Министерство сельского хозяйства

Российской Федерации

Технологический институт- филиал ФГБОУ ВО Ульяновский ГАУ

Н.В. Шевченко

Элементы теории вероятностей и математической статистики

краткий курс лекций

Димитровград -2021

Шевченко, Н.В., Элементы теории вероятностей и математической статистики: краткий курс лекций/ Н.В. Шевченко. – Димитровград: Технологический институт – филиал УлГАУ, 2021. - 102 с.

Рецензенты: Шигапов Ильяс Исхакович, доктор технических наук, профессор, заведующий кафедрой «Технологии производства переработки и экспертизы продукции АПК» ФГБОУ Ульяновский ГАУ им. П.А. Столыпина

Джабраилов Тайяр Акбер Оглы - кандидат физико-математических наук, доцент кафедры «Математика и физика» ФГБОУ Ульяновский ГАУ им. П.А. Столыпина

Элементы теории вероятностей и математической статистики: курс лекций предназначен для подготовки бакалавров очной и заочной форм обучения.

Утверждено
на заседании кафедры «Эксплуатация мобильных машин и социально-гуманитарных дисциплин» технологического института –филиала УлГАУ
Протокол № 10 от 11 мая, 2021 г.

ОГЛАВЛЕНИЕ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

Основные понятия

Наблюдаемые нами события (явления) можно подразделить на следующие три вида: достоверное, невозможное и случайное.

Достоверным называют событие, которое обязательно произойдет, если будет осуществлена определенная совокупность условий *S*. Например, если в сосуде содержится вода при нормальном атмосферном давлении и температуре 20°, то событие «вода в сосуде находится в жидком состоянии» есть достоверное. В этом примере заданное атмосферное давление и температура воды составляют совокупность условий *S*.

Невозможным называют событие, которое заведомо не произойдет, если будет осуществлена совокупность условий *S*. Например, событие «вода в сосуде находится в твердом состоянии» заведомо не произойдет при осуществлении условий *S*.

Случайным называется событие, которое при осуществлении совокупности условий *S* может либо произойти, либо не произойти. Например, если брошена монета, то она может упасть так, что сверху будет либо «герб», либо «решетка». Каждое случайное событие, в частности выпадение «герба», есть следствие действия очень многих причин, которые также являются случайными. Поэтому теория вероятностей не ставит задачу предсказать, произойдет единственное событие или нет, но при многократном повторении одного испытания в одних и тех же условиях, можно установить определенные закономерности. Установление этих закономерностей и занимается теория вероятностей.

Предметом теории вероятностей является изучение вероятностных закономерностей однородных случайных массовых событий.

Виды случайных событий

События называют несовместными, если появление одного из них исключает появление других событий в одном и том же испытании. Несколько событий образуют полную группу, если в результате испытания появится хотя

бы одно из них. Если события, образующие полную группу попарно несовместны, то в результате испытания появится одно и только одно из этих событий.

События называют равновозможными, если есть основание считать, что ни одно из них не является более возможным, чем другое.

Классическое определение вероятности

Каждый из возможных результатов испытания называется элементарным исходом. Элементарные исходы обозначим $\omega_1,\omega_2,...\omega_n$. Те элементарные исходы, в которых интересующее нас событие *A* наступает, называются благоприятствующими этому событию. Таким образом, событие *A* наступает, если результатом испытания является один безразлично какой из элементарных исходов, благоприятствующих событию *A*.

Отношение числа благоприятствующих событию *A* элементарных исходов к их общему числу называют вероятностью события *A* и обозначают *P*(*A*). Итак:

$$P(A)=\frac{m}{n}$$

Здесь предполагается, что элементарные исходы:

1. несовместны;
2. равновозможны;
3. образуют полную группу.

Из определения вероятности вытекают следующие ее свойства.

1. Вероятность достоверного события равна единице. Действительно, если событие достоверно, то каждый элементарный исход испытания благоприятствует событию. В этом случае *m* = *n*, следовательно

$$P(A)=\frac{m}{n}=\frac{n}{n}=1$$

2. Вероятность невозможного события равно нулю. Если событие невозможно, то ни один из элементарных исходов испытания не благоприятствует событию. В этом случае $m = 0$, следовательно

$$P(A) = \frac{m}{n} = \frac{0}{n} = 0$$

3. Вероятность случайного события есть положительное число, заключенное между нулем и единицей. В этом случае $0 < m < n$, таким образом: $0 < P(A) = \frac{m}{n} < 1.$

Пример 1. Монета бросается два раза. Какова вероятность: 1)выпадения герба хотя бы один раз (событие *A*); 2) двукратного выпадения герба (событие *B*)?

Равновозможными элементарными исходами здесь являются: *ГГ*, *ГР*, *РГ*, *РР*; число их $n = 4$. Событию *A* благоприятствуют: *ГГ*, *ГР*, *РГ*, число их $m = 3$. Следовательно, $P(A) = \frac{m}{n} = \frac{3}{4}$.

Событию *B* благоприятствует один исход: *ГГ* ($m'=1$). Поэтому $P(B) = \frac{m'}{n} = \frac{1}{4}$.

Пример 2. Игральная кость бросается два раза. Какова вероятность того, что сумма выпавших очков равна 6 (событие *A*)?

Равновозможными элементарными исходами здесь являются пары (*x*, *y*), где *x*, *y* принимают значения 1, 2, 3, 4, 5, 6, общее число элементарных исходов $n = 36$. Событию *A* благоприятствуют пары (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) число которых $m = 5$. Следовательно $P(A) = \frac{m}{n} = \frac{5}{36}$.

Элементы комбинаторики

Рассмотрим совокупность *n* различных элементов a_1, a_2, … a_n. Произвольную упорядоченную выборку из этих элементов называют соединением.

Определение 1. Размещениями из *n* элементов по *m* ($m \le n$) называют их соединения, каждое из которых содержит ровно *m* различных элементов (выбранных из данных элементов) и которые отличаются либо самими элемента-

ми, либо порядком элементов. Определим число A_n^m размещений из n элементов $a_1, a_2,\ldots a_n$ по m. Сначала определим $a_{\alpha 1}$ -первый элемент размещения. Очевидно, из данной совокупности n элементов его можно выбрать n различными способами. После выбора первого элемента $a_{\alpha 1}$ для второго элемента $a_{\alpha 2}$ остается $(n - 1)$ способ выбора и т.д., так как каждый такой выбор дает новое размещение, то все эти выборы можно комбинировать между собой. Поэтому имеем:

$$A_n^m = n(n-1)\ldots[n-(m-1)].$$

Вводя в обозначение факториала $n! = 1\cdot 2\cdot 3\cdot\ldots\cdot n$, получим: $A_n^m = \dfrac{n!}{(n-m)!}$

Определение 2. Соединение из n элементов, каждое из которых содержит все n элементов, называются перестановками. Число перестановок из n элементов обозначим:

$$P_n = A_n^n = n(n-1)(n-2)\ldots[n-(n-1)] = n!$$

Определение 3. Сочетаниями из n элементов по m называют такие соединения, каждое из которых содержит ровно m данных элементов и которые отличаются хотя бы одним элементом. (Порядок элементов не имеет значения)

Обозначим через C_n^m число сочетаний из n элементов по m. Рассмотрим все допустимые сочетания наших элементов $a_{\alpha 1}, a_{\alpha 2}, \ldots a_{\alpha m}$. Делая в каждом из них $m!$ возможных перестановок их элементов, получим все размещения из n элементов по m. Таким образом, имеем формулу $C_n^m \cdot m! = A_n^m$. Отсюда

$$C_n^m = \frac{A_n^m}{m!} = \frac{n(n-1)\ldots[n-(m-1)]}{m!}, \qquad C_n^m = \frac{n!}{m!(n-m)!}, \qquad C_n^m = C_n^{n-m},$$

если $m = 0$, то $C_n^0 = 1$.

Замечание. При решении задач комбинаторики используют следующие правила:

Правило суммы. Если некоторый объект A может быть выбран из совокупности объектов m способами, а другой объект B может быть выбран n способами, то выбрать либо A либо B можно $n + m$ способами.

Правило произведения. Если объект A можно выбрать из совокупности объектов m способами, а объект B можно выбрать n способами, то пара объектов (A, B) может быть выбрана $m \cdot n$ способами.

Относительная частота. Устойчивость относительной частоты

Если определение множества Ω пространства элементарных исходов испытания затруднено, тогда вероятность события определяют как относительную частоту появления события A в n испытаниях $\mu_n(A) = \frac{m}{n}$.

Относительная частота обладает свойством устойчивости, колеблясь около некоторого постоянного числа. Таким образом, это постоянное число и есть вероятность появления события A: $P(A) = W(A)$.

Пример. На данной территории в течение 10 лет произошло 14 сильных и 55 слабых землетрясений. Найти относительную частоту появления сильных землетрясений.

Число всех землетрясений $n = 14 + 55 = 69$, число наступления события А (сильных землетрясений) $m = 14$.

$$W(A) = \frac{14}{69} = 0{,}2029.$$

Теорема сложения вероятностей

Суммой двух событий называют событие, состоящее в появлении события A или события B, или обоих этих событий. Например, если из орудия произведено два выстрела, A – попадание при первом выстреле, B – попадание при втором выстреле, тогда $A + B$ – попадание при первом выстреле, или при втором выстреле, или в обоих выстрелах. Произведением двух событий A и B называют событие $A \cdot B$, состоящее в совместном появлении этих событий.

Теорема. Вероятность суммы двух событий равна сумме вероятностей этих событий минус вероятность произведения этих событий:

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

Доказательство. Пусть n – число элементарных исходов опыта, m_1 – число элементарных исходов опыта, благоприятствующих событию A, m_2 – число элементарных исходов опыта, благоприятствующих событию B, m_3 – число элементарных исходов опыта, благоприятствующих одновременному появлению событию $A \cdot B$.

Тогда $m_1 + m_2 - m_3$ – число элементарных исходов опыта благоприятствующих появлению события $A + B$.

Значит,

$$P(A+B) = \frac{m_1 + m_2 - m_3}{n} = \frac{m_1}{n} + \frac{m_2}{n} - \frac{m_3}{n} = P(A) + P(B) - P(A \cdot B).$$

Если события A и B несовместны, т.е. одновременно произойти не могут, тогда $m_3 = 0$:

$$P(A \cdot B) = \frac{m_3}{n} = 0$$

и формула упрощается: $P(A + B) = P(A) + P(B)$.

Противоположные события

Противоположными называют два единственно возможных события, образующих полную группу. Если одно обозначено A, тогда другое принято обозначать $\overline{A}$.

Теорема. Вероятность противоположного события $P(\overline{A}) = 1 - P(A)$.

Доказательство. Пусть n – число элементарных исходов опыта, m – число элементарных исходов опыта, благоприятствующих событию A, тогда остальные $(n - m)$ исходов благоприятствующих событию $\overline{A}$. Таким образом:

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

Теорема умножения вероятностей

Определение 1. Произведением двух событий A и B называют событие $A \cdot B$, состоящее в одновременном появлении этих событий.

Случайное событие мы определили как событие, которое происходит или не происходит при осуществлении определенного комплекса условий S. Если при вычислении вероятности события никаких других условий кроме S не налагается, то вероятность события называется безусловной.

Определение 2. Условной вероятностью $P_A(B)$ называют вероятность события B, вычисленную в предположении, что событие A уже наступило.

Теорема умножения вероятностей. Вероятность совместного действия появления двух событий равна произведению вероятностей одного из них на условную вероятность другого, вычисленную в предположении, что первое уже наступило:

$$P(A \cdot B) = P(A) \cdot P_A(B).$$

Доказательство. Пусть n – число элементарных исходов опыта, m – число элементарных исходов опыта, благоприятствующих событию A, k – число элементарных исходов опыта, благоприятствующих событию $A \cdot B$. Тогда

$$P(A \cdot B) = \frac{k}{n} = \frac{k}{n} \cdot \frac{m}{m} = \frac{m}{n} \cdot \frac{k}{m} = P(A) P_A(B)$$

События A и B называются независимыми, если

$$P_A(B) = P(B) \text{ и } P_B(A) = P(A).$$

Если событие A и B независимы, то $P(A \cdot B) = P(A) \cdot P(B)$.

Замечание. Легко доказать, что: $P(A \cdot B) = P(B) \cdot P_B(A)$. Таким образом:

$$P(A) \cdot P_A(B) = P(B) \cdot P_B(A).$$

Вероятность появления хотя бы одного события

Пусть в результате испытания могут появиться n событий, независимых в совокупности, либо некоторые из них (в частности, только одно или ни одного), причем вероятности появления каждого из этих событий известны.

Теорема. Вероятность появления хотя бы одного из событий $A_1, A_2, \ldots A_n$, независимых в совокупности, равно разности между единицей и произведением вероятностей противоположных событий: $P(A) = 1 - P(\overline{A}_1)P(\overline{A}_2)\ldots P(\overline{A}_n)$.

Доказательство следует из того факта, что события A и $\overline{A}_1\overline{A}_2\ldots\overline{A}_n$ противоположны.

Пример 1. В урне находятся 2 белых, 3 красных и 5 синих одинаковых по размеру шаров. Какова вероятность, что шар случайным образом извлеченный из урны будет цветным (не белым)?

Пусть A – извлечение из урны красного шара, B – извлечение из урны синего шара. Тогда $A + B$ – извлечение из урны цветного шара:

$$P(A) = \frac{3}{10}, \quad P(B) = \frac{5}{10}$$

и

$$P(A + B) = P(A) + P(B) = \frac{3}{10} + \frac{5}{10} = \frac{8}{10}.$$

Пример 2. Вероятность поражения цели первым стрелком (событие A) равна 0,9, а вероятность поражения вторым стрелком (событие B) равна 0,8. Какова вероятность того, что цель будет поражена хотя бы одним стрелком (событие C)?

Очевидно событие $\overline{C}$ - оба стрелка промахнулись:

$$\overline{C} = \overline{A} \cdot \overline{B},$$

$P(\overline{C}) = P(\overline{A})P(\overline{B})$ = [1 - $P(A)$][1 - $P(B)$] = (1 – 0,9)(1 – 0,8) = 0,1·0,2 = 0,02.

Отсюда $P(C) = 1 - P(\overline{C}) = 1 - 0{,}02 = 0{,}98$.

Формула полной вероятности

Пусть событие A может наступить при условии появления одного из несовместных событий $H_1, H_2, \ldots, H_n$, которые образуют полную группу. Пусть известны вероятности этих событий и условные вероятности: $P_{H_i}(A)$, где $i = \overline{1,n}$.

Теорема. Вероятность события A, которое может произойти лишь при условии появления одного из несовместных событий $H_1, H_2, \ldots, H_n$, образующих полную группу, вычисляется по формуле:

$$P(A) = P(H_1)\,P_{H_1}(A) + P(H_2)\,P_{H_2}(A)\ldots + P(H_n)\,P_{H_n}(A).$$

Доказательство. Рассмотрим сумму событий $H_1 + H_2 \ldots + H_n = D$, где D – достоверное событие. Это следует из того факта, что $H_1, H_2, \ldots, H_n$ – полная группа событий. Тогда

$$A \cdot D = A \cdot (H_1 + H_2 \ldots + H_n) = A \cdot H_1 + A \cdot H_2 \ldots + A \cdot H_n.$$

Но произведение событий $A \cdot D = A$ в силу того, что событие D достоверное и наступление события $A \cdot D$ зависит только от появления события A. Таким образом:

$$P(A) = P(A \cdot D) = P(A \cdot H_1 + A \cdot H_2 \ldots + A \cdot H_n) = P(A \cdot H_1) + P(A \cdot H_2) \ldots + P(A \cdot H_n),$$

тогда как события $H_1 + H_2, \ldots + H_n$ попарно несовместны, тогда и события $A \cdot H_1$, $A \cdot H_2 \ldots$, $A \cdot H_n$ также попарно несовместны. Вычислим:

$$P(A \cdot H_i) = P(H_i) \cdot P_{H_i}(A).$$

Следовательно $P(A) = \sum P(H_i) P_{H_i}(A)$. События $H_1, H_2, \ldots, H_n$ называют гипотезами.

Формула Байеса

Рассмотрим ту же самую модель. Вероятность события определим по формуле полной вероятности. Допустим, что произведено испытание, в резуль-

тате которого появилось событие A. Будем искать условные вероятности $P_A(H_1)$, …, $P_A(H_n)$.

По теореме умножения вероятностей имеем:

$$P(A \cdot H_i) = P(A) \cdot P_A(H_i) = P(H_i) \cdot P_{H_i}(A).$$

Отсюда $$P_A(H_i) = \frac{P(H_i) P_{H_i}(A)}{P(A)},$$

заменив $P(A)$ по формуле полной вероятности получим

$$P_A(H_i) = \frac{P(H_i) P_{H_i}(A)}{\sum_{k=1}^{n} P(H_k) P_{H_k}(A)}, \; i = \overline{1,n}.$$

Полученные формулы называют формулами Байеса.

Пример 1. Керн скважины 1 представлен тремя ящиками, в каждом из которых четыре отделения, в трех отделениях лежат алевролиты, в одном – песчаники.

Керн со скважины 2 представлен пятью ящиками, в каждом из которых шесть отделений, в четырех отделениях лежат алевролиты, в двух – песчаники.

Керн со скважины 3 представлен двумя ящиками, в каждом из которых имеется, пять отделений, содержащих только алевролиты.

Для анализа отобран наугад керн из одного отделения одного ящика. Определить вероятность того, что проба представлена алевролитами.

Решение. События H_1, H_2, H_3 заключающиеся в том, что проба взята из скважины 1, 2, 3, соответственно, образуют полную группу. Событие A состоит в том, что проба представлена алевролитами только при наступлении H_1 или H_2 или H_3:

$$P_{H_1}(A) = \frac{3}{4}, \; P_{H_2}(A) = \frac{4}{6}, \; P_{H_3}(A) = 1, \; P(H_1) = \frac{3}{10}, \; P(H_2) = \frac{5}{10}, \; P(H_3) = \frac{2}{10}.$$

Полученные значения подставим в формулу полной вероятности:

$$P(A) = P(H_1)P_{H_1}(A) + P(H_2)P_{H_2}(A) + P(H_3)P_{H_3}(A) =$$
$$= \frac{3}{10}\cdot\frac{3}{4} + \frac{5}{10}\cdot\frac{4}{6} + \frac{2}{10}\cdot 1 = 0{,}7583 \quad .$$

Пример 2. На геологической карте данный район разбит на 25 равных по площади участков, в том числе 6 несмежных участков распространения юрских отложений. Точки для бурения двух скважин выбирают наугад поочередно, но так, чтобы на один и тот же участок не попали обе скважины. Определить вероятность того, что точка для бурения второй скважины попадет на участок юрских отложений, если точка для первой скважины попала на участок распространения юрских отложений.

Решение. A – событие, состоящее в том, что точка для бурения первой скважины попала на один из участков распространения юрских отложений,

B – событие, состоящее в том, что точка для бурения второй скважины попала на другой из участков отложений этого же возраста,

$A\cdot B$ – событие, состоящее в одновременном появлении событий A и B:

$$P(A\cdot B) = \frac{C_6^2}{C_{25}^2} = \frac{1}{20}, \quad P(A) = \frac{6}{25}.$$

Вероятность события B, если A уже произошло $P_A(B)$, т.е. если точка бурения первой скважины уже попала на участок юрских отложений, вычисляется по формуле умножения вероятностей: $P_A(B) = \dfrac{P(A\cdot B)}{P(A)} = \dfrac{\frac{1}{20}}{\frac{6}{25}} = \dfrac{5}{24}$.

Схема Бернулли

Событие A называется независимым в данной системе испытаний, если вероятность этого события в каждом из них не зависит от исхода других испытаний. Серия независимых повторных испытаний, в каждом из которых данное событие A имеет одну и ту же вероятность $P(A) = p$ не зависящую от номера испытания, называется схемой Бернулли. Таким образом, в схеме Бернулли для каждого испытания имеются только два исхода:

1) событие A - «успех»: $P(A) = p$;

2) событие $\overline{A}$ - «неудача»: $P(\overline{A}) = 1 - p = q$.

Рассмотрим задачу. Найти вероятность того, что при n независимых испытаниях событие A появится ровно m раз $P_n(m)$. Благоприятные серии испытаний здесь имеют вид: $A_{\alpha 1}, A_{\alpha 2}, \ldots A_{\alpha n}$, где $A_{\alpha i} = A$ или $\overline{A}$ ($i = 1, 2, \ldots n$), причем событие A встречается ровно m раз, а событие $\overline{A}$ ровно $n - m$ раз. Так как испытания независимы, то вероятность реализации одной такой благоприятной серии равна $p^m \cdot q^{n-m}$.

Все благоприятные серии получаются в результате выбора различных m номеров испытаний из общего количества n номеров и, следовательно, число их равно C_n^m. Отсюда, применяя теорему сложения вероятностей для случая несовместных событий, для вероятности появления события A точно m раз при n испытаниях получим формулу Бернулли:

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

Эта формула также называется биноминальной, так как её правая часть представляет собой $(m + 1)$ член бинома Ньютона:

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

Локальная теорема Лапласа

Если число испытаний n велико, то вычисления становятся затруднительными. Лаплас получил важную приближенную формулу для вероятностей $P_n(m)$, если n большое число.

Теорема. Пусть $p = P(A)$ – вероятность события A. Тогда вероятность того, что в условиях схемы Бернулли событие A при n независимых испытаниях появится точно m раз, выражается приближенной формулой Лапласа:

$$P_n(m) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{t^2}{2}}, \text{ где } q = 1-p,\ t = \frac{m - np}{\sqrt{npq}}.$$

Интегральная теорема Лапласа

Поставлен вопрос: какова вероятность $P_n(m_1, m_2)$ того, что в n независимых испытаниях событие A появится не менее m_1 раз и не более m_2 раза?

На основании теоремы сложения вероятностей для несовместных событий получим:

$$P_n(m_1, m_2) = \sum_{m=m_1}^{m_2} P_n(m).$$

Отсюда, используя локальную теорему Лапласа, приближенно будем иметь:

$$P_n(m_1, m_2) = \sum_{m=m_1}^{m_2} \frac{1}{\sqrt{npq}} \varphi_0(t_m),$$

где

$$t_m = \frac{m - np}{\sqrt{npq}} \ (m_1 < m < m_2) \text{ и } \varphi_0(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Обозначим $\Delta t_m = t_{m+1} - t_m = \frac{m+1-np}{\sqrt{npq}} - \frac{m-np}{\sqrt{npq}} = \frac{1}{\sqrt{npq}}.$

Следовательно,

$$P_n(m_1, m_2) \approx \sum_{m=m_1}^{m_2} \varphi_0(t_m) \Delta t_m.$$

Полученная сумма является интегральной для функции $\varphi_0(t)$ на отрезке $t_{m1} < t < t_{m2}$. При $n \to \infty$ т.е. при $\Delta t_m \to 0$ её предел есть определенный интеграл. Поэтому, считая n достаточно большим, получим формулу:

$$P_n(m_1, m_2) \approx \int_{tm_1}^{tm_2} \varphi_0(t) dt = \frac{1}{\sqrt{2\pi}} \int_{tm_1}^{tm_2} e^{-\frac{t^2}{2}} dt,$$

где

$$t_{m1} = \frac{m_1 - np}{\sqrt{npq}}, \qquad t_{m2} = \frac{m_2 - np}{\sqrt{npq}}.$$

Введем интеграл вероятностей

$$\Phi(x) \approx \int_0^x \varphi_0(t)dt = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt,$$

тогда на основании формулы Ньютона-Лейбница, будем иметь:

$$P_n(m_1,m_2) \approx \Phi(t_{m2}) - \Phi(t_{m1}).$$

Функция $\Phi(x)$ обладает следующими свойствами:

1) $\Phi(0) = 0$

2) $\Phi(+\infty) = 1/2$

3) $\Phi(-x) = -\Phi(x)$.

При $x > 3$ с точностью до тысячных можно принять $\Phi(x) = 0{,}500$.

Теорема Пуассона

Пусть производится серия n независимых испытаний ($n = 1, 2, 3\ldots$), причем вероятность появления данного события A в этой серии $p_n = P(A) > 0$ зависит от её номера n и стремится к нулю при $n \to \infty$ (последовательность «редких событий»).

Предположим, что для каждой серии среднее значение числа появлений события A постоянно, т.е. $p_n = \mu = \text{const}$. Отсюда $p_n = \frac{\mu}{n}$.

На основании биноминальной формулы для вероятности появления события A в n-ой серии равно m раз имеем:

$$P_n(m) = C_n^m p_n^m (1-p_n)^{n-m} = C_n^m \left(\frac{\mu}{n}\right)^m \left(1-\frac{\mu}{n}\right)^{n-m}$$

Пусть m фиксировано и $n \to \infty$. Тогда

$$C_n^m \left(\frac{\mu}{n}\right)^m = \frac{n(n-1)(n-2)\ldots[n-(m-1)]}{m!n^m}\mu^m =$$

$$= \frac{\mu^m}{m!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{m-1}{n}\right) \to \frac{\mu^m}{m!}$$

Кроме того, используя второй замечательный предел, получим:

$$\left(1-\frac{\mu}{n}\right)^{n-m}=\left[\left(1-\frac{\mu}{n}\right)^{\frac{n}{\mu}}\right]^{\mu}\left(1-\frac{\mu}{n}\right)^{-m}\to e^{-\mu}\cdot 1=e^{-\mu},\ \text{если } n\to\infty.$$

Таким образом:

$$\lim_{n\to\infty}P_n(m)=\lim_{n\to\infty}C_n^m\left(\frac{\mu}{n}\right)^n\cdot\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^{n-m}=\frac{\mu^m}{m!}e^{-\mu}.$$

Если n велико, то вероятность $P_n(m)$ сколь угодно мало отличается от своего предела. Отсюда, при больших n для искомой вероятности $P_n(m)$ имеем приближенную формулу Пуассона: $P_n(m)\approx\frac{\mu^m}{m!}e^{-\mu}$.

Формулу Пуассона можно применять в случаях, когда число испытаний n «велико», вероятность события $p_n = p$ «мала».

Пример 1. Найти вероятность того, что при десятикратном бросании монеты герб выпадает ровно пять раз. Пусть A - выпадение герба при однократном бросании монеты

$$P(A)=\frac{1}{2},\quad P_{10}(5)=C_{10}^5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{10-5}=\frac{10!}{5!\cdot 5!}\left(\frac{1}{2}\right)^{10}=\frac{252}{1024}\approx 0{,}25.$$

Пример 2. Вероятность поражения цели стрелком при одиночном выстреле p = 0,2. Какова вероятность того, что при 100 выстрелах цель будет поражена ровно 20 раз?

Здесь $p = 0{,}2$; $q = 0{,}8$; $n = 100$; $m = 20$; $\sqrt{npq}=\sqrt{100\cdot 0{,}2\cdot 0{,}8}=4$,

$$t=\frac{m-np}{\sqrt{npq}}=\frac{20-100\cdot 0{,}2}{4}=0;\ \varphi_0(0)=\frac{1}{\sqrt{2\pi}}e^0\approx 0{,}4,$$

$$P_{100}(20)\approx 0{,}40\cdot\frac{1}{4}=0{,}10.$$

Естественно, что вероятность такого события мала, так как это событие достаточно редкое. Например, вероятность события $\{15 < m < 25\}$, включающего 11 значений m близка к единице: $P\{15<m<25\}=\sum_{k=15}^{25}P_{100}(k)$

Пример 3. При выработке некоторой массовой продукции вероятность появления одного нестандартного изделия составляет 0,01. Какова вероятность того, что в партии из 100 изделий этой продукции *2* изделия будет нестандартным?

Здесь $p = 0{,}01$ мало, а число $n = 100$ велико, причем $\mu = n \cdot p = 100 \cdot 0{,}01 = 1$.

Используя закон Пуассона $P_{100}(2) \approx \frac{\mu^2}{2!} e^{-\mu} = \frac{1}{2} e^{-1} \approx 0{,}184$

Случайная величина

Величина называется случайной, если она принимает свои значения в зависимости от исходов некоторого испытания (опыта), причем для каждого элементарного исхода она имеет единственное значение. Случайная величина называется дискретной, если множество всех возможных значений её конечно (или счётно). Геометрически множество всех возможных значений дискретной случайной величины представляет конечную систему точек числовой оси. пусть X – дискретная случайная величина, возможными и единственно возможными значениями которой являются числа x_1, x_2, ...x_n. Обозначим через $p_i = P\{X = x_i\}; i = \overline{1,n}$. События $X = x_i$ $(i = \overline{1,n})$, очевидно, образуют полную группу событий, поэтому $p_1 + p_2 + \ldots + p_n = 1$.

Определение. Соответствие между всеми возможными значениями дискретной случайной величины и их вероятностями называется законом распределения данной случайной величины.

Математическое ожидание дискретной случайной величины

Математическим ожиданием дискретной случайной величины называют сумму произведений всех её возможных значений на их вероятности. Пусть случайная величина имеет закон распределения:

X	x_1	x_2	x_3	……	x_n
P	p_1	p_2	p_3	……	p_n

Тогда $M(X) = \sum_{i=1}^{n} x_i p_i$.

Свойства математического ожидания

1. Математическое ожидание постоянной величины равно самой постоянной: $M(C) = C$. Будем рассматривать постоянную величину как дискретную случайную величину, которая имеет одно возможное значение C с вероятностью единица.

Следовательно $M(C) = C \cdot 1 = C$.

2. Постоянный множитель можно выносить за знак математического ожидания: $M(CX) = CM(X)$.

Пусть случайная величина задана законом распределения вероятностей

X	x_1	x_2	x_3	……	x_n
P	p_1	p_2	p_3	……	p_n

Тогда $C \cdot X$ будет иметь закон:

CX	Cx_1	Cx_2	Cx_3	……	Cx_n
P	p_1	p_2	p_3	……	p_n

$$M(CX) = C \cdot x_1 \cdot p_1 + C \cdot x_2 \cdot p_2 + \ldots + C \cdot x_n \cdot p_n = C\,[x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_n \cdot p_n] = C \cdot M(X).$$

Определение. Случайные величины называются независимыми, если закон распределения одной из них не зависит от того, какие возможные значения приняла другая величина. В противном случае случайные величины называются зависимыми. Несколько случайных величин называются взаимно независимыми, если законы распределения любого числа из них не зависят от того, какие возможные значения приняли остальные величины.

3. Математическое ожидание суммы двух случайных величин равно сумме математических ожиданий слагаемых: $M(X + Y) = M(X) + M(Y)$.

Доказательство. Под суммой случайных величин $X + Y$ понимается случайная величина Z, значениями которой являются допустимые суммы $z_{ij} = x_i + y_j$, где x_i, y_j все возможные значения соответствующих случайных величин X и Y, причем

$$\pi_{ij} = P\{Z = z_{ij}\} = P\{X = x_i\}P\{Y = y_j / X = x_i\}.$$

Таким образом

$$M(Z) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\pi_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j)\pi_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i\pi_{ij} + y_j\pi_{ij}) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i\pi_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j\pi_{ij} = \sum_{i=1}^{n} x_i \sum_{j=1}^{m}\pi_{ij} + \sum_{j=1}^{n} y_j \sum_{i=1}^{n}\pi_{ij}\,.$$

Рассмотрим сумму $\sum_{j=1}^{m}\pi_{ij} = \sum_{j=1}^{m} P\{X = x_i / Y = y_j\} = P\{X = x_i\} = p_i$

$$\sum_{i=1}^{n}\pi_{ij} = \sum_{o=1}^{n} P\{X = x_i / Y = y_j\} = P\{Y = y_j\} = p_j^*\,.$$

Таким образом: $M(X+Y) = \sum_{i=1}^{n} x_i p_i + \sum_{j=1}^{m} y_j p_j^* = M(X) + M(Y)$.

Следствие. Если C = const, то $M(X+C) = M(X) + C$.

4. Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:

$$M(X\cdot Y) = M(X)\cdot M(Y).$$

Пусть независимые случайные величины X и Y заданы своим законами распределения вероятностей:

X	x_1	x_2
P	p_1	p_2

Y	y_1	y_2
G	g_1	g_2

Тогда закон распределения случайной величины $X\cdot Y$ задается следующим образом:

$X\cdot Y$	$x_1 y_1$	$x_2 \cdot y_1$	$x_1 \cdot y_2$	$x_2 \cdot y_2$
P	$p_1 \cdot g_1$	$p_2 \cdot g_1$	$p_1 \cdot g_2$	$p_2 \cdot g_2$

$$M(X\cdot Y) = x_1 \cdot y_1 \cdot p_1 \cdot g_1 + x_2 \cdot y_1 \cdot p_2 \cdot g_1 + x_1 \cdot y_2 \cdot p_1 \cdot g_2 + x_2 \cdot y_2 \cdot p_2 \cdot g_2 =$$

$$= y_1 \cdot g_1 \cdot (x_1 \cdot p_1 + x_2 \cdot p_2) + y_2 \cdot g_2 \cdot (x_1 \cdot p_1 + x_2 \cdot p_2) =$$

$$= (x_1 \cdot p_1 + x_2 \cdot p_2) \cdot (y_1 \cdot g_1 + y_2 \cdot g_2) = M(X) \cdot M(Y)$$

Следствие. Математическое ожидание произведения нескольких независимых случайных величин равно произведению математических ожиданий этих величин.

Дисперсия дискретной случайной величины

Пусть X и Y – случайные величины, которые имеют одинаковые математические ожидания, но различные возможные значения:

X	-0,01	0,01
P	0,5	0,5

Y	-100	100
G	0,5	0,5

$M(X) = -0{,}5 \cdot 0{,}01 + 0{,}5 \cdot 0{,}01 = 0$; $M(Y) = -100 \cdot 0{,}5 + 100 \cdot 0{,}5 = 0$

Таким образом, зная лишь математическое ожидание случайной величины, еще нельзя судить ни о том, какие возможные значения она может принимать, ни о том, как они рассеяны вокруг математического ожидания. Мерой рассеяния случайной величины является дисперсия.

Пусть X – случайная величина, $M(X)$ – ёе математическое ожидание. Рассмотрим в качестве новой случайной величины разность $X - M(X)$, которая называется отклонением случайной величины.

Определение. Дисперсией случайной величины X называется математическое ожидание квадрата уклонения случайной величины X.

Пусть закон распределения случайной величины X известен:

X	x_1	x_2	x_3	…	x_n
P	p_1	p_2	p_3	…	p_n

Обозначим $M(X) = m$. Тогда закон распределения уклонения случайной величины X будет такой:

$X - M(X)$	$x_1 - m$	$x_2 - m$	$x_3 - m$	……	$x_n - m$
P	p_1	p_2	p_3	……	p_n

Запишем закон распределения квадрата уклонения

$(X - m)^2$	$(x_1 - m)^2$	$(x_2 - m)^2$	$(x_3 - m)^2$	……	$(x_n - m)^2$
P	p_1	p_2	p_3	……	p_n

Таким образом, дисперсия $D(X)$ будет вычисляться по формуле:

$$D(X) = M(X - m)^2 = \sum (x_i - m)^2 p_i$$

Свойства дисперсии

Опираясь на свойства математического ожидания, вычислим

$$D(X) = M(X - m)^2 = M[X^2 - 2 \cdot m \cdot X + m^2] =$$

$$= M(X^2) - 2 \cdot m \cdot M(X) + M(m^2) = M(X^2) - 2 \cdot m \cdot m + m^2 =$$

$$= M(X^2) - 2 \cdot m^2 + m^2 = M(X^2) - m^2 = M(X^2) - [M(X)]^2.$$

Таким образом, получена ещё одна формула вычисления дисперсии

$$D(X) = M(X^2) - [M(X)]^2.$$

1. Дисперсия постоянной величины равна нулю $D(C) = 0$.

$$D(C) = M(C^2) - [M(C)]^2 = C^2 - C^2 = 0.$$

2. Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат.

$D(CX) = C^2 D(X)$; $D(CX) = M(C^2X^2) - [M(CX)]^2 = C^2[M(X^2) - (M(X))^2] = C^2 D(X)$.

3. Дисперсия суммы двух независимых случайных величин равна сумме дисперсий этих величин.

$$D(X + Y) = M(X + Y)^2 - [M(X + Y)]^2 = M(X^2 + 2 \cdot X \cdot Y + Y^2) - [M(X) + M(Y)]^2 =$$

$$= M(X^2) + 2 \cdot M(X \cdot Y) + M(Y^2) - [M(X)]^2 - 2 \cdot [M(X)][M(Y)]^2 - [M(Y)]^2 =$$

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

$$= D(X) + D(Y).$$

Следствие. $D(X + C) = D(X)$, где $C = const$.

$$D(X + C) = D(X) + D(C) = D(X) \text{ т.к. } D(C) = 0.$$

4. $D(X - Y) = D(X) + D(Y)$.

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

Математическое ожидание и дисперсия случайной величины, распределенной по биноминальному закону

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A постоянна и равна P. Чему равно среднее число появления события A в этих испытаниях.

Пусть случайная величина X – число появления события A в n независимых испытаниях. Очевидно, в одном испытании X_1 равно нулю или единице, а число X определяется количеством появлений события A в каждом из n испытаний. Таким образом,

$$X = X_1 + X_2 + \ldots + X_n; \quad q + p = 1;$$

X_k	0	1
P	Q	p

$$M(X_k) = 0 \cdot q + 1 \cdot p = p; \; k = \overline{1,n};$$

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

Вычислим $M(X^2)$:

X^2	0	1
P	Q	p

$M(X^2) = p$. Вычислим дисперсию в единичном испытании:

$$D(X_k) = M(X^2) - [M(X)]^2 = p - [p]^2 = p[1-p] = p \cdot q.$$

$$D(X_1 + X_2 + \ldots + X_n) = D(X_1) + D(X_2) + \ldots + D(X_n) = n \cdot p \cdot q.$$

Таким образом, математическое ожидание и дисперсия в схеме Бернулли вычисляются по формулам: $M(X) = n \cdot p$, $D(X) = n \cdot p \cdot q$.

Определение. Среднеквадратическим отклонением случайной величины X называется корень квадратный из дисперсии $\sigma(X) = \sqrt{D(X)}$.

$$\sigma(X_1 + X_2 + \ldots X_n) = \sqrt{D(X_1 + X_2 + \ldots X_n)} = \sqrt{D(X_1) + D(X_2) + \ldots + D(X_n)} =$$

$$= \sqrt{\sigma_1^2(X) + \sigma_2^2(X) + \ldots + \sigma_n^2(X)},$$

если случайные величины $X_1, X_2 \ldots X_n$ взаимно независимы.

Одинаково распределенные взаимно независимые случайные величины

Рассмотрим n взаимно независимых случайных величин X_1, $X_2, \ldots X_n$, которые имеют одинаковые распределения, следовательно, и одинаковые характеристики (математическое ожидание и дисперсию). Обозначим среднее арифметическое рассматриваемых величин через $\overline{X}$:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Пусть $M(X_1) = M(X_2) = \ldots M(X_n) = a$, $D(X_1) = D(X_2) = \ldots = D(X_n) = \sigma^2$.

1. Математическое ожидание среднего арифметического одинаково распределенных взаимно независимых случайных величин равно математическому ожиданию a каждой из величин:

$$M(\overline{X}) = M\left(\frac{X_1 + X_2 + \ldots X_n}{n}\right) = \frac{1}{n}[M(X_1) + \ldots + M(X_n)] = \frac{1}{n} \cdot n \cdot a = a$$

2. Дисперсия среднего арифметического n одинаково распределенных взаимно независимых случайных величин в n раз меньше дисперсии σ^2 каждой из величин:

$$D(\overline{X}) = D\left(\frac{X_1 + X_2 + \ldots X_n}{n}\right) = \frac{1}{n^2}[D(X_1) + \ldots + D(X_n)] = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Пример. Обычно для измерения некоторой физической величины производят несколько измерений, а затем находят среднее арифметическое полученных чисел, которое принимают за приближенное значение измеряемой величины. Предполагая, что измерения производят в одних и тех же условиях, доказать:

а) среднее арифметическое дает результат более надежный, чем отдельные измерения;

б) с увеличением числа измерений надежность этого результата возрастает.

Решение. Мы вправе рассматривать возможные результаты n отдельных измерений в качестве случайных величин $X_1, X_2,\ldots X_n$ (индекс указывает номер измерения). Эти величины имеют одинаковое распределение, т.к. измерения производятся по одной и той же методике и теми же приборами, а, следовательно, и одинаковые числовые характеристики, кроме того, они взаимно независимы, т.к. результат каждого отдельного измерения не зависит от остальных. На основании свойства 2 мы можем утверждать, что среднее арифметическое рассматриваемых величин имеет меньшее рассеяние, чем каждая отдельная величина. Иначе говоря, среднее арифметическое оказывается более близким к истинному значению. Кроме того, с увеличением числа измерений среднее арифметическое всё менее отличается от истинного значения измеряемой величины. Таким образом, увеличивая число измерений, получают более надежный результат.

Неравенство Чебышева

Рассмотрим дискретную случайную величину X, заданную таблицей распределения:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Докажем следующую *теорему*: Вероятность того, что отклонение случайной величины X от её математического ожидания по абсолютной величине меньше положительного числа ε, не меньше, чем $1-D(X)/\varepsilon^2$.

Доказательство. Вычислим $D(X)=\sum_{i=1}^{n}[x_i-M(X)]^2\,p_i$.

Из этой суммы выбросим те слагаемые, у которых $|x_i-M(X)|\langle\varepsilon$, таким образом

$$D(X)=\sum_{i=1}^{n}[x_i-m]^2\,p_i \geq \sum_{i:|x_i-m|\rangle\varepsilon}[x_i-m]^2\,p_i \geq \varepsilon^2\cdot\sum_{i:|x_i-m|\rangle\varepsilon}p_i=\varepsilon^2 P\{|X-m|\rangle\varepsilon\}.$$

Таким образом, $D(X) \geq \varepsilon^2 \cdot P\{|X_i - m| \geq \varepsilon\}$ или $P\{|X_i - m| \geq \varepsilon\} \leq \frac{D(X)}{\varepsilon^2}$.

Оценим вероятность противоположного события: $P\{|X_i - m| < \varepsilon\} \geq 1 - \frac{D(X)}{\varepsilon^2}$.

Теорема Чебышева.

Если X_1, X_2,…X_n – попарно независимые случайные величины, причем дисперсии их равномерно ограничены (не превышают постоянное число с), то, как бы ни было мало положительное число ε, вероятность неравенства

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

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

Таким образом $\lim\limits_{n \to \infty} P\left\{ \left| \frac{X_1 + X_2 ... + X_n}{n} - \frac{M(X_1) + M(X_2) ... + M(X_n)}{n} \right| < \varepsilon \right\} = 1$

Доказательство. Рассмотрим

$$\overline{X} = \frac{X_1 + X_2 ... + X_n}{n}; \quad M(\overline{X}) = \frac{M(X_1) + M(X_2) ... + M(X_n)}{n}$$

Запишем неравентсво Чебышева:

$$P\left\{ \left| \frac{X_1 + X_2 ... + X_n}{n} - \frac{M(X_1) + M(X_2) ... + M(X_n)}{n} \right| < \varepsilon \right\} \geq 1 - \frac{D\left(\frac{X_1 + X_2 ... + X_n}{n} \right)}{\varepsilon^2}$$

Воспользуемся свойствами дисперсии:

$$D\left(\frac{X_1 + X_2 ... + X_n}{n} \right) = \frac{1}{n^2} [D(X_1) ... + D(X_n)] \leq \frac{c \cdot n}{n^2} = \frac{c}{n}$$

Следовательно:

$$P\left\{ \left| \frac{X_1 + X_2 ... + X_n}{n} - \frac{M(X_1) + M(X_2) ... + M(X_n)}{n} \right| \right\} \geq 1 - \frac{c}{n\varepsilon^2} \underset{n \to \infty}{\to} 1.$$

Итак, среднее арифметическое достаточно большого числа независимых случайных величин (дисперсии которых равномерно ограничены) утрачивает характер случайной величины.

Теорема Бернулли

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A равна p. Можно ли предвидеть, какова будет относительная частота появлений события.

Ответ на этот вопрос дает теорема Бернулли (1713 г.). Если в каждом из n независимых испытаний вероятность p появления события A постоянна, то как угодно близка к единице вероятность того, что отклонение относительной частоты от вероятности p по абсолютной величине будет сколь угодно малым, если число испытаний будет достаточно велико:

$$P\left\{\left|\frac{m}{n}-p\right|<\varepsilon\right\}=1.$$

Доказательство следует из теоремы Чебышева для случайной величины

$$X=\frac{X_1+X_2\ldots+X_n}{n},$$

т.к. случайные величины X_1, X_2,…X_n попарно независимы и дисперсии их ограничены.

Функция распределения вероятностей случайной величины

Рассмотрим случайную величину X, возможные значения которой заполняют сплошь интервал (a, b) или всю числовую ось. Такая случайная величина называется непрерывной. Очевидно, непрерывную случайную величину нельзя задавать в виде перечня всех её значений и соответствующих вероятностей, как для дискретной случайной величины. Поэтому возникает необходимость ввести универсальный способ определения случайной величины.

Пусть x – действительное число. Вероятность события, состоящего в том, что X примет значение меньшее x, т.е. вероятность события $\{X < x\}$, обозначим

через $F(x)$. Разумеется, если x изменяется, то изменяется и $F(x)$, т.е. $F(x)$ является функцией аргумента x: $F(x) = P\{X < x\}$.

Случайную величину называют непрерывной, если её функция распределения есть непрерывная, кусочно-дифференцируемая функция с непрерывной производной.

Свойства функции распределения

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

Следует из того, что вероятность всегда неотрицательное число, не превышающее единицы.

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

Доказательство. Пусть $x_2 > x_1$. Событие, состоящее в том, что X примет значение, меньшее x_2, можно рассматривать как сумму несовместных событий:

1) X примет значение меньшее x_1 с вероятностью $P\{X < x_1\}$;

2) X примет значение, удовлетворяющее неравенству $x_1 \le X \le x_2$, с вероятностью $P\{x_1 \le X \le x_2\}$. По теореме сложения вероятностей имеем:

$$P\{X < x_2\} = P\{X < x_1\} + P\{x_1 \le X < x_2\}.$$

Отсюда: $P\{X < x_2\} - P\{X < x_1\} = P\{x_1 \le X < x_2\}$ или

$$F(x_2) - F(x_1) = P\{x_1 \le X < x_2\} \ge 0, \qquad (*)$$

так как любая вероятность есть число неотрицательное. Таким образом,

$$F(x_2) - F(x_1) \ge 0 \quad \text{или} \quad F(x_2) \ge F(x_1).$$

Следствие 1.

$$P\{a \le X \le b\} = F(b) - F(a). \qquad (**)$$

Это важное следствие вытекает из формулы (*), если $x_2 = b$, $x_1 = a$.

Следствие 2. Вероятность того, что непрерывная случайная величина X примет одно определенное значение равно нулю.

Доказательство. Положив в формуле $a = x$, $b = x + \Delta x$, получим

$$P\{x < X < x + \Delta x\} = F(x + \Delta x) - F(x).$$

Устремим $\Delta x \to 0$. Так как X непрерывная случайная величина, то функция $F(x)$ непрерывна, таким образом

$$\lim_{\Delta x \to 0} \Delta F(x) = \lim_{\Delta x \to 0} [F(x + \Delta x) - F(x)] = 0 .$$

Следовательно,

$$P\{X = x_1\} = \lim_{\Delta x \to 0} P\{x_1 \le X < x_1 + \Delta x\} = \lim_{x \to 0} [F(x_1 + \Delta x) - F(x_1) = 0$$

3. Если возможные значения случайной величины принадлежат интервалу (a, b), то:

   1) $F(x) = 0$ при $x \le a$; 2) $F(x) = 1$, при $x \ge b$.

Следствие. Если возможные значения непрерывной случайной величины расположены на всей оси x, то справедливы следующие предельные соотношения:

$$\lim_{x \to -\infty} F(x) = 0; \quad \lim_{x \to \infty} F(x) = 1 .$$

Пример. Дискретная случайная величина задана таблицей распределения

X	1	4	8
P	0,3	0,1	0,6

Найти функцию распределения и построить её график.

Решение. Если $x \le 1$, то $F(x) = 0$.

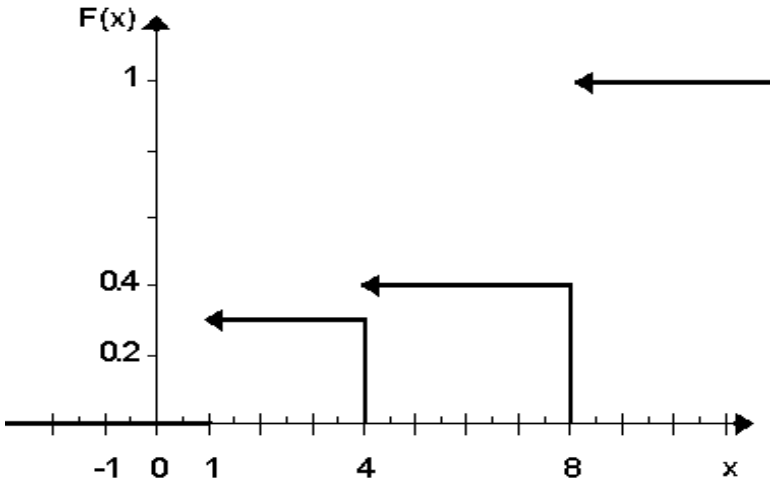
Если $1 < x \le 4$, то $F(x) = 0{,}3$.

Если $4 < x \le 8$, то $F(x) = 0{,}4$.

Действительно, в этом случае $F(x) = P\{x = 1\} + P\{x = 4\} = 0{,}3 + 0{,}1 = 0{,}4$.

Если $x > 8$, то $F(x) = 1$, так как событие $\{x \le 8\}$ достоверно, следовательно, $P\{x \le 8\} = 1$.

$$F(x)=\begin{cases}0, & \text{при} \quad x \le 1\\ 0{,}3 & \text{при} \quad 1 < x \le 4\\ 0{,}4 & \text{при} \quad 4 < x \le 8\\ 1 & \text{при} \quad x > 8\end{cases}$$



Плотность распределения вероятностей непрерывной случайной величины

Плотностью распределения вероятностей непрерывной случайной величины X называют функцию $f(x)$ – первую производную от функции распределения $F(x)$:

$$F'(x) = f(x).$$

Зная плотность распределения, функцию распределения можно определить:

$$F(x) = \int_{-\infty}^{x} f(t)dt\,.$$

Заметим, что для описания распределения вероятностей дискретной случайной величины плотность распределения неприменима.

Теорема. Вероятность того, что непрерывная случайная величина X примет значение, принадлежащее интервалу (a, b), равна определенному интегралу от плотности распределения в пределах от a до b.

Доказательство.

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

Замечание. Если $f(x)$ – четная функция и концы интервала симметричны относительно начала координат, то

$$P\{-a < X < a\} = P\{|x| < a\} = 2\int_0^a f(x)dx.$$

Свойства плотности распределения

1. Плотность распределения неотрицательная функция: $f(x) \geq 0$.

Доказательство. Функция распределения $F(x)$ неубывающая функция, следовательно:

$$F'(x) = f(x) \geq 0.$$

2. Несобственный интеграл от плотности распределения в пределах от -∞ до +∞ равен единице. $\int_{-\infty}^{+\infty} f(x)dx = 1.$

Доказательство.

$$\int_{-\infty}^{+\infty} f(x)dx = P\{x < +\infty\} = 1$$

Числовые характеристики непрерывных случайных величин

1. Математическим ожиданием непрерывной случайной величины X, возможные значения которой принадлежат интервалу $[a, b]$, называют определенный интеграл

$$M(x) = \int_a^b x\ f(x)dx.$$

Если возможные значения принадлежат всей числовой оси, то:

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

2. Дисперсией непрерывной случайной величины X называют математическое ожидание её квадрата уклонения:

$$D(x) = \int_a^b [x - M(x)]^2\ f(x)dx;\ \text{или}\ D(x) = \int_{-\infty}^{+\infty} [x - M(x)]^2\ f(x)dx.$$

Среднеквадратичное уклонение $\sigma(x) = \sqrt{D(x)}$.

Замечание. Легко доказать, что свойства математического ожидания и дисперсии дискретных случайных величин сохраняются и для непрерывных случайных величин.

Нормальное распределение

Нормальным называется распределение вероятностей непрерывной случайной величины, которое описывается плотностью:

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

Мы видим, что нормальное распределение определяется двумя параметрами a и σ. Покажем, что математическое ожидание случайной величины распределенной по нормальному закону равно a, а дисперсия – σ^2.

$$M(x)=\int_{-\infty}^{+\infty}x\ f(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}xe^{-\frac{(x-a)^2}{2\sigma^2}}dx=\left[\begin{array}{l}\text{сделаем замену}\\ \dfrac{x-a}{\sigma}=t;\ dt=\dfrac{dx}{\sigma}\\ -\infty<t<+\infty\end{array}\right]=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(t\sigma+a)e^{-\frac{t^2}{2}}dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t\sigma\, e^{-\frac{t^2}{2}}dt+\frac{a}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=$$

$$=\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}d\left(\frac{t^2}{2}\right)+\frac{a}{\sqrt{2\pi}}\sqrt{2\pi}=\frac{\sigma}{\sqrt{2\pi}}\lim_{A\to\infty}e^{-\frac{t^2}{2}}\Big|_{-A}^{A}+a=a.$$

В силу того, что $\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=\sqrt{2\pi}$,

$$\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}d\left(\frac{t^2}{2}\right)=\lim_{A\to\infty}\int_{-A}^{+A}e^{-\frac{t^2}{2}}d\left(\frac{t^2}{2}\right)=\lim_{A\to\infty}e^{-\frac{t^2}{2}}\Big|_{-A}^{A}=\lim_{A\to\infty}[e^{-\frac{A^2}{2}}-e^{-\frac{A^2}{2}}]=0.$$

Вычислим дисперсию

$$D(x) = M(x-a)^2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-a)^2 e^{\frac{(x-a)^2}{2\sigma^2}} = \left[\begin{array}{l}\frac{x-a}{\sigma} = t;\ x-a = t\sigma \\ \frac{dx}{\sigma} \equiv dt\end{array}\right] =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t^2\sigma^2 e^{-\frac{t^2}{2}}dt = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}}dt = -\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} te^{-\frac{t^2}{2}}d\left(-\frac{t^2}{2}\right) =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} t\,d\left(e^{-\frac{t^2}{2}}\right) = \frac{-\sigma^2}{\sqrt{2\pi}}\lim_{A\to\infty}\int_{-A}^{+A} t\,d\left(e^{-\frac{t^2}{2}}\right) = \frac{-\sigma^2}{\sqrt{2\pi}}\lim_{A\to\infty}\left[te^{-\frac{t^2}{2}}\Big|_{-A}^{+A} - \int_{-A}^{+A} e^{-\frac{t^2}{2}}dt\right] =$$

$$= \frac{-\sigma^2}{\sqrt{2\pi}}\lim_{A\to\infty}\left[\frac{2A}{e^{\frac{A^2}{2}}} - \int_{-A}^{+A} e^{-\frac{t^2}{2}}dt\right] = \frac{-\sigma^2}{\sqrt{2\pi}}(-\sqrt{2\pi}) = \sigma^2$$

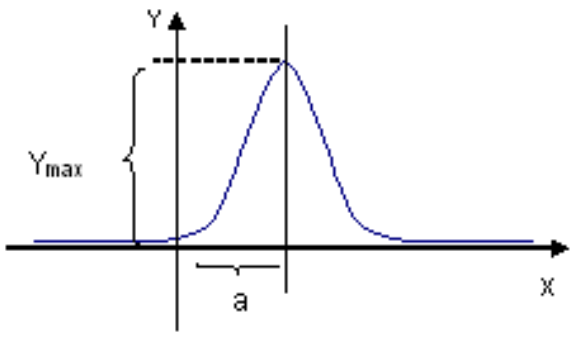
В силу того, что $\lim\limits_{A\to\infty}\frac{2A}{e^{A^2/2}} = \lim\limits_{A\to\infty}\frac{2}{e^{A^2/2}A} = 0$ по правилу Лопиталя, $\lim\limits_{A\to\infty}\int_{-A}^{+A} e^{-\frac{t^2}{2}}dt = \sqrt{2\pi}$. Исследуем функцию $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$ на экстремум:

$$f'(x) = \frac{-1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}\frac{2(x-a)}{2\sigma^2} = \frac{-(x-a)}{\sigma^3\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}};$$

$$f'(x) = 0 \text{ при } x = a.$$

Так как $f'(x) > 0$ при $x < a$ и $f'(x) < 0$ при $x > a$, то этот экстремум является максимумом этой функции. При $x = a$ функция $y = f(x)$ принимает значение $y_{\max}\big|_{x=a} = \frac{1}{\sqrt{2\pi}\sigma}$.

Кроме того, $x = a$ является осью симметрии функции $y = f(x)$.



При любых значениях параметров a и σ площадь ограниченная нормальной кривой и осью OX остается равной единице.

Заметим, что при $a = 0$, $\sigma = 1$ нормальный закон распределения называется основным.

Вероятность попадания в заданный интервал нормально распределенной случайной величины

Введем интеграл с переменным верхним пределом, который называется функцией Лапласа:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz .$$

Вычислим вероятность попадания случайной величины X в заданный интервал (α, β):

$$P\{\alpha < x < \beta\} = \int_\alpha^\beta f(t)dt = \frac{1}{\sqrt{2\pi}\sigma} \int_\alpha^\beta e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \left[\begin{array}{l} \text{замена} \\ \dfrac{t-a}{\sigma} = z \\ t = \sigma\, z + a \\ dt = \sigma\, dz \end{array}\right] =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}} e^{-\frac{z^2}{2}} \sigma\, dz = \frac{1}{\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}} e^{-\frac{z^2}{2}} dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{0} e^{-\frac{z^2}{2}} dz + \frac{1}{\sqrt{2\pi}} \int_0^{\frac{\beta-a}{\sigma}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{0} e^{-\frac{z^2}{2}} dz + \Phi\left(\frac{\beta-a}{\sigma}\right)$$

В первом интеграле поменяем пределы интегрирования, знак интеграла сменится на противоположный:

$$P\{\alpha < x < \beta\} = -\frac{1}{\sqrt{2\pi}} \int_0^{\frac{\alpha-a}{\sigma}} e^{-\frac{z^2}{2}} dz + \Phi\left(\frac{\beta-a}{\sigma}\right) = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right).$$

Используем полученную формулу для вычисления вероятности заданного отклонения:

$$P\{|x-a|<S\}=P\{-S<x-a<S\}=P\{a-S<x<a+S\}=$$

$$=\Phi\left(\frac{a+S-a}{\sigma}\right)-\Phi\left(\frac{a-S-a}{\sigma}\right)=\Phi\left(\frac{S}{\sigma}\right)-\Phi\left(\frac{-S}{\sigma}\right)=2\Phi\left(\frac{S}{\sigma}\right),$$

учитывая, что $\Phi\left(-\frac{S}{\sigma}\right)=-\Phi\left(\frac{S}{\sigma}\right)$.

Отсюда получим правило ***«трех сигм»***:

$$P\{|x-a|<3\sigma\}=2\Phi\left(\frac{3\sigma}{\sigma}\right)=2\Phi(3)=2\cdot 0{,}49865=0{,}9973.$$

Таким образом, если случайная величина распределена нормально, то абсолютная величина её отклонения от математического ожидания не превосходит утроенного среднего квадратического отклонения с вероятностью близкой к единице.

Понятие о центральной предельной теореме

Известно, что нормально распределенные случайные величины широко распространены на практике. Объяснение этому явлению было дано русским математиком А.М. Ляпуновым (центральная предельная теорема):

Если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то X имеет распределение, близкое к нормальному.

Пример. При проведении геохимических поисков на одном из участков было встречено значение концентрации 0,1% интересующего нас элемента. Необходимо оценить является ли встреченное значение x = 0,1% аномалией или это случайное отклонение фоновых концентраций, если известно, что среднее значение и стандартное отклонение фоновых концентраций соответственно равны: $\bar{x}=0.003\%; S_x=0.02\%$.

Задачу решить в двух вариантах:

1. Закон распределения фоновых концентраций является нормальным.

2. Закон распределения фоновых концентраций неизвестен, но есть предположение, что он не является нормальным.

Решение.

Вариант 1.

$$P\{x > t_x\} = 1 - F(t_x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t_x} e^{-\frac{z^2}{2}} dz = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \int_{0}^{t_x} e^{-\frac{z^2}{2}} dz =$$

$$= 1 - 0{,}5 - \int_{0}^{t_x} e^{-\frac{z^2}{2}} dz = 0{,}5 - \int_{0}^{t_x} e^{-\frac{z^2}{2}} dz,$$

где $t_x = \frac{x - \bar{x}}{\sigma_x} = \frac{0{,}1 - 0{,}03}{0{,}02} = \frac{0{,}07}{0{,}02} = 3{,}5.$

Таким образом: $P\{x > t_x\} = 0{,}5 - \Phi(3{,}5) = 0{,}0003.$

Вероятность встречи фоновых концентраций мала, поэтому встреченную концентрацию можно считать аномальной, $P \le 0{,}0816$.

Вариант 2. Воспользуемся неравенством Чебышева:

$$P\left\{\left|\frac{x - \bar{x}}{\sigma}\right| \ge t\right\} \le \frac{1}{t^2},$$

где t –любое число. В данном случае его можно определить из выражения, стоящего в фигурных скобках $t = \frac{|x - \bar{x}|}{\sigma} = \frac{0{,}1 - 0{,}03}{0{,}02} = 3{,}5.$

Таким образом, встреченную концентрацию можно считать аномальной, т.к. $P \le 0{,}0816$.

Система двух случайных величин

Будем обозначать через (X, Y) двумерную случайную величину. Каждую из величин X и Y называют составляющей (компонентой). Обе величины X и Y рассматриваемые одновременно, образуют систему двух случайных величин.

Закон распределения вероятностей дискретной двумерной случайной величины

Законом распределения вероятностей дискретной двумерной случайной величины называют перечень возможных значений этой величины, т.е. (x_i, y_j) и их вероятностей $p_{ij} = P(x_i, y_j)$, где $i = 1,2,\ldots n$; $j = 1,2,\ldots m$. Обычно закон распределения задают в виде таблицы с двойным входом, причем

$$\sum_i^n \sum_j^m P_{ij} = 1; \quad \sum_j^m P_{ij} = P\{X = x_i\}; \quad \sum_i^n P_{ij} = P\{Y = y_j\}.$$

Y	X					
	x_1	x_2	…	x_i	…	x_n
y_1	$p(x_1,y_1)$	$p(x_2,y_1)$		$p(x_i,y_1)$	…	$p(x_n,y_1)$
…	…	…	…	…	…	…
y_j	$p(x_1,y_j)$	$p(x_2,y_j)$		$p(x_i,y_j)$	…	$p(x_n,y_j)$
	…	…	…	…	…	…
y_m	$p(x_1,y_m)$	$p(x_2,y_m)$		$p(x_i,y_m)$	…	$p(x_n,y_m)$

Функция распределения двумерной случайной величины

Рассмотрим двумерную случайную величину (X, Y) дискретную или непрерывную. Вероятность события состоящего в том, что X примет значение, меньше x и при этом Y примет значение меньшее y, обозначается через $F(x, y)$. Если X, Y будет изменяться, то будет изменяться и $F(x, y)$. Таким образом, $F(x, y)$ – функция от x и y. Функцией распределения двумерной случайной величины (X, Y) называют функцию $F(x, y)$, определяющую для каждой пары чисел (x, y) вероятность того, что X примет значение, меньшее x, при этом Y примет значе-

ние меньшее y: $F(x, y) = P\{X < x, Y < y\}$.

Свойства функции распределения двумерной случайной величины

1. $0 < F(x, y) \le 1$. Свойство вытекает из определения функции $F(x,y)$ как вероятности.

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

   $F(x, y_2) \ge F(x, y_1)$, если $y_2 > y_1$.

Событие $\{X < x_2, Y < y\} = \{X < x_1, Y < y\} + \{x_1 \le X < x_2, Y < y\}$.

Очевидно $P\{x_1 < X < x_2, Y < y\} > 0$. Таким образом:

$$P\{X < x_2, Y < y\} = P\{X < x_1, Y < y\} + P\{x_1 \le X < x_2, Y < y\},$$

$$F(x_2, y) = F(x_1, y) + P\{x_1 \le X < x_2, Y < y\},$$

или

$$F(x_2, y) - F(x_1, y) = P\{x_1 \le X < x_2, Y < y\} > 0.$$

Отсюда $F(x_2, y) > F(x_1, y)$.

3. Имеют место предельные соотношения:

   а) $F(-\infty, y) = 0$; б) $F(x, -\infty) = 0$; в) $F(-\infty, -\infty) = 0$; г) $F(+\infty,+\infty) = 1$.

Доказательство

   а) $F(-\infty, y) = P\{X < -\infty, Y < y\} = 0$; б) $F(x, -\infty) = P\{X < x, Y < -\infty\} = 0$;

   в) $F(-\infty,-\infty) = P\{X < -\infty, Y < -\infty\} = 0$; г) $F(+\infty,+\infty) = P\{X < +\infty, Y < +\infty\} = 1$.

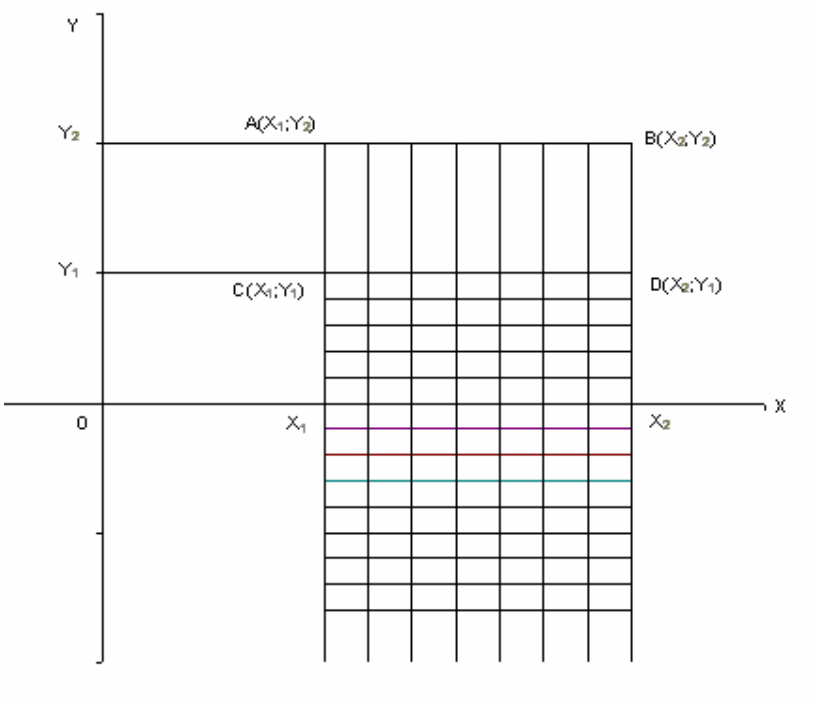
4. При $y \to +\infty$ функция распределения не зависит от Y:

   $F(x, +\infty) = P\{X < x, Y < +\infty\} = F_1(x)$; $F(+\infty, y) = P\{X < +\infty, Y < y\} = F_2(y)$.

5. Вероятность попадания случайной точки в прямоугольник

$$P\{x_1 \le X < x_2, y_1 \le Y < y_2\} = [F(x_2,y_2) - F(x_1,y_2)] - [F(x_2,y_1) - F(x_1,y_1)].$$

Легко доказать геометрически. Это вероятность попадания в прямоугольник *ABCD*. Исходную вероятность можно найти как вероятность попадания случайной точки в полуполосу *AB* минус вероятность попадания случайной точки в полуполосу *CD*, т.е.: $[F(x_2,y_2) - F(x_1,y_2)] - [F(x_2,y_1) - F(x_1,y_1)]$.



Плотность совместного распределения вероятностей непрерывной двумерной случайной величины

Плотностью совместного распределения вероятностей $f(x,y)$ двумерной непрерывной случайной величины (X, Y) называют вторую смешанную частную производную от функции распределения: $f(x,y)=\dfrac{\partial^2 F(x,y)}{\partial x \partial y}$.

Тогда $F(x,y)=\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{y} f(x,y)dxdy$.

Свойства функции $f(x,y)$

1. $f(x,y)\geq 0$
2. $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)dxdy=1$
3. $\int\limits_{-\infty}^{+\infty} f(x,y)dy=f_1(x);\cdots\int\limits_{-\infty}^{+\infty} f(x,y)dx=f_2(y)$

Условные законы распределения составляющих системы дискретных величин

Известно, что если события A и B зависимы, то условная вероятность события B отличается от его безусловной вероятности. В этом случае

$$P_A(B) = P(AB)/P(A).$$

Аналогичное положение имеет место и для случайных величин. Для того, чтобы охарактеризовать зависимость между составляющими двумерной случайной величины, введем понятие условного распределения.

Допустим, что в результате испытания величина Y приняла значение $Y = y_j$, при этом X может принять одно из своих возможных значений $x_1, x_2,\ldots x_n$. В этом случае условные вероятности составляющей будем обозначать $P(x_i/y_j)$, $i = 1,2,\ldots n$. Условным распределением составляющей X при $Y = y_j$ называют совокупность значений:

$$P(x_1/y_j), P(x_2/y_j),\ldots P(x_n/y_j);\quad P(x_i / y_j) = \frac{P(x_i, y_j)}{P(y_j)},\ i = 1,2,\ldots n.$$

Пусть $(X,\ Y)$ – непрерывная двумерная случайная величина. Условной плотностью $\varphi(x/y)$ распределения составляющих X при данном значении $Y = y$ называют отношение плотности совместного распределения $f(x,y)$ системы $(X,\ Y)$ к плотности распределения $f_2(y)$ составляющей Y:

$$\varphi(x/y) = f(x,y)/\ f_2(y).$$

Аналогично,

$$\psi(y/x) = f(x,y)/\ f_1(x);\quad \varphi(x/y) = \frac{f(x,y)}{\int\limits_{-\infty}^{+\infty} f(x,y)dx};\quad \psi(y/x) = \frac{f(x,y)}{\int\limits_{-\infty}^{+\infty} f(x,y)dy}.$$

При этом $\int\limits_{-\infty}^{+\infty}\varphi(x/y)dx = 1$, $\int\limits_{-\infty}^{+\infty}\psi(y/x)dy = 1.$

Условное математическое ожидание

Условным математическим ожиданием дискретной случайной величины Y при $X = x$ (x – определенное возможное значение случайной величины X) называют произведение возможных значений Y на их условные вероятности

$$M(Y/X = x) = \sum y_j P(y_j / x).$$

Для непрерывных величин

$$M(Y / X = x) = \int Y \psi(y / x) dy .$$

$M(Y / x) = f(x)$ - функция от x, так как зависит от x, $f(x)$ называют функцией регрессии Y на X.

Зависимые и независимые случайные величины

Теорема. Для того, чтобы случайные величины X и Y были независимы, необходимо и достаточно, чтобы функция распределения системы (X, Y) была равна произведению функций распределения составляющих

$$F(x, y) = F_1(x) F_2(y) .$$

Необходимость. Поскольку X и Y независимы

$$P\{X < x, Y < y\} = P\{X < x\} P\{Y < y\} = F_1(x) F_2(y) .$$

По определению $P\{X < x, Y < y\} = F(x, y)$.

Таким образом $F(x, y) = F_1(x) F_2(y)$.

Достаточность. Пусть $F(x, y) = F_1(x) F_2(y)$ и $F(x, y) = P\{X < x, Y < y\}$,

$$F_1(x) = P\{X < x\},\ F_2(y) = P\{Y < y\} .$$

Таким образом $P\{X < x, Y < y\} = P\{X < x\} P\{Y < y\}$.

Следствие. Для того, чтобы непрерывные случайные величины были независимы, необходимо и достаточно, чтобы плотность совместного распределения системы (X, Y) была равна произведению плотностей распределений составляющих

$$f(x, y) = f_1(x) f_2(y) .$$

Доказательство.

Необходимость:

$$F(x, y) = F_1(x) F_2(y) \text{ и } \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial F_1(x)}{\partial x} \cdot \frac{\partial F_2(y)}{\partial y} = f_1(x) f_2(y) .$$

Таким образом $f(x, y) = f_1(x) f_2(y)$.

Достаточность. Пусть $f(x, y) = f_1(x) f_2(y)$.

Интегрируя это равенство по x и по y, получим:

$$\int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dxdy = \int_{-\infty}^{x} f_1(x)dx \int_{-\infty}^{y} f_2(y)dy \text{ и } F(x,y) = F_1(x)F_2(y).$$

Числовые характеристики системы двух случайных величин. Корреляционный момент. Коэффициент корреляции.

Корреляционным моментом μ_{xy} случайных величин X, Y называют математическое ожидание произведения отклонений этих величин:

$$\mu_{xy} = M\{[X - M(X)][Y - M(Y)]\}.$$

Для вычисления корреляционного момента дискретных величин используют формулу

$$\mu_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{m}[x_i - M(X)][y_j - M(Y)]P_{ij}$$

для непрерывных величин

$$\mu_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[X - M(X)][Y - M(Y)]f(x,y)dxdy.$$

Корреляционный момент служит для характеристики связи между величинами X и Y.

<u>*Теорема.*</u> Корреляционный момент двух независимых случайных величин X и Y равен нулю.

Доказательство. Так как X и Y независимы, то и случайные величины $X - M(x), Y - M(y)$ также независимы, следовательно:

$$\mu_{xy} = M\{[X - M(X)][Y - M(Y)]\} = M[X - M(x)]M[Y - M(y)] = 0.$$

Если корреляционный момент μ_{xy} не равен нулю, то его величина характеризует степень зависимости случайных величин X, Y. Для удобства сравнения вводят коэффициент корреляции

$$r_{xy} = \mu_{xy}/(\sigma_x\sigma_y).$$

Свойства коэффициента корреляции

1. $|\mu_{xy}| \le \sqrt{D_x D_y}$;

2. $|r_{xy}| \le 1$.

Две случайные величины X и Y называют коррелированными, если их корреляционный момент (или коэффициент корреляции) отличен от нуля; X и Y называют некоррелированными величинами, если их корреляционный момент равен нулю. Если две величины зависимы, то они могут быть как коррелированны, так и не коррелированны. Итак, из коррелированности двух случайных величин следует их зависимость, но из зависимости ещё не следует коррелированность.

Линейная регрессия. Прямые линии среднеквадратической регрессии.

Рассмотрим двумерную случайную величину (X, Y), где X и Y - зависимые случайные величины. Представим одну величину как функцию другой:

$$y \approx g(x) = \alpha \mathrm{X} + \beta,$$

где α, β - параметры, подлежащие определению. Это можно сделать с помощью метода наименьших квадратов. Функцию $g(\mathrm{X}) = \alpha \mathrm{X} + \beta$ называют «наилучшим приближением» Y в смысле метода наименьших квадратов, если $M[\mathrm{Y} - g(\mathrm{X})]^2$ принимает наименьшее значение, функцию $g(\mathrm{X})$ называют регрессией Y на X.

Теорема. Линейная среднеквадратическая регрессия Y на X имеет вид

$$g(\mathrm{X}) = m_y + r\frac{\sigma_y}{\sigma_x}(\mathrm{X} - m_x),$$

где $m_x = M(\mathrm{X}), m_y = M(\mathrm{Y}), \sigma_x = \sqrt{D(x)}, \sigma_y = \sqrt{D(y)}$, $r = M_{xy}/(\sigma_x \sigma_y)$ - коэффициент корреляции X и Y.

Доказательство.

Введем в рассмотрение функцию

$$F(\alpha, \beta) = M[\mathrm{Y} - \alpha - \beta \mathrm{X}]^2.$$

Учитывая

$$M(\mathrm{X}-m_x)=M(Y-m_y)=0 \quad \text{и} \quad M[(\mathrm{X}-m_x)\cdot(\mathrm{Y}-m_y)]=M_{xy}=r\sigma_x\sigma_y,$$

получим:

$$F(\alpha,\beta)=\sigma_y^2+\beta^2\sigma_x^2-2r\beta\sigma_x\sigma_y+(m_y-\alpha-\beta m_x)^2 \ .$$

Исследуем функцию $F(\alpha,\beta)$ на экстремум, для чего приравняем нулю частные производные:

$$\left\{\begin{array}{l}\dfrac{\partial F}{\partial\alpha}=-2(m_y-\alpha-\beta m_x)=0\\ \dfrac{\partial F}{\partial\beta}=2\beta\sigma_x^2-2r\sigma_x\sigma_y)=0\end{array}\right\}$$

Решая систему, получим $\beta=r\dfrac{\sigma_y}{\sigma_x}$, $\alpha=m_y-r\dfrac{\sigma_y}{\sigma_x}m_x$.

Легко убедиться, что при данных значениях α,β рассматриваемая функция принимает наименьшее значение. Коэффициент $\beta=r\dfrac{\sigma_y}{\sigma_x}$ называют коэффициентом регрессии Y на X, а прямую $Y-m_y=r\dfrac{\sigma_y}{\sigma_x}(X-m_x)$ называют прямой среднеквадратической регрессии Y на X.

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Основные понятия

Математическая статистика – это прикладной раздел теории вероятностей, занимающийся обработкой статистических данных, для того чтобы получить научно-обоснованные выводы.

Пусть требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака, характеризующего эти объекты. Иногда проводят сплошное обследование, т.е. обследуют каждый из объектов совокупности относительно некоторого признака. Но сплошное обследование сопряжено с определенными трудностями, например с большим количеством объектов исследования или с уничтожением объекта в результате исследования. Поэтому чаще проводят выборочное обследование, т.е. отбирают случайным образом ограниченное число объектов и подвергают их изучению.

Выборкой называют совокупность случайно отобранных объектов.

Генеральной совокупностью называют совокупность объектов, из которых производится выборка.

Объемом совокупности называют число объектов данной совокупности.

Повторной называется выборка, при которой отобранный объект перед отбором следующего возвращается в генеральную совокупность.

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

Выборка должна правильно представлять пропорции генеральной совокупности, выборка должна быть репрезентативной (представительной).

Пример: Из района изучения отбирается наугад n образцов, для которых определяется значение данного признака (например, битумонасыщенность):

$$\underbrace{x_1,\ x_2,\ldots,\ x_n}_{выборка объема\ n}\ .$$

Выборка должна отражать все закономерности изучаемого признака, свойства, т.е. должна быть представительной. Представительность достигается способом отбора:

- Пусть изучается порода одной скважины. Отбор производится через 20 см в глубину, исходя из насыщения пород нефтью или битумом или каждый третий образец после разбиения на интервалы. Этот способ – механический. Интервал (шаг) отбора устанавливается экспериментатором исходя из цели и задачи работы.
- Пусть имеется N пронумерованных образцов, а для изучения необходимо отобрать наугад k образцов. В таких случаях, пользуясь таблицей случайных чисел, берут подряд k чисел по таблице, начиная с любого места. Образцы, детали с полученными номерами отбираются для изучения. Такой способ называется простым случайным отбором.
- Пусть генеральная совокупность разбита на подмножества –

А) по предположению неоднородности или

Б) по районам исследования или

В) по пластам залегания породы.

Тогда целесообразно по каждому множеству провести простой случайный отбор для того, чтобы в выборку попали элементы из каждого множества. Такой способ выделения выборки называется типическим.

Выборка может включить все результаты исследования проб за определенный промежуток времени. В этом случае говорят, что произведен серийный отбор продукции.

Сама выборка является случайной системой относительно генеральной совокупности. Поэтому возникает проблема: как оценить параметры распределения и как установить закон распределения случайной величины по выборке. Для этого определяются характеристики выборки.

Статистическое распределение выборки.

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 - n_2 раз,…, x_k - n_k раз и $\sum_{i=1}^{k} n_i = n$ - объем выборки. Наблю-

даемые значения x_k называются вариантами, а последовательность вариант, записанных в возрастающем порядке - вариационным рядом. Числа наблюдений называются частотами, а отношение частот к объему выборки – относительными частотами.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение может быть задано в виде последовательности интервалов и соответствующих им частот (*Таблица 1*).

Таблица 1.

x_1	x_2	…	x_k
n_1	n_2	…	n_k
$\frac{n_1}{n}$	$\frac{n_2}{n}$		$\frac{n_m}{n}$

- варианты

- частоты

- относительные частоты

Величина интервала (шаг) оценивается по формуле

$$d = \frac{x_{\max} - x_{\min}}{1 + 3{,}332 \cdot \lg n},$$

где $x_{\max}$, $x_{\min}$ - крайние члены вариационного ряда, n - объем выборки.

Эмпирическая функция распределения

Пусть известно статистическое распределение частот количественного признака X. Эмпирической функцией распределения называют функцию $F^*(x) = \frac{n_x}{n}$, где n_x - число вариант $X < x$, n - объем выборки, x – произвольное значение аргумента.

Свойства функции $F^*(x)$:

1. Значение функции $F^*(x)$ принадлежит интервалу [0;1].
2. $F^*(x)$ - неубывающая функция.

Пример. Пусть в *Таблице 2* задано распределение дискретной случайной величины X и ее относительные частоты. Построить эмпирическую функцию.

Решение:

Графическую иллюстрацию дискретной таблицы (*Рис.1*) называют полигоном относительных частот.

Таблица 2.

№	1	2	3	4
X	2	4	7	8
W	0,2	0,1	0,3	0,4

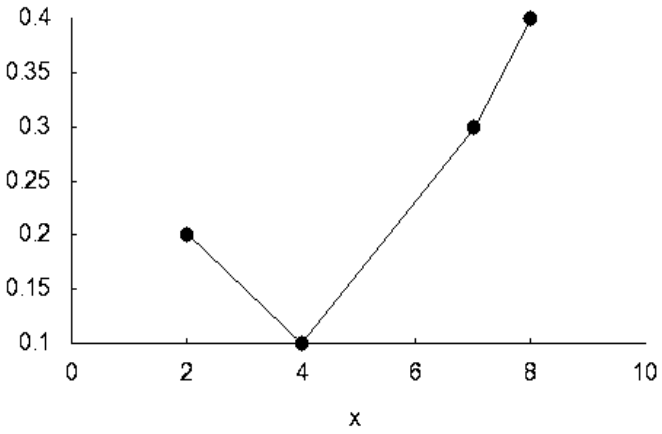


Рис. 1

Эмпирическая функция распределения $F^*(x)$ случайной величины в этом случае будет иметь вид:

Таблица 3.

$X < x$	$F^*(x)$	$\sum_{x_i < x} w_i$
$x \le 2$	0	0
$x \le 4$	0,2	w_1
$x \le 7$	0,3	$w_1 + w_2$
$x \le 8$	0,6	$w_1 + w_2 + w_3$
$x > 8$	1	$w_1 + w_2 + w_3 + w_4$

Построим графическое изображение эмпирической функции:

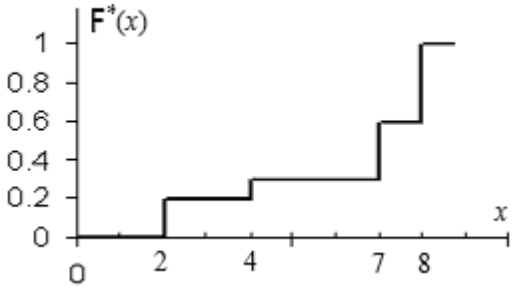


Рис. 2

Гистограмма (эмпирический закон распределения непрерывного признака X).

В случае непрерывного признака целесообразно строить гистограмму, для чего интервал, в котором заключены все наблюдаемые значения признака, разбиваются на несколько частичных интервалов длиной h и находят для каждого частичного интервала h_i сумму всех частот вариант, попавших в i – тый интервал. В результате получают интервальную таблицу.

Таблица 4.

$[b_1, b_2]$	$(b_2, b_3]$	…	$(b_m, b_{m+1}]$
n_1	n_2	…	n_m
$\frac{n_1}{n}$	$\frac{n_2}{n}$		$\frac{n_m}{n}$

где n – число всех измерений, m – количество интервалов, n_i -сумма частот, приходящихся на i-ый интервал, $w_i = \frac{n_i}{n}$ - относительная частота.

Гистограммой частот называют фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частот). Площадь i – того прямоугольника равна

$$S_i = \frac{n_i}{h} \cdot h = n_i,$$

$$\sum_i S_i = \sum_i n_i = n \text{ - объем выборки.}$$

Гистограмма относительных частот строится по тем же правилам, только высота прямоугольников определяется как $\frac{w_i}{h}$ - плотность относительной частоты.

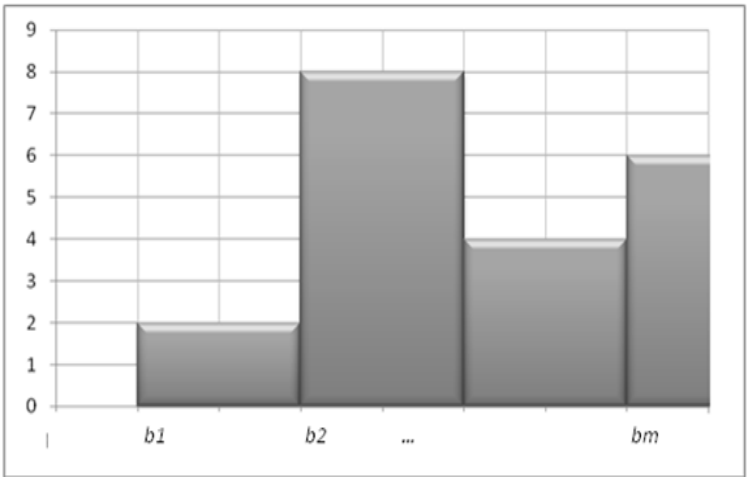


Рис.2

Задача 1: В *Таблице 5* дано содержание битума в породах песчаной пачки Шешминского горизонта по ряду площадей восточного борта Мелекесской впадины (относительные единицы). Построить вариационный ряд и статистическое распределение.

Таблица 5.

№ п/п	№ анализа	Скважина, площадь	*X* (отн.ед.)
1	59	Ашальчинская, 67	0,265
2	57	Ашальчинская, 62	0,262
3	94	Кармалинская, 74	0,127
4	35	Подлесная, 46	0,142
5	140	Сарабикуловская,10	0,122
6	15	Сарабикуловская, 2	0,22
7	14	Сарабикуловская, 7	0,03
8	12	Шугуровская, 5	0,34
9	13	Шугуровская, 7	0,03
10	141	Шугуровская, 16	0,143
11	2	Сугушлинская, 119	0,03
12	3	Сугушлинская, 119	0,06
13	135	Сугушлинская, 132	0,088
14	137	Сугушлинская, 132	0,008
15	8	Сугушлинская, 105	0,07
16	9	Сугушлинская, 105	0,04
17	134	Сугушлинская, 124	0,057
18	130	Сугушлинская, 123	0,077

Решение: Находим в таблице максимальное и минимальное значения вариационного ряда $x_{\max} = 0{,}34;\ \ x_{\min} = 0{,}008$. Объем выборки равен 18.

Таблица 6.

№ интервала	*Интервалы*	*Частоты* n_i
1	0,00 – 0,05	5
2	0,05 – 0,10	5
3	0,10 – 0,15	4
4	0,15 – 0,20	0
5	0,20 – 0,25	2
6	0,25 – 0,30	1
7	0,30 – 0,35	1

Вычислим размах и величину интервала

$$x_{\max} - x_{\min} = 0{,}34 - 0{,}008 = 0{,}332,$$

$$d = \frac{x_{\max} - x_{\min}}{1 + 3{,}332 \cdot \lg n} = \frac{0{,}332}{1 + 3{,}332 \cdot \lg 18} = 0{,}06.$$

Для удобства вычислений возьмем d=0,05.

Разобьем вариационный ряд на интервалы [0; 0,05], (0,05; 0,10], … , (0,30; 0,35]. Подсчитаем число значений, попадающих в каждый интервал.

Упражнение1. Определение содержания битума в породе из песчаной пачки Шешминского горизонта по 22 пробам бурения скважины на восточном борту Мелекесской впадины дало следующие результаты:

Таблица 7.

0,265	0,220	0,030	0,040	0,080	0,242	0,030	0,060	0,057	0,170	0,127
0,34	0,088	0,077	0,142	0,030	0,008	0,169	0,122	0,143	0,070	0,183

Построить вариационный ряд и статистическое распределение с величиной d=0,05.

Задача 2: В *Таблице 8* приведены значения проницаемости *X* в *мд* пласта а горизонта $Д_1$ по 100 скважинам Зеленогорской площади Ромашкинского месторождения и их логарифмы. Построить статистическое распределение для значений проницаемости их логарифмов. Построить гистограммы частот.

Таблица 8

x_i	$Lg\ x_i$	x_i	$Lg\ x_i$	x_i	$Lg\ x_i$	x_i	$Lg\ x_i$	x_i	$Lg\ x_i$
20	**1,301**	**80**	1,903	125	2,097	200	2,301	320	2,505
25	1,398	85	1,929	135	2,130	200	2,301	320	2,505
31	1,491	90	1,954	140	2,146	200	2,301	320	2,505
40	1,602	90	1,954	140	2,146	200	2,301	340	2,531
42	1,623	90	1,954	140	2,146	210	2,322	350	2,544
43	1,633	91	1,959	145	2,161	210	2,322	375	2,574
45	1.653	93	1,968	150	2,176	215	2,332	380	2,580
50	1,699	100	2,000	150	2,176	220	2,342	380	2,580
50	1,699	100	2,000	150	2,176	225	2,352	380	2,580
60	1,778	115	2,061	160	2,204	230	2,362	400	2,602
60	1,778	115	2,061	165	2,217	240	2,380	400	2,602
60	1,778	115	2,061	170	2,230	250	2,398	420	2,623
70	1,845	120	2,079	170	2,230	270	2,431	450	2,653
70	1,845	120	2,079	170	2,230	270	2,431	470	2,672
70	1,845	120	2,079	175	2,243	280	2,447	550	2,740
75	1,875	125	2,097	180	2,255	281	2,449	580	2,763
80	1,903	125	2,097	180	2,255	300	2,477	650	2,813
80	1,903	125	2,097	185	2,267	300	2,477	650	2,813
80	1,903	125	2,097	190	2,279	300	2,477	650	2,813
80	1,903	125	2,097	200	2,301	320	2,505	**700**	**2,845**

Решение: 1. Для вариационного ряда X (проницаемости) находим крайние члены ряда $x_{\max} = 700;\ \ x_{\min} = 20$. Объем выборки равен 100. Вычислим размах и величину интервала

$$x_{\max} - x_{\min} = 680,$$

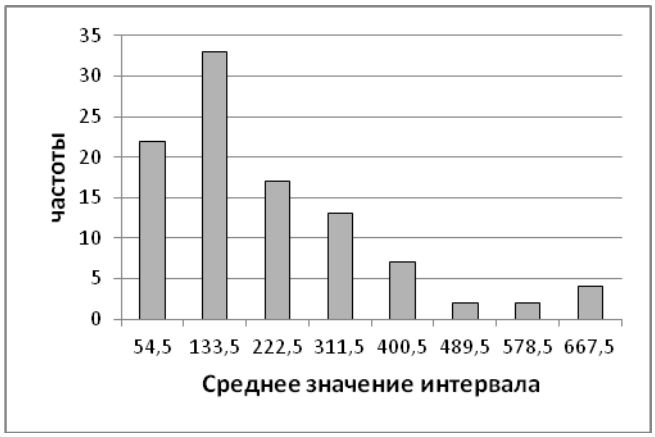
$$d = \frac{x_{\max} - x_{\min}}{1 + 3{,}332 \cdot \lg n} = \frac{680}{1 + 3{,}332 \cdot \lg 100} \approx 89.$$

Для проведения дальнейшей статистической обработки (определение среднего значения, функции распределения и т.д.) необходимо определить среднее значение интервала - $(a+b)/2$ - полусумму граничных значений интервала. Правильное применение интервалов позволяет построить компактный и наглядный сгруппированный вариационный ряд (*Таблица 9*).

Таблица 9

№	интервалы		(a+b)/2	частоты
	a	*b*		
1	20	89	54,5	22
2	89	178	133,5	33
3	178	267	222,5	17
4	267	356	311,5	13
5	356	445	400,5	7
6	445	534	489,5	2
7	534	623	578,5	2
8	623	712	667,5	4

Гистограмма 1



2. Вторая часть задания включает построение статистического распределения логарифма проницаемости (*Таблица 8).* Находим крайние значения вариационного ряда $Lg\ x_{\max} = 2{,}845;\ \ Lg\ x_{\min} = 1{,}301$. Рассчитаем шаг

$$d = \frac{2{,}845 - 1{,}301}{1 + 3{,}332 \cdot 2} \approx 0{,}2.$$

Представим статистическое распределение LgX и его графическое представление – гистограмму в зависимости от величины шага А) *d*=0,2; Б) *d*=0,21; В) *d*=0,3:

Таблица 10А

№	интервалы		(a+b)/2	часто-ты
	a	*b*		
1	1,301	1,501	1,401	3
2	1,501	1,701	1,601	6
3	1,701	1,901	1,801	7
4	1,901	2,101	2,001	25
5	2,101	2,301	2,201	23
6	2,301	2,501	2,401	15
7	2,501	2,701	2,601	15
8	2,701	2,901	2,801	6

А) *d*=0,2 *Гистограмма 2А*

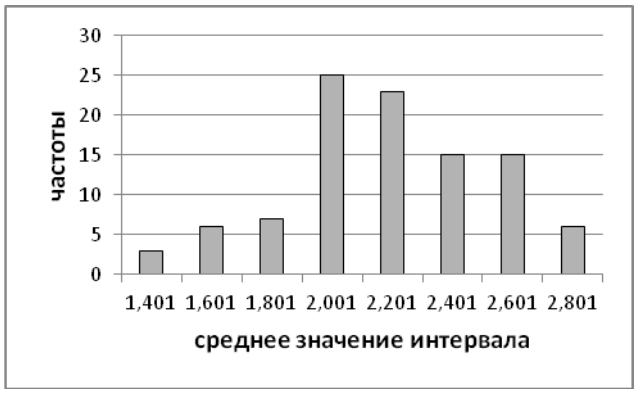


Таблица 10Б

№	интервалы		$(a+b)/2$	частоты
	a	*b*		
1	1,301	1,511	1,406	3
2	1,511	1,721	1,616	6
3	1,721	1,931	1,826	13
4	1,931	2,141	2,036	20
5	2,141	2,351	2,246	26
6	2,351	2,561	2,456	17
7	2,561	2,771	2,666	11
8	2,771	2,981	2,876	4

Б) d=0,21 *Гистограмма 2Б*

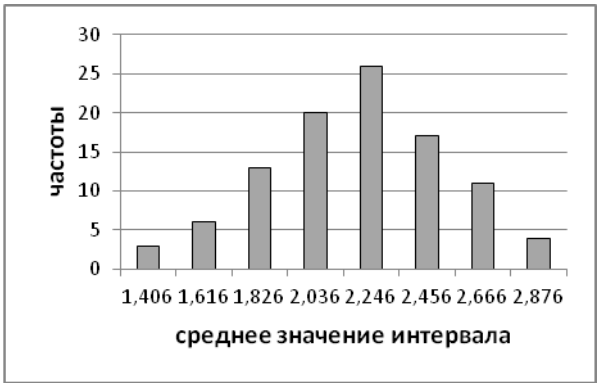
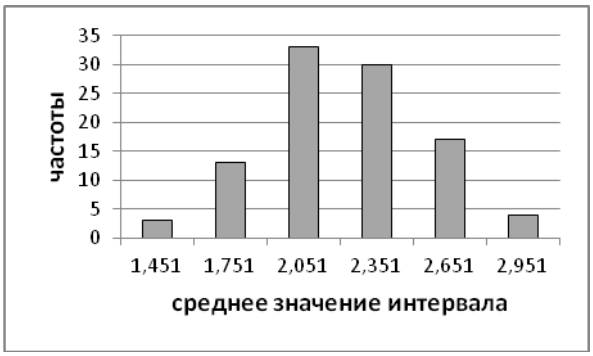


Таблица 10В

№	интервалы		$(a+b)/2$	частоты
	a	*b*		
1	1,301	1,601	1,451	3
2	1,601	1,901	1,751	13
3	1,901	2,201	2,051	33
4	2,201	2,501	2,351	30
5	2,501	2,801	2,651	17
6	2,801	3,101	2,951	4

В) d=0,2 *Гистограмма 2В*



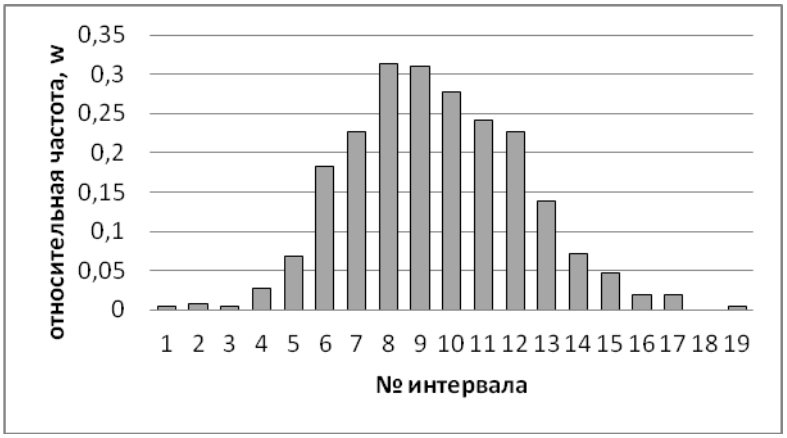
Задача 3: В *Таблице 11* дан ряд распределения мощности коллекторов горизонта $Д_1$ Ромашкинского месторождения по 552 скважинам на Миннибаевской, Абдрахмановской и Павловской площадях. Построить гистограмму относительных частот.

Таблица 11

№	интервалы		m_i
	a	b	
1	0	2	1
2	2	4	2
3	4	6	1
4	6	8	7
5	8	10	17
6	10	12	46
7	12	14	57
8	14	16	79
9	16	18	78
10	18	20	70

№	интервалы		m_i
	a	b	
11	20	22	61
12	22	24	57
13	24	26	35
14	26	28	18
15	28	30	12
16	30	32	5
17	32	34	5
18	34	36	0
19	36	38	1

Решение: В данном статистическом распределении 19 интервалов. Определяем мощность выборки $n=\sum_{i=1}^{19} m_i = 252$. Интервал известен и равен d=2. Значения частот известны, относительные частоты вычисляются как $w_i = \frac{m_i}{n}$.



Гистограмма 3

Статистические оценки параметров распределения.

Пусть требуется изучить количественный признак генеральной совокупности. Обычно в распоряжении исследователя имеются лишь данные выборки $x_1,\ x_2,...,x_n$, полученные в результате n наблюдений. Через эти данные и выражается оцениваемый параметр, который называют статистической оценкой искомого

параметра. Таким образом, статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин, которая в свою очередь также является случайной величиной.

Несмещенной называют статистическую оценку θ^*, математическое ожидание которой равно оцениваемому параметру θ при любом объеме выборки $M(\theta^*) = \theta$.

Эффективной называют статистическую оценку, которая (при заданном объеме выборки n) имеет наименьшую возможную дисперсию.

Состоятельной называют оценку, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру.

Выборочной средней $\bar{x}_B$ называют среднее арифметическое значение признака выборочной совокупности

$$\bar{x}_B = (x_1 + x_2 + \ldots + x_n) \cdot \frac{1}{n},$$

если же значение признака $x_1,\ x_2, \ldots, x_k$ имеют частоты $n_1,\ n_2, \ldots, n_k$, $\sum_{i=1}^{k} n_i = n$, тогда

$$\bar{x}_B = (x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_k \cdot n_k) \cdot \frac{1}{n}.$$

Вычислим

$$M(\bar{x}_B) = (M(x_1) \cdot n_1 + M(x_2) \cdot n_2 + \ldots + M(x_k) \cdot n_k) \cdot \frac{1}{n} =$$
$$= \frac{1}{n} \cdot (a \cdot n_1 + a \cdot n_2 + \ldots + a \cdot n_k) = \frac{a}{n} \cdot n = a.$$

Таким образом, выборочная средняя является несмещенной оценкой. Используя неравенство Чебышева, можно доказать, что оценка $\bar{x}_B$ является состоятельной оценкой.

Выборочной дисперсией D_B называется среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего арифметического значения

$$D_B = \sum_{i=1}^{n} (x_i - \bar{x}_B)^2 \cdot \frac{1}{n}.$$

Если значения же $x_1,\ x_2,...,x_k$ имеют частоты $n_1,\ n_2,...,n_k$, $\sum_{i=1}^{k} n_i = n$, то

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

Выборочным среднеквадратическим отклонением называется $\sigma_B = \sqrt{D_B}$.

Легко доказать, что $D_B = (\overline{x^2})_B - (\bar{x}_B)^2$.

Выборочная оценка является смещенной оценкой генеральной дисперсии, именно:

$M(D_B) = \frac{n-1}{n} D$, где D - генеральная дисперсия.

Введем понятие исправленной дисперсии

$$s^2 = \frac{n}{n-1} D_B = \left[\sum_{i=1}^{k} n_i (x_i - \bar{x}_B)^2 \right] \cdot \frac{1}{n-1}.$$

Доверительный интервал.

Точечной называют оценку, которая определяется одним числом. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т.е. приводить к грубым ошибкам. По этой причине пользуются интервальными оценками.

Пусть количественный признак *X* распределен нормально, причем среднеквадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание *a* по выборочной средней $\bar{x}_B$. Если случайная величина *X* распределена нормально, то выборочная средняя $\bar{x}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения:

$$M(\bar{x}) = a, \quad \sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

Потребуем, чтобы выполнялось условие:

$$P\{|\bar{x} - a| < s\} = \gamma,$$

где γ - заданная надежность.

Если X распределена нормально, то вероятность неравенства $|\bar{x} - a| < s$ вычисляется следующим образом:

$$P\{|\bar{x} - a| < s\} = 2\Phi\left(\frac{s}{\sigma(\bar{x})}\right).$$

Теперь вычислим требуемую вероятность, где $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

$$P\{|\bar{x} - a| < s\} = 2\Phi\left(\frac{s}{\sigma/\sqrt{n}}\right) = 2\Phi\left(\frac{s\sqrt{n}}{\sigma}\right) = 2\Phi(t),$$

здесь $t = \frac{s\sqrt{n}}{\sigma}$, отсюда $s = \frac{t\sigma}{\sqrt{n}}$. Таким образом:

$$P\left\{|\bar{x} - a| < \frac{t\sigma}{\sqrt{n}}\right\} = 2\Phi(t) = \gamma.$$

Вероятность γ задана заранее, она является надежностью полученного результата, $\Phi(t) = \frac{\gamma}{2}$, по таблице для функции Лапласа $\Phi(t)$ находим значение аргумента *t*, подставив *t* в неравенство получаем доверительный интервал:

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}}.$$

Некоторые распределения, связанные с нормальным распределением.

1. *Распределение χ^2 (хи-квадрат).*

Случайной величиной χ^2_k (хи-квадрат с *k* степенями свободы) называется сумма квадратов *k* независимых случайных величин x_1, x_2,..., x_k с одним и тем же простейшим нормальным распределением *N*(0,1). Плотность распределения $\chi^2_k = x_1^2 + x_2^2 + \ldots + x_k^2$ зависит, очевидно, от числа *k*.

$$P_{\chi^2_k}(u) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)} e^{-\frac{u}{2}} u^{\frac{k}{2}-1} \quad (u>0),$$

а центр распределения равен

$$M(\chi_k^2) = \frac{a}{\lambda} = \frac{k/2}{1/2} = k,$$

где $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$ - гамма функция Эйлера.

Это распределение введено Пирсоном. Оно применяется в качестве критерия согласия Пирсона и в задачах математической обработки результатов измерений.

2. *Распределение Стьюдента.*

Распределением Стьюдента называют распределение отношения $T = \frac{X}{\sqrt{\chi_k^2 / k}}$, где величина X распределена нормально $N(0,1)$, а независимая от нее величина $u = \chi_k^2$ имеет плотность $P_{\chi_k^2}(u)$. Параметр k называется числом степеней свободы, закон распределения:

$$P_T(t) = \frac{1}{\sqrt{\pi k}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \qquad (-\infty < t < +\infty), \qquad \text{где } M(T) = 0.$$

Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ.

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднеквадратическое отклонение неизвестно. Рассмотрим случайную величину $T = \frac{\bar{x} - a}{s/\sqrt{n}}$, которая имеет распределение Стьюдента с $k=n-1$ степенями свободы (т.к. здесь добавляется еще одна формула для выборочной дисперсии), где $\bar{X}$ – выборочная средняя, s – исправленное среднеквадратическое отклонение. Напомним, что распределение Стьюдента определяется объемом выборки n и не зависит от a и σ, что является существенной особенностью, т.к. эти величины нам неизвестны

$$P\left\{\frac{|\bar{x} - a|}{s/\sqrt{n}} < t_\gamma\right\} = 2\int_0^{t_\gamma} P_T(t,n)dt = \gamma,$$

в силу того, что плотность распределения четная функция. Определив t_γ по таблицам распределения Стьюдента, получим доверительный интервал:

$$P\left\{\bar{x}-\frac{t_\gamma s}{\sqrt{n}}<a<\bar{x}+\frac{t_\gamma s}{\sqrt{n}}\right\}=\gamma,$$

покрывающий параметр a с надежностью γ.

Замечание: При возрастании n распределение Стьюдента стремится к нормальному распределению. Поэтому при $n>30$ можно вместо распределения Стьюдента пользоваться нормальным распределением.

Задача 4: Вычислить точечные оценки математического ожидания и дисперсии, а также найти доверительный интервал, соответствующий доверительной вероятности P=0,95 для распределения содержания железа в руде по данным *Таблицы 12*. Допустить, что содержание железа подчиняется нормальному закону.

Таблица 12.

Интервалы %	Частоты, m_i
28-32	1
32-36	9
36-40	29
40-44	55
44-48	72
48-52	56
52-56	27
56-60	7
60-64	1

Решение: Из значений случайной величины, распределенной по нормальному закону с неизвестными математическим ожиданием и дисперсией, сделана выборка, представленная в виде интервалов и частот. Объем выборки равен сумме всех частот $n=\sum_{i=1}^{9} m_i = 257$. Для проведения статистических вычислений определяем середины интервалов, и все промежуточные вычисления запишем в *Таблице 13.*

Таблица 13.

Исходные данные				Вычисления		
№	Интервалы, %		Частоты, m_i	$x_i=\frac{a_1+a_2}{2}$	$x_i \cdot m_i$	$(x_i-\bar{x})^2 \cdot m_i$
	a_1	a_2				
1	28	32	1	30	30	251,54
2	32	36	9	34	306	1265,92
3	36	40	29	38	1102	1791,57
4	40	44	55	42	2310	819,44
5	44	48	72	46	3312	1,41
6	48	52	56	50	2800	959,85
7	52	56	27	54	1458	1789,04
8	56	60	7	58	406	1031,67
9	60	64	1	62	62	260,50
Объем выборки $n=\sum_{i=1}^{9} m_i = 257$						

Среднее значение $\bar{x} = \frac{\sum_{i=1}^{9} x_i \cdot m_i}{n} = 45{,}86$
Исправленная дисперсия $s^2 = \frac{\sum_{i=1}^{9} (x_i - \bar{x})^2 \cdot m_i}{n-1} = 31{,}92$

Несмещенной оценкой математического ожидания в этом случае является среднее значение $\bar{x} = \frac{\sum_{i=1}^{9} x_i \cdot m_i}{n} = 45{,}86$, а для дисперсии – исправленная несмещенная оценка

$$s^2 = \frac{\sum_{i=1}^{9} (x_i - \bar{x})^2 \cdot m_i}{n-1} = 31{,}92.$$

1) Для вычисления интервальной оценки математического ожидания a используется случайная величина $t = \frac{\bar{x} - a}{s/\sqrt{n}}$, подчиняющаяся распределению Стьюдента со степенью свободы $k = n–1 = 257–1 = 256$.

Из таблиц Стьюдента для $\gamma = 0{,}95$ (т.е. для $q = 1 - \gamma = 0{,}05$) и 256 степеней свободы находим $t_\gamma = 1{,}96$. Значит, с доверительной вероятностью $\gamma = 0{,}95$ величина t находится в интервале (-1,96; 1,96), т.е.

$$-t_\gamma < \frac{\bar{x} - a}{s/\sqrt{n}} < t_\gamma \qquad \text{или} \qquad -1{,}96 < \frac{45{,}86 - a}{\sqrt{31{,}92}/\sqrt{257}} < 1{,}96.$$

Следовательно, доверительный интервал равен

$$45{,}17 < a < 46{,}55.$$

2) Для вычисления интервальной оценки дисперсии с надежностью $\gamma = 0{,}95$ используется случайная величина $\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$, подчиняющаяся χ^2 распределению со степенями свободы $k = n–1$. Доверительные пределы χ_1^2, χ_2^2 интервала $\chi_1^2 < \chi^2 < \chi_2^2$ находим из условия

$$\int_0^{\chi_1^2} P_{\chi^2}(t)\, dt = \frac{1-\gamma}{2}, \quad \int_{\chi_2^2}^{\infty} P_{\chi^2}(t)\, dt = \frac{1-\gamma}{2},$$

где $P_{\chi^2}(t)$ - плотность вероятности хи-квадрат распределения. С вероятностью γ находим, что

$$\chi_1^2 < \frac{(n-1)\,s^2}{\sigma^2} < \chi_2^2 \quad \text{или} \quad \frac{(n-1)\,s^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)\,s^2}{\chi_1^2}.$$

С помощью таблиц для распределения хи-квадрат с k=256 и $q = \frac{1-\gamma}{2} = 0{,}025$ находим $\chi_1^2(\gamma) = 210$ и для $1-q = 1-0{,}025 = 0{,}975$ находим $\chi_2^2(\gamma) = 280$. Следовательно, доверительный интервал с надежностью $\gamma = 0{,}95$ для несмещенной исправленной дисперсии $s^2 = 31{,}92$ будет равен

$$\frac{(n-1)\,s^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)\,s^2}{\chi_1^2} \text{ или } \frac{256 \cdot 31{,}92}{280} < \sigma^2 < \frac{256 \cdot 31{,}92}{210} \text{ или } 29{,}184 < \sigma^2 < 38{,}912.$$

Задача 5: В *Таблице 14* дано распределение пористости K породы одного из месторождений Татарстана. Найти интервальную оценку математического ожидания с доверительной вероятностью $\gamma = 0{,}95$, используя

$$\gamma = 0{,}95 = \frac{2}{\sqrt{2\pi}} \int_0^{t_\gamma} e^{-t^2/2} dt\,.$$

Таблица 14

Середина интервала	Частота, m_i	Середина интервала	Частота, m_i	Середина интервала	Частота, m_i
10	4	18	51	26	10
12	9	20	45	28	5
14	26	22	32		
16	36	24	19		

Решение: Вычисления будем производить в *Таблице 15*, записывая

1) исходные данные: порядковый номер, середина интервала K_i, частота m_i - вариационный ряд;

2) промежуточные вычисления: $K_i m_i$ и $(K_i - K_{cp})^2 m_i$ - взвешенные значения.

Таблица 15.

Исходные данные			Вычисления	
№	K_i	Частот, m_i	$K_i m_i$	$(K_i - K_{cp})^2 m_i$
1	10	4	40	321,86
2	12	9	108	421,55
3	14	26	364	610,04
4	16	36	576	291,16
5	18	51	918	36,32
6	20	45	900	60,15
7	22	32	704	318,75
8	24	19	456	505,13
9	26	10	260	512,10
10	28	5	140	419,17
$n = \sum_{i=1}^{9} m_i = 237$				
$K_{cp} = \overline{K} = \frac{\sum_{i=1}^{10} K_i \cdot m_i}{n} = 18{,}84$				
$s^2 = \frac{\sum_{i=1}^{10} (K_i - \overline{K})^2 m_i}{n-1} = 14{,}78$				

Объем выборки n=237 равен сумме всех частот, математическое ожидание случайной величины пористости K_i определяется как взвешенное среднее выборки $K_{cp} = \overline{K} = 18{,}84$. Дисперсия выборки неизвестна, вычисляем исправленную средне взвешенную дисперсию

$$s^2 = \frac{\sum_{i=1}^{10} (K_i - \overline{K})^2 m_i}{n-1} = 14{,}78$$

и среднеквадратичное отклонение $s = \sqrt{s^2} = 3{,}84$. Для вычисления интервальной оценки математического ожидания используется случайная величина $t = \frac{\overline{x} - a}{s/\sqrt{n}}$, подчиняющаяся распределению Стьюдента со степенью свободы

$$k = n–1 = 237–1 = 236.$$

Из заданной вероятности $\gamma = 0{,}95$ по таблице Стьюдента для $k = n–1 = 236$ и $q = 1 - \gamma = 0{,}05$ находим $t_\gamma = 1{,}96$, вычисляем точность оценивания математического ожидания, подставляя значения $s = 3{,}84$; $\sqrt{n} = \sqrt{237} = 15{,}39$ в формулу

$$-t_\gamma < \frac{\overline{K} - a}{s/\sqrt{n}} < t_\gamma \qquad \text{или} \qquad -1{,}96 < \frac{18{,}84 - a}{3{,}84/15{,}39} < 1{,}96.$$

Следовательно, доверительный интервал с надежностью $\gamma = 0{,}95$ равен

$$18{,}35 < a < 19{,}33.$$

Задача 6: Для определения петрографического типа неогеновых лав одного из районов России отобрано и проанализировано содержание SiO_2 (%)30 образцов. Содержание SiO_2 приведено в *Таблице 16.*

Таблица 16

SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2	SiO_2
59,5	69,2	69,2	61,2	71,4	67,5	72,5	67,8	63,7	56,6
66,8	61,2	62,4	69,3	67,7	65,3	64,6	61,6	79,2	63,8
60,5	66,3	71,6	64,6	63,6	69,9	63,1	73,2	65,8	60,7

Как известно, вулканические породы классифицируются по содержанию SiO_2 на типы пород *Таблица 16а*. Определить доверительный интервал математического ожидания с доверительной вероятностью 0,95. По доверительному интервалу отнести исследуемые лавы к определенному типу пород.

Таблица 16а

Содержание SiO_2	Типы пород	Содержание SiO_2	Типы пород
47,0 – 52,0	Бальзаты	63,0 – 68,5	Дациты
52,0 – 57,2	Андезито-бальзаты	68,5 – 70,5	Липарито-дациты
57,2 – 62,1	Андезиты	более 70,5	Липариты
62,1 –63,0	Андезит-дациты		

Решение: В каждом интервале определим число частот случайных величин и запишем результаты в *Таблице 17*, вычислим середины интервалов x_i, взвешенные среднit в интервале $x_i \cdot m_i$ и $(x_i - \bar{x})^2 \cdot m_i$. Определяем статистические параметры

– объем выборки $n = 30$;

– взвешенное среднее $\bar{x} = 66$;

– несмещенную дисперсию $s = 24{,}45$.

Для вычисления интервальной оценки математического ожидания используется случайная величина $t = \dfrac{\bar{x} - a}{s/\sqrt{n}}$, подчиняющаяся распределению Стьюдента со степенью свободы $k = n–1 = 29$.

Таблица 17.

№	Интервалы		m_i	$x_i = \frac{a+b}{2}$	$x_i \cdot m_i$	$(x_i - \bar{x})^2 \cdot m_i$
	a	b				
1	47	52	0	49,5	0	0
2	52	57,2	1	54,6	54,6	117,72
3	57,2	63	6	59,65	357,9	201,84
4	63	68,5	14	65,75	920,5	1,26
5	68,5	70,5	4	69,5	278	65,61
6	70,5		5	73,6	352,5	127,51
$n = \sum_{i=1}^{9} m_i = 30$						
$\bar{x} = \frac{\sum_{i=1}^{9} x_i \cdot m_i}{n} = 66,0$						
$s^2 = \frac{\sum_{i=1}^{9} (x_i - \bar{x})^2 \cdot m_i}{n-1} = 24,45$						

Из заданной вероятности $\gamma = 0,95$ по таблице Стьюдента для k = 29 и $q = 1 - \gamma = 0,05$ находим $t_\gamma = 2,05$, вычисляем точность оценивания математического ожидания, подставляя значения $s = 24,45$; $\sqrt{n} = \sqrt{30} = 5,385$ в формулу

$$-t_\gamma < \frac{\bar{x} - a}{s/\sqrt{n}} < t_\gamma \qquad \text{или} \qquad -2,05 < \frac{66,0 - a}{4,94/5,385} < 2,05.$$

Следовательно, доверительный интервал с надежностью $\gamma = 0,95$ равен

$$63,87 < a < 67,03.$$

Этот интервал сравниваем с интервалами *Таблицы 16а* и устанавливаем, что наша выборка относится к дацитам.

Критерий согласия χ^2 (критерий согласия Пирсона).

Все рассмотренные выше задачи решались в предположении нормального распределения результатов эксперимента. Но иногда это предположение приходится подвергать сомнению. Если гистограмма эмпирического распределения заметно отличается от кривых нормального распределения, то возникает вопрос, можно ли объяснить это отличие случайными ошибками эксперимента. В противном случае надо искать другой закон распределения, более согласующийся с результатами эксперимента. Надежным способом проверки соответствия результатов эксперимента

предполагаемому теоретическому распределению *N(a, σ)* является критерий согласия χ^2, разработанный английским ученым К. Пирсоном. Изложим этот критерий в применении к проверке гипотезы о нормальном распределении.

Разобьем ось *OX* на *l* интервалов:

$$(-\infty, x_1),\ (x_1, x_2), ..., (x_{l-1}, +\infty)$$

и проведем *n* независимых измерений эмпирических значений исследуемой величины. Подсчитаем число m_i результатов, попавших в *i*-тый интервал, и вычислим по формуле:

$$p'_i = P\{x_i < x < x_{i+1}\} = P\left\{\frac{x_i - a}{\sigma} < x < \frac{x_{i+1} - a}{\sigma}\right\} = \left(\Phi(t_2) - \Phi(t_1)\right),$$

где $t_1 = \frac{x_1 - a}{\sigma}$, $t_2 = \frac{x_2 - a}{\sigma}$, $\qquad m'_i = p'_i \cdot n$.

Таким образом, мы определили теоретические частоты. Теоретические частоты также можно определить следующим образом:

$$m'_i = \frac{n \cdot d}{s} f(t_i),$$

где *d* – длина интервала, $t_i = \frac{x_i - \bar{x}}{s}$, $f(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$, $\bar{x}$ - выборочная средняя и s^2 - выборочная дисперсия.

Как следует из теоремы Муавра-Лапласа, при большом числе испытаний *n* каждая величина m_i имеет асимптотически нормальное распределение с центром np_i и стандартом $\sqrt{np_i q_i}$, $q_i = 1 - p_i$. Поэтому распределение нормированных величин

$$y_i = \frac{m_i - np_i}{\sqrt{np_i q_i}},\ i = 1,2,\ldots,l$$

будет близко к простейшему нормальному распределению. Если бы величины $y_1, y_2, ..., y_l$ были независимыми, то распределение суммы их квадратов было бы близко χ^2 распределению. Но эти величины связаны линейной зависимостью

$$\sum_{i=1}^{l} y_i \sqrt{np_i q_i} = \sum_{i=1}^{l} m_i - n\sum_{i=1}^{l} p_i = n - n = 0.$$

Оказывается, что если каждый квадрат y_i^2 умножить на q_i, то распределение суммы:

$$\sum_{i=1}^{l} y_i^2 q_i = \sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i} \qquad (*)$$

будет стремиться к χ^2 распределению с $l-1$ степенью свободы при $n \to \infty$.

По распределению Пирсона находят критическое значение t_γ, для которого

$$P\{u > u_1\} = \int_{t_\gamma}^{\infty} P_{\chi^2}(u)\, du = 1 - \gamma \quad (k = l - 1),$$

где γ - заданная надежность вывода (и, значит, $1-\gamma$ - пренебрежимо малая вероятность). Если сумма (*) окажется больше этого критического значения, то с надежностью γ можно считать, что проверяемое нормальное распределение не согласуется с результатами эксперимента, т.е. гипотезу о нормальном распределении признака *X* следует отвергнуть. Число степеней свободы находят по формуле $k = l - 1 - r$, где l - число интервалов, r - число параметров предполагаемого распределения, которые оцениваются по данным выборки.

Задача 7: В качестве примера рассмотрим проверку гипотезы о нормальном распределении логарифма проницаемости ($y = \lg x$) пласта горизонта *Д1* по данным 100 скважин (*Таблица 18*).

Таблица 18

Исходные данные			Вычисления								
№	Интервалы $y_i = \lg x$	m_i	Середин. интерв. y_i	$y_i \cdot m_i$	$(y_i - \bar{y})^2 \cdot m_i$	$\lvert t_i \rvert$	$f(t_i)$	m_{iT}	m'_{iT}	m_{iT_2}	χ_i^2
1	1,3-1,6	3	1,45	4,35	1,7787	2,27	0,0303	2,7	3		
2	1,6-1,9	13	1,75	22,75	2,8717	1,39	0,1518	13,4	13	16	0
3	1,9-2,2	33	2,05	67,65	0,9537	0,50	0,3521	31,1	31	31	0,1290
4	2,2-2,5	30	2,35	70,5	0,5070	0,38	0,3712	32,8	33	33	0,2727
5	2.5-2,8	17	2,65	45,05	3,1433	1,27	0,1781	15,9	16	20	0,05
6	2,8-3,1	4	2,95	11,8	2,136	2,15	0,0396	3,5	4		
$n = \sum_{i=1}^{6} m_i = 100$			$\bar{y} = \frac{\sum_{i=1}^{6} y_i \cdot m_i}{n} = 2,22$		$s^2 = \frac{\sum_{i=1}^{6} (y_i - \bar{y})^2 \cdot m_i}{n-1} = 0,1150$			99,4	100	100	0,4518

Для того, чтобы проверить гипотезу о нормальном распределении генеральной совокупности:

1) Вычисляем объем выборки *n*, среднее значение в каждом интервале y_i, взвешенное среднее - $y_i m_i$, выборочное среднее $\bar{y} = 2,22$ и дисперсию $s^2 = 0,1150$.
2) По предполагаемому закону распределения вычисляем теоретические частоты

$$m_{iT} = \frac{n \cdot d}{s} f(t_i),$$

где $t_i = \frac{y_i - \bar{y}}{s}$, $f(t_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t_i^2}{2}\right)$, n=100 – объем выборки, d=0,3 – длина интервала, $s = \sqrt{s^2} = \sqrt{0,1150} = 0,3391$ - среднеквадратичное отклонение. Закон нормального распределения является четной функцией, поэтому ограничимся вычислением $\lvert t_i \rvert$. Вычисленные m_{iT} округляются так, чтобы выполнялось условие $\sum m'_{iT} = n$. Если m_{iT} и m'_{iT} меньше пяти, то их группируют с соседними частотами (см. *Таблицу 18* столбец m_{iT_2}).

3) Вычисляем $\chi_i^2 = \dfrac{(m_i - m_{iT_2})^2}{m_{iT_2}}$ и, суммируя, определяем наблюдаемое значение критерия согласия $\chi^2 = \sum_{i=1}^{l} \dfrac{(m_i - m_{iT_2})^2}{m_{iT_2}} = 0{,}4518,$ где $l = 4$ – число интервалов после группировки.

4) Для степеней свободы $k = l - 3$ и уровня значимости $q = 0{,}05$ находим по таблице $\chi^2_{q,k} = \chi^2_{0{,}05} = 3{,}841$. Если $\chi^2 < \chi^2_{q,k}$, то предполагаемый закон принимается как не противоречащий результатам эксперимента. Для нашего случая:

$$\chi^2 = 0{,}4518 < \chi^2_{q,k} = 3{,}841,$$

т.е. закон нормального распределения можно принять в качестве статистической модели распределения логарифма проницаемости пород. При этом $\sigma = 0{,}3391$, $M(y) = 2{,}22$, т.е.

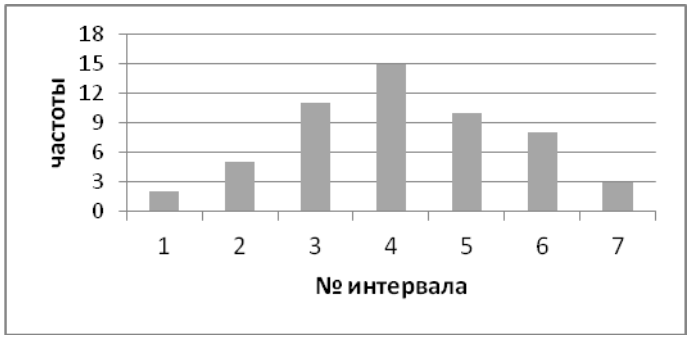
$$f(y) = \frac{1}{0{,}3391\sqrt{2\pi}} \exp\left(-\frac{(y - 2{,}22)^2}{2 \cdot 0{,}1150} \right).$$

Задача 8: Дано распределение содержания битума в породах уфимского яруса восточного борта Мелекесской впадины (*Таблица 19*). Проверить гипотезу о нормальном законе распределения содержания битума в породах по критерию Пирсона.

Таблица 19

№	Интервалы		m_i
	a	*b*	
1	0	0,05	2
2	0,05	0,1	5
3	0,1	0,15	11
4	0,15	0,2	15
5	0,2	0,25	10
6	0,25	0,3	8
7	0,3	0,35	3

Гистограмма частот



Решение: Построим гистограмму частот содержания битума.

Для того, чтобы проверить гипотезу о нормальном распределении генеральной совокупности проведем промежуточные вычисления и запишем их в *Таблицу 20*.

Таблица 20.

Исходные данные			Вычисления								
№	Интервал	m_i	Серед. интерв. y_i	$y_i \cdot m_i$	$(y_i - \bar{y})^2 \cdot m_i$	$\lvert t_i \rvert$	$f(t_i)$	m_{iT}	m'_{iT}	m_{iT_2}	χ_i^2
1	0 - 0,05	2	0,025	0,05	0,0496	2,1205	0,0421	1,5	2		
2	0,05 - 0,1	5	0,075	0,375	0,0577	1,4469	0,1400	5,1	5	8	0,125
3	0,1 - 0,15	11	0,125	1,375	0,0363	0,7734	0,2958	10,8	11	11	0,000
4	0,15 - 0,2	15	0,175	2,625	0,0008	0,0998	0,3970	14,4	14	14	0,071
5	0,2 - 0,25	10	0,225	2,25	0,0181	0,5738	0,3384	12,3	12	12	0,333
6	0,25 - 0,3	8	0,275	2,2	0,0686	1,2474	0,1832	6,7	7	9	0,444
7	0,3 - 0,35	3	0,325	0,975	0,0610	1,9209	0,0630	2,3	2		
$n = \sum_{i=1}^{6} m_i = 54$			$\bar{y} = \frac{\sum_{i=1}^{6} y_i \cdot m_i}{n} = 0{,}182$		$s^2 = \frac{\sum_{i=1}^{6} (y_i - \bar{y})^2 \cdot m_i}{n-1} = 0{,}0055$			53,1	53	54	0,974

Для того, чтобы проверить гипотезу о нормальном распределении генеральной совокупности:

1) Вычисляем объем выборки n, среднее значение в каждом интервале y_i, взвешенное среднее - $y_i m_i$, выборочное среднее $\bar{y} = 0{,}182$ и дисперсию $s^2 = 0{,}0055$.

2) По предполагаемому закону распределения вычисляем теоретические частоты

$$m_{iT} = \frac{n \cdot d}{s} f(t_i),$$

где $t_i = \frac{y_i - \bar{y}}{s}$, $f(t_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t_i^2}{2}\right)$, n=54 – объем выборки, d=0,05 – длина интервала, $s = \sqrt{s^2} = \sqrt{0{,}0055} = 0{,}0742$ - среднеквадратичное отклонение. Закон нормального распределения является четной функцией, поэтому ограничимся вычислением $|t_i|$. Вычисленные m_{iT} округляются так, чтобы выполнялось условие $\sum m'_{iT} = n$. Если m_{iT} и m'_{iT} меньше пяти, то их группируют с соседними частотами (см. *Таблицу 20* столбец m_{iT_2}).

3) Вычисляем $\chi_i^2 = \frac{(m_i - m_{iT_2})^2}{m_{iT_2}}$ и, суммируя, определяем наблюдаемое значение критерия согласия $\chi^2 = \sum_{i=1}^{l} \frac{(m_i - m_{iT_2})^2}{m_{iT_2}} = 0{,}974,$ где $l = 5$ – число интервалов после группировки.

4) Для степеней свободы $k = l - 3 = 2$ и уровня значимости $q = 0{,}05$ находим по таблице $\chi^2_{q,k} = \chi^2_{0{,}05} = 6$. Т.к.

$$\chi^2 = 0{,}974 < \chi^2_{q,k} = 6,$$

то закон нормального распределения можно принять в качестве статистической модели распределения содержания битума в породах. При этом $\sigma = 0{,}0742$, $M(y) = 0{,}182$, т.е.

$$f(y) = \frac{1}{0{,}0742\sqrt{2\pi}} \exp\left(-\frac{(y - 0{,}182)^2}{2 \cdot 0{,}0742}\right).$$

Упражнение2. По результатам *Таблицы 21* проверить гипотезу о нормальном законе распределения содержания железа в железной руде по критерию Пирсона.

Таблица 21

интервалы		частоты	Интервалы		частоты	интервалы		частоты
28	32	1	40	44	55	52	56	27
32	36	9	44	48	72	56	60	7
36	40	29	48	52	56	60	64	1

Упражнение 3. В *Таблице 22* дано распределение мощности (в см) пласта. Проверить гипотезу о нормальном законе распределения мощности пласта по критерию Пирсона.

Таблица 22

Интервалы (см)	частоты	интервалы(см)	частоты	интервалы(см)	частоты
40 – 50	2	80 – 90	33	120 – 130	86
50 – 60	5	90 – 100	32	130 – 140	19
60 – 70	16	100 – 110	21	140 – 150	9
70 – 80	74	110 – 120	71	150 – 160	2

Коэффициент линейной корреляции

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины y от одной или нескольких случайных величин.

Пусть изучается система количественных признаков (X, Y). В результате n независимых опытов получены n пар чисел (x_1, y_1), (x_2, y_2),..., (x_n, y_n). Найдем по данным наблюдениям выборочное уравнение прямой линии среднеквадратической регрессии.

Для определенности будем искать уравнение $\bar{y}_x = kx + b$ регрессии Y на X. Исщем выборочное уравнение регрессии y на X вида:

$$y = \rho_{xy} x + b.$$

Подберем параметры ρ_{xy} и b так, чтобы сумма квадратов отклонений была минимальной. Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих параметров

$$F(\rho_{xy}, b) = \sum_{i=1}^{n} (\bar{y}_i - y_i)^2 \text{ или } F(\rho_{xy}, b) = \sum_{i=1}^{n} (\rho_{xy}\ x_i + b - y_i)^2$$

Чтобы найти минимальное значение этой функции, приравняем к нулю частные производные:

$$\frac{\partial F}{\partial \rho_{xy}} = 2\sum_{i=1}^{n}(\rho_{xy}\ x_i + b - y_i)\, x_i = 0,$$

$$\frac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(\rho_{xy}\ x_i + b - y_i) = 0.$$

Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно ρ_{xy} и b.

$$\left(\sum_{i=1}^{n} x_i^2\right)\rho_{xy} + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} x_i\, y_i\,,$$

$$\left(\sum_{i=1}^{n} x_i\right)\rho_{xy} + nb = \sum_{i=1}^{n} y_i.$$

Решив эту систему, найдем ρ_{xy} и b:

$$\rho_{xy} = \left(n\sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) \Bigg/ \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right),$$

$$b = \left(n\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i\right) \Bigg/ \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right).$$

Задача 9: В 21 образце гранитоидов Урала определено процентное содержание фтора в биотите (*X*) и сфене (*Y*) результаты приведены в *Таблице 23.* Найти уравнение связи в виде $y = ax + b$ и коэффициент линейной связи (корреляции).

Таблица 23.

X (биотит)	*Y* (сфен)	*X* (биотит)	*Y* (сфен)	*X* (биотит)	*Y* (сфен)
0,25	0,27	0,88	0,48	1,57	0,77
0,26	0,36	0,98	0,52	1,72	1,17
0,72	0,52	0,98	0,60	1,98	1,30
0,78	0,53	1,01	0,63	2,12	0,81
0,80	0,37	1,09	0,43	2,28	1,29
0,84	0,51	1,16	0,59	2,59	1,25
0,86	0,43	1,42	0,77	3	1,36

Решение: Вычисляются величины:

1) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$,

2) $S_x^2 = \overline{x^2} - \bar{x}^2$, $S_y^2 = \overline{y^2} - \bar{y}^2$, $S_x = \sqrt{S_x^2}$, $S_y = \sqrt{S_y^2}$,

3) $r_{xy} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x S_y}$,

4) $a = r_{xy} \cdot \dfrac{S_x}{S_y}$, $b = \bar{y} - r_{xy} \cdot \dfrac{S_x}{S_y} \bar{x}$.

Уравнение связи имеет вид:

$$y - \bar{y} = r_{xy} \cdot \frac{S_x}{S_y}(x - \bar{x}).$$

Величина r_{xy} называется коэффициентом линейной корреляции и является мерой связи. Если связь функциональная, то $|r_{xy}| = 1$, если $r_{xy} = 0$, то связь отсутствует. При $|r_{xy}| < 1$ связь является вероятностной (стохастической). При проверке гипотезы $H_0: \rho = 0$ используется критерий Стьюдента, если *x* и *y* подчиняются закону нормального распределения.

По данным *Таблицы 23* получаем (промежуточные вычисления приведены в *Таблице 24*):

1) $\bar{x} = 1{,}2995$; $\bar{y} = 0{,}7124$; $\overline{xy} = 1{,}1532$; $\bar{x}^2 = 2{,}2072$; $\bar{y}^2 = 0{,}6240$.

2) $S_x^2 = 0{,}5185$; $S_y^2 = 0{,}1165$; $S_x = 0{,}7201$; $S_y = 0{,}3413$.

3) $r_{xy} = 0{,}9255$; $a = 0{,}4387$; $b = 0{,}1423$.

Таблица 24.

№	x_i	y_i	x_i^2	y_i^2	$x_i\, y_i$
1	0,25	0,27	0,0625	0,0729	0,0675
2	0,26	0,36	0,0676	0,1296	0,0936
3	0,72	0,52	0,5184	0,2704	0,3744
4	0,78	0,53	0,6084	0,2809	0,4134
5	0,8	0,37	0,64	0,1369	0,296
6	0,84	0,51	0,7056	0,2601	0,4284
7	0,86	0,43	0,7396	0,1849	0,3698
8	0,88	0,48	0,7744	0,2304	0,4224
9	0,98	0,52	0,9604	0,2704	0,5096

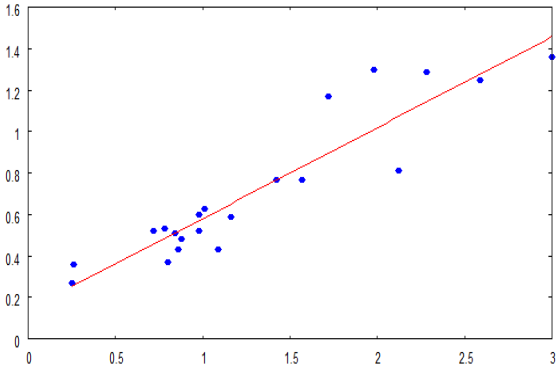
10	0,98	0,6	0,9604	0,36	0,588
11	1,01	0,63	1,0201	0,3969	0,6363
12	1,09	0,43	1,1881	0,1849	0,4687
13	1,16	0,59	1,3456	0,3481	0,6844
14	1,42	0,77	2,0164	0,5929	1,0934
15	1,57	0,77	2,4649	0,5929	1,2089
16	1,72	1,17	2,9584	1,3689	2,0124
17	1,98	1,3	3,9204	1,69	2,574
18	2,12	0,81	4,4944	0,6561	1,7172
19	2,28	1,29	5,1984	1,6641	2,9412
20	2,59	1,25	6,7081	1,5625	3,2375
21	3	1,36	9	1,8496	4,08
Среднее	1,2995	0,7124	2,2072	0,624	1,1532

Уравнение связи имеет вид

$$y = 0{,}4387x + 0{,}1423;$$

$$t = \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} \cdot \sqrt{n-2} =$$

$$= \frac{0{,}9255}{\sqrt{1-(0{,}9255)^2}} \sqrt{19} = 10{,}7.$$



Степень свободы для критерия Стьюдента равна $k = n - 2 = 19$. Выбирая уровень значимости $q = 0{,}05$, по таблице находим $t_{q,k} = t_{0{,}05} = 2{,}09$. Так как $t = 10{,}7 > 2{,}09$, то гипотеза H_0: $\rho = 0$ отвергается, т.е. связь является устойчивой.

Задача 10: В *Таблице 25* приведено содержание в процентах *Al* и *Fe* в 7 лунных пробах. Вычислить коэффициент корреляции и уравнение связи в виде $y = a \cdot x + b$. По *t* -критерию Стьюдента проверить гипотезу об отсутствии связи между *Al* и *Fe* в 7 лунных пробах.

Таблица 25

X (*Al*,%)	5,9	4	4	5,4	6,2	5,7	6,0
Y (*Fe*,%)	14,7	15,7	15,4	15,2	13,2	14,8	13,8

Решение:

Вся процедура принятия решения состоит из трех частей:

1) Вычисляем выборочные средние случайных величин *X* (*Al*) и *Y* (*Fe*) – промежуточные результаты запишем в *Таблице 26.*

Таблица 26.

№	x_i (*Al*)	y_i (*Fe*)	x_i^2	y_i^2	$x_i\, y_i$
1	5,9	14,7	34,81	216,09	86,73
2	4,0	15,7	16	246,49	62,8
3	4,0	15,4	16	237,16	61,6
4	5,4	15,2	29,16	231,04	82,08
5	6,2	13,2	38,44	174,24	81,84
6	5,7	14,8	32,49	219,04	84,36
7	6,0	13,8	36	190,44	82,8
Среднее	$\bar{x} = 5{,}31$	$\bar{y} = 14{,}69$	$\overline{x^2} = 28{,}99$	$\overline{y^2} = 216{,}36$	$\overline{xy} = 77{,}46$

2) Находим дисперсию $S_x^2 = \overline{x^2} - \bar{x}^2 = 0{,}7441$; среднеквадратическое отклонение $S_x = 0{,}8626$ выборки *x* и, соответственно, $S_y^2 = \overline{y^2} - \bar{y}^2 = 0{,}6869$ и $S_y = 0{,}8288$ выборки *y*. Коэффициент корреляции равен

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x S_y} = -0{,}81896,$$

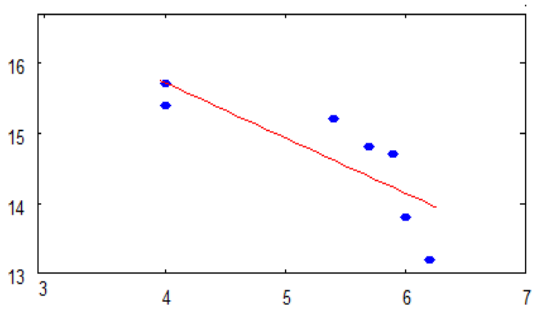
$$a = r_{xy} \cdot \frac{S_x}{S_y} = -0{,}7869,$$

$$b = \bar{y} - r_{xy} \cdot \frac{S_x}{S_y} \bar{x} = 14{,}69 + 0{,}81896 \cdot \frac{0{,}8288}{0{,}8626} \cdot 5{,}31 = 18{,}8675.$$

Таким образом, получено уравнение зависимости $y = -0{,}7869 \cdot x + 18{,}8675$ между содержанием *Y* (*Fe*) и *X* (*Al*) в 7 лунных пробах с $r_{xy} = -0{,}81896$.

3) Необходимо проверить гипотезу об отсутствии связи $H_0: \rho = 0$. Для этого определяем наблюдаемую величину

$$t = \frac{|r_{xy}|}{\sqrt{1-r_{xy}^2}} \cdot \sqrt{n-2} = \frac{|-0{,}81896|}{\sqrt{1-(-0{,}81896)^2}} \sqrt{7-2} = 3{,}19 .$$



Вычисляем степень свободы $k = n - 2 = 5$ и для уровня значимости $q = 0{,}05$ по таблице распределения Стьюдента находим критическую точку $t_{q,k} = 2{,}57$ и сопоставляем с эмпирическим значением. Так как $t = 3{,}19 > 2{,}57$, то $H_0: \ \rho = 0$ отвергается, т.е. X и Y имеют устойчивую связь.

Упражнение 4. В 10 пробах угля определено значение удельного веса (X г/см^3) и зольности (Y %) *Таблица 27* . По критерию Стьюдента проверить гипотезу об отсутствии связи между X и Y. Вычислить параметры a и b в уравнении связи

$$y = a \cdot x + b,$$

если эта связь имеется.

Таблица 27.

X	1,2	1,3	1.5	1,3	1,7	1,4	1,5	1,8	1,4	1,6
Y	4	7	24	5	32	20	25	36	15	24

Упражнение 5. На месторождениях колумбита для ряда образцов определялось содержание X – ZrO_2 и Y – Nb_2O_5. Получены следующие результаты (*Таблица 28*). Найти параметры уравнения $y = a \cdot x + b$, проверить гипотезу $H_0: \ \rho = 0$.

Таблица 28.

X	0,02	0,2	0,6	0,5	0,3	0,3	0,7	0,5	0,7	0,4	0,4	0,9
Y	0,06	0,06	0,22	0,18	0,14	0,06	0,26	0,12	0,3	0,14	0,18	0,34

Задача 11: При проверке гидрогеологических исследований в профиле пробурено 12 скважин (№) и выполнены опытные работы. Для оценки эффективности метода необходимо установить, существует ли зависимость между электрическим сопротивлением пород $Y(\rho_k,$ см) и относительной мощностью $X(m_2,$ %) горизонта гравийно-галечных отложений, к которым приурочены основные водоносные горизонты. Результаты приведены в *Таблице 29*.

Таблица 29.

№	1	2	3	4	5	6	7	8	9	10	11	12
X	67	80	40	24	25	38	18	72	44	51	76	50
Y	253	115	126	82	66	25	44	180	32	319	421	51

Вычислить параметры уравнения связи $y = a \cdot x + b$ и коэффициент корреляции. По критерию Стьюдента выяснить, существенна ли зависимость между электрическим сопротивлением и мощностью горизонта.

Решение: Для того, чтобы при заданном уровне значимости q проверить гипотезу о характере зависимости между электрическим сопротивлением и мощностью горизонта, необходимо:

1) Вычислить выборочные средние случайных величин $X\,(m)$ и $Y\,(\rho)$ – промежуточные результаты записать в *Таблице 30.*

Таблица 30.

№	$x_i\,(m)$	$y_i\,(\rho)$	x_i^2	y_i^2	$x_i\,y_i$
1	67	253	4489	64009	16951
2	80	115	6400	13225	9200
3	40	126	1600	15876	5040
4	24	82	576	6724	1968
5	25	66	625	4356	1650
6	38	25	1444	625	950
7	18	44	324	1936	792
8	72	180	5184	32400	12960
9	44	32	1936	1024	1408
10	51	319	2601	101761	16269

11	76	421	5776	177241	31996
12	50	51	2500	2601	2550
среднее значение	$\bar{x} = 48{,}75$	$\bar{y} = 142{,}83$	$\overline{x^2} = 2787{,}92$	$\overline{y^2} = 35148{,}17$	$\overline{xy} = 8477{,}83$

2) Оценить дисперсию и среднеквадратическое отклонение выборки $S_x^2 = 411{,}35;\ S_y^2 = 14746{,}81;\ S_x = 20{,}28;\ S_y = 121{,}44$. Подставить полученные результаты в формулу коэффициента корреляции $r_{xy} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x S_y} = 0{,}61$, параметров $a = r_{xy} \cdot \dfrac{S_x}{S_y} = 3{,}68,\ b = \bar{y} - r_{xy} \cdot \dfrac{S_x}{S_y} \bar{x} = -36{,}68$ для уравнения прямой. Записать уравнение связи уравнение связи $y = -0{,}7869 \cdot x + 18{,}8675$ между электрическим сопротивлением Y (ρ, ом) и мощностью горизонта X (m, %) в 12 скважинах с коэффициентом корреляции $r_{xy} = 0{,}61$.

3) Вычислить наблюдаемую величину критерия

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n-2} = \frac{0{,}61}{\sqrt{1 - 0{,}61^2}} \sqrt{12-2} = 2{,}21.$$

Для степени свободы $k = n - 2 = 10$ и для уровня значимости $q = 0{,}05$ по таблице распределения Стьюдента определить критическую точку $t_{q,k} = 2{,}23$.

4) Сопоставить теоретическое и эмпирическое значения критерия согласия по Стьюденту. Так как $t = 2{,}21 < 2{,}23$, то гипотеза об отсутствии связи H_0: $\rho = 0$ принимается, т.е. X и Y имеют не устойчивую связь с коэффициентом корреляции $r_{xy} = 0{,}61$.

Ответ: связь между электрическим сопротивлением и мощностью пласта неустойчивая при уровне значимости $q = 0{,}05$.

Упражнение 6. Для горизонта $Д_1$ Туймазинского месторождения установлено, что вариации ***V*** мощности меняются с изменением площади ***S*** расположения скважин фиксированного количества. При изучении харак-

тера этой зависимости коэффициенты вариации мощности вычислялись по данным 50 скважин для площадей, равных в условных единицах 1,2, 3, 4, 5, 6, 7, 8, 9, 10. Результаты исследований приведены в *Таблице 31*. Вычислить коэффициент корреляции и параметры уравнения $V = a \cdot S + b$.

Таблица 31.

S	1	2	3	4	5	6	7	8	9	10
V	46,3	59,5	45,5	76,9	51,7	82,7	78,9	72,0	87,7	98,0

Проверка гипотез о равенстве математических ожиданий и дисперсиях случайных величин.

Часто при геологических исследованиях требуется выяснить по характеристикам двух выборок равенство или близость неизвестных параметров случайных величин. Такая задача возникает при совместном изучении различных пластов по фиксированному признаку, при сравнении лабораторных методов исследования пород. Выясним сказанное на примере по проверке гипотез о равенстве математических ожиданий и дисперсий двух случайных величин.

Задача 12: С целью оценки влияния метода определения для первого пласта Верхнебашкирского горизонта Бахметьевского месторождения пористость определялась как по керну (*X*) так и по каротажу (*Y*). Результаты приведены в *Таблице 32*. Выяснить является ли существенным влияние метода определения на величину получаемых результатов по пористости.

Таблица 32.

X(%)	22,3	23	11,5	26	27							
Y(%)	14	14	18	18	23	21	30	14	14,5	22	27	28

Решение: Чтобы выяснить является ли существенным влияние метода определения на величину получаемых результатов, необходимо проверить гипотезы $H_0: M(X) = M(Y)$ и $H_0: \sigma_x^2 = \sigma_y^2$. Для этого:

Таблица 33.

n_x	x_i (%)	n_y	y_i (%)	x_i^2	y_i^2
1	22,3	1	14	497,29	196
2	23	2	14	529	196
3	11,5	3	18	132,25	324
4	26	4	18	676	324
5	27	5	23	729	529
		6	21		441
		7	30		900
		8	14		196
		9	14,5		210,25
		10	22		484
		11	28		784
		12	27		729
$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = 21{,}96$		$\bar{y} = \frac{\sum_{i=1}^{12} y_i}{12} = 20{,}29$		$\overline{x^2} = \frac{\sum_{i=1}^{5} x_i^2}{5} = 512{,}71$	$\overline{y^2} = \frac{\sum_{i=1}^{12} y_i^2}{12} = 442{,}77$

1) Определяем объемы $n_x = 5$ и $n_y = 12$ выборки, соответственно для *X* и *Y*. Затем, вычисляем выборочные средние, промежуточные результаты, которых размещены в *Таблице 33*.

2) Вычисляем дисперсии

$$S_x^2 = \overline{x^2} - \bar{x}^2 = 512{,}71 - 21{,}96^2 = 30{,}466;$$
$$S_y^2 = \overline{y^2} - \bar{y}^2 = 442{,}77 - 20{,}29^2 = 31{,}09;$$

3) Вычисляем наблюдаемую величину критерия согласия распределения Стьюдента:

$$t = (\bar{x} - \bar{y}) \cdot \sqrt{\frac{n_x + n_y - 2}{n_x S_x^2 + n_y S_y^2}} \cdot \sqrt{\frac{n_x n_y}{n_x + n_y}} =$$
$$= (21{,}96 - 20{,}29) \cdot \sqrt{\frac{5 + 12 - 2}{5 \cdot 30{,}466 + 12 \cdot 31{,}09}} \cdot \sqrt{\frac{5 \cdot 12}{5 + 12}} = 0{,}53;$$

Критерий Фишера: $F = \frac{S_1^2}{S_2^2}$, где $S_1^2 > S_2^2$, т.е. $S_1^2 = S_y^2$, $S_2^2 = S_x^2$ и

$$F = \frac{31{,}09}{30{,}466} = 1{,}02.$$

Если X и Y подчиняются нормальному закону, то при условии $M(X)=M(Y)$ и $\sigma_x^2=\sigma_y^2$ случайная величина t подчиняется распределению Стьюдента со степенью свободы $k=n_x+n_y-2$, а случайная величина $\frac{n_x}{n_y}\cdot F$ при условии $\sigma_x^2=\sigma_y^2$ подчиняется распределению Фишера со степенями свободы $k_1=n_y-1=12-1=11$и $k_2=n_x-1=5-1=4$. Сначала по критерию Фишера проверяем гипотезу $H_0:\ \sigma_x^2=\sigma_y^2$. Выбираем уровень значимости $q=0{,}05$. Потом, по таблице находим $F_{q,k_1,k_2}=5{,}91$. Сравниваем F с F_{q,k_1,k_2}: 1,02<5,91. Таким образом, гипотеза H_0 принимается, т.е. дисперсии вносимыми методами существенно не отличаются. Если $F > F_{q,k_1,k_2}$, то H_0 должна быть отвергнута.

Если гипотеза о равенстве $\sigma_x^2=\sigma_y^2$ принята, как в данном случае, то проверяется гипотеза $H_0: M(X)=M(Y)$. Для этого по таблице находят $t_{q,k}$ и сравнивают с t. Если $t<t_{q,k}$,то гипотеза о равенстве математических ожиданий принимается, если же $t>t_{q,k}$,то гипотеза отвергается. В нашем случае $k=n_x+n_y-2=5+12-2=15$, уровень значимости $q=0{,}05$, значит $t_{q,k}$=2,13 и $t=0{,}53<t_{q,k}=2{,}13$. Таким образом, гипотеза $H_0: M(X)=M(Y)$ принимается, т.е. средние значения, определяемые двумя методами, существенно не отличаются.

ВЫВОД: методы определения пористости по керну и каротажу дают одинаковые результаты.

Упражнение 7. Для подтверждения выводов, полученных по результатам *Таблицы 32*, было выполнено определение пористости по керну (*X*) так и по каротажу (*Y*) для пятого пласта Бахметьевского месторождения (*Таблица 34*). Проверить гипотезы $H_0: M(X)=M(Y)$ и $H_0:\ \sigma_x^2=\sigma_y^2$.

Таблица 34

X(%)	25	23,5	25			
Y(%)	7	2,5	20	11,5	12	3,2

Корреляционная таблица.

Для наглядности представления характера связи между случайными величинами *X* и *Y* , а так же с целью изучения закона распределения одной случайной величины в зависимости от значения другой, составляют корреляционную таблицу. Корреляционной таблицей называют совокупность частот совместного наблюдения пар значений (*x*,*y*) случайных величин *X* и *Y*. При составлении такой таблицы для каждого интервала статистического распределения одной случайной величины строят статистическое распределение другой случайной величины. Рассмотрим таблицу распределения проницаемости (*X*) и амплитуды спонтанной поляризации (*Y*)

Таблица 35

Y/X	**0 - 100**	**100 - 200**	**200 - 300**	**300 -400**	**400 -500**	**500 - 600**	**600 - 700**	**700 - 800**	**800 - 900**	**900 - 1000**	**1000-1100**	m_{yj}	y_j	y_j^2
25 - 35	1			1								1	30	900
35 - 45	1	2	2	1								6	40	1600
45 - 55		4	7	10	4	1	1	2				29	50	2500
55 - 65		2	4	4	12	6	8	1				37	60	3600
65 - 75				3	7	4	6	2	2		2	26	70	4900
75 - 85								1	1	2	0	4	80	6400
85 - 95								2		3	0	5	90	8100
x_i	50	150	250	350	450	550	650	750	850	950	1050	**109**		
m_{xi}	2	8	13	19	23	11	15	8	3	5	2	**109**		
y_i	70	400	670	1020	1410	690	950	560	220	430	140			

В случае линейной связи $y = a \cdot x + b$ частоты в таблице располагаются внутри полосы, направленной вдоль диагонали матрицы. Если коэффициент линейной корреляции $r_{xy} > 0$(корреляция положительная), то частоты

располагаются вдоль главной диагонали матрицы. Если $r_{xy} < 0$, то частоты располагаются вдоль второй диагонали матрицы. В случае нелинейной корреляции частоты располагаются внутри кривой полосы. Вычисление коэффициента линейной корреляции r_{xy} производится по следующей схеме:

1) Для каждой случайной величины *X* и *Y* определяем число интервалов $k = 11$ и $l = 11$ ($i = 1,\ldots,k$ и $j = 1,\ldots,l$ соответственно).

2) Середина интервала принимается за значения (x_i, y_j) $i = 1,\ldots,k$; $j = 1,\ldots,l$.

3) В клетке пересечения *i*-столбца и *j*-строки записываем частоту и обозначаем ее через m_{ij}.

4) Для нахождения статистического распределения *X* по таблице распределения вычисляют сумму $m_{xi} = \sum_{j=1}^{l} m_{ij}$ и записывают в *i*-ом столбце. Также строят распределение *Y*. Сумма частот m_{ij} по каждой j-ой строке является частотой значения y_j случайной величины *Y*. Объем выборки равен

$$n = \sum_{i=1}^{k} n_{xi} = \sum_{j=1}^{l} n_{yj} = 109.$$

5) Определяем выборочные средние

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot m_{xi}}{n} = 483{,}9, \quad \bar{y} = \frac{\sum_{j=1}^{l} y_j \cdot m_{yj}}{n} = 60{,}2,$$

$$\overline{x^2} = \frac{\sum_{i=1}^{k} x_i^2 \cdot m_{xi}}{n} = 284701{,}8, \quad \overline{y^2} = \frac{\sum_{j=1}^{l} y_j^2 \cdot m_{yj}}{n} = 3766{,}9,$$

$$\overline{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij} \cdot x_i \cdot y_j}{n} = 31027{,}5.$$

6) Вычисляем дисперсии $S_x^2 = \overline{x^2} - \bar{x}^2 = 5049{,}2;\quad S_y^2 = \overline{y^2} - \bar{y}^2 = 144{,}9$ и среднеквадратические отклонения

$$S_x = \sqrt{5049{,}2} = 224{,}7;\quad S_y = \sqrt{144{,}9} = 12{,}0.$$

7) Коэффициент линейной корреляции $r_{xy} = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x S_y} = 0{,}703.$

Таким образом, величины *X* и *Y* коррелированны.

Энтропия газа как логарифма вероятности наивероятнейшего распределения молекул.

Пусть в некотором объеме *V* заключен совершенный газ имеющий абсолютную температуру *T*. Если этому газу сообщить обратимым путем некоторое количество энергии в виде тепла dQ, тот величина

$$dS = \frac{dQ}{T} \qquad (1)$$

будет являться дифференциалом энтропии *S*. Для интегрирования уравнения (1) выразим dQ через *T*. Пусть *E* есть энергия, которой обладает газ в объеме *V*, а PdV - затрачиваемая внешняя работа, связанная со сжатием газа, причем *P* – давление. Тогда $dQ = dE + PdV$. Согласно кинетической теории *E* пропорционально числу молекул газа *N* и температуре *T*, а именно:

$$E = \frac{3}{2} NKT.$$

Давление *P* пропорционально *T* и числу молекул, приходящихся на единицу объема, а именно:

$$P = \frac{N}{V} KT.$$

Таким образом,

$$dQ = \frac{3}{2} NK\ dT + NKT \frac{dV}{V},$$

тогда

$$S = \int \frac{dQ}{T} = NK\left(\frac{3}{2}\ln T + \ln V\right) + const \quad . \qquad (2)$$

Будем считать, что каждая молекула независимо от других может оказаться в любой части объема V с пропорциональной объему этой части вероятностью.

Разобьем весь объем V на части $V_1, V_2, ..., V_S$ и найдем вероятность того, что V_1 будет содержать n_1 молекул, V_2 - n_2 молекул,…, V_S - n_S молекул. По обобщенному биноминальному распределению находим:

$$P_{n_1 \ldots n_S} = \frac{N!}{n_1! .. n_S!} P_1^{n_1} \ldots P_S^{n_S} ,$$

где $P_\nu = \frac{V_\nu}{V}$, $n_1 + n_2 + \ldots + n_S = N$.

$$\ln P_{n_1 \ldots n_S} = \ln N! - \sum \ln n_\nu ! + \sum n_\nu \log P_\nu \quad . \qquad (3)$$

Найдем максимальное значение функции $\ln P_{n_1 \ldots n_S}$. Для этого составим функцию Лагранжа:

$$F = \ln N! - \sum \ln n_\nu ! + \sum n_\nu \log P_\nu + \lambda \sum n_\nu \quad .$$

Вычислим частные производные и приравняем их к нулю:

$$\frac{\partial F}{\partial n_\nu} = -\frac{\partial \ln n_\nu !}{\partial n_\nu} + \ln P_\nu + \lambda = 0 \quad .$$

Воспользуемся формулой Стирлинга:

$$\ln m! \sim \int_1^m \ln x dx = m(\ln m - 1) \quad \text{и} \quad \frac{\partial \ln m!}{\partial m} \approx \ln m \quad .$$

Тогда уравнение (3) приобретает вид:

$$-\ln n_\nu + \ln P_\nu + \lambda = 0 ,$$

откуда

$$\ln \frac{n_\nu}{P_\nu} = \lambda = const,$$

т.е. n_ν пропорциональны P_ν, но $n_\nu = NP_\nu = N\frac{V_\nu}{V}$ (P_ν - частота, N –объем выборки).

Таким образом, наивероятнейшее распределение молекул газа такое, при котором частичные объемы содержат пропорциональные им числа молекул.

ПРИЛОЖЕНИЯ

I. Таблицы

Таблица значений плотности стандартного нормального распределения

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

x	0	1	2	3	4	5	6	7	8	9
0,0	0,398942	0,398922	0,398862	0,398763	0,398623	0,398444	0,398225	0,397966	0,397668	0,397330
0,1	0,396953	0,396536	0,396080	0,395585	0,395052	0,394479	0,393868	0,393219	0,392531	0,391806
0,2	0,391043	0,390242	0,389404	0,388529	0,387617	0,386668	0,385683	0,384663	0,383606	0,382515
0,3	0,381388	0,380226	0,379031	0,377801	0,376537	0,375240	0,373911	0,372548	0,371154	0,369728
0,4	0,36827	0,366782	0,365263	0,363714	0,362135	0,360527	0,358890	0,357225	0,355533	0,353812
0,5	0,352065	0,350292	0,348493	0,346668	0,344818	0,342944	0,341046	0,339124	0,337180	0,335213
0,6	0,333225	0,331215	0,329184	0,327133	0,325062	0,322972	0,320864	0,318737	0,316593	0,314432
0,7	0,312254	0,310060	0,307851	0,305627	0,303389	0,301137	0,298872	0,296595	0,294305	0,292004
0,8	0,289692	0,287369	0,285036	0,282694	0,280344	0,277985	0,275618	0,273244	0,270864	0,268477
0,9	0,266085	0,263688	0,261286	0,258881	0,256471	0,254059	0,251644	0,249228	0,246809	0,24439
1,0	0,241971	0,239551	0,237132	0,234714	0,232297	0,229882	0,227470	0,22506	0,222653	0,220251
1,1	0,217852	0,215458	0,213069	0,210686	0,208308	0,205936	0,203571	0,201214	0,198863	0,196520
1,2	0,194186	0,19186	0,189543	0,187235	0,184937	0,182649	0,180371	0,178104	0,175847	0,173602
1,3	0,171369	0,169147	0,166937	0,164740	0,162555	0,160383	0,158225	0,15608	0,153948	0,151831
1,4	0,149727	0,147639	0,145564	0,143505	0,14146	0,139431	0,137417	0,135418	0,133435	0,131468
1,5	0,129518	0,127583	0,125665	0,123763	0,121878	0,120009	0,118157	0,116323	0,114505	0,112704
1,6	0,110921	0,109155	0,107406	0,105675	0,103961	0,102265	0,100586	0,098925	0,097282	0,095657
1,7	0,094049	0,092459	0,090887	0,089333	0,087796	0,086277	0,084776	0,083293	0,081828	0,08038
1,8	0,07895	0,077538	0,076143	0,074766	0,073407	0,072065	0,070740	0,069433	0,068144	0,066871
1,9	0,065616	0,064378	0,063157	0,061952	0,060765	0,059595	0,058441	0,057304	0,056183	0,055079
2,0	0,053991	0,052919	0,051864	0,050824	0,04980	0,048792	0,047800	0,046823	0,045861	0,044915
2,1	0,043984	0,043067	0,042166	0,041280	0,040408	0,039550	0,038707	0,037878	0,037063	0,036262
2,2	0,035475	0,034701	0,033941	0,033194	0,03246	0,031740	0,031032	0,030337	0,029655	0,028985
2,3	0,028327	0,027682	0,027048	0,026426	0,025817	0,025218	0,024631	0,024056	0,023491	0,022937
2,4	0,022395	0,021862	0,021341	0,020829	0,020328	0,019837	0,019356	0,018885	0,018423	0,017971
2,5	0,017528	0,017095	0,016670	0,016254	0,015848	0,015449	0,015060	0,014678	0,014305	0,01394
2,6	0,013583	0,013234	0,012892	0,012558	0,012232	0,011912	0,011600	0,011295	0,010997	0,010706
2,7	0,010421	0,010143	0,009871	0,009606	0,009347	0,009094	0,008846	0,008605	0,00837	0,00814
2,8	0,007915	0,007697	0,007483	0,007274	0,007071	0,006873	0,006679	0,006491	0,006307	0,006127
2,9	0,005953	0,005782	0,005616	0,005454	0,005296	0,005143	0,004993	0,004847	0,004705	0,004567
3,0	0,004432	0,004301	0,004173	0,004049	0,003928	0,003810	0,003695	0,003584	0,003475	0,00337
3,1	0,003267	0,003167	0,00307	0,002975	0,002884	0,002794	0,002707	0,002623	0,002541	0,002461

x	0	1	2	3	4	5	6	7	8	9
3,2	0,002384	0,002309	0,002236	0,002165	0,002096	0,002029	0,001964	0,001901	0,001840	0,001780
3,3	0,001723	0,001667	0,001612	0,001560	0,001508	0,001459	0,001411	0,001364	0,001319	0,001275
3,4	0,001232	0,001191	0,001151	0,001112	0,001075	0,001038	0,001003	0,000969	0,000936	0,000904
3,5	0,000873	0,000843	0,000814	0,000785	0,000758	0,000732	0,000706	0,000681	0,000657	0,000634
3,6	0,000612	0,00059	0,000569	0,000549	0,000529	0,000510	0,000492	0,000474	0,000457	0,000441
3,7	0,000425	0,000409	0,000394	0,000380	0,000366	0,000353	0,000340	0,000327	0,000315	0,000303
3,8	0,000292	0,000281	0,000271	0,000260	0,000251	0,000241	0,000232	0,000223	0,000215	0,000207
3,9	0,000199	0,000191	0,000184	0,000177	0,000170	0,000163	0,000157	0,000151	0,000145	0,000139
4,0	0,000134	0,000129	0,000124	0,000119	0,000114	0,000109	0,000105	0,000101	0,000097	0,000093

Таблица значений функции Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	X	$\Phi(x)$	x	$\Phi(x)$
0,00	0,00000	0,50	0,19146	1,00	0,34134	1,50	0,43319	2,00	0,47725	3,00	0,49865
0,01	0,00399	0,51	0,19497	1,01	0,34375	1,51	0,43448	2,02	0,47831	3,05	0,49886
0,02	0,00798	0,52	0,19847	1,02	0,34614	1,52	0,43574	2,04	0,47932	3,10	0,49903
0,03	0,01197	0,53	0,20194	1,03	0,34849	1,53	0,43699	2,06	0,48030	3,15	0,49918
0,04	0,01595	0,54	0,20540	1,04	0,35083	1,54	0,43822	2,08	0,48124	3,20	0,49931
0,05	0,01994	0,55	0,20884	1,05	0,35314	1,55	0,43943	2,10	0,48214	3,25	0,49942
0,06	0,02392	0,56	0,21226	1,06	0,35543	1,56	0,44062	2,12	0,48300	3,30	0,49952
0,07	0,02790	0,57	0,21566	1,07	0,35769	1,57	0,44179	2,14	0,48382	3,35	0,49960
0,08	0,03188	0,58	0,21904	1,08	0,35993	1,58	0,44295	2,16	0,48461	3,40	0,49966
0,09	0,03586	0,59	0,22240	1,09	0,36214	1,59	0,44408	2,18	0,48537	3,45	0,49972
0,10	0,03983	0,60	0,22575	1,10	0,36433	1,60	0,44520	2,20	0,48610	3,50	0,49977
0,11	0,04380	0,61	0,22907	1,11	0,36650	1,61	0,44630	2,22	0,48679	3,55	0,49981
0,12	0,04776	0,62	0,23237	1,12	0,36864	1,62	0,44738	2,24	0,48745	3,60	0,49984
0,13	0,05172	0,63	0,23565	1,13	0,37076	1,63	0,44845	2,26	0,48809	3,65	0,49987
0,14	0,05567	0,64	0,23891	1,14	0,37286	1,64	0,44950	2,28	0,48870	3,70	0,49989
0,15	0,05962	0,65	0,24215	1,15	0,37493	1,65	0,45053	2,30	0,48928	3,75	0,49991
0,16	0,06356	0,66	0,24537	1,16	0,37698	1,66	0,45154	2,32	0,48983	3,80	0,49993
0,17	0,06749	0,67	0,24857	1,17	0,37900	1,67	0,45254	2,34	0,49036	3,85	0,49994
0,18	0,07142	0,68	0,25175	1,18	0,38100	1,68	0,45352	2,36	0,49086	3,90	0,49995
0,19	0,07535	0,69	0,25490	1,19	0,38298	1,69	0,45449	2,38	0,49134	3,95	0,49996
0,20	0,07926	0,70	0,25804	1,20	0,38493	1,70	0,45543	2,40	0,49180	4,00	0,49997
0,21	0,08317	0,71	0,26115	1,21	0,38686	1,71	0,45637	2,42	0,49224	4,05	0,49997

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,22	0,08706	0,72	0,26424	1,22	0,38877	1,72	0,45728	2,44	0,49266	4,10	0,49998
0,23	0,09095	0,73	0,26730	1,23	0,39065	1,73	0,45818	2,46	0,49305	4,15	0,49998
0,24	0,09483	0,74	0,27035	1,24	0,39251	1,74	0,45907	2,48	0,49343	4,20	0,49999
0,25	0,09871	0,75	0,27337	1,25	0,39435	1,75	0,45994	2,50	0,49379	4,25	0,49999
0,26	0,10257	0,76	0,27637	1,26	0,39617	1,76	0,46080	2,52	0,49413	4,30	0,49999
0,27	0,10642	0,77	0,27935	1,27	0,39796	1,77	0,46164	2,54	0,49446	4,35	0,49999
0,28	0,11026	0,78	0,28230	1,28	0,39973	1,78	0,46246	2,56	0,49477	4,40	0,49999
0,29	0,11409	0,79	0,28524	1,29	0,40147	1,79	0,46327	2,58	0,49506	4,45	0,50000
0,30	0,11791	0,80	0,28814	1,30	0,40320	1,80	0,46407	2,60	0,49534	4,50	0,50000
0,31	0,12172	0,81	0,29103	1,31	0,40490	1,81	0,46485	2,62	0,49560	4,55	0,50000
0,32	0,12552	0,82	0,29389	1,32	0,40658	1,82	0,46562	2,64	0,49585	4,60	0,50000
0,33	0,12930	0,83	0,29673	1,33	0,40824	1,83	0,46638	2,66	0,49609	4,65	0,50000
0,34	0,13307	0,84	0,29955	1,34	0,40988	1,84	0,46712	2,68	0,49632	4,70	0,50000
0,35	0,13683	0,85	0,30234	1,35	0,41149	1,85	0,46784	2,70	0,49653	4,75	0,50000
0,36	0,14058	0,86	0,30511	1,36	0,41309	1,86	0,46856	2,72	0,49674	4,80	0,50000
0,37	0,14431	0,87	0,30785	1,37	0,41466	1,87	0,46926	2,74	0,49693	4,85	0,50000
0,38	0,14803	0,88	0,31057	1,38	0,41621	1,88	0,46995	2,76	0,49711	4,90	0,50000
0,39	0,15173	0,89	0,31327	1,39	0,41774	1,89	0,47062	2,78	0,49728	4,95	0,50000
0,40	0,15542	0,90	0,31594	1,40	0,41924	1,90	0,47128	2,80	0,49744	5,00	0,50000
0,41	0,15910	0,91	0,31859	1,41	0,42073	1,91	0,47193	2,82	0,49760		
0,42	0,16276	0,92	0,32121	1,42	0,42220	1,92	0,47257	2,84	0,49774		
0,43	0,16640	0,93	0,32381	1,43	0,42364	1,93	0,47320	2,86	0,49788		
0,44	0,17003	0,94	0,32639	1,44	0,42507	1,94	0,47381	2,88	0,49801		
0,45	0,17364	0,95	0,32894	1,45	0,42647	1,95	0,47441	2,90	0,49813		
0,46	0,17724	0,96	0,33147	1,46	0,42785	1,96	0,47500	2,92	0,49825		
0,47	0,18082	0,97	0,33398	1,47	0,42922	1,97	0,47558	2,94	0,49836		
0,48	0,18439	0,98	0,33646	1,48	0,43056	1,98	0,47615	2,96	0,49846		
0,49	0,18793	0,99	0,33891	1,49	0,43189	1,99	0,47670	2,98	0,49856		

Критические точки распределения Стьюдента.

k \ α	**0,1**	**0,05**	**0,02**	**0,01**	**0,001**
1	6,3138	12,7062	31,8205	63,6567	636,6192
2	2,9200	4,3027	6,9646	9,9248	31,5991
3	2,3534	3,1824	4,5407	5,8409	12,924
4	2,1318	2,7764	3,7469	4,6041	8,6103
5	2,0150	2,5706	3,3649	4,0321	6,8688
6	1,9432	2,4469	3,1427	3,7074	5,9588
7	1,8946	2,3646	2,9980	3,4995	5,4079
8	1,8595	2,3060	2,8965	3,3554	5,0413
9	1,8331	2,2622	2,8214	3,2498	4,7809
10	1,8125	2,2281	2,7638	3,1693	4,5869
11	1,7959	2,2010	2,7181	3,1058	4,4370
12	1,7823	2,1788	2,6810	3,0545	4,3178
13	1,7709	2,1604	2,6503	3,0123	4,2208
14	1,7613	2,1448	2,6245	2,9768	4,1405
15	1,7531	2,1314	2,6025	2,9467	4,0728
16	1,7459	2,1199	2,5835	2,9208	4,0150
17	1,7396	2,1098	2,5669	2,8982	3,9651
18	1,7341	2,1009	2,5524	2,8784	3,9216
19	1,7291	2,0930	2,5395	2,8609	3,8834
20	1,7247	2,0860	2,5280	2,8453	3,8495
21	1,7207	2,0796	2,5176	2,8314	3,8193
22	1,7171	2,0739	2,5083	2,8188	3,7921
23	1,7139	2,0687	2,4999	2,8073	3,7676
24	1,7109	2,0639	2,4922	2,7969	3,7454
25	1,7081	2,0595	2,4851	2,7874	3,7251
26	1,7056	2,0555	2,4786	2,7787	3,7066
27	1,7033	2,0518	2,4727	2,7707	3,6896
28	1,7011	2,0484	2,4671	2,7633	3,6739
29	1,6991	2,0452	2,4620	2,7564	3,6594
30	1,6973	2,0423	2,4573	2,7500	3,6460
35	1,6896	2,0301	2,4377	2,7238	3,5911
40	1,6839	2,0211	2,4233	2,7045	3,5510
45	1,6794	2,0141	2,4121	2,6896	3,5203

k \ α	0,1	0,05	0,02	0,01	0,001
50	1,6759	2,0086	2,4033	2,6778	3,4960
55	1,6730	2,004	2,3961	2,6682	3,4764
60	1,6706	2,0003	2,3901	2,6603	3,4602
70	1,6669	1,9944	2,3808	2,6479	3,4350
80	1,6641	1,9901	2,3739	2,6387	3,4163
90	1,6620	1,9867	2,3685	2,6316	3,4019
100	1,6602	1,9840	2,3642	2,6259	3,3905
110	1,6588	1,9818	2,3607	2,6213	3,3812
120	1,6577	1,9799	2,3578	2,6174	3,3735
∞	1,6448	1,9600	2,3263	2,5758	3,2905

Таблица критических точек распределения Пирсона.

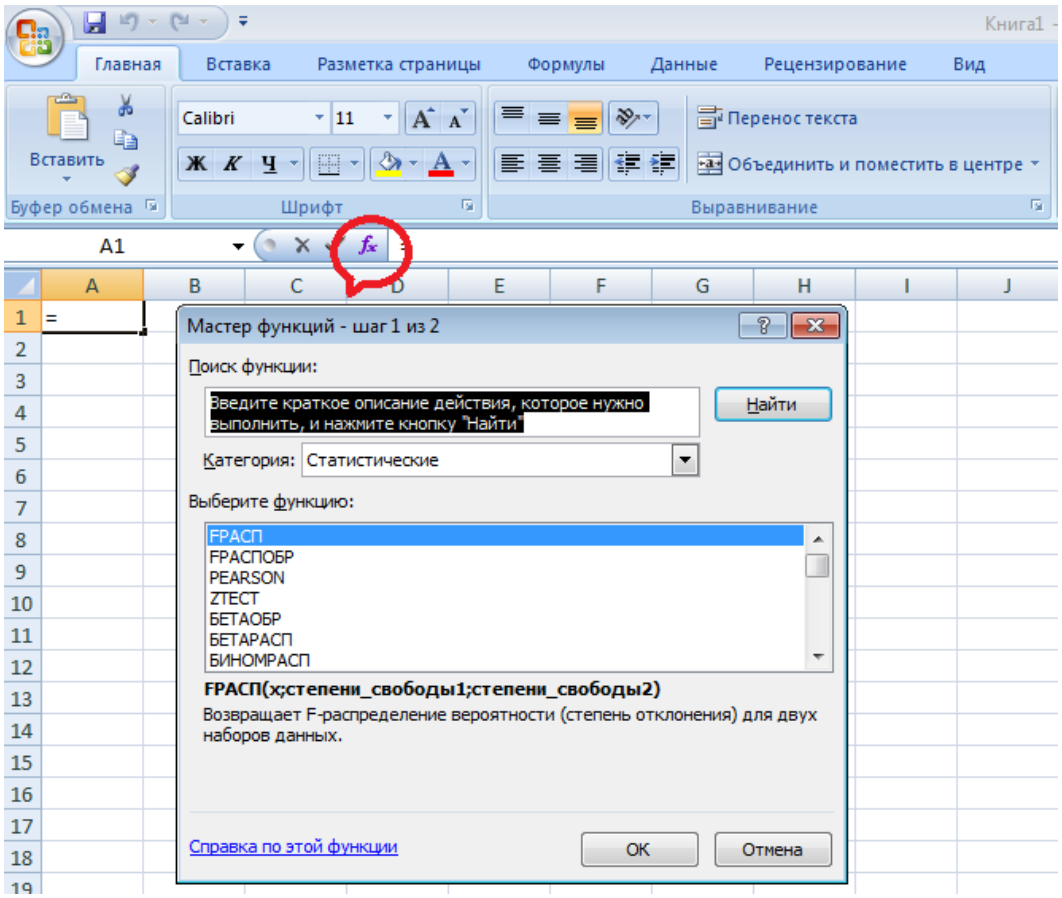
k /α	0,01	0,025	0,05	0,95	0,975	0,99
1	6,63490	5,02389	3,84146	0,00393	0,00098	0,00016
2	9,21034	7,37776	5,99146	0,10259	0,05064	0,02010
3	11,34487	9,34840	7,81473	0,35185	0,21580	0,11483
4	13,2767	11,14329	9,48773	0,71072	0,48442	0,29711
5	15,08627	12,8325	11,0705	1,14548	0,83121	0,55430
6	16,81189	14,44938	12,59159	1,63538	1,23734	0,87209
7	18,47531	16,01276	14,06714	2,16735	1,68987	1,23904
8	20,09024	17,53455	15,50731	2,73264	2,17973	1,64650
9	21,66599	19,02277	16,91898	3,32511	2,70039	2,08790
10	23,20925	20,48318	18,30704	3,94030	3,24697	2,55821
11	24,72497	21,92005	19,67514	4,57481	3,81575	3,05348
12	26,21697	23,33666	21,02607	5,22603	4,40379	3,57057
13	27,68825	24,7356	22,36203	5,89186	5,00875	4,10692
14	29,14124	26,11895	23,68479	6,57063	5,62873	4,66043
15	30,57791	27,48839	24,99579	7,26094	6,26214	5,22935
16	31,99993	28,84535	26,29623	7,96165	6,90766	5,81221
17	33,40866	30,19101	27,58711	8,67176	7,56419	6,40776
18	34,80531	31,52638	28,86930	9,39046	8,23075	7,01491
19	36,19087	32,85233	30,14353	10,11701	8,90652	7,63273

k /α	0,01	0,025	0,05	0,95	0,975	0,99
20	37,56623	34,16961	31,41043	10,85081	9,59078	8,26040
21	38,93217	35,47888	32,67057	11,59131	10,2829	8,89720
22	40,28936	36,78071	33,92444	12,33801	10,98232	9,54249
23	41,63840	38,07563	35,17246	13,09051	11,68855	10,19572
24	42,97982	39,36408	36,41503	13,84843	12,40115	10,85636
25	44,31410	40,64647	37,65248	14,61141	13,11972	11,52398
26	45,64168	41,92317	38,88514	15,37916	13,84391	12,19815
27	46,96294	43,19451	40,11327	16,15140	14,57338	12,87850
28	48,27824	44,46079	41,33714	16,92788	15,30786	13,56471
29	49,58788	45,72229	42,55697	17,70837	16,04707	14,25645
30	50,89218	46,97924	43,77297	18,49266	16,79077	14,95346
31	52,19139	48,23189	44,98534	19,28057	17,53874	15,65546
32	53,48577	49,48044	46,19426	20,07191	18,29076	16,36222
33	54,77554	50,72508	47,39988	20,86653	19,04666	17,07351
34	56,06091	51,96600	48,60237	21,66428	19,80625	17,78915
35	57,34207	53,20335	49,80185	22,46502	20,56938	18,50893
36	58,61921	54,43729	50,99846	23,26861	21,33588	19,23268
37	59,89250	55,66797	52,19232	24,07494	22,10563	19,96023
38	61,16209	56,89552	53,38354	24,8839	22,87848	20,69144
39	62,42812	58,12006	54,57223	25,69539	23,65432	21,42616
40	63,69074	59,34171	55,75848	26,5093	24,43304	22,16426
41	64,95007	60,56057	56,94239	27,32555	25,21452	22,90561
42	66,20624	61,77676	58,12404	28,14405	25,99866	23,65009
43	67,45935	62,99036	59,30351	28,96472	26,78537	24,39760
44	68,70951	64,20146	60,48089	29,78748	27,57457	25,14803
45	69,95683	65,41016	61,65623	30,61226	28,36615	25,90127
46	71,20140	66,61653	62,82962	31,43900	29,16005	26,65724
47	72,44331	67,82065	64,00111	32,26762	29,95620	27,41585
48	73,68264	69,02259	65,17077	33,09808	30,75451	28,17701
49	74,91947	70,22241	66,33865	33,93031	31,55492	28,94065
50	76,15389	71,42020	67,50481	34,76425	32,35736	29,70668

Статистика в Excel

Microsoft Excel дает широкие возможности для анализа статистических данных.

Исходные данные для анализа могут быть заданы на рабочем листе в виде списка значений. Ячейки нумеруются одновременным заданием имени столбца и номером элемента в этом столбце. Список реализованных в EXCEL статистических команд можно получить, нажав значок f_x и выбрав там категорию **Статистические**. Выделив нужную статистическую функцию, можно получить по ней справку .



Приведем основные статистические функции:

СРЗНАЧ()	вычисляет среднее значение для последовательности чисел: суммируются числовые значения в интервале ячеек и результат делится на количество

	этих значений. Эта функция игнорирует пустые, логические и текстовые ячейки.
МЕДИАНА()	вычисляет медиану множества чисел. Медиана – это число, являющееся серединой множества: количества чисел, меньшие и большие медианы, равны. Если количество чисел или ячеек четное, то результатом будет среднее двух чисел в середине множества.
МОДА()	возвращает наиболее часто встречающееся значение во множестве чисел.
МАКС()	возвращает наибольшее значение среди заданных чисел.
МИН()	возвращает минимальное значение среди заданных чисел.
СУММПРОИЗВ()	возвращает сумму произведений соответствующих членов двух и более массивов-аргументов (но не более 30 аргументов). Встречающиеся в аргументах нечисловые значения интерпретируются нулями.
СУММКВ()	возвращает сумму квадратов аргументов.
ДИСП(), ***ДИСПР(),*** ***СТАНДОТКЛОН(),*** ***СТАНДОТКЛОНП()***	предназначены для вычисления дисперсии и стандартного отклонения чисел в интервале ячеек. Перед тем как вычислять дисперсию и стандартное отклонение набора данных, нужно определить, представляют ли эти данные генеральную совокупность или выборку из генеральной совокупности. В случае выборки из генеральной совокупности следует

	использовать функции *ДИСП()* и *СТАНДОТКЛОН()*, а в случае генеральной совокупности – функции *ДИСПР()* и *СТАНДОТЛОНП()*.
СУММСУММКВ()	вычисляет сумму сумм квадратов соответствующих элементов в массивах.
СУММКВРАЗН()	вычисляет сумму квадратов разности соответствующих элементов в массивах.
НОРМРАСП() ***и*** ***НОРМОБР()***	вычисляют значения для нормального распределения и для обратной ему функции. Например, команда НОРМОБР(0.2,3,0.5) дает значение, соответствующее значению вероятности 0.2 для нормального закона распределения со средним значением (математическим ожиданием), равным 3 и среднеквадратичным отклонением, равным 0.5.
ПУАССОН(), СТЬЮДРАСП(), БИНОМРАСП().	вычисляют значения для распределений Пуассона, Стьюдента и биноминального распределения соответственно.
ПИРСОН() или ***КОРРЕЛ()***	вычисляет коэффициент корреляции (здесь он фигурирует как коэффициент Пирсона).
КВПИРСОН()	дает квадрат коэффициента корреляции.
НАКЛОН()	дает коэффициент a для уравнения линейной регрессии $y=ax+b$.
ОТРЕЗОК()	дает коэффициент b для уравнения линейной регрессии $y=ax+b$.

ЛИНЕЙН()	позволяет не только находить линейную регрессию, но и вычислять различные дополнительные параметры для ее анализа, а также проводить и кратную регрессию.
СТЬЮДРАСПОБР()	Вычисляет критические точки распределения Стьюдента. Например, *СТЬЮДРАСПОБР(0,05;25)* дает значение критической точки распределения Стьюдента при уровне значимости q=0,05 и при числе степеней свободы k=25
ХИ2ОБР()	Вычисляет критические точки распределения Пирсона (хи-квадрат).

Это далеко не все встроенные статистические функции. Если же их оказывается недостаточно, следует обратиться к **Пакету анализа**.

Пакет анализа является дополнением и содержит набор функций и инструментов, расширяющих встроенные аналитические возможности Excel. Предварительно его необходимо настроить:

в Excel 2003 необходимо в меню выбрать команду **Сервис** -> **Надстройки** и поставить галочку напротив **Пакета анализа**. Теперь в меню Сервис появится команда **Анализ данных**.

в Excel 2007 необходимо щелкнуть по кнопке **Офис**, далее по кнопе **Параметры Excel**, выбрать **Надстройки**, в нижней части окна в поле **Управления** выбрать **Надстройки Excel**, щелкнуть по кнопке **Перейти,** поставить галочку напротив **Пакета анализа.** На вкладке **Данные** появится команда **Анализ данных.**

При выполнении команды **Анализ данных** вызывается диалоговое окно, в котором выбирается режим **Описательная статистика**. Для каж-

дого набора входных данных в выходном интервале строится таблица со следующей информацией: Среднее, Стандартная ошибка, Медиана, Мода, Стандартное отклонение, Дисперсия выборки, Эксцесс, Асимметричность, Интервал, Минимум, Максимум, Сумма, Счет, Наибольший (k), Наименьший (k) (для любого заданного k) и Уровень надежности (доверительный интервал). Статистической обработке подвергается один или несколько наборов данных, располагаемых в интервале, ссылка на который задается в поле *Входной интервал*. Переключатель *Группирование* дает возможность уточнить, как размещаются данные: по столбцам или по строкам. При установленном флажке *Итоговая статистика* создается подробная выходная таблица, установив соответствующие флажки, можно поместить в нее дополнительные данные.

Список рекомендуемой литературы:

1. Колмогоров А.Н. Основные понятия теории вероятностей. М.: Наука. -1974. -117с.
2. Гнеденко Б.В., Хинчин А.Я. Элементарное введение в теорию вероятностей. М.: Наука. -1970. -166с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа. -2001. -479 с.
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшая школа. -2000. -400 с.
5. Методические указания по курсу: Теория вероятностей. Ч.1 – Казань: Изд-во Казанского государственного университета, 2008. – 48с.
6. Боев Г.П. Теория вероятностей. М.: Государственное издательство теоретической литературы. – 1950. – 368с.

Наталья Владимировна Шевченко

Элементы теории вероятностей и математической статистики:

Краткий курс лекций

для подготовки бакалавров очной и заочной форм обучения. –

Димитровград: Технологический институт – филиал УлГАУ, 2021. - 102 с.