Основы математической статистики

Горшков А.В.

Курс «Основы статистики » для студентов факультета «Связи с общественностью» рассчитан на то, чтобы дать представление об основных задачах, методах и подходах статистики, ее основах. Предполагается, что студенты знают основы курса высшей математики (элементы математического анализа: теория предела, производная, интегралы, в том числе несобственные) в пределах курса высшей математики для студентов гуманитарных факультетов университетов.

Курс состоит из двух частей. Первая - элементы теории вероятностей. Вторая - основы математической статистики. Первая часть необходима для более глубокого и полного понимания основных задач и методов статистики.

Объем курса 36 часов лекций. Отчетность – зачет.

литература

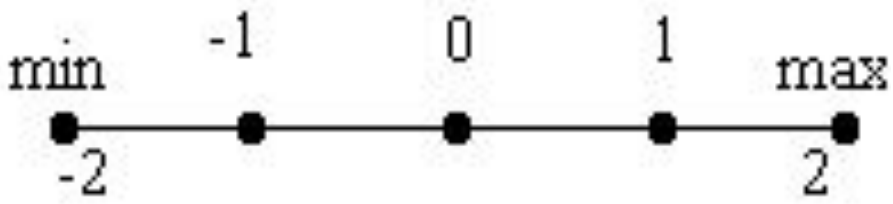
1. Шолохович Ф.А. Высшая математика в кратком изложении. Екатеринбург, УрГУ, 2003
2. Турецкий В.Я. Высшая математика. Екатеринбург, 1997.
3. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая Школа, 2001, 479 с.
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. Учебное пособие для студентов вузов. М.: "Высшая Школа", 1999.

Математическая статистика позволяет обрабатывать результаты опытов, измерений и т.д. Математическая статистика использует методы теории вероятности.

Теория вероятностей определяет законы случайности.

Шкалы измерений

- Номинальная. Позволяет различать предметы, например по цвету
- Дихотомическая.
- Ранговая. Позволяет упорядочить предметы
- Интервальная



- Относительная ноль соответствует полному отсутствию свойства (качества)

Случайные события

- Событие называется *детерминированным*, если в результате опыта оно происходит или не происходит наверняка. В детерминированном случае мы точно знаем, что данная причина приведет к единственному, вполне определенному следствию.
- Событие называется *случайным*, если в результате опыта мы не можем заранее предсказать - произойдет событие или нет. При этом предполагается, что опыт можно повторять неограниченное число раз при неизменных условиях.
- События, исход которых нельзя предсказать, но и невозможно повторять многократно, называются *неопределенными*.

- События A и B называются *несовместными*, если появление одного исключает появление другого.
- Событие B *следует* из события A, если событие B происходит всегда, когда произошло событие A .

  Это обозначается тем же символом, что и подмножество: $A \subset B$.
- Будем говорить о равенстве двух событий A и B, если из A следует B и из B следует A.
- Событие называется *невозможным*, если оно не может произойти никогда при данных условиях.
- Событие называется *достоверным*, если оно происходит всегда при данных условиях.

- Пусть случайный эксперимент проводится раз n, и событие A произошло m раз. Тогда говорят, что *относительная частота* события A есть $v(A)=m/n$.
- Частота события связана с его вероятностью. Относительную частоту называют еще *эмпирической вероятностью* потому, что по частоте события мы оцениваем возможность его появления в будущем.
- *Для любого случайного события A*

$$0 \le P_n(A) \le 1$$

n - количество случайных экспериментов.

Алгебра событий

- *Суммой* двух событий A и B называется событие $A+B$, состоящее в том, что произошло событие A или событие B. В данном случае "или" употребляется в не исключающем значении: A или B означает, что произошло событие A, или событие B или оба этих события одновременно.

