Лекция 1.

Предмет теории вероятностей. Случайные события. Алгебра событий. Относительная частота и вероятность случайного события. Полная группа событий. Классическое определение вероятности. Основные свойства вероятности. Основные формулы комбинаторики.

В различных разделах науки и техники нередко возникают ситуации, когда результат каждого из многих проводимых опытов заранее предугадать невозможно, однако можно исследовать закономерности, возникающие при проведении серии опытов. Нельзя, например, точно сказать, какая сторона монеты окажется сверху при данном броске: герб или цифра – но при большом количестве бросков число выпадений герба приближается к половине количества бросков; нельзя заранее предсказать результат одного выстрела из данного орудия по данной цели, но при большом числе выстрелов частота попадания приближается к некоторому постоянному числу. Исследование вероятностных закономерностей массовых однородных явлений составляет предмет **теории вероятностей**.

Основным интуитивным понятием классической теории вероятностей является **случайное событие**.

События, которые могут произойти в результате опыта, можно подразделить на три вида:

а) **достоверное событие** – событие, которое всегда происходит при проведении опыта;

б) **невозможное событие** – событие, которое в результате опыта произойти не может;

в) **случайное событие** – событие, которое может либо произойти, либо не произойти.

Например, при броске игральной кости достоверным событием является выпадение числа очков, не превышающего 6, невозможным – выпадение 10 очков, а случайным – выпадение 3 очков.

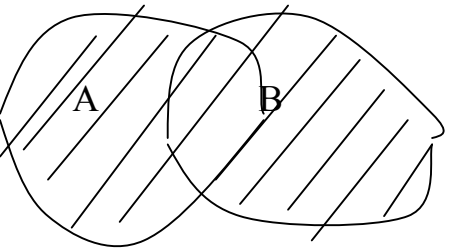
Алгебра событий.

Определение 1.1. **Суммой A+B** двух событий A и B называют событие, состоящее в том, что произошло хотя бы одно из событий A и B. **Суммой нескольких событий**, соответственно, называется событие, заключающееся в том, что произошло хотя бы одно из этих событий.

Пример 1. Два стрелка делают по одному выстрелу по мишени. Если событие A – попадание первого стрелка, а событие B – второго, то сумма A+B – это хотя бы одно попадание при двух выстрелах.

Пример 2. Если при броске игральной кости событием A_i назвать выпадение i очков, то выпадение нечетного числа очков является суммой событий $A_1+A_2+A_3$.

Назовем все возможные результаты данного опыта его *исходами* и предположим, что множество этих исходов, при которых происходит событие A (исходов, *благоприятных* событию A), можно представить в виде некоторой области на плоскости. Тогда множество исходов, при которых произойдет событие A+B, является объединением множеств исходов, благоприятных событиям A или B (рис. 1).



A + B

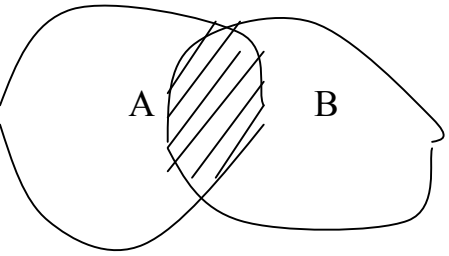
Рис.1.

Определение 1.2. **Произведением АВ** событий А и В называется событие, состоящее в том, что произошло и событие А, и событие В. Аналогично **произведением нескольких событий** называется событие, заключающееся в том, что произошли все эти события.

Пример 3. В примере 1 (два выстрела по мишени) событием АВ будет попадание обоих стрелков.

Пример 4. Если событие А состоит в том, что из колоды карт извлечена карта пиковой масти, а событие В – в том, что из колоды вынута дама, то событием АВ будет извлечение из колоды дамы пик.

Геометрической иллюстрацией множества исходов опыта, благоприятных появлению произведения событий А и В, является пересечение областей, соответствующих исходам, благоприятным А и В.



AB

Рис.2.

Определение 1.3. **Разностью А\В** событий А и В называется событие, состоящее в том, что А произошло, а В – нет.

Пример 5. Вернемся к примеру 1, где А\ В – попадание первого стрелка при промахе второго.

Пример 6. В примере 4 А\В – извлечение из колоды любой карты пиковой масти, кроме дамы. Наоборот, В \А – извлечение дамы любой масти, кроме пик.

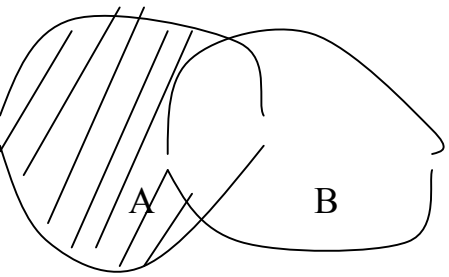


Рис.3.

Введем еще несколько категорий событий.

Определение 1.4. События A и B называются **совместными**, если они могут произойти оба в результате одного опыта. В противном случае (то есть если они не могут произойти одновременно) события называются **несовместными.**

Примеры: совместными событиями являются попадания двух стрелков в примере 1 и появление карты пиковой масти и дамы в примере 4; несовместными – события $A_1 - A_6$ в примере 2.

Замечание 1. Если изобразить графически области исходов опыта, благоприятных несовместным событиям, то они не будут иметь общих точек.

Замечание 2. Из определения несовместных событий следует, что их произведение является невозможным событием.

Определение 1.5. Говорят, что события A_1, A_2,…,A_n образуют **полную группу**, если в результате опыта обязательно произойдет хотя бы одно из событий этой группы.

Замечание. В частности, если события, образующие полную группу, попарно несовместны, то в результате опыта произойдет *одно и только одно* из них. Такие события называют **элементарными событиями**.

Пример. В примере 2 события $A_1 - A_6$ (выпадение одного, двух,…, шести очков при одном броске игральной кости) образуют полную группу несовместных событий.

Определение 1.6. События называются **равновозможными**, если нет оснований считать, что одно из них является более возможным, чем другое.

Примеры: выпадение любого числа очков при броске игральной кости, появление любой карты при случайном извлечении из колоды, выпадение герба или цифры при броске монеты и т.п.

Классическое определение вероятности.

При изучении случайных событий возникает необходимость количественно сравнивать возможность их появления в результате опыта. Например, при последовательном извлечении из колоды пяти карт более возможна ситуация, когда появились карты разных мастей, чем появление пяти карт одной масти; при десяти бросках монеты более

возможно чередование гербов и цифр, нежели выпадение подряд десяти гербов, и т.д. Поэтому с каждым таким событием связывают по определенному правилу некоторое число, которое тем больше, чем более возможно событие. Это число называется **вероятностью события** и является вторым основным понятием теории вероятностей.
Отметим, что само понятие вероятности, как и понятие случайного события, является аксиоматическим и поэтому не поддается строгому определению. То, что в дальнейшем будет называться различными определениями вероятности, представляет собой способы вычисления этой величины.

Определение 1.7. Если все события, которые могут произойти в результате данного опыта,
а) попарно несовместны;
б) равновозможны;
в) образуют полную группу,
то говорят, что имеет место **схема случаев.**

Можно считать, что случаи представляют собой все множество исходов опыта. Пусть их число равно n (число возможных исходов), а при m из них происходит некоторое событие А (число благоприятных исходов).

Определение 1.8. **Вероятностью события А** называется отношение числа исходов опыта, благоприятных этому событию, к числу возможных исходов:

$$p(A) = \frac{m}{n} \quad - \qquad (1.1)$$

- *классическое определение вероятности*.

Свойства вероятности.
Из определения 1.8 вытекают следующие свойства вероятности:

Свойство 1. Вероятность достоверного события равна единице.
Доказательство. Так как достоверное событие всегда происходит в результате опыта, то все исходы этого опыта являются для него благоприятными, то есть $m = n$, следовательно,
$P(A) = 1$.

Свойство 2. Вероятность невозможного события равна нулю.
Доказательство. Для невозможного события ни один исход опыта не является благоприятным, поэтому $m = 0$ и $p(A) = 0$.

Свойство 3. Вероятность случайного события есть положительное число, заключенное между нулем и единицей.
Доказательство. Случайное событие происходит при некоторых исходах опыта, но не при всех, следовательно, $0 < m < n$, и из (1.1) следует, что $0 < p(A) < 1$.

Пример. Из урны, содержащей 6 белых и 4 черных шара, наудачу вынут шар. Найти вероятность того, что он белый.
Решение. Будем считать элементарными событиями, или исходами опыта, извлечение из урны каждого из имеющихся в ней шаров. Очевидно, что эти события удовлетворяют

всем условиям, позволяющим считать их схемой случаев. Следовательно, число возможных исходов равно 10, а число исходов, благоприятных событию А (появлению белого шара) – 6 (таково количество белых шаров в урне). Значит,

$$p(A) = \frac{m}{n} = \frac{6}{10} = 0{,}6.$$

Относительная частота. Статистическое определение вероятности.

Классическое определение вероятности применимо только для очень узкого класса задач, где все возможные исходы опыта можно свести к схеме случаев. В большинстве реальных задач эта схема неприменима. В таких ситуациях требуется определять вероятность собы-тия иным образом. Для этого введем вначале понятие **относительной частоты W(A)** события А как отношения числа опытов, в которых наблюдалось событие А, к общему количеству проведенных испытаний:

$$W(A) = \frac{M}{N}, \qquad (1.2)$$

где N – общее число опытов, M – число появлений события A.
Большое количество экспериментов показало, что если опыты проводятся в одинаковых условиях, то для большого количества испытаний относительная частота изменяется мало, колеблясь около некоторого постоянного числа. Это число можно считать вероятностью рассматриваемого события.

Определение 1.9. **Статистической вероятностью события** считают его относительную частоту или число, близкое к ней.

Замечание 1. Из формулы (1.2) следует, что свойства вероятности, доказанные для ее классического определения, справедливы и для статистического определения вероят-ности.
Замечание 2. Для существования статистической вероятности события A требуется:

1) возможность производить неограниченное число испытаний;
2) устойчивость относительных частот появления A в различных сериях достаточно большого числа опытов.

Замечание 3. Недостатком статистического определения является неоднозначность статистической вероятности.

Пример. Если в задаче задается вероятность попадания в мишень для данного стрелка (скажем, p = 0,7), то эта величина получена в результате изучения статистики большого количества серий выстрелов, в которых этот стрелок попадал в мишень около семидесяти раз из каждой сотни выстрелов.

Основные формулы комбинаторики.

При вычислении вероятностей часто приходится использовать некоторые формулы *комбинаторики* – науки, изучающей комбинации, которые можно составить по определенным правилам из элементов некоторого конечного множества. Определим основные такие комбинации.

Определение 1.10. **Перестановки** – это комбинации, составленные из всех n элементов данного множества и отличающиеся только порядком их расположения. Число всех возможных перестановок

$$P_n = n! \qquad (1.3)$$

Пример. Сколько различных списков (отличающихся порядком фамилий) можно составить из 7 различных фамилий?
Решение. $P_7 = 7! = 2\cdot3\cdot4\cdot5\cdot6\cdot7 = 5040$.

Определение 1.11. **Размещения** – комбинации из m элементов множества, содержащего n различных элементов, отличающиеся либо составом элементов, либо их порядком. Число всех возможных размещений

$$A_n^m = n(n-1)(n-2)...(n-m+1). \qquad (1.4)$$

Пример. Сколько возможно различных вариантов пьедестала почета (первое, второе, третье места), если в соревнованиях принимают участие 10 человек?
Решение. $A_{10}^3 = 10\cdot9\cdot8 = 720.$

Определение 1.12. **Сочетания** – неупорядоченные наборы из m элементов множества, содержащего n различных элементов (то есть наборы, отличающиеся только составом элементов). Число сочетаний

$$C_n^m = \frac{n!}{m!(n-m)!}. \qquad (1.5)$$

Пример. В отборочных соревнованиях принимают участие 10 человек, из которых в финал выходят трое. Сколько может быть различных троек финалистов?
Решение. В отличие от предыдущего примера, здесь не важен порядок финалистов, следовательно, ищем число сочетаний из 10 по 3:

$$C_{10}^3 = \frac{10!}{3!\cdot7!} = \frac{8\cdot9\cdot10}{6} = 120.$$

Лекция 2.
Геометрические вероятности. Теорема сложения вероятностей. Противоположные события. Условные вероятности. Теорема умножения вероятностей. Независимые события. Вероятность появления хотя бы одного события.

Одним из недостатков классического определения вероятности является то, что оно неприменимо к испытаниям с бесконечным количеством исходов. В таких случаях можно воспользоваться понятием **геометрической вероятности**.

Пусть на отрезок L наудачу брошена точка. Это означает, что точка обязательно попадет на отрезок L и с равной возможностью может совпасть с любой точкой этого отрезка. При этом вероятность попадания точки на любую часть отрезка L не зависит от расположения этой части на отрезке и пропорциональна его длине. Тогда вероятность того, что брошен-ная точка попадет на отрезок l, являющийся частью отрезка L, вычисляется по формуле:

$$p = \frac{l}{L}, \qquad (2.1)$$

где l – длина отрезка l, а L – длина отрезка L.

Можно дать аналогичную постановку задачи для точки, брошенной на плоскую область S и вероятности того, что она попадет на часть этой области s:

$$p = \frac{s}{S}, \qquad (2.1`)$$

где s – площадь части области, а S – площадь всей области.

В трехмерном случае вероятность того, что точка, случайным образом расположенная в теле V, попадет в его часть v, задается формулой:

$$p = \frac{v}{V}, \qquad (2.1``)$$

где v – объем части тела, а V – объем всего тела.

Пример 1. Найти вероятность того, что точка, наудачу брошенная в круг, не попадет в правильный шестиугольник, вписанный в него.

Решение. Пусть радиус круга равен R , тогда сторона шестиугольника тоже равна R. При этом площадь круга $S = \pi R^2$, а площадь шестиугольника $s = \frac{3\sqrt{3}}{2} R^2$. Следовательно,

$$p = \frac{S - s}{S} = \frac{\pi R^2 - \frac{3\sqrt{3}}{2} R^2}{\pi R^2} = \frac{\pi - 3\sqrt{3}}{2\pi} \approx 0{,}174.$$

Пример 2. На отрезок AB случайным образом брошены три точки: C, D и M. Найти вероятность того, что из отрезков AC, AD и AM можно построить треугольник.

Решение. Обозначим длины отрезков AC, AD и AM через x, y и z и рассмотрим в качестве возможных исходов множество точек трехмерного пространства с координатами (x, y, z). Если принять длину отрезка равной 1, то эти множество возможных исходов представляет собой куб с ребром, равным 1. Тогда множество благоприятных исходов состоит из точек, для координат которых выполнены неравенства треугольника: $x + y > z$, $x + z > y$, $y + z > x$. Это часть куба, отрезанная от него плоскостями $x + y = z$, $x + z = y$, $y + z = x$

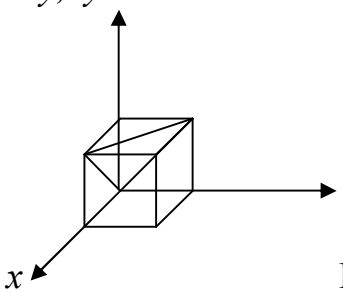


Рис.1.

(одна из них, плоскость $x + y = z$, проведена на рис.1). Каждая такая плоскость отделяет от куба пирамиду, объем которой равен $\frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6}$. Следовательно, объем оставшейся части

$v = 1 - 3 \cdot \frac{1}{6} = \frac{1}{2}$. Тогда $p = \frac{v}{V} = \frac{1}{2} : 1 = \frac{1}{2}$.

Теорема сложения вероятностей.

Теорема 2.1 (теорема сложения). Вероятность $p(A + B)$ суммы событий A и B равна

$$P(A + B) = p(A) + p(B) - p(AB). \qquad (2.2)$$

Доказательство.

Докажем теорему сложения для схемы случаев. Пусть n – число возможных исходов опыта, m_A – число исходов, благоприятных событию A, m_B – число исходов, благопри-ятных событию B, а m_{AB} – число исходов опыта, при которых происходят оба события (то есть исходов, благоприятных произведению AB). Тогда число исходов, при которых имеет место событие $A + B$, равно $m_A + m_B - m_{AB}$ (так как в сумме $(m_A + m_B)$ m_{AB} учтено дважды: как исходы, благоприятные A, и исходы, благоприятные B). Следовательно, вероятность суммы можно определить по формуле (1.1):

$$p(A+B) = \frac{m_A + m_B - m_{AB}}{n} = \frac{m_A}{n} + \frac{m_B}{n} - \frac{m_{AB}}{n} = p(A) + p(B) - p(AB),$$

что и требовалось доказать.

Следствие 1. Теорему 2.1 можно распространить на случай суммы любого числа событий. Например, для суммы трех событий A, B и C

$$P(A + B + C) = p(A) + p(B) + p(C) - p(AB) - p(AC) - p(BC) + p(ABC) \qquad (2.3)$$

и т.д.

Следствие 2. Если события A и B несовместны, то $m_{AB} = 0$, и, следовательно, вероятность суммы несовместных событий равна сумме их вероятностей:

$$P(A + B) = p(A) + p(B). \qquad (2.4)$$

Определение 2.1. **Противоположными событиями** называют два несовместных события, образующих полную группу. Если одно из них назвать A, то второе принято обозначать $\overline{A}$.

Замечание. Таким образом, $\overline{A}$ заключается в том, что событие A не произошло.

Теорема 2.2. Сумма вероятностей противоположных событий равна 1:

$$p(A) + p(\overline{A}) = 1. \qquad (2.5)$$

Доказательство.

Так как A и $\overline{A}$ образуют полную группу, то одно из них обязательно произойдет в результате опыта, то есть событие $A + \overline{A}$ является достоверным. Следовательно, $P(A + \overline{A}) = 1$. Но, так как A и $\overline{A}$ несовместны, из (2.4) следует, что $P(A + \overline{A}) = p(A) + p(\overline{A})$. Значит, $p(A) + p(\overline{A}) = 1$, что и требовалось доказать.

