МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**Московский государственный университет экономики,
статистики и информатики
Московский международный университет эконометрики,
информатики, финансов и права**

**Мхитарян В.С.
Трошин Л.И.
Адамова Е.В.
Шевченко
Бамбаева Н.Я.**

Теория вероятностей и математическая статистика

Москва 2001

СОДЕРЖАНИЕ

ЭЛЕМЕНТЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

1.1. Случайные события и вероятности

1.1.1. Случайные события

Одним из основных понятий теории вероятностей является случайное событие. *Случайным событием* называется событие, которое должно либо произойти, либо не произойти при выполнении некоторого комплекса условий.

В дальнейшем вместо “выполнение некоторого комплекса условий” и “случайное событие” будем употреблять выражения “произведено испытание”, “событие” и “результат испытания”.

Случайные события обычно обозначаются заглавными буквами латинского алфавита: A, B, C, ... Зафиксируем некоторое испытание, то есть комплекс условий, и будем рассматривать некоторую систему S событий A, B, C.

Укажем некоторые соотношения, которые могут существовать между событиями системы S.

1. Если в результате испытания при каждом появлении события A наступает событие B, то говорят, что A является частным случаем B, и записывают этот факт в виде

$$A \subset B.$$

2. Если $A \subset B$ и $B \supset A$, то A=B. События A и B называются равносильными, если при каждом испытании они оба наступают, либо не наступают.

3. Произведением событий A и B называется такое событие AB, которое заключается в совместном наступлении этих событий.

4. Суммой событий A и B называется такое событие A+B, которое заключается в наступлении по крайней мере одного из этих событий.

5. Событие U называется достоверным, если оно с необходимостью должно произойти при каждом испытании. Событие V называется невозможным, если оно не происходит ни при каком испытании. Все достоверные события равносильны, то же самое относится и к невозможным событиям.

6. Событие $\overline{A}$ называется противоположным событию A /и наоборот/, если для них одновременно выполняются неравенства

$$A + \overline{A} = U \; ; \; A\overline{A} = V.$$

7. События A и B называются несовместимыми, если их совместное наступление неосуществимо, т. е. если

$$AB=V.$$

8. События A_1, A_2, ... A_n образуют полную группу попарно несовместных событий, если события A_i и A_j при $i \neq j$ несовместимы и хотя бы одно из событий A_1, A_2, ... A_n непременно должно произойти. Иными словами, полная группа попарно несовместных событий A_1, A_2, ... A_n удовлетворяют двум условиям:

$A_1+A_2+...+A_n=U$ /полная группа/

$A_i \cdot A_j=V$, $i \neq j$ /попарная несовместимость/

Введенные операции над событиями удовлетворяют следующим правилам:

а) A+B=B+A; A+V=A; A+U=U; A+A=A;

б) AB=BA; AU=A; AV=V; A·A=A;

в) (A+B)C=AC+BC.

Одним из наглядных представлений случайных событий и операций над ними являются так называемые диаграммы Виена. Пусть внутри квадрата, изображенного на рис. 1.1. наудачу выбирается точка, не лежащая ни на одной из нарисованных окружностей. Обозначим через A и B соответствующий выбор точки в левом и правом кругах. Области, заштрихованные на рис. 1.1. изображают соответственно

события $A, \overline{A}, B, \overline{B}, A+B, AB$. По диаграммам Виена легко проверяются правила сложения и умножения событий.

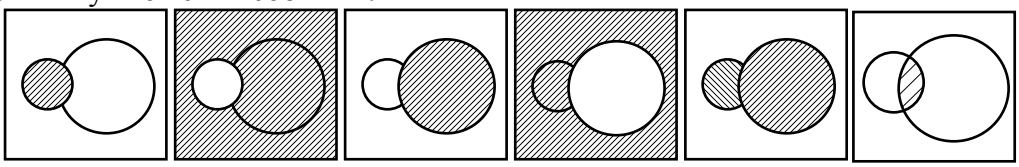


Рис. 1.1. Диаграмма Виенна

1.1.2. Классическое определение вероятности

Классическое определение вероятности исходит из некоторой системы равновозможных (равновероятных) событий, которые формально не определяются.

Рассмотрим полную группу попарно несовместных равновозможных событий E_1, E_2, ..., E_N. Добавим к этим N невозможное событие V и сложные события, образованные с помощью операции сложения любого числа и любых номеров событий E_1, E_2, ..., E_N. Полученная система событий называется полем событий S. Система S исчерпывается конечным числом событий, если считать равносильные события просто тождественно равными друг другу.

Пусть, например, полная группа попарно несовместимых равновозможных событий состоит из двух событий E_1 и E_2. Тогда система S содержит следующие четыре события: V, E_1, E_2, $E_1+E_2=U$. Если же полная группа попарно несовместимых равновозможных событий состоит из трех событий E_1, E_2, E_3, то система S содержит восемь событий: V, E_1, E_2, E_3, E_1+E_2, E_1+E_3, E_2+E_3, $E_1+E_2+E_3=U$.

Назовем для краткости событие E_i (i=1,2, ... ,N) возможным случаем. Пусть событие A является некоторым событием системы S, тогда A представляется в виде суммы некоторых возможных случаев E_i. Слагаемые E_i, входящие в разложение A, назовем случаями, благоприпятствующими событию A, а их число обозначим буквой M.

Определение. Вероятность P(A) события A равняется отношению числа возможных случаев, благоприпятствующих событию A, к числу всех возможных случав, то есть

$$P(A) = \frac{M}{N}. \qquad (1.1)$$

Из определения вероятности следует, что для вычисления P(A) требуется прежде всего выяснить, какие события в условиях данной задачи, являются возможными случаями, затем подсчитать число возможных случаев, благоприятствующих событию A, число всех возможных случаев и найти отношение числа благоприятствующих случаев к числу всех возможных.

Пример 1.1. На семи карточках написаны: а, а, о, с, т, т, ч. Какова вероятность того, что при произвольном порядке расположения этих карточек в ряд будет составлено слово “частота”?

Решение. Нумеруем данные карточки. Возможными случаями считаются любые расположения этих карточек в ряд. Следовательно, число всех возможных случаев N есть число перестановок, составленных из семи элементов, то есть N=7!=5040. Благоприятствующими возможными случаями для события A, вероятность которого требуется найти, будут те перестановки, у которых на первом

месте стоит буква “ч”, на втором - “а”, на третьем - “с”, на четвертом - “т”, на пятом - “с”, на шестом - “т” и на седьмом - “а”. На втором и седьмом местах буква “а” может появится 2! способами в зависимости от номера, присвоенного карточке с буквой “а”. Следовательно, различных перестановок, благоприятствующих появлению слова “частота” и отличающихся только номерами карточек с буквой “а”, будет 2!. То же самое можно сказать о букве “т”. Число перестановок, благоприятствующих появлению слова “частота” и отличающихся как номерами карточек с буквой “а”, так и номерами карточек с буквой “т”, будет равно 2! × 2!. Итак, число случаев, благоприятствующих событию А, равно M=2! × 2!. В результате получаем искомую вероятность:

$$P(A)=\frac{M}{N}=\frac{2!\cdot 2!}{7!}=\frac{4}{5040}=\frac{1}{1260}.$$

Пример 1.2. Известно, что среди 11 приборов имеется 3 непроверенных. Какова вероятность при случайном безвозвратном отборе 5 приборов обнаружить среди них 2 непроверенных.

Решение. Перенумеруем все 11 приборов. Возможными случаями будем считать соединения по пять приборов из 11, отличающихся только номерами приборов, входящих в каждое соединение. Отсюда следует, что число всех возможных случаев будет равно числу сочетаний из 11 элементов по 5 элементов:

$$\mathrm{N}=\mathrm{C}_{11}^{5}=\frac{11\cdot 10\cdot 9\cdot 8\cdot 7}{1\cdot 2\cdot 3\cdot 4\cdot 5}=462.$$

Для подсчета возможных благоприятствующих случаев учитываем, что 2 непроверенных из 3 непроверенных приборов можно извлечь $\mathrm{C}_3^2=3$ способами. Кроме того, 3 непроверенных прибора можно выбрать из 8 имеющихся проверенных $\mathrm{C}_8^3=\frac{8\cdot 7\cdot 6}{1\cdot 2\cdot 3}=56$ различными способами. Каждый вариант из двух непроверенных приборов комбинируется с каждым вариантом из трех проверенных, следовательно, число возможных случаев M, благоприятствующих событию А, вероятность которого требуется найти, равно $\mathrm{C}_3^2\cdot\mathrm{C}_8^3=3\cdot 56=168$. Отсюда

$$m(A)=\frac{168}{462}=\frac{4}{11}.$$

Пример 1.3. В лифт восьмиэтажного дома вошли три человека. Каждый из них с одинаковой вероятностью выходит на любой из трех этажей, начиная с третьего. Найти вероятность того, что все пассажиры лифта выйдут на разных этажах.

Решение. Возможными случаями в данном примере считаются любые мыслимые распределения, отличающиеся не только количеством, но и индивидуальностью пассажиров лифта, выходящих на том или ином этаже. Так как любой человек может выйти на каждом из шести (от третьего до восьмого) этажей, всех возможных случаев будет $\mathrm{N}=6^3=216$. Для подсчета благоприятствующих случаев предположим сначала, что пассажиры выходят по одному на фиксированных этажах. Общее число таких случаев равно 3!. Теперь обратим внимание на тот факт, что общее число сочетаний из 6 этажей по три этажа равно $\mathrm{C}_6^3=\frac{6!}{3!\cdot 3!}$. Следовательно, число благоприятствующих случаев M равно $\mathrm{C}_6^3\cdot 3!$, то есть равно числу размещений из 6 элементов по 3 – A_6^3. Итак,

$$\mathrm{P}(\mathrm{A})=\frac{\mathrm{A}_6^3}{6^3}=\frac{5}{9}$$

Рассмотрим некоторые свойства вероятностей, вытекающие из классического определения.

1. Вероятность достоверного события равна единице. Достоверное событие U обязательно происходит при испытании, поэтому все возможные случаи являются для него благоприятствующими и

$$P(U) = \frac{N}{N} = 1.$$

2. Вероятность невозможного события равна нулю. Число благоприятствующих случаев для невозможного события равна нулю (M=0), поэтому $P(V) = \frac{0}{N} = 0$.

3. Вероятность события есть число, заключенное между нулем и единицей.

В силу того, что дробь $\frac{M}{N}$ не может быть числом отрицательным и большим единицы, справедливо неравенство:

$$0 \le m(A) \le 1.$$

1.1.3. Статистическое определение вероятности

Следует отметить, что классическое определение вероятности имеет существенный недостаток, заключающийся в том, что в практических задачах не всегда можно найти разумный способ выделения “равносильных случаев”. Например, затруднительно определить вероятность того, что ребенок, который должен родиться, окажется мальчиком, или определить вероятность брака в партии деталей. Из-за указанного недостатка наряду с классическим пользуются статистическим определением вероятности, опирающимся на понятие частоты (или частости).

Если классическое определение вероятности исходит из соображений равновозможности событий при некоторых испытаниях, то статистически вероятность определяется из опыта, наблюдения результатов испытания.

Назовем число m появления события A при n испытаниях частотой, а отношение $\frac{m}{n}$ - частостью (относительной частотой) события.

Например, пусть испытание состоит в подбрасывании монеты, а событием является появление герба. Приведем результаты трех опытов, произведенных известными статистиками Бюффоном и К.Пирсоном.

Число подбрасываний	**Частоты появления герба**	**Частость**
4040	2048	0,5080
12000	6019	0,5016
24000	12012	0,5005

Как видно, относительные частоты незначительно уклоняются от вероятности 0,5, вычисленной на основе классического определения вероятности.

Тот факт, что при большем числе испытаний относительная частота событий остается почти постоянной, приводит к предположению о наличии объективных закономерностей, характеризующих это событие и не зависящих от испытателя.

Вероятностью случайного события А можно назвать некоторую постоянную, являющуюся числовой характеристикой, мерой объективной возможности этого события, около которой колеблется относительная частота.

Статистическое определение вероятности заключается в том, что за вероятность события А принимается относительная частота или число, близкое к ней. При этом требуется, чтобы в неизменных условиях было проведено достаточно большое число испытаний, независимых друг от друга, в каждом из которых может произойти или не произойти событие А.

К недостаткам статистического определения вероятности следует отнести то, что оно носит описательный, а не формально-математический характер; кроме того, такое определение не показывает реальных условий, при которых наблюдается устойчивость частот.

1.1.4. Понятие об аксиоматическом определении вероятности

Классическое и статистическое определения вероятности в совокупности до некоторой степени компенсируют друг друга и лишены недостатков, присущих им в отдельности.

Точным, строгим с математической точки зрения является аксиоматическое определение вероятности. Такое построение теории вероятностей опирается на теорию меры и интегрирования и исходит из некоторого списка не определяемых формально основных понятий и аксиом, на основе которого все дальнейшие понятия отчетливо определяются , а дальнейшие предложения доказываются.

В настоящее время в теории вероятностей принята система аксиом, сформулированная академиком А.Н. Колмогоровым.

Основным понятием аксиоматики является элементарное событие. Рассматривается множество всех элементарных событий U. Выбирается некоторая система S подмножеств этого множества. Элементы множества S определяются как случайные события или события. События подчиняются следующим аксиомам.

1. Если А и В - события, то А, АВ и А+В - тоже события.
2. Каждому событию А соответствует неотрицательное число Р(А), называемое вероятностью события А.
3. Достоверное событие U является событием с вероятностью, равной единице, то есть Р(U)=1.
4. Если события A_1, A_2, ... A_n попарно несовместимы, то $\sum_{i=1}^{n} A_i$ также является событием и вероятность его равна сумме вероятностей этих событий:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right) \qquad (1.2)$$

Из аксиом и определений выводятся другие свойства вероятностей.

1.1.5. Теоремы сложения и умножения вероятностей

На основании классического определения вероятностей можно доказать теоремы о вычислении вероятностей сложных событий.

Теорема сложения вероятностей для несовместимых событий

Если событие А является суммой несовместимых событий В и С, входящих в поле событий S, то вероятность суммы этих событий равна сумме их вероятностей:

$$P\left(A\right) = P\left(B\right) + P\left(C\right). \qquad (1.3)$$

Доказательство. Пусть событию В благоприятствует M_B, а событию С-M_C событий E_i системы S. В силу несовместимости событий В и С случай E_i,

благоприятствующий B, не может быть благоприятствующим C и наоборот. Следовательно, событию A благоприятствуют $M=M_B+M_C$ случаев из общего числа N случаев, откуда

$$P(A)=\frac{M}{N}=\frac{M_B+M_C}{N}=\frac{M_B}{N}+\frac{M_C}{N}=P(B)+P(C).$$

Следствие. Вероятность события $\overline{A}$, противоположного событию A, равна единице без вероятности события A:

$$P(\overline{A})=1-P(A). \qquad (1.4)$$

Доказательство. События A и $\overline{A}$ несовместимы и в сумме составляют достоверное событие U. Применяя теорему сложения вероятностей, получим:

$$P(U)=P(A)+P(\overline{A})$$

Так как вероятность достоверного события равна единице, получим:

$$P(\overline{A})=1-P(A).$$

Пример 1.4. Каждое из трех несовместимых событий A, B и C происходит соответственно с вероятностями 0,01; 0,02 и 0,03. Найти вероятность того, что в результате опыта не произойдет ни одного события.

Решение. Найдем вероятность того, что в результате опыта произойдет хотя бы одно из событий A, B и C, то есть найдем вероятность суммы событий Д=A+B+C. Так как по условию события A, B и C несовместимы,

$$m(Д)=m(A)+m(B)+m(C)=0{,}06.$$

Событие, вероятность которого требуется найти в задаче, является противоположным событию Д. Следовательно, искомая вероятность равна:

$$m(Д)=1-m(Д)=0{,}94.$$

Два события A и B называются зависимыми, если вероятность одного из них зависит от наступления или не наступления другого. В случае зависимых событий вводится понятие условной вероятности события.

Условной вероятностью P(A/B) события A называется вероятность события A, вычисленная при условии, что событие B произошло. Аналогично через P(B/A) обозначается условная вероятность события B при условии, что A наступило.

Безусловная вероятность события A отличается от условной вероятности этого события. Например, пусть брошены две монеты и требуется определить вероятность того, что появится два “орла” (событие A), если известно, что на первой монете появится “орел” (событие B). Все возможные случаи следующие: (орел, решка), (орел, орел), (решка, орел), (решка, решка), в скобках на первом месте указана сторона первой монеты, на втором месте - второй монеты.

Если речь идет о безусловной вероятности событий A, то N=4, M=1 и P(A)=0,25. Если же событие B произошло, то число благоприятствующих A случаев остается тем же самым M=1, а число возможных случаем N=2: (орел, орел), (орел, решка). Следовательно, условная вероятность A при условии, что B наступило, есть P(A/B)=0,5.

Теорема умножения вероятностей зависимых событий. Вероятность совместного наступления двух зависимых событий равна вероятности одного события, умноженной на условную вероятность другого события при условии, что первое произошло:

$$P(A\cdot B)=P(A)\cdot P(B/A)=P(B)\cdot P(A/B) \qquad (1.5)$$

Доказательство. Пусть событию A благоприятствуют m случаев, событию B благоприятствуют k случаев и событию AB благоприятствуют r случаев. Очевидно, r

≤ m и r ≤ k. Обозначим через N число всех возможных случаев, тогда $P(A \cdot B) = \frac{r}{N}$; $\mathrm{P(A)} = \frac{m}{N}$ или $\mathrm{P(B)} = \frac{k}{N}$. Если событие A произошло, то осуществится один из m случаев, ему благоприятствующих. При таком условии событию B благоприятствуют r и только r случаев, благоприятствующих AB. Следовательно, $P(B/A) = \frac{r}{m}$. Точно так же $\mathrm{P(A/B)} = \frac{\mathrm{r}}{\mathrm{k}}$. Подставляя соответствующие обозначения в очевидные равенства

$$\frac{\mathrm{r}}{\mathrm{N}} = \frac{\mathrm{m}}{\mathrm{N}} \cdot \frac{\mathrm{r}}{\mathrm{m}} = \frac{\mathrm{k}}{\mathrm{N}} \cdot \frac{\mathrm{r}}{\mathrm{k}},$$

получим: $P(A \cdot B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$.

Говорят, что событие A независимо от события B, если имеет место равенство P(A/B)=P(A).

Следствие 1. Вероятность совместного наступления двух независимых событий равна произведению вероятностей этих событий (теорема умножения для независимых событий):

$$m(A \cdot e) = m(A) \cdot m(e) \qquad (1.6)$$

Доказательство. Пусть A не зависит от B, тогда согласно теореме умножения вероятностей и равенству P(A/B)=P(A), получим P(AB)=P(B)・P(A) или P(AB)=P(A) × P(B), так что следствие доказано.

Кроме того, имеем равенство:

$$P(A) \cdot P(B/A) = P(A) \cdot P(B),$$

откуда P(B/A)=P(B), т.е. свойство независимости событий взаимно: если A не зависит от B, то B не зависит от A.

Следствие 2. Вероятность суммы двух событий равна сумме вероятностей этих событий без вероятности совместного их наступления (теорема сложения для любых событий), т.е. если A и B - любые события, совместимые или несовместимые, то

$$\mathrm{P(A+B) = P(A) + P(B) - P(A \cdot B)} \qquad (1.7)$$

Доказательство. Рассмотрим следующие представления событий A+B и B:

$$\mathrm{A + B = A + \overline{A} \cdot B;} \quad \mathrm{B = A \cdot B + \overline{A} \cdot B}$$

Поскольку в правых частях представлены несовместимые события, то, применяя теорему сложения вероятностей, получим:

$$P(A+B) = P(A) + P(\overline{A} \cdot B); \quad \mathrm{P(B)} = P(A \cdot B) + P(\overline{A} \cdot B),$$

откуда следует:

$$P(A+B) = P(A) + P(B) - P(AB).$$

Отметим, что если события A и B несовместимы, то совместное наступление их невозможно: AB=V и P(AB)=P(V)=0, так что

$$\mathrm{P(A+B)=P(A)+P(B).}$$

Следствие 3. Пусть производится n одинаковых независимых испытаний, при каждом из которых событие A появляется с вероятностью p. Тогда вероятность появления события A хотя бы один раз при этих испытаниях равна $1-(1-p)^n$.

Доказательство. Обозначим через A_i появление события A в i-м испытании (i=1,2,...,n). Тогда событие B, состоящее в появлении события A в n испытаниях хотя бы один раз, запишется в виде

$$B = A_1 + A_2 + \ldots + A_n = \sum_{i=1}^{n} A_i .$$

Рассмотрим событие $\overline{B}$, заключающееся в том, что при n испытаниях событие A не появится ни разу, тогда

$$\overline{B} = \overline{A_1} \cdot \overline{A_2} \cdot \ldots \cdot \overline{A_n} .$$

Так как $B + \overline{B} = U$, получим, что

$$P(B) = 1 - P(\overline{B}) = 1 - P(\overline{A_1} \cdot \overline{A_2} \cdot \ldots \cdot \overline{A_n}).$$

Так как для любых i события A_i не зависят от остальных, окончательно получим

$$m(e) = 1 - \prod_{i=1}^{n} P(\overline{A_i}) = 1 - (1-p)^n$$

Пример 1.5. Вероятность попадания стрелка в мишень при каждом выстреле равна 0,8. Найти вероятность того, что после двух выстрелов мишень окажется поврежденной.

Решение. Обозначим через A_1 событие, заключающееся в попадании в мишень при первом выстреле, а через A_2 - при втором выстреле. Тогда $A_1 \times A_2$ является событием, означающим попадание в мишень при обоих выстрелах. Событие A, вероятность которого требуется найти в задаче, является суммой события A_1 и A_2. Применяя теоремы сложения и умножения вероятностей для совместимых независимых событий A_1 и A_2 получим

$$P(A) = P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 \cdot A_2) =$$
$$= P(A_1) + P(A_2) - P(A_1) \cdot P(A_2)$$

Подставляя значение $m(A_1) = m(A_2) = 0{,}8$, будем иметь

$$m(A) = 0{,}8 + 0{,}8 - 0{,}8^2 = 0{,}96.$$

Искомую вероятность можно найти иначе: события A, заключающиеся в попадании в мишень хотя бы при одном выстреле, и $\overline{A_1} \cdot \overline{A_2}$, означающее непопадание в мишень ни при одном выстреле, являются противоположными, поэтому, применяя теорему умножения вероятностей, вычислим вероятность попадания хотя бы при одном выстреле.

Так как $m(\overline{A_1}) = m(\overline{A_2}) = 1 - 0{,}8 = 0{,}2$, искомая вероятность равна

$$P(A) = 0{,}96$$

Пример 1.6. Вероятность попадания стрелка в цель при одном выстреле равна 0,2. Сколько выстрелов должен сделать стрелок, чтобы с вероятностью не менее 0,9 попасть в цель хотя бы один раз?

Решение. Обозначим через событие A попадания стрелка в цель хотя бы один раз при n выстрелах. Так как события, состоящие в попадании в цель при первом, втором и т.д. выстрелах независимы, искомая вероятность равна

$$m(A) = 1 - p(\overline{A}_1 \cdot \overline{A}_2 \cdots \overline{A}_N) = 1 - P(\overline{A}_1) \cdot P(\overline{A}_2) \cdots P(\overline{A}_N).$$

По условию $P(A) \geq 0{,}9$ и

$$m(A_1) = m(A_2) = \ldots = m(A_N) = 0{,}2,$$

следовательно, $m(\overline{A}_1) = m(\overline{A}_2) = \ldots = m(\overline{A}_N) = 1 - 0{,}2 = 0{,}8$

и в результате получим

$$m(A) = 1 - 0{,}8^N \geq 0{,}9$$

Отсюда $0{,}8^N \le 0{,}1$. Прологарифмировав это неравенство и учитывая, что $N \cdot \lg 0{,}8 \le N \cdot \lg 0{,}1$, получим

$N \ge \dfrac{\lg 0{,}1}{\lg 0{,}8} = \dfrac{-1}{-0{,}0969} = 10{,}3$, то есть,

следовательно, стрелок должен произвести не менее 11 выстрелов.

1.1.6. Формулы полной вероятности и вероятности гипотез

Рассмотрим полную группу n попарно несовместимых событий $A_1, A_2, ..., A_n$, то есть

$$U = A_1 + A_2 + \ldots + A_n \text{ и } A_i \cdot A_j = V \text{ при } i \ne j$$

и некоторое событие B. Возьмем произведение события U на событие B:

$$U \cdot B = (A_1 + A_2 + \ldots + A_n) \cdot B$$

и, применяя свойства операций над событиями, получим

$$B = A_1B + A_2B + \ldots + A_nB.$$

События $A_i \cdot B$ и $A_j \cdot B$ при $i \ne j$ попарно несовместимы, так как $(A_j \cdot B) \cdot (A_i \cdot B) = (A_i A_j) \cdot B = VB = V$. По теореме сложения вероятностей для несовместимых событий получим

$$P(B) = \sum_{i=1}^{n} P(A_i \cdot B),$$

далее, применяя теорему умножения, окончательно будем иметь

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i) \qquad (1.8)$$

Итак, вероятность P(B) события B, которое может произойти только совместно с одним из событий $A_1, A_2, \ldots, A_n$, образующих полную группу попарно несовместимых событий, определяется последней формулой, носящий название формулы полной вероятности.

Пример 1.7. В магазин поступила новая продукция с трех предприятий. Процентный состав этой продукции следующий: 20% - продукция первого предприятия, 30% - продукция второго предприятия, 50% - продукция третьего предприятия; далее, 10% продукции первого предприятия высшего сорта, на втором предприятии - 5% и на третьем - 20% продукции высшего сорта. Найти вероятность того, что случайно купленная новая продукция окажется высшего сорта.

Решение. Обозначим через B событие, заключающееся в том, что будет куплена продукция высшего сорта, через A_1, A_2 и A_3 обозначим события, заключающиеся в покупке продукции, принадлежащей соответственно первому, второму и третьему предприятиям. Очевидно,

$$B = A_1B + A_2B + A_3B,$$

и можно применить формулу полной вероятности, причем в наших обозначениях

$$P(A_1) = 0{,}2 \quad P(B/A_1) = 0{,}1$$

$$P(A_2) = 0{,}3 \quad P(B/A_2) = 0{,}05$$

$$P(A_3) = 0{,}5 \quad P(B/A_3) = 0{,}2$$

Подставляя эти значения в формулу полной вероятности, получим искомую вероятность

$$m(e) = 0{,}2 \cdot 0{,}1 + 0{,}3 \cdot 0{,}05 + 0{,}5 \cdot 0{,}2 = 0{,}135,$$

Пусть, как и при выводе формулы полной вероятности, событие B может наступить в различных условиях, относительно существа которых можно сделать n предположений, гипотез: A_1, A_2 , A_3...A_n. Вероятности $P(A_1)$, $P(A_2)$...$P(A_n)$ этих гипотез известны до испытания, и, кроме того, известна вероятность $P(B/A_i)$, сообщаемая событию B гипотезой A_i. Пусть после проведенного испытания событие B наступило, требуется при этом условии найти вероятность гипотезы A_i.

Воспользуемся для вывода формулы искомой вероятности теоремой умножения:

$$P(A_i \cdot B) = P(B) \cdot P(A_i / B) = P(A_i) \cdot P(B / A_i),$$

откуда

$$m(A_i / B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}.$$

Подставив в знаменатель этой формулы правую часть формулы полной вероятности (1.8), окончательно будем иметь:

$$P(A_i / B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}, \quad \mathrm{i} = 1,2,...,\mathrm{n}$$

Полученные формулы носят название формул вероятности гипотез, или формул Бейеса.

Пример 1.8. В течение месяца в порт нефтеперерабатывающего завода приходят независимо друг от друга два танкера одинакового тоннажа. Технико-экономические условия для данного завода таковы, что завод может выполнить месячный заказ, если придет хотя бы один из этих танкеров в течении первых пяти суток месяца; завод не выполнит заказ, если в начале месяца не придет ни один танкер. Вероятность прихода каждого танкера в течение первых пяти суток постоянна и равна 0,1. Доставленная в начале месяца нефть обеспечивает выполнение плана с вероятностью 0,05, если придет только один танкер, и с вероятностью 0,2, если придут оба танкера. Завод выполнил план. Указать при этом условии число танкеров, прибывших в течении первых пяти суток месяца, вероятность которого наибольшая.

Решение. Обозначим через E_1 событие, заключающееся в том, что в начале месяца пришел первый танкер, а через E_2 - второй. Пусть гипотеза A_1 состоит в том, что в первые пять суток пришел только один танкер, тогда, согласно правилам операций над событиями, имеем:

$$A_1 = E_1 \cdot \overline{E_2} + \overline{E_1} \cdot E_2$$

Пусть, далее, A_2 - гипотеза, заключающаяся в приходе в начале планируемого периода обоих танкеров, тогда

$$\mathrm{A_2 = E_1 \cdot E_2},$$

и, наконец, A_3 - гипотеза, состоящая в том, что не пришел ни один танкер в начале месяца, тогда

$$A_3 = \overline{E_1} \cdot \overline{E_2}.$$

Найдем вероятности этих гипотез исходя из условий $P(E_1) = P(E_2) = 0{,}1;\ \mathrm{P}(\overline{\mathrm{E_1}}) = P(\overline{E_2}) = 0{,}9$ и применяя теоремы сложения и умножения вероятностей:

$$P(A_1) = P(\overline{E_1}) \cdot P(E_2) + P(E_1) \cdot P(\overline{E_2}) = 0{,}18;$$
$$P(A_2) = P(E_1) \cdot P(E_2) = 0{,}01;$$
$$P(A_3) = P(\overline{E_1}) \cdot P(\overline{E_2}) = 0{,}81.$$

Обозначим через B событие, заключающееся в выполнении заказа заводом, тогда

$$B = BA_1 + BA_2 + BA_3,$$

причем согласно условиям задачи

$$P(B/A_1) = 0{,}05;\ P(B/A_2) = 0{,}2 \text{ и } P(B/A_3) = 0.$$

По формуле полной вероятности получим

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i) = 0{,}18 \cdot 0{,}05 + 0{,}01 \cdot 0{,}2 + 0{,}81 \cdot 0 = 0{,}011.$$

Теперь вычислим вероятности всех гипотез при условии, что событие B имело место, применяя формулы Бейса:

$$P(A_1/B) = \frac{0{,}18 \cdot 0{,}05}{0{,}011} = \frac{9}{11};$$
$$P(A_2/B) = \frac{0{,}01 \cdot 0{,}2}{0{,}011} = \frac{2}{11};$$
$$P(A_3/B) = \frac{0{,}81 \cdot 0}{0{,}011} = 0.$$

Сравнивая полученные вероятности, заключаем: если завод выполнил заказ, то вероятнее всего за счет того, что пришел в первые пять суток планируемого периода один танкер.

1.1.7. Повторение испытаний. Формула Бернулли

При решении вероятностных задач часто приходится сталкиваться с ситуациями, в которых одно и тоже испытание (опыт) испытания повторяется многократно.

Поставим задачу общем виде. Пусть в результате испытания возможны два исхода: либо появится событие A, либо противоположное ему событие $\overline{A}$. Проведем n испытаний Бернулли. Это означает, что все n испытаний независимы; вероятность появления события A в каждом отдельно взятом или единичном испытании постоянна и от испытания к испытанию не изменяется (т.е. испытания проводятся в одинаковых условиях). Обозначим вероятность P(A) появления события A единичном испытании буквой p, т.е. P(A) = p, а вероятность $P(\overline{A})$ - буквой q, т.е. $P(\overline{A}) = 1 - P(\overline{A}) = 1\text{-}p = q$.

Найдем вероятность $P_n(m)$ наступления события A ровно m раз (ненаступления n-m раз) в этих n испытаниях. Отметим, что здесь не требуется появление m раз события A в определенной последовательности.

Обозначим: A_i - появление события A в i-м опыте; $\overline{A}_i$ - непоявление события A в i-м опыте, где i=1, 2, 3, ..., n.

Для одного испытания возможны следующие два исхода: A, $\overline{A}$. Вероятности этих исходов выпишем в виде следующей таблицы:

События	A	$\overline{A}$
Вероятность	p	q

Очевидно, $P_1(1) = p$; $P_1(0) = q$ и $P_1(1)+P_1(0)=(p+q)^1 = 1$.

Для двух испытаний возможно следующие $4 = 2^2$ исхода: A_1A_2, $\overline{A}_1A_2$, $A_1\overline{A}_2$, $\overline{A}_1\overline{A}_2$Вероятность этих исходов также запишем в виде таблицы:

События	A_1A_2	$\overline{A}_1A_2$	$A_1\overline{A}_2$	$\overline{A}_1\overline{A}_2$
Вероятность	p^2	pq	pq	q^2

Очевидно, $P_2(2) = p^2$, $P_2(1) =P(A_1\overline{A}_2)+P(\overline{A}_1A_2) = 2pq$, $P_2(0) = q^2$ и $P_2(2)+P_2(1)+P_2(0) = p^2+2pq+q^2 = (p+q)^2 = 1$.

Для трех испытаний возможны следующие $8 = 2^3$ исходов $A_1A_2A_3$, $\overline{A}_1A_2A_3$, $A_1\overline{A}_2A_3$, $A_1A_2\overline{A}_3$, $\overline{A}_1\overline{A}_2A_3$, $\overline{A}_1A_2\overline{A}_3$, $A_1\overline{A}_2\overline{A}_3$, $\overline{A}_1\overline{A}_2\overline{A}_3$. Вероятности этих исходов запишем в виде таблицы:

События	$A_1A_2A_3$	$\overline{A}_1A_2A_3$	$A_1\overline{A}_2A_3$	$A_1A_2\overline{A}_3$	$\overline{A}_1\overline{A}_2A_3$
Вероятность	p^3	p^2q	p^2q	p^2q	pq^2

События	$\overline{A}_1A_2\overline{A}_3$	$A_1\overline{A}_2\overline{A}_3$	$\overline{A}_1\overline{A}_2\overline{A}_3$
Вероятность	pq^2	pq^2	q^3

Очевидно, $P_3(3) = p^3$, $P_3(2) = P(\overline{A}_1A_2A_3)+P(A_1\overline{A}_2A_3)+P(A_1A_2\overline{A}_3)=3p^2q$, $P_3(1) = P(\overline{A}_1\overline{A}_2A_3)+P(\overline{A}_1A_2\overline{A}_3)+P(A_1\overline{A}_2\overline{A}_3) = 3pq^2$, $P_3(0) = q^3$ и $P_3(3)+P_3(2)+P_3(1)+P_3(0) = p^3+3p^2q+3pq^2+q^3 = (p+q)^3 = 1$. Анализируя эти случаи, можно сделать общий вывод: вероятность $P_n(m)$ пропорциональна произведению p^mq^{n-m}, причем коэффициент пропорциональности равен C_n^m, т.е.

$$P_n(m) = C_n^m p^m q^{n-m} = \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

Полученную формулу называют формулой Бернулли.