- Сложение событий удовлетворяет коммутативному и ассоциативному законам:

$$A + B = B + A$$

$$A + (B + C) = (A + B) + C$$

- Коммутативность и ассоциативность позволяют складывать любое число событий в любом порядке. Свойство: *из события A следует сумма этого события с любым бытием B :*

$$A \subset A + B$$

Произведением двух событий A и B называется событие , состоящее в том, что события A и B произошли одновременно. Умножение событий так же, как и сложение, коммутативно и ассоциативно:

$$A(BC) = (AB)C$$

$$AB = BA$$

Свойство: *из события AB следуют событие A и событие B*

$$AB \subset A \qquad \text{и} \qquad AB \subset B$$

Сложение и умножение событий удовлетворяют двум дистрибутивным законам.

$$A(B + C) = AB + AC$$

$$(A + B)(A + C) = A(A + B + C) + BC = A + BC$$

Две теоремы о вероятности суммы событий и произведении

1. Если события несовместны, то вероятность суммы событий равна сумме вероятностей:

$$P(A+B) = P(A) + P(B)$$

2. Если события независимы, то вероятность произведения событий равна произведению вероятностей:

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

Обобщение этих теорем докажем позже.

Примеры.

1. Подбрасываем кубик. Всего исходов 6. Какова вероятность, что выпадет четное число?

Благоприятны исходы 2, 4, 6. Всего 3. 3/6

2. Какова вероятность, что при первом броске выпадет 3, во втором 4?

Здесь вероятность произведения событий: P(3)=1/6, P(4)=1/6

$$\frac{1}{6}\frac{1}{6}=\frac{1}{36}$$

3. Какова вероятность, что выпадет 3 или 5?

Здесь вероятность суммы несовместных событий. P=1/3

4. Стрелок стреляет по мишени 4 раза. Вероятность попадания в одном выстреле 0.8. Считаем, что у него хорошие нервы – каждый следующий выстрел не зависит от предыдущего. Какова вероятность, что он промахнется ровно 1 раз?

Решение. Могут произойти следующие события:
A1 промах в 1 выстреле, A2 промах во 2, A3 - в 3,или A4 - в 4.
Следовательно, событие, состоящее в одном промахе, можно представить как сумму событий A= A1 + A2 +A3 +A4. Но события очевидно, несовместны и P(A)= P(A1) + P(A2)+P(A3) +P(A4). – вероятность суммы событий равна сумме вероятностей.
Но каждое событие Ai состоит в том, что одновременно, в одной серии выстрелов, произошли 4 события, причем эти события независимы по условию: Ai= Ai1Ai2 Ai3 Ai4. Но вероятность произведения независимых событий равна произведению вероятностей.
Нужно найти вероятность промаха. В одном выстреле стрелок может либо попасть, либо промахнуться. Следовательно, эти события образуют полный набор и они несовместны. Но тогда вероятность промаха 1-0.8=0.2.
P(Ai)=P(Ai1)P(Ai2)P(Ai3)P(Ai4.)=0.8*0.8*0.8*0.2=0.1024. Всего 0.4096

Разностью событий *A* и *B* называется событие *A - B ,* состоящее в том, что произошло событие *A* и не произошло событие *B*.
Событие *B* называется *противоположным* событию *A* , если оно состоит в том, что не произошло событие *A*

Элементарные исходы

1. *не представимы в виде суммы двух других*
2. *попарно несовместны*
3. *никакие другие исходы в результате опыта произойти не могут*

События образуют полный набор, если они несовместны, а их сумма есть достоверное событие. Полный набор исходов называют также пространством элементарных исходов и обозначают обычно буквой Ω.

Комбинаторика

Число перестановок $n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$

Число расстановок $A_n^m = \frac{n!}{(n-m)!}$

Число сочетаний $C_n^m = \frac{n!}{m!(n-m)!}$

Например, число перестановок из 6 предметов

1x2x3x4x5x6=120.

Число расстановок из 6 предметов на 4 места

120/(6-4)!=120/2!=120/2=60.