Замечание. В ряде задач проще искать не вероятность заданного события, а вероятность события, противоположного ему, а затем найти требуемую вероятность по формуле (2.5).

Пример. Из урны, содержащей 2 белых и 6 черных шаров, случайным образом извлекаются 5 шаров. Найти вероятность того, что вынуты шары разных цветов.

Решение. Событие $\overline{A}$, противоположное заданному, заключается в том, что из урны вынуто 5 шаров одного цвета, а так как белых шаров в ней всего два, то этот цвет может быть только черным. Множество возможных исходов опыта найдем по формуле (1.5):

$$n = C_8^5 = \frac{8!}{5!\cdot 3!} = \frac{6\cdot 7\cdot 8}{6} = 56,$$

а множество исходов, благоприятных событию $\overline{A}$ - это число возможных наборов по 5 шаров только из шести черных:

$$m_{\overline{A}} = C_6^5 = 6.$$

Тогда $p(\overline{A}) = \frac{6}{56} = \frac{3}{28}$, а $p(A) = 1 - \frac{3}{28} = \frac{25}{28}$.

Теорема умножения вероятностей

Определение 2.2. Назовем **условной вероятностью *p*(*B*/*A*) события *B*** вероятность события *B* при условии, что событие *A* произошло.

Замечание. Понятие условной вероятности используется в основном в случаях, когда осуществление события *A* изменяет вероятность события *B*.

Примеры:

1) пусть событие *A* – извлечение из колоды в 32 карты туза, а событие *B* – то, что и вторая вынутая из колоды карта окажется тузом. Тогда, если после первого раза карта была возвращена в колоду, то вероятность вынуть вторично туз не меняется: $p(B) = p(A) = \frac{4}{32} = \frac{1}{8} = 0{,}125$. Если же первая карта в колоду не возвращается, то осуществление события *A* приводит к тому, что в колоде осталась 31 карта, из которых только 3 туза. Поэтому $p(B/A) = \frac{3}{31} \approx 0{,}097$.
2) если событие *A* – попадание в самолет противника при первом выстреле из орудия, а *B* – при втором, то первое попадание уменьшает маневренность самолета, поэтому *p*(*B*/*A*) увеличится по сравнению с *p*(*A*).

Теорема 2.3 (теорема умножения). Вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого при условии, что первое событие произошло:

$$p(AB) = p(A) \cdot p(B/A). \qquad (2.6)$$

Доказательство.

Воспользуемся обозначениями теоремы 2.1. Тогда для вычисления *p*(*B*/*A*) множеством возможных исходов нужно считать m_A (так как *A* произошло), а множеством благоприятных исходов – те, при которых произошли и *A*, и *B* (m_{AB}). Следовательно,

$p(B/A) = \frac{m_{AB}}{m_A} = \frac{m_{AB}}{n} \cdot \frac{n}{m_A} = p(AB) : p(A)$, откуда следует утверждение теоремы.

Пример. Для поражения цели необходимо попасть в нее дважды. Вероятность первого попадания равна 0,2, затем она не меняется при промахах, но после первого попадания увеличивается вдвое. Найти вероятность того, что цель будет поражена первыми двумя выстрелами.

Решение. Пусть событие *A* – попадание при первом выстреле, а событие *B* – попадание при втором. Тогда $p(A) = 0{,}2$, $p(B/A) = 0{,}4$, $p(AB) = 0{,}2 \cdot 0{,}4 = 0{,}08$.

Следствие. Если подобным образом вычислить вероятность события *BA*, совпадающего с событием *AB*, то получим, что $p(BA) = p(B) \cdot p(A/B)$. Следовательно,

$$p(A) \cdot p(B/A) = p(B) \cdot p(A/B). \qquad (2.7)$$

Определение 2.3. Событие *B* называется **независимым** от события *A*, если появление события *A* не изменяет вероятности *B*, то есть $p(B/A) = p(B)$.

Замечание. Если событие B не зависит от A, то и A не зависит от B. Действительно, из (2.7) следует при этом, что $p(A) \cdot p(B) = p(B) \cdot p(A/B)$, откуда $p(A/B) = p(A)$. Значит, **свойство независимости событий взаимно**.

Теорема умножения для независимых событий имеет вид:

$$p(AB) = p(A) \cdot p(B), \qquad (2.8)$$

то есть вероятность произведения независимых событий равна произведению их вероятностей.

При решении задач теоремы сложения и умножения обычно применяются вместе.

Пример. Два стрелка делают по одному выстрелу по мишени. Вероятности их попадания при одном выстреле равны соответственно 0,6 и 0,7. Найти вероятности следующих событий:

A – хотя бы одно попадание при двух выстрелах;

B – ровно одно попадание при двух выстрелах;

C – два попадания;

D – ни одного попадания.

Решение. Пусть событие H_1 – попадание первого стрелка, H_2 – попадание второго. Тогда $A = H_1 + H_2$, $B = H_1 \cdot \overline{H}_2 + \overline{H}_1 \cdot H_2$, $C = H_1 \cdot H_2$, $D = \overline{H}_1 \cdot \overline{H}_2$. События H_1 и H_2 совместны и независимы, поэтому теорема сложения применяется в общем виде, а теорема умножения – в виде (2.8). Следовательно, $p(C) = 0{,}6 \cdot 0{,}7 = 0{,}42$, $p(A) = 0{,}6 + 0{,}7 - 0{,}42 = 0{,}88$, $p(B) = 0{,}6 \cdot 0{,}3 + 0{,}7 \cdot 0{,}4 = 0{,}46$ (так как события $H_1 \cdot \overline{H}_2$ и $\overline{H}_1 \cdot H_2$ несовместны), $p(D) = 0{,}4 \cdot 0{,}3 = 0{,}12$. Заметим, что события A и D являются противоположными, поэтому $p(A) = 1 - p(D)$.

Вероятность появления хотя бы одного события.

Теорема 2.4. Вероятность появления хотя бы одного из попарно независимых событий $A_1, A_2, \ldots, A_n$ равна

$$p(A) = 1 - q_1 q_2 \ldots q_n, \qquad (2.9)$$

где q_i – вероятность события $\overline{A}_i$, противоположного событию A_i.

Доказательство.

Если событие A заключается в появлении хотя бы одного события из $A_1, A_2, \ldots, A_n$, то события A и $\overline{A}_1 \overline{A}_2 \ldots \overline{A}_n$ противоположны, поэтому по теореме 2.2 сумма их вероятностей равна 1. Кроме того, поскольку $A_1, A_2, \ldots, A_n$ независимы, то независимы и $\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_n$, следовательно, $p(\overline{A}_1 \overline{A}_2 \ldots \overline{A}_n) = p(\overline{A}_1) p(\overline{A}_2) \ldots p(\overline{A}_n) = q_1 q_2 \ldots q_n$. Отсюда следует справедливость формулы (2.9).

Пример. Сколько нужно произвести бросков монеты, чтобы с вероятностью не менее 0,9 выпал хотя бы один герб?

Решение. Вероятность выпадения герба при одном броске равна вероятности противоположного события (выпадения цифры) и равна 0,5. Тогда вероятность выпадения хотя бы одного герба при n выстрелах равна $1 - (0{,}5)^n$. Тогда из решения неравенства $1 - (0{,}5)^n > 0{,}9$

следует, что $n > \log_2 10 \geq 4$.

Лекция 3.

Формула полной вероятности и формула Байеса. Схема и формула Бернулли. Приближение Пуассона для схемы Бернулли.

Определение 3.1. Пусть событие A может произойти только совместно с одним из событий H_1, H_2,…, H_n, образующих полную группу несовместных событий. Тогда события H_1, H_2,…, H_n называются **гипотезами.**

Теорема 3.1. Вероятность события A, наступающего совместно с гипотезами H_1, H_2,…, H_n, равна:

$$p(A) = \sum_{i=1}^{n} p(H_i)p(A/H_i), \qquad (3.1)$$

где $p(H_i)$ – вероятность i- й гипотезы, а $p(A/H_i)$ – вероятность события A при условии реализации этой гипотезы. Формула (3.1) носит название **формулы полной вероятности**.

Доказательство.

Можно считать событие A суммой попарно несовместных событий AH_1, AH_2,…, AH_n. Тогда из теорем сложения и умножения следует, что

$$p(A) = p(AH_1 + AH_2 + ... + AH_n) = p(AH_1) + p(AH_2) + ... + p(AH_n) = \sum_{i=1}^{n} p(H_i)p(A/H_i),$$

что и требовалось доказать.

Пример. Имеются три одинаковые урны с шарами. В первой из них 3 белых и 4 черных шара, во второй – 2 белых и 5 черных, в третьей – 10 черных шаров. Из случайно выбранной урны наудачу вынут шар. Найти вероятность того, что он белый.

Решение. Будем считать гипотезами H_1, H_2 и H_3 выбор урны с соответствующим номером. Так как по условию задачи все гипотезы равновозможны, то $p(H_1) = p(H_2) = p(H_3) = \frac{1}{3}$. Найдем условную вероятность A при реализации каждой гипотезы: $p(A/H_1) = \frac{3}{7}$,

$p(A/H_2) = \frac{2}{7}$, $p(A/H_3) = 0$. Тогда $p(A) = \frac{1}{3}\cdot\frac{3}{7} + \frac{1}{3}\cdot\frac{2}{7} + \frac{1}{3}\cdot 0 = \frac{5}{21} \approx 0{,}238.$

Формула Байеса (теорема гипотез).

Пусть известен результат опыта, а именно то, что произошло событие A. Этот факт может изменить априорные (то есть известные до опыта) вероятности гипотез. Например, в предыдущем примере извлечение из урны белого шара говорит о том, что этой урной не могла быть третья, в которой нет белых шаров, то есть $p\ (H_3/A) = 0$. Для переоценки вероятностей гипотез при известном результате опыта используется **формула Байеса**:

$$p(H_i / A) = \frac{p(H_i)p(A/H_i)}{p(A)}. \qquad (3.2)$$

Действительно, из (2.7) получим, что $p(A)p(H_i / A) = p(H_i)p(A/H_i),$ откуда следует справедливость формулы (3.2).

Пример. После двух выстрелов двух стрелков, вероятности попаданий которых равны 0,6 и 0,7, в мишени оказалась одна пробоина. Найти вероятность того, что попал первый стрелок.

Решение.

Пусть событие A – одно попадание при двух выстрелах, а гипотезы: H_1 – первый попал, а второй промахнулся, H_2 – первый промахнулся, а второй попал, H_3 – оба попали, H_4 – оба промахнулись. Вероятности гипотез: $p(H_1) = 0{,}6\cdot0{,}3 = 0{,}18$, $p(H_2) = 0{,}4\cdot0{,}7 = 0{,}28$, $p(H_3) = 0{,}6\cdot0{,}7 = 0{,}42$, $p(H_4) = 0{,}4\cdot0{,}3 = 0{,}12$. Тогда $p(A/H_1) = p(A/H_2) = 1$, $p(A/H_3) = p(A/H_4) = 0$. Следовательно, полная вероятность $p(A) = 0{,}18\cdot1 + 0{,}28\cdot1 + 0{,}42\cdot0 + 0{,}12\cdot0 = 0{,}46$. Применяя формулу Байеса, получим:

$$p(H_1 / A) = \frac{0{,}18 \cdot 1}{0{,}46} = \frac{9}{23} \approx 0{,}391.$$

Схема повторения испытаний. Формула Бернулли.

Рассмотрим серию из n испытаний, в каждом из которых событие A появляется с одной и той же вероятностью p, причем результат каждого испытания не зависит от результатов остальных. Подобная постановка задачи называется **схемой повторения испытаний**. Найдем вероятность того, что в такой серии событие A произойдет ровно $к$ раз (неважно, в какой последовательности). Интересующее нас событие представляет собой сумму равно-вероятных несовместных событий, заключающихся в том, что A произошло в некоторых $к$ испытаниях и не произошло в остальных $n - к$ испытаниях. Число таких событий равно числу сочетаний из n по $к$, то есть $C_n^к$, а вероятность каждого из них: $p^k q^{n-k}$, где $q = 1 - p$ – вероятность того, что в данном опыте A не произошло. Применяя теорему сложения для несовместных событий, получим **формулу Бернулли**:

$$p_n(k) = C_n^k \cdot p^k \cdot q^{n-k}. \qquad (3.3)$$

Пример. Для получения приза нужно собрать 5 изделий с особым знаком на этикетке. Найти вероятность того, что придется купить 10 изделий, если этикетки с этим знаком имеют 5% изделий.

Решение. Из постановки задачи следует, что последнее купленное изделие имеет особый знак. Следовательно, из предыдущих девяти эти знаки имели 4 изделия. Найдем вероят-ность этого по формуле Бернулли: $p_9(4) = C_9^4 \cdot (0{,}05)^4 \cdot (0{,}95)^5 = 0{,}0006092.$ Тогда $p = 0{,}0006092\cdot0{,}05 = 0{,}0000304$.

Приближение Пуассона для схемы Бернулли.

Формула Бернулли требует громоздких расчетов при большом количестве испытаний. Можно получить более удобную для расчетов приближенную формулу, если при большом числе испытаний вероятность появления A в одном опыте мала, а произведение $np = \lambda$ сохраняет постоянное значение для разных серий опытов (то есть среднее число появле-ний события A в разных сериях испытаний остается неизменным). Применим формулу Бернулли:

$$p_n(k) = \frac{n(n-1)(n-2)...(n-k+1)}{k!} p^k (1-p)^{n-k} = \frac{n(n-1)...(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Найдем предел полученного выражения при $n \to \infty$:

$$p_n(k) \approx \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-k}\right) = \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} = \frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1.$$

Таким образом, **формула Пуассона**

$$p_n(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (3.4)$$

позволяет найти вероятность *к* появлений события *A* для массовых (*n* велико) и редких (*p* мало) событий.

Лекция 4.

Случайные величины. Закон распределения и функция распределения дискретной случайной величины. Биномиальное распределение и распределение Пуассона.

Наряду с понятием случайного события в теории вероятности используется и более удобное понятие *случайной величины*.

Определение 4.1. **Случайной величиной** называется величина, принимающая в результате опыта одно из своих возможных значений, причем заранее неизвестно, какое именно.

Будем обозначать случайные величины заглавными буквами латинского алфавита (*X, Y,Z,…*), а их возможные значения – соответствующими малыми буквами (x_i, y_i,…).

Примеры: число очков, выпавших при броске игральной кости; число появлений герба при 10 бросках монеты; число выстрелов до первого попадания в цель; расстояние от центра мишени до пробоины при попадании.

Можно заметить, что множество возможных значений для перечисленных случайных величин имеет разный вид: для первых двух величин оно конечно (соответственно 6 и 11 значений), для третьей величины множество значений бесконечно и представляет собой множество натуральных чисел, а для четвертой – все точки отрезка, длина которого равна радиусу мишени. Таким образом, для первых трех величин множество значений из отдельных (дискретных), изолированных друг от друга значений, а для четвертой оно представляет собой непрерывную область. По этому показателю случайные величины подразделяются на две группы: дискретные и непрерывные.

Определение 4.2. Случайная величина называется **дискретной**, если она принимает отдельные, изолированные возможные значения с определенными вероятностями.

Определение 4.3. Случайная величина называется **непрерывной**, если множество ее возможных значений целиком заполняет некоторый конечный или бесконечный промежуток.

Дискретные случайные величины.

Для задания дискретной случайной величины нужно знать ее возможные значения и вероятности, с которыми принимаются эти значения. Соответствие между ними называется **законом распределения** случайной величины. Он может иметь вид таблицы, формулы или графика.

Таблица, в которой перечислены возможные значения дискретной случайной величины и соответствующие им вероятности, называется **рядом распределения**:

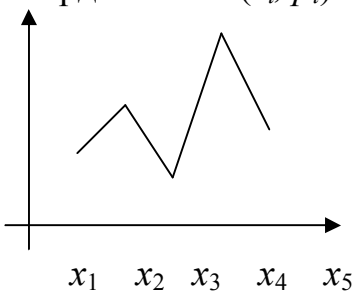
x_i	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

Заметим, что событие, заключающееся в том, что случайная величина примет одно из своих возможных значений, является достоверным, поэтому $\sum_{i=1}^{n(\infty)} p_i = 1.$

Пример. . Два стрелка делают по одному выстрелу по мишени. Вероятности их попадания при одном выстреле равны соответственно 0,6 и 0,7. Составить ряд распределения случайной величины X – числа попаданий после двух выстрелов.
Решение. Очевидно, что X может принимать три значения: 0, 1 и 2. Их вероятности найдены в примере, рассмотренном в лекции 3. Следовательно, ряд распределения имеет вид:

x_i	0	1	2
p_i	0,12	0,46	0,42

Графически закон распределения дискретной случайной величины можно представить в виде **многоугольника распределения** – ломаной, соединяющей точки плоскости с координатами (x_i, p_i).



Функция распределения.

Определение 4.4. **Функцией распределения *F*(*x*)** случайной величины X называется вероятность того, что случайная величина примет значение, меньшее x:

$$F(x) = p(X < x). \qquad (4.1)$$

Свойства функции распределения.

1) $0 \le F(x) \le 1$.

Действительно, так как функция распределения представляет собой вероятность, она может принимать только те значения, которые принимает вероятность.

2) Функция распределения является неубывающей функцией, то есть $F(x_2) \ge F(x_1)$ при $x_2 > x_1$. Это следует из того, что $F(x_2) = p(X < x_2) = p(X < x_1) + p(x_1 \le X < x_2) \ge F(x_1)$.
3) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$. В частности, если все возможные значения X лежат на интервале $[a, b]$, то $F(x) = 0$ при $x \le a$ и $F(x) = 1$ при $x \ge b$. Действительно, $X < a$ – событие невозможное, а $X < b$ – достоверное.
4) Вероятность того, что случайная величина примет значение из интервала $[a, b]$, равна разности значений функции распределения на концах интервала:

$$p\ (\ a < X < b\ ) = F(b) - F(a).$$

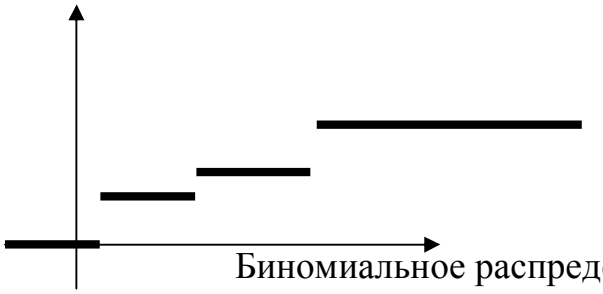
Справедливость этого утверждения следует из определения функции распределения (см. свойство 2).

