый ноутбук на продолжительность работы невозможно. Тогда как получить представление о качестве изготавливаемых ноутбуков? Для этого достаточно иметь сведения о качестве небольшого числа ноутбуков, отобранных случайно. Тогда по продолжительности работы отобранных приборов можно судить о качестве всей партии.

Совокупность всех возможных значений, или реализаций, исследуемых случайных величин называется *генеральной совокупностью*. Она может состоять из конечного или бесконечного множества значений, называемых *элементами генеральной совокупности*.

Выборочной совокупностью (или просто *выборкой*) называется совокупность элементов случайно отобранных из генеральной совокупности.

Объемом совокупности (генеральной или выборочной) называют число элементов этой совокупности.

Метод, основанный на том, что по данным обследования выборки, выделенной из генеральной совокупности, делается заключение о всей генеральной совокупности, называется *выборочным методом*.

Задача математической статистики состоит в исследовании свойств выборки и обобщении этих свойств на всю генеральную совокупность. Полученный при этом вывод называется *статистическим*.

Основное требование к выборке: она должна хорошо представлять генеральную совокупность, т.е. быть *репрезентативной* (*представительной*). Выборка будет репрезентативной, если её осуществлять случайным образом.

При составлении выборки можно поступать двумя способами: после того как объект отобран и над ним произведено наблюдение, он может быть возвращен либо не возвращен в генеральную совокупность. В соответствии со сказанным выборки подразделяются на повторные и бесповторные.

Повторной называют выборку, при которой отобранный элемент (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при которой отобранный элемент в генеральную совокупность не возвращается.

На практике чаще используется бесповторная выборка.

Кроме того, различают следующие способы составления выборки: а) простой (случайный), б) механический, в) типический, г) серийный.

Так, если занумеровать все элементы генеральной совокупности и затем изготовить карточки с такими же номерами, тщательно перемешать их и отобрать пачку карточек, то элементы генеральной совокупности с номерами извлечённых карточек образуют *простую (случайную) выборку*. Здесь возможно повторная и бесповторная выборка.

Если элементы генеральной совокупности выбираются через определённый интервал, то такая выборка называется *механической*. Например, при анализе качества ноутбуков, сходящих с конвейера, отбирается каждый 25 ноутбук.

Предположим теперь, что генеральную совокупность разбили на несколько неперекрывающихся групп и из каждой группы отобраны в случайном порядке объекты. Это *типический способ (районированная или стратифицированная выборка)* составления выборки. Типическим отбором пользуются тогда, когда обследуемый признак заметно колеблется в различных типических частях генеральной совокупности. Например, при определении рейтинга кандидатов в президенты на выборах страну делят на округа и в каждом округе определяется рейтинг кандидатов в президенты.

Наконец, *серийная (гнездовая или кластерная) выборка* образуется следующим образом. Генеральная совокупность делится на неперекрывающиеся группы. После этого случайным образом отбираются некоторые группы. Полученная выборка будет *серийной*.

На практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы. Например, иногда разбивают генеральную совокупность на серии одинакового объема, затем простым

случайным отбором выбирают несколько серий и, наконец, из каждой серии простым случайным отбором извлекают отдельные объекты.

Разумеется, если бы мы могли провести сплошное обследование всех элементов генеральной совокупности, то не нужно было бы применять никакие статистические методы, и саму математическую статистику можно было бы отнести к чисто теоретическим наукам. Однако такой полный контроль невозможен по следующим причинам. Во-первых, часто испытание сопровождается разрушением испытуемого объекта; в этом случае мы имеем выборку без повторения. Во-вторых, обычно необходимо исследовать весьма большое количество объектов, что просто невозможно физически, и т.д.