Пример 1.9. Монету бросают 6 раз. Какова вероятность того, что 4 раза выпадет “орел”.

Решение. Обозначим: количество испытаний $n = 6$; число поступлений события “выпадет орел” $m = 4$; вероятность поступления события “выпадет орел” $p = 0,5$; тогда $q = 1-p = 0,5$.

По формуле Бернулли получаем

$$P_6(4) = C_6^4 \cdot 0,5^4 \cdot 0,5^2 = \frac{6!}{4!\cdot 2!} 0,5^6.$$

1.1.8. Локальная и интегральная теоремы Лапласа

Пусть производится n одинаковых независимых испытаний с вероятностью появления события в каждом испытании, равной p. Тогда вероятность частоты m наступления события A определяется, как было показано ранее по формуле Бернулли:

$$P_n(m) = C_n^m p^m q^{n-m}$$

Вычисление по этой формуле трудно практически осуществить при $n>20$.

Муавром и Лапласом была получена асимптотическая формула, позволяющая найти указанную вероятность. Теорема, выражающая эту формулу, носит название локальной теоремы Муавра-Лапласа .

Если производится n одинаковых испытаний, в каждом из которых вероятность появления события равна p, то вероятность того, что данное событие появится m раз, определяется по формуле

$$P_n(m) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{1}{2}\left(\frac{m-np}{\sqrt{npq}}\right)^2}.$$

Эта теорема дает приближение биномиального закона распределения к нормальному при $n \to \infty$ и p, значительно отличающемся от нуля и единицы. Для практических расчетов удобнее представлять полученную формулу в виде

$$P_n(m) \approx \frac{1}{\sqrt{npq}} \varphi(t), \qquad (1.9)$$

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}};$$

$$t = \frac{m - np}{\sqrt{npq}}.$$

Если $m=m_0=np$, то вероятность наиболее вероятной частоты находится по формуле

$$P(m_0) \approx \frac{1}{\sqrt{2\pi npq}}.$$

Пример 1.10. Для мастера определенной квалификации вероятность изготовить деталь отличного качества равна 0,75. За смену он изготовил 400 деталей. Найти вероятность того, что в их числе 250 деталей отличного качества.

Решение. По условию n = 400, p = 0,75 , q = 0,25 и m = 280, откуда

$$t = \frac{m - np}{\sqrt{npq}} = \frac{280 - 300}{75} = -2{,}31.$$

По таблицам $\varphi(t)$ найдем $\varphi(-2.31)= \varphi(2.31) = 0{,}0277$.

Искомая вероятность равна

$$P_{400}(280) = \frac{1}{\sqrt{npq}} \varphi(t) = \frac{0{,}0277}{75} \approx 0{,}0032.$$

Для вычисления вероятности того, что частота m, подчиненная биномиальному закону распределения, заключена между данными значениями m_1 и m_2, применяют интегральную теорему Лапласа, выраженную асимптотической формулой

$$P(m_1 \le m \le m_2) \approx \frac{1}{2}\left[\Phi(t_2) - \Phi(t_1)\right],$$

где

$$t_1 = \frac{m_1 - np}{\sqrt{npq}}; \quad t_2 = \frac{m_2 - np}{\sqrt{npq}}.$$

Формулу, выражающую интегральную терему Лапласа, можно получить из закона нормального распределения (см. далее), положив

$X = m$; $x_1=m_1$; $x_2=m_2$; $\mu = np$; $\sigma = \sqrt{npq}$.

При больших значениях n наиболее вероятная частота m_0 совпадает с математическим ожиданием (см. далее) частоты. Поэтому для нахождения вероятности того, что абсолютная величина отклонения частоты от наиболее вероятной частоты не превосходит заданного числа $\varepsilon > 0$, применим формулу закона нормального распределения

$P(|X-\mu|<t\sigma) = \Phi(t)$,

где

$X = m$; $\mu = M(m) = m_0$; $\sigma = \sqrt{npq}$; $t\sigma = \varepsilon$.

Заметим, что, пользуясь теоремами Лапласа, можно находить вероятность того, что частость $\frac{m}{n}$ примет заданное значение.

1.1.9. Формула Пуассона

Если вероятность события p (или q) в отдельном испытании близка к нулю (такие события называются редкими), то даже при большом числе испытаний n, но небольшой величине произведения np (меньше 10) вероятности $P_n(m)$, полученные по формуле (1.9), недостаточно близки к их истинным значениям. В таких случаях применяют другую асимптотическую формулу - формулу Пуассона, справедливость которой доказывает следующая теорема.

Теорема. Если вероятность p наступления события A в каждом испытании постоянно близка к нулю, число независимых испытаний n достаточно велико, произведение $np = \lambda$ (см. далее), то вероятность $P_n(m)$ того, что в n независимых испытаниях событие A наступит m раз приближенно равна $\frac{\lambda^m e^{-\lambda}}{m!}$, т.е.

$$P_n(m) \approx P(m) = \frac{\lambda^m e^{-\lambda}}{m!}$$

Для вычисления вероятности $P_n(m)$ воспользуемся формулой Бернулли. Имеем

$$P_n(m) = \frac{n!}{m!(n-m)!} P^m q^{n-m} = \frac{n(n-1)(n-2)\ldots(n-m+1)}{m!} p^m q^{n-m}$$

Так как, по условию, $np=\lambda$, то $p=\lambda/n$. Тогда

$$P_n(m) = \frac{n(n-1)(n-2)\ldots(n-m+1)}{m!}\left(\frac{\lambda}{n}\right)^m \left(1-\frac{\lambda}{n}\right)^{n-m}$$

и $\lim\limits_{n\to\infty} P_n(m) = P(m)$

Условия теоремы требуют, чтобы вероятность события p была мала, а число испытаний n велико. Обычно указанную формулу используют, когда $n \geq 10$, лучше $n \geq 100$, а $np<10$.

Пример 1.11. Некоторое электронное устройство выходит из строя, если откажет определенная микросхема. Вероятность ее отказа в течение 1 ч работы устройства равна 0,004. Какова вероятность того, что за 100 ч работы устройства придется пять раз менять микросхему?

Решение. По условию, n=1000, p=0,004, а $\lambda=np=1000\cdot 0{,}004=4<10$. Для нахождения вероятности $P_{1000}(5)$ воспользуемся формулой Пуассона, так как условия ее применения выполнены. Имеем

$$P_{1000}(5) \approx \frac{4^5 e^{-4}}{5!} \approx 0{,}1563.$$

1.2. Случайные величины и их числовые характеристики

1.2.1. Случайная величина и ее распределение

Случайная величина является одним из основных понятий теории вероятностей. Рассмотрим некоторые примеры, разъясняющие смысл случайной величины.

При последовательном бросании монеты несколько раз число появлений “орла” является переменной величиной, принимающей значения 0,1,2,... в зависимости от случайных обстоятельств.

Интервал времени между двумя последовательными появлениями автобуса на данной остановке также является переменной величиной, подверженной различным колебаниям в зависимости от многих причин, учесть которые мы не в состоянии.

Рассматриваемая в этих примерах переменная величина обладает характерной особенностью. Хотя мы можем указать область ее возможных значений, однако мы не можем заранее знать, какое конкретное значение примет эта переменная величина, так как оно зависит от случая и меняется от испытания к испытанию.

Переменную величину, обладающую указанной особенностью, называют случайной величиной. Для изучения случайной величины необходимо не только указать область ее возможных значений, но и то, как часто принимается этой величиной определенное значение, то есть вероятность этих значений.

Соответствие между областью возможных значений случайной величины и множеством вероятностей этих значений носит название закона распределения случайной величины.

В зависимости от характера области возможных значений можно выделить два вида случайных величин: дискретные и непрерывные. Функцию, устанавливающую соответствие между областью возможных значений и множеством вероятностей для каждого вида случайных величин, можно задать разными способами.

Будем обозначать случайные величины большими латинскими буквами *X, Y, T, ...* , а соответствующие значения, которые они принимают, малыми буквами *x, y, t, ...*

Случайная величина называется дискретной, если она принимает конечное или счетное число значений. Дискретная случайная величина задается с помощью ряда распределения - функции, ставящей в соответствие каждому возможному значению случайной величины определенную вероятность. Таким образом, ряд распределения - это конечное или счетное множество пар элементов:

$$\{x_i, p_i\},\ i=1,2, \ldots ; \qquad p_i=P(X= x_i) .$$

Так как случайная величина X примет обязательно какое-нибудь из своих возможных значений x_i , сумма вероятностей p_i всех возможных значений равно единице, то есть $\sum_{i=1}^{n} p_i = 1$ для случайной величины, принимамающей конечное число n возможных значений, и $\sum_{i=1}^{\infty} p_i = 1$ для дискретной случайной величины, принимающей счетное число значений.

Ряд распределения удобно изображать в виде таблицы:

$$X = \begin{Bmatrix} x_1 & x_2 & \ldots & x_i & \ldots \\ p_1 & p_2 & \ldots & p_i & \ldots \end{Bmatrix}$$

в верхней строке которой указаны возможные значения x_i дискретной случайной величины X, а в нижней - соответственно вероятности того, что X примет значение x_i.

Графическое изображение ряда распределения называется многоугольником /полигоном/ распределения. Для его построения возможные значения x_i случайной величины откладываются по оси абсцисс, а вероятности - по оси ординат; точки с координатами (x_i, p_i) соединяются отрезками /рис.1.2/.

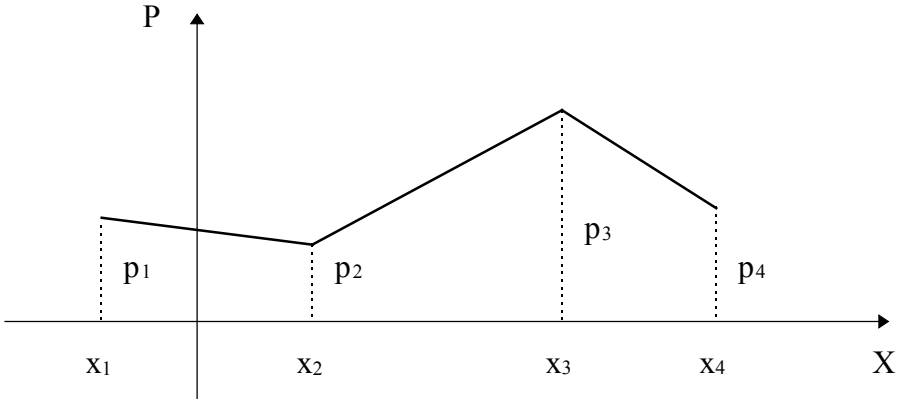


Рис. 1.2

Пример 1.12. Найти ряд распределения случайной величины, являющейся частотой выпадения “орла” при трех бросаниях монеты. Построить полигон распределения вероятностей.

Решение. Возможные значения частоты X выпадения “орла” следующие: 0, 1, 2, 3. Соответствующие вероятности нетрудно подсчитать путем учета благоприятствующих каждому значению частоты случаев при числе всех возможных случаев, равных 8:

$$P(X=0)=\frac{1}{8}; \qquad P(X=1)=\frac{3}{8}; \qquad P(X=2)=\frac{3}{8}; \qquad P(X=3)=\frac{1}{8};$$

Таким образом,

$$X=\begin{Bmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{Bmatrix}$$

Полигон распределения показан на рисунке 1.3.

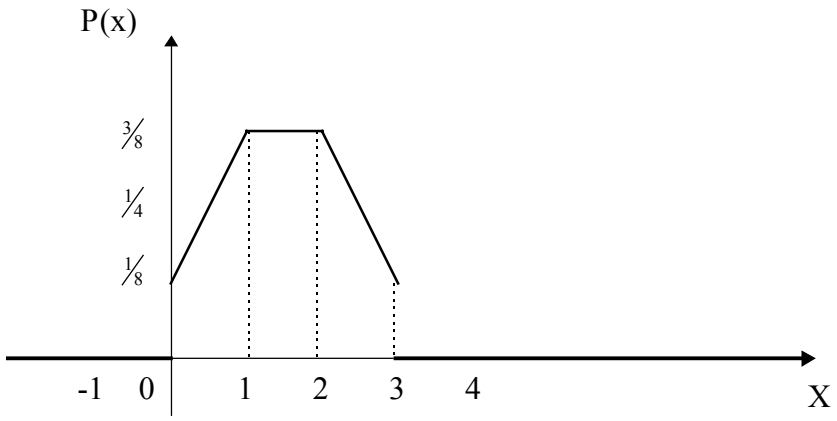


Рис. 1.3

Перейдем теперь к понятию непрерывной случайной величины.

Непрерывная случайная величина принимает возможные значения, заполняющие сплошь заданный интервал, причем для любого x из этого интервала существует предел:

$$p(x) = \lim_{\Delta x \to 0} \frac{P(x < X < x + \Delta x)}{\Delta x}$$

Функция *p(x)* /иногда обозначаемая через *f(x)* / называется плотностью распределения /дифференциальным законом распределения/. Из приведенного определения вытекают следующие свойства плотности распределения:

1. $p(x) \geq 0$;
2. При любых x_1 и x_2, входящих в заданный интервал, удовлетворяет равенству

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)dx$$

3. $\int_{-\infty}^{+\infty} p(x)dx = 1$.

Заметим, что для удобства изучения непрерывных случайных величин плотность распределения определяют не на конечном интервале возможных значений случайной величины, а на всей действительной числовой прямой, полагая, естественно, *p(x)* тождественно равной нулю для x , лежащих вне интервала возможных значений случайной величины.

График плотности распределения носит название кривой распределения (рис. 1.4).

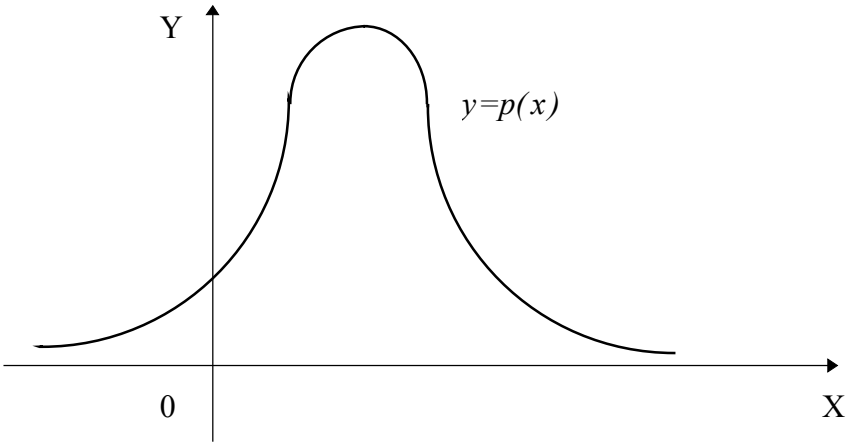


Рис. 1.4

Пример 1.13. Интервал времени между моментами прихода автобусов к остановке равновозможен в пределах от нуля до пяти минут, Найти плотность распределения вероятностей интервала времени, построить кривую распределения и определить вероятность того, что этот интервал будет находиться в пределах от одной до трех минут.

Решение. Согласно условиям задачи можно считать, что вероятность попадания интервала X в пределы от x до $x+\Delta x$ пропорциональна отношению $\frac{\Delta x}{5}$, то есть

$$P(x < X < x + \Delta x) = \frac{\Delta x}{5}$$

Отсюда

$$\mathrm{p(x)} = \lim_{\Delta \mathrm{x} \to 0} \frac{\mathrm{P(x < X < x + \Delta x)}}{\Delta \mathrm{x}} = \lim \frac{\mathrm{k} \cdot \frac{\Delta \mathrm{x}}{5}}{\Delta \mathrm{x}} = \frac{\mathrm{k}}{5}$$

Теперь найдем значение параметра k. Из свойства 3 плотности вероятностей

$$\int_0^5 \frac{k}{5} dx = 1$$

откуда k = 1.

Итак, плотность распределения для любого действительного x задается следующим образом:

$$\mathrm{p(x)} = \begin{cases} 0 & \text{при } \mathrm{x} \le 0; \\ \frac{1}{5} & \text{при } 0 < \mathrm{x} \le 5; \\ 0 & \text{при } \mathrm{x} > 5. \end{cases}$$

Кривая распределения изображена на рисунке 1.5

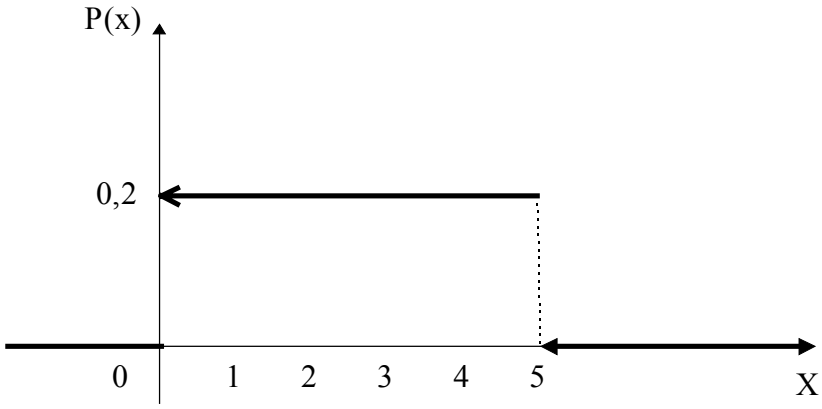


Рис. 1.5

Найдем вероятность того, что интервал времени будет заключен в пределах от одной до трех минут, заметив, что по свойству 2:

$$P(1 < X < 3) = \int_1^3 \frac{1}{5} dx = \frac{3}{5} - \frac{1}{5} = \frac{2}{5}$$

Наиболее общим способом задания различных по своей природе случайных величин является функция распределения случайной величины /интегральный закон распределения/.

Функцией распределения *F(x)* случайной величины X , принимающей любое действительное значение x , называется вероятность того, что случайная величина X примет значение, меньшее, чем x , то есть

$$F(x)=P(X<x)$$

Для дискретной случайной величины функция *F(x)* вычисляется по формуле

$$F(x)= \sum_{x_i<x} p_i$$

где суммирование ведется по всем значениям i , для которых $x_i < x$.

Для непрерывной случайной величины интегральный закон выражается формулой:

$$F(x) = \int_{-\infty}^{x} p(z)dz$$

где функция p(z) является плотностью распределения.

Функия распределения обладает следующими основными свойствами:

1. $P(x_1 \le X < x_2) = F(x_2) - F(x_1)$;

2. $F(x_2) \ge F(x_1)$, если $x_2 > x_1$;

3. $\lim\limits_{x\to-\infty} F(x)=0$;

4. $\lim\limits_{x\to+\infty} F(x)=1$;

5. $dF(x) = p(x)dx$ (для непрерывной случайной величины).

График функции распределения для непрерывной случайной величины называется интегральной кривой расрпределения и имеет, например, вид, указанный на рис. 1.6

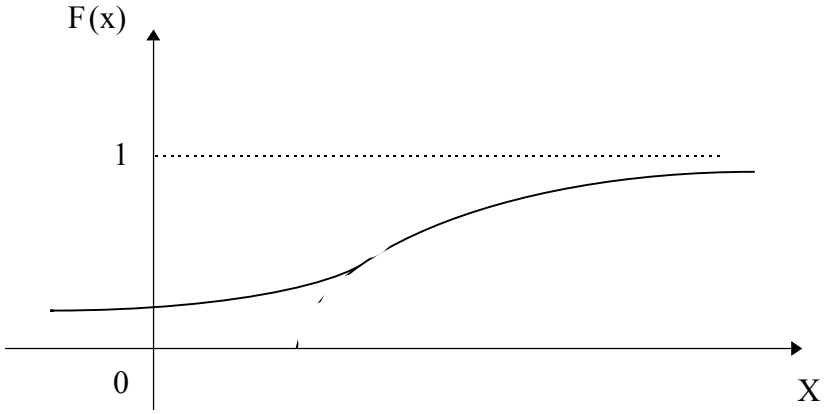


Рис. 1.6

Пример 1.14. Построить функцию и график распределения для случайной величины примера 1.9

Решение. Интегральная функция имет вид

$$F(x)=\begin{cases} 0 & при \quad x\le 0 \\ 0{,}125 & при \quad 0<x\le 1 \\ 0{,}125+0{,}375=0{,}5 & при \quad 1<x\le 2 \\ 0{,}5+0{,}375=0{,}875 & при \quad 2<x\le 3 \\ 0{,}875+0{,}125=1 & при \quad x>3 \end{cases}$$

График функции распределения для дискретной случайной величины представляет собой ступенчатую разрывную линию, непрерывную слева (рис. 1.7).

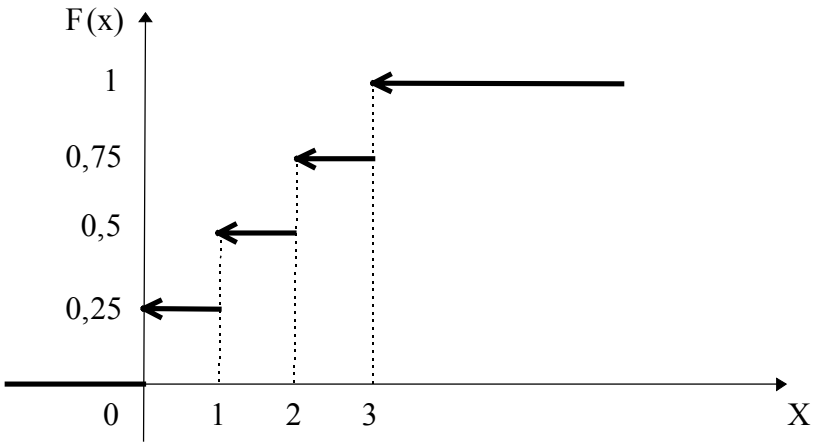


Рис. 1.7

Пример 1.15. Построить интегральный закон распределения и интегральную кривую для случайной величины примера 1.10.

Решение. Исходя из формулы

$$F(x) = \int_{-\infty}^{x} p(z)dz$$

будем иметь:

1. при $x \le 0$

$$F(x) = \int_{-\infty}^{x} p(z)dz = 0;$$

2. при $0 < x \le 5$

$$F(x) = \int_{-\infty}^{x} p(z)dz = \int_{-\infty}^{0} 0 \cdot dz + \int_{0}^{x} \frac{1}{5} dz = 0{,}2x;$$

3. при $x > 5$

$$F(x) = \int_{-\infty}^{x} p(z)dz = \int_{-\infty}^{0} 0 \cdot dz + \int_{0}^{5} \frac{1}{5} dz + \int_{5}^{x} 0 \cdot dz = 1.$$

Таким образом,

$$F(x) = \begin{cases} 0 & при \quad x \le 5 \\ 0{,}2 & при \quad 0 < x \le 5 \\ 1 & при \quad x > 5 \end{cases}$$

Интегральная кривая имеет вид (рис. 1.8)

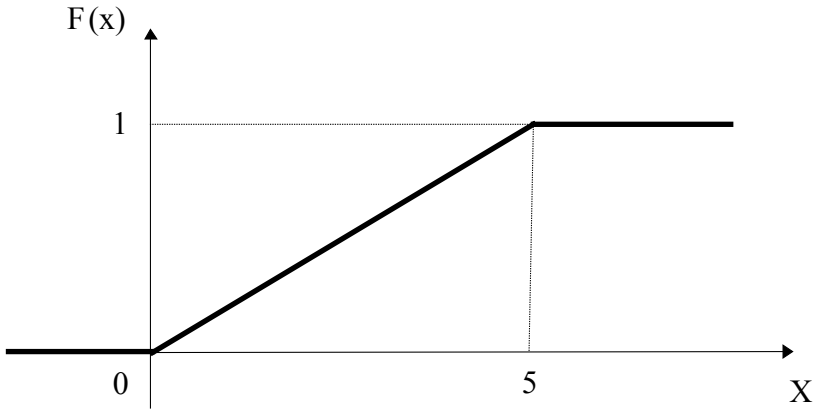


Рис. 1.8

1.2.2. Математическое ожидание и дисперсия случайной величины

Для практического применения не всегда необходимо иметь полное представление о случайной величине, достаточно знать некоторые ее числовые характеристики, дающие суммарное представление о случайной величине.

К таким характеристикам прежде всего относятся математическое ожидание и дисперсия.

Математическое ожидание /среднее значение/ *M(X)* дискретной случайной величины *X* определяется по формуле

$$M(X) = \sum_{i=1}^{\infty} x_i p_i \qquad (1.10)$$

где символ ∞ заменяется числом n , если случайная величина имеет конечное число n значений, и ряд $\sum_{i=1}^{\infty} x_i p_i$ сходится абсолютно.

Если случайная величина *X* непрерывна и *p(x)* - ее плотность распределения, то математическим ожиданием случайной величины называется интеграл

$$M(X) = \int_{-\infty}^{\infty} x p(x)\, dx \qquad (1.11)$$

в тех случаях, когда существует интеграл

$$\int_{-\infty}^{\infty} |x| \cdot p(x)\, dx .$$

Пример 1.16. Найти математическое ожидание случайных величин, рассмотренных в примерах 1.12 и 1.13.

Решение. Для числа появлений “орла” имеем следующий ряд распределения:

$$X = \begin{Bmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{Bmatrix},$$

так что среднее число появлений “орла” при трех бросаниях монеты следующее:

$$\mathrm{M(x)} = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1{,}5$$

Для интервала времени между двумя появлениями автобуса на остановке плотность распределения имеет вид

$$\mathrm{P(x)} = \begin{cases} 0 \quad \text{при} & \mathrm{x} < 5 \\ 0{,}2 \quad \text{при} & 0 < \mathrm{x} < 5 \\ 1 \quad \text{при} & \mathrm{x} > 5 \end{cases}$$

Среднее значение интервала времени получаем равным:

$$\mathrm{M(X)} = \int_{-\infty}^{\infty} \mathrm{xp(x)dx} = \int_{-\infty}^{0} \mathrm{x} \cdot 0 \cdot \mathrm{dx} + \int_{0}^{5} \mathrm{x} \cdot 0{,}2 \cdot \mathrm{dx} + \int_{5}^{\infty} \mathrm{x} \cdot 0 \cdot \mathrm{dx} = 0{,}2 \cdot \int_{0}^{5} \mathrm{x\,dx} = 2{,}5$$

Дисперсия *D(x)* случайной величины *X* характеризует средний разброс, рассеяние значений случайной величины около математического ожидания.

Дисперсией случайной величины называется математическое ожидание квадрата отклонения случайной величины от математического ожидания, то есть

$$\mathrm{D[X]=M[X-M(X)]^2}$$

Пусть имеется дискретная случайная величина *X* , заданная рядом распределения:

$$X = \begin{Bmatrix} x_1 & x_2 & \dots & x_k & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{Bmatrix}$$

Рассмотрим случайную величину *X-M(X)*, равную разности случайной величины *X* и постоянной величины *M(X)* и называемую отклонением *X* от *M(X)* . Ряд распределения для отклонения имеет следующий вид:

$$X - M(X) = \begin{Bmatrix} x_1 - M(X) & x_2 - M(X) & \dots & x_k - M(X) & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{Bmatrix}$$

так как случайная величина *X-M(X)* принимает значение x_k-*M(X)* тогда и только тогда, когда *X* принимает значение x_k , следовательно, вероятность значений x_k и x_k-*M(X)* - одна и та же и равна p_k.

Далее рассмотрим случайную величину, равную квадрату отклонения случайной величины *X* от ее математического ожидания *M(X)*. Рассуждая, как

выше, получим следующий ряд распределения для $\left[X - M(X)\right]^2$, если $\left|x_k - M(X)\right| \neq \left|x_l - M(X)\right|$ для любых $k \neq l$

$$\left[X - M(X)\right]^2 = \begin{Bmatrix} \left[x_1 - M(X)\right]^2 & \left[x_2 - M(X)\right]^2 & \dots & \left[x_k - M(X)\right]^2 & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{Bmatrix}$$

Тогда дисперсия вычисляется по формуле:

$$\mathrm{DX} = \sum_{\mathrm{k}=1}^{\infty} \left[\mathrm{x_k} - \mathrm{M(X)}\right]^2 \cdot \mathrm{p_k} \quad (1.12)$$

Заметим, что правая часть формулы для дисперсии верна и в случае, когда $\left|\mathrm{x_k} - \mathrm{M(X)}\right| = \left|\mathrm{x_1} - \mathrm{M(X)}\right|$ для некоторых $k \neq l$, хотя ряд для $\left[X - M(X)\right]^2$ будет отличаться от написанного выше. Отличие состоит в том, что $\left|\mathrm{x_k} - \mathrm{M(X)}\right| = \left|\mathrm{x_1} - \mathrm{M(X)}\right|$ соответствует одно значение $\left[x_k - M(X)\right]^2 = \left[x_l - M(X)\right]^2$ с вероятностью p_k+p_l , так как, если $\left[X - M(X)\right]^2$ примет это значение, то *X-M(X)* примет значение либо x_k*-M(X)* либо x_l*-M(X).*

Для непрерывной случайной величины дисперсия определяется по формуле

$$D(X) = M\left[X - M(X)\right]^2 = \int_{-\infty}^{\infty} \left[X - M(X)\right]^2 p(x)\, dx \qquad (1.13)$$

Пример 1.17. Найти дисперсию случайных величин, приведенных в примерах 1.12 и 1.13.

Решение. Напишем ряд распределения для квадрата отклонений от числа выпадений орла от среднего значения, равного 1,5:

$$\left[X - M(X)\right]^2 = \begin{Bmatrix} (0-1{,}5)^2 & (1-1{,}5)^2 & (2-1{,}5)^2 & (3-1{,}5)^2 \\ 1/8 & 3/8 & 3/8 & 1/8 \end{Bmatrix} = \begin{Bmatrix} 0{,}25 & 2{,}25 \\ 0{,}75 & 0{,}25 \end{Bmatrix}$$

Затем вычислим дисперсию:

$$\mathrm{D(X)} = 0{,}25 \cdot 0{,}75 + 2{,}25 \cdot 0{,}25 = 0{,}75$$

Дисперсия для интервала времени между двумя появлениями автобуса найдем по формуле для дисперсии непрерывной случайной величины, имея *M(X)=2,5* :

$$\mathrm{D(X)} = \int_{-\infty}^{\infty} \left(\mathrm{x} - 2{,}5\right)^2 \mathrm{p(x)}\, \mathrm{dx} = \int_{0}^{5} \left(\mathrm{x} - 2{,}5\right)^2 \cdot \frac{1}{5}\, \mathrm{dx} = \frac{1}{5}\left(\frac{\mathrm{x}^3}{3} - \frac{5\mathrm{x}^2}{2} + \frac{25}{4}\mathrm{x} \right)\Bigg|_0^5 = \frac{25}{12}$$

Как не трудно заметить, если случайная величина выражена в некоторых единицах измерения, то дисперсия имеет наименование, выраженное в квадратных единицах. Для удобства представления случайной величины через свои

характеристики вводят понятие среднего квадратического отклонения $\sigma(x)$, равного арифметическому корню из дисперсии:

$$\sigma(x) = \sqrt{M\left[X - M(X)\right]^2}. \quad (1.14)$$

1.2.3. Основные свойства математического ожидания и дисперсии

Доказательства рассматриваемых свойств будем проводить для дискретных случайных величин.

Свойство 1. Математическое ожидание постоянной равно этой постоянной.

Доказательство. Постоянную C можно рассматривать как дискретную случайную величину, принимающую единственное значение c с вероятностью единица, поэтому $\mathrm{M(C) = c \cdot 1 = c}$.

Свойство 2. Математическое ожидание суммы случайных величин равно сумме их математических ожиданий:

$$M(X+Y) = M(X) + M(Y). \quad (1.15)$$

Доказательство. Пусть случайные величины X и Y имеют соответственно следующие ряды распределения:

$$X = \begin{Bmatrix} x_1 & x_2 & \dots & x_k & \dots & x_n \\ p_1 & p_2 & \dots & p_k & \dots & p_n \end{Bmatrix}, \qquad Y = \begin{Bmatrix} y_1 & y_2 & \dots & y_l & \dots & y_m \\ p_1 & p_2 & \dots & p_l & \dots & p_m \end{Bmatrix}$$

Напишем ряд распределения для суммы $X+Y$.

Возможные значения случайной величины $X+Y$ есть следующие:

$$\mathrm{x_1 + y_1,\ \ x_2 + y_2,\ \ \dots,\ \ x_1 + y_l,\ \dots,\ x_1 + y_m,\ \ x_2 + y_1,\ x_2 + y_2,\ \dots,\ x_2 + y_l,}$$
$$\mathrm{\dots,\ \ x_2 + y_m,\ \ \dots,\ \ x_k + y_1,\ \ \dots,\ \ x_n + y_m}\ .$$

Более компактная запись возможных значений выглядит так:

$$x_k + y_l \quad , \qquad \left(k = 1 \div n;\ \ l = 1 \div m\right).$$

Обозначим вероятность того, что X примет значение x_k, а Y - значение y_l через p_{kl}, тогда:

$$X + Y = \begin{Bmatrix} x_k + y_l \\ p_{kl} \end{Bmatrix}, \qquad \left(k = 1 \div n;\ \ l = 1 \div m\right).$$

Рассмотрим событие $X+Y=x_k+Y$ и найдем вероятность этого события. Это событие происходит тогда и только тогда, когда Y принимает одно из значений $y_1, y_2, \dots, y_l, \dots, y_m$, причем события

x_k+y_1, x_k+y_2, ..., x_k+y_m попарно несовместимы. Следовательно, можно применить формулу вероятности суммы:

$$P(X+Y=x_k+Y)=\sum_{l=1}^{m} p_{kl}\ .$$

С другой стороны, $P(X+Y=x_k+Y)=P(X=x_k)$ и $P(X=x_k)=p_k$, следовательно

$$\sum_{l=1}^{m} p_{kl}=p_k\ .$$

Аналогично доказывается формула

$$\sum_{k=1}^{n} p_{kl}=p_l\ .$$

По определению математического ожидания

$$\mathrm{M(X+Y)}=\sum_{\mathrm{k,l}}(\mathrm{x_k+y_1})\cdot\mathrm{p_{kl}}=\sum_{\mathrm{k=1}}^{\mathrm{n}}\sum_{\mathrm{l=1}}^{\mathrm{m}}(\mathrm{x_k+y_1})\cdot\mathrm{p_{kl}}=$$

$$\sum_{\mathrm{k=1}}^{\mathrm{n}}\mathrm{x_k}\left(\sum_{\mathrm{l=1}}^{\mathrm{m}}\mathrm{p_{kl}}\right)+\sum_{\mathrm{l=1}}^{\mathrm{m}}\mathrm{y_l}\left(\sum_{\mathrm{k=1}}^{\mathrm{n}}\mathrm{p_{kl}}\right)=\sum_{\mathrm{k=1}}^{\mathrm{n}}\mathrm{x_k p_k}+\sum_{\mathrm{l=1}}^{\mathrm{m}}\mathrm{y_l p_l}=\mathrm{M(X)+M(Y)}\ .$$

Следствие. Математическое ожидание суммы конечного числа случайных величин равно сумме их математических ожиданий.

$$M(X_1+X_2+\ldots+X_n)=M(X_1)+M(X_2)+\ldots+M(X_n). \qquad (1.16)$$

Доказательство. Применяя свойство 2 и метод математической индукции, получим

$$\mathrm{M(X_1+X_2+\ldots+X_n)=M(X_1)+M(X_2+\ldots+X_n)=M(X_1)+}$$
$$\mathrm{+M(X_2)+M(X_3+..+X_n)=M(X_1)+M(X_2)+\ldots+M(X_n)}.$$

Свойство 3. Математическое ожидание произведения независимых случайных величин X и Y равно произведению математических ожиданий этих величин: $\mathrm{MXY=MX\cdot MY}$. Пусть случайные величины X и Y заданы рядами распределения. Ряд распределения для произведения случайных величин выглядит следующим образом:

$$XY=\begin{Bmatrix} x_k y_l \\ p_{kl} \end{Bmatrix},\qquad (k=1\div n;\quad l=1\div m).$$

Причем в силу независимости случайных величин X и Y события $(X=x_k)$ и $(Y=y_l)$ независимы, следовательно, по теореме умножения вероятностей независимых событий получим $\mathrm{p_{kl}=p_k\cdot p_l}$.

По определению математического ожидания

$$M(XY) = \sum_{k,l} x_k y_l p_{kl} = \sum_{k=1}^{n}\sum_{l=1}^{m} x_k y_l p_k p_l =$$

$$\left(\sum_{k=1}^{n} x_k p_k\right)\cdot\left(\sum_{l=1}^{m} y_l p_l\right) = M(X)\cdot M(Y)$$

Следствие. Постоянный множитель можно выносить за знак математического ожидания:

$$M(cX)=cM(X)\ .$$

Доказательство. Постоянную c можно рассматривать как случайную величину, причем c и X - независимые случайные величины, поэтому

$$M(cX) = M(c)\cdot M(X) = cM(X)\ . \qquad (1.17)$$

Свойство 4. Дисперсия постоянной величины равна нулю.

Доказательство. Согласно свойству 1

$$D(c) = M\left[c - M(c)\right]^2 = M(c-c)^2 = M(0) = 0\ .$$

Свойство 5. Постоянную величину можно вынести за знак дисперсии, предварительно возведя ее в квадрат, т.е.

$$D(cX)=c^2D(X)\ . \qquad (1.18)$$

Доказательство. В силу следствия из свойства 3 имеем:

$$D(cX) = M\left[cX - M(cX)\right]^2 = M\left[cX - cM(X)\right]^2 = M\left\{c^2\left[c - M(X)\right]^2\right\} =$$

$$c^2 M\left[X - M(X)\right]^2 = c^2 D(X).$$

Свойство 6. Дисперсия суммы независимых случайных величин X и Y равна сумме их дисперсии:

$$D(X+Y)=D(X)+D(Y)\ . \qquad (1.19)$$

Доказательство. По определению дисперсии и по свойству 2 получим:

$$D(X+Y) = M\left[(X+Y) - M(X+Y)\right]^2 = M\left\{\left[X - M(X)\right] + \left[Y - M(Y)\right]\right\}^2 =$$
$$D(X) + D(Y) + 2M\left\{\left[X - M(X)\right]\cdot\left[Y - M(Y)\right]\right\}.$$

Величины X и Y независимы, поэтому величины X-$M(X)$ и Y-$M(Y)$ также независимы, следовательно:

$$\mathrm{M}\left\{\left[\mathrm{X}-\mathrm{M}(\mathrm{X})\right]\cdot\left[\mathrm{Y}-\mathrm{M}(\mathrm{Y})\right]\right\}=\mathrm{M}[\mathrm{X}-\mathrm{M}(\mathrm{X})]\cdot\mathrm{M}[\mathrm{Y}-\mathrm{M}(\mathrm{Y})]=$$
$$\left[\mathrm{M}(\mathrm{X})-\mathrm{M}(\mathrm{X})\right]\cdot[\mathrm{M}(\mathrm{Y})-\mathrm{M}(\mathrm{Y})]=0\cdot 0=0\,.$$

Следствие. Если x_1 , x_2 , ... , x_n - случайные величины, каждая из которых независима от суммы остальных, то

$$D(X_1+X_2+\ldots+X_n)=D(X_1)+D(X_2)+\ldots+D(X_n)\,. \qquad (1.20)$$

Пусть дана случайная величина X , имеющая математическое ожидание $M(X)$ и среднее квадратическое отклонение $\sigma(x)\neq 0$, тогда случайная величина

$$T=\frac{X-M(X)}{\sigma(x)}$$

называется стандартизованной (нормированной). Такая случайная величина обладает тем свойством, что ее математическое ожидание равно нулю, а дисперсия равна единице.

Действительно,

$$M(T)=M\left[\frac{X-M(X)}{\sigma(x)}\right]=\frac{M[X-M(X)]}{\sigma(x)}=0\,;$$

$$D(T)=D\left[\frac{X-M(X)}{\sigma(x)}\right]=\frac{D(X)+D(-M(X))}{\sigma^2(x)}=\frac{D(X)}{D(X)}=1\,.$$

1.2.4. Моменты случайной величины

Для характеристики случайной величины, кроме математичекого ожидания и дисперсии, применяюся и моменты.

Моментом k - порядка называется математическое ожидание k - й степени отклонения случайной величины X от некоторой постоянной c .

Если в качестве с берется нуль, моменты называют начальными, то есть

$$\nu_k=M(X)^k\,. \qquad (1.21)$$

Если $c=M(X)$, то моменты называются центральными, то есть

$$\mu_k=M\,[X-M\,(X)]^k\,. \qquad (1.22)$$

В формулах, определеяющих начальные и центральные моменты, нижние индексы указывают порядок момента.

С помощью свойств математического ожидания легко показать, что

$$\nu_0=\mu_0=1\,; \qquad \mu_1=0\,;$$
$$\nu_1=M(X)\,; \qquad \mu_2=D(X)\,.$$

Выведем формулу, связывающую центральные моменты с начальными:

$$\mu_k = M[X - M(X)]^k = \sum_{m=0}^{k} C_k^m \left[-M(X)\right]^{k-m} \cdot M[X^m] = \sum_{m=0}^{k} C_k^m \left[-M(X)\right]^m \cdot \nu_m .$$

Из $\nu_1 = M(X)$ следует:

$$\mu_k = \sum_{m=2}^{k} (-1)^{k-m} C_m^k \, \nu_m \, \nu_1^{k-m} + (-1)^{k-1}(k-1)\nu_1^k . \qquad (1.23)$$

В частности, для первых четырех моментов выведенная формула дает следующие равенства:

$$\left.\begin{array}{c} \mu_0 = 1 \\ \mu_1 = 0 \\ \mu_2 = \nu_2 - \nu_1^2 \\ \mu_3 = \nu_3 - 3\nu_1 \nu_2 + 2\nu_1^3 \\ \mu_4 = \nu_4 - 4\nu_1 \nu_3 + 6\nu_1^2 \nu_2 - 3\nu_1^4 \end{array}\right\} . \qquad (1.24)$$

Первые моменты играют важную роль в статистике при нахождении параметров функции распределения.

Формула

$$D(X) = \mu_2 = \nu_2 - \nu_1^2 \qquad (1.25)$$

употребляется для вычисления дисперсии.

Пример 1.18. Вычислить начальный и центральный моменты третьего порядка для случайных величин, рассмотренных в примерах 1.12 и 1.13.

Решение. Для вычисления моментов дискретной случайной величины, числа появлений “орла” (пример 1.12), удобно воспользоваться схемой, указанной в таблице.

x_i	p_i	$x_i p_i$	$x_i^2 p_i$	$x_i^3 p_i$
0	0,125	0	0	0
1	0,375	0,375	0,375	0,375
2	0,375	0,750	1,500	3,000
3	0,125	0,375	1,125	3,375
Итого	1	1,500	3,000	6,750

Теперь воспользуемся формулой

$$\mu_3 = \nu_3 - 3\nu_1 \nu_2 + 2\nu_1^3$$

и получим

$$\mu_3 = 6{,}75 - 3 \cdot 1{,}5 \cdot 3 + 2 \cdot (1{,}5)^2 = 0 .$$

Для вычисления центрального момента третьего порядка непрерывной случайной величины - интервала времени между двумя появлениями автобуса (пример 4.2) удобнее пользоваться формулой, непосредственно определяющей центральные моменты:

$$\mu_3 = \int_{-\infty}^{\infty} [X - M(X)]^3 p(x)\, dx .$$

Так как M(X)=2,5 , то

$$\mu_3 = \int_0^5 (x - 2{,}5)^3 \cdot 0{,}2\, dx .$$

Подстановкой $\frac{x - 2{,}5}{5} = z$ мы приводим этот интеграл к виду

$$\mu = 5^2 \int_{-0{,}5}^{0{,}5} z^3 dz .$$

Так как под интегралом стоит нечетная функция, а пределы интегрирования равны по абсолютной величине и противоположны по знаку, интеграл равен нулю, следовательно, $\mu_3 = 5^2 \cdot 0 = 0$.

В заключение заметим, что если кривая распределения *p(x)* непрерывной случайной величины *X* симметрично расположена относительно оси, проходящей через *M(X)*, то все центральные моменты нечетного порядка равны нулю. То же самое заключение можно сделать по поводу дискретной случайной величины *X*, если ее полигон симметричен относительно оси, проходящей через среднее значение случайной величины.

1.2.5. Биномиальный закон распределения

Биноминальное распределение представляет собой распределение вероятностей возможных чисел появления события A при n независимых испытаний, в каждом из которых событие A может осуществиться с одной и той же вероятностью P(A) = p = const. Кроме события A может произойти также противоположное событие $\overline{A}$, вероятность которого P($\overline{A}$) = 1-p = q.

Вероятности любого числа событий соответствуют членам разложения бинома Ньютона в степени, равной числу испытаний:

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{1 \cdot 2} p^{n-1}q^2 + \ldots + c_n^m p^m q^{n-m} + \ldots + npq^{n-1} + q^n$$

где p^n - вероятность того, что при n испытаниях событие A наступит n раз;

q^n - вероятность того, что при n испытаниях событие A не наступит ни разу;

$c_n^m p^m q^{n-m}$ - вероятность того, что при n испытаниях событие A наступит m раз, а событие $\overline{A}$ наступит n-m раз;

c_n^m - число сочетаний (комбинаций) появления события A и $\overline{A}$.

Таким образом, вероятность осуществления события A m раз при n независимых испытаниях с одинаковой вероятностью p можно рассчитать по формуле общего члена разложения бинома Ньютона:

$$P_{m,n} = \frac{n!}{m!(n-m)!} p^m q^{n-m}.$$

Эта формула называется формулой Бернулли. Ее целесообразно использовать при небольшом числе испытаний, порядка $n \le 8$.

Сумма вероятностей всех комбинаций равна единице,

$$\sum_{m=0}^{n} Pm, n = 1$$

как сумма вероятностей единственно возможных и несовместных событий (комбинаций), составляющих полную группу событий.

Ряд распределения вероятностей случайной величины X = m записывают следующим образом:

$$X = \begin{Bmatrix} m, \\ c_n^m p^m (1-p)^{n-m} \end{Bmatrix},$$

где n - число независимых испытаний;
m=0÷n - частота появления события A в n независимых испытаниях.

Числовые характеристики биноминального распределения:
M(m)=np - математическое ожидание частоты появлений события A при n независимых испытаниях;
D(m)=npq - дисперсия частоты появления события A;
$\sigma(m) = \sqrt{npq}$ - среднее квадратическое отклонение частоты.

Когда число испытаний n велико, то для вычисления вероятности комбинаций используется локальная теорема Лапласа:

$$P_{m,n} = \frac{1}{\sqrt{npq}} \varphi(t),$$

где $\varphi(t) = \frac{1}{\sqrt{npq}} e^{-t^2/2}$ - нормированная плотность нормального распределения;

$t = \frac{m-np}{\sqrt{npq}}$ - нормированное значение частоты.

1.2.6. Нормальный закон распределения

Нормальное распределение - наиболее часто встречающийся вид распределения. С ним приходится сталкиваться при анализе погрешностей измерений, контроле технологических процессов и режимов, а также при анализе и прогнозировании различных явлений в в экономике, социологии, демографии и других областях знаний.

Наиболее важным условием возникновения нормального распределения является формирование признака как суммы большого числа взаимно независимых слагаемых, ни одно из которых не характеризуется исключительно большой по сравнению с другими дисперсией. В производственных условиях такие предпосылки в основном соблюдаются.

Главная особенность нормального распределения состоит в том, что оно является предельным, к которому приближаются другие распределения.

Нормальным называется распределение, функция плотности вероятности которого имеет вид

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

где μ - математическое ожидание случайной величины;
σ^2 - дисперсия случайной величины, характеристика рассеяния значений случайной величины около математического ожидания.

1.3. Закон больших чисел

1.3.1. Принцип практической невозможности маловероятных событий. Формулировка закона больших чисел

Ранее было отмечено, что нельзя предвидеть, какое из возможных значений примет случайная величина, так как мы не можем учесть все обстоятельства, от которых зависит это событие. Однако в некоторых случаях можно указать вероятность такого события.

Опыт подсказывает нам, что события, вероятность наступления которых мала, редко происходят, а события, имеющие вероятность, близкую к единице, почти обязательно происходят.

Принцип, заключающийся в том, что маловероятные события на практике рассматриваются как невозможные, носит название “принципа практической невозможности маловероятных событий”. События, происходящие с вероятностями, весьма близкими к единице, считаются практически достоверными (принцип практической достоверности). Сколь мала или сколь велика должна быть вероятность события, зависит от практического применения, от важности этого события.

Следовательно одной из основных задач теории вероятностей является установление закономерностей, происходящих с вероятностями близкими к единице. Эти закономерности должны учитывать совместное влияние большого числа независимо (или слабо зависимо) действующих факторов. При этом каждый фактор в отдельности характеризуется незначительным воздействием. Всякое предложение, устанавливающее отмеченные выше закономерности, называется законом больших чисел. Законом больших чисел, по определению проф. А.Я.Хиничина, следует назвать общий принцип, в силу которого совокупное действие большого числа факторов приводит при некоторых весьма общих условиях к результату, почти не зависящему от случая.

Некоторые конкретные условия, при которых выполняется закон больших чисел, указаны в теоремах Чебышева, Бернули, Пуасона и Ляпунова.

1.3.2. Лемма Маркова. Неравенство и теорема Чебышева. Теоремы Бернулли и Пуассона

Лемма Маркова

Пусть X - случайная величина, принимающая лишь неотрицательные значения. Тогда можно получить следующее неравенство:

$$\mathrm{P}\left(\mathrm{X} \geq \tau\right) \leq \frac{\mathrm{M}(\mathrm{X})}{\tau} \quad (\tau>0 \text{ любое})$$

Доказательство. Для определенности предположим, что X - непрерывная случайная величина с плотностью p(x). По определению математического ожидания получаем

$$M(X) = \int_0^{\infty} xp(x)dx$$

Далее будем иметь

$$\mathrm{M}\left(\mathrm{X}\right) = \int_0^{\tau} \mathrm{xp}\left(\mathrm{x}\right)\mathrm{dx} + \int_{\tau}^{\infty} \mathrm{xp(x)dx}$$

Оба слагаемых в правой части не отрицательны, поэтому

$$\mathrm{M}\left(\mathrm{X}\right) \geq \int_{\tau}^{\infty} \mathrm{xp(x)dx},$$

но теперь $\mathrm{x} \geq \tau$, и следовательно,

$$\int_{\tau}^{\infty} \mathrm{xp}\left(\mathrm{x}\right)\mathrm{dx} \geq \int_{\tau}^{\infty} \tau\mathrm{p}\left(\mathrm{x}\right)\mathrm{dx} = \tau\int_{\tau}^{\infty} \mathrm{p}\left(\mathrm{x}\right)\mathrm{dx} = \tau\mathrm{P}\left(\mathrm{X} \geq \tau\right)$$

Таким образом,

$$\mathrm{M}(\mathrm{X}) \geq \tau\mathrm{P}(\mathrm{X} \geq \tau)$$

Так как $\tau>0$, получим

$$\mathrm{P}(\mathrm{X} \geq \tau) \leq \frac{\mathrm{M}(\mathrm{X})}{\tau}$$

Рассмотрим теперь случайную величину X, имеющую математическое ожидание M(X) и дисперсию D(X). Оценим вероятность события, заключающегося в том, что отклонение X-M(X) не превысит по абсолютной величине положительного числа ε. Оценка указанной вероятности получается с помощью неравенства Чебышева.

Неравенство Чебышева.

Вероятность того, что отклонение случайной величины X от ее математического ожидания по обсолютной величине меньше положительного числа ε, не меньше, чем $1-\frac{D(X)}{\varepsilon^2}$, то есть

$$P\left(\left|X - M(X)\right| < \varepsilon\right) \geq 1 - \frac{D(X)}{\varepsilon^2} \quad (1.26)$$

Доказательство. Приведем доказательство для дискретной (конечной) случайной величины X:

$$X = \begin{Bmatrix} x_1\ \mathrm{x}_2 \ldots x_k\ \mathrm{x}_{\mathrm{k+1}} \ldots x_n \\ p_1\ \mathrm{p}_2 \ldots p_k\ \mathrm{p}_{\mathrm{k+1}} \ldots p_n \end{Bmatrix}$$

Рассмотрим случайную величину $\left|X - M(X)\right|$. Тогда ее ряд распределения имеет вид

$$\left|\mathrm{X}-\mathrm{M}(\mathrm{X})\right| = \begin{Bmatrix} \left|\mathrm{X}_1-\mathrm{M}(\mathrm{X})\right| & \left|\mathrm{X}_2-\mathrm{M}(\mathrm{X})\right| \ldots \left|\mathrm{X}_k-\mathrm{M}(\mathrm{X})\right| & \left|\mathrm{X}_{k+1}-\mathrm{M}(\mathrm{X})\right| \ldots \left|\mathrm{X}_n-\mathrm{M}(\mathrm{X})\right| \\ \mathrm{p}_1 & \mathrm{p}_2 \quad\quad \mathrm{p}_k & \mathrm{p}_{k+1} \quad\quad \mathrm{p}_n \end{Bmatrix}$$

Не ограничивая общность рассуждения, можно предположить, что первые к значений случайной величины $|X-M(X)|$ меньше заданного ε, а остальные значения не меньше ε. Тогда на основании теоремы сложения вероятностей получим следующую формулу:

$$P\left(|X-M(X)|<\varepsilon\right)=1-\sum_{i=k+1}^{n}P_i$$

Чтобы найти $\sum_{i=k+1}^{n}P_i$, запишем формулу D(X) в виде

$$D(X)=\sum_{i=1}^{k}\left[x_i-M(x)\right]^2 p_i+\sum_{i=k+1}^{n}\left[x_i-M(X)\right]^2 p_i$$

Опуская в правой части этого равенства первую сумму и заменяя во второй сумме $\left[x_i-M(X)\right]^2$ меньшей величиной ε, получим неравенство

$$D(X)\geq\sum_{i=k+1}^{n}\varepsilon^2 p_i$$

Из этого неравенства следует:

$$\sum_{i=k+1}^{n}p_i\leq\frac{D(X)}{\varepsilon^2}$$

Подставляя правую часть (1.28) в (1.26), окончательно получим

$$P\left(|X-M(X)|<\varepsilon\right)\geq 1-\frac{D(X)}{\varepsilon^2},$$

что и требовалось доказать.

Рассмотрим достаточно большое число n независимых случайных величин $X_1, X_2, \ldots X_n$. Если дисперсии их ограничены числом с, то событие, заключающееся в том, что отклонение среднего арифметического этих случайных величин от среднего арифметического их математических ожиданий будет по абсолютной величине сколь угодно малым, является почти достоверным. Это предложение, относящиеся к закону больших чисел, доказал П.Л. Чебышев.

Теорема Чебышева. Если $X_1, X_2, \ldots X_n$ попарно независемые случайные величины, причем дисперсии их не привышают постоянноо числа с, то как бы мало ни было положительное число ε, вероятность неравенства

$$\left|\frac{X_1+X_2+\ldots+X_n}{n}-\frac{M\left(X_1\right)+M\left(X_2\right)+\ldots+M\left(X_n\right)}{n}\right|<\varepsilon$$

будет как угодно близка к единице, если число n случайных величин достаточно велико.

Используя понятие предела, можно в условиях теоремы записать:

$$\lim_{n\to\infty}P\left(\left|\frac{\sum_{i=1}^{n}x_i}{n}-\frac{\sum_{i=1}^{n}M\left(X_i\right)}{n}\right|<\varepsilon\right)=1.$$

Вместо последней записи часто кратко говорят, что суммы $\frac{1}{n}\sum_{i=1}^{n}\left[X_i - M\left(X_i\right)\right]$ сходятся по вероятности к нулю.

Доказательство. Рассмотрим случайную величину $X = \frac{\sum_{i=1}^{n} x_i}{n}$. На основании свойств математического ожидания и дисперсии можно записать:

$$M(X) = \frac{1}{n}\sum_{i=1}^{n} M\left(X_i\right)$$

$$D(X) = \frac{1}{n^2}\sum_{i=1}^{n} D(X)$$

По условию теоремы $D(X_i) \le c$, поэтому $D(X) \le \frac{nc}{n^2} = \frac{c}{n}$.

Теперь можно воспользоваться неравенством Чебышева:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M\left(X_i\right)\right| < \varepsilon\right) \ge 1 - \frac{D(X)}{\varepsilon^2} \ge 1 - \frac{c}{n\varepsilon^2}$$

Переходя к пределу при $n \to \infty$, будем иметь:

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M\left(X_i\right)\right| < \varepsilon\right] \ge 1$$

Так как вероятность не может быть больше единицы, этот предел равен единице, что и требовалось доказать.

Из теоремы Чебышева следует утверждение, заключающиеся в том, что среднее арифметическое достаточно большого числа независимых случайных величин, имеющих ограниченные дисперсии, утрачивает случайный характер и становится детерминированной величиной.

Пример 1.19. Дисперсия каждой из 6250 независимых случайных величин не превосходит 9. Оценить вероятность того, что абсолютная величина отклонения средней арифметической этих случайных величин от средней арифметической их математических ожиданий не превысит 0,6.

Решение. Согласно теореме Чебышева искомая вероятность P не меньше $1 - \frac{c}{n\varepsilon^2}$. По условиям задачи c=9, n=6250, ε=0,6, следовательно, P≥0,996.

Отметим некоторые важные частные случаи теоремы Чебышева.

Теорема Бернулли. Пусть производится n независимых испытаний, в каждом из которых вероятность появления события постоянна и равна p. Тогда каково бы ни было ε>0,

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1,$$

где $\frac{m}{n}$ - частость появления события A.

Доказательство. Для доказательства рассмотрим случайную величину $X_i=m_i$, являющуюся числом наступления события A в I испытании, так что $m=m_1+m_2+\ldots+m_i+\ldots+m_n$, и случайные величины m_i попарно независимы. Ранее было

показано, что $M(m_i)=p$ и $D(m_i)=pq$. Так как $pq \le \frac{1}{4}$, то дисперсии случайных величин m_i ограничены одним и тем же числом $c=\frac{1}{4}$, следовательно, получаем все условия, при которых справедлива теорема Чебышева и окончательно получим

$$P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right) \ge 1-\frac{pq}{n\varepsilon^2},$$

откуда

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$$

Пример 1.20. На предприятии, выпускающем кинескопы, 0,8 всей продукции выдерживает гарантийный срок службы. С вероятностью, превышающей 0,95, найти пределы, в которых находится доля кинескопов, выдерживающих гарантийный срок, из партии 8000 кинескопов.

Решение. Применяем теорему Бернулли при n=8000, P≥0,95, p=0,8 и q=0,2. Подставляя в равенство

$$1-\frac{pq}{n\varepsilon^2}=0{,}95$$

p,q и n, найдем ε=0,02. Из неравенства получим $0{,}78<\frac{m}{n}<0{,}82$

Теорема Пуассона. Если в последовательности независимых испытаний появление события A в К-ом испытании равна $p_к$, то

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-\frac{\sum_{k=1}^{n} p_k}{n}\right|<\varepsilon\right)=1,$$

где m есть случайная величина, равная числу появлений событя A в первых n испытаниях.

Доказательство. Пусть случайная величина $X_к=m_k$ означает число появления события A в к-м испытании. Тогда $m=\sum_{k=1}^{n} m_k$, $D(m_k)=p_k q_k \le \frac{1}{4}$, и случайные величины m_k попарно независимы. Таким образом, теорема Пуассона является частным случаем теоремы Чебышева. На основании свойств математического ожидания и дисперсии случайной величины $X=\frac{1}{n}\sum_{k=1}^{n} X_k$ получим следующие формулы:

$$M(X)=\frac{1}{n}\sum_{k=1}^{n} M(m_k)=\frac{1}{n}\sum_{k=1}^{n} p_k=\overline{p};$$

$$D(X)=\frac{1}{n}\sum D(m_k)=\frac{1}{n^2}\sum_{k=1}^{n} p_k q_k=\frac{1}{n}\frac{\sum_{k=1}^{n} p_k q_k}{n}=\frac{\overline{pq}}{n}$$

Подставляя эти формулы в неравенство Чебышева, получаем неравенство, выражающее теорему Пуассона:

$$P\left(\left|\frac{m}{n}-\bar{p}\right|<\varepsilon\right)\geq 1-\frac{\overline{pq}}{n\varepsilon^2}$$

Пример 1.21. Произведено 900 независимых испытаний, причем в 450 из этих испытаний вероятность появления события А равна 2/3, в 200 - 0,5, в 160 - 0,3 и в 90 - 0,4. Найти оценку вероятности того, что частость появления события А отклоняется по абсолютной величине от средней вероятности не больше, чем на 0,1.

Решение. Применяем теорему Пуассона. Находим $\bar{p}$ и $\overline{pq}$:

$$\bar{p}=\frac{\sum_{k=1}^{n}p_k}{n}=\frac{2/3\cdot 450+0{,}5\cdot 200+0{,}3\cdot 160+0{,}4\cdot 90}{900}=\frac{121}{225}$$

$$\overline{pq}=\frac{\sum_{k=1}^{n}p_k q_k}{n}=\frac{2/3\cdot 1/3\cdot 450+0{,}5\cdot 0{,}5\cdot 200+0{,}3\cdot 0{,}7\cdot 160+0{,}4\cdot 0{,}6\cdot 90}{900}=0{,}228$$

Подставляя в правую часть неравенства

$$P\left(\left|\frac{m}{n}-\bar{p}\right|<\varepsilon\right)\geq 1-\frac{\overline{pq}}{n\varepsilon^2}$$

значения $\bar{p}$, $\overline{pq}$, ε и n, получим P≥0,97.

Теорема Бернулли является частным случаем теоремы Пуассона.

В самом деле, если вероятность появления данного события в каждом испытании постоянна: $p_1=p_2=\ldots p_n=p$, то $\bar{p}=p$ и $\overline{pq}=pq$

Замечание. В тех случаях, когда вероятность появления события в каждом испытании не известна, за верхнюю границу дисперсии принимают c=1/4, т.е.

$$P\left(\left|\frac{m}{n}-\bar{p}\right|<\varepsilon\right)\geq 1-\frac{1}{4n\varepsilon^2}$$

Теорема Лапласа.

Теоремы Чебышева, Бернулли, Пуассона устанавливают нижнюю границу вероятности, что часто бывает недостаточно. В некоторых случаях важно знать достаточно точное значение вероятности. Этому требованию отвечают так называемые предельные теоремы закона больших чисел, указывающие асимптотические формулы для вероятностей неравенства $\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}M\left(X_i\right)\right|<\varepsilon$ относительно n случайных величин X_i.

Мы уже знаем, что вероятность неравенства $\left|\frac{m}{n}-p\right|<\varepsilon$ вычисляется по интегральной теорема Лапласа, а именно

$$P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=\Phi(t)$$

где $\varepsilon=t\sqrt{\frac{pq}{n}}$.

Следовательно, достаточно точным выражением теоремы Бернулли является интегральная теорема Лапласа. Асимптотическую формулу для теоремы Чебышева доказал его ученик А.М. Ляпунов. Приведем теорему Ляпунова без доказательства.

1.3.3. Центральная предельная теорема

Теорема Ляпунова. Рассмотрим n независимых случайных величин X_1, $X_2,\ldots,X_n$, удовлетворяющих условиям:

1) все величины имеют определенные математические ожидания и конечные дисперсии;
2) ни одна из величин не выделяется резко от остальных по своим значениям.

Тогда при неограниченном возрастании n распределение случайной величины $\frac{1}{n}\sum_{i=1}^{n} X_i$ приближается к нормальному закону.

Таким образом, имеем следующую асимптотическую формулу:

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < t\sqrt{\frac{\overline{D(X)}}{n}}\right] = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{u^2}{2}}du = \Phi(t),$$

где $\overline{D(X)} = \frac{1}{n}\sum_{i=1}^{n} D(X_i)$

Пример 1.22. Дисперсия каждой из 400 независимых случайных величин равна 25. Найти вероятность того, что абсолютная величина отклонения средней арифметической случайных величин от средней арифметической их математических ожиданий не превысит 0,5.

Решение. Применим теорему Ляпунова. По условию задачи n=400, $D(X_i)$=25, следовательно, $\overline{D(X)}$=25 и ε=0,5. Подставляя эти данные в формулу $\varepsilon = t\sqrt{\frac{\overline{D(X)}}{n}}$, получим t=2 откуда P=Φ(2)=0,9545.

Тест

1. Каково максимальное значение вероятности произведения противоположных событий?
   а) 0,5
   б) 0,25
   в) 1,0
   г) 0,64

2. Чему равна вероятность достоверного события?
   а) 0,5
   б) 0
   в) 1,0
   г) 0,25

3. Монета подбрасывается 2 раза. Какова вероятность выпадения “орла” один раз.
   а) 1,0
   б) 0
   в) 0,025
   г) 0,5

4. Монета была подброшена 10 раз. “Герб” выпал 4 раза. Какова частость (относительная частота) выпадания “герба”?
   а) 0

б) 0,4
в) 0,5
г) 0,6

5. Консультационный пункт института получает пакеты с контрольными работами студентов из городов А, В и С. Вероятность получения пакета из города А равна 0,7, из города В - 0,2. Какова вероятность того, что очередной пакет будет получен из города С?
   а) 0,14
   б) 0,1
   в) 0,86
   г) 0,9

6. Какова вероятность выигрыша хотя бы одной партии у равносильного противника в матче, состоящем из трех результативных партий?
   а) 0,875
   б) 1
   в) 0,375
   г) 0,333

7. Если вероятность наступления события в каждом испытании постоянна, но мала, а число испытаний велико, то для нахождения вероятности того, что событие А произойдет m раз в n испытаниях следует использовать:
   а) формулу Бернулли;
   б) локальную теорему Муавара-Лапласа;
   в) формулу Пуассона;
   г) теорему умножения вероятностей.
8. Чему равно математическое ожидание случайной величины У=2Х+1, если математическое ожидание случайной величины Х равно 5?
   а) 10
   б) 6
   в) 21
   г)11

9. Чему равна дисперсия случайной величины У=2Х+1, если дисперсия случайной величины Х равна 2?
   а) 4
   б) 5
   в) 8
   г) 9

10. Какое из положений закона больших чисел оценивает вероятность отклонения случайной величины х от ее математического ожидания?
    а) Неравенство Чебышева
    б) Теорема Бернулли
    в) Теорема Чебышева
    г) Лемма Маркова

2. СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.1. Понятие о статистической оценке параметров

Методы математической статистики используются при анализе явлений, обладающих свойством **статистической устойчивости**. Это свойство заключается в том, что, хотя результат X отдельного опыта не может быть предсказан с достаточной точностью, значение некоторой функции $\theta_n^* = \theta_n^*(x_1, x_2, \ldots, x_n)$ от результатов наблюдений при неограниченном увеличении объема выборки теряет свойство случайности и сходится по вероятности к некоторой неслучайной величине θ.

Рассмотрим некоторые понятия.

Генеральной совокупностью X называют множество результатов всех мыслимых наблюдений, которые могут быть сделаны при данном комплексе условий.

В некоторых задачах генеральную совокупность рассматривают как случайную величину X.