Число сочетаний из 6 предметов по 4

120/(4!(6-4)!)=120/(4!2!)=5x6/2=15

Например, число перестановок из 6 предметов

1x2x3x4x5x6=720.

Число сочетаний из 6 предметов по 4

$$C_6^4 = \frac{6!}{4!(6-4)!} = \frac{6!}{4!\cdot 2!} = \frac{6\cdot 5\cdot 4!}{4!\cdot 2!} = \frac{6\cdot 5}{2} = 15$$

Классическое определение

- *Вероятностью P(A)* события называется отношение числа благоприятных исходов *m(A) к* общему числу несовместных равновозможных исходов:

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

Свойства вероятности.

- I. *Для любого случайного события A* $0 \leq P(A) \leq 1$
- 2. *Пусть события A и B несовместны. Тогда* $P(A+B)=P(A)+P(B)$

Например: бросание кубика. Всего исходов 6, число исходов, благоприятных выпадению четного числа – 3. P(A)=1/2

Пример. В корзине 15 шаров. Из низ 5 белых и 10 черных. Какова вероятность вытащить 3 белых шара?

Какова вероятность вытащить 3 черных шара?

Общее число исходов – число сочетаний из 15 по 3:

$$C_{15}^{3} = \frac{15!}{3!\cdot 12!} = \frac{13\cdot 14\cdot 15}{2\cdot 3} = 455$$

Число благоприятных исходов - число сочетаний из 10 по 3

$$C_{10}^{3} = \frac{10!}{3!\cdot 7!} = \frac{8\cdot 9\cdot 10}{2\cdot 3} = 120$$

Вероятность события 120/455=0,2637

Пример. Из колоды в 36 карт вытаскиваем 4. Какова вероятность, что среди них 2 короля?

Решаем по классической схеме.

1. Общее число равновозможных исходов определяем с помощью комбинаторики.

$$N = \frac{36!}{4!(36-32)!} = \frac{36!}{4!\cdot 32!}$$

Сокращаем 36! и 32!: по определению 36!=36*35*34*33*32! тогда

$$N = \frac{36\cdot 35\cdot 34\cdot 33}{1\cdot 2\cdot 3\cdot 4} = 58905$$

Считаем число благоприятных исходов: каждой паре королей нужно добавить пару «не королей»

$$M = \frac{4!}{2!2!}\frac{32!}{2!30!} = 6\cdot 31\cdot 16 = 2976 \qquad p = \frac{2976}{58905} = 0.0505$$

Вероятность суммы событий

Даны два события A и B. Подсчитаем вероятность суммы событий по классической схеме.

m(A) – число исходов, благоприятных только событию A.

m(B) – число исходов, благоприятных только событию B.

m(AB) – число исходов, благоприятных событию A и событию B.

Тогда сумме событий благоприятно m(A)+ m(B)+ m(AB) исходов и вероятность суммы событий равна

$$P(A+B)=\frac{m(A)+m(B)+m(AB)}{N} \qquad (1)$$

N – общее число исходов. С другой стороны

$$P(A)=\frac{m(A)+m(AB)}{N} \quad P(B)=\frac{m(B)+m(AB)}{N}$$

$$P(A)+P(B)=\frac{m(A)+m(B)+2m(AB)}{N} \qquad (2)$$

Вычтем из (2) выражение (1)

$$P(A)+P(B)-P(A+B)=\frac{m(AB)}{N}$$

Отсюда находим формулу вероятности суммы событий

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

Пример. Два стрелка независимо стреляют по мишени. Первый попадает с вероятностью 0.8, второй 0.7. Какова вероятность, что попадет хотя бы один?