п.2 Статистическое распределение выборки

Как правило, результаты эксперимента или наблюдения дискретных случайных величин (первичные данные) сводятся в таблицу, в первой строке которой записывается номер i эксперимента, а во второй – соответствующий признак x_i, называемый *вариантой случайной величины*. Таблицы такого вида называются *статистическими рядами несгруппированных данных*. Таблица может включать данные о нескольких признаках (несколько видов вариант), но часто ограничиваются данными об одном признаке.

Статистический ряд несгруппированных данных не позволяет проводить содержательный анализ. Учитывая, что нередко статистические исследования охватывают совокупность численностью десятки и сотни тысяч объектов, возникает необходимость упорядочения первичных данных. Для этого используются статистические методы ранжирования и группировки. Иногда этих приёмов обработки статистических данных достаточно для последующего анализа. Чаще приходится прибегать к более сложным методам, но и тогда предварительное упорядочение является обязательной операцией.

Ранжированием называется расположение элементов совокупности в порядке возрастания или убывания величины соответствующих им вариантов.

Статистический ряд, расположенный по возрастанию вариант, называется *вариационным рядом*.

Ранжированный перечень содержит список элементов совокупности упорядоченный по возрастанию. Каждому элементу (и соответствующему ему варианту) приписан ранг – номер занимаемого им места. Одинаковые варианты получают одинаковый ранг.

После ранжирования данных легко заметить, что некоторые варианты повторяются несколько раз. Если представить совокупность в виде таблицы , в которой записано сколько раз встречаются совокупности с одинаковой вариантой, она станет ещё более обозримой и удобной для анализа по сравнению с ранжированным рядом. Этот приём называется дискретной группировкой.

Дискретной группировкой называется распределение совокупности вариантов по группам, содержащим одинаковые варианты.

Число, показывающее сколько раз (как часто) некоторый вариант x_i встречается в совокупности, называется *частотой* n_i (*абсолютной частотой*) данного варианта. Сумма всех частот равняется количеству элементов совокупности (*объему выборки*), т.е.

$$\sum n_i = n. \quad (1)$$

Относительной частотой (частостью) ω_i некоторого варианта x_i называется доля этого варианта в общем количестве данных, т.е. отношение частоты к объему выборки:

$$\omega_i = \frac{n_i}{n}. \quad (2)$$

Для удобства относительную частоту часто выражают в процентах, умножая результат на 100.

Соответствие между вариантами и их частотами (относительными частотами) называется *статистическим распределением выборки*.

Одновременно с понятием частоты и относительной частоты в сгруппированных совокупностях применяются понятия накопленной частоты и относительной частоты.

Накопленной частотой $n_i^{íàê}$ некоторого варианта x_i называется количество элементов ранжированной в порядке возрастания совокупности, имеющих значение признака меньше или равное данному:

$$n_i^{íàê} = n_1 + n_2 + \ldots + n_i. \quad (3)$$

Накопленной относительной частотой $\omega_i^{íàê}$ некоторого варианта x_i называется отношение накопленной частоты этого варианта $n_i^{íàê}$ к объему выборки:

$$\omega_i^{íàê} = \frac{n_i^{íàê}}{n}. \quad (4)$$

В тех случаях, когда число различных вариантов в совокупности велико или вариация является непрерывной при обработке статистических данных используется интервальная группировка.

Интервальной группировкой называется распределение совокупности вариантов на группы вариантов, лежащих в определённых границах.

Статистическая таблица, получаемая в результате интервальной группировки, называется *интервальным вариационным рядом*.

Максимальное значение варианта для конкретного интервала называется *верхней границей* $x_{i(\max)}$, а минимальное – *нижней границей* интервала $x_{i(\min)}$. *Величина интервала* – разность между верхней и нижней границами интервала:

$$h_i = x_{i(\max)} - x_{i(\min)}. \quad (5)$$

Понятия частоты, относительной частоты, накопленной частоты и накопленной относительной частоты интервального вариационного ряда аналогичны соответствующим понятиям дискретного вариационного ряда, но относятся не к отдельному признаку. А ко всему интервалу.