Для дискретной случайной величины значение $F(x)$ в каждой точке представляет собой сумму вероятностей тех ее возможных значений, которые меньше аргумента функции.

Пример. Найдем $F(x)$ для предыдущего примера:

$$F(x) = \begin{cases} 0, & x \le 0 \\ 0{,}12, & 0 < x \le 1 \\ 0{,}12 + 0{,}46 = 0{,}58, & 1 < x \le 2 \\ 0{,}58 + 0{,}42 = 1, & x > 2 \end{cases}$$

Соответственно график функции распределения имеет ступенчатый вид:



Биномиальное распределение.

Вернемся к схеме независимых испытаний и найдем закон распределения случайной величины X – числа появлений события A в серии из n испытаний. Возможные значения A: 0, 1, …, n. Соответствующие им вероятности можно вычислить по формуле Бернулли:

$$p(X = k) = C_n^k p^k q^{n-k} \qquad (4.2)$$

(p – вероятность появления A в каждом испытании).

Такой закон распределения называют **биномиальным**, поскольку правую часть равенства (4.2) можно рассматривать как общий член разложения бинома Ньютона:

$$(p+q)^n = C_n^n p^n + C_n^{n-1} p^{n-1} q + \ldots + C_n^k p^k q^{n-k} + \ldots + C_n^0 q^n.$$

Пример. Составим ряд распределения случайной величины X – числа попаданий при 5 выстрелах, если вероятность попадания при одном выстреле равна 0,8.

$p(X{=}0) = 1\cdot(0{,}2)^5 = 0{,}00032$; $p(X{=}1) = 5\cdot0{,}8\cdot(0{,}2)^4 = 0{,}0064$; $p(X{=}2) = 10\cdot(0{,}8)^2\cdot(0{,}2)^3 = 0{,}0512$; $p(X{=}3) = 10\cdot(0{,}8)^3\cdot(0{,}2)^2 = 0{,}2048$; $p(X{=}4) = 5\cdot(0{,}8)^4\cdot0{,}2 = 0{,}4096$; $p(X{=}5) = 1\cdot(0{,}8)^5 = 0{,}32768$. Таким образом, ряд распределения имеет вид:

x	0	1	2	3	4	5
p	0.00032	0.0064	0.0512	0.2048	0.4096	0.32728

Распределение Пуассона.

Рассмотрим дискретную случайную величину X, принимающую только целые неотрицательные значения (0, 1, 2,…, m,…), последовательность которых не ограничена. Такая случайная величина называется распределенной **по закону Пуассона**, если вероятность того, что она примет значение m, выражается формулой:

$$p(X=m)=\frac{a^m}{m!}e^{-a}, \qquad (4.3)$$

где a – некоторая положительная величина, называемая *параметром* закона Пуассона. Покажем, что сумма всех вероятностей равна 1:

$$\sum_{m=0}^{\infty} p(X=m)=e^{-a}\sum_{m=0}^{\infty}\frac{a^m}{m!}=e^{-a}\cdot e^{a}=1$$

(использовано разложение в ряд Тейлора функции e^x).

Рассмотрим типичную задачу, приводящую к распределению Пуассона. Пусть на оси абсцисс случайным образом распределяются точки, причем их распределение удовлетворяет следующим условиям:

1) вероятность попадания некоторого количества точек на отрезок длины l зависит только от длины отрезка и не зависит от его расположения на оси (то есть точки распределены с одинаковой средней плотностью);
2) точки распределяются независимо друг от друга (вероятность попадания какого-либо числа точек на данный отрезок не зависит от количества точек, попавший на любой другой отрезок);
3) практическая невозможность совпадения двух или более точек.

Тогда случайная величина X – число точек, попадающих на отрезок длины l – распределена по закону Пуассона, где a – среднее число точек, приходящееся на отрезок длины l.

Замечание. В лекции 3 говорилось о том, что формула Пуассона выражает биномиальное распределение при большом числе опытов и малой вероятности события. Поэтому закон Пуассона часто называют *законом редких явлений*.

Лекция 5.

Функция распределения и плотность распределения непрерывной случайной величины, их взаимосвязь и свойства. Равномерное распределение вероятностей.

Определение и свойства функции распределения сохраняются и для непрерывной случайной величины, для которой функцию распределения можно считать одним из видов задания закона распределения. Но для непрерывной случайной величины вероятность каждого отдельного ее значения равна 0. Это следует из свойства 4 функции распределения: $p(X=a)=F(a)-F(a)=0$. Поэтому для такой случайной величины имеет смысл говорить только о вероятности ее попадания в некоторый интервал.

Вторым способом задания закона распределения непрерывной случайной величины является так называемая плотность распределения (плотность вероятности, дифференциальная функция).

Определение 5.1. Функция $f(x)$, называемая **плотностью распределения** непрерывной случайной величины, определяется по формуле:

$$f(x)=F'(x), \qquad (5.1)$$

то есть является производной функции распределения.

Свойства плотности распределения.

1) $f(x)\ge 0$, так как функция распределения является неубывающей.
2) $F(x)=\int_{-\infty}^{x} f(t)dt$, что следует из определения плотности распределения.

3) Вероятность попадания случайной величины в интервал $(a,\ b)$ определяется формулой $p(a < X < b) = \int_a^b f(x)dx.$

Действительно, $p(a < X < b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_a^b f(x)dx.$

4) $\int_{-\infty}^{+\infty} f(x)dx = 1$ (условие нормировки). Его справедливость следует из того, что $\int_{-\infty}^{+\infty} f(x)dx = F(+\infty),$ а $\lim_{x\to+\infty} F(x) = 1.$

5) $\lim_{x\to\pm\infty} f(x) = 0,$ так как $F(x) \to const$ при $x \to \pm\infty.$

Таким образом, график плотности распределения представляет собой кривую, располо-женную выше оси Ox, причем эта ось является ее горизонтальной асимптотой при $x \to \pm\infty$ (последнее справедливо только для случайных величин, множеством возможных значений которых является все множество действительных чисел). Площадь криволинейной трапеции, ограниченной графиком этой функции, равна единице.

Замечание. Если все возможные значения непрерывной случайной величины сосредоточе-ны на интервале $[a,\ b]$, то все интегралы вычисляются в этих пределах, а вне интервала $[a,\ b]$ $f(x) \equiv 0.$

Пример 1. Плотность распределения непрерывной случайной величины задана формулой

$$f(x) = \frac{C}{1+x^2},\ -\infty < x < +\infty.$$

Найти: а) значение константы C; б) вид функции распределения; в) $p(-1 < x < 1)$.

Решение. а) значение константы C найдем из свойства 4:

$$\int_{-\infty}^{+\infty} \frac{C}{1+x^2}dx = Carctgx\Big|_{-\infty}^{+\infty} = C\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = C\pi = 1,\ \text{откуда}\ C = \frac{1}{\pi}.$$

б) $F(x) = \frac{1}{\pi}\int_{-\infty}^{x} \frac{1}{1+t^2}dt = \frac{1}{\pi}arctg\ t\Big|_{-\infty}^{x} = \frac{1}{\pi}\left(arctgx + \frac{\pi}{2}\right) = \frac{1}{\pi}arctgx + \frac{1}{2}.$

в) $p(-1 < x < 1) = \frac{1}{\pi}\int_{-1}^{1} \frac{1}{1+x^2}dx = \frac{1}{\pi}arctgx\Big|_{-1}^{1} = \frac{1}{\pi}\left(\frac{\pi}{4} + \frac{\pi}{4}\right) = 0{,}5.$

Пример 2. Функция распределения непрерывной случайной величины имеет вид:

$$F(x) = \begin{cases} 0,\ x \le 2 \\ \dfrac{x-2}{2},\ 2 < x \le 4 \\ 1,\ x > 4. \end{cases}$$

Найти плотность распределения.

Решение.

$$f(x)=\begin{cases} 0', & x\le 2 \\ \left(\dfrac{x-2}{2}\right)', & 2<x\le 4 \\ 1', & x>4 \end{cases}=\begin{cases} 0, & x\le 2 \\ 0{,}5, & 2<x\le 4 \\ 0, & x>4. \end{cases}$$

Равномерный закон распределения.

Часто на практике мы имеем дело со случайными величинами, распределенными определенным типовым образом, то есть такими, закон распределения которых имеет некоторую стандартную форму. В прошлой лекции были рассмотрены примеры таких законов распределения для дискретных случайных величин (биномиальный и Пуассона). Для непрерывных случайных величин тоже существуют часто встречающиеся виды закона распределения, и в качестве первого из них рассмотрим равномерный закон.

Определение 5.2. Закон распределения непрерывной случайной величины называется **равномерным**, если на интервале, которому принадлежат все возможные значения случайной величины, плотность распределения сохраняет постоянное значение ($f(x)$ = const при $a \le x \le b$, $f(x) = 0$ при $x < a$, $x > b$.

Найдем значение, которое принимает $f(x)$ при $x \in [a,b]$. Из условия нормировки следует, что $\int_a^b f(x)dx = \int_a^b c dx = c(b-a) = 1$, откуда $f(x) = c = \dfrac{1}{b-a}$.

Вероятность попадания равномерно распределенной случайной величины на интервал $[\alpha, \beta]$ $(a \le \alpha < \beta \le b)$ равна при этом $\int_\alpha^\beta \dfrac{1}{b-a}dx = \dfrac{\beta-\alpha}{b-a}$.

Вид функции распределения для нормального закона: $F(x)=\begin{cases} 0, & x<a \\ \dfrac{x-a}{b-a}, & a\le x\le b \\ 1, & x>b. \end{cases}$

Пример. Автобусы некоторого маршрута идут с интервалом 5 минут. Найти вероятность того, что пришедшему на остановку пассажиру придется ожидать автобуса не более 2 минут.

Решение. Время ожидания является случайной величиной, равномерно распределенной в интервале [0, 5]. Тогда $f(x) = \dfrac{1}{5}$, $p(0 \le x \le 2) = \dfrac{2}{5} = 0{,}4$.

Лекция 6.

Нормальный закон распределения вероятностей. Нормальная кривая. Функция Лапласа. Вычисление вероятности попадания в заданный интервал нормальной случайной величины. Правило трех сигм. Показательное распределение. Функция надежности. Показательный закон надежности.

Определение 6.1. Непрерывная случайная величина называется распределенной по **нормальному закону**, если ее плотность распределения имеет вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}. \qquad (6.1)$$

Замечание. Таким образом, нормальное распределение определяется двумя параметрами: a и σ.

График плотности нормального распределения называют **нормальной кривой (кривой Гаусса)**. Выясним, какой вид имеет эта кривая, для чего исследуем функцию (6.1).

1) Область определения этой функции: (-∞, +∞).
2) $f(x) > 0$ при любом x (следовательно, весь график расположен выше оси Ox).
3) $\lim\limits_{|x|\to\infty} f(x)=0,$ то есть ось Ox служит горизонтальной асимптотой графика при $x\to\pm\infty$.
4) $f'(x)=-\frac{x-a}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}=0$ при $x = a$; $f'(x)>0$ при $x > a$, $f'(x)<0$ при $x < a$. Следовательно, $\left(a,\frac{1}{\sigma\sqrt{2\pi}}\right)$ - точка максимума.
5) $F(x – a) = f(a – x)$, то есть график симметричен относительно прямой $x = a$.
6) $f''(x)=-\frac{1}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}\left(1-\frac{(x-a)^2}{\sigma^2}\right)=0$ при $x=a\pm\sigma$, то есть точки $\left(a\pm\sigma,\frac{1}{\sigma\sqrt{2\pi e}}\right)$ являются точками перегиба.

Примерный вид кривой Гаусса изображен на рис.1.

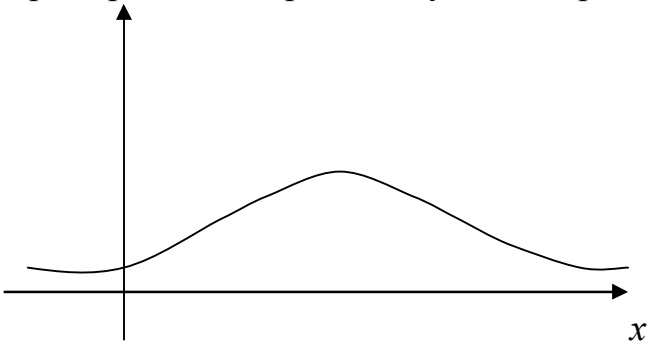


Рис.1.

Найдем вид функции распределения для нормального закона:

$$F(x)=\int_{-\infty}^{x}f(t)dt=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-a)^2}{2\sigma^2}}dt. \qquad (6.2)$$

Перед нами так называемый «неберущийся» интеграл, который невозможно выразить через элементарные функции. Поэтому для вычисления значений $F(x)$ приходится пользоваться таблицами. Они составлены для случая, когда $a = 0$, а $\sigma = 1$.

Определение 6.2. Нормальное распределение с параметрами $a = 0$, $\sigma = 1$ называется **нормированным**, а его функция распределения

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt \quad - \qquad (6.3)$$

- **функцией Лапласа**.

Замечание. Функцию распределения для произвольных параметров можно выразить через функцию Лапласа, если сделать замену: $t = \frac{x-a}{\sigma}$, тогда $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-a}{\sigma}} e^{-\frac{t^2}{2}} dt$.

Найдем вероятность попадания нормально распределенной случайной величины на заданный интервал:

$$p(\alpha < x < \beta) = F(\beta) - F(\alpha) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \qquad (6.4)$$

Пример. Случайная величина X имеет нормальное распределение с параметрами $a = 3$, $\sigma = 2$. Найти вероятность того, что она примет значение из интервала (4, 8).
Решение.

$$p(4 < x < 8) = F(8) - F(4) = \Phi\left(\frac{8-3}{2}\right) - \Phi\left(\frac{4-3}{2}\right). = \Phi(2{,}5) - \Phi(0{,}5) = 0{,}9938 - 0{,}6915 = 0{,}3023.$$

Правило «трех сигм».

Найдем вероятность того, что нормально распределенная случайная величина примет значение из интервала (a - 3σ, a + 3σ):

$$p(a - 3\sigma < x < a + 3\sigma) = \Phi(3) - \Phi(-3) = 0{,}9986 - 0{,}0014 = 0.9973.$$

Следовательно, вероятность того, что значение случайной величины окажется *вне* этого интервала, равна 0,0027, то есть составляет 0,27% и может считаться пренебрежимо малой. Таким образом, на практике можно считать, что *все* возможные значения нормально распределенной случайной величины лежат в интервале (a - 3σ, a + 3σ).

Полученный результат позволяет сформулировать **правило «трех сигм»**: *если случайная величина распределена нормально, то модуль ее отклонения от $x = a$ не превосходит 3σ.*

Показательное распределение.

Определение 6.3. **Показательным (экспоненциальным)** называют распределение вероятностей непрерывной случайной величины X, которое описывается плотностью

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0. \end{cases} \qquad (6.5)$$

В отличие от нормального распределения, показательный закон определяется только одним параметром λ. В этом его преимущество, так как обычно параметры распределения заранее не известны и их приходится оценивать приближенно. Понятно, что оценить один параметр проще, чем несколько.

Найдем функцию распределения показательного закона:

$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{0} 0 \cdot dt + \lambda \int_{0}^{x} e^{-\lambda t} dt = 1 - e^{-\lambda x}$. Следовательно,

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases} \qquad (6.6)$$

Теперь можно найти вероятность попадания показательно распределенной случайной величины в интервал (a, b):

$$p(a < x < b) = e^{-\lambda a} - e^{-\lambda b}. \qquad (6.7)$$

Значения функции e^{-x} можно найти из таблиц.

Функция надежности.

Пусть *элемент* (то есть некоторое устройство) начинает работать в момент времени $t_0 = 0$ и должен проработать в течение периода времени t. Обозначим за T непрерывную случайную величину – время безотказной работы элемента, тогда функция $F(t) = p(T > t)$ определяет вероятность отказа за время t. Следовательно, вероятность безотказной работы за это же время равна

$$R(t) = p(T > t) = 1 - F(t). \qquad (6.8)$$

Эта функция называется **функцией надежности**.

Показательный закон надежности.

Часто длительность безотказной работы элемента имеет показательное распределение, то есть

$$F(t) = 1 - e^{-\lambda t}.$$

Следовательно, функция надежности в этом случае имеет вид:

$$R(t) = 1 - F(t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}.$$

Определение 6.4. **Показательным законом надежности** называют функцию надежности, определяемую равенством

$$R(t) = e^{-\lambda t}, \qquad (6.9)$$

где λ – интенсивность отказов.

Пример. Пусть время безотказной работы элемента распределено по показательному закону с плотностью распределения $f(t) = 0{,}1\ e^{-0{,}1t}$ при $t \geq 0$. Найти вероятность того, что элемент проработает безотказно в течение 10 часов.

Решение. Так как $\lambda = 0{,}1$, $R(10) = e^{-0{,}1 \cdot 10} = e^{-1} = 0{,}368$.

Лекция 7.

Основные числовые характеристики дискретных и непрерывных случайных величин: математическое ожидание, дисперсия и среднее квадратическое отклонение. Их свойства и примеры.

Закон распределения (функция распределения и ряд распределения или плотность вероятности) полностью описывают поведение случайной величины. Но в ряде задач достаточно знать некоторые числовые характеристики исследуемой величины (например, ее среднее значение и возможное отклонение от него), чтобы ответить на поставленный вопрос. Рассмотрим основные числовые характеристики дискретных случайных величин.

Математическое ожидание.

Определение 7.1. **Математическим ожиданием** дискретной случайной величины называ-ется сумма произведений ее возможных значений на соответствующие им вероятности:

$$M(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n\,. \qquad (7.1)$$

Если число возможных значений случайной величины бесконечно, то $M(X) = \sum_{i=1}^{\infty} x_i p_i$, если полученный ряд сходится абсолютно.

Замечание 1. Математическое ожидание называют иногда *взвешенным средним*, так как оно приближенно равно среднему арифметическому наблюдаемых значений случайной величины при большом числе опытов.

Замечание 2. Из определения математического ожидания следует, что его значение не меньше наименьшего возможного значения случайной величины и не больше наибольше-го.

Замечание 3. Математическое ожидание дискретной случайной величины есть *неслучайная* (постоянная) величина. В дальнейшем увидим, что это же справедливо и для непрерывных случайных величин.

Пример 1. Найдем математическое ожидание случайной величины X – числа стандартных деталей среди трех, отобранных из партии в 10 деталей, среди которых 2 бракованных. Составим ряд распределения для X. Из условия задачи следует, что X может принимать значения 1, 2, 3. $p(1) = \frac{C_8^1 \cdot C_2^2}{C_{10}^3} = \frac{1}{15}$, $p(2) = \frac{C_8^2 \cdot C_2^1}{C_{10}^3} = \frac{7}{15}$, $p(3) = \frac{C_8^3}{C_{10}^3} = \frac{7}{15}$.

Тогда

$$M(X) = 1 \cdot \frac{1}{15} + 2 \cdot \frac{7}{15} + 3 \cdot \frac{7}{15} = 2{,}4.$$

Пример 2. Определим математическое ожидание случайной величины X – числа бросков монеты до первого появления герба. Эта величина может принимать бесконечное число значений (множество возможных значений есть множество натуральных чисел). Ряд ее распределения имеет вид:

X	1	2	…	n	…
p	0,5	$(0,5)^2$	…	$(0,5)^n$	…

Тогда $M(X) = \sum_{n=1}^{\infty} \frac{n}{2^n} = \frac{1}{2} + 2 \cdot \left(\frac{1}{2}\right)^2 + 3 \cdot \left(\frac{1}{2}\right)^3 + \ldots + n \cdot \left(\frac{1}{2}\right)^n + \ldots = \sum_{n=1}^{\infty} \frac{1}{2^n} + \frac{1}{2}\sum_{n=1}^{\infty} \frac{1}{2^n} + \ldots +$

$+ \frac{1}{2^n}\sum_{n=1}^{\infty} \frac{1}{2^n} + \ldots = 1 \cdot \left(1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots\right) = 1 \cdot 2 = 2$ (при вычислении дважды использовалась формула суммы бесконечно убывающей геометрической прогрессии: $S = \frac{b_1}{1-q}$, откуда $\frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 1,\ 1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots = 2$).

Свойства математического ожидания.

1) Математическое ожидание постоянной равно самой постоянной:

$$M(C) = C. \qquad (7.2)$$

Доказательство. Если рассматривать C как дискретную случайную величину, принимающую только одно значение C с вероятностью $p = 1$, то $M(C) = C \cdot 1 = C$.

2) Постоянный множитель можно выносит за знак математического ожидания:

$$M(CX) = C\,M(X). \qquad (7.3)$$

Доказательство. Если случайная величина X задана рядом распределения

x_i	x_1	x_2	…	x_n
p_i	p_1	p_2	…	p_n

то ряд распределения для CX имеет вид:

Cx_i	Cx_1	Cx_2	…	Cx_n
p_i	p_1	p_2	…	p_n

Тогда $M(CX) = Cx_1p_1 + Cx_2p_2 + \ldots + Cx_np_n = C(x_1p_1 + x_2p_2 + \ldots + x_np_n) = CM(X)$.

Определение 7.2. Две случайные величины называются **независимыми**, если закон распределения одной из них не зависит от того, какие значения приняла другая. В противном случае случайные величины **зависимы**.

Определение 7.3. Назовем **произведением независимых случайных величин *X* и *Y*** случайную величину XY, возможные значения которой равны произведениям всех возможных значений X на все возможные значения Y, а соответствующие им вероятности равны произведениям вероятностей сомножителей.

3) Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:

$$M(XY) = M(X)M(Y). \qquad (7.4)$$

Доказательство. Для упрощения вычислений ограничимся случаем, когда X и Y принимают только по два возможных значения:

x_i	x_1	x_2
p_i	p_1	p_2

y_i	y_1	y_2
g_i	g_1	g_2

Тогда ряд распределения для XY выглядит так:

XY	x_1y_1	x_2y_1	x_1y_2	x_2y_2
p	p_1g_1	$p_2\,g_1$	p_1g_2	p_2g_2

Следовательно, $M(XY) = x_1y_1 \cdot p_1g_1 + x_2y_1 \cdot p_2g_1 + x_1y_2 \cdot p_1g_2 + x_2y_2 \cdot p_2g_2 = y_1g_1(x_1p_1 + x_2p_2) + y_2g_2(x_1p_1 + x_2p_2) = (y_1g_1 + y_2g_2)(x_1p_1 + x_2p_2) = M(X) \cdot M(Y)$.

Замечание 1. Аналогично можно доказать это свойство для большего количества возможных значений сомножителей.

Замечание 2. Свойство 3 справедливо для произведения любого числа независимых случайных величин, что доказывается методом математической индукции.

Определение 7.4. Определим **сумму случайных величин *X* и *Y*** как случайную величину $X + Y$, возможные значения которой равны суммам каждого возможного

значения X с каждым возможным значением Y; вероятности таких сумм равны произведениям вероятностей слагаемых (для зависимых случайных величин – произведениям вероятности одного слагаемого на условную вероятность второго).

4) Математическое ожидание суммы двух случайных величин (зависимых или незави-симых) равно сумме математических ожиданий слагаемых:

$$M(X + Y) = M(X) + M(Y). \qquad (7.5)$$

Доказательство.

Вновь рассмотрим случайные величины, заданные рядами распределения, приведенными при доказательстве свойства 3. Тогда возможными значениями $X + Y$ являются $x_1 + y_1, x_1 + y_2, x_2 + y_1, x_2 + y_2$. Обозначим их вероятности соответственно как p_{11}, p_{12}, p_{21} и p_{22}. Найдем $M(X+Y) = (x_1 + y_1)p_{11} + (x_1 + y_2)p_{12} + (x_2 + y_1)p_{21} + (x_2 + y_2)p_{22} =$
$= x_1(p_{11} + p_{12}) + x_2(p_{21} + p_{22}) + y_1(p_{11} + p_{21}) + y_2(p_{12} + p_{22})$.

Докажем, что $p_{11} + p_{22} = p_1$. Действительно, событие, состоящее в том, что $X + Y$ примет значения $x_1 + y_1$ или $x_1 + y_2$ и вероятность которого равна $p_{11} + p_{22}$, совпадает с событием, заключающемся в том, что $X = x_1$ (его вероятность – p_1). Аналогично доказывается, что $p_{21} + p_{22} = p_2$, $p_{11} + p_{21} = g_1$, $p_{12} + p_{22} = g_2$. Значит,

$M(X + Y) = x_1p_1 + x_2p_2 + y_1g_1 + y_2g_2 = M(X) + M(Y)$.

Замечание. Из свойства 4 следует, что сумма любого числа случайных величин равна сумме математических ожиданий слагаемых.

Пример. Найти математическое ожидание суммы числа очков, выпавших при броске пяти игральных костей.

Найдем математическое ожидание числа очков, выпавших при броске одной кости: $M(X_1) = (1 + 2 + 3 + 4 + 5 + 6)\cdot\frac{1}{6} = \frac{7}{2}$. Тому же числу равно математическое ожидание числа очков, выпавших на любой кости. Следовательно, по свойству 4 $M(X) = 5\cdot\frac{1}{6} = \frac{5}{6}$.

Дисперсия.

Для того, чтобы иметь представление о поведении случайной величины, недостаточно знать только ее математическое ожидание. Рассмотрим две случайные величины: X и Y, заданные рядами распределения вида

X	49	50	51
p	0,1	0,8	0,1

Y	0	100
p	0,5	0,5

Найдем $M(X) = 49\cdot0,1 + 50\cdot0,8 + 51\cdot0,1 = 50$, $M(Y) = 0\cdot0,5 + 100\cdot0,5 = 50$. Как видно, математические ожидания обеих величин равны, но если для X $M(X)$ хорошо описывает поведение случайной величины, являясь ее наиболее вероятным возможным значением (причем остальные значения ненамного отличаются от 50), то значения Y существенно отстоят от $M(Y)$. Следовательно, наряду с математическим ожиданием желательно знать, насколько значения случайной величины отклоняются от него. Для характеристики этого показателя служит дисперсия.

Определение 7.5. **Дисперсией (рассеянием)** случайной величины называется математическое ожидание квадрата ее отклонения от ее математического ожидания:

$$D(X) = M(X - M(X))^2. \qquad (7.6)$$

Пример.

Найдем дисперсию случайной величины X (числа стандартных деталей среди отобранных) в примере 1 данной лекции. Вычислим значения квадрата отклонения каждого возможно-го значения от математического ожидания:
$(1 - 2{,}4)^2 = 1{,}96$; $(2 - 2{,}4)^2 = 0{,}16$; $(3 - 2{,}4)^2 = 0{,}36$. Следовательно,

$$D(X) = 1{,}96 \cdot \frac{1}{15} + 0{,}16 \cdot \frac{7}{15} + 0{,}36 \cdot \frac{7}{15} = \frac{28}{75} \approx 0{,}373.$$

Замечание 1. В определении дисперсии оценивается не само отклонение от среднего, а его квадрат. Это сделано для того, чтобы отклонения разных знаков не компенсировали друг друга.
Замечание 2. Из определения дисперсии следует, что эта величина принимает только неотрицательные значения.
Замечание 3. Существует более удобная для расчетов формула для вычисления дисперсии, справедливость которой доказывается в следующей теореме:

Теорема 7.1. $D(X) = M(X^2) - M^2(X)$. (7.7)
Доказательство.
Используя то, что $M(X)$ – постоянная величина, и свойства математического ожидания, преобразуем формулу (7.6) к виду:
$D(X) = M(X - M(X))^2 = M(X^2 - 2X \cdot M(X) + M^2(X)) = M(X^2) - 2M(X) \cdot M(X) + M^2(X) =$
$= M(X^2) - 2M^2(X) + M^2(X) = M(X^2) - M^2(X)$, что и требовалось доказать.

Пример. Вычислим дисперсии случайных величин X и Y, рассмотренных в начале этого раздела. $M(X) = (49^2 \cdot 0{,}1 + 50^2 \cdot 0{,}8 + 51^2 \cdot 0{,}1) - 50^2 = 2500{,}2 - 2500 = 0{,}2$.
$M(Y) = (0^2 \cdot 0{,}5 + 100^2 \cdot 0{,}5) - 50^2 = 5000 - 2500 = 2500$. Итак, дисперсия второй случайной величины в несколько тысяч раз больше дисперсии первой. Таким образом, даже не зная законов распределения этих величин, по известным значениям дисперсии мы можем утверждать, что X мало отклоняется от своего математического ожидания, в то время как для Y это отклонение весьма существенно.

Свойства дисперсии.

1) Дисперсия постоянной величины C равна нулю:
$$D(C) = 0. \quad (7.8)$$
Доказательство. $D(C) = M((C - M(C))^2) = M((C - C)^2) = M(0) = 0$.
2) Постоянный множитель можно выносить за знак дисперсии, возведя его в квадрат:
$$D(CX) = C^2 D(X). \quad (7.9)$$
Доказательство. $D(CX) = M((CX - M(CX))^2) = M((CX - CM(X))^2) = M(C^2(X - M(X))^2) =$
$= C^2 D(X)$.
3) Дисперсия суммы двух независимых случайных величин равна сумме их дисперсий:
$$D(X + Y) = D(X) + D(Y). \quad (7.10)$$
Доказательство. $D(X + Y) = M(X^2 + 2XY + Y^2) - (M(X) + M(Y))^2 = M(X^2) + 2M(X)M(Y) +$
$+ M(Y^2) - M^2(X) - 2M(X)M(Y) - M^2(Y) = (M(X^2) - M^2(X)) + (M(Y^2) - M^2(Y)) = D(X) + D(Y)$.
Следствие 1. Дисперсия суммы нескольких взаимно независимых случайных величин равна сумме их дисперсий.
Следствие 2. Дисперсия суммы постоянной и случайной величин равна дисперсии случайной величины.

4) Дисперсия разности двух независимых случайных величин равна сумме их дисперсий:

$$D(X - Y) = D(X) + D(Y). \qquad (7.11)$$

Доказательство. $D(X - Y) = D(X) + D(-Y) = D(X) + (-1)^2 D(Y) = D(X) + D(X)$.

Дисперсия дает среднее значение квадрата отклонения случайной величины от среднего; для оценки самого отклонения служит величина, называемая средним квадратическим отклонением.

Определение 7.6. **Средним квадратическим отклонением** σ случайной величины X называется квадратный корень из дисперсии:

$$\sigma = \sqrt{D(X)}\,. \qquad (7.12)$$

Пример. В предыдущем примере средние квадратические отклонения X и Y равны соответственно $\sigma_x = \sqrt{0{,}2} \approx 0{,}447; \quad \sigma_y = \sqrt{2500} = 50.$

Лекция 8

Числовые характеристики непрерывных случайных величин.

Распространим определения числовых характеристик случайных величин на непрерывные случайные величины, для которых плотность распределения служит в некотором роде аналогом понятия вероятности.

Определение 7.7. **Математическим ожиданием** непрерывной случайной величины называется

$$M(X) = \int_{-\infty}^{+\infty} x f(x)dx. \qquad (7.13)$$

Замечание 1. Общее определение дисперсии сохраняется для непрерывной случайной величины таким же, как и для дискретной (опр. 7.5), а формула для ее вычисления имеет вид:

$$D(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X). \qquad (7.14)$$

Среднее квадратическое отклонение вычисляется по формуле (7.12).

Замечание 2. Если все возможные значения непрерывной случайной величины не выходят за пределы интервала [*a, b*], то интегралы в формулах (7.13) и (7.14) вычисляются в этих пределах.

Пример. Плотность распределения случайной величины X имеет вид:

$$f(x) = \begin{cases} 0, & x < 2 \\ -\dfrac{3}{4}(x^2 - 6x + 8), & 2 \le x \le 4 \\ 0, & x > 4. \end{cases}$$

Найти $M(X)$, $D(X)$, σ.

Решение. $M(X) = -\dfrac{3}{4}\int_2^4 x(x^2 - 6x + 8)dx = -\dfrac{3}{4}\left(\dfrac{x^4}{4} - 2x^3 + 4x^2\right)\Bigg|_2^4 = 3;$

$$D(X) = -\frac{3}{4}\int_2^4 x^2(x^2 - 6x + 8)dx - 9 = -\frac{3}{4}\left(\frac{x^5}{5} - \frac{3x^4}{2} + \frac{8x^3}{3}\right)\Bigg|_2^4 - 9 = 9{,}2 - 9 = 0{,}2; \quad \sigma = \sqrt{0{,}2} \approx 0{,}447$$

Числовые характеристики случайных величин, имеющих некоторые стандартные законы распределения.

1. **Биномиальное распределение.**

Для дискретной случайной величины X, представляющей собой число появлений события A в серии из n независимых испытаний (см. лекцию 6), $M(X)$ можно найти, используя свойство 4 математического ожидания. Пусть X_1 – число появлений A в первом испытании, X_2 – во втором и т.д. При этом каждая из случайных величин X_i задается рядом распределения вида

X_i	0	1
p_i	q	p

Следовательно, $M(X_i) = p$. Тогда $M(X) = \sum_{i=1}^{n} M(X_i) = \sum_{i=1}^{n} p = np.$

Аналогичным образом вычислим дисперсию: $D(X_i) = 0^2 \cdot q + 1^2 \cdot p - p^2 = p - p^2 = p(1-p)$,

откуда по свойству 4 дисперсии $D(X) = \sum_{i=1}^{n} D(X_i) = np(1-p) = npq.$

2. **Закон Пуассона.**

Если $p(X = m) = \frac{a^m}{m!}e^{-a}$, то $M(X) = \sum_{m=1}^{\infty} m\frac{a^m}{m!}e^{-a} = ae^{-a}\sum_{m=1}^{\infty}\frac{a^{m-1}}{(m-1)!} = ae^{-a}e^{a} = a$ (использовалось разложение в ряд Тейлора функции e^x).

Для определения дисперсии найдем вначале $M(X^2) = \sum_{m=1}^{\infty} m^2\frac{a^m}{m!}e^{-a} = a\sum_{m=1}^{\infty} m\frac{a^{m-1}}{(m-1)!}e^{-a} =$

$$= a\sum_{m=1}^{\infty}\big((m-1)+1\big)\frac{a^{m-1}}{(m-1)!}e^{-a} = a\left(\sum_{m=1}^{\infty}(m-1)\frac{a^{m-1}}{(m-1)!}e^{-a} + \sum_{m=1}^{\infty}\frac{a^{m-1}}{(m-1)!}e^{-a}\right) = a(a+1).$$

Поэтому $D(X) = a^2 + a - a^2 = a$.

Замечание. Таким образом, обнаружено интересное свойство распределения Пуассона: математическое ожидание равно дисперсии (и равно единственному параметру a, определяющему распределение).

3. **Равномерное распределение.**

Для равномерно распределенной на отрезке $[a, b]$ непрерывной случайной величины

$$M(X) = \int_a^b x\frac{1}{b-a}dx = \frac{x^2}{2(b-a)}\Bigg|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$ то есть математическое ожидание равномерно распределенной случайной величины равно абсциссе середины отрезка $[a, b]$.

Дисперсия

$$D(X) = \int_a^b x^2\frac{1}{b-a}dx - \frac{(a+b)^2}{4} = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} =$$

$= \frac{(b-a)^2}{12}$.

4. **Нормальное распределение**.

Для вычисления математического ожидания нормально распределенной случайной величины воспользуемся тем, что *интеграл Пуассона* $\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$.

$$M(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-a)^2}{2\sigma^2}} dx = (z = \frac{x-a}{\sigma}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + a) e^{-\frac{z^2}{2}} dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \sigma z e^{-\frac{z^2}{2}} dz + \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = 0 + \frac{a}{\sqrt{2\pi}} \sqrt{2\pi} = a$$ (первое слагаемое равно 0, так как подынтегральная функция нечетна, а пределы интегрирования симметричны относительно нуля).

$$D(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z \cdot z e^{-\frac{z^2}{2}} dz = (u = z, dv = z e^{-\frac{z^2}{2}}) =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left(-z \cdot e^{-\frac{z^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \right) = \frac{\sigma^2}{\sqrt{2\pi}} \left(-0 + \sqrt{2\pi} \right) = \sigma^2.$$

Следовательно, параметры нормального распределения (a и σ) равны соответственно математическому ожиданию и среднему квадратическому отклонению исследуемой случайной величины.

Лекция 9

Основные понятия математической статистики. Генеральная совокупность и выборка. Вариационный ряд, статистический ряд. Группированная выборка. Группированный статистический ряд. Полигон частот. Выборочная функция распределения и гистограмма.

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений. Двумя основными задачами математической статистики являются:

- определение способов сбора и группировки этих статистических данных;
- разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т.д.;

б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых можно сделать прогноз относительно исследуемого признака этих объектов.
Определим основные понятия математической статистики.
Генеральная совокупность – все множество имеющихся объектов.
Выборка – набор объектов, случайно отобранных из генеральной совокупности.
Объем генеральной совокупности *N* и объем выборки *n* – число объектов в рассматриваемой совокупности.
Виды выборки:
Повторная – каждый отобранный объект перед выбором следующего возвращается в генеральную совокупность;
Бесповторная – отобранный объект в генеральную совокупность не возвращается.

Замечание. Для того, чтобы по исследованию выборки можно было сделать выводы о поведении интересующего нас признака генеральной совокупности, нужно, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была **репрезентативной** (представительной). Учитывая закон больших чисел, можно утверждать, что это условие выполняется, если каждый объект выбран случайно, причем для любого объекта вероятность попасть в выборку одинакова.

Первичная обработка результатов.

Пусть интересующая нас случайная величина X принимает в выборке значение x_1 n_1 раз, $x_2 - n_2$ раз, …, $x_к - n_к$ раз, причем $\sum_{i=1}^{k} n_k = n,$ где n – объем выборки. Тогда наблюдаемые значения случайной величины $x_1, x_2, \ldots, x_к$ называют **вариантами**, а $n_1, n_2, \ldots, n_к$ – **частотами**. Если разделить каждую частоту на объем выборки, то получим **относительные частоты** $w_i = \frac{n_i}{n}$. Последовательность вариант, записанных в порядке возрастания, называют **вариационным** рядом, а перечень вариант и соответствующих им частот или относительных частот – **статистическим рядом**:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
w_i	w_1	w_2	…	w_k

Пример.

При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1,1,4,0,1,2,1,2,2,0,5,3,3,1,0,2,2,3,4,1.Составим вариационный ряд: 0,1,2,3,4,5. Статистический ряд для абсолютных и относительных частот имеет вид:

x_i	0	1	2	3	4	5
n_i	3	6	5	3	2	1
w_i	0,15	0,3	0,25	0,15	0,1	0,05

Если исследуется некоторый непрерывный признак, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку**. Для ее получения интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот вариант,

попавших в i-й интервал. Составленная по этим результатам таблица называется **группированным статистическим рядом**:

Номера интервалов	1	2	…	k
Границы интервалов	$(a,\ a+h)$	$(a+h,\ a+2h)$	…	$(b-h,\ b)$
Сумма частот вариант, попав-ших в интервал	n_1	n_2	…	n_k

Полигон частот. Выборочная функция распределения и гистограмма.

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики. Один из них – **полигон частот**: ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2),…, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i – на оси ординат. Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим **полигон относительных частот** (рис.1)**.**

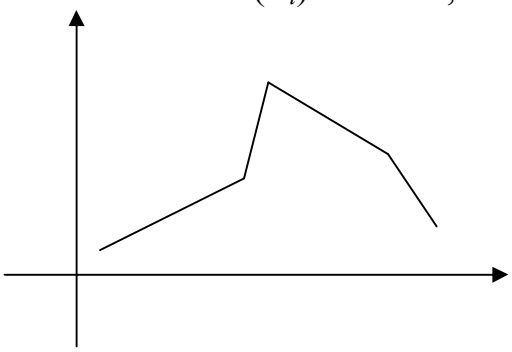


Рис. 1.

По аналогии с функцией распределения случайной величины можно задать некоторую функцию, относительную частоту события $X < x$.

Определение 15.1. **Выборочной (эмпирической) функцией распределения** называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$. Таким образом,

$$F^*(x) = \frac{n_x}{n}, \qquad (15.1)$$

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

Замечание. В отличие от эмпирической функции распределения, найденной опытным путем, функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*. $F(x)$ определяет вероятность события $X < x$, а $F^*(x)$ – его относительную частоту. При достаточно больших n, как следует из теоремы Бернулли, $F^*(x)$ стремится по вероятности к $F(x)$.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами $F(x)$, а именно:

1) $0 \le F^*(x) \le 1$.
2) $F^*(x)$ – неубывающая функция.
3) Если x_1 – наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшая варианта, то $F^*(x) = 1$ при $x > x_к$.

Для непрерывного признака графической иллюстрацией служит **гистограмма**, то есть ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами – отрезки длиной n_i/h (гистограмма частот) или w_i/h (гистограмма относительных частот). В первом случае площадь гистограммы равна объему выборки, во втором – единице (рис.2).

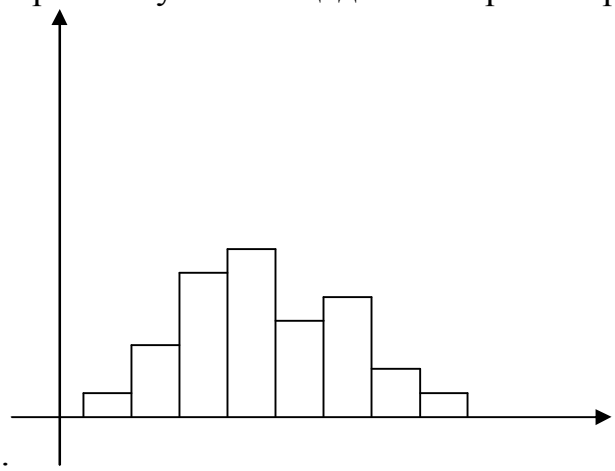


Рис.2.

Лекция10

Числовые характеристики статистического распределения: выборочное среднее, оценки дисперсии, оценки моды и медианы, оценки начальных и центральных моментов. Статистическое описание и вычисление оценок параметров двумерного случайного вектора.

Одна из задач математической статистики: по имеющейся выборке оценить значения числовых характеристик исследуемой случайной величины.

Определение 16.1. **Выборочным средним** называется среднее арифметическое значений случайной величины, принимаемых в выборке:

$$\overline{x}_B = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n} = \frac{\sum_{i=1}^{k} n_i x_i}{n}, \qquad (16.1)$$

где x_i – варианты, n_i - частоты.

Замечание. Выборочное среднее служит для оценки математического ожидания исследуемой случайной величины. В дальнейшем будет рассмотрен вопрос, насколько точной является такая оценка.

Определение 16.2. **Выборочной дисперсией** называется

$$D_B = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_B)^2}{n} = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x}_B)^2}{n}, \qquad (16.2)$$

а **выборочным средним квадратическим отклонением** –

$$\sigma_B = \sqrt{D_B}. \qquad (16.3)$$

Так же, как в теории случайных величин, можно доказать, что справедлива следующая формула для вычисления выборочной дисперсии:

$$D = \overline{x^2} - (\overline{x})^2. \qquad (16.4)$$

Пример 1. Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
n_i	3	8	7	2

$$\bar{x}_B = \frac{2\cdot 3 + 5\cdot 8 + 7\cdot 7 + 8\cdot 2}{20} = 5{,}55; \quad D_B = \frac{4\cdot 3 + 25\cdot 8 + 49\cdot 7 + 64\cdot 2}{20} - 5{,}55^2 = 3{,}3475; \quad \sigma_B = \sqrt{3{,}3475} =$$

Другими характеристиками вариационного ряда являются:

- **мода M_0** – варианта, имеющая наибольшую частоту (в предыдущем примере $M_0 = 5$).
- **медиана m_e** - варианта, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно ($n = 2k + 1$), то $m_e = x_{k+1}$, а при четном $n = 2k$ $m_e = \frac{x_k + x_{k+1}}{2}$. В частности, в примере 1 $m_e = \frac{5+7}{2} = 6$.

Оценки начальных и центральных моментов (так называемые эмпирические моменты) определяются аналогично соответствующим теоретическим моментам:

- **начальным эмпирическим моментом порядка k** называется

$$M_k = \frac{\sum n_i x_i^k}{n}. \qquad (16.5)$$

В частности, $M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_B$, то есть начальный эмпирический момент первого порядка равен выборочному среднему.

- **центральным эмпирическим моментом порядка k** называется

$$m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}. \qquad (16.6)$$

В частности, $m_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_B$, то есть центральный эмпирический момент второго порядка равен выборочной дисперсии.

Лекция 11

Основные свойства статистических характеристик параметров распределения: несмещенность, состоятельность, эффективность. Несмещенность и состоятельность выборочного среднего как оценки математического ожидания. Смещенность выборочной дисперсии. Пример несмещенной оценки дисперсии. Асимптотически несмещенные оценки. Способы построения оценок: метод наибольшего правдоподобия, метод моментов, метод квантили, метод наименьших квадратов, байесовский подход к получению оценок.

Получив статистические оценки параметров распределения (выборочное среднее, выбороч-ную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближе-нием соответствующих характеристик генеральной совокупности. Определим требования, которые должны при этом выполняться.

Пусть Θ^* - статистическая оценка неизвестного параметра Θ теоретического распределения. Извлечем из генеральной совокупности несколько выборок одного и того же объема n и вычислим для каждой из них оценку параметра Θ: $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Тогда оценку Θ^*

можно рассматривать как случайную величину, принимающую возможные значения $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Если математическое ожидание Θ* не равно оцениваемому параметру, мы будем получать при вычислении оценок систематические ошибки одного знака (с избытком, если $M(\Theta^*) > \Theta$, и с недостатком, если $M(\Theta^*) < \Theta$). Следовательно, необходимым условием отсутствия систе-матических ошибок является требование $M(\Theta^*) = \Theta$.

Определение 17.2. Статистическая оценка Θ* называется **несмещенной**, если ее математичес-кое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

$$M(\Theta^*) = \Theta. \qquad (17.1)$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако несмещенность не является достаточным условием хорошего приближения к истинному значению оцениваемого параметра. Если при этом возможные значения Θ* могут значительно отклоняться от среднего значения, то есть дисперсия Θ* велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Определение 17.2. Статистическая оценка называется **эффективной**, если она при заданном объеме выборки *n* имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема к статистическим оценкам предъявляется еще и требование состоятельности.

Определение 17.3. **Состоятельной** называется статистическая оценка, которая при $n \to \infty$ стре-мится по вероятности к оцениваемому параметру (если эта оценка несмещенная, то она будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к 0).

Убедимся, что $\bar{x}_B$ представляет собой несмещенную оценку математического ожидания $M(X)$.

Будем рассматривать $\bar{x}_B$ как случайную величину, а $x_1, x_2, \ldots, x_n$, то есть значения исследуемой случайной величины, составляющие выборку, – как независимые, одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$, имеющие математическое ожидание a. Из свойств математического ожидания следует, что

$$M(\overline{X}_B) = M\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = a.$$

Но, поскольку каждая из величин $X_1, X_2, \ldots, X_n$ имеет такое же распределение, что и генеральная совокупность, $a = M(X)$, то есть $M(\overline{X}_B) = M(X)$, что и требовалось доказать.

Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания. Если предположить, что $X_1, X_2, \ldots, X_n$ имеют ограниченные дисперсии, то из теоремы Чебышева следует, что их среднее арифметическое, то есть $\overline{X}_B$, при увеличении *n* стремится по вероятности к математическому ожиданию *a* каждой их величин, то есть к $M(X)$. Следовательно, выборочное среднее есть состоятельная оценка математического ожидания.

В отличие от выборочного среднего, выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности. Можно доказать, что

$$M(D_B) = \frac{n-1}{n} D_\Gamma , \qquad (17.2)$$

где D_Γ – истинное значение дисперсии генеральной совокупности. Можно предложить другую оценку дисперсии – **исправленную дисперсию s^2**, вычисляемую по формуле

$$s^2 = \frac{n}{n-1} D_B = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1} . \qquad (17.3)$$

Такая оценка будет являться несмещенной. Ей соответствует **исправленное среднее квадратическое отклонение**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2}{n-1}} . \qquad (17.4)$$

Определение 17.4. Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки $x_1, x_2, \ldots, x_n$

$$\lim_{n\to\infty} \frac{x_1 + x_2 + \ldots + x_n}{n} = X , \qquad (17.5)$$

где X – истинное значение исследуемой величины.

Лекция 12

Интервальное оценивание неизвестных параметров. Точность оценки, доверительная вероятность (надежность), доверительный интервал. Построение доверительных интервалов для оценки математического ожидания нормального распределения при известной и при неизвестной дисперсии. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, что приводит к грубым ошибкам. Поэтому в таком случае лучше пользоваться *интервальными оценками*, то есть указывать интервал, в который с заданной вероятностью попадает истинное значение оцениваемого параметра. Разумеется, чем меньше длина этого интервала, тем точнее оценка параметра. Поэтому, если для оценки Θ^* некоторого параметра Θ справедливо неравенство $|\Theta^* - \Theta| < \delta$, число $\delta > 0$ характеризует **точность оценки** (чем меньше δ, тем точнее оценка). Но статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.
Определение 18.1. **Надежностью (доверительной вероятностью)** оценки Θ^* параметра Θ называется вероятность γ того, что выполняется неравенство $|\Theta^* - \Theta| < \delta$. Если заменить это неравенство двойным неравенством $-\delta < \Theta^* - \Theta < \delta$, то получим:

$$p\,(\Theta^* - \delta < \Theta < \Theta^* + \delta) = \gamma .$$

Таким образом, γ есть вероятность того, что Θ попадает в интервал $(\Theta^* - \delta, \Theta^* + \delta)$.
Определение 18.2. **Доверительным** называется интервал, в который попадает неизвестный параметр с заданной надежностью γ.

Построение доверительных интервалов.

1. Доверительный интервал для оценки математического ожидания нормального распределения при известной дисперсии.

Пусть исследуемая случайная величина X распределена по нормальному закону с известным средним квадратическим σ, и требуется по значению выборочного среднего $\bar{x}_B$ оценить ее математическое ожидание a. Будем рассматривать выборочное среднее $\bar{x}_B$ как случайную величину $\bar{X}$, а значения вариант выборки $x_1, x_2, \ldots, x_n$ как одинаково распределенные независимые случайные величины $X_1, X_2, \ldots, X_n$, каждая из которых имеет математическое ожидание a и среднее квадратическое отклонение σ. При этом $M(\bar{X}) = a$, $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ (используем свойства математического ожидания и дисперсии суммы независимых случайных величин). Оценим вероятность выполнения неравенства $|\bar{X} - a| < \delta$. Применим формулу для вероятности попадания нормально распределенной случайной величины в заданный интервал:

$p(|X - a| < \delta) = 2\Phi\left(\frac{\delta}{\sigma}\right)$. Тогда , с учетом того, что $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, $p(|\bar{X} - a| < \delta) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) =$

$=2\Phi(t)$, где $t = \frac{\delta\sqrt{n}}{\sigma}$. Отсюда $\delta = \frac{t\sigma}{\sqrt{n}}$, и предыдущее равенство можно переписать так:

$$p\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}} < a < \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma\,. \qquad (18.1)$$

Итак, значение математического ожидания a с вероятностью (надежностью) γ попадает в интервал $\left(\bar{x}_B - \frac{t\sigma}{\sqrt{n}}; \bar{x}_B + \frac{t\sigma}{\sqrt{n}}\right)$, где значение t определяется из таблиц для функции Лапласа так, чтобы выполнялось равенство $2\Phi(t) = \gamma$.

Пример. Найдем доверительный интервал для математического ожидания нормально распреде-ленной случайной величины, если объем выборки $n = 49$, $\bar{x}_B = 2{,}8$, $\sigma = 1{,}4$, а доверительная вероятность $\gamma = 0{,}9$.

Определим t, при котором $\Phi(t) = 0{,}9:2 = 0{,}45$: $t = 1{,}645$. Тогда

$2{,}8 - \frac{1{,}645 \cdot 1{,}4}{\sqrt{49}} < a < 2{,}8 + \frac{1{,}645 \cdot 1{,}4}{\sqrt{14}}$, или $2{,}471 < a < 3{,}129$. Найден доверительный интервал, в который попадает a с надежностью 0,9.

2. Доверительный интервал для оценки математического ожидания нормального распределения при неизвестной дисперсии.

Если известно, что исследуемая случайная величина X распределена по нормальному закону с неизвестным средним квадратическим отклонением, то для поиска доверительного интервала для ее математического ожидания построим новую случайную величину

$$T = \frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}}, \qquad (18.2)$$

где $\bar{x}_B$ - выборочное среднее, s – исправленная дисперсия, n – объем выборки. Эта случайная величина, возможные значения которой будем обозначать t, имеет распределение Стьюдента (см. лекцию 12) с $k = n - 1$ степенями свободы.

Поскольку плотность распределения Стьюдента $s(t,n) = B_n\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$, где $B_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}$, явным образом не зависит от *a* и σ, можно задать вероятность ее попадания в некоторый интервал (- t_γ , t_γ), учитывая четность плотности распределения, следующим образом: $p\left(\left|\frac{\bar{x}_B - a}{\frac{s}{\sqrt{n}}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} s(t,n)dt = \gamma$. Отсюда получаем:

$$p\left(\bar{x}_B - \frac{t_\gamma s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma s}{\sqrt{n}}\right) = \gamma. \qquad (18.3)$$

Таким образом, получен доверительный интервал для *a*, где t_γ можно найти по соответствую-щей таблице при заданных *n* и γ.

Пример. Пусть объем выборки *n* = 25, $\bar{x}_B$ = 3, *s* = 1,5. Найдем доверительный интервал для *a* при γ = 0,99. Из таблицы находим, что t_γ (*n* = 25, γ = 0,99) = 2,797. Тогда $3 - \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}} < a < 3 + \frac{2{,}797 \cdot 1{,}5}{\sqrt{25}}$, или 2,161< *a* < 3,839 – доверительный интервал, в который попадает *a* с вероятностью 0,99.

3. Доверительные интервалы для оценки среднего квадратического отклонения нормального распределения.

Будем искать для среднего квадратического отклонения нормально распределенной случайной величины доверительный интервал вида (*s* – *δ, s* +*δ*), где *s* – исправленное выборочное среднее квадратическое отклонение, а для δ выполняется условие: *p* (|σ – *s*| < δ) = γ.

Запишем это неравенство в виде: $s\left(1-\frac{\delta}{s}\right) < \sigma < s\left(1+\frac{\delta}{s}\right)$ или, обозначив $q = \frac{\delta}{s}$,

$$s(1-q) < \sigma < s(1+q). \qquad (18.4)$$

Рассмотрим случайную величину χ, определяемую по формуле

$$\chi = \frac{s}{\sigma}\sqrt{n-1},$$

которая распределена по закону «хи-квадрат» с *n*-1 степенями свободы (см. лекцию 12). Плотность ее распределения

$$R(\chi,n) = \frac{\chi^{n-2}e^{-\frac{\chi^2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}$$

не зависит от оцениваемого параметра σ, а зависит только от объема выборки *n*. Преобразуем неравенство (18.4) так, чтобы оно приняло вид $\chi_1 < \chi < \chi_2$. Вероятность выполнения этого неравенства равна доверительной вероятности γ, следовательно, $\int_{\chi_1}^{\chi_2} R(\chi,n)d\chi = \gamma$. Предположим, что *q* < 1, тогда неравенство (18.4) можно записать так:

$$\frac{1}{s(1+q)} < \frac{1}{\sigma} < \frac{1}{s(1-q)},$$

или, после умножения на $s\sqrt{n-1}$, $\frac{\sqrt{n-1}}{1+q} < \frac{s\sqrt{n-1}}{\sigma} < \frac{\sqrt{n-1}}{1-q}$. Следовательно, $\frac{\sqrt{n-1}}{1+q} < \chi < \frac{\sqrt{n-1}}{1-q}$.

Тогда $\int_{\frac{\sqrt{n-1}}{1+q}}^{\frac{\sqrt{n-1}}{1-q}} R(\chi, n)d\chi = \gamma$. Существуют таблицы для распределения «хи-квадрат», из которых можно найти q по заданным n и γ, не решая этого уравнения. Таким образом, вычислив по выборке значение s и определив по таблице значение q, можно найти доверительный интервал (18.4), в который значение σ попадает с заданной вероятностью γ.

Замечание. Если $q > 1$, то с учетом условия $\sigma > 0$ доверительный интервал для σ будет иметь границы

$$0 < \sigma < s(1+q). \qquad (18.5)$$

Пример.
Пусть $n = 20$, $s = 1{,}3$. Найдем доверительный интервал для σ при заданной надежности $\gamma = 0{,}95$. Из соответствующей таблицы находим $q\ (n = 20, \gamma = 0{,}95\ ) = 0{,}37$. Следовательно, границы доверительного интервала: 1,3(1-0,37) = 0,819 и 1,3(1+0,37) = 1,781. Итак, $0{,}819 < \sigma < 1{,}781$ с вероятностью 0,95.

Лекция 13

Статистическая проверка статистических гипотез. Общие принципы проверки гипотез. Понятия статистической гипотезы (простой и сложной), нулевой и конкурирующей гипотезы, ошибок первого и второго рода, уровня значимости, статистического критерия, критической области, области принятия гипотезы. Наблюдаемое значение критерия. Критические точки.

Определение 19.1. **Статистической гипотезой** называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений.

Определение 19.2. **Нулевой (основной)** называют выдвинутую гипотезу H_0.
Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Пример. Пусть H_0 заключается в том, что математическое ожидание генеральной совокупности $a = 3$. Тогда возможные варианты H_1: а) $a \neq 3$; б) $a > 3$; в) $a < 3$.

Определение 19.3. **Простой** называют гипотезу, содержащую только одно предположение, **сложной** – гипотезу, состоящую из конечного или бесконечного числа простых гипотез.

Пример. Для показательного распределения гипотеза H_0: $\lambda = 2$ – простая, H_0: $\lambda > 2$ – сложная, состоящая из бесконечного числа простых (вида $\lambda = c$, где c – любое число, большее 2).

В результате проверки правильности выдвинутой нулевой гипотезы (такая проверка называется **статистической**, так как производится с применением методов математической статистики) возможны ошибки двух видов: **ошибка первого рода**,

состоящая в том, что будет отвергнута правильная нулевая гипотеза, и **ошибка второго рода**, заключаю-щаяся в том, что будет принята неверная гипотеза.
Замечание. Какая из ошибок является на практике более опасной, зависит от конкретной задачи. Например, если проверяется правильность выбора метода лечения больного, то ошибка первого рода означает отказ от правильной методики, что может замедлить лечение, а ошибка второго рода (применение неправильной методики) чревата ухудшением состояния больного и является более опасной.

Определение 19.4. Вероятность ошибки первого рода называется **уровнем значимости α.**

Основной прием проверки статистических гипотез заключается в том, что по имеющейся выборке вычисляется значение некоторой случайной величины, имеющей известный закон распределения.

Определение 19.5. **Статистическим критерием** называется случайная величина K с известным законом распределения, служащая для проверки нулевой гипотезы.

Определение 19.6. **Критической областью** называют область значений критерия, при которых нулевую гипотезу отвергают, **областью принятия гипотезы** – область значений критерия, при которых гипотезу принимают.

Итак, процесс проверки гипотезы состоит из следующих этапов:

1) выбирается статистический критерий K;
2) вычисляется его наблюдаемое значение $K_{набл}$ по имеющейся выборке;
3) поскольку закон распределения K известен, определяется (по известному уровню значимости α) **критическое значение** $k_{кр}$, разделяющее критическую область и область принятия гипотезы (например, если $p(K > k_{кр}) = \alpha$, то справа от $k_{кр}$ располагается критическая область, а слева – область принятия гипотезы);
4) если вычисленное значение $K_{набл}$ попадает в область принятия гипотезы, то нулевая гипотеза принимается, если в критическую область – нулевая гипотеза отвергается.

Различают разные виды критических областей:
- **правостороннюю** критическую область, определяемую неравенством $K > k_{кр}$ ($k_{кр} > 0$);
- **левостороннюю** критическую область, определяемую неравенством $K < k_{кр}$ ($k_{кр} < 0$);
- **двустороннюю** критическую область, определяемую неравенствами $K < k_1$, $K > k_2$ ($k_2 > k_1$).

Определение 19.7. **Мощностью критерия** называют вероятность попадания критерия в критическую область при условии, что верна конкурирующая гипотеза.

Если обозначить вероятность ошибки второго рода (принятия неправильной нулевой гипотезы) β, то мощность критерия равна 1 – β. Следовательно, чем больше мощность критерия, тем меньше вероятность совершить ошибку второго рода. Поэтому после выбора уровня значимости следует строить критическую область так, чтобы мощность критерия была максимальной.

Критерий для проверки гипотезы о вероятности события.

Пусть проведено n независимых испытаний (n – достаточно большое число), в каждом из которых некоторое событие A появляется с одной и той же, но неизвестной вероятностью p, и найдена относительная частота $\frac{m}{n}$ появлений A в этой серии испытаний. Проверим при заданном уровне значимости α нулевую гипотезу H_0, состоящую в том, что вероятность p равна некоторому значению p_0.

Примем в качестве статистического критерия случайную величину

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}, \qquad (19.1)$$

имеющую нормальное распределение с параметрами $M(U) = 0$, $\sigma(U) = 1$ (то есть нормированную). Здесь $q_0 = 1 - p_0$. Вывод о нормальном распределении критерия следует из теоремы Лапласа (при достаточно большом n относительную частоту можно приближенно считать нормально распределенной с математическим ожиданием p и средним квадрати-ческим отклонением $\sqrt{\frac{pq}{n}}$).

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1) Если H_0: $p = p_0$, а H_1: $p \neq p_0$, то критическую область нужно построить так, чтобы вероятность попадания критерия в эту область равнялась заданному уровню значимости α. При этом наибольшая мощность критерия достигается тогда, когда критическая область состоит из двух интервалов, вероятность попадания в каждый из которых равна $\frac{\alpha}{2}$. Поскольку U симметрична относительно оси Oy, вероятность ее попадания в интервалы (-∞; 0) и (0; +∞) равна 0,5, следовательно, критическая область тоже должна быть симметрична относительно Oy. Поэтому $u_{кр}$ определяется по таблице значений функции Лапласа из условия $\Phi(u_{кр}) = \frac{1-\alpha}{2}$, а критическая область имеет вид $(-\infty; -u_{кр}) \cup (u_{кр}; +\infty)$.

Замечание. Предполагается, что используется таблица значений функции Лапласа, заданной в виде $\Phi(x) = \int_0^x e^{-\frac{t^2}{2}} dt$, где нижний предел интегрирования равен 0, а не -∞.

Функция Лапласа, заданная таким образом, является нечетной, а ее значения на 0,5 меньше, чем значения стандартной функции $\Phi(x)$ (см. лекцию 6).

Далее нужно вычислить наблюдаемое значение критерия:

$$U_{набл} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}} \quad . \qquad (19.2)$$

Если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается.
Если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

2) Если конкурирующая гипотеза H_1: $p > p_0$, то критическая область определяется неравенством $U > u_{кр}$, то есть является правосторонней, причем $p(U > u_{кр}) = \alpha$. Тогда $p(0 < U < u_{кр}) = \frac{1}{2} - \alpha = \frac{1-2\alpha}{2}$. Следовательно, $u_{кр}$ можно найти по таблице значений

функции Лапласа из условия, что $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$. Вычислим наблюдаемое значение критерия по формуле (19.2).

Если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается.

Если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

3) Для конкурирующей гипотезы H_1: $p < p_0$ критическая область является левосторонней и задается неравенством $U <$- $u_{кр}$, где $u_{кр}$ вычисляется так же, как в предыдущем случае.

Если $U_{набл} >$ - $u_{кр}$, то нулевая гипотеза принимается.

Если $U_{набл} <$ - $u_{кр}$, то нулевая гипотеза отвергается.

Пример. Пусть проведено 50 независимых испытаний, и относительная частота появления события A оказалась равной 0,12. Проверим при уровне значимости α = 0,01 нулевую гипотезу H_0: $p = 0{,}1$ при конкурирующей гипотезе H_1: $p > 0{,}1$. Найдем $U_{набл} = \frac{(0{,}12-0{,}1)\sqrt{50}}{\sqrt{0{,}1 \cdot 0{,}9}} = 0{,}471.$ Критическая область является правосторонней, а $u_{кр}$ нахо-дим из равенства $\Phi(u_{кр}) = \frac{1-2\cdot 0{,}01}{2} = 0{,}49.$ Из таблицы значений функции Лапласа определяем $u_{кр} = 2{,}33$. Итак, $U_{набл} < u_{кр}$, и гипотеза о том, что $p = 0{,}1$, принимается.

Критерий для проверки гипотезы о математическом ожидании.

Пусть генеральная совокупность X имеет нормальное распределение, и требуется проверить предположение о том, что ее математическое ожидание равно некоторому числу a_0. Рассмотрим две возможности.

1) Известна дисперсия σ^2 генеральной совокупности. Тогда по выборке объема n найдем выборочное среднее $\bar{x}_B$ и проверим нулевую гипотезу H_0: $M(X) = a_0$.

Учитывая, что выборочное среднее $\bar{X}$ является несмещенной оценкой $M(X)$, то есть $M(\bar{X}) = M(X)$, можно записать нулевую гипотезу так: $M(\bar{X}) = a_0$. Для ее проверки выберем критерий

$$U = \frac{\bar{X} - a_0}{\sigma(\bar{X})} = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}. \qquad (19.3)$$

Это случайная величина, имеющая нормальное распределение, причем, если нулевая гипотеза справедлива, то $M(U) = 0$, $\sigma(U) = 1$.

Выберем критическую область в зависимости от вида конкурирующей гипотезы:

- если H_1: $M(\bar{X}) \neq a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \frac{1-\alpha}{2}$, критическая область двусторонняя, $U_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$, и, если $|U_{набл}| < u_{кр}$, то нулевая гипотеза принимается; если $|U_{набл}| > u_{кр}$, то нулевая гипотеза отвергается.

- если H_1: $M(\bar{X}) > a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$, критическая область правосторонняя, и, если $U_{набл} < u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} > u_{кр}$, то нулевая гипотеза отвергается.

- если H_1: $M(\overline{X}) < a_0$, то $u_{кр}$: $\Phi(u_{кр}) = \frac{1-2\alpha}{2}$, критическая область левосторонняя, и, если $U_{набл} > - u_{кр}$, то нулевая гипотеза принимается; если $U_{набл} < - u_{кр}$, то нулевая гипотеза отвергается.

2) Дисперсия генеральной совокупности неизвестна.

В этом случае выберем в качестве критерия случайную величину

$$T = \frac{(\overline{X} - a_0)\sqrt{n}}{S}, \qquad (19.4)$$

где S – исправленное среднее квадратическое отклонение. Такая случайная величина имеет распределение Стьюдента с $k = n - 1$ степенями свободы. Рассмотрим те же, что и в предыдущем случае, конкурирующие гипотезы и соответствующие им критические области. Предварительно вычислим наблюдаемое значение критерия:

$$T_{набл} = \frac{(\overline{x}_B - a_0)\sqrt{n}}{S}. \qquad (19.5)$$

- если H_1: $M(\overline{X}) \neq a_0$, то критическая точка $t_{двуст.кр.}$ находится по таблице критических точек распределения Стьюдента по известным α и $k = n - 1$.

Если $|T_{набл}| < t_{двуст.кр.}$, то нулевая гипотеза принимается.

Если $|T_{набл}| > t_{двуст.кр.}$, то нулевая гипотеза отвергается.

- если H_1: $M(\overline{X}) > a_0$, то по соответствующей таблице находят $t_{правост.кр.}(\alpha, k)$ – критическую точку правосторонней критической области. Нулевая гипотеза принимается, если $T_{набл} < t_{правост.кр.}$.

- при конкурирующей гипотезе H_1: $M(\overline{X}) < a_0$ критическая область является левосторонней, и нулевая гипотеза принимается при условии $T_{набл} > - t_{правост.кр.}$. Если $T_{набл} < - t_{правост.кр.}$., нулевую гипотезу отвергают.

Критерий для проверки гипотезы о сравнении двух дисперсий.

Пусть имеются две нормально распределенные генеральные совокупности X и Y. Из них извлечены независимые выборки объемов соответственно n_1 и n_2, по которым вычислены исправленные выборочные дисперсии s_X^2 и s_Y^2. Требуется при заданном уровне значимос-ти α проверить нулевую гипотезу H_0: $D(X) = D(Y)$ о равенстве дисперсий рассматривае-мых генеральных совокупностей. Учитывая несмещенность исправленных выборочных дисперсий, можно записать нулевую гипотезу так:

$$H_0: M(s_X^2) = M(s_Y^2). \qquad (19.6)$$

Замечание. Конечно, исправленные дисперсии, вычисленные по выборкам, обычно оказываются различными. При проверке гипотезы выясняется, является ли это различие незначимым и обусловленным случайными причинами (в случае принятия нулевой гипотезы) или оно является следствием того, что сами генеральные дисперсии различны.

В качестве критерия примем случайную величину

$$F = \frac{S_\sigma^2}{S_M^2} - \qquad (19.6)$$

- отношение большей выборочной дисперсии к меньшей. Она имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки,

по которой вычислена большая исправленная дисперсия, а n_2 – объем второй выборки.
Рассмотрим два вида конкурирующих гипотез:
- пусть H_1: $D(X) > D(Y)$. Наблюдаемым значением критерия будет отношение большей из исправленных дисперсий к меньшей: $F_{набл} = \frac{s_\sigma^2}{s_M^2}$. По таблице критических точек распределения Фишера-Снедекора можно найти критическую точку $F_{набл}(\alpha; k_1; k_2)$. При $F_{набл} < F_{кр}$ нулевая гипотеза принимается, при $F_{набл} > F_{кр}$ отвергается.
- если H_1: $D(X) \neq D(Y)$, то критическая область является двусторонней и определяется неравенствами $F < F_1$, $F > F_2$, где $p(F < F_1) = p(F > F_2) = \alpha/2$. При этом достаточно найти правую критическую точку $F_2 = F_{кр}(\frac{\alpha}{2}, k_1, k_2)$. Тогда при $F_{набл} < F_{кр}$ нулевая гипотеза принимается, при $F_{набл} > F_{кр}$ отвергается.

Лекция 13

Критерий Пирсона для проверки гипотезы о виде закона распределения случайной величины. Проверка гипотез о нормальном, показательном и равномерном распределениях по критерию Пирсона.

В предыдущей лекции рассматривались гипотезы, в которых закон распределения генеральной совокупности предполагался известным. Теперь займемся проверкой гипотез о предполагаемом законе неизвестного распределения, то есть будем проверять нулевую гипотезу о том, что генеральная совокупность распределена по некоторому известному закону. Обычно статистические критерии для проверки таких гипотез называются **критериями согласия**.

Критерий Пирсона.

Достоинством критерия Пирсона является его универсальность: с его помощью можно проверять гипотезы о различных законах распределения.

1. **Проверка гипотезы о нормальном распределении.**

Пусть получена выборка достаточно большого объема n с большим количеством различных значений вариант. Доя удобства ее обработки разделим интервал от наименьшего до наибольшего из значений вариант на s равных частей и будем считать, что значения вари ант, попавших в каждый интервал, приближенно равны числу, задающему середину интервала. Подсчитав число вариант, попавших в каждый интервал, составим так называе-мую сгруппированную выборку:

варианты……..x_1 x_2 … x_s
частоты……n_1 n_2 … n_s ,

где x_i – значения середин интервалов, а n_i – число вариант, попавших в i-й интервал (эмпи-рические частоты).
По полученным данным можно вычислить выборочное среднее $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B. Проверим предположение, что генеральная совокупность распределена по нормальному закону с параметрами $M(X) = \bar{x}_B$, $D(X) = \sigma_B^2$. Тогда можно найти количество чисел из выборки объема n, которое должно оказаться в каждом интер-

вале при этом предположении (то есть теоретические частоты). Для этого по таблице значений функции Лапласа найдем вероятность попадания в *i*-й интервал:

$$p_i = \Phi\left(\frac{b_i - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{a_i - \bar{x}_B}{\sigma_B}\right),$$

где a_i и b_i - границы *i*-го интервала. Умножив полученные вероятности на объем выборки п, найдем теоретические частоты: $n_i = n \cdot p_i$. Наша цель – сравнить эмпирические и теоретические частоты, которые, конечно, отличаются друг от друга, и выяснить, являются ли эти различия несущественными, не опровергающими гипотезу о нормальном распределении исследуемой случайной величины, или они настолько велики, что противоречат этой гипотезе. Для этого используется критерий в виде случайной величины

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}. \qquad (20.1)$$

Смысл ее очевиден: суммируются части, которые квадраты отклонений эмпирических частот от теоретических составляют от соответствующих теоретических частот. Можно доказать, что вне зависимости от реального закона распределения генеральной совокупности закон распределения случайной величины (20.1) при $n \to \infty$ стремится к закону распределения χ^2 (см. лекцию 12) с числом степеней свободы $k = s - 1 - r$, где r – число параметров предполагаемого распределения, оцененных по данным выборки. Нормальное распределение характеризуется двумя параметрами, поэтому $k = s - 3$. Для выбранного критерия строится правосторонняя критическая область, определяемая условием

$$p(\chi^2 > \chi^2_{kp}(\alpha, k)) = \alpha, \qquad (20.2)$$

где α – уровень значимости. Следовательно, критическая область задается неравенством $\chi^2 > \chi^2_{kp}(\alpha, k)$, а область принятия гипотезы - $\chi^2 < \chi^2_{kp}(\alpha, k)$.

Итак, для проверки нулевой гипотезы H_0: генеральная совокупность распределена нормально – нужно вычислить по выборке наблюдаемое значение критерия:

$$\chi^2_{набл} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'}, \qquad (20.1`)$$

а по таблице критических точек распределения χ^2 найти критическую точку $\chi^2_{кр}(\alpha, k)$, используя известные значения α и $k = s - 3$. Если $\chi^2_{набл} < \chi^2_{kp}$ - нулевую гипотезу принимают, при $\chi^2_{набл} > \chi^2_{kp}$ ее отвергают.

2. **Проверка гипотезы о равномерном распределении**.

При использовании критерия Пирсона для проверки гипотезы о равномерном распределе-нии генеральной совокупности с предполагаемой плотностью вероятности

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

необходимо, вычислив по имеющейся выборке значение $\bar{x}_B$, оценить параметры a и b по формулам:

$$a^* = \bar{x}_B - \sqrt{3}\sigma_B, \quad b^* = \bar{x}_B + \sqrt{3}\sigma_B, \qquad (20.3)$$

где a^* и b^* - оценки a и b. Действительно, для равномерного распределения $M(X) = \frac{a+b}{2}$, $\sigma(x) = \sqrt{D(X)} = \sqrt{\frac{(a-b)^2}{12}} = \frac{a-b}{2\sqrt{3}}$, откуда можно получить систему для определения a^* и b^*: $\begin{cases} \frac{b^*+a^*}{2} = \bar{x}_B \\ \frac{b^*-a^*}{2\sqrt{3}} = \sigma_B \end{cases}$, решением которой являются выражения (20.3).

Затем, предполагая, что $f(x) = \frac{1}{b^*-a^*}$, можно найти теоретические частоты по формулам

$$n_1' = np_1 = nf(x)(x_1 - a^*) = n \cdot \frac{1}{b^*-a^*}(x_1 - a^*);$$

$$n_2' = n_3' = ... = n_{s-1}' = n \cdot \frac{1}{b^*-a^*}(x_i - x_{i-1}),\ \ i = 1,2,...,s-1;$$

$$n_s' = n \cdot \frac{1}{b^*-a^*}(b^* - x_{s-1}).$$

Здесь s – число интервалов, на которые разбита выборка.

Наблюдаемое значение критерия Пирсона вычисляется по формуле (20.1`), а критическое – по таблице с учетом того, что число степеней свободы $k = s - 3$. После этого границы критической области определяются так же, как и для проверки гипотезы о нормальном распределении.

3. **Проверка гипотезы о показательном распределении.**

В этом случае, разбив имеющуюся выборку на равные по длине интервалы, рассмотрим последовательность вариант $x_i^* = \frac{x_i + x_{i+1}}{2}$, равноотстоящих друг от друга (считаем, что все варианты, попавшие в i – й интервал, принимают значение, совпадающее с его серединой), и соответствующих им частот n_i (число вариант выборки, попавших в i – й интервал). Вычислим по этим данным $\bar{x}_B$ и примем в качестве оценки параметра λ величину $\lambda^* = \frac{1}{\bar{x}_B}$. Тогда теоретические частоты вычисляются по формуле

$$n_i' = n_i p_i = n_i p(x_i < X < x_{i+1}) = n_i(e^{-\lambda x_i} - e^{-\lambda x_{i+1}}).$$

Затем сравниваются наблюдаемое и критическое значение критерия Пирсона с учетом того, что число степеней свободы $k = s - 2$.

Лекция 14
Корреляционный анализ.

Проверка гипотезы о значимости выборочного коэффициента корреляции.

Рассмотрим выборку объема n, извлеченную из нормально распределенной двумерной генеральной совокупности (X, Y). Вычислим выборочный коэффициент корреляции r_B. Пусть он оказался не равным нулю. Это еще не означает, что и коэффициент корреляции генеральной совокупности не равен нулю. Поэтому при заданном уровне значимости α возникает необходимость проверки нулевой гипотезы H_0:

$r_г = 0$ о равенстве нулю гене-рального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$. Таким образом, при принятии нулевой гипотезы X и Y некоррелированы, то есть не связаны линейной зависимостью, а при отклонении H_0 они коррелированы.

В качестве критерия примем случайную величину

$$T = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}}, \qquad (21.1)$$

которая при справедливости нулевой гипотезы имеет распределение Стьюдента (см. лекцию 12) с $k = n - 2$ степенями свободы. Из вида конкурирующей гипотезы следует, что критическая область двусторонняя с границами $\pm t_{кр}$, где значение $t_{кр}(\alpha, k)$ находится из таблиц для двусторонней критической области.

Вычислив наблюдаемое значение критерия

$$T_{набл} = \frac{r_B\sqrt{n-2}}{\sqrt{1-r_B^2}}$$

и сравнив его с $t_{кр}$, делаем вывод:

- если $|T_{набл}| < t_{кр}$ – нулевая гипотеза принимается (корреляции нет);
- если $|T_{набл}| > t_{кр}$ – нулевая гипотеза отвергается (корреляция есть).

Ранговая корреляция.

Пусть объекты генеральной совокупности обладают двумя качественными признаками (то есть признаками, которые невозможно измерить точно, но которые позволяют сравнивать объекты между собой и располагать их в порядке убывания или возрастания качества). Договоримся для определенности располагать объекты в порядке ухудшения качества.

Пусть выборка объема n содержит независимые объекты, обладающие двумя качественными признаками: A и B. Требуется выяснить степень их связи между собой, то есть установить наличие или отсутствие **ранговой корреляции**.

Расположим объекты выборки в порядке ухудшения качества по признаку A, предполагая, что все они имеют различное качество по обоим признакам. Назовем место, занимаемое в этом ряду некоторым объектом, его **рангом x_i:** $x_1 = 1, x_2 = 2,\ldots, x_n = n$.

Теперь расположим объекты в порядке ухудшения качества по признаку B**,** присвоив им ранги y_i , где номер i равен порядковому номеру объекта по признаку A, а само значение ранга равно порядковому номеру объекта по признаку B. Таким образом, получены две последовательности рангов:

по признаку A … $x_1, x_2,\ldots, x_n$

по признаку B … $y_1, y_2,\ldots, y_n$.

При этом, если, например, $y_3 = 6$, то это означает, что данный объект занимает в ряду по признаку A третье место, а в ряду по признаку B – шестое.

Сравним полученные последовательности рангов.

1. Если $x_i = y_i$ при всех значениях i, то ухудшение качества по признаку A влечет за собой ухудшение качества по признаку B, то есть имеется «полная ранговая зависимость».
2. Если ранги противоположны, то есть $x_1 = 1, y_1 = n$; $x_2 = 2, y_2 = n - 1;\ldots, x_n = n, y_n = 1$, то признаки тоже связаны: ухудшение качества по одному из них приводит к улучшению качества по другому («противоположная зависимость»).
3. На практике чаще всего встречается промежуточный случай, когда ряд y_i не монотонен. Для оценки связи между признаками будем считать ранги $x_1, x_2,\ldots, x_n$

возможными значениями случайной величины X, а $y_1, y_2, \ldots, y_n$ – возможными значениями случайной величины Y. Теперь можно исследовать связь между X и Y, вычислив для них выборочный коэффициент корреляции

$$r_B = \frac{\sum n_{uv} uv - n\bar{u}\bar{v}}{n\sigma_u \sigma_v}, \qquad (21.2)$$

где $u_i = x_i - \bar{x},\ v_i = y_i - \bar{y}$ (условные варианты). Поскольку каждому рангу x_i соответствует только одно значение y_i, то частота любой пары условных вариант с одинаковыми индексами равна 1, а с разными индексами – нулю. Кроме того, из выбора условных вариант следует, что $\bar{u} = \bar{v} = 0$, поэтому формула (21.2) приобретает более простой вид:

$$r_B = \frac{\sum u_i v_i}{n\sigma_u \sigma_v}. \qquad (21.3)$$

Итак, требуется найти $\sum u_i v_i,\ \sigma_u$ и σ_v.

Можно показать, что $\sum u_i^2 = \sum v_i^2 = \frac{n^3 - n}{12}$. Учитывая, что $\bar{x} = \bar{y}$, можно выразить $\sum u_i v_i$ через разности рангов $d_i = x_i - y_i = u_i - v_i$. После преобразований получим: $\sum u_i v_i = \frac{n^3 - n}{12} - \sum \frac{d_i^2}{2}$, $\sigma_u = \sigma_v = \sqrt{\frac{n^2 - 1}{12}}$, откуда $n\sigma_u \sigma_v = \frac{n^3 - n}{12}$. Подставив эти результаты в (21.3), получим **выборочный коэффициент ранговой корреляции Спирмена**:

$$\rho_B = 1 - \frac{6\sum d_i^2}{n^3 - n}. \qquad (21.4)$$

Свойства выборочного коэффициента корреляции Спирмена.

1. Если между A и B имеется «полная прямая зависимость», то есть ранги совпадают при всех i, то $\rho_B = 1$. Действительно, при этом $d_i = 0$, и из формулы (21.4) следует справедливость свойства 1.
2. Если между A и B имеется «противоположная зависимость», то $\rho_B = -1$. В этом случае, преобразуя $d_i = (2i - 1) - n$, найдем, что $\sum d_i^2 = \frac{n^3 - n}{3}$, тогда из (21.4)

   $$\rho_B = 1 - \frac{6(n^3 - n)}{3(n^3 - n)} = 1 - 2 = -1.$$
3. В остальных случаях $-1 < \rho_B < 1$, причем зависимость между A и B тем меньше, чем ближе $|\rho_B|$ к нулю.

Итак, требуется при заданном уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции Спирмена $\rho_г$ при конку-рирующей гипотезе H_1: $\rho_г \neq 0$. Для этого найдем критическую точку:

$$T_{кр} = t_{кр}(\alpha, k)\sqrt{\frac{1 - \rho_B^2}{n - 2}}, \qquad (21.5)$$

где n – объем выборки, ρ_B – выборочный коэффициент ранговой корреляции Спирмена, $t_{кр}$ (α, k) – критическая точка двусторонней критической области, найденная по таблице критических точек распределения Стьюдента, число степеней свободы $k = n - 2$.
Тогда, если $|\rho_B| < T_{кр}$, то нулевая гипотеза принимается, то есть ранговая корреляционная связь между признаками незначима.

Если $|\rho_B| > T_{кр}$, то нулевая гипотеза отвергается, и между признаками существует значимая ранговая корреляционная связь.

Можно использовать и другой коэффициент – коэффициент ранговой корреляции Кендалла. Рассмотрим ряд рангов $y_1, y_2, \ldots, y_n$, введенный так же, как и ранее, и зададим величины R_i следующим образом: пусть правее y_1 имеется R_1 рангов, больших y_1; правее y_2 – R_2 рангов, больших y_2 и т.д. Тогда, если обозначить $R = R_1 + R_2 + \ldots + R_{n-1}$, то **выборочный коэффициент ранговой корреляции Кендалла** определяется формулой

$$\tau_B = \frac{4R}{n(n-1)} - 1, \qquad (21.6)$$

где n – объем выборки.

Замечание. Легко убедиться, что коэффициент Кендалла обладает теми же свойствами, что и коэффициент Спирмена.

Для проверки нулевой гипотезы H_0: $\tau_г = 0$ (генеральный коэффициент ранговой корреляции Кендалла равен нулю) при альтернативной гипотезе H_1: $\tau_г \neq 0$ необходимо найти критическую точку:

$$T_{кр} = z_{kp}\sqrt{\frac{2(2n+5)}{9n(n-1)}}, \qquad (21.7)$$

где n – объем выборки, а $z_{кр}$ – критическая точка двусторонней критической области, определяемая из условия $\Phi(z_{kp}) = \frac{1-\alpha}{2}$ по таблицам для функции Лапласа.

Если $|\tau_B| < T_{кр}$, то нулевая гипотеза принимается (ранговая корреляционная связь между признаками незначима).

Если $|\tau_B| > T_{кр}$, то нулевая гипотеза отвергается (между признаками существует значимая ранговая корреляционная связь).

Лекция 15
Регрессионный анализ.

Рассмотрим выборку двумерной случайной величины (X, Y). Примем в качестве оценок условных математических ожиданий компонент их условные средние значения, а именно: **условным средним** $\bar{y}_x$ назовем среднее арифметическое наблюдавшихся значений Y, соответствующих $X = x$. Аналогично **условное среднее** $\bar{x}_y$ **-** среднее арифметическое наблюдавшихся значений X, соответствующих $Y = y$. В лекции 11 были выведены уравнения регрессии Y на X и X на Y:

$$M(Y/x) = f(x), \quad M(X/y) = \varphi(y).$$

Условные средние $\bar{y}_x$ и $\bar{x}_y$ являются оценками условных математических ожиданий и, следовательно, тоже функциями от x и y, то есть

$$\bar{y}_x = f^*(x) - \qquad (22.1)$$

- выборочное уравнение регрессии *Y* на *X*,

$$\bar{x}_y = \varphi^*(y) - \qquad (22.2)$$

- выборочное уравнение регрессии *X* на *Y*.

Соответственно функции $f^*(x)$ и $\varphi^*(y)$ называются **выборочной регрессией *Y* на *X* и *X* на *Y***, а их графики – **выборочными линиями регрессии**. Выясним, как определять параметры выборочных уравнений регрессии, если сам вид этих уравнений известен.

Пусть изучается двумерная случайная величина (X, Y), и получена выборка из n пар чисел (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Будем искать параметры прямой линии среднеквадратической регрессии Y на X вида

$$Y = \rho_{yx}x + b\ , \qquad (22.3)$$

Подбирая параметры ρ_{yx} и b так, чтобы точки на плоскости с координатами (x_1, y_1), (x_2, y_2), …, (x_n, y_n) лежали как можно ближе к прямой (22.3). Используем для этого метод наименьших квадратов и найдем минимум функции

$$F(\rho,b) = \sum_{i=1}^{n}(Y_i - y_i)^2 = \sum_{i=1}^{n}(\rho x_i + b - y_i)^2\ . \qquad (22.4)$$

Приравняем нулю соответствующие частные производные:

$$\begin{cases} \dfrac{\partial F}{\partial \rho} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0 \\ \dfrac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0 \end{cases}.$$

В результате получим систему двух линейных уравнений относительно ρ и b:

$$\begin{cases} \left(\sum x^2\right)\rho + \left(\sum x\right)b = \sum xy \\ \left(\sum x\right)\rho + nb = \sum y \end{cases}. \qquad (22.5)$$

Ее решение позволяет найти искомые параметры в виде:

$$\rho_{xy} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2};\ \ b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n\sum x^2 - \left(\sum x\right)^2}\ . \qquad (22.6)$$

При этом предполагалось, что все значения X и Y наблюдались по одному разу.
Теперь рассмотрим случай, когда имеется достаточно большая выборка (не менее 50 значений), и данные сгруппированы в виде *корреляционной таблицы*:

Y	X				
	x_1	x_2	…	x_k	n_y
y_1	n_{11}	n_{21}	…	n_{k1}	$n_{11}+n_{21}+\ldots+n_{k1}$
y_2	n_{12}	n_{22}	…	n_{k2}	$n_{12}+n_{22}+\ldots+n_{k2}$
…	…	…	…	…	……..
y_m	n_{1m}	n_{2m}	…	n_{km}	$n_{1m}+n_{2m}+\ldots+n_{km}$
n_x	$n_{11}+n_{12}+\ldots+n_{1m}$	$n_{21}+n_{22}+\ldots+n_{2m}$	…	$n_{k1}+n_{k2}+\ldots+n_{km}$	$n=\sum n_x = \sum n_y$

Здесь n_{ij} – число появлений в выборке пары чисел (x_i, y_j).

Поскольку $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$, $\overline{x^2} = \frac{\sum x^2}{n}$, заменим в системе (22.5) $\sum x = n\bar{x}$, $\sum y = n\bar{y}$, $\sum x^2 = n\overline{x^2}$, $\sum xy = \sum n_{xy} xy$, где n_{xy} – число появлений пары чисел (x, y). Тогда система (22.5) примет вид:

$$\begin{cases} (n\overline{x^2})\rho_{yx} + (n\bar{x})b = \sum n_{xy} xy \\ (\bar{x})\rho_{yx} + b = \bar{y} \end{cases} . \qquad (22.7)$$

Можно решить эту систему и найти параметры ρ_{yx} и b, определяющие выборочное уравнение прямой линии регрессии:

$$\bar{y}_x = \rho_{yx}\bar{x} + b .$$

Но чаще уравнение регрессии записывают в ином виде, вводя **выборочный коэффициент корреляции**. Выразим b из второго уравнения системы (22.7):

$$b = \bar{y} - \rho_{yx}\bar{x} .$$

Подставим это выражение в уравнение регрессии: $\bar{y}_x - \bar{y} = \rho_{yx}(x - \bar{x})$. Из (22.7)

$$\rho_{yx} = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n(\overline{x^2} - (\bar{x})^2)} = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x^2} , \qquad (22.8)$$

где $\tilde{\sigma}_x^2 = \overline{x^2} - (\bar{x})^2$. Введем понятие **выборочного коэффициента корреляции**

$$r_B = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x \tilde{\sigma}_y}$$

и умножим равенство (22.8) на $\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}$: $\rho_{yx}\frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = r_B$, откуда $\rho_{yx} = r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}$. Используя это соотношение, получим выборочное уравнение прямой линии регрессии Y на X вида

$$\bar{y}_x - \bar{y} = r_B\frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \bar{x}) . \qquad (22.9)$$

Лекция 16
Однофакторный дисперсионный анализ.

Пусть генеральные совокупности $X_1, X_2, \ldots, X_p$ распределены нормально и имеют одинаковую дисперсию, значение которой неизвестно. Найдем выборочные средние по выборкам из этих генеральных совокупностей и проверим при заданном уровне значимости нулевую гипотезу H_0: $M(X_1) = M(X_2) = \ldots = M(X_p)$ о равенстве всех математических ожиданий. Для решения этой задачи применяется метод, основанный на сравнении дисперсий и названный поэтому **дисперсионным анализом**.

Будем считать, что на случайную величину X воздействует некоторый качественный фактор F, имеющий p уровней: $F_1, F_2, \ldots, F_p$. Требуется сравнить «факторную дисперсию», то есть рассеяние, порождаемое изменением уровня фактора, и «остаточную дисперсию», обусловленную случайными причинами. Если их различие значимо, то фактор существенно влияет на X и при изменении его уровня групповые средние различаются значимо.

Будем считать, что количество наблюдений на каждом уровне фактора одинаково и равно q. Оформим результаты наблюдений в виде таблицы:

Номер испытания	Уровни фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповое среднее	$\bar{x}_{гр1}$	$\bar{x}_{гр2}$	…	$\bar{x}_{грp}$

Определим общую, факторную и остаточную суммы квадратов отклонений от среднего:

$$S_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x})^2 \text{ -} \qquad (23.1)$$

- общая сумма квадратов отклонений наблюдаемых значений от общего среднего $\bar{x}$;

$$S_{факт} = q\sum_{j=1}^{p}(\bar{x}_{грj} - \bar{x})^2 \text{ -} \qquad (23.2)$$

- факторная сумма отклонений групповых средних от общей средней, характеризующая рассеяние между группами;

$$S_{ост} = \sum_{i=1}^{q}(x_{i1} - \bar{x}_{гр1})^2 + \sum_{i=1}^{q}(x_{i2} - \bar{x}_{гр2})^2 + ... + \sum_{i=1}^{q}(x_{ip} - \bar{x}_{грp})^2 \text{ -} \qquad (23.3)$$

- остаточная сумма квадратов отклонений наблюдаемых значений группы от своего группового среднего, характеризующая рассеяние внутри групп.

Замечание. Остаточную сумму можно найти из равенства

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{факт}} .$$

Вводя обозначения $R_j = \sum_{i=1}^{q} x_{ij}$, $P_j = \sum_{i=1}^{q} x_{ij}^2$, получим формулы, более удобные для расчетов:

$$S_{общ} = \sum_{j=1}^{p} P_j - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq}, \qquad (23.1`)$$

$$S_{факт} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{\left(\sum_{j=1}^{p} R_j\right)^2}{pq} . \qquad (23.2`)$$

Разделив суммы квадратов на соответствующее число степеней свободы, получим общую, факторную и остаточную дисперсии:

$$s^2_{общ} = \frac{S_{общ}}{pq-1}, \; s^2_{факт} = \frac{S_{факт}}{p-1}, \; s^2_{ост} = \frac{S_{ост}}{p(q-1)} . \qquad (23.4)$$

Если справедлива гипотеза H_0, то все эти дисперсии являются несмещенными оценками генеральной дисперсии. Покажем, что проверка нулевой гипотезы сводится к сравнению факторной и остаточной дисперсии по критерию Фишера-Снедекора (см. лекцию 12).

1. Пусть гипотеза H_0 правильна. Тогда факторная и остаточная дисперсии являются несмещенными оценками неизвестной генеральной дисперсии и, следовательно, различаются незначимо. Поэтому результат оценки по критерию Фишера-Снедекора F покажет, что нулевая гипотеза принимается. Таким образом, если верна гипотеза о равенстве математических ожиданий генеральных совокупностей, то верна и гипотеза о равенстве факторной и остаточной дисперсий.

2. Если нулевая гипотеза неверна, то с возрастанием расхождения между математическими ожиданиями увеличивается и факторная дисперсия, а вместе с ней и отношение $F_{набл} = \frac{s^2_{факт}}{s^2_{ост}}$. Поэтому в результате $F_{\text{набл}}$ окажется больше $F_{\text{кр}}$, и гипотеза о равенстве дисперсий будет отвергнута. Следовательно, если гипотеза о равенстве математических ожиданий генеральных совокупностей ложна, то ложна и гипотеза о равенстве факторной и остаточной дисперсий.

Итак, метод дисперсионного анализа состоит в *проверке по критерию F нулевой гипотезы о равенстве факторной и остаточной дисперсий.*

Замечание. Если факторная дисперсия окажется меньше остаточной, то гипотеза о равенстве математических ожиданий генеральных совокупностей верна. При этом нет необходимости использовать критерий *F*.

Если число испытаний на разных уровнях различно (q_1 испытаний на уровне F_1, q_2 – на уровне F_2, …, q_p - на уровне F_p), то

$$S_{общ} = (P_1 + P_2 + ... + P_p) - (R_1 + R_2 + ... + R_p),$$

где $P_j = \sum_{i=1}^{q_j} x_{ij}^2$ – сумма квадратов наблюдавшихся значений признака на уровне F_j,

$R_j = \sum_{i=1}^{q_j} x_{ij}$ – сумма наблюдавшихся значений признака на уровне F_j. При этом объем выборки, или общее число испытаний, равен $n = q_1 + q_2 + ... + q_p$.

Факторная сумма квадратов отклонений вычисляется по формуле

$$S_{факт} = \left(\frac{R_1^2}{q_1} + \frac{R_2^2}{q_2} + ... + \frac{R_p^2}{q_p} \right) - \frac{(R_1 + R_2 + ... + R_p)^2}{n}.$$

Остальные вычисления проводятся так же, как в случае одинакового числа испытаний:

$$S_{ост} = S_{общ} - S_{факт},\ s^2_{факт} = \frac{S_{факт}}{p-1},\ s^2_{ост} = \frac{S_{ост}}{n-p}.$$

Лекция 17

Моделирование случайных величин методом Монте-Карло (статистических испытаний).

Задачу, для решения которой применяется метод Монте-Карло, можно сформулировать так: требуется найти значение *a* изучаемой случайной величины. Для его определения выбирается случайная величина *X*, математическое ожидание которой равно *a*, и для выборки из *n* значений *X*, полученных в *n* испытаниях, вычисляется выборочное среднее:

$$\bar{x} = \frac{\sum x_i}{n},$$

которое принимается в качестве оценки искомого числа *a*:

$$a \approx a^* = \bar{x}.$$

Этот метод требует проведения большого числа испытаний, поэтому его иначе называют **методом статистических испытаний**. Теория метода Монте-Карло исследует,

как наиболее целесообразно выбрать случайную величину X, как найти ее возможные значения, как уменьшить дисперсию используемых случайных величин, чтобы погрешность при замене a на a^* была возможно меньшей.

Поиск возможных значений X называют **разыгрыванием случайной величины**. Рассмотрим некоторые способы разыгрывания случайных величин и выясним, как оценить допускаемую при этом ошибку.

Оценка погрешности метода Монте-Карло.

Если поставить задачу определения верхней границы допускаемой ошибки с заданной доверительной вероятностью γ, то есть поиска числа δ, для которого

$$p(|\overline{X} - a| \le \delta) = \gamma,$$

то получим известную задачу определения доверительного интервала для математического ожидания генеральной совокупности (см. лекцию 18). Воспользуемся результатами решения этой задачи для следующих случаев:

1) случайная величины X распределена нормально и известно ее среднее квадратическое отклонение. Тогда из формулы (18.1) получаем: $\delta = \frac{t\delta}{\sqrt{n}}$, где n – число испытаний, σ - известное среднее квадратическое отклонение, а t – аргумент функции Лапласа, при котором $\Phi(t) = \gamma/2$.
2) Случайная величина X распределена нормально с неизвестным σ. Воспользуемся формулой (18.3), из которой следует, что $\delta = \frac{t_\gamma s}{\sqrt{n}}$, где s – исправленное выборочное среднее квадратическое отклонение, а t_γ определяется по соответствующей таблице.
3) Если случайная величина распределена по иному закону, то при достаточно большом количестве испытаний ($n > 30$) можно использовать для оценки δ предыдущие формулы, так как при $n \to \infty$ распределение Стьюдента стремится к нормальному, и границы интервалов, полученные по формулам (18.1) и (18.3), различаются незначительно.

Разыгрывание случайных величин.

Определение 24.1. **Случайными числами** называют возможные значения r непрерывной случайной величины R, распределенной равномерно в интервале (0; 1).

1. Разыгрывание дискретной случайной величины.

Пусть требуется разыграть дискретную случайную величину X, то есть получить последовательность ее возможных значений, зная закон распределения X:

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n .

Рассмотрим равномерно распределенную в (0, 1) случайную величину R и разобьем интервал (0, 1) точками с координатами $p_1,\ p_1 + p_2,\ \ldots, p_1 + p_2 + \ldots + p_{n-1}$ на n частичных интервалов $\Delta_1, \Delta_2, \ldots, \Delta_n$, длины которых равны вероятностям с теми же индексами.

Теорема 24.1. Если каждому случайному числу $r_j (0 \le r_j < 1)$, которое попало в интервал Δ_i, ставить в соответствие возможное значение x_i, то разыгрываемая величина будет иметь заданный закон распределения:

X x_1 x_2 … x_n
p p_1 p_2 … p_n .

Доказательство.

Возможные значения полученной случайной величины совпадают с множеством x_1 , x_2 ,… x_n, так как число интервалов равно n, а при попадании r_j в интервал Δ_i случайная величина может принимать только одно из значений x_1 , x_2 ,… x_n.

Так как R распределена равномерно, то вероятность ее попадания в каждый интервал равна его длине, откуда следует, что каждому значению x_i соответствует вероятность p_i. Таким образом, разыгрываемая случайная величина имеет заданный закон распределения.

Пример. Разыграть 10 значений дискретной случайной величины X, закон распределения которой имеет вид:

X	2	3	6	8
p	0,1	0,3	0,5	0,1

Решение. Разобьем интервал (0, 1) на частичные интервалы: Δ_1- (0; 0,1), Δ_2 – (0,1; 0,4), Δ_3 - (0,4; 0,9), Δ_4 – (0,9; 1). Выпишем из таблицы случайных чисел 10 чисел: 0,09; 0,73; 0,25; 0,33; 0,76; 0,52; 0,01; 0,35; 0,86; 0,34. Первое и седьмое числа лежат на интервале Δ_1, следовательно, в этих случаях разыгрываемая случайная величина приняла значение x_1 = 2; третье, четвертое, восьмое и десятое числа попали в интервал Δ_2, что соответствует x_2 = 3; второе, пятое, шестое и девятое числа оказались в интервале Δ_3 – при этом $X = x_3 = 6$; на последний интервал не попало ни одного числа. Итак, разыгранные возможные значения X таковы: 2, 6, 3, 3, 6, 6, 2, 3, 6, 3.

2. Разыгрывание противоположных событий.

Пусть требуется разыграть испытания, в каждом из которых событие A появляется с известной вероятностью p. Рассмотрим дискретную случайную величину X, принимающую значения 1 (в случае, если событие A произошло) с вероятностью p и 0 (если A не произошло) с вероятностью $q = 1 – p$. Затем разыграем эту случайную величину так, как было предложено в предыдущем пункте.

Пример. Разыграть 10 испытаний, в каждом из которых событие A появляется с вероятностью 0,3.

Решение. Для случайной величины X с законом распределения

X	1	0
p	0,3	0,7

получим интервалы Δ_1 – (0; 0,3) и Δ_2 – (0,3; 1). Используем ту же выборку случайных чисел, что и в предыдущем примере, для которой в интервал Δ_1 попадают числа №№1,3 и 7, а остальные – в интервал Δ_2. Следовательно, можно считать, что событие A произошло в первом, третьем и седьмом испытаниях, а в остальных – не произошло.

3. Разыгрывание полной группы событий.

Если события A_1, A_2, …, A_n, вероятности которых равны p_1 , p_2 ,… p_n, образуют полную группу, то для из разыгрывания (то есть моделирования последовательности их появлений в серии испытаний) можно разыграть дискретную случайную величину X с законом распределения X 1 2 … n, сделав это так же, как в пункте 1. При этом считаем, что

p p_1 p_2 … p_n

если X принимает значение $x_i = i$, то в данном испытании произошло событие A_i.

4. Разыгрывание непрерывной случайной величины.

а) Метод обратных функций.

Пусть требуется разыграть непрерывную случайную величину X, то есть получить последовательность ее возможных значений x_i ($i = 1, 2, \ldots, n$), зная функцию распределения $F(x)$.

Теорема 24.2. Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F(x)$, соответствующее r_i , является корнем уравнения

$$F(x_i) = r_i. \qquad (24.1)$$

Доказательство.

Так как $F(x)$ монотонно возрастает в интервале от 0 до 1, то найдется (причем единственное) значение аргумента x_i , при котором функция распределения примет значение r_i . Значит, уравнение (24.1) имеет единственное решение: $x_i = F^{-1}(r_i)$, где F^{-1}-функция, обратная к F. Докажем, что корень уравнения (24.1) является возможным значением рассматриваемой случайной величины X. Предположим вначале, что x_i – возможное значение некоторой случайной величины ξ, и докажем, что вероятность попадания ξ в интервал (c, d) равна $F(d) - F(c)$. Действительно, $c < x_i < d \Leftrightarrow F(c) < r_i < F(d)$ в силу монотонности $F(x)$ и того, что $F(x_i) = r_i$. Тогда

$c < \xi < d \Leftrightarrow F(c) < R < F(d)$, следовательно, $p(c < \xi < d) = p(F(c) < R < F(d)) = F(d) - F(c).$

Значит, вероятность попадания ξ в интервал (c, d) равна приращению функции распределения $F(x)$ на этом интервале, следовательно, $\xi = X$.

Пример.

Разыграть 3 возможных значения непрерывной случайной величины X, распределенной равномерно в интервале (5; 8).

Решение.

$F(x) = \frac{x-5}{3}$, то есть требуется решить уравнение $\frac{x_i - 5}{3} = r_i,\ x_i = 3r_i + 5.$ Выберем 3 случайных числа: 0,23; 0,09 и 0,56 и подставим их в это уравнение. Получим соответствующие возможные значения X: $x_1 = 5{,}69;\ x_2 = 5{,}27;\ x_3 = 6{,}68.$

б) Метод суперпозиции.

Если функция распределения разыгрываемой случайной величины может быть представлена в виде линейной комбинации двух функций распределения:

$$F(x) = C_1F_1(x) + C_2F_2(x)\ \ (C_{1,2} > 0), \qquad (24.2)$$

то $C_1 + C_2 = 1$, так как при $x \to \infty$ $F(x) \to 1$.

Введем вспомогательную дискретную случайную величину Z с законом распределения

Z	1	2
p	C_1	C_2

. Выберем 2 независимых случайных числа r_1 и r_2 и разыграем возможное значение Z по числу r_1 (см. пункт 1). Если $Z = 1$, то ищем искомое возможное значение X из уравнения $F_1(x) = r_2$, а если $Z = 2$, то решаем уравнение $F_2(x) = r_2$.

Можно доказать, что при этом функция распределения разыгрываемой случайной величины равна заданной функции распределения.

в) Приближенное разыгрывание нормальной случайной величины.

Так как для R, равномерно распределенной в (0, 1), $M(R) = \frac{1}{2},\ D(R) = \frac{1}{12}$, то для суммы n независимых, равномерно распределенных в интервале (0,1) случайных величин

$\sum_{j=1}^{n} R_j$ $M\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{2}$, $D\left(\sum_{j=1}^{n} R_j\right) = \frac{n}{12}$, $\sigma = \sqrt{\frac{n}{12}}$. Тогда в силу центральной предельной теоремы нормированная случайная величина $\frac{\sum_{j=1}^{n} R_j - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$ при $n \to \infty$ будет иметь распределение, близкое к нормальному, с параметрами $a = 0$ и $\sigma = 1$. В частности, достаточно хорошее приближение получается при $n = 12$: $\sum_{j=1}^{12} R_j - 6$.

Итак, чтобы разыграть возможное значение нормированной нормальной случайной величины x, надо сложить 12 независимых случайных чисел и из суммы вычесть 6.