Используем полученную формулу:

0.8+0.7-0.8*0.7=0.94

Условная вероятность

Пусть известно, что в результате эксперимента произошло событие *B*. Зная это, мы хотим подсчитать вероятность некоторого события *A*. Такую вероятность (при условии, что произошло событие *B*) называют *условной вероятностью* события *A* и обозначают P(A|B). Число исходов, благоприятных *B* обозначим через *m(B)*. Условная вероятность равна

$$P(A \mid B) = \frac{m(AB)}{m(B)}$$

Если разделить числитель и знаменатель на общее число исходов N, мы придем к формуле

$$P(A \mid B) = \frac{m(AB)/N}{m(B)/N} = \frac{P(AB)}{P(B)}$$

Свойства условной вероятности.

1. $P(A|A)=1$.
2. *Если* $B \subset A$ *, то* $P(A|B)=1$.
3. *Для любого события B с ненулевой вероятностью*
   $P(\Omega \bullet B)=1$, $P(0 \bullet B)=0$.
4. *Если события* B_1 *и* B_2 *несовместны, то*
   $$P(A|(B_1+B_2))=P(A|B_1)+P(A|B_2)$$
5. Теорема умножения
   $$P(AB)=P(A|B)P(B).$$

Определение независимости событий.

Говорят, что событие *A не зависит от* события *B,* если

$P(A|B)=P(A)$.

Формула полной вероятности

Теорема. *Пусть события $B_1, B_2, \ldots, B_n$ попарно несовместны и событие A содержится в их сумме:*

$$A \subset B_1+B_2+\ldots+B_n.$$

Умножим это соотношение на событие A. Тогда в левой части получим $A^2=A$, в правой

$$A(B_1+B_2+\ldots+B_n)$$

и

$$A=A(B_1+B_2+\ldots+B_n)$$

Вложение перешло в равенство. Вычислим вероятность от обеих частей

$$P(A)=P(AB_1)+P(AB_2)+\ldots+P(AB_n)$$

Вероятности произведений заменим $P(AB_1)=P(A|B_1)P(B_1)$

Тогда вероятность события A можно вычислить по следующей формуле:

$$P(A)=P(A|B_1)P(B_1)+ P(A|B_2)P(B_2)+\ldots+ P(A|B_n)P(B_n)$$

Пример. Имеется 3 группы корзин.

Корзин типа B_1 – 5

Корзин типа B_2 – 3

Корзин типа B_3 – 2

В каждой корзине типа B_1 10 белых 5 черных шаров.

В каждой корзине типа B_2 5 белых 10 черных шаров.

В каждой корзине типа B_3 10 белых 15 черных шаров.

Какова вероятность из выбранной наугад корзины выбрать белый шар? Решение.

Вероятность выбрать белый шар из B_1 $P(A|B_1)=10/15$

Вероятность выбрать белый шар из B_2 $P(A|B_2)=5/15$

Вероятность выбрать белый шар из B_3 $P(A|B_3)=10/25$

Вероятность выбрать корзину первой группы $P(B_1)=5/10$

Вероятность выбрать корзину второй группы $P(B_2)=3/10$

Вероятность выбрать корзину третьей группы $P(B_3)=2/10$

Подставляем числа в формулу

$$\frac{10}{15}\frac{5}{10}+\frac{5}{15}\frac{3}{10}+\frac{10}{25}\frac{2}{10}=0.513333$$

Формула Байеса

Теорема. *Пусть события* $B_1, B_2, \ldots, B_n$ *попарно несовместны и событие* A *содержится в их сумме:* $A \subset B_1+B_2+\ldots+B_n$. *Тогда при* $k=1,2,\ldots,n$ *справедлива формула*

$$P\left(B_k \middle| A\right) = \frac{P\left(B_k\right)P\left(A \middle| B_k\right)}{\sum_{i=1}^{n} P\left(B_i\right)P\left(A \middle| B_i\right)}$$

Пример

Те же самые условия. Случайно извлечен белый шар. Какова вероятность, что он извлечен из первой группы корзин?

Т.е. Мы фактически хотим найти вероятность $P(B_1|A)$.