Ещё одним способом группировки совокупности является *комбинационная группировка* – распределение совокупности на группы по сочетанию (комбинации) нескольких признаков.

п. 4 Графическое изображение вариационных рядов

Для наглядности рассмотрения статистических данных ряды распределения представляются в графической форме. Наиболее широко используются следующие виды графического изображения вариационных рядов в прямоугольной системе координат: полигон, гистограмма, кумулятивная кривая.

Эти графики дают возможность представить характер варьирования значений признака, выявить состав изучаемой совокупности, её структуру и структурные сдвиги. При нанесении на единую координатную сетку, возможно сравнение нескольких вариационных рядов.

Полигоном (многоугольником) распределения называется графическое изображение вариационного ряда в прямоугольной системе координат, при котором величины признака (варианты) x_i откладываются на оси абсцисс, а частоты (или относительные частоты) на оси ординат.

Таким образом, *полигон частот* представляет собой ломанную, отрезки которой соединяют точки $M_1(x_1,\ n_1)$, $M_2(x_2,\ n_2)$, …, $M_k(x_k,\ n_k)$. *Полигон относительных частот* есть ломанная, отрезки которой соединяют точки $M_1(x_1,\ \omega_1)$, $M_2(x_2,\ \omega_2)$, …, $M_k(x_k,\ \omega_k)$. Крайние точки M_1 и M_2, если они не лежат на оси абсцисс, обычно также соединяют со смежными точками $M_0(x_0,\ 0)$, $M_{k+1}(x_{k+1},\ 0)$.

Гистограммой вариационного ряда называется графическое изображение интервального вариационного ряда в виде прямоугольников, основания которых – отрезки оси абсцисс, соответствующие интервалам изменения признака, а высоты пропорциональны плотностям частот (или относительных частот) интервалов.

В случае непрерывных интервалов гистограмма частот строится следующим образом (см. Рисунок 1): на оси абсцисс наносится шкала для интервалов, на оси ординат – для плотностей частот интервалов $\frac{n_i}{h_i}$. Из всех точек на оси абсцисс восстанавливаются перпендикуляры, на которых последовательно, начиная с первого, откладываются значения плотности частот интервалов.

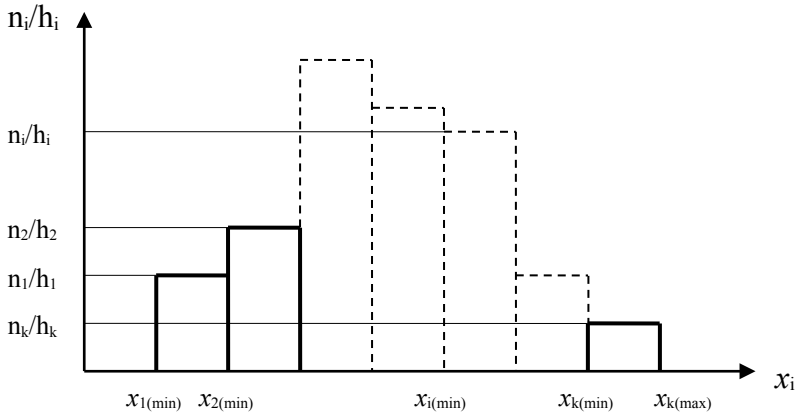

Рисунок 1 – Построение гистограммы

Кумулятивная кривая (кумулята) это графическое изображение вариационного ряда, составленное по последовательно суммированным, т.е. накопленным частотам (или относительным частотам).

При построении кумулятивной кривой дискретного вариационного ряда на ось абсцисс наносят значения варианты, ординатами служат нарастающие итоги частот (или относительных частот). Ломаная линия, соединяющая вершины ординат образует кумулятивную кривую.

п. 3 Эмпирическая функция распределения

Пусть известно статистическое распределение частот случайной величины X.

Эмпирической функцией распределения (или функцией распределения выборки) называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события X<x:

$$F^*(x) = \frac{n_x}{n}, \quad (6)$$

где n_x – число вариант, меньших x; n – объем выборки.

Таким образом, для того чтобы найти, например, $F^*(x_2)$, надо число вариант, меньших x_2, разделить на объем выборки: $F^*(x_2) = \frac{n_{x_2}}{n}$.

Из определения эмпирической функции следует, что $F^*(x)$ обладает всеми свойствами функции распределения F(x), а именно:

1) значения функции $F^*(x)$ принадлежат интервалу [0,1];
2) $F^*(x)$ - неубывающая функция;
3) если a – наименьшая, а b – наибольшая варианта, то $F^*(x) = 0$ при $x \le a$ и $F^*(x) = 1$ при $x > b$.

Функцию распределения $F(x)$, в отличие от эмпирической функции $F^*(x)$, называют теоретической функцией распределения. Различие между

эмпирической и теоретической функцией распределения состоит в том, что первая определяет относительную частоту события X<*x*, а вторая – вероятность того же события.

Построим эмпирическую функцию распределения. Из свойств функции $F^*(x)$ и данных таблицы получаем:

$$F^*(x)=\begin{cases} 0 \text{ при } x \le 15, \\ 0{,}2 \text{ при } 15 < x \le 16, \\ 0{,}3 \text{ при } 16 < x \le 17, \\ 0{,}4 \text{ при } 17 < x \le 18, \\ 0{,}5 \text{ при } 18 < x \le 19, \\ 0{,}6 \text{ при } 19 < x \le 20, \\ 0{,}8 \text{ при } 20 < x \le 21, \\ 0{,}9 \text{ при } 21 < x \le 26, \\ 1 \text{ при } x > 26. \end{cases}$$

График функции $F^*(x)$ изображён на рисунке 2.

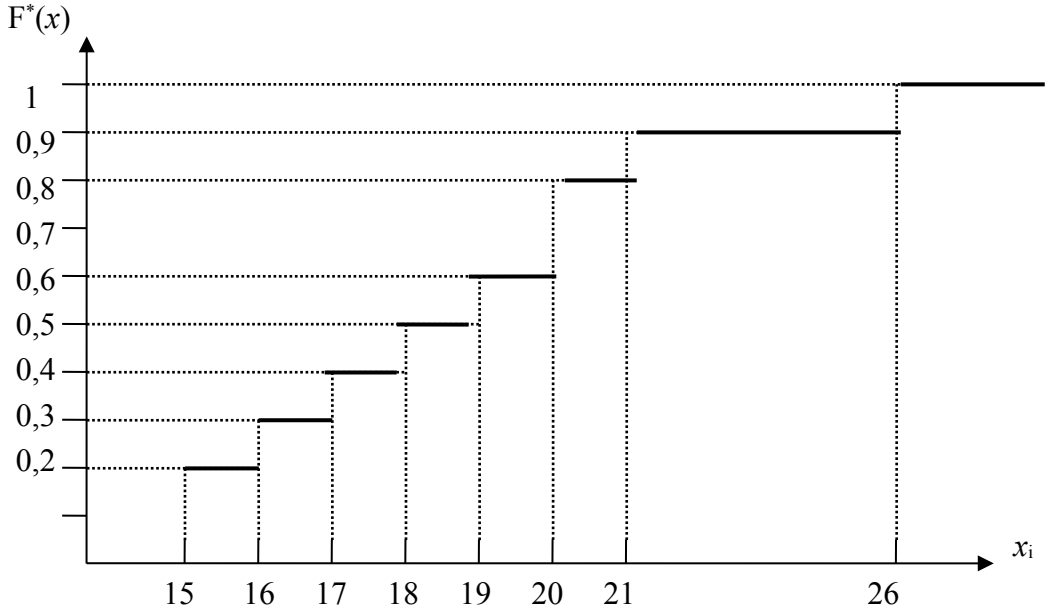

Рисунок 2 – График эмпирической функции распределения