Выборочной совокупностью (выборкой) называют множество результатов, случайно отобранных из генеральной совокупности.

Выборка должна быть репрезентативной, т.е. правильно отражать пропорции генеральной совокупности. Это достигается случайностью отбора, когда все объекты генеральной совокупности имеют одинаковую вероятность быть отобранными.

Задачи математической статистики практически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам случайной выборки.

Параметры генеральной совокупности есть постоянные величины, а выборочные характеристики (статистики) - случайные величины.

В самом общем смысле статистическое оценивание параметров распределения можно рассматривать как совокупность методов, позволяющих делать научно обоснованные выводы о числовых параметрах генеральной совокупности по случайной выборке из нее.

Сформулируем задачу статистической оценки параметров в общем виде. Пусть X - случайная величина, подчиненная закону распределения $F(x,\theta)$, где θ - параметр распределения, числовое значение которого неизвестно. Исследовать все элементы генеральной совокупности для вычисления параметра θ не представляется возможным, поэтому о данном параметре пытаются судить по выборкам из генеральной совокупности.

Всякую однозначно определенную функцию результатов наблюдений, с помощью которой судят о значении параметра θ, называют оценкой (или статистикой) параметра θ.

Рассмотрим некоторое множество выборок объемом n каждая. Оценку параметра θ, вычисленную по i-ой выборке, обозначим через $\tilde{\theta}_i$. Так как состав выборки случаен, то можно сказать, что $\tilde{\theta}_i$ примет неизвестное заранее числовое значение, т.е. является случайной величиной. Известно, что случайная величина определяется соответствующим законом распределения и числовыми характеристиками, следовательно, и выборочную оценку также можно описывать законом распределения и числовыми характеристиками.

Основная задача теории оценивания состоит в том, чтобы произвести выбор оценки $\tilde{\theta}_n$ параметра θ, позволяющей получить хорошее приближение оцениваемого параметра.

2.2. Законы распределения выборочных характеристик, используемые при оценке параметров

2.2.1. Распределение средней арифметической

Пусть из генеральной совокупности X, имеющей нормальный закон распределения $N(\mu;\sigma)$ с математическим ожиданием μ и средним квадратическим отклонением σ, взята случайная выборка объемом n и определена средняя арифметическая

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad (2.1)$$

где x_i - результат i-го наблюдения.

Здесь и в дальнейшем будем рассматривать выборку объема n, т.е. последовательность наблюдений $X_1, X_2, \ldots X_n$, как систему независимых, одинаково распределенных случайных величин с распределением $N(\mu;\sigma)$.

Таким образом, если случайная величина X распределена нормально, то средняя арифметическая распределена также нормально с параметрами $\mu;\ \frac{\sigma}{\sqrt{n}}$, т.е. $\overline{X} \in N(\mu, \frac{\sigma}{\sqrt{n}})$

Откуда следует, что

$$M\overline{x} = \mu;\ \ D\overline{x} = \frac{\sigma^2}{n} \quad \text{и} \quad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}. \qquad (2.2)$$

Для одинаково распределенных и взаимно независимых случайных величин дисперсия распределения средней арифметической в n раз меньше дисперсии случайной величины X.

2.2.2. Распределение Пирсона (χ^2 - хи квадрат)

Если $X_1, X_2, \ldots, X_n$ есть ряд независимых, нормированных, нормально распределенных случайных величин N(0,1), т.е. $MX_i=0$ и $Dx_i=1$ для i=1, 2, ..., ν, то случайная величина

$$U^2 = \sum_{i=1}^{\nu} X_i^2 \qquad (2.3)$$

имеет распределение χ^2 с ν степенями свободы, где ν - единственный параметр распределения χ^2, характеризующий число независимых случайных величин в выражении (2.3).

В таблицах приложения для различных ν приводятся числа, вероятность превышения которых случайной величиной U^2 равна заданному значению уровня значимости $\alpha=1-\gamma$.

Отметим, что математическое ожидание случайной величины U^2 равно числу степеней свободы ν, а дисперсия - удвоенному числу степеней свободы

$$MU^2=\nu; \qquad DU^2=2\nu \qquad (2.4)$$

Распределение Пирсона используется для построения доверительного интервала для генеральной дисперсии σ^2.

2.2.3. Распределение Стьюдента (t - распределение)

В 2.2.1. был рассмотрен закон распределения средней арифметической $\overline{X}$, зависящей от среднего квадратичного отклонения σ генеральной совокупности.

Однако во многих практических приложениях математической статистики параметр σ, как правило, не известен. В этой связи возникает задача определения закона распределения $\overline{X}$, не зависящего от σ, которую решил английский статистик Госсет, публиковавшийся под псевдонимом Стьюдент. Распределение Стьюдента находит широкое применение в теории статистического оценивания параметров генеральной совокупности и в проверке статистических гипотез.

Дадим определение случайной величины, имеющей распределение Стьюдента.

Если случайная величина Z имеет нормированное распределение N(0;1), а величина U^2 имеет распределение χ^2 с ν степенями свободы, причем Z и U взаимно независимы, то случайная величина

$$T = \frac{Z}{U}\sqrt{\nu}$$

имеет t распределение с ν степенями свободы.

Наибольшее применение на практике находят таблицы, в которых даны значения $t(\alpha,\nu)$, соответствующие заданному числу степеней свободы ν=1, 2, ... , и уровню значимости α, т.е. вероятности выполнения неравенства $P[|T|>t(\alpha,\nu)]=\alpha$.

Если из генеральной совокупности X с нормальным законом распределения $N(\mu;\sigma)$ взята случайная выборка объемом n, то статистика

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n-1} \qquad (2.5)$$

имеет распределение Стьюдента с ν=n-1 степенями свободы.

Распределение Стьюдента (t - распределение) используется при интервальной оценке математического ожидания при неизвестном значении среднего квадратического отклонения σ.

Теория статистического оценивания рассматривает два основных вида оценок параметров распределений: точечные и интервальные оценки.

2.3. Точечные оценки параметров распределений

Точечной оценкой называют некоторую функцию результатов наблюдения $\theta_n(x_1, x_2, \dots, x_n)$, значение которой принимается за наиболее приближенное в данных условиях к значению параметра θ генеральной совокупности.

Примером точечных оценок являются $\overline{X}$, S^2, S и др., т.е. оценки параметров одним числом.

Из точечных оценок в приложениях математической статистики часто используют начальные

$$\tilde{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \qquad (2.6)$$

где $\tilde{\nu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

и центральные $\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$, (2.7)

моменты до четвертого порядка включительно, т.е. k=1,2,3,4.

2.3.1. Основные свойства точечной оценки

Основная проблема точечной оценки заключается в выборе возможно лучшей оценки, отвечающей требованиям несмещенности, эффективности и состоятельности.

Точечную оценку $\tilde{\theta}_n$ называют **несмещенной**, если ее математическое ожидание равно оцениваемому параметру:

$$M\tilde{\theta}_n = \theta. \qquad (2.8)$$

Выполнение требования несмещенности оценки гарантирует отсутствие ошибок в оценке параметра одного знака.

Эффективной называют несмещенную выборочную оценку, обладающую наименьшей дисперсией среди всех возможных несмещенных оценок параметра θ для данного объема выборки n и функции распределения вероятности F(X,θ) генеральной совокупности.

Точечная оценка $\tilde{\theta}_n$ параметра θ называется **состоятельной**, если при $n \to \infty$ оценка $\tilde{\theta}_n$ сходится по вероятности к оцениваемому параметру, т.е. выполняется условие

$$\lim_{n\to\infty} P\left\{\left|\tilde{\theta}_n - \theta\right| < \varepsilon\right\} = 1 \qquad \text{для любого } \varepsilon>0. \qquad (2.9)$$

Следует отметить, что при состоятельности оценки оправдывается увеличение объема наблюдений, так как при этом становится маловероятным допущение значительных ошибок при оценивании.

2.3.2. Точечные оценки основных параметров распределений

Наиболее важными числовыми характеристиками случайной величины являются математическое ожидание и дисперсия.

Рассмотрим вопрос о том, какими выборочными характеристиками лучше всего в смысле несмещенности, эффективности и состоятельности оцениваются математическое ожидание и дисперсия.

1. Средняя арифметическая $\overline{X}$, вычисленная по n независимым наблюдениям над случайной величиной X, которая имеет математическое ожидание $M(x)=\mu$ и дисперсию $D(x)=\sigma^2$, является несмещенной и состоятельной оценкой этого параметра.

2. Если случайная величина X распределена нормально с параметрами $N(\mu,\sigma)$, то несмещенная оценка $\overline{X}$ математического ожидания MX имеет минимальную дисперсию, равную $D\overline{X} = \frac{\sigma^2}{n}$, поэтому средняя арифметическая $\overline{X}$ в этом случае является также эффективной оценкой математического ожидания.

3. Если случайная подборка состоит из n независимых наблюдений над случайной величиной X с математическим ожиданием MX и дисперсией $DX=\sigma^2$, то выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ не является несмещенной оценкой генеральной дисперсии σ^2.

Несмещенной оценкой дисперсии генеральной совокупности σ^2 является исправленная выборочная дисперсия

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad (2.10)$$

где дробь $\frac{n}{n-1}$ называется поправкой Бесселя. При малых значениях n поправка Бесселя довольно значительно отличается от единицы, с увеличением значений n она стремиться к единице. При n>50 практически нет разницы между оценками S^2 и $\hat{S}^2$. Оценки S^2 и $\hat{S}^2$ являются состоятельными оценками генеральной дисперсии σ^2.

4. Если известно значение математического ожидания μ, то несмещенной, состоятельной и эффективной оценкой генеральной дисперсии является выборочная оценка

$$S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad (2.11)$$

5. Если случайная величина X имеет биноминальное распределение, то несмещенной и состоятельной оценкой генеральной доли P является частость события (статистическая доля ω).

2.4. Интервальные оценки параметров распределений

При выборке небольшого объема точечная оценка $\tilde{\theta}_n$ может существенно отличаться от истинного значения параметра, т.е. приводить к грубым ошибкам. Поэтому в случае малой выборки часто используют интервальные оценки.

Интервальной оценкой называют числовой интервал $(\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)})$, определяемый по результатам выборки, относительно которого можно утверждать с определенной, близкой к единице вероятностью, что он заключает в себе значение оцениваемого параметра генеральной совокупности, т.е.

$$P(\tilde{\theta}_n^{(1)} \le \theta \le \tilde{\theta}_n^{(2)}) = \gamma, \qquad (2.12)$$

где $\tilde{\theta}_n^{(1)}$ и $\tilde{\theta}_n^{(2)}$ называют также нижней и верхней границами доверительного интервала параметра θ.

Вероятность $\gamma=1-\alpha$ принято называть доверительной вероятностью. Выбор значения доверительной вероятности следует производить исходя из конкретной задачи.

Чтобы получить представление о точности и надежности оценки $\tilde{\theta}_n$ параметра θ, можно для каждой близкой к единице вероятности γ указать такое значение Δ, что

$$P\left(\left|\tilde{\theta}_n - \theta\right| < \Delta\right) = P(\tilde{\theta}_n - \Delta \le \theta \le \tilde{\theta}_n + \Delta) = \gamma. \qquad (2.13)$$

Оценка $\tilde{\theta}_n$ будет тем точнее, чем меньше для заданной доверительной вероятности γ будет Δ. Из соотношения (2.13) следует, что вероятность того, что доверительный интервал $(\tilde{\theta}_n - \Delta; \tilde{\theta}_n + \Delta)$ со случайными границами накроет неизвестный параметр θ, равна γ. Величину Δ, равную половине ширины h доверительного интервала называют точностью оценки. В общем случае границы интервала $\tilde{\theta}_n - \Delta$ и $\tilde{\theta}_n + \Delta$ есть некоторые функции от результатов наблюдений X_1, X_2, ..., X_n. Вследствие случайного характера выборки при многократном ее повторении будут изменяться как положение, так и величина доверительного интервала

Рассмотрим теперь правила построения доверительных интервалов для некоторых параметров распределений.

2.4.1. Интервальные оценки для генеральной средней

Правила построения доверительного интервала для математического ожидания зависит от того, известна или не известна дисперсия генеральной совокупности σ^2.

Пусть из генеральной совокупности X с нормальным законом распределения $N(\mu;\ \sigma)$ и известным генеральным средним квадратическим отклонением σ взята случайная выборка $X_1,\ X_2,\ ...,\ X_n$ объемом n и вычислено $\overline{X}$. Требуется найти интервальную оценку для μ. Используем среднюю арифметическую $\overline{X}$, которая имеет нормальное распределение с параметрами $N(\mu;\sigma/\sqrt{n})$.

Тогда статистика $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}$ имеет нормированное нормальное распределение с параметрами N(0;1). Вероятность любого отклонения $\left|\overline{X}-\mu\right|$ может быть вычислена по интегральной теореме Лапласа для интервала, симметричного относительно μ, по формуле

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right|<t\right)=\Phi(t) \qquad (2.14)$$

Задавая определенную доверительную вероятность γ по таблице интегральной функции вероятностей $\Phi(t)$, можно определить значение t_γ. Для оценки математического ожидания преобразуем формулу (2.14)

$$P\left(\left|\overline{X}-\mu\right|<t_\gamma\frac{\sigma}{\sqrt{n}}\right)=\Phi(t) \qquad (2.15)$$

и далее будем иметь

$$P\left(\overline{X}-t_\gamma\frac{\sigma}{\sqrt{n}}\le\mu\le\overline{X}+t_\gamma\frac{\sigma}{\sqrt{n}}\right)=\gamma. \qquad (2.16)$$

Интервал, определенный по (2.16), и представляет собой доверительный интервал для математического ожидания μ.

Точность оценки равна

$$\Delta=t_\gamma\frac{\sigma}{\sqrt{n}}. \qquad (2.17)$$

Формула (2.17) в практических приложениях занимает особое место. По этой формуле можно, например, вычислить объем случайной выборки n, необходимый для оценки нормальной средней с заданной надежностью γ и точностью Δ, а также при заданной точности Δ и известном объеме выборки n можно определить надежность (вероятность) γ.

Нижняя и верхняя границы доверительного интервала равны

$$\mu_{min}=\overline{X}-\Delta\ ;\ \ \mu_{max}=\overline{X}+\Delta. \qquad (2.18)$$

Ширина доверительного интервала равна

$$h=\mu_{max}-\mu_{min}=2\Delta. \qquad (2.19)$$

Предположим теперь, что генеральная совокупность X распределена по нормальному закону $N(\mu;\sigma)$ с неизвестным средним квадратическим отклонением σ.

В этом случае для построения интервальной оценки генеральной средней μ используется статистика $T=\dfrac{\overline{X}-\mu}{S}\sqrt{n-1}$, имеющая распределение Стьюдента с числом степеней свободы $\nu = n-1$.

Предполагается, что средняя арифметическая $\overline{x}$ и выборочное среднее квадратическое отклонение S определены по результатам выборки объемом n из генеральной совокупности X.

По таблице t - распределения (Стьюдента) для $\nu = n-1$ степеней свободы находим значение $t_{\alpha,\nu}$, для которого справедливо равенство

$$P\left(\bar{x} - t_\alpha \frac{S}{\sqrt{n-1}} \le \mu \le \bar{x} + t_\alpha \frac{S}{\sqrt{n-1}}\right) = \gamma, \qquad (2.20)$$

где точность оценки генеральной средней равна

$$\Delta = t_\alpha \frac{S}{\sqrt{n-1}}. \qquad (2.21)$$

При достаточно больших n различия между доверительными интервалами, определенными по формулам (2.16) и (2.20), мало, так как при $n \to \infty$ распределение Стьюдента стремится к нормальному распределению.

Пример 2.1. По результатам n = 10 наблюдений установлено, что средний темп роста акций предприятий отрасли равен $\bar{X}$ = 104,4%. В предположении, что ошибки наблюдений распределены по нормальному закону со средним квадратическим отклонением σ = 1%, определить надежность γ = 0,95 интервальную оценку для генеральной средней μ.

Решение. Поскольку параметр σ нам известен, интервальную оценку будем искать согласно (2.16).

По таблице интегральной функции Лапласа Ф(t) из условия γ = 0,95 найдем t_γ = 1,96.

Тогда точность оценки равна

$$\Delta = t_\gamma \frac{\sigma}{\sqrt{n}} = 1{,}96 \cdot \frac{1}{\sqrt{10}} = 1{,}96 \cdot \frac{1}{3{,}16} = 0{,}62.$$

Отсюда доверительный интервал имеет вид

$$104{,}4 - 0{,}62 \le \mu \le 104{,}4 + 0{,}62$$

и окончательно

$$103{,}78 \le \mu \le 105{,}02 \ (\%).$$

Пример 2.2. Средняя урожайность пшеницы на 17 опытных участках области составила $\bar{X}$ = 25 ц/га, а S = 2 ц/га. Найти с надежностью 0,9 границы доверительного интервала для оценки генеральной средней.

Решение. Так как σ нам неизвестно, то интервальную оценку генеральной средней μ будем искать согласно (2.20).

Из таблиц t-распределения для числа степеней свободы ν=n-1= =16 и $\alpha = 1-\gamma$ = 1-0,9 = 0,1 найдем t_α =1,746.

Тогда точность оценки согласно (2.20) равна

$$\Delta = t_\alpha \frac{S}{\sqrt{n-1}} = 1{,}746 \cdot \frac{2}{\sqrt{16}} = 0{,}873.$$

Отсюда доверительный интервал равен

$$25 - 0{,}873 \le \mu \le 25 + 0{,}873$$

и окончательно

$$24{,}127 \le \mu \le 25{,}873 \ (\text{ц/га}).$$

2.4.2. Интервальные оценки для генеральной дисперсии и среднего квадратического отклонения

Пусть из генеральной совокупности X, распределенной по нормальному закону $N(\mu;\sigma)$, взята случайная выборка объемом n и вычислена выборочная

дисперсия S^2. Требуется определить с надежностью γ интервальные оценки для генеральной дисперсии σ^2 и среднего квадратического отклонения σ.

Построение доверительного интервала для генеральной дисперсии основывается на том, что случайная величина $\frac{nS_*^2}{\sigma^2}$ имеет распределение Пирсона (χ^2) с $\nu = n$ степенями свободы, а величина $\frac{nS^2}{\sigma^2}$ имеет распределение Пирсона с $\nu = n-1$ степенями свободы.

Подробно рассмотрим построение доверительного интервала для второго случая, так как он наиболее часто встречается на практике.

Для выбранной доверительной вероятности $\gamma = 1-\alpha$, учитывая, что $\frac{nS^2}{\sigma^2}$ имеет распределение χ^2 с $\nu = n-1$ степенями свободы, можно записать

$$P\left(\chi_1^2 \le \frac{nS^2}{\sigma^2} \le \chi_2^2\right) = 1-\alpha.$$

Далее по таблице χ^2 - распределения нужно выбрать такие два значения χ_1^2 и χ_2^2, чтобы площадь, заключенная под дифференциальной функцией распределения χ^2 между χ_1^2 и χ_2^2, была равна $\gamma = 1-\alpha$.

Обычно χ_1^2 и χ_2^2 выбирают так, чтобы

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_1^2) = \frac{\alpha}{2}, \qquad (2.22)$$

т.е. площади, заштрихованные на рис. 2.1 были равны между собой.

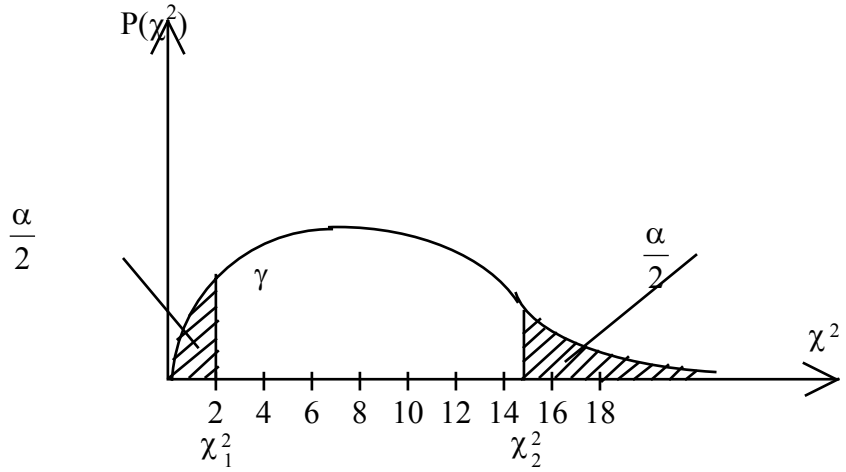


Рис. 2.1

Тогда имеем

$$P\left(\chi_1^2 \le \frac{nS^2}{\sigma^2} \le \chi_2^2\right) = 1 - P(\chi^2 < \chi_1^2) - P(\chi^2 > \chi_2^2). \qquad (2.23)$$

Так как таблица χ^2 - распределения содержит лишь $P(\chi^2 > \chi_{\alpha,\nu}^2)$, то для вычисления $P(\chi^2 < \chi_1^2)$ запишем следующее тождество:

$$P(\chi^2 < \chi_1^2) = 1 - P(\chi^2 > \chi_1^2). \qquad (2.24)$$

Подставив (2.24) в (2.23), получим

$$P(\chi_1^2 \le \frac{nS^2}{\sigma^2} \le \chi_2^2) = P(\chi^2 > \chi_1^2) - P(\chi^2 > \chi_2^2) = \gamma$$

и окончательно

$$\gamma = P(\chi_1^2) - P(\chi_2^2). \qquad (2.25)$$

Формула (2.25) используется при решении обратной задачи - нахождении доверительной вероятности по заданному доверительному интервалу генеральной дисперсии.

Причем $P(\chi_1^2) = P(\chi^2 > \chi_1^2) = 1 - \frac{\alpha}{2};$

$$P(\chi_2^2) = P(\chi^2 > \chi_2^2) = \frac{\alpha}{2}. \qquad (2.26)$$

Преобразуем двойное неравенство в (2.23):

$$\chi_1^2 \le \frac{nS^2}{\sigma^2} \le \chi_2^2, \qquad (2.27)$$

окончательно получим

$$\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}. \qquad (2.28)$$

Это и есть доверительный интервал для генеральной дисперсии, когда неизвестно значение генеральной средней и по выборке объемом n вычисляется выборочная дисперсия S^2.

Ширина доверительного интервала для генеральной дисперсии равна

$$h = \sigma^2_{max} - \sigma^2_{min} = \frac{nS^2}{\chi_1^2} - \frac{nS^2}{\chi_2^2}. \qquad (2.29)$$

Доверительный интервал для генерального среднего квадратического отклонения σ при $n \le 30$ равен

$$\frac{\sqrt{n}S}{\chi_2} \le \sigma \le \frac{\sqrt{n}S}{\chi_1}. \qquad (2.30)$$

При достаточно больших объемах выборки ($n > 30$) доверительный интервал для генерального среднего квадратического отклонения определяется по формуле

$$\frac{\sqrt{2n}}{\sqrt{2n-3} + t_\gamma} \cdot S \le \sigma \le S \cdot \frac{\sqrt{2n}}{\sqrt{2n-3} - t_\gamma}, \qquad (2.31)$$

где t - нормированное значение нормальной случайной величины, соответствующее заданной надежности γ и определяемое по таблице функции Лапласа $\Phi(t)$.

Пример 2.3. По результатам контроля $n = 9$ деталей вычислено выборочное среднее квадратическое отклонение $S = 5$ мм. В предположении, что ошибка изготовления деталей распределена нормально, определить с надежностью $\gamma = 0{,}95$ доверительный интервал для параметра σ.

Решение. Так как $n < 30$, то используется χ^2 распределение. Согласно (2.26)

$$P(\chi^2 > \chi_1^2) = 1 - \frac{\alpha}{2} = 1 - \frac{0{,}05}{2} = 0{,}975;$$

$$P(\chi^2 > \chi_2^2) = \frac{\alpha}{2} = \frac{0{,}05}{2} = 0{,}025.$$

По таблице χ^2 - распределения для числа степеней свободы $\nu = n - 1 = 8$ и найденных вероятностей 0,975 и 0,025 определяем, что $\chi_1^2 = 2{,}180$ и $\chi_2^2 = 17{,}535$.

Вычисляем $\chi_1 = \sqrt{2{,}18} = \mathbf{1{,}47}$ **и** $\chi_1 = \sqrt{17{,}535} = 4{,}19$.

Доверительный интервал (2.30) равен

$$\frac{\sqrt{9}\cdot 5}{4{,}19} \le \sigma \le \frac{\sqrt{9}\cdot 5}{1{,}47}$$

и окончательно

$$3{,}58 \le \sigma \le 10{,}2 \text{ (мм)}.$$

2.4.3. Интервальные оценки для генеральной доли

Пусть в n независимых испытаниях некоторое событие A, вероятность появления которого в каждом испытании равна p, имело место m раз, где $0 \le m \le n$, тогда границы доверительного интервала для генеральной доли определяются из уравнений

$$\sum_{i=m}^{n} c_n^i \cdot p_1^i (1-p_1)^{n-i} = \frac{\gamma}{2};$$

$$\sum_{i=0}^{m} c_n^i \cdot p_2^i (1-p_2)^{n-i} = \frac{\gamma}{2}.$$

Эти уравнения решаются приближенно. Для различных значений m, n и надежности γ могут быть найдены p_1 и p_2. Могут быть составлены специальные таблицы.

При достаточно больших n ($n > 30$) можно считать, что частость $\omega = \frac{m}{n}$ имеет приближенно нормальное распределение с параметрами $N\left(p; \sqrt{\frac{pq}{n}}\right)$. В этом случае доверительный интервал для генеральной доли p определяется соотношением

$$\frac{m}{n} - t_\gamma \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} \le p \le \frac{m}{n} + t_\gamma \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}, \qquad (2.32)$$

где t_γ определяется по таблице интегральной функции Лапласа $\Phi(t)$:

$\frac{m}{n}$ - частость события A;

$\left(1-\frac{m}{n}\right)$ - частость противоположного события $\overline{A}$;

n - объем выборки.

Точность оценки генеральной доли p равна

$$\Delta = t_\gamma \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}. \quad (2.33)$$

Пример 2. При испытании зерна на всхожесть из $n = 400$ зерен проросло $m = 384$. С надежностью $\gamma = 0{,}98$ определить доверительный интервал для генеральной доли p.

Решение. По таблице интегральной функции Лапласа из условия $\gamma = \Phi(t_\gamma) = 0{,}98$ определяем $t_\gamma = 3{,}06$.

Учитывая, что $\frac{m}{n} = \frac{384}{400} = 0{,}96$, определим точность оценки

$$\Delta = 3{,}06\sqrt{\frac{0{,}96\cdot 0{,}04}{400}} = \frac{3{,}06}{20}\sqrt{0{,}0384} = 0{,}153\cdot 0{,}196 = 0{,}03$$

Доверительный интервал равен

$$0{,}96 - 0{,}03 \le p \le 0{,}96 + 0{,}03$$

и окончательно

$$0{,}93 \le p \le 0{,}99$$

В заключении приведем табл. 2.1, в которой укажем формулы, используемые при интервальном оценивании основных параметров распределений.

Таблица 2.1

Основные формулы, используемые при интервальном оценивании параметров распределений

Оцениваемый параметр	Условия оценки	Используемое распределение	Основные формулы	Доверительный интервал
μ	σ известно	$\Phi(t)$	$\gamma \leftrightarrow t_\gamma;$ $\Delta = t_\gamma \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm \Delta$
	σ не известно	St	$\left.\begin{matrix}\alpha \\ \nu = n-1\end{matrix}\right\} \leftrightarrow t_\alpha;$ $\Delta = t_\alpha \frac{S}{\sqrt{n-1}}$	
σ^2	$n \le 30$ μ известно	χ^2	$\left.\begin{matrix}1-\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_1^2;$ $\left.\begin{matrix}\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_2^2;$ $\gamma = P(\chi_1^2) - P(\chi_2^2)$	$\frac{nS_*^2}{\chi_2^2} \le \sigma^2 \le \frac{nS_*^2}{\chi_1^2}$
	$n \le 30$ μ не известно	χ^2	$\left.\begin{matrix}1-\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_1^2;$ $\left.\begin{matrix}\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_2^2;$ $\gamma = P(\chi_1^2) - P(\chi_2^2)$	$\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}$
	$n > 30$	$\Phi(t)$	$\gamma \to t_\gamma$	$\frac{\sqrt{2n}}{\sqrt{2n-3}+t_\gamma}\cdot S \le \sigma \le$
p	$n \to \infty$	$\Phi(t)$	$\gamma \leftrightarrow t_\gamma$ $\Delta = t_\gamma\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}$	$\frac{m}{n} \pm \Delta$

Таблица 2.1

Основные формулы, используемые при интервальном

оценивании параметров распределений

Оцениваемый параметр	Условия оценки	Используемое распределение	Основные формулы	Доверительный интервал
μ	σ известно	$\Phi(t)$	$\gamma \leftrightarrow t_\gamma;$ $\Delta = t_\gamma \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm \Delta$
	σ не известно	St	$\left.\begin{matrix}\alpha \\ \nu = n-1\end{matrix}\right\} \leftrightarrow t_\alpha;$ $\Delta = t_\alpha \frac{S}{\sqrt{n-1}}$	
σ^2	$n \le 30$ μ известно	χ^2	$\left.\begin{matrix}1-\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_1^2;$ $\left.\begin{matrix}\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_2^2;$ $\gamma = P(\chi_1^2) - P(\chi_2^2)$	$\frac{nS_*^2}{\chi_2^2} \le \sigma^2 \le \frac{nS_*^2}{\chi_1^2}$
	$n \le 30$ μ не известно	χ^2	$\left.\begin{matrix}1-\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_1^2;$ $\left.\begin{matrix}\frac{\alpha}{2} \\ \nu = n-1\end{matrix}\right\} \leftrightarrow \chi_2^2;$ $\gamma = P(\chi_1^2) - P(\chi_2^2)$	$\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}$
	$n > 30$	$\Phi(t)$	$\gamma \rightarrow t_\gamma$	$\frac{\sqrt{2n}}{\sqrt{2n-3}+t_\gamma} \cdot S \le \sigma \le$
p	$n \to \infty$	$\Phi(t)$	$\gamma \leftrightarrow t_\gamma$ $\Delta = t_\gamma \sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}$	$\frac{m}{n} \pm \Delta$

Пояснения к табл. 2.1.

1. Стрелка вправо ($\rightarrow$) означает порядок решения "прямой" задачи, т.е. определения доверительного интервала по заданной доверительной вероятности.

2. Стрелка вправо ($\leftarrow$) означает порядок решения "обратной" задачи, т.е. определения доверительной вероятности γ по заданному доверительному интервалу.

3. $\Phi(t)$, $S(t)$ и χ^2 - соответствующие таблицы законов распределения: нормального, Стьюдента, Пирсона.

4. Δ - точность оценки соответствующих параметров.

5. $h = 2\Delta$ - ширина доверительного интервала параметров μ или p.

Тест

1. Какая статистика является несмещенной оценкой математического ожидания:

1. $S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$
2. $M_3^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}$
3. $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
4. $M_4^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}$

2. Какая статистика является несмещенной оценкой генеральной дисперсии:

1. $S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$
2. $\hat{S}^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
3. $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
4. $M_3^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}$

3. Какая оценка параметра является несмещенной:

1. Если дисперсия оценки является минимальной.
2. Если математическое ожидание оценки равно значению оцениваемого параметра
3. Если математическое ожидание оценки меньше значения оцениваемого параметра
4. Если расстояние между оценкой и параметром не превышает 3σ

4. Для расчета интервальной оценки математического ожидания μ по выборке объема n, при известной дисперсии, точность оценки определяется по формуле:

1. $\Delta = t_\gamma \sqrt{\frac{1}{n-3}}$
2. $\Delta = t_\gamma \frac{S}{\sqrt{n-1}}$
3. $\Delta = t_\gamma \frac{\sigma}{\sqrt{n}}$
4. $\Delta = t_\gamma \sqrt{\frac{1}{n-4}}$

5. Для расчета нижней границы доверительного интервала математического ожидания μ, при неизвестной дисперсии, используют формулу:

1. $\frac{nS^2}{X_1^2}$
2. $\overline{x} - t_\gamma \frac{\sigma}{\sqrt{n}}$
3. $\overline{x} - t_\alpha \sqrt{\frac{1}{n-3}}$
4. $\overline{x} - t_\alpha \frac{S}{\sqrt{n-1}}$

6. Для расчета верхней границы доверительного интервала генеральной дисперсии σ^2, если объем выборки составляет $n \le 30$, используют формулу:

1. $\frac{nS^2}{\chi_1^2}$
2. $\overline{x} - t_\gamma \frac{\sigma}{\sqrt{n}}$
3. $\frac{nS^2}{\chi_2^2}$
4. $\frac{\sqrt{2n}}{\sqrt{2\nu - 1} - t} \cdot S$

7. С вероятностью $\gamma = 0{,}95$ найти нижнюю границу доверительного интервала для математического ожидания μ случайной величины x, если $n = 9$, $\overline{x} = 44$, $S = 3$

1. 31,25
2. 41,55
3. 46,41
4. 32,75

8. С вероятностью $\gamma = 0{,}95$ найти нижнюю границу доверительного интервала для генерального среднего квадратического отклонения σ случайной величины X, если $n = 9$, $S = 3$

1. 1,65
2. 3,35
3. 2,15
4. 4,75

9. Определить доверительную вероятность γ интервальной оценки математического ожидания μ случайной величины X, если точность оценки равна $\Delta = 2{,}45$ найдена по выборке, с характеристиками: $n = 9$, $\overline{x} = 44$, $S = 3$

1. 0,99
2. 0,76
3. 0,87
4. 0,95

10. Определить доверительную вероятность γ интервальной оценки генеральной дисперсии σ^2, случайной величины X, если верхняя граница интервала равна 37,21 , а n = 9 и S = 3

1. 0,99
2. 0,76
3. 0,87
4. 0,95

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Проверка статистической гипотезы и статистического критерия

Статистическая проверка гипотез тесно связана с теорией оценивания параметров распределений. В экономике, технике, естествознании, медицине, демографии и т.д. часто для выяснения того или иного случайного явления прибегают к высказыванию гипотез (предположений), которые можно проверить статистически, т.е. опираясь на результаты наблюдений в случайной выборке.

Статистической гипотезой называют любое предположение о виде неизвестного закона распределения случайной величины или значении его параметров.

Статистическую гипотезу, однозначно определяющую закон распределения, называют **простой**, в противном случае ее называют **сложной**.

Например, статистической является гипотеза о том, что распределение производительности труда рабочих, выполняющих одинаковую работу в одинаковых организационно-технических условиях, имеет нормальный закон распределения, или статистической является также гипотеза о том, что средние размеры деталей, производимых на однотипных, параллельно работающих станках, не различаются между собой.

Основные принципы проверки статистических гипотез состоят в следующем. Пусть $f(X,\theta)$ - закон распределения случайной величины X, зависящей от одного параметра θ. Предположим, что необходимо проверить гипотезу о том, что $\theta = \theta_0$, где θ_0 - определенное число. Назовем эту гипотезу нулевой (проверяемой) и обозначим ее через H_0.

Нулевой гипотезой H_0 называют выдвинутую гипотезу, которую необходимо проверить.

Конкурирующей (альтернативной) гипотезой H_1 называют гипотезу, противоположную нулевой.

Таким образом, задача заключается в проверке гипотезы H_0 относительно конкурирующей гипотезы H_1 на основании выборки, состоящей из n независимых наблюдений X_1, X_2, ... , X_n над случайной величиной X. Следовательно, все возможное множество выборок объемом n можно разделить на два непересекающихся подмножества (обозначим их через Q и W) таких, что проверяемая гипотеза H_0 должна быть отвергнута, если наблюдаемая выборка попадает в подмножество W, и принята если наблюдаемая выборка принадлежит подмножеству Q.

Подмножество W называют **критической областью**, Q - **областью допустимых значений**.

Вывод о принадлежности данной выборки к соответствующему подмножеству делают по статистическому критерию.

Статистическим критерием называют однозначно определенное правило, устанавливающее условия, при которых проверяемую гипотезу H_0 следует либо отвергнуть либо не отвергнуть.

Основой критерия является специально составленная выборочная характеристика (статистика) $Q^* = f(X_1, X_2, ..., X_n)$, точное или приближенное распределение которой известно.

Основные правила проверки гипотезы состоят в том, что если наблюдаемое значение статистики критерия попадает в критическую область, то гипотезу отвергают, если же оно попадает в область допустимых значений, то гипотезу не отвергают (или принимают).

Такой принцип проверки гипотезы не дает логического доказательства или опровержения гипотезы. При использовании этого принципа возможны четыре случая:

- гипотеза H_0 верна и ее принимают согласно критерию;
- гипотеза H_0 неверна и ее отвергают согласно критерию;
- гипотеза H_0 верна но ее отвергают согласно критерию;

т.е. допускается ошибка, которую принято называть **ошибкой первого рода**;

- гипотеза H_0 неверна и ее принимают согласно критерию,

т.е. допускается **ошибка второго рода**.

Уровнем значимости $\alpha = 1-\gamma$ называют вероятность совершить ошибку первого рода , т.е. вероятность отвергнуть нулевую гипотезу H_0, когда она верна. С уменьшением α возрастает вероятность ошибки второго рода β.

Мощностью критерия $(1 - \beta)$ называют вероятность того, что нулевая гипотеза H_0 будет отвергнута, если верна конкурирующая гипотеза H_1, т.е. вероятность не допустит ошибку второго рода.

Обозначим через $P(Q^* \in W/H)$ вероятность попадания статистики критерия Q^* в критическую область W, если верна соответствующая гипотеза H.

Тогда требования к критической области аналитически можно записать следующим образом:

$$\left.\begin{aligned} P(Q^* \in W/H_0) &= \alpha \\ P(Q^* \in W/H_1) &= \max \end{aligned}\right\}, \qquad (3.1)$$

где H_0 - нулевая гипотеза;

H_1 - конкурирующая гипотеза.

Второе условие выражает требование максимума мощности критерия.

Из условий (3.1) следует, что критическую область нужно выбирать так, чтобы вероятность попадания в нее была бы минимальной (равной α), если верна нулевая гипотеза H_0, и максимальной в противоположном случае.

В зависимости от содержания конкурирующей гипотезы H_1 выбирают правостороннюю, левостороннюю или двустороннюю критические области.

Границы критической области при заданном уровне значимости α находят из соотношений:

при правосторонней критической области

$$P(Q^* > Q_{кр}) = \alpha; \qquad (3.2)$$

при левосторонней критической области

$$P(Q^* < Q_{кр}) = \alpha; \qquad (3.3)$$

при двусторонней критической области

$$P(Q^* > Q_{кр.пр.}) = \frac{\alpha}{2};$$

$$P(Q^* < Q_{кр.лев.}) = \frac{\alpha}{2}. \quad (3.4)$$

где $Q_{кр.лев.}$ - левосторонняя, а $Q_{кр.пр.}$ - правосторонняя граница критической области.

Следует иметь ввиду, что статистические критерии не доказывают справедливости гипотезы, а лишь устанавливают на принятом уровне значимости ее согласие или несогласие с результатом наблюдений.

При проверки статистических гипотез наряду с известными уже нам законами распределения используется распределение Фишера-Снедекора (F- распределение).

3.2. Распределение Фишера-Снедекора

Во многих задачах математической статистики, особенно в дисперсионном анализе в проверке статистических гипотез, важную роль играет F - распределение. Это распределение отношения двух выборочных дисперсий впервые было исследовано английским статистиком Р. Фишером. Однако оно нашло широкое применение в статистических исследованиях лишь после того, как американский статистик Дж. Снедекор составил таблицы для данного распределения. В этой связи F - распределение называют распределением Фишера-Снедекора.

Пусть имеем две независимые случайные величины X и Y, подчиняющиеся нормальному закону распределения. Произведены две независимые выборки объемами n_1 и n_2, и вычислены выборочные дисперсии S_1^2 и S_2^2. Известно, что случайные величины $U_1^2 = \frac{n_1 S_1^2}{\sigma_1^2}$ и $U_2^2 = \frac{n_2 S_2^2}{\sigma_2^2}$ имеют χ^2 - распределение с соответственно $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$ степенями свободы. Случайная величина

$$F = \frac{U_1^2 / \nu_1}{U_2^2 / \nu_2} \qquad (3.5)$$

имеет F - распределение с ν_1 и ν_2 степенями свободы. Причем $U_1^2 \geq U_2^2$, так что $F \geq 1$.

Закон распределения случайной величины F не зависит от неизвестных параметров (μ_1, σ_1^2) и (μ_2, σ_2^2) а зависит лишь от числа наблюдений в выборках n_1 и n_2. Составлены таблицы распределения случайной величины F, в которых различным значениям уровня значимости α и различным сочетаниям величин ν_1 и ν_2 соответствуют такие значения $F(\alpha, \nu_1, \nu_2)$, для которых справедливо равенство $P[F > F(\alpha, \nu_1, \nu_2)] = \alpha$.

3.3. Гипотезы о генеральных средних нормально распределенных совокупностей

3.3.1. Проверка гипотезы о значении генеральной средней

Пусть из генеральной совокупности X, значения признака которой имеют нормальный закон распределения с параметрами $N(\mu, \sigma)$ при неизвестном математическом ожидании μ и неизвестной дисперсии σ^2, взята случайная выборка объемом n и вычислена выборочная средняя арифметическая $\overline{x}$, а μ_0 и μ_1 - определенные значения параметра μ. Для проверки нулевой гипотезы H_0: $\mu = \mu_0$ при конкурирующей гипотезе H_1: $\mu = \mu_1$ используют статистику

$$t_H = \frac{\overline{x} - \mu_0}{\sigma} \sqrt{n}, \qquad (3.6)$$

которая при выполнении нулевой гипотезы имеет нормированное нормальное распределение N(0;1).

Согласно требованию к критической области при $\mu_1 > \mu_0$ выбирают правостороннюю критическую область, при $\mu_1 < \mu_0$ - левостороннюю, а при $\mu_1 \neq \mu_0$ - двустороннюю критическую область.

Границы критической области $t_{кр}$ определяют по интегральной функции Лапласа $\Phi(t)$ из условий:

в случае правосторонней и левосторонней критической областей

$$\Phi(t_{кр}) = 1 - 2\alpha,$$

где $\Phi(t_{кр}) = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{z^2}{2}}dz$ - интегральная функция Лапласа;

в случае двусторонней критической области

$$\Phi(t_{кр}) = 1 - \alpha. \qquad (3.8)$$

При проверке гипотезы о значении генеральной средней H_0: $\mu = \mu_0$ при неизвестной генеральной дисперсии σ^2 используют статистику

$$t_H = \frac{\overline{x} - \mu_0}{S}\sqrt{n-1}, \qquad (3.9)$$

которая при выполнении нулевой гипотезы H_0 имеет распределение Стьюдента (t - распределение) с $\nu = n-1$ степенями свободы.

Границы критической области $t_{кр}$ определяют по таблице t - распределения для заданного уровня значимости α (при двусторонней симметричной критической области) или 2α (при правосторонней и левосторонней критических областях) и числа степеней свободы $\nu = n - 1$.

Правила проверки гипотезы сводятся к следующему:

1) при левосторонней критической области, если $t_H \geq -t_{кр}$, нулевая гипотеза H_0 не отвергается;

2) при правосторонней критической области, если $t_H < -t_{кр}$, нулевая гипотеза H_0 не отвергается;

3) при двусторонней критической области, если $|t_H| \leq -t_{кр}$, нулевая гипотеза H_0 не отвергается;

В противном случае нулевая гипотеза H_0 отвергается с вероятностью ошибки α.

3.3.2. Проверка гипотезы о равенстве генеральных средних двух номинальных совокупностей

Пусть X и Y - нормальные генеральные совокупности с известными генеральными дисперсиями σ_1^2 и σ_2^2 и неизвестными математическими ожиданиями μ_x и μ_y. Из генеральных совокупностей взяты две независимые выборки объемами n_1 и n_2 и вычислены средние арифметические $\overline{x}$ и $\overline{y}$. Для проверки гипотезы о равенстве генеральных средних $H_0: \mu_x = \mu_y$ используют статистику

$$t_H = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (3.10)$$

которая при выполнении нулевой гипотезы имеет нормированный нормальный закон распределения N(0;1).

Выбор критической области зависит от содержания конкурирующей гипотезы H_1. Согласно требованию к критической области при H_1: $\mu_x > \mu_y$ выбирают

правостороннюю, при H_1: $\mu_x < \mu_y$ - левостороннюю, а при H_1: $\mu_x \neq \mu_y$ - двустороннюю критические области.

Границы критических областей находят по интегральной функции Лапласа из условий (3.7) и (3.8).

При неизвестных генеральных дисперсиях либо требуется достаточно большой объем выборки для надежной и точной оценки, либо требуется, чтобы эти дисперсии были одинаковы, в противном случае известные критерии малоэффективны.

Если генеральные дисперсии равны $\sigma_1^2 = \sigma_2^2$, то для проверки гипотезы H_0: $\mu_x = \mu_y$ используют статистику

$$t_H = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}, \qquad (3.11)$$

имеющую распределение Стьюдента с $\nu = n_1 + n_2 - 2$ степенями свободы. Вид критической области зависит, как обычно, от конкурирующей гипотезы.

Границы критической области ($t_{кр}$) находят по таблице распределения Стьюдента при двусторонней симметричной критической области для заданного уровня значимости α, а при правосторонней и левосторонней критических областях при 2α.

Правила проверки гипотезы H_0: $\mu_x = \mu_y$ такие же, как гипотезы H_0: $\mu = \mu_0$. Гипотеза H_0 отвергается при $|t_H| > t_{кр}$.

3.4. Гипотезы о генеральных дисперсиях нормально распределенных генеральных совокупностях

3.4.1. Проверка гипотезы о значении генеральной дисперсии

Пусть из генеральной совокупности, значения признака которой распределены по нормальному закону с неизвестной дисперсией σ^2, взята случайная выборка из n независимых наблюдений и вычислена выборочная дисперсия S^2.

Требуется проверить нулевую гипотезу H_0: $\sigma^2 = \sigma_0^2$, где σ_0^2 - определенное заданное значение генеральной дисперсии. Для проверки нулевой гипотезы используют статистику

$$U_H^2 = \frac{nS^2}{\sigma_0^2}, \qquad (3.12)$$

которая при выполнении гипотезы H_0 имеет распределение χ^2 с $\nu = n - 1$ степенями свободы.

Как было сказано ранее, в зависимости от конкурирующей гипотезы выбирают правостороннюю, левостороннюю или двустороннюю критическую область.

Границы критической области $\chi^2_{кр}$ определяют по таблице распределения Пирсона χ^2.

Рассмотрим три случая:

1. Если $H_1: \sigma^2_1 > \sigma^2_0$, то выбирают правостороннюю критическую область и $\chi_{кр}^2$ находят из условия

$$P\left[U_H^2 > \chi_{кр}^2(\alpha, n-1)\right] = \alpha,$$

где $\chi_{кр}^2$ (α, n-1) - табличное значение χ^2 , найденное для уровня значимости α и числа степеней свободы $\nu = n - 1$.

Правила проверки гипотезы заключается в следующем:

1) если $U_Н^2 \le \chi_{кр}^2$, то нулевая гипотеза не отвергается;

2) если $U_Н^2 > \chi_{кр}^2$, то нулевая гипотеза отвергается;

2. Если H_0: $\sigma_1^2 \ne \sigma_0^2$, то строят двустороннюю симметричную критическую область и ее границы $\chi_{кр.лев}^2$ и $\chi_{кр.пр}^2$ находят из условий

$$P\left[U_Н^2 > \chi_{кр.лев}^2\left(1-\frac{\alpha}{2};n-1\right)\right] = 1-\frac{\alpha}{2};$$

$$P\left[U_Н^2 > \chi_{кр.пр.}^2\left(\frac{\alpha}{2};n-1\right)\right] = \frac{\alpha}{2}. \qquad (3.14)$$

Правила проверки гипотезы заключаются в следующем:

1) если $\chi_{кр.лев}^2 \le U^2{}_Н \le \chi_{кр.пр}^2$, то гипотеза не отвергается;

2) если $U_Н^2 < \chi_{кр.лев}^2$ или $U_Н^2 > \chi_{кр.пр}^2$, то гипотеза отвергается;

3. Если H_1: $\sigma_1^2 < \sigma_0^2$, то строят левостороннюю критическую область и $\chi_{к р}^2$ находят из условия

$$P\left[U_Н^2 > \chi_{кр.}^2\left(1-\alpha;n-1\right)\right] = 1-\alpha. \qquad (3.15)$$

Правила проверки гипотезы заключаются в следующем:

1) если $U_Н^2 \ge \chi_{кр.}^2$, то гипотеза не отвергается;

2) если $U_Н^2 < \chi_{кр.}^2$, то гипотеза отвергается;

3.4.2. Проверка гипотезы о равенстве генеральных дисперсий двух нормальных совокупностей

Пусть X и Y - генеральные совокупности, значения признаков которых распределены по нормальному закону с дисперсиями σ_1^2 и σ_2^2. Из этих совокупностей взяты две независимые выборки объемами n_1 и n_2 и вычислены исправленные выборочные дисперсии $\hat{S}^2{}_1$ и $\hat{S}^2{}_2$, причем $\hat{S}_1^2 > \hat{S}_2^2$.

Требуется проверить нулевую гипотезу H_0: $\sigma_1^2 = \sigma_2^2$ против конкурирующей гипотезы H_1: $\sigma_1^2 > \sigma_2^2$. Основу критерия для проверки нулевой гипотезы составляет статистика

$$F_H = \frac{\hat{S}_1^2}{\hat{S}_2^2}, \qquad (3.16)$$

где $\hat{S}_1^2 > \hat{S}_2^2$, которая при выполнении нулевой гипотезы имеет распределение Фишера-Снедекора (F- распределение) со степенями свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$, где ν_1 - число степеней свободы числителя, а ν_2 - число степеней свободы знаменателя (меньшей дисперсии).

Для проверки гипотезы выбирают правостороннюю критическую область. Границу критической области $F_{кр.}$ определяют по таблице F - распределения из условия

$$P\left[F_Н > F_{кр.}\left(\alpha;n_1-1;n_2-1\right)\right] = \alpha. \qquad (3.17)$$

Правила проверки гипотезы заключаются в следующем:

1) если $F_H \le F_{кр.}$, то гипотеза не отвергается;

2) если $F_H > F_{кр.}$, то гипотеза отвергается.

3.4.3. Проверка гипотезы об однородности ряда дисперсий

При сравнении более двух генеральных дисперсий применяют два наиболее часто употребляемых критерия: критерий Бартлета и критерий Кохрана.

Критерий Бартлета применятся при проверке гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$ по выборкам разного объема $n_1 \ne n_2 \ne \ldots \ne n_1$.

В качестве выборочной характеристики Барлет предложил использовать статистику

$$U_H^2 = \frac{\nu \ln \hat{S}_{-p}^2 - \sum_{\ell=1}^{\ell} \nu_i \ln \hat{S}_i^2}{1 - \frac{1}{3(\ell-1)}\left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\nu}\right]}, \quad (3.16)$$

где $\nu_i = n_i - 1$ - число степеней свободы i-ой выборки;

$\hat{S}_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$ - исправленная дисперсия i-ой выборки;

x_{ij} - результат j-ого наблюдения в i-ой выборки;

$\bar{x}_i$ - средняя арифметическая i-ой выборки;

l - число выборок;

$\nu = \sum_{i=1}^{l} \nu_i$ - сумма чисел степеней свободы l выборок;

$\hat{S}_{-p}^2 = \frac{\sum_{i=1}^{\ell} \hat{S}_i^2 \cdot \nu_i}{\sum_{i=1}^{\ell} \nu_i}$ - среднее значение исправленной дисперсии по всем l выборкам;

При выполнении нулевой гипотезы и при $\nu_i > 3$ статистика U_H^2 приближенно имеет распределение χ^2 с числом степеней свободы $\nu = l - 1$.

Для проверки нулевой гипотезы строят правостороннюю критическую область, границы которой $\chi_{кр.}^2$ определяют по таблице χ^2 - распределения из условия:

$$P\left[U_H^2 > \chi_{кр.}^2(\alpha; l-1)\right] = \alpha \quad (3.19)$$

Критерий Бартлета весьма чувствителен к отклонениям законов распределения случайных величин X_i от нормального закона распределения.

Критерий Кохрана применяется при проверке гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$ по выборкам одинакового объема n, взятым соответственно из нормальных генеральных совокупностей.

Для проверки нулевой гипотезы Кохран предложил критерий, основанный на статистике

$$G_H = \frac{\hat{S}^2_{\max}}{\sum_{i=1}^{\ell} \hat{S}_i^2}, \qquad (3.20)$$

которая при выполнении нулевой гипотезы имеет G - распределение с числом степеней свободы ν = n - 1 и числа сравниваемых совокупностей l , где $\hat{S}^2_{\max}$ - наибольшая из исправленных выборочных дисперсий.

Для проверки нулевой гипотезы также строят правостороннюю критическую область, границу которой $G_{кр}$ определяют по таблице G - распределения из условия

$$P\left[G_{Н} > G_{кр.}\left(\alpha; n-1; l\right)\right] = \alpha \qquad (3.21)$$

Правила проверки гипотезы заключаются в следующем:

1) если $G_{Н} \le G_{кр.}$ - то нулевая гипотеза не отвергается;

2) если $G_{Н} > G_{кр.}$ - то нулевая гипотеза отвергается;

3.5. Гипотеза об однородности ряда вероятностей

Пусть $X_1, X_2, ..., X_l$ - l генеральных совокупностей, каждая из которых характеризуется неизвестным параметром P_i, где P_i - вероятность появления события A в соответствующей выборке.

Требуется по результатам выборочных наблюдений проверить нулевую гипотезу о равенстве вероятностей появления события A в генеральных совокупностях, т.е. H_0: $p_1 = p_2 = \ldots = p_l$.

Для проверки гипотезы используется статистика

$$U_H^2 = \frac{1}{\tilde{p}(1-\tilde{p})} \sum_{i=1}^{\ell} (\tilde{p}_i - \tilde{p})^2 \cdot n_i \,. \quad (3.22)$$

где $\omega_i = \frac{m_i}{n_i}$ - частость появления события A в i-ой выборке;

m_i - частота появления события A в i-ой выборке;

n_i - объем i-ой выборки;

l - число выборок;

$\tilde{p} = \frac{\sum m_i}{\sum n_i}$ - частость появления события A во всех выборках;

$\tilde{q} = 1 - \tilde{p}$ - частота появления события $\overline{A}$ во всех выборках;

Статистика $U^2_{Н}$ при выполнении нулевой гипотезы имеет асимптотическое χ^2 - распределение с ν = l - 1 степенями свободы.

Для проверки нулевой гипотезы строят **правостороннюю** критическую область, границу которой определяют из условия

$$P\left[U^2_{Н} > \chi^2_{кр.}\left(\alpha; l-1\right)\right] = \alpha \,. \quad (3.23)$$

Правила проверки гипотезы заключаются в следующем:

1) если $U^2_{Н} \le \chi^2_{кр.}$, то гипотеза не отвергается;

2) если $U^2_{Н} > \chi^2_{кр.}$, то нулевая гипотеза отвергается.

При решении задач проверки статистических гипотез необходимо в первую очередь уяснить содержание проверяемой H_0 и конкурирующей H_1 гипотез, так как от этого зависит выбор алгоритма (формулы) для вычисления наблюдаемого

значения критерия. От содержания конкурирующей гипотезы зависит также выбор вида критической области.

В таблице 3.1 приведены основные формулы, используемые при проверке гипотез о значении параметров распределений.

Пример 3.1. Точность работы автоматической линии проверяют по дисперсии контролируемого признака, которая не должна превышать 0,1 мм^2. По результатам выборочного контроля получены следующие данные:

Контролируемый размер, x_i^1, мм	43,0	43,5	43,8	44,4	44,6
Частота m_i	3	7	10	8	2

Требуется проверить на уровне значимости 0,01, обеспечивает ли линия требуемую точность.

Решение. Задача состоит в проверке гипотезы о значении генеральной дисперсии H_0: $\sigma_0^2 \le 0{,}1$. Автоматическая линия не обеспечивает требуемой точности, если H_1^2: $\sigma_1^2 > \sigma_0^2$, следовательно в данном случае строится правосторонняя критическая область.

Наблюдаемое значение критерия вычисляем по формуле $U_{\text{Н}}^2 = \dfrac{nS^2}{\sigma_0^2}$, следовательно, по данным вариационного ряда сначала необходимо вычислить выборочную дисперсию, для чего определяем среднюю арифметическую и средний квадрат по условным вариантам, принимая $x_0 = 43{,}0$.

x_i	m_i	$x_i^1 = x_i - 43{,}0$	$x_i^1 \cdot m_i$	$(x_i^1)^2 \cdot m_i$
43,0	3	0	0	0
43,5	7	0,5	3,5	1,75
43,8	10	0,8	8,0	6,40
44,4	8	1,4	11,2	15,68
44,6	2	1,6	3,2	5,12
	30	-	25,9	29,95

$$\overline{x}^1 = \frac{\sum x_i^1 \cdot m_i}{\sum m_i} = \frac{25{,}9}{30} = 0{,}863;$$

$$\overline{(x')^2} = \frac{28{,}95}{30} = 0{,}965;$$

$$S^2 = \overline{(x')^2} - (\overline{x'})^2 = 0{,}965 - 0{,}863^2 = 0{,}965 - {,}0745 = 0{,}22 \text{м м}^2.$$

Вычисляем наблюдаемое значение критерия

$$U_{\text{Н}}^2 = \frac{30 \cdot 0{,}22}{0{,}1} = \frac{6{,}6}{0{,}1} = 66.$$

По таблице χ^2 - распределения при заданном уровне значимости $\alpha = 0{,}01$ и $\nu = n - 1 = 30 - 1 = 29$ определяем $\chi_{\text{кр.}} = 49{,}588$.

Сравнивая $U_{\text{Н}}^2$ и $\chi_{\text{кр.}}^2$, получаем $U_{\text{Н}}^2 (= 66{,}0) > \chi_{\text{кр.}}^2 (= 49{,}588)$, т.е. нулевая гипотеза H_o отвергается; так как генеральная дисперсия не равна 0,1 , автоматическая линия не обеспечивает заданную точность и требуется ее регулировка.

Пример 3.2. Во время экзамена студентам были предложены задачи из семи разделов изучаемого курса. Результаты экзамена представлены в таблице.

Требуется на уровне значимости 0,1 проверить гипотезу о том, что вероятность решения задачи не зависит от того, к какому разделу он относится.

Решение. Задача заключается в проверке гипотезы об однородности ряда вероятностей: H_0: $p_1 = p_2 = \ldots = p_7$.

Номер раздела курса	1	2	3	4	5	6	7	∑
Число предложенных задач n_i	165	66	270	160	80	350	150	1241
Доля решенных задач	0,855	0,509	0,522	0,484	0,860	0,412	0,42	-
Число решенных задач m_i	140	34	141	77	69	144	63	688

Наблюдаемое значение критерия вычисляется по формуле (3.22). Сначала необходимо определить среднюю частность решенных задач по всем семи разделам курса:

$$m_1 = n_1 \cdot \tilde{p}_1 = 165 \cdot 0.855 = 140.$$

$$\tilde{p} = \frac{\sum m_i}{\sum n_i} = \frac{140+34+141+77+69+144+63}{165+66+270+160+80+350+150} = \frac{668}{1241} = 0{,}538.$$

Вычисляем необходимое значение критерия

$$U_H^2 = \frac{1}{\tilde{p}(1-\tilde{p})}\sum_{i=1}^{\ell}(\tilde{p}_i - \tilde{p})^2 \cdot n_i = \frac{1}{0{,}538 \cdot 0{,}462}(0{,}855-0{,}538)^2 \cdot 66 +$$

$$+ (0{,}522-0{,}538)^2 \cdot 270 + (0{,}484-0{,}538)^2 \cdot 160 + (0{,}860-0{,}538)^2 \cdot 350 +$$

$$(0{,}420-0{,}538)^2 \cdot 150 = 4{,}023 \cdot (16{,}58 + 0{,}06 + 0{,}07 + 0{,}47 + 8{,}29 + 5{,}57 + 2{,}09) =$$

$$= 4{,}023 \cdot 33{,}13 = 133{,}28.$$

По таблице χ^2 - распределения при заданном уровне значимости $\alpha = 0{,}1$ и $\nu = n - 1 = 6$ определяем $\chi^2_{кр.} = 10{,}645$.

Так как $U^2_{Н} = 133{,}28 > \chi^2_{кр.} = 10{,}645$, нулевая гипотеза отвергается, т.е. ряд вероятностей неоднороден, разделы данного курса студентами усвоены с разной вероятностью.

3.6. Вычисление мощности критерия

Мощность критерия (1 - β) может быть вычислена только при проверке простых статистических гипотез: гипотезы о значении генеральной средней $H_0 : \mu = \mu_0$ и гипотезы о значении генеральной дисперсии $H_0: \sigma^2 = \sigma_0^2$ и только при односторонней критической области.

3.6.1. Мощность критерия при проверке гипотезы о значении генеральной средней

Если известна генеральная дисперсия σ^2, то при проверке гипотезы $H_0 : \mu = \mu_0$ используется нормальное распределение. Для вычисления мощности критерия при односторонней конкурирующей гипотезе применяется формула

$$1-\beta=\frac{1}{2}\left[1+\Phi\left(\frac{|\mu_0-\mu_1|}{\sigma}\sqrt{n}-t_{кр.}\right)\right], \qquad (3.24)$$

где $t_{кр.}=\Phi^{-1}(1-2\alpha)$, (3.25)

т.е. $t_{кр.}$ определяется по таблице функции Лапласа Φ(t) по вероятности $(1-2\alpha)$.

Если генеральная дисперсия неизвестна, то мощность критерия определяется по формулам:

$$1-\beta=1-\frac{1}{2}St\left(\frac{|\mu_1-\mu_0|}{S}\sqrt{n-1}-t_{кр};n-1\right), \qquad (3.26)$$

где $t_{кр}=St^{-1}(2\alpha;n-1)$, (3.27)

т.е. $t_{кр.}$ определяется по таблице распределения Стьюдента по вероятности 2α и $\nu=n-1$.

3.6.2. Мощность критерия при проверке гипотезы о значении генеральной дисперсии

При проверки гипотезы H_0: $\sigma^2=\sigma_0^2$ мощность критерия вычисляется с использованием распределения Пирсона χ^2.

Если H_1: $\sigma_1^2<\sigma_0^2$, то мощность критерия вычисляется по формуле

$$1-\beta=P\left[U_Н^2<\chi^2_{кр.}(1-\alpha;n-1)/H_1\right]=1-P\left[U_Н^2>\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{кр.}(1-\alpha;n-1)\right] \qquad (3.28)$$

Если H_1: $\sigma_1^2>\sigma_0^2$, то мощность критерия вычисляется по формуле

$$1-\beta=P\left[U_Н^2>\chi^2_{кр.}(\alpha;n-1)/H_1\right]=P\left[U_Н^2>\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{кр.}(\alpha;n-1)\right] \qquad (3.29)$$

Пример 3.3. По результатам 7 независимых измерениях диаметра поршня одним и тем же прибором в предположении, что ошибки измерений имеют нормальное распределение, была проведена на уровне значимости 0,05 гипотеза H_0: $\sigma_0^2=0{,}02$ мм^2 при конкурирующей гипотезе H_1: $\sigma_1^2=0{,}05$ мм^2. Гипотеза H_0 не отвергнута. Вычислить мощность критерия.

Решение. Согласно H_1: $\sigma_1>\sigma_0^2$ стоится правосторонняя критическая область.

По таблице χ^2 - распределения на уровне значимости $\alpha=0{,}05$ и при числе степеней свободы $\nu=n-1=6$ определяем $\chi^2_{кр.}=12{,}592$.

Вычисляем $\chi^2_{кр.}(0{,}05;\ \nu=6/H_1)=\frac{\sigma_0^2}{\sigma_1^2}\cdot\chi^2_{кр.}(0{,}05;\ \nu=6)=\frac{0{,}02}{0{,}05}\cdot 12{,}592=5{,}037$.

По $\chi^2_{кр.}/H_1=5{,}037$ и числу степеней свободы $\nu=n-1=6$ по таблице χ^2 - распределения определяем $P\left[U_Н^2>\chi^2_{кр.}/H_1\right]=1-\beta=0{,}541$.

Пример 3.4. При испытаниях были получены значения максимальной скорости самолета: 423, 426, 420, 425, 421, 423, 432, 427, 439, 435 м/с. Сделав предположение, что максимальная скорость самолета есть нормальная случайная величина, проверить гипотезу H_0 : $\mu_0=430$ м/с при конкурирующей гипотезе H_1: $\mu_1=420$ м/с и вычислить мощность критерия при $\alpha=0{,}005$.

Решение. По измененным вариантам $x_i = x_i^1 - 420$ определим среднюю арифметическую $\bar{x}$ и средний квадрат $\left(\overline{x^2}\right)$ условного ряда распределения:

$$\bar{x} = \frac{71}{10} = 7{,}1;$$

$$\overline{x^2} = \frac{859}{10} = 85{,}9.$$

x_i^1	x_i	x_i^2
423	3	9
426	6	36
420	0	0
425	5	25
421	1	1
423	3	9
432	12	144
427	7	49
430	19	361
435	15	225
$\sum$	71	859

Вычисляем выборочную дисперсию:

$$S^2 = \left(S^1\right)^2 = 85{,}9 - 7{,}1^2 = 85{,}9 - 50{,}41 = 35{,}49;$$

$$S = \sqrt{35{,}49} = 5{,}96 \text{ м/с}.$$

Так как генеральная дисперсия не известна, то при вычислении мощности критерия используется распределение Стьюдента.

При $H_0 : \mu_1 < \mu_0$ строится левосторонняя критическая область. По таблице распределения Стьюдента по вероятности $2\alpha = 0{,}01$ и $\nu = n - 1 = 9$ определяем $t_{кр.}/H_0 = 1{,}833$.

Вычисляем 1-β по формуле (3.26):

Обозначим

$$t_{кр.}/H_1 = -t_{кр.}/H_0 + \frac{|\mu_1 - \mu_0|}{S}\sqrt{n-1} = -1{,}833 + \frac{|430 - 420|}{5{,}96}\sqrt{9} =$$

$$= -1{,}833 + 5{,}034 = 3{,}201.$$

По $t_{кр.}/H_1 = 3{,}201$ и $\nu = n - 1 = 9$ по таблице t - распределения определяем $St\left(\left[t_{кр.}/H_1\right]\right) = 0{,}01.$

Вычисляем мощность критерия

$$1 - \beta = 1 - \frac{1}{2}St\left(\left[t_{кр.}/H_1\right]\right) = 1 - \frac{1}{2}\cdot 0{,}01 = 1 - 0{,}005 = 0{,}995.$$

3.7. Гипотезы о виде законов распределения генеральной совокупности

3.7.1. Основные понятия

Проверка гипотез о виде законов распределения генеральной совокупности осуществляется с помощью критериев согласия.

Проверяемая (нулевая) гипотеза утверждает, что полученная выборка взята из генеральной совокупности, значения признака в которой распределены по предлагаемому теоретическому закону (нормальному, биноминальному или другому) с параметрами, либо оцениваемыми по выборке, либо заранее известными.

Математически, нулевую гипотезу можно записать в следующем виде:

$$H_0: \frac{m_1}{n_1} = p_1, \frac{m_2}{n_2} = p_2, \ldots, \frac{m_l}{n_l} = p_l,$$

где $\frac{m_i}{n_i} = \tilde{p}_i$ - относительная частота (частость, доля) i-го интервала вариационного ряда или i-го варианта, принимаемого случайной величиной X;

P_i - вероятность попадания случайной величины в i-й интервал или вероятность того, что дискретная случайная величина примет i-ое значение ($X = x_i$);

$i = \overline{1, \ell}$ - номер интервала или значения случайной величины;

n - объем выборки.

Критерий состоит в том, что выбранная некоторая случайная величина Y является мерой расхождения (рассогласования) между вариационным рядом и предполагаемым теоретическим распределением. При проверке нулевой гипотезы заранее задается уровень значимости α (α = 0,1; 0,05; 0,01; 0,001). Затем на основании закона распределения случайной величины находится такое значение $Y_{кр}$, что

$$P(Y>Y_{кр}) = \alpha. \quad (3.30)$$

Критическое значение $Y_{кр}$ обычно находят по таблице соответствующей функции распределения. Далее вычисляется на основании выборки наблюдаемое значение статистики критерия Y_H. Наконец, сравниваются два значения: Y_H и $Y_{кр}$. Если $Y_H > Y_{кр}$, то нулевая гипотеза отвергается. Если $Y_H \le Y_{кр}$, то нулевая гипотеза не отвергается, т.е. в этом случае отклонения от предполагаемого теоретического закона считаются незначимыми - данные наблюдений не противоречат гипотезе о виде закона распределения.

Можно осуществлять проверку гипотезы о виде закона распределения в другом порядке: по наблюдаемому значению критерия Y_H определить, пользуясь соответствующей таблицей, $\alpha_H = P(Y>Y_H)$. Если $\alpha_H \le \alpha$, то отклонения значимы и гипотеза отвергается; если же $\alpha_H > \alpha$, то гипотеза не отвергается.

3.7.2. Критерий Пирсона

Критерий Пирсона или критерий χ^2 (хи - квадрат) имеет наибольшее применение при проверке согласования теоретической и эмпирических кривых распределения. Наблюдаемое значение критерия ($Y = \chi^2_H$) вычисляется по следующей формуле:

$$\chi^2_H = \sum_{i=1}^{l} \frac{(m_{эi} - m_{тi})^2}{m_{тi}}, \quad (3.31)$$

где $m_{эi}$ - эмпирическая частота i-го интервала (варианта);

$m_{тi}$ - теоретическая частота i-го интервала (варианта);

l - число интервалов (вариантов).

Как известно χ^2 - распределение зависит от числа степеней свободы, это число находится по формуле

$$\nu = l - r - 1, \quad (3.32)$$

где r - число параметров предполагаемого теоретического закона, использованных для вычисления теоретических частот и оцениваемых по выборке.

По теоретическим соображениям при расчете $\chi^2_{\text{Н}}$ не следует исходить из слишком малых значений $m_{\text{т}i}$. Поэтому рекомендуется объединять соседние интервалы (варианты) таким образом, чтобы $m_{\text{Т}i} > (5 \div 10)$ для объединенных интервалов. Кроме того, объем выборки должен быть достаточно велик ($n \geq 50$) и $\sum m_{\text{т}i} = \sum m_{\text{э}i}$.

В случае нормального закона распределения расчет теоретической кривой распределения $\varphi(x)$ производится при условии, что статистические характеристики $(\bar{x}; S)$ приравниваем числовым характеристикам нормального закона (μ; σ), поэтому r = 2 и число степеней свободы ν = l -3.

Вероятности попадания случайной величины X в соответствующие интервалы вычисляется по интегральной теореме Лапласа

$$p_i = P(a_i < x < b_i) = \frac{1}{2}\left[\Phi(t_{2i}) - \Phi(t_{1i})\right], \qquad (3.33)$$

где $t_{1i} = \dfrac{a_i - \bar{x}}{S}$; $t_{2i} = \dfrac{b_i - \bar{x}}{S}$.

В случае биномиального закона распределения расчет теоретической кривой распределения производится при условии, что статистическая доля (частость) приравнивается вероятности p появления интересующего нас события A, поэтому r = 1 и число степеней свободы ν = l -2.

Вероятность p_i того, что случайная величина X принимает значение $x_i = m$, где $m = \overline{o,n}$, определяется по формуле Бернулли

$$p_i = P(X = x_i) = P(X = m) = C_n^m \bar{\omega}^m \cdot (1 - \bar{\omega})^{n-m}, \qquad (3.34)$$

где $\bar{\omega} = \dfrac{\sum_{i=1}^{k} m_i x_i}{k \cdot n}$ - средняя частость проявления появления события во всех k выборках;

n - число испытаний в каждой выборке.

В случае закона Пуассона расчет теоретической кривой распределения производится при условии, что средняя интенсивность $\bar{\lambda}$ приравнивается математическому ожиданию M(x), поэтому r = 1 и ν = l -2.

Вероятность p_i того, что случайная величина X принимает значение $x_i = m$, определяется по формуле Пуассона

$$p_i = P(X = x_i) = P(X = m) = \frac{\bar{\lambda}^m}{m!} e^{-\bar{\lambda}}, \qquad (3.35)$$

где $\bar{\lambda} = \dfrac{\sum_{i=0}^{k} m_i x_i}{\sum_{i=1}^{k} m_i}$ - средняя интенсивность.

m_i - частота появления значения x_i; i=1, 2, ... , к.

При проверке гипотез о виде законов распределения могут быть использованы и другие критерии согласия: Колмогорова, Романовского, Ястремского и др.

Пример 3.5. По данным примера 1.2 рассчитать теоретические частоты в предположении нормального закона распределения; результаты вычислений приводятся в следующей таблице.

Интервалы	3,65-3,75	3,75-3,85	3,85-3,95	3,95-4,05	4,05-4,15	4,15-4,25	4,25-4,35
$m_{эi}$	1	6	11	15	9	6	2
$m_{тi}$	2	5	11	14	11	5	2

На уровне значимости 0,05 проверить гипотезу о нормальном законе распределения.

Решение. Вычисляем наблюдаемое значение критерия $\chi^2_н$ по формуле (3.31). Результаты вычислений представим в виде таблицы.

Интервалы	$m_{эi}$	$m_{тi}$	$(m_{эi} - m_{тi})^2$	$\frac{(m_{эi} - m_{тi})^2}{m_{тi}}$
3,65-3,75 3,75-3,85	1 6	2 5	0	0
3,85-3,95	11	11	0	0
3,95-4,05	15	14	1	$\frac{1}{14} = 0{,}071$
4,05-4,15	9	11	4	$\frac{4}{11} = 0{,}364$
4,15-4,25 4,25-4,35	6 2	5 2	1	$\frac{1}{7} = 0{,}143$
$\sum$	50	50	-	$\chi^2_н = 0{,}578$

По таблице χ^2 - распределения на уровне значимости 0,05 и числе степеней свободы $\nu = 1 - 3 = 5 - 3 = 2$ определим $\chi^2_{кр} = 5{,}991$. Так как $\chi^2_н = 0{,}578 < \chi^2_{кр} = 5{,}991$, нулевая гипотеза H_0 не отвергается, т.е. производительность труда для данной совокупности подчиняется нормальному закону распределения.

Пример 3.6. Даны следующие числа рождения мальчиков у 50 матерей, родивших четыре раза:

3	1	0	2	1	2	1	3	3	3
2	3	2	2	1	2	1	3	2	3
3	0	1	1	2	2	1	0	3	2
0	2	2	2	3	3	2	4	3	3
2	1	1	2	2	3	3	2	3	4

Проверить на уровне 0,01 гипотезу о биноминальном законе распределения.

Решение. Всего 50 матерей родили $N = k \cdot n = 50 \cdot 4 = 200$ детей. Случайной величиной X является число мальчиков в семьях из 4 детей. Построим вариационный ряд:

x_i	0	1	2	3	4	$\sum$
$m_{эi}$	4	10	18	16	2	50

Эмпирическими частотами $m_{эi}$ являются числа матерей, родивших определенное число мальчиков.

Рассчитаем среднюю частоту рождения мальчика:

$$\overline{\omega} = \frac{\sum_{i=0}^{4} m_i x_i}{к \times n} = \frac{0\cdot 4 + 1\cdot 10 + 2\cdot 18 + 3\cdot 16 + 4\cdot 2}{50\cdot 4} = \frac{102}{200} = 0{,}51.$$

По формуле (3.34) вычислим вероятности комбинаций рождения мальчика (и девочки) в семьях из 4 детей:

$m = 0;\ P_{0;4} = (1-\overline{\omega})^4 = 0{,}49^4 = 0{,}0576;$

$m = 1;\ P_{1;4} = n\cdot\overline{\omega}(1-\overline{\omega})^3 = 4\cdot 0{,}51\cdot 0{,}49^3 = 0{,}2401;$

$m = 2;\ P_{2;4} = \frac{n(n-1)}{1\cdot 2}\overline{\omega}^2(1-\overline{\omega})^2 = 6\cdot 0{,}51^3\cdot 0{,}49^2 = 0{,}3747;$

$m = 3;\ P_{3;4} = n\cdot\overline{\omega}^3(1-\overline{\omega}) = 4\cdot 0{,}51^3\cdot 0{,}49 = 0{,}2600;$

$m = 4;\ P_{4;4} = \overline{\omega}^4 = 0{,}51^4 = 0{,}0676.$

Итого: $\sum_{m=0}^{4} P_{m,n} = 1{,}000$

Теоретические частоты равны $m_{тi} = k\cdot p_i$.

Рассчитаем наблюдаемое значение критерия $\chi^2_Н$.

По таблице χ^2 - распределения на уровне значимости $\alpha = 0{,}01$ и при числе степеней свободы $\nu = 1 - 2 = 3 - 2 = 1$ определяем $\chi^2_{кр} = 6{,}635$.

Так как $\chi^2_Н = 0{,}370 < \chi^2_{кр} = 6{,}635$, нулевая гипотеза не отвергается, т.е. число мальчиков в семье из 4 детей данной совокупности подчиняется биноминальному закону распределения.

(к примеру 3.6)

x_i	$m_{эi}$	$m_{тi}$	$(m_{эi} - m_{тi})^2$	$\frac{(m_{эi} - m_{тi})^2}{m_{тi}}$
0	4	3	1	$\frac{1}{15} = 0{,}067$
1	10	12		
2	18	19	1	$\frac{1}{19} = 0{,}053$
3	16	13	1	$\frac{4}{16} = 0{,}250$
4	2	3		
$\sum$	50	50	-	$\chi^2_Н = 0{,}370$

Пример 3.7. Число рабочих, не выполнивших сменного задания в 100 выборках по 20 рабочих, приводится в таблице:

Число рабочих x_i	0	1	2	3	4
Число выборок m_i	85	11	3	1	0

На уровне значимости 0,05 проверить гипотезу о законе Пуассона.

Решение. Определяем среднюю интенсивность числа рабочих, не выполнивших сменного задания, на одну выборку:

$$\overline{\lambda} = \frac{\sum_{i=0}^{к} m_i x_i}{\sum_{i=1}^{к} m_i} = \frac{0 \cdot 85 + 1 \cdot 11 + 3 \cdot 2 + 1 \cdot 3 + 4 \cdot 0}{100} = \frac{20}{100} = 0{,}2$$

По таблице $e^{-\overline{\lambda}}$ определяем $e^{-0,2} = 0{,}8187$.

По формуле (3.35) вычисляем вероятности:

$$P_0 = \frac{\overline{\lambda}}{0!} e^{-\overline{\lambda}} = \frac{1}{1} \cdot 0{,}8187 = 0{,}8187;$$

$$P_1 = \frac{\overline{\lambda}^1}{1!} e^{-\overline{\lambda}} = 0{,}2 \cdot 0{,}8187 = 0{,}1637;$$

$$P_2 = \frac{0{,}2^2}{2!} \cdot 0{,}8187 = 0{,}02 \cdot 0{,}8187 = 0{,}0164;$$

$$P_3 = \frac{0{,}2^3}{3!} \cdot 0{,}8187 = \frac{0{,}008}{6} \cdot 0{,}8187 = 0{,}0011.$$

Вычисляем наблюдаемое значение критерия:

x_i	$m_{эi}$	$m_{тi}$	$(m_{эi} - m_{тi})^2$	$\frac{(m_{эi} - m_{тi})^2}{m_{тi}}$
0	85	82	9	$\frac{9}{82} = 0{,}108$
1	11	16	25	$\frac{25}{16} = 1{,}563$
2	3	2	4	$\frac{4}{2} = 2{,}0$
3	1	0		
$\sum$	100	100	-	$\chi^2_Н = 3{,}671$

По таблице χ^2 - распределения на уровне значимости 0,05 и при числе степеней свободы $\nu = 1 - 2 = 3 - 2 = 1$ определяем $\chi^2_{кр} = 12{,}706$.

Так как $\chi^2_Н (= 3{,}671) < \chi^2_{кр} (= 12{,}706)$, нулевая гипотеза H_0 не отвергается, т.е. число рабочих, не выполнивших сменного задания, подчиняется закону Пуассона.

Таблица 3.1

Основные формулы, используемые при проверке гипотез о значении параметров распределений

№ пп	H_0	Условия проверки	Используемое распределение	Формулы для вычисления наблюдаемого значения параметров	H_1	Порядок определения критического значения критериев	Правила проверки
1	2	3	4	5	6	7	8
1	$\mu=\mu_0$	σ^2 известна	$\Phi(t)$	$t_H = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n}$	$\mu_1<\mu_0;\ \mu_1>\mu_0$	$(1-2\alpha)\to t_{кр}$	$\lvert t_H\rvert > t_{кр} \to H_0$ отвергается с вероятностью ошибки α
					$\mu_1\neq\mu_0$	$(1-\alpha)\to t_{кр}$	
		σ^2 не известна	$S(t)$	$t_H = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n-1}$	$\mu_1<\mu_0;\ \mu_1>\mu_0$	$\left.\begin{matrix}2\alpha\\ \nu=n-1\end{matrix}\right\}\to t_{кр}$	
					$\mu_1\neq\mu_0$	$\left.\begin{matrix}\alpha\\ \nu=n-1\end{matrix}\right\}\to t_{кр}$	
2	$\mu_X=\mu_Y$	σ_1^2 и σ_2^2 известны	$\Phi(t)$	$t_H = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$\mu_x<\mu_y;\ \mu_x>\mu_y$	$(1-2\alpha)\to t_{кр}$	$\lvert t_H\rvert \le t_{кр} \to H_0$ не отвергается
					$\mu_x\neq\mu_y$	$(1-\alpha)\to t_{кр}$	
		σ_1^2 и σ_2^2 не известны, но $\sigma_1^2=\sigma_2^2$	$S(t)$	$t_H = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}}\sqrt{\frac{n_1\cdot n_2}{n_1+n_2}}$	$\mu_x<\mu_y;\ \mu_x>\mu_y$	$\left.\begin{matrix}2\alpha\\ \nu=n_1+n_2-2\end{matrix}\right\}\to t_{кр}$	
					$\mu_x\neq\mu_y$	$\left.\begin{matrix}\alpha\\ \nu=n_1+n_2-2\end{matrix}\right\}\to t_{кр}$	
					$\sigma_1^2<\sigma_0^2$	$\left.\begin{matrix}1-\alpha\\ \nu=n-1\end{matrix}\right\}\to \chi^2_{кр}$	$U_H^2 \ge \chi^2_{кр} \to H_0$ не отвергается

3	$\sigma_1^2 = \sigma_0^2$		χ^2	$U_H^2 = \dfrac{nS^2}{\sigma_0^1}$	$\sigma_1^2 \neq \sigma_0^2$	$\left.\begin{matrix} 1-\dfrac{\alpha}{2} \\ \nu = n-1 \end{matrix}\right\} \to \chi^2_{кр.лев}$ $\left.\begin{matrix} \dfrac{\alpha}{2} \\ \nu = n-1 \end{matrix}\right\} \to \chi^2_{кр.пр.}$	$\chi^2_{кр.лев} \le U^2_H \le \chi^2_{кр.пр.}$ $\to H_0$ не отвергается При $U^2_H \le \chi^2_{кр.лев}$ или $U^2_H > \chi^2_{кр.пр.} \to$ $\to H_0$ отвергается
					$\sigma_1^2 > \sigma_0^2$	$\left.\begin{matrix} \alpha \\ \nu = n-1 \end{matrix}\right\} \to \chi^2_{кр}$	$U^2_H \le \chi^2_{кр} \to H_0$ не отвергается
4	$\sigma_1^2 = \sigma_2^2$	$S_1^2 > S_2^2$	F	$F_H = \dfrac{\hat{S}_1^2}{\hat{S}_2^2}$	$\sigma_1^2 > \sigma_2^2$	$\left.\begin{matrix} \alpha \\ \nu_1 = n_1 - 1 \\ \nu_2 = n_2 - 1 \end{matrix}\right\} \to F_{кр}$	$F_H \le F_{кр} \to H_0$ не отвергается
5	$\sigma_1^2 = \sigma_2^2 =$	$n_1 \neq n_2 \neq ...$ $... \neq n_l$ $n_i > 4$	χ^2	$U_H^2 = \dfrac{\gamma \ln \hat{S}^2_{-p} - \sum_{i=1}^{\ell} \gamma_i \ln \hat{S}_i^2}{1 + \dfrac{1}{3(\ell-1)}\left[\sum_{i=1}^{\ell} \dfrac{1}{\gamma_i} - \dfrac{1}{\gamma}\right.}$	$\sigma^2 / H_1 > \sigma^2_{max}$	$\left.\begin{matrix} \alpha \\ l-1 \end{matrix}\right\} \to \chi^2_{кр}$	$U^2_H \le \chi^2_{кр} \to H_0$ не отвергается
	$... = \sigma_l^2$	$n_1 = n_2 = ...$ $... = n_l$	G	$G_H = \dfrac{\hat{S}^2_{max}}{\sum_{i=1}^{\ell} \hat{S}_i^2}$	$\sigma^2 / H_1 > \sigma^2_{max}$	$\left.\begin{matrix} \alpha \\ n-1 \\ l \end{matrix}\right\} \to G_{кр}$	$G_H \le G_{кр} \to H_0$ не отвергается
6	$p_1 = p_2 =$ $... = p_l$	$n \to \infty$	χ^2	$U_H^2 = \dfrac{1}{\tilde{p}(1-\tilde{p})} \sum_{i=1}^{\ell} (\tilde{p} - \tilde{p}_i)^2 n$	$P / H_1 > P_{max}$	$\left.\begin{matrix} \alpha \\ l-1 \end{matrix}\right\} \to \chi^2_{кр}$	$U^2_H \le \chi^2_{кр} \to H_0$ не отвергается

Тест

1. Что называют ошибкой первого рода:

а) Гипотеза H_0 верна и ее принимают согласно критерию;

б) Гипотеза H_0 верна и ее отвергают согласно критерию;

в) Гипотеза H_0 не верна и ее отвергают согласно критерию;

г) Гипотеза H_0 не верна и ее принимают согласно критерию;

2. Что называют мощностью критерия:

а) Вероятность, с которой статистика критерия должна попасть в критическую область, если верна гипотеза H_0;

б) Вероятность, с которой статистика критерия должна попасть в критическую область, если верна гипотеза H_1;

в) Вероятность, с которой статистика критерия должна попасть в область принятия гипотезы, если верна гипотеза H_0;

г) Вероятность, с которой статистика критерия должна попасть в область принятия гипотезы, если верна гипотеза H_1;

3. Когда при проверке гипотезы $H_0 : \mu = \mu_0$ против $H_1 : \mu = \mu_1$ следует выбрать правостороннюю критическую область

а) $H_1 : \mu_1 < \mu_0$;

б) $H_1 : \mu_1 > \mu_0$;

в) $H_1 : \mu_1 \neq \mu_0$,

г) $H_1 : \mu_1 = \mu_0$,

4. Пусть статистика критерия θ_n^* имеет нормальное распределение. Какое условие является исходным для расчета значения $\theta_{кр}$ границы правосторонней критической области.

а) $P\left(\theta_n^* < \theta_{кр}\right) = \alpha$;

б) $P\left(\left|\theta_n^*\right| > \theta_{кр}\right) = \dfrac{\alpha}{2}$;

в) $P\left(\theta_n^* > \theta_{кр}\right) = \alpha$;

г) $P\left(\left|\theta_n^*\right| < \theta_{кр}\right) = \dfrac{\alpha}{2}$.

5. Какая статистика используется при проверке гипотезы H_0: $\sigma^2 = \sigma_0^2$:

а) $t = \dfrac{\overline{x} - \mu}{\sigma}\sqrt{n}$;

б) $t = \dfrac{\overline{x} - \mu}{S}\sqrt{n-1}$;

в) $\chi^2 = \dfrac{nS^2}{\sigma^2}$;

г) $F = \dfrac{\hat{S}_1^2}{\hat{S}_2^2}$.

6. Для проверки какой гипотезы используется статистика $\frac{\overline{x}-\mu}{S}\sqrt{n-1}$:

а) H_0: $\sigma_1^2 = \sigma_2^2$;
б) H_0: $\mu = \mu_0$;
в) H_0: $\sigma^2 = \sigma_0^2$
г) H_0: $\mu_1 = \mu_2$.

7. Чему равна граница критической области при проверке на уровне значимости α = 0,0344, гипотезы H_0 : μ = 50, если H_1: μ = 52, σ =3:

à) 2,15;
б) 1,97;
в) 1,82;
г) 2,88.

8. Чему равна граница критической области при проверке на уровне значимости α = 0,05, гипотезы H_0 : μ = 50, если H_1: μ_1 = 52, S = 3, n =9:

à) 1,78;
б) 1,86;
в) 2,15;
г) 2,88.

9. Для проверки какой гипотезы применяется критерий Кохрана

а) H_0: $\sigma^2 = \sigma_0^2$
б) H_0: $\mu_1 = \mu_2$;
в) H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$, если $n_1 = n_2 = \ldots = n_l$;
г) H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_l^2$, если $n_1 \neq n_2 \neq \ldots \neq n_l$.

10. Какому закону распределения должна подчиняться статистика $\frac{\overline{x}-\mu}{S}\sqrt{n-1}$, при справедливости гипотезы H_0:

а) Нормальному;
б) Фишера-Снедекора;
в) Стьюдента;
г) Пирсона.

4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

4.1. Задачи и проблемы корреляционного анализа

Главной задачей корреляционного анализа является оценка взаимосвязи между переменными величинами на основе выборочных данных.

Различают два вида зависимостей между экономическими явлениями: функциональную и стохастическую. При функциональной зависимости имеет место однозначность отображения множества значений изучаемых величин, т.е. существует

правило y=f(x) - соответствия независимой переменной x и зависимой переменной y. В экономике примером функциональной связи может служить зависимость производительности труда от объема произведенной продукции и затрат рабочего времени.

При изучении массовых явлений зависимость между наблюдаемыми величинами проявляется часто лишь в случае, когда число единиц изучаемой совокупности достаточно велико. При этом каждому фиксированному значению аргумента соответствует определенный закон распределения значений функции и, наоборот, заданному значению зависимой переменной соответствует закон распределения объясняющий переменной. Например, при изучении потребления электроэнергии y в зависимости от объема производства x каждому значению x соответствует множество значений y и наоборот. В этом случае можно констатировать наличие стохастической (корреляционной) связи между переменными.

Множественность результатов при анализе связи x и y объясняется прежде всего тем, что зависимая переменная y испытывает влияние не только фактора x, но и целого ряда других факторов, которые не учитываются. Кроме того, влияние выделенного фактора может быть не прямым, а проявляется через цепочку других факторов.

При изучении **корреляционной зависимости** между переменными возникают следующие задачи:

1. Измерение силы (тесноты) связи.
2. Отбор факторов, оказывающих наиболее существенное влияние на результат признак.
3. Обнаружение неизвестных причин связей.
4. Построение корреляционной модели и оценка ее параметров.
5. Проверка значимости параметров связи.
6. Интервальное оценивание параметров связи.

Пусть из генеральной совокупности, которую образуют “к” признаков, являющихся случайными величинами, сделана выборка объемом n, тогда выборка будет представлять собой n независимо наблюдаемых к-мерных точек (векторов): $(x_{i1}, x_{i2},\ldots,x_{ij},\ldots x_{ik})$, где $i=1\div n$, а каждая координата x_{ij} наблюдаемой точки является вариантом соответствующего признака x_j $(j=1\div k)$ генеральной совокупности, изучаемой с точки зрения взаимозависимости к признаков.

В настоящее время при построении корреляционных моделей исходят из условия нормальности многомерного закона распределения генеральной совокупности. Эти условия обеспечивают линейный характер связи между изучаемыми признаками, что делает правомерным использование в качестве показателей тесноты связи: парного, частного и множественного коэффициентов корреляции.

На практике не всегда строго соблюдаются предпосылки корреляционного анализа: один из признаков оказывается величиной не случайной или признаки не имеют совместного нормального распределение. Для изучения связи между признаками в этом случае существует общий показатель зависимости признаков, который называется корреляционным отношением.

В практике статистического анализа возможны случаи, когда с помощью корреляционных моделей обнаруживают достаточно сильную “зависимость” признаков, в действительности не имеющих причинной связи друг с другом. Такие корреляции называют ложными.

4.2. Двумерная корреляционная модель

Рассмотрим случай изучения корреляционной зависимости между двумя признаками Y и X. Построение двумерной корреляционной модели предполагает, что закон распределения двумерной случайной величины в генеральной совокупности является нормальным, а выборка репрезентативной.

Плотность двумерного нормального закона распределения определяется пятью параметрам:

$MX=\mu_x$ - математическое ожидание X;

$MY=\mu_y$ - математическое ожидание Y;

$DX=\sigma^2_x$ - дисперсия X;

$DY=\sigma^2_y$ - дисперсия Y;

$\rho = M\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y-\mu_y}{\sigma_y}\right]$ - парный коэффициент корреляции, характеризует тесноту линейной связи между величинами X и Y.

В двумерной корреляционной модели используется так же, как мера тесноты связи, ρ^2 - коэффициент детерминации, указывающий долю дисперсии одной случайной величины, обусловленную вариацией другой.

Для получения точечных оценок параметров двумерной корреляционной модели обычно используют метод моментов, т.е. в качестве точечных оценок неизвестных начальных моментов первого и второго порядков генеральной совокупности берутся соответствующие выборочные моменты, и расчеты производят в соответствии со следующими формулами:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - оценка для μ_x;

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - оценка для μ_y;

$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$ - оценка для $M(x^2)$;

$\overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2}{n}$ - оценка для $M(y^2)$;

$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ - оценка для M(xy);

$S_x^2 = \overline{x^2} - \left(\bar{x}\right)^2$ - оценка для σ_x^2;

$S_y^2 = \overline{y^2} - \left(\bar{y}\right)^2$ - оценка для σ_y^2;

$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x \cdot S_y}$ - оценка для ρ.

Полученные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ также обладают свойствами несмещенности и эффективности. Следует отметить, что в

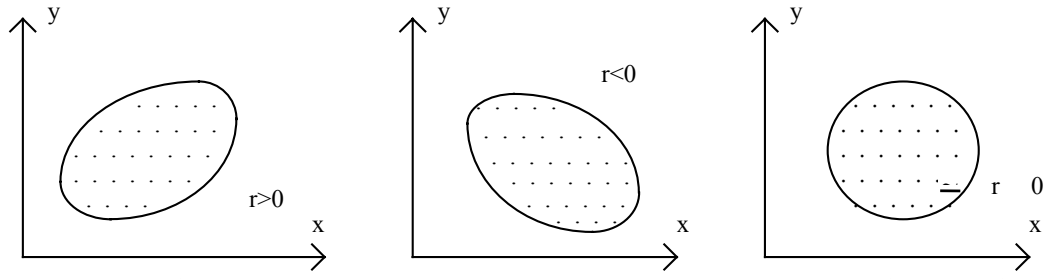
корреляционной модели распределение выборочных средних $\bar{x}$ и $\bar{y}$ не зависит от законов распределения S^2_x, S^2_y, r.

Парный коэффициент корреляции ρ в силу своих свойств является одним из самых распространенных способов измерения связи между случайными величинами в генеральной совокупности; для выборочных данных используется эмпирическая мера связи r.

Коэффициент корреляции не имеет размерности и, следовательно, его можно сопоставлять для разных статистических рядов. Величина его лежит в пределах (-1 до +1). Значение ρ=±1 свидетельствует о наличии функциональной зависимости между рассматриваемыми признаками. Если ρ=0, можно сделать вывод, что линейная связь между x и y отсутствует, однако это не означает, что они статистически независимы. В этом случае не отрицается возможность существования иной формы зависимости между переменными. Положительный знак коэффициента корреляции указывает на положительную корреляцию, т.е. все данные наблюдения лежат на прямой с положительным углом наклона в плоскости xy и с увеличением x растет y. Когда x уменьшается, то y уменьшается. Отрицательный знак коэффициента свидетельствует об отрицательной корреляции. Чем ближе значение |r| к единице, тем связь теснее, приближение |r| к нулю означает ослабление линейной зависимости между переменными. При |r|=1 корреляционная связь перерождается в функциональную.

На практике при изучении зависимости между двумя случайными величинами используют поле корреляции, с помощью которого при минимальных затратах труда и времени можно установить наличие корреляционной зависимости.

Поле корреляции представляет собой диаграмму, на которой изображается совокупность значений двух признаков. Каждая точка этой диаграммы имеет координаты (x_i, y_i), соответствующие размерам признаков в i-м наблюдении. Три варианта распределения точек на поле корреляции показаны на рисунках 1.4.1; 1.4.2; 1.4.3. На первом из них основная масса точек укладывается в эллипсе, главная диагональ которого образует положительный угол с осью X. Это график положительной корреляции. Второй вариант распределения соответствует отрицательной корреляции. Равномерное распределение точек в пространстве (XY) свидетельствует об отсутствии корреляционной зависимости.(рис. 1.4.3.)



Если наблюдаемые значения Y и X представляет собой выборку из двумерного нормального распределения, то формально можно рассматривать два уравнения регрессии:

$$M(Y/X) = \beta_0 + \beta_1 x \quad \text{и} \quad M(X/Y) = \alpha_0 + \alpha_1 y$$

В двумерном корреляционном анализе, обычно строят корреляционную таблицу, поле корреляции, рассчитывают точечные оценки параметров корреляционной модели, оценивают уравнения регрессии, проверяют значимость параметров связи и для значимых параметров строят интервальные оценки, не разделяя при этом задачи корреляционного и регрессионного анализа.

Имея оценки параметров модели $\bar{x}$, $\bar{y}$, S_x, S^2_y, r, можно рассчитать оценки уравнений регрессии в соответствии с формулой для генеральной регрессии

$$M(y/x) - M(y) = \beta_{yx}[x - M(x)],$$

где $\beta_{yx} = \rho\frac{\sigma_y}{\sigma_x}$ - коэффициент регрессии y на x, оценка здесь $\hat{y}/x - \bar{y} = b_{yx}(x - \bar{x})$,

где $b_{yx} = r\frac{S_y}{S_x}$ - оценка генерального коэффициента регрессии β_{yx}.

Аналогичные формулы расчета справедливы для оценки уравнения регрессии x на y:

$M(x/y) - M(x) = \beta_{xy}[y - M(y)]$ - генеральная регрессия x на y,

где $\beta_{xy} = \rho\frac{\sigma_x}{\sigma_y}$ - коэффициент регрессии x на y,

$\hat{x}/y - \bar{x} = b_{xy}(y - \bar{y})$ - оценка генерального коэффициента регрессии β_{xy}.

Можно показать, что формулы $M(y/x) - M(y) = \beta_{yx}[x - M(x)]$ и $M(y/x) = \beta_0 + \beta_1 x$ идентичны. Из формулы

$$M(y/x) = \beta_{yx}x - \beta_{yx}M(x) + M(y)$$

полагая, что $\beta_{yx}=\beta_1$, а $-\beta_{yx}M(ч)+M(y)= \beta_0$ запишем: $M(y/x) = \beta_0 + \beta_1 x$

Аналогично можно показать идентичность формул попарно:

$M(x/y) - M(x) = \beta_{xy}[y - M(y)]$ и $M(x/y) = \alpha_0 + \alpha_1 y$;

$\hat{y}/x - \bar{y} = b_{yx}(x - \bar{x})$ и $\hat{y} = b_0 + b_1 x$;

$\hat{x}/y - \bar{x} = b_{xy}(y - \bar{y})$ и $\hat{x} = a_0 + a_1 y$

В двумерной модели параметрами связи являются коэффициент корреляции ρ (или коэффициент детерминации ρ^2) и коэффициенты регрессии β_{yx}, β_{xy}, которые обычно бывают неизвестны.

По результатам выборки рассчитывают их точечные оценки, соответственно r, b_y, b_x, проверяют гипотезу о значимости (существенности) параметров. В двумерной модели достаточно проверить значимость только коэффициента корреляции. Проверяется гипотеза H_0: $\rho=0$. Если на уровне значимости α гипотеза отвергнется, то коэффициент корреляции считается значимым и рассчитанное по выборке значение r может быть использовано в качестве его точечной оценки. Если коэффициент корреляции оказывается незначимым, то гипотеза не отвергается и на практике обычно принимают, что x и y в генеральной совокупности линейно независимым.

Доказано, что если верна гипотеза H_0: $\rho=0$, то статистика $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ имеет распределение Стьюдента с $\nu=n-2$ числом степеней свободы. По таблице распределения Стьюдента были определены значения статистики $t_{табл}(\alpha;\nu=n-2)$ для α=0,001; 0,01; 0,02; 0,05 и рассчитаны соответственно границы для r (таблицы Фишера-Иейтса). Таким образом, для проверки гипотезы H_0: $\rho=0$ находят $r_{табл}(\alpha, \nu=n-2)$ и сравнивают его с $r_{набл}$, рассчитанным по выборочным данным. Если $|r_{набл.}| \geq r_{табл}$, то гипотеза H_0 отвергается на уровне значимости α, если $|r_{набл.}| \leq r_{табл}$, то гипотеза не отвергается.

При n>100, считая распределение статистики нормированным нормальным, проверяют гипотезу H_0: $\rho=0$ исходя из условия, что при справедливой гипотезе выполняется равенство

$P(|t|\le t_{табл})=\gamma=\Phi(t_{табл})$, т.е. если $|t|\le t_{табл}$, то гипотеза H_0 не отвергается. Статистика $r\sqrt{n-1}$, если n>100, также имеет нормированный нормальный закон распределения при справедливости H_0: $\rho=0$ и этим можно пользоваться для проверки значимости коэффициента корреляции.

Для двумерной корреляционной модели, если отвергается гипотеза H_0: $\rho=0$, то параметры связи ρ, β_{yx}, β_{xy} считаются значимыми и для них имеет смысл найти интервальные оценки, для чего нужно знать закон распределения выборочных оценок параметров.

Плотность вероятности выборочного коэффициента корреляции имеет сложный вид, поэтому используют специально подобранные функции от выборочного коэффициента корреляции, которые подчиняются хорошо изученным законам, например нормальному или Стьюдента.

При нахождении доверительного интервала для коэффициента корреляции ρ чаще используют преобразование Фишера:

$$z=\frac{1}{2}\ln\frac{1+r}{1-r}$$

Эта статистика уже при n>10 распределена приблизительно нормально, с параметрами $M(z)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)}$.

По таблице z - преобразования Фишера для выборочного коэффициента r находят соответствующее ему z_r и находят интервальную оценку для M(z) из условия:

$$P\left(z_r-t_\gamma\sqrt{\frac{1}{n-3}}\le M(z)\le z_r+t_\gamma\sqrt{\frac{1}{n-3}}\right)=\gamma=\Phi\left(t_\gamma\right)$$

где t_γ находят по таблице интегральной функции Лапласа

$$\Phi(t)=\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt$$ для данного $\gamma=1-\alpha$.

Получив доверительный интервал: $z_{min}\le M(z)\le z_{max}$, с помощью таблицы z - преобразования Фишера получают интервальную оценку: $r_{min}\le\rho\le r_{max}$, где r_{min} и r_{max} выбираются с учетом того, что z - функция нечетная, а поправочным членом $\frac{\rho}{2(n-1)}$ пренебрегают.

Для значимых коэффициентов регрессии β_{yx} и β_{xy} с надежностью $\gamma=1-\alpha$. Находят интервальные оценки из условия, что статистики

$$t=\left(b_{yx}-\beta_{yx}\right)\frac{S_x\sqrt{n-2}}{S_y\sqrt{1-r^2}};$$

$$t=\left(b_{xy}-\beta_{xy}\right)\frac{S_y\sqrt{n-2}}{S_x\sqrt{1-r^2}}$$

имеют распределение Стьюдента с ν=n-2 степенями свободы и, следовательно, из условия P(|t|<=t_α)=γ можно рассчитать интервальные оценки

$$b_{yx} - t_\alpha \frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}} \le \beta_{yx} \le b_{yx} + t_\alpha \frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}};$$

$$b_{xy} - t_\alpha \frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}} \le \beta_{xy} \le b_{xy} + t_\alpha \frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}}$$

где t_α определяется по таблице Стьюдента для данного α=1-γ и ν=n-2.

Пример 4.1. На основании выборочных данных о производительности труда (x) и себестоимости продукции (y), полученных с однотипных предприятий за месяц и представленных в таблице 4.1, найти: а) точную оценку коэффициента корреляции между x и y, проверить его значимость при α=0,05 и найти интервальную оценку коэффициента корреляции при γ=0,95; б) оценку уравнения регрессии, характеризующего зависимость себестоимости продукции от производительности труда.

Таблица 4.1

производительность труда x	5	4	3	20	10	15
себестоимость продукции y	7	10	12	2	5	4

Решение

Составим вспомогательную таблицу 4.2

Таблица 4.2

	x_i	y_i	x_iy_i	x_i^2	y_i^2
	5	7	35	25	49
	4	10	40	16	100
	3	12	36	9	144
	20	2	40	400	4
	10	5	50	100	25
	15	4	60	225	9
Σ	57	40	261	775	331

а) Выборочный парный коэффициент корреляции рассчитывается по формуле

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x S_y} = \frac{43{,}5 - 9{,}5 \cdot 6{,}67}{6{,}24 \cdot 3{,}27} = -0{,}97$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}57 = 9{,}5$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}40 = 6{,}67$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{6}261 = 43{,}5$$

$$S_x = \sqrt{S_x^2} = \sqrt{\overline{x^2} - \left(\overline{x}\right)^2} = \sqrt{129{,}17 - 90{,}25} = \sqrt{38{,}92} = 6{,}24$$

$$S_y = \sqrt{S_y^2} = \sqrt{\overline{y^2} - \left(\overline{y}\right)^2} = \sqrt{55{,}17 - 44{,}49} = \sqrt{10{,}68} = 3{,}27$$

Для проверки значимости коэффициента корреляции сформулируем статистическую гипотезу H_0: $\rho=0$. По таблице Фишера-Йейтса находим $r_{табл}$ ($\alpha=0{,}05$; $\upsilon=n-2=4$)=0,811. Сравнение $|r_{набл}|=0{,}97$ с $r_{табл}=0{,}811$ свидетельствует о том , что нулевая гипотеза отвергается и, следовательно, коэффициент корреляции ρ значим.

Интервальную оценку для ρ рассчитаем с помощью z - преобразований Фишера. По таблице значений статистики $z = \frac{1}{2}\ln\frac{1+r}{1-r}$ находим $z_r=0{,}97$. Из условия, что $\gamma=\Phi(t_\gamma)=0{,}95$, находим по таблице интегральной функции Лапласа $t_\gamma=1{,}96$. Тогда интегральная оценка для MZ_r определяется:

$$-2{,}0923 - 1{,}96\sqrt{\frac{1}{6}} \le MZ_r \le -2{,}0925 + 1{,}96\sqrt{\frac{1}{6}}$$

$$-2{,}8925 \le MZ_r \le -1{,}2922$$

Воспользовавшись таблицей z - преобразования Фишера, перейдем от z к ρ и найдем интегральную оценку с надежностью $\gamma=0{,}95$:

$$-0{,}994 \le \rho \le -0{,}86$$

б) Для нахождения оценок уравнения регрессии себестоимости продукции от производительности труда $\hat{y}=b_0+b_1x$, воспользуемся формулой

$$b_1 = b_{yx} = r\frac{S_y}{S_x} = -0{,}97\frac{3{,}27}{6{,}24} = -0{,}51$$

Тогда используя $\hat{y} - \overline{y} = b_1\left(x - \overline{x}\right)$, находим

$$\hat{y} = 11{,}515 - 0{,}51x$$

4.3. Трехмерная корреляционная модель

На примере трехмерной генеральной совокупности достаточно четко можно продемонстрировать основные задачи и особенности многомерного корреляционного анализа.

Пусть признаки X, Y, Z образуют трехмерную нормально распределенную генеральную совокупность, которая определяется девятью параметрами:

- тремя математическими ожиданиями

$$MX = \mu_x \quad MY = \mu_y \quad MZ = \mu_z \qquad (4.9.)$$

- тремя дисперсиями

$$DX = \sigma_x^2 \quad DY = \sigma_y^2 \quad DZ = \sigma_z^2 \qquad (4.10.)$$

- тремя парными коэффициентами корреляции

$$\rho_{xy} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y}\right]; \quad \rho_{xz} = M\left[\frac{x-\mu_x}{\sigma_x}\cdot\frac{z-\mu_z}{\sigma_z}\right]; \quad \rho_{yz} = M\left[\frac{y-\mu_y}{\sigma_y}\cdot\frac{z-\mu_z}{\sigma_z}\right]$$

Следует отметить, что частные одномерные (X, Y, Z) и двумерные [(X, Y), (X,Z), (Y, Z)] распределения компонент, а так же условные распределения при

фиксированных одной [(X,Y)/Z; (X,Z)/Y; (Y,Z)/X] и двух [X/(Y,Z); Y/(X,Z); z/(X,Y)] переменных являются нормальными. Поэтому поверхности и линии регрессии являются плоскостями и прямыми соответственно.

Для изучения разнообразия связей между тремя случайными величинами рассчитывают не только парные, но частные и множественные коэффициенты корреляции (детерминации)

Частные коэффициенты корреляции между двумя случайными величинами при фиксированной третьей (в силу их независимости от фиксированных переменных) характеризуют тесноту связи между этими двумя величинами при исключении из рассмотрения фиксированной третьей величины. Поэтому, если парный коэффициент корреляции между теми же двумя случайными величинами оказался больше соответствующего частного коэффициента, то можно сделать вывод о том, что третья фиксированная величина усиливает взаимосвязь между изучаемыми величинами, т.е. более высокое значение парного коэффициента обусловлено присутствием третьей величины. Более низкое значение парного коэффициента корреляции в сравнении с соответствующими частными свидетельствует об ослаблении связи между изучаемыми величинами действием фиксируемой величины.

Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции, т.к. он является коэффициентом корреляции условного двумерного распределения.

Для трехмерной модели можно рассчитать три частных коэффициента корреляции:

$$\rho_{xy/z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{\left(1-\rho_{xz}^2\right)\cdot\left(1-\rho_{yz}^2\right)}};$$

$$\rho_{xz/y} = \frac{\rho_{xz} - \rho_{xy} \cdot \rho_{zy}}{\sqrt{\left(1-\rho_{xy}^2\right)\cdot\left(1-\rho_{zy}^2\right)}}; \qquad (4.11.)$$

$$\rho_{yz/x} = \frac{\rho_{yz} - \rho_{yx} \cdot \rho_{zx}}{\sqrt{\left(1-\rho_{yx}^2\right)\cdot\left(1-\rho_{zy}^2\right)}}.$$

Множественный коэффициент корреляции в трехмерной нормальной совокупности служит мерой связи между одной случайной величиной и совместным действием двух остальных. Для трехмерной корреляционной модели можно рассчитать три множественных коэффициента корреляции:

$$R_x = R_{x/yz} = \sqrt{\frac{\rho_{xy}^2 + \rho_{xz}^2 - 2\rho_{xy}\rho_{xz}\rho_{yz}}{1-\rho_{yz}^2}};$$

$$R_y = R_{y/xz} = \sqrt{\frac{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}}{1-\rho_{xz}^2}}; \qquad (4.12.)$$

$$R_z = R_{z/xy} = \sqrt{\frac{\rho_{zx}^2 + \rho_{zy}^2 - 2\rho_{zx}\rho_{zy}\rho_{xy}}{1-\rho_{xy}^2}}.$$

По величине множественный коэффициент корреляции заключен между нулем и единицей. Если $R_x=1$, то связь между величинами X и (Y, Z) является функциональной, линейной: точки (x, y, z) расположены в плоскости регрессии X на (Y, Z). Если $R_x=0$, то одномерная случайная величина X и двумерная случайная величина (Y, Z) являются независимыми (в силу нормальности распределения). Множественный коэффициент детерминации R_x^2 показывает долю дисперсии случайной величины X, обусловленную изменением величины (Y, Z).

Множественный коэффициент корреляции может увеличиваться при введении в модель дополнительных признаков и не увеличится при исключении некоторых признаков из модели. Наибольшему множественному коэффициенту детерминации соответствует большие частные коэффициенты детерминации. Например, если $R_x^2>R_z^2$ и $R_x^2>R_y^2$, то

$$\rho_{xz/y} > \rho_{zy/x}$$
$$\rho^2_{xy/z} > \rho^2_{zy/x}$$

При фиксировании одной случайной величины трехмерное нормальное распределение превращается в двумерное нормальное распределение, определяемое пятью параметрами. Если фиксирована случайная величина z, то двумерное нормальное распределение (X,Y/Z) характеризуется следующими параметрами:

$$\mu_{x/z} = \mu_x + \rho_{zx}\frac{\sigma_x}{\sigma_z}\left(z-\mu_z\right)$$

$$\mu_{y/z} = \mu_z + \rho_{zy}\frac{\sigma_y}{\sigma_x}\left(z-\mu_z\right)$$

$$\sigma^2_{x/z} = \sigma^2_x\left(1-\rho^2_{zx}\right) \qquad (4.13.)$$

$$\sigma^2_{y/z} = \sigma^2_y\left(1-\rho^2_{zy}\right)$$

$$\rho_{xy/z} = \frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\left(1-\rho^2_{xz}\right)\left(1-\rho^2_{yz}\right)}$$

Зависимость между величинами X и У при фиксированном значении случайной величины Z выражается прямыми регрессиями в плоскости Z=z:

$$\begin{aligned} &M\left(Y/X\right)/Z-\mu_{y/z} = \beta_{yx/z}\left(X-\mu_{x/z}\right);\\ &M\left(X/Y\right)/Z-\mu_{x/z} = \beta_{xy/z}\left(Y-\mu_{y/z}\right)\end{aligned} \qquad (4.14.)$$

Коэффициенты частной регрессии рассчитывают в соответствии с формулами:

$$\begin{aligned} &\beta_{yx/z} = \rho_{xy/z}\frac{\sigma_{y/z}}{\sigma_{x/z}} = \frac{\beta_{yx}-\beta_{yz}\beta_{zx}}{1-\beta_{xz}\beta_{zx}};\\ &\beta_{xy/z} = \rho_{xy/z}\frac{\sigma_{x/z}}{\sigma_{y/z}} = \frac{\beta_{xy}-\beta_{xz}\beta_{zy}}{1-\beta_{yz}\beta_{zy}},\end{aligned} \qquad (4.15.)$$

причем

$$\rho^2_{xy/z} = \beta_{xy/z}\beta_{yx/z};$$

для расчета условных средних квадратических отклонений используют формулы:

$$\begin{aligned} \sigma_{y/zx} &= \sigma_{y/z}\sqrt{1-\rho^2_{xy/z}} = \sigma_{y/x}\sqrt{1-\rho_{yz/x}{}^2}; \\ \sigma_{x/yz} &= \sigma_{x/z}\sqrt{1-\rho^2_{xy/z}} = \sigma_{x/y}\sqrt{1-\rho^2_{xz/y}}. \end{aligned} \qquad (4.16.)$$

Условное распределение при фиксировании величины (X, У) будет одномерным Z/(X,Y), которое характеризуется условным математическим ожиданием

$$M_z/(X,Y) = M(Z/X)/y = M(Z/Y)/x$$

и условной дисперсией $Dz/(X,Y) = \sigma^2_{z/xy}$

Плоскость регрессии Z на (X, У) будет получена при изменении точки (X, У):

$$MZ/(X,Y) - \mu_z = \beta_{zx/y}(X-\mu_x) + \beta_{zy/x}(Y-\mu_y)$$

Остаточная дисперсия относительно плоскости регрессии рассчитывается в соответствии с формулой

$$\sigma^2_{z/xy} = \sigma^2_{z/y}(1-\rho^2_{zx/y}) = \sigma^2_{z/x}(1-\rho^2_{yz/x})$$

Для оценки девяти параметров трехмерной корреляционной модели используют следующие формулы:

$$\mu_x \to \bar{x} = \frac{\sum x}{n};\ \sigma^2_x \to S^2_x = \overline{x^2} - (\bar{x})^2;\ \rho_{xy} \to r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x S_y};$$

$$\mu_y \to \bar{y} = \frac{\sum y}{n};\ \sigma^2_y \to S^2_y = \overline{y^2} - (\bar{y})^2;\ \rho_{xz} \to r_{xz} = \frac{\overline{xz} - \bar{x}\cdot\bar{z}}{S_x S_z};$$

$$\mu_z \to \bar{z} = \frac{\sum z}{n};\ \sigma^2_z \to S^2_z = \overline{z^2} - (\bar{z})^2;\ \rho_{yz} \to r_{yz} = \frac{\overline{xz} - \bar{x}\cdot\bar{z}}{S_x S_z};$$

Оценки условных средних квадратических отклонений при фиксировании одной переменной, частных коэффициентов корреляции, условных средних квадратических отклонений при двух фиксированных переменных и множественных коэффициентов корреляции рассчитываются в соответствии со следующими формулами:

$$S_{x/y} = S_x\sqrt{1-r_{xy}^2};\quad \mathbf{S_{x/z}} = S_x\sqrt{1-r_{xz}^2};\quad \mathbf{S_{y/z}} = S_y\sqrt{1-r_{yz}^2};$$

$$S_{y/x} = S_y\sqrt{1-r_{xy}^2};\quad \mathbf{S_{z/x}} = S_z\sqrt{1-r_{xz}^2};\quad \mathbf{S_{z/y}} = S_z\sqrt{1-r_{yz}^2};$$

$$r_{xy/z} = \frac{r_{xy} - r_{xz}\cdot r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}};$$

$$r_{xz/y} = \frac{r_{xz} - r_{xy}\cdot r_{yz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{yz}^2\right)}};$$

$$r_{yz/x} = \frac{r_{yz} - r_{xy}\cdot r_{xz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{xz}^2\right)}};$$

$$S_{x/yz} = S_{x/y}\sqrt{1-r_{xz/y}^2};\ \mathbf{S_{y/zx}} = S_{y/z}\sqrt{1-r_{yx/z}^2};\ \mathbf{S_{z/xy}} = S_{z/x}\sqrt{1-r_{yz/x}^2};$$

$$r_{x/yz}^2 = 1-\frac{S_{x/yz}^2}{S_x^2};\ \mathbf{r_{y/xz}^2} = 1-\frac{S_{y/xz}^2}{S_y^2};\ \mathbf{r_{z/xy}^2} = 1-\frac{S_{z/xy}^2}{S_z^2}$$

Для проверки значимости параметров связи трехмерной корреляционной модели формулируется нулевая гипотеза о равенстве нулю проверяемого параметра. Если на уровне значимости α гипотеза отвергается, то с надежностью $\gamma=1-\alpha$ можно утверждать, что параметр значимо отличается от нуля. Если же гипотеза принимается, то параметр связи незначим.

В трехмерном корреляционном анализе проверяется значимость только частных и множественных коэффициентов корреляции или коэффициентов детерминации. Коэффициенты регрессии одновременно равны нулю или отличны от нуля вместе с соответствующими коэффициентами корреляции (детерминации).

Проверка значимости парных коэффициентов корреляции для трехмерной модели обычно не проводится. Чтобы установить значимость частного коэффициента корреляции, необходимо на выбранном уровне значимости α проверить гипотезу H_0: $\rho_{частн}=0$. В основе критерия используемого для проверки этой гипотезы, лежит статистика

$$t = \frac{r_{частн}}{\sqrt{1-r_{частн}^2}}\sqrt{n-3}$$

которая при справедливости нулевой гипотезы подчиняется распределению Стьюдента с числом степеней свободы $\nu=n-3$. Для упрощения процедуры проверки значимости разработаны таблицы, где табулирован $r_{табл}$ (α, $\nu=n-3$) в соответствии с перечисленными условиями. Если $|r_{частн}|>r_{табл}(\alpha, \nu)$, то $\rho_{частн}$ считается значимым на уровне α. В противоположном случае , когда H_0: $\rho_{частн}=0$ не отвергнется, следует считать, что между соответствующими признаками связь отсутствует, либо провести анализ на основе другой выборки.

Основу критерия оценки значимости множественного коэффициента детерминации, а также и корреляции $r_{мн}$ составляет статистика

$$F_{\text{набл}} = \frac{r_{\text{мн}}^2 / 2}{\left(1 - r_{\text{мн}}^2\right) / \left(n - 3\right)},$$

которая при справедливости нулевой гипотезы H_o: $R^2=0$ имеет распределение Фишера. По таблице распределения Фишера определяют $F_{табл}(\alpha;\ \nu_1=2;\ \nu_2=n-3)$ и сравнивают с $F_{набл}$. Если $F_{набл} > F_{табл}$ то гипотеза отвергается и, следовательно, R^2 значимо отличается от нуля.

Осуществляя проверку значимости коэффициентов связи трехмерной корреляционной модели, следует учитывать , что если, например, R_z незначим, то коэффициенты $\rho_{zx/y}$ и $\rho_{zy/x}$ становятся незначимыми. Или, если $\rho_{zx/y}$ незначим, то множественный коэффициент корреляции незначимо отличается от абсолютной величины парного коэффициента корреляции $R_z=|\rho_{zy}|$.

Для значимых множественных коэффициентов корреляции можно получить оценки уравнения регрессии. Для значимого R_z оценкой соответствующего уравнения регрессии будет

$$\overline{z / (x, y)} - \bar{z} = b_{zx/y}\left(x - \bar{x}\right) + b_{xy/z}\left(y - \bar{y}\right),$$

где $b_{zx/y} = r_{zx/y}\dfrac{S_{z/y}}{S_{x/y}}$ и $b_{zy/x} = r_{zy/x}\dfrac{S_{z/x}}{S_{y/x}}$

частные коэффициенты регрессии.

Для значимых параметров связи имеет смысл определить границы доверительного интервала с надежностью $\gamma=1-\alpha$. Исходным равенством интервального оценивания $\rho_{частн}$ служит

$$P(r_{\text{частн min}} \le \rho_{\text{частн}} \le r_{\text{частн max}}) = \gamma$$

Для получения более точных значений доверительного интервала в данном случае используется z-преобразование Фишера, так как статистика $z = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$ имеет приблизительно нормальный закон распределения с параметрами $M_z \approx \dfrac{1}{2}\ln\dfrac{1+\rho}{1-\rho}$ и $D_z \cong \dfrac{1}{n-4}$. Поэтому первоначально определяют границы доверительного интервала для M_z исходя из равенства

$$P\left(Z_{\text{гчастн}} - t_\gamma \frac{1}{\sqrt{n-4}} \le M_z \le Z_{\text{частн}}(\quad) + t_\gamma \frac{1}{\sqrt{n-4}}\right) = \gamma = \Phi(t)$$

где $Z^{\text{гчастн}}$ определяется по таблице z - преобразования Фишера для рассчитанного $r_{частн}$.

t_γ находится по таблице интегральной функции Лапласа для заданного значения γ.

Зная границы интервальной оценки для M_z по таблице преобразования Фишера получают доверительные границы для $\rho_{частн.}$

Для значимых частных и множественных коэффициентов детерминации существуют более предпочтительные точечные оценки, чем выборочно коэффициенты:

$$\frac{(n-2)r^2_{\text{частн}} - 1}{n-3}$$ - оценка для $\rho^2_{\text{частн}}$;

$$\frac{(n-1)r_{\text{мн}}{}^2 - 2}{n-3}$$ - оценка для $\rho^2_{\text{мн}}$.

Интервальные оценки для коэффициента плоскости регрессии можно найти решением относительно оцениваемого коэффициента регрессии неравенства $|t| \le t(\alpha; \nu=n-3)$,
где

$$t = \frac{\left(b_{zx/y} - \beta_{zx/y}\right) S_{x/y}\sqrt{n-3}}{S_{z/y}\sqrt{1-r^2_{zx/y}}}$$

$$t = \frac{\left(b_{zy/x} - \beta_{zy/x}\right) S_{y/x}\sqrt{n-3}}{S_{z/x}\sqrt{1-r^2_{zy/x}}}$$

- статистики, подчиняющиеся распределению Стьюдента с числом степеней свободы $\nu=n-3$;
$t(\alpha; \nu=n-3)$ - определяют по таблице Стьюдента.

Пример 4.2
С целью изучения эффективности производства продукции была отобрана группа 25 однотипных предприятий. На основании полученной выборки для трех показателей (X - производительность труда, У - фондоотдача, Z - материалоемкость продукции) были вычислены величины:

$\bar{x}$=6,06 $\bar{y}$=2,052
$\bar{z}$=24,32 S_x=0,7782
r_{xy}=0,9016392 r_{xz}=-0,8770319 r_{yz}=-0,8899999 S_y=0,7925
S_z=3,7086

Требуется рассчитать оценки частных и множественных коэффициентов корреляции, проверить на уровне α=0,05 их значимость, для значимых частных коэффициентов корреляции рассчитать интервальные оценки с надежностью γ=0,95

Решение. Для расчета частных коэффициентов корреляции воспользуемся формулами

$$r_{xy/z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1-r^2_{xz}\right)\left(1-r^2_{yz}\right)}} = \frac{0{,}9016392 - 0{,}8770319 \cdot 0{,}8899999}{0{,}203078 \cdot 0{,}2079002} = 0{,}5526811;$$

$$r_{xz/y} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{\left(1-r^2_{xy}\right)\left(1-r^2_{yz}\right)}} = -0{,}3782736;$$

$$r_{yz/x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{\left(1-r^2_{xy}\right)\left(1-r^2_{xz}\right)}} = -0{,}4775473;$$

Множественные коэффициенты корреляции можно вычислить по формулам через парные коэффициенты, например

$$r_{x/yz} = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{\left(1 - r_{yz}^2\right)}} = 0{,}9163587, \text{ или через коэффициенты}$$

детерминации в соответствии с формулами:

$$r_{x/yz}^2 = 1 - \frac{S_{x/yz}^2}{S_x^2} = 1 - \frac{0{,}0970695}{0{,}6056} = 0{,}8397136;$$

$$r_{x/yz} = 0{,}9163588;$$

$$r_{y/xz} = 0{,}9249889;$$

$$r_{z/xy} = 0{,}9065585,$$

где $S_{x/yz}^2$ рассчитана в соответствии с формулами (1.28).

Проверку значимости множественных коэффициентов корреляции сделаем с помощью статистики:

$$F_{\text{набл (x)}} = \frac{r_{x/yz}^2 / 2}{\left(1 - r_{x/yz}^2\right)/\left(n - 3\right)} = \frac{0{,}8397136 \cdot 22}{2 \cdot 0{,}1602864} = 57{,}627157;$$

$$F_{\text{набл (y)}} = \frac{r_{y/xz}^2 / 2}{\left(1 - r_{y/zz}^2\right)/\left(n - 3\right)} = \frac{0{,}8556046 \cdot 22}{2 \cdot 0{,}1443954} = 65{,}1797121;$$

$$F_{\text{набл (z)}} = \frac{r_{z/xy}^2 / 2}{\left(1 - r_{z/xy}^2\right)/\left(n - 3\right)} = 50{,}745165$$

Сравнивая $F_{набл}$ с $F_{кр}(0{,}05;\ 2;\ 22)=3{,}44$, найденым по таблице распределения Фишера для $\alpha=0{,}05$, $\nu_1=2$; $\nu_2=n-3=22$, делаем вывод, что все множественные коэффициенты корреляции $r_{x/yz}$, $r_{y/xz}$, $r_{z/xy}$ генеральной совокупности значимо отличаются от нуля.

Для проверки значимости частных коэффициентов корреляции по таблице Фишера-Иейтса находим $r_{кр}(0{,}05;\ 25)=0{,}381$ и $r_{кр}(0{,}05;\ 20)=0{,}423$, тогда с помощью линейной интерполяции рассчитаем

$$r_{кр}(0{,}05;\ 22) = 0{,}381 + \frac{0{,}423 - 0{,}381}{25 - 20} \cdot \left(25 - 22\right) = 0{,}4062.$$

Так как наблюдаемые значения $|r_{xy/z}|$ и $|r_{yz/x}|$ больше, чем $r_{кр}(0{,}05;\ 22)$, то с вероятностью ошибки 0,05 гипотеза о равенстве нулю генеральных частных коэффициентов корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$ отвергается. Для частного коэффициента корреляции $\rho_{xz/y}$ гипотеза H: $\rho_{xz/y}=0$ не отвергается, т.к. $r_{xz/y}=0{,}381$ меньше $r_{кр}(0{,}05;\ 22)=0{,}4062$. Для значимых частных коэффициентов корреляции $\rho_{xy/z}$ и $\rho_{yz/x}$ с надежностью $\gamma=0{,}95$ найдем интервальные оценки с помощью z - преобразования

Фишера. По таблице значений статистики $z = \frac{1}{2}\ln\frac{1+r}{1-r}$ находим для $r_{xy/z}$=0,55 соответствующее ему z_r=0,6184, тогда

$$P\left(0{,}6184 - t_\gamma\sqrt{\frac{1}{n-4}} \le Mz_{yz/x} \le -0{,}523 + t_\gamma\sqrt{\frac{1}{n-4}}\right) = 0{,}95,$$

где t_γ=1,96 найдено по таблице значений интегральной функции Лапласа для $\Phi(t_\gamma)$=0,95;

$$\sqrt{\frac{1}{n-4}} = \sqrt{\frac{1}{21}} = 0{,}4277$$

следовательно,

$$1{,}1907 \le MZ_{xy/z} \le 1{,}0461$$

$$-0{,}9507 \le MZ_{yz/x} \le -0{,}0953$$

По таблице z - преобразования совершим переход к интервальным оценкам ρ:

$$0{,}19 \le \rho_{xy/z} \le 0{,}78$$

$$-0{,}1 \le \rho_{yz/x} \le 0{,}74$$

На основании полученных расчетов можно сделать вывод, что существует тесная взаимосвязь каждого из исследуемых показателей эффективности работы с другими, т.е. все множественные коэффициенты детерминации значимы и превышают 0,8.

Особенно тесная связь между фондоотдачей и двумя остальными показателями. Изменение фондоотдачи в среднем на 85,6% объясняется изменением производительности труда и материалоемкости. При увеличении производительности труда на 1 тыс. руб. фондоотдача увеличивается в среднем на 0,55 руб на рубль основных производственных фондов; при уменьшении материалоемкости на 1% фондоотдача увеличивается в среднем на 0,48 руб.

Взаимосвязь между материалоемкостью и производительностью труда не доказана (без учета фондоотдачи). Однако можно сказать, что фондоотдача усиливает связь между материалоемкостью и производительностью труда, т.к. $|r_{xz}|>|r_{xz/y}|$.

4.4. Методы оценки корреляционных моделей.

Для оценки параметров корреляционных моделей в основном используют три метода: моментов, максимального правдоподобия и наименьших квадратов.

Метод моментов был предложен К.Пирсоном. В соответствии с ним первые q моментов случайной величины X приравниваются q выборочным моментам , полученным по экспериментальным данным. Теоретическим обоснованием метода моментов служит закон больших чисел, согласно которому для рассматриваемого случая при большом объеме выборки выборочные моменты близки к моментам генеральной совокупности.

Для двумерной корреляционной модели согласно методу моментов неизвестное ожидание оценивается средним арифметическим (выборочным начальным моментом первого порядка), а дисперсия - выборочной дисперсией

(выборочным центральным моментом второго порядка). Коэффициент корреляции ρ оценивается выборочным коэффициентом r, который является функцией выборочных начальных моментов первого порядка самих случайных величин и их произведения.

Метод моментов дает возможность получать состоятельные оценки, т.е. надежность выводов, сделанных при его использовании, зависит от количества наблюдений. Использование метода моментов на практике приводит к сравнительно простым вычислениям.

Метод максимального правдоподобия, предложенный английским математиком Р.А.Фишером, часто приводит к более сложным вычислениям, чем метод моментов, однако оценки, получаемые с его помощью, как правило, оказываются более надежными и особенно предпочтительными в случае малого числа наблюдений.

Метод максимального правдоподобия для оценки математического ожидания предполагает использование средней арифметической, которая обладает свойствами несмещенности, состоятельности и эффективности.

Дисперсию генеральной совокупности согласно методу максимального правдоподобия рекомендуется оценивать выборочной дисперсией, которая удовлетворяют лишь условию состоятельности. Использование исправленной дисперсии позволяет иметь оценку дисперсии, удовлетворяющую условиям несмещенности и состоятельности.

Применение метода максимального правдоподобия часто приводит к решению сложных систем уравнений, поэтому метод наименьших квадратов, использование которого связано с более простыми выкладками, имеет большое практическое применение. Основоположниками этого метода являются Лекандр, Р.Андрейн, Гаусс.

Основная идея метода наименьших квадратов сводится к тому, чтобы в качестве оценки неизвестного параметра принимать значение, которое минимизируют сумму квадратов отклонений между оценкой и параметром для всех наблюдений.

Так как нормальный закон распределения генеральной совокупности является исходной предпосылкой построения корреляционных моделей, метод наименьших квадратов и метод максимального правдоподобия дают одинаковые результаты.

В анализе двумерной корреляционной модели обычно оценку уравнения регрессии производят с помощью метода наименьших квадратов.

4.5. Ранговая корреляция.

Для изучения взаимосвязи признаков, не поддающихся количественному измерению, используются различные показатели ранговой корреляции. В этой случае элементы совокупности располагают в определенном порядке в соответствии с некоторыми признаками (качественным и количественным) , т.е. производят ранжирование. При этом каждому объекту присваивается порядковый номер, называемый рангом. Например, элементу с наименьшим значением признака присваивается ранг 1, следующему за ним элементу - ранг 2 и т.д. Элементы можно располагать также в порядке убывания значений признака. Если объекты ранжированы по двум признакам, то можно изменить силу связи между признаками, основываясь на значениях рангов.

Коэффициент ранговой корреляции Спирмэна является парным, и его использование не связано с предпосылкой нормальности распределения исходных данных

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

где d - разность значений рангов, расположенных в двух рядах у одного и того же объекта.

Величина r_s для двух рядов, состоящих из n рангов, зависит только от Σd^2. Если два ряда полностью совпадают, то $\Sigma d^2=0$ и, следовательно $r_s=1$, т.е. при полной прямой связи $r_s=1$. При полной обратной связи, когда ранги двух рядов расположены в обратном порядке, $r_s=-1$. При отсутствии корреляции между рангами $r_s=0$.

Пример 4.3. При ранжировании оценок на вступительных экзаменах и средних баллов за первую экзаменационную сессию одних и тех же лиц получены следующие ранги:

таблица 4.3

Ранг	студент	А	Б	В	Г	Д	Е	Ж	З	И	К
	вступит. экзамен	2	5	6	1	4	10	7	8	3	9
	экзамен. сессия	3	6	4	1	2	7	8	10	5	9
d		-1	-1	2	0	2	3	-1	-2	-2	0
d^2		1	1	4	0	4	9	1	4	4	0

Из данных таблицы 4.3 следует: $\Sigma d^2=28$; $r_s=1-\frac{6\cdot 28}{10\left(10-1\right)}=0{,}83$,

что свидетельствует о достаточно высокой связи между изучаемыми признаками.

Для измерения тесноты связи между признаками, не поддающимися точной количественной оценке, используются и другие коэффициенты, например коэффициент Кэндела, конкордации, ассоциации, контингенции и др.

4.6. Нелинейная парная корреляция

Для изучения связи между признаками, которая выражается нелинейной функцией, используется более общий, чем коэффициент корреляции, показатель тесноты связи - корреляционное отношение.

Нелинейная (или криволинейная) связь между двумя величинами - это такая связь , при которой равномерным изменениям одной величины соответствует неравномерные изменения другой, причем эта неравномерность имеет определенный закономерный характер.

Использование корреляционного отношения основано на разложении общей дисперсии зависимой переменной на составляющие: дисперсию, характеризующую влияние объясняющей переменной, и дисперсию, характеризующую влияние неучтенных и случайных факторов:

$$S_y^2 = S_{y/x}^2 + S_{ост}^2$$

где

S^2_y - общая дисперсия зависимой переменной, т.е. дисперсия относительно среднего значения;

$S^2_{y/x}$ - дисперсия функции регрессии относительно среднего значения зависимой переменной, характеризующая влияние объясняющей переменной;

$S^2_{ост}$ - дисперсия зависимой переменной y относительно функции регрессии, т.е. остаточная дисперсия.

Корреляционной отношение выборочных данных определяется по формуле

$$\eta_{yx} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})2}}$$

$$\text{или } \eta_{yx} = \sqrt{1 - \frac{S^2_{ост}}{S^2_y}}$$

Влияние корреляционного отношения заключено в пределах

$$0 \le \eta_{yx} \le 1$$

Если дисперсия $S^2_{y/x}$, обусловленная зависимостью величины y от объясняющей переменной x, равно общей дисперсии S^2_y (а это возможно лишь при наличие функциональной связи), то $\eta_{yx}=1$. Если же остаточная (т.е. необъясненная) дисперсия $S^2_{ост}$ равна общей дисперсии S^2_y, то $\eta_{yx}=0$, т.е. корреляционная связь отсутствует.

В предыдущей главе было отмечено, что линейный коэффициент парной корреляции является симметричной функцией относительно x и y. Следует подчеркнуть, что этим свойством не обладают корреляционное отношение, т.е. $\eta_{xy}\neq\eta_{yx}$. Для линейной связи $\eta_{xy}=\eta_{yx}=r_{xy}$ Поэтому величину η^2-r^2 можно использовать для характеристики нелинейности связи между переменными.

В качестве одного из самых простых критериев оценки нелинейности связи можно использовать следующий:

$$K_н = \frac{\sqrt{n}}{0{,}67449} \cdot \frac{1}{2}\sqrt{\eta^2_{yx} - r^2_{yx}}$$

Если значение $K_н>2{,}5$, то корреляционную связь можно считать нелинейной.

Проверка и построение доверительных интервалов для корреляционного отношения генеральной совокупности осуществляются так же, как аналогичные процедуры для линейного коэффициента парной корреляции:

$$t_н = \frac{\eta_{yx}}{\sqrt{1-\eta^2_{yx}}}\sqrt{n-2};$$

$t_{кр}$ находится по таблице распределения Стьюдента из условия

$$\left.\begin{matrix}\alpha \\ \nu = n-2\end{matrix}\right\} \to t_{кр}$$

доверительный интервал имеет вид

$$\eta - t_\gamma \sqrt{\frac{1-\eta^2}{n-3}} \le \eta \le \eta + t_\gamma \sqrt{\frac{1-\eta^2}{n-3}}$$

где t_γ находится по таблице интегральной функции Лапласа с учетом уровня доверительной вероятности γ.

Следует обратить внимание на то, что использование корреляционного отношения η имеет смысл только для функций криволинейных, но линейных относительно параметров. Для функций, нелинейных относительно параметров [типа $\tilde{y} = f(x)$], корреляционное отношение не может служить точным измерителем тесноты связи.

Тест

1. В каких пределах изменяется парный коэффициент корреляции?

а) $0 \le \rho_{xy} \le 1$
б) $-1 \le \rho_{xy} \le 1$
в) $-\infty \le \rho_{xy} \le +\infty$
г) $0 \le \rho_{xy} \le \infty$

2. В каких пределах изменяется множественный коэффициент корреляции?

а) $0 \le \rho_{y/xz} \le 1$
б) $-1 \le \rho_{y/xz} \le 1$
в) $-\infty \le \rho_{y/xz} \le +\infty$
г) $0 \le \rho_{y/xz} \le \infty$

3. Если парный коэффициент корреляции по модулю больше модуля соответствующего частного (например $|\rho_{xy}| > |\rho_{xy/z}|$) и коэффициенты не имеют разных знаков, то это означает, что:

а) фиксируемая переменная z ослабляет корреляционную связь;
б) фиксируемая переменная усиливает связь между x и y;
в) фиксируемая переменная не связана с факторами x и y;
г) возможен любой из первых трех исходов.

4. Коэффициент детерминации между x и y характеризует:

а) долю дисперсии y, обусловленную влиянием не входящих в модель факторов;
б) долю дисперсии y, обусловленную влиянием x;
в) долю дисперсии x, обусловленную влиянием не входящих в модель факторов;
г)направление зависимости между x и y.

5. Парный коэффициент корреляции между факторами равен 1. Это означает:

а) наличие нелинейной функциональной связи;
б) отсутствие связи;
в) наличие функциональной связи;
г) отрицательную линейную связь.

6. На основании 20 наблюдений выяснено, что выборочная доля дисперсии случайной величины y, вызванной вариацией x, составит 64%. Чему равен выборочный парный коэффициент корреляции:

а) 0,64;
б) 0,36;
в) 0,8;
г) 0,8 или -0,8.

7. По данным выборочного обследования группы предприятий было установлено, что выборочная доля дисперсии прибыли y, вызванная влиянием неучтенных в модели факторов, кроме фондовооруженности x, составляет 19%. Чему равен выборочный коэффициент детерминации:

а) 0,9
б) -0,9
в) 0,81
г) 0,19

8. По результатам выборочных наблюдений были получены выборочные коэффициенты регрессии: b_{yx}=-0,5 и b_{xy}=-1,62. Чему равен выборочный парный коэффициент корреляции?

а) 0,81
б) 0,9
в) -0,9
г) 0,19

9. Частный коэффициент корреляции оценивает:

а) тесноту связи между двумя переменными при фиксированном значении остальных;
б) тесноту связи между двумя переменными;
в) тесноту связи между тремя переменными;
г) свободное влияние нескольких переменных на одну.

10. Множественный коэффициент корреляции оценивает:

а) долю дисперсии одной переменной, обусловленную влиянием остальных переменных, включенных в модель;
б) степень совокупного влияния нескольких переменных на одну;
в) тесноту нелинейной связи между переменными;
г) тесноту связи между двумя переменными при фиксированном значении остальных.

5. РЕГРЕССИОННЫЙ АНАЛИЗ

5.1. Задачи регрессионного анализа

Понятия регрессии и корреляции непосредственно связаны между собой, но при этом существует четкое различие между ними. В корреляционном анализе

оценивается сила стохастической связи, в регрессионном анализе исследуются ее формы.

Под регрессионным анализом обычно понимают метод стохастического анализа зависимости случайной величины Y от переменных x_j (j=1, 2, ..., k), рассматриваемых как неслучайные величины, независимо от истинного закона распределения x_j.

С помощью уравнения регрессии $\hat{y} = f(x_1, x_2, ..., x_ђ)$, применяемого для экономического анализа, можно измерить влияние отдельных факторов на зависимую переменную, что делает анализ конкретным, существенно повышает его познавательную ценность, уравнения регрессии также применяются в прогнозных работах.

Построение уравнения регрессии предполагает решение двух основных задач.

Первая задача заключается в выборе независимых переменных, оказывающих существенное влияние на зависимую величину, а также в определении вида уравнения регрессии.

Вторая задача построения уравнения регрессии - оценивание параметров (коэффициентов) уравнения. Она решается с помощью того или иного математико-статистического метода обработки данных. В связи с тем, что оценки параметров уравнения являются выборочными характеристиками, в процессе оценивания необходимо проводить статистическую проверку существенности полученных параметров.

Выбор уравнения регрессии осуществляется в соответствии с экономической сущностью изучаемого явления. Процессы, где влияние факторов-аргументов происходит с постоянным ускорением или замедлением, описываются параболическими кривыми. Иногда в экономике для описания зависимостей используются и более сложные виды функций, например, логистические, если процесс сначала ускоренно развивается, а затем после достижения некоторого уровня затухает и приближается к некоему пределу.

Наиболее простыми видами зависимости являются линейные, или приводимые к ним.

На практике чаще встречаются следующие виды уравнений регрессии:

- $\tilde{y} = \beta_0 + \beta_1 x$ - двумерное линейное;
- $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$ - полиномиальное;
- $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$ - гиперболическое;
- $\tilde{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ - линейное многомерное;
- $\tilde{y} = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \ldots x_k^{\beta_k}$ - степенное.

Линейной с точки зрения регрессионного анализа называется - модель, линейная относительно неизвестных параметров β_j.

Будем рассматривать модель, зависящую линейно как от параметров β_j так и от переменных x_j.

Так как теория линейных моделей разработана наиболее полно, то на практике степенные уравнения регрессии часто преобразуют к линейному путем логарифмирования:

$\lg \tilde{y} = \lg \beta_0 + \beta_1 \lg x_1 + \beta_2 \lg x_2 + \ldots + \beta_k \lg x_k$.

С помощью подстановок $\lg x_j = u_j$; $\lg \tilde{y} = \tilde{z}$ и $\lg \beta_0 = \beta_0'$ приходят к получению линейного уравнения регрессии:

$\widetilde{Z} = \beta_0^{/} + \beta_1 u_1 + \beta_2 u_2 + \ldots + \beta_k u_k$.

Путем подстановок $\frac{1}{x} = u$ и $x^j = u_j$ гиперболическое и полиномиальное уравнения так же могут быть преобразованы в линейные.

Предполагается, что случайная величина Y имеет нормальный закон распределения с условным математическим ожиданием $\widetilde{Y}$, являющимся функцией аргументов x_j (j=1, 2, ..., k), и постоянной, не зависящей от аргументов дисперсии σ^2.

В общем виде линейная связь регрессионного анализа может быть представлена в следующем виде:

$$\widetilde{Y} = \sum_{j=1}^{n} \beta_j \ \varphi_j \ (x_1, x_2, \ldots, x_k) + \varepsilon,$$

где:

- φ_j - некоторая функция переменных $x_1, x_2, \ldots, x_k$;
- ε - случайная величина с нулевым математическим ожиданием $M(\varepsilon)=0$ и дисперсией $D(\varepsilon)=\sigma^2$;
- β_j - коэффициенты уравнения регрессии.

Оценка неизвестных параметров β_j (j = 1, 2, 3, ..., k) по результатам выборки объемом n является основной задачей регрессионного анализа.

Для оценки неизвестных параметров уравнение регрессии чаще всего используют метод наименьших квадратов, который позволяет получить несмещенные оценки. В случае линейной модели b_j будут несмещенными оценками с минимальной дисперсией параметров β_j:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k .$$

5.2. Исходные предпосылки регрессионного анализа и свойства оценок

Применение методов наименьших квадратов для нахождения оценок параметров простой множественной регрессии предполагает выполнение некоторых предпосылок, касающихся прежде всего случайной переменной ε в уравнении $y=x\beta+\varepsilon$, учитывающей ошибки измерения и ошибки спецификации. Эти предпосылки не определяются объемом выборки и числом включенных в анализ переменных.

1. Полагаем, что при заданных значениях переменных на переменную Y не оказывают влияния никакие другие систематически действующие факторы и случайности, учитываемые с помощью ε, т.е. $M(\varepsilon)=0$. Отсюда следует, что средний уровень переменной Y определяется только функцией $\hat{y} = x\beta$ и возмущающая переменная ε не коррелирует со значениями регрессии.

2. Дисперсия случайной переменной ε должна быть для всех ε_i одинакова и постоянна: $M(\varepsilon_i^2) = \sigma_\varepsilon^2$. Это свойство переменной ε называется гомоскедастичностью и означает, что неучтенные факторы и модели оказывают одинаковое влияние.

3. Значение случайной переменной ε попарно не коррелированны, т.е. $M(\varepsilon_i \varepsilon_{i-l})=0$ (для $l \neq 0$). В случае, когда исходные данные представляют собой временные ряды, выполнение этой предпосылки свидетельствуют об отсутствии автокорреляции возмущающей переменной ε. Обобщая вторую и третью предпосылки, можно записать:

$$M(\varepsilon\varepsilon^{T}) = \sigma_{\varepsilon}^{2}E = \begin{bmatrix} \sigma_{\varepsilon}^{2} & 0 & . & 0 \\ 0 & \sigma_{\varepsilon}^{2} & & \\ . & . & . & . \\ 0 & 0 & . & \sigma_{\varepsilon}^{2} \end{bmatrix}.$$

4. Число наблюдений должно превышать число параметров. Согласно этой предпосылке, между объясняющими переменными не должно быть линейной зависимости, т.е. предполагается отсутствие мультиколлинеарности.

5. Объясняющие переменные не должны коррелировать с возмущающей переменной ε, т.е. $M(x\varepsilon)=0$. Отсюда следует, что переменные x_j (j=1, 2, ..., k) объясняют переменную y, а переменная y не объясняет переменные x_j (j=1, 2, ..., k).

6. Возмущающая переменная распределена нормально, не оказывает никакого существенного влияния на переменную у и представляет собой суммарный эффект от большого числа незначительных некоррелированных влияющих факторов. Одновременно эта предпосылка означает, что зависимая переменная у или переменные y и x_j (j=1, 2, ..., k) распределены нормально. Оценки параметров регрессии являются функциями от наблюдаемых значений и зависят также от применяемых методов оценки. Метод наименьших квадратов - один из наиболее распространенных. Исходя из того, что статистическая оценка в отличие от оцениваемых параметров является случайной величиной с определенным распределением вероятностей, считают, что распределение этой случайной величины зависит от закона распределения возмущающей переменной ε.

Метод наименьших квадратов (МНК) дает хорошее приближение оценок b_j к истинным значениям параметров β_j.

5.3. Двумерная линейная регрессионная модель.

Рассмотрим простейшую двумерную модель регрессионного анализа:

$$\tilde{y} = M(y / x = x) = \beta_0 + \beta_1 x. \qquad (5.1)$$

Выражение (5.1) называется функцией регрессии у на х. Определению подлежат параметры уравнения регрессии β_0 и β_1, называемые коэффициентами регрессии, а также $\sigma^2_{ост}$ - остаточная дисперсия.

Остаточной дисперсией называется та часть вариации зависимой переменной, которую нельзя объяснить воздействием объясняющей переменной. Именно поэтому остаточная дисперсия может быть использована для оценки качества модели, точности подбора функции, полноты набора объясняющих переменных.

Для нахождения оценок параметров уравнения регрессии чаще всего используется метод наименьших квадратов. Обозначим оценки параметров уравнения регрессии β_0 и β_1 как b_0 и b_1. В соответствии с методом наименьших квадратов оценки b_0 и b_1 можно получить из условия минимизации суммы квадратов ошибок оценивания, т.е. суммой квадратов отклонений фактических значений зависимой переменной от расчетных ее значений, полученных на основе уравнения регрессии

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x)^2 \to \min, \qquad (5.2)$$

где $\hat{y}_i = b_0 + b_1 x_i$.

Значения $\hat{y}_i$ называются расчетными; они представляют собой значения зависимой переменной при заданном значении объясняющей переменной и в предположении, что последняя является единственной причиной изменения у, а

ошибка оценки равна нулю. Разброс фактических значений $\hat{y}_i$ вокруг y_i обусловлен воздействием множества случайных факторов. Разность $(y_i - \hat{y}_i)$ называется остатком и дает количественную оценку значения ошибки, т.е. показывает воздействие возмущающей переменной.

Для того, чтобы найти минимум функции (5.2), сначала рассчитывают частные производные первого порядка, затем каждую из них приравнивают к нулю и решают полученную систему уравнений.

На основе изложенного выведем теперь оценки коэффициентов регрессии:

$$\frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i);$$

$$2(\sum_{i=1}^{n} y_i - nb_0 - b_1\sum_{i=1}^{n} x_i) = 0,$$

откуда

$$nb_0 + b_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i);$$

$$2(\sum_{i=1}^{n} x_i y_i - b_0\sum_{i=1}^{n} x_i - b_1\sum_{i=1}^{n} x_i^2) = 0,$$

откуда

$$b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Итак, получили систему двух линейных уравнений, которая называется системой нормальных уравнений:

$$\begin{cases} nb_0 + b_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad (5.3)$$

Решим систему относительно b_0 и b_1.

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \cdot n} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (5.4)$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1\frac{\sum_{i=1}^{n} x_i}{n} = \overline{y} - b_1\overline{x}. \qquad (5.5)$$

Оценку остаточной дисперсии можно получить, используя формулу

$$\hat{S}_{ост}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \qquad (5.6)$$

Следует отметить, что оценки b_0 и b_1 коэффициентов регрессии β_0 и β_1, полученных по методу наименьших квадратов, обладает минимальной дисперсией среди всех возможных в классе линейных оценок.

Свободный член b_0 определяет точку пересечения линии регрессии с осью ординат (рис 5.1). Поскольку b_0 является средним значением y в точке x=0, экономическая интерпретация его вряд ли возможна. Поэтому на практике обычно больший интерес вызывает коэффициент регрессии b_1.

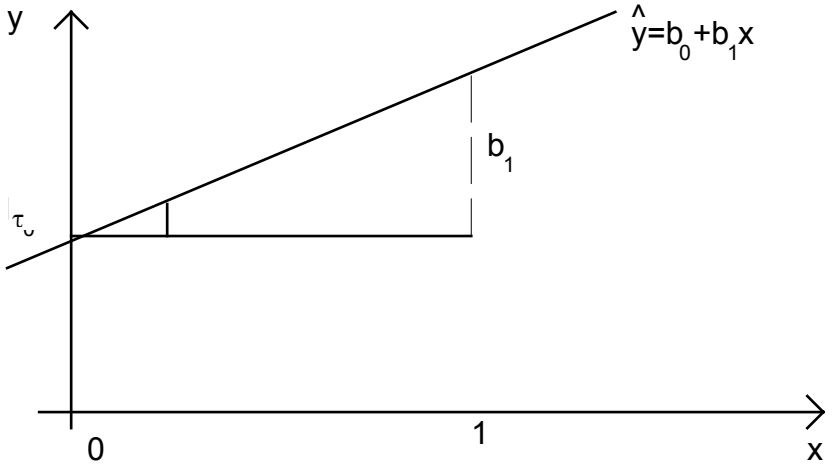


Рис. 5.1. Регрессионная прямая и ее параметры

Коэффициент регрессии b_1 характеризует наклон прямой, описываемой уравнением, к оси абсцисс. Если обозначить угол, образуемый этой прямой и осью ox как τ, то $b_1=tg\tau$. Коэффициент регрессии b_1 показывает среднюю величину изменения зависимой переменной y при изменении объясняющей переменной x на единицу собственного изменения. Знак при b_1 указывает направление этого изменения. Если коэффициент регрессии имеет отрицательный знак, то это говорит об отрицательной регрессии, при которой увеличение значений объясняющей переменной ведет к убыванию значения y. Если коэффициент регрессии имеет положительный знак, то это говорит о положительной регрессии, означающей, что при увеличении значений объясняющей переменной увеличиваются и значения зависимой переменной.

Коэффициент b_0 имеет размерность зависимой переменной. Размерность коэффициента регрессии b_1 представляет собой отношение размерности зависимой переменной к размерности объясняющей переменной.

После того, как модель построена, то есть найдены ее параметры, необходимо проверить ее адекватность исходным данным, а также полученную точность.

При соблюдении всех предпосылок регрессионного анализа можно проверить значимость уравнения регрессии, для чего следует проверить нулевую гипотезу $H_0 : \beta_1=0$. В основе проверки лежит идея дисперсионного анализа, состоящая в разложении дисперсии на составляющие. В регрессионном анализе общая сумма $Q_{общ}$ квадратов отклонений зависимой переменной разлагается на сумму квадратов Q_R отклонений, обусловленных регрессией, которая характеризует воздействие объясняющей переменной, и сумму квадратов $Q_{ост}$ отклонений относительно плоскости регрессии, характеризующую воздействие неучтенных в модели или случайных факторов. При этом $Q_{общ}=Q_R + +Q_{ост}$, где $Q_{общ} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Разложим $Q_{общ}$ на составляющие, прибавив и вычтя предварительно $\hat{y}_i$:

$$Q_{общ} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left[(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\right]^2 =$$

$$= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

Покажем, что последнее слагаемое равно 0. Для этого учтем (5.2) и (5.5) запишем:

$$(\hat{y}_i - \bar{y}) = (b_0 + b_1 x_i) - (b_0 + b_1 \bar{x}) = b_1 (x_i - \bar{x})$$

и $$(y_i - \hat{y}_i) = y_i - b_0 - b_1 x_i = y_i - (\bar{y} - b_1 \bar{x}) - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$$

Тогда получим с учетом (5.4)

$$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

Откуда:

$$Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (5.7)$$

$$Q_{ост} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5.8)$$

Понятно, что чем меньше $Q_{ост}$, т.е. меньше воздействие неучтенных в модели или случайных факторов, тем точнее соответствует модель фактическим данным.

Для проверки гипотезы используется F-критерий, основанный на статистике

$$F_н = \frac{Q_R / 1}{Q_{ост} / n - 2}, \qquad (5.9)$$

который имеет распределение Фишера-Снедекора с числом степеней свободы ν_1=1 и ν_2=n-2.

Задавшись уровнем значимости α и соответствующим числом степеней свободы (используя таблицу F-распределения Фишера-Снедекора), находим $F_{кр}$, удовлетворяющее условию $P(F_н > F_{кр}) \le \alpha$.

Если $F_н > F_{кр}$, нулевая гипотеза отвергается и уравнение регрессии считается значимым. При $F_н \le F_{кр}$ оснований для отклонения гипотезы нет.

Если уравнение регрессии значимо, то представляет интерес определение с надежностью γ интервальных оценок параметров β_0, β_1, $\tilde{y}$:

$$\beta_0 \in \left[b_0 + t_\gamma \frac{\hat{S}_{ост} \sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]; \qquad (5.10)$$

$$\beta_1 \in \left[b_1 \pm t_\gamma \frac{\hat{S}_{ост}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]; \qquad (5.11)$$

$$\tilde{y} \in \left[(b_0 + b_1 x_0) \pm t_\gamma \hat{S}_{ост} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right]. \qquad (5.12)$$

Доверительную оценку с надежностью γ для интервала предсказания в точке $x=x_0$ определяют по формуле (здесь $x_0 \neq x_i$, где i=1,2,...,n):

$$\tilde{y}_{n+1} \in \left[(b_0 + b_1 x_0) \pm t_\gamma \hat{S}_{\text{ост}} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 1} \right], \qquad (5.13)$$

где t_γ определяют по таблице t-распределения Стьюдента при α=1-γ и ν=n-2.

Одной из наиболее эффективных оценок адекватности построенной модели является коэффициент детерминации r^2, определяемый как:

$$r^2 = \frac{Q_R / 1}{\frac{Q_{общ}}{n-1}} = \frac{\hat{S}_R^2}{\hat{S}_{общ}^2}. \qquad (5.14)$$

Отношение (5.14) показывает, какая часть общей дисперсии зависимой переменной y обусловлена вариацией объясняющей переменной x. Чем больше доля дисперсии $\hat{S}_R^2$ в общей дисперсии $\hat{S}_{общ}^2$, тем лучше выбранная функция аппроксимирует фактические данные. При этом выбранная функция тем лучше определена, чем меньше величина $\hat{S}_{общ}^2$, т.е. чем меньше эмпирические значения отклоняются от расчетной линии регрессии.

Величина коэффициента детерминации находится в интервале $0 \leq r^2 \leq 1$. Если r^2=0, то это означает, что вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели факторов. В этом случае линия регрессии будет параллельна оси абсцисс: $y_i = \bar{y}$ - и никакой причинно-следственной связи не будет наблюдаться.

Если r^2=1, то все фактические значения y_i лежат на линии регрессии, т.е. $y_i = \hat{y}_i$. В этом случае говорят о строгой линейной функциональной связи между зависимой и объясняющей переменными.

При расчете коэффициента детерминации удобно пользоваться видоизмененной формулой

$$r^2 = \frac{\left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)^2}{\left(n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \right)\left(n\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i \right)} \qquad (5.15)$$

Легко заметить, что r^2 является квадратом выборочного коэффициента корреляции r. Величина 1-r^2 характеризует долю общей дисперсии зависимой переменной, объясненную воздействием неучтенных в модели и случайных факторов.

Поясним это на примере. Для проведения экономического анализа было случайным образом отобрано 71 предприятие хлебопекарной промышленности. Следует оценить зависимость между x - долей активной части в стоимости основных

промышленно-производственных фондов, %; y - выработкой товарной продукции на одного работающего, тыс. руб.

По исходным данным определим вспомогательные величины:

Σx_i=1911,9; Σy_i=1037,5; $\Sigma x_i y_i$=29296,89; Σx_i^2=58317,27; Σy_i^2=16391,56.

Определим оценки параметров, уравнения регрессии, для чего воспользуемся формулами 5.4 и 5.5

$$b_1 = \frac{412{,}632 - 26{,}298 \cdot 14{,}613}{96{,}243} - \frac{19{,}133}{96{,}243} = 0{,}199;$$

$$b_0 = 14{,}613 - 0{,}199 \cdot 26{,}928 = 9{,}254$$

Таким образом, получим $\hat{y}$ =9,254+0,199x.

Проверим значимость полученного уравнения, для чего определим Q_R и $Q_{ост}$ по формулам (5.7) и (5.8).

Q_R = 269,29;

$Q_{ост}$ = 964,03.

Тогда $F_H = \frac{269{,}29}{964{,}03:69} = 19{,}27$

Найдем $F_{кр}$ из условия α=0,05; ν_1=1; ν_2=69 по таблице Фишера - Снедекора. $F_{кр}$ = 4.

Уравнение оказывается статически значимым (нулевая гипотеза отвергается).

В результате статистического моделирования получено уравнение регрессии $\hat{y} = 9{,}254 + 0{,}199x$ зависимости выработки товарной продукции на одного работающего от доли активной части основных промышленно-производственных фондов.

Коэффициент регрессии b_1 = 0,199 показывает, что при изменении доли активной части фондов на 1% выработка товарной продукции на одного работающего увеличивается на 0,199 тыс. руб., или на 199 рублей. Коэффициент детерминации r^2 =0,468^2 =0,291, т.е. 21,9% вариации зависимой переменной объясняется вариацией доли активной части фондов, а 78,1% вариации вызвано воздействием неучтенных в модели и случайных факторов. Поэтому очевидно, что для характеристики выработки товарной продукции данная модель малопригодна.

Для сравнительного анализа влияния разных факторов и устранения различий в единицах их измерения используется коэффициент эластичности

$$Э = b_1 \frac{\overline{x}}{\overline{y}} = 0{,}199 \frac{26{,}928}{14{,}613} = 0{,}367.$$

Он означает, что при изменении (увеличении) доли активной части фондов на 1% выработка товарной продукции увеличивается на 0,367%.

Для устранения различий в степени колеблемости переменных в экономическом анализе используются β-коэффициенты:

$$\beta^{CT} = \frac{b_1 S_x}{S_y} = 0{,}199 \frac{9{,}81}{4{,}164} = 0{,}47.$$

Величина β^{CT} коэффициента свидетельствует о том, что при увеличении доли активной части фондов на одно среднеквадратическое отклонение выработка товарной продукции увеличится примерно на 0,5 среднеквадратического отклонения.

Таким образом, в результате экономической интерпретации выясняется, что модель недостаточно адекватно отражает исследуемый процесс, поэтому требуется дополнительный содержательный анализ по выявлению факторов, оказывающих существенное влияние на производительность труда.

Тест

1. Уравнение регрессии имеет вид $\tilde{y} = 5{,}1 - 1{,}7x$. На сколько единиц своего измерения в среднем изменится *y* при увеличении *x* на 1 единицу своего измерения:

а) Увеличится на 1,7;
б) Не изменится;
в) Уменьшится на 1,7;
г) Увеличится на 3,4.

2. Статистика $\dfrac{Q_R / 1}{Q_{ост} / n-2}$ имеет распределение:

а) Фишера-Снедекора;
б) Фишера-Йейтса;
в) Стьюдента;
г) Пирсона.

3. Несмещенная оценка остаточной дисперсии в двумерной регрессионной модели рассчитывается по формуле:

а) $\hat{S}^2_{ост} = \dfrac{1}{n-2} Q_{ост}$;

б) $\hat{S}^2_{ост} = \dfrac{1}{n-1} Q_{ост}$;

в) $\hat{S}^2_{ост} = \dfrac{1}{n} Q_{ост}$;

г) $\hat{S}^2_{ост} = \dfrac{1}{n-3} Q_{ост}$.

4. При интервальной оценке коэффициентов регрессии t_α определяется по таблице:

а) Нормального распределения;
б) Распределения Стьюдента;
в) Распределения Фишера-Снедекора;
г) *Z*-преобразования Фишера.

5. Согласно методу наименьших квадратов в качестве оценок параметров β_0 и β_1 следует использовать такие значения *b1* и *b2*, которые минимизируют сумму квадратов отклонений:

а) фактических значений зависимой переменной от ее среднего значения;
б) фактических значений объясняемой переменной от ее среднего значения;
в) расчетных значений зависимой переменной от ее среднего значения;
г) фактических значений зависимой переменной от ее расчетных значения.

6. Какой коэффициент указывает в среднем процент изменения результативного показателя *y* при увеличении аргумента *x* на 1 процент:

а) Бета-коэффициент;
б) коэффициент эластичности;
в) коэффициент детерминации;
г) коэффициент регрессии.

7. Линейное относительно аргумента уравнение регрессии имеет вид:

а) $\tilde{y} = \beta_0 + \beta_1 x + \beta_2 x^2$;

б) $\tilde{y} = \beta_0 + \beta_1 x$;

в) $\tilde{y} = \beta_0 + \beta_1 \frac{1}{x}$;

г) $\tilde{y} = \beta_0 x_1^{\beta_1}$.

8. При проверке гипотезы H_0 : β_1=0 оказалось, что $F_{набл} > F_{кр}$. Справедливо следующее утверждение:

а) β_1=0;

б) $\beta_1 \neq 0$;

в) $\beta_1 \neq 0$ с вероятностью ошибки α;

г) β_1=0 с вероятностью ошибки α.

9. Оценку b_1 коэффициента β_1 находят по формуле:

а) $b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S^2}$;

б) $b_1 = \frac{\bar{x}\,\bar{y} - \overline{xy}}{S_y^2}$;

в) $b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_x^2}$;

г) $b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{S_y^2}$.

10. Какая из следующих формул справедлива?

а) $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0$;

б) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 0$;

в) $\sum_{i=1}^{n} |\hat{y}_i - \bar{y}| = 0$;

г) $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$.

6. Выводы

Математическая статистика - это наука, занимающаяся анализом результатов наблюдений или опытов и построением оптимальных математико-статистических моделей массовых повторяющихся явлений, изменчивость которых обусловлена рядом неуправляемых факторов.

Методы математической статистики расширяют возможности научного прогнозирования и принятия решений в задачах, где параметры модели не могут быть известны или контролируемы с достаточной точностью. Ее методы универсальны и в настоящее время широко используются в экономике, технике,

социологии, демографии, медицине и других отраслях народного хозяйства при анализе явлений, обладающих тем свойством, что хотя результат отдельного опыта не может быть однозначно определен, но значения результатов наблюдения обладают свойством статистической устойчивости.

Задачи математической статистики практически сводятся к обоснованному суждению об объективных свойствах генеральной совокупности по результатам случайной выборки из нее. Причем выборка называется случайной, когда каждый элемент генеральной совокупности имеет одинаковую вероятность быть отобранным.

Необходимость выборочного обследования при решении практических задач может быть связана со следующими причинами:

- генеральная совокупность настолько многочисленна, что проведение обследования всех элементов совокупности (сплошное обследование) слишком трудоемко. С такой ситуацией приходится встречаться при контроле продукции крупносерийного и массового производства;
- при бесконечной генеральной совокупности, когда даже весьма большое множество наблюдений не исчерпывает всей совокупности;
- в процессе проведения испытания происходит разрушение испытуемых образцов (например, испытание на определение срока службы изделия, предела прочности и т.д.);
- встречаются обстоятельства, когда мы располагаем результатами испытания всей совокупности, реально существующей на данный момент, но рассматриваем их как выборку из гипотетической генеральной сверхсовокупности. Так поступают в тех случаях, когда хотят выявить общую закономерность, по отношению к которой имеющаяся совокупность представляется лишь частным случаем. Например, на протяжении ряда лет установлено, что доля мальчиков среди новорожденных составляла 0,513 общего числа родившихся в стране. Это данные сплошного обследования. Но если нас интересует общая закономерность соотношения полов среди новорожденных и мы хотим распространить полученные результаты на последующие годы, то эти данные следует рассматривать как выборку из некоторой бесконечной сверхсовокупности.

Перед экономистом, инженером, организатором производства ежедневно возникают вопросы, связанные с расчетом экономической эффективности различных мероприятий.

В связи с этим квалифицированному специалисту необходимо не только иметь правильные качественные представления об основных направлениях развития экономики, но и уметь учитывать сложное взаимосвязанное многообразие факторов, оказывающих заметное воздействие на производительность труда, объем производства, расход сырья и других видов ресурсов. Сложность состоит в том, что на строго детерминированные /определенные/ процессы и явления, определяющие развитие производства, накладываются случайные влияния. Следовательно, нельзя проводить ответственные экономические и технологические исследования без учета действия случайных факторов, без знания основ теории вероятностей и математической статистики - дисциплин, занимающихся нахождением численных закономерностей массовых случайных явлений.

Теория вероятностей с помощью специфических средств математического анализа раскрывает переход от случайного в единичных явлениях к объективной закономерности в массе таких явлений. Правила, выясняющие этот переход, составляют математическое содержание закона больших чисел. Знание закономерностей, которым подчиняются массовые случайные явления, позволяет

предвидеть, как эти явления будут развиваться. Таким образом, теория вероятностей - это математическая дисциплина, изучающая случайные явления и выясняющая закономерности, которым подчинены случайные явления при массовом их повторении.

Более широкому внедрению методов математической статистики в социально-экономических исследованиях способствует успешное развитие электронно-вычислительной техники. ПЭВМ дают возможность решать сложные социально-экономические задачи в режиме диалога экономиста-исследователя и машины.