Решение. $P(A)=0.513333$ (из предыдущей задачи)
$P(B_1)=5/10$, $P(A|B_1)=10/15$. После подстановки получим

$$\frac{\frac{5}{10}\frac{10}{15}}{0.513333} = 0.64928$$

Схема испытаний Бернулли

- Пусть в результате некоторого случайного испытания может произойти или не произойти определенное событие А. Если событие произошло, будем называть испытание успешным, а само событие – *успехом.* Испытание повторяется n раз. При этом соблюдаются следующие условия:
- вероятность успеха $P(A)=p$ в каждом испытании одна и та же;
- результат любого испытания не зависит от исходов предыдущих испытаний.

Такая последовательность испытаний с двумя исходами (успех/ неуспех) называется *последовательностью независимых испытаний Бернулли* или – *схемой испытаний Бернулли.*

Формула Бернулли вероятности k успехов в n независимых испытаниях. В серии из n испытаний должно *одновременно* произойти k успехов и n-k - «неуспехов». Вероятность успеха p, «неуспеха» q=1-p, так как для одного испытания события «успех/неуспех» образуют полную набор. Но тогда вероятность одной серии равна

$$p^k q^{n-k}$$

Сколько может быть различных серий? Очевидно, сколькими способами мы можем расставить k успехов на n мест. Это число расстановок. Но все успехи одинаковы. Следовательно, число серий равно числу сочетаний из n по k. Серии между собой, очевидно, несовместны, так как отличаются положением хотя бы одного успеха. Следовательно, вероятность суммы равна сумме вероятностей:

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

Пример. Подбрасываем 10 раз кубик. Какова вероятность, что пятерка выпадет ровно 4 раза?

Решение. Схема испытаний Бернулли. p=1/6, q=1-1/6=5/6.

$$P = C_{10}^{4}\left(\frac{1}{6}\right)^{4}\left(\frac{5}{6}\right)^{6} = \frac{10!}{4!\cdot 6!}\frac{5^{6}}{6^{10}} = \frac{7\cdot 8\cdot 9\cdot 10}{2\cdot 3\cdot 4}\frac{5^{6}}{6^{10}} = 0.0542659$$

Задача 2. Какова вероятность, что число успехов будет от k_1 до k_2?

Очевидно, что цепочки с числом успехов k_1, $k_1+1,\ldots,k_2$ представляют собой несовместные события. Но тогда вероятность суммы равна сумме вероятностей и

$$P = \sum_{m=k_1}^{k_2} C_n^m p^m q^{n-m}$$

Пример. Какова вероятность, что из 10 бросаний монеты орел выпадет от 4 до 6 раз?

Решение. Очевидно, что события выпал орел 4 раза, 5 раз, 6 раз несовместны. Следовательно, вероятность суммы равна сумме вероятностей:

$$P = C_{10}^4\left(\frac{1}{2}\right)^{10} + C_{10}^5\left(\frac{1}{2}\right)^{10} + C_{10}^6\left(\frac{1}{2}\right)^{10} = \frac{672}{1024} = 0.6563$$

Локальная теорема Лапласа

Если вероятность p появления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность того, что событие A появится в n испытаниях ровно k раз, приближенно равна

$$B_{n,p}(k) \approx \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

где $x = (k - np) / \sqrt{npq}$

Интегральная теорема Лапласа

Если вероятность p появления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность того, что событие A появится в n испытаниях от k_1 до k_2 раз, приближенно равна

$$B_{n,p}(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-z^2/2} dz$$

где $x_i = (k_i - np) \Big/ \sqrt{npq}$

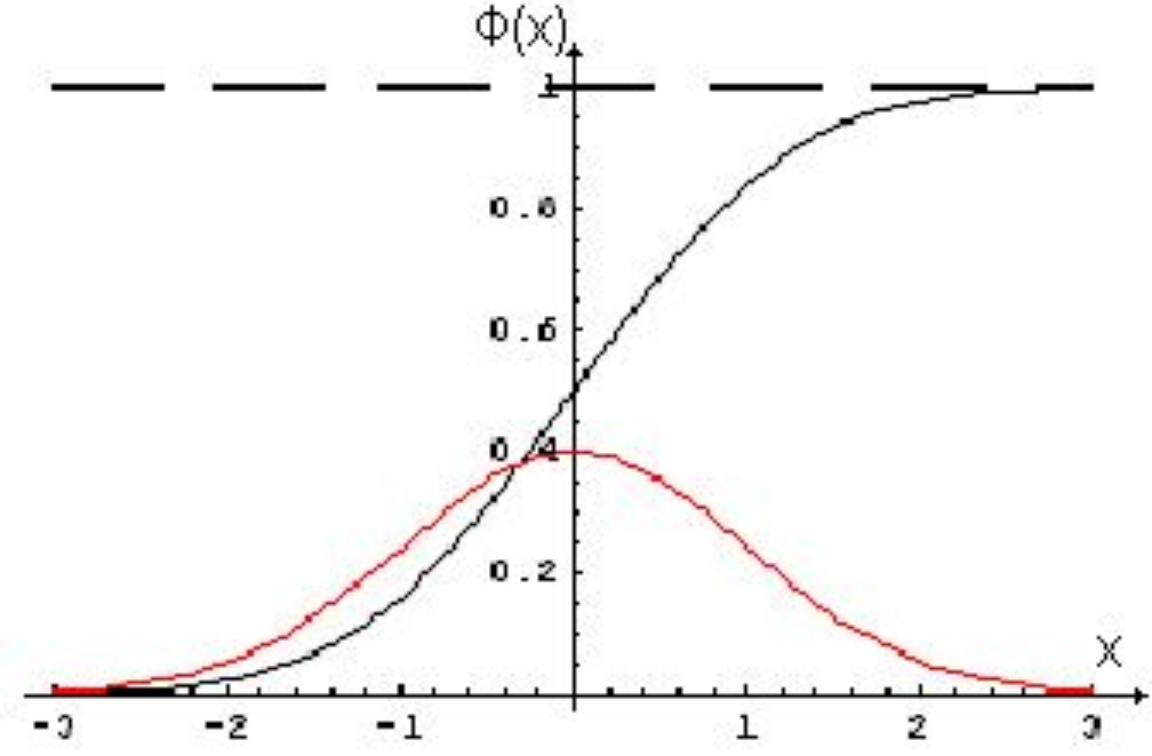
Для вычислений по формуле имеются таблицы. В таблицах приведены значения функции

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

для положительных значений аргумента. Значения для отрицательных значений аргумента вычисляются по формуле

$$\Phi(-x)=1-\Phi(x)$$

На рисунке приведены графики функций Φ(x) (черным цветом) и подынтегральной функции (красным цветом)



Если $-x_1 = x_2 = x$, то справедлива формула

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

$$= \Phi(x) - (1 - \Phi(x)) = 2\Phi(x) - 1$$

Пример 1. Какова вероятность, что из 100 подбрасываний кубика четверка выпадет ровно 30 раз.

n=100, m=30, p=1/6, q=1-1/6=5/6

$$x = \frac{m - np}{\sqrt{npq}} = \frac{30 - 100/6}{\sqrt{100 * 5/36}} = 3.57771$$

$$B_{100,1/6}(30) \approx \frac{1}{\sqrt{100 * 5/36}} \frac{1}{\sqrt{2\pi}} e^{-\frac{3.5778^2}{2}}$$

$$= 0.000177865$$

Пример 2. Какова вероятность, что из 100 подбрасываний кубика 4 выпадет от m_1=15 до m_2=25 раз.

$$x_1 = \frac{m_1 - np}{\sqrt{npq}} = \frac{15 - 100/6}{\sqrt{100 \cdot 5/36}} = -0.447214$$

$$x_2 = \frac{m_2 - np}{\sqrt{npq}} = \frac{25 - 100/6}{\sqrt{100 \cdot 5/36}} = 2.23607$$

$$p = \frac{1}{\sqrt{2\pi}} \int_{-0.447214}^{2.23607} e^{-z^2/2} dz =$$

$$= \Phi(2.23607) - \Phi(-0.447214) = \Phi(2.23607) + \Phi(0.447214) - 1$$

$$\Phi(2.23607) = 0.9911, \quad \Phi(0.447214) = 0.6736$$

$$p = 0.6647$$

Задача 3 . Сколько раз нужно подбросить кубик, чтобы частота отличалась от вероятности не более чем на 0.005 с вероятностью 0.9?

Решение. Основная формула.

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

В нашем случае p=1/6, $\varepsilon=0.005$, γ=0.9.

$$P\left(\left|\frac{m-np}{\sqrt{npq}}\right|<\frac{\varepsilon n}{\sqrt{npq}}\right)=\gamma$$

Роль x_1, x_2 теоремы Лапласа играют $\pm\varepsilon\sqrt{\frac{n}{pq}}$

Из теоремы Лапласа $2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)-1=\gamma$

Подставив числа, получим уравнение относительно n

$$2\Phi\left(0.005\sqrt{\frac{n}{5/36}}\right)-1=0.9$$

$$\Phi\left(0.01341\sqrt{n}\right)=0.95$$

По таблице ищем значение аргумента функции Лапласа такое, что ее значение равно 0.95. Это 1.65. Отсюда находим n

$$0.01341\sqrt{n}=1.65 \qquad \sqrt{n}=122.98$$

n=15125

Конечная случайная величина

- Часто исход случайного эксперимента выражается некоторым числом. Когда каждому элементарному исходу случайного эксперимента мы ставим в соответствие некоторое число x_k, то мы определяем на множестве событий некоторую числовую функцию**.** Набор чисел может быть конечным или бесконечным: это зависит от количества элементарных исходов эксперимента. Неформально говоря, такое число, принимающее случайные значения, и называется случайной величиной.
- Например. Подбрасываем монетку. Поставим в соответствие «орлу» - 1 , «решке» - 0. Можно наоборот: «орлу» - 0 , «решке» - 1. А можно симметрично: «орлу» - -1 , «решке» - 1.
- Каждой грани кубика ставим в соответствие число очков, которое нарисовано на грани.
- Картам, в зависимости от игры назначают то или иное число очков.

- Случайная величина, принимающая конечное число значений, называется *конечной случайной величиной.* Пусть пространство элементарных исходов конечно: $\Omega = \bullet \omega_1, \omega_2, \ldots, \omega_n \bullet$. Вероятность P любого случайного события, связанного с данным экспериментом, полностью определяется набором неотрицательных чисел $p_i = P(\omega_i)$, $i = 1,2,\ldots,n$, таких, что $p_1 + p_2 + \ldots + p_n = 1$.

Такое вероятностное пространство можно представить с помощью таблицы

$$\begin{pmatrix} \omega_1 & \omega_2 & \dots & \omega_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

Функцию $\xi(\omega)$, заданную на конечном числе аргументов, также задаем табличным способом:

$$\begin{pmatrix} \omega_1 & \omega_2 & \dots & \omega_n \\ x_1 & x_2 & \dots & x_n \end{pmatrix}$$

Будем предполагать, что все числа x_k различны. Случайная величина принимает значение x_k , если произошел исход ω_k, вероятность которого равна p_k Точнее: вероятность события $\{\xi(\omega_к)=x_k\}$ равна p_k Конечная случайная величина полностью определяется своими значениями и их вероятностями.

Поэтому таблица

$$\begin{pmatrix} \xi_1 & \xi_2 & \dots & \xi_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad p_1 + p_2 + \ldots + p_n = 1$$

часто отождествляется с самой случайной величиной и называется *законом распределения* конечной случайной величины. Часто закон распределения записывают короче

$$\xi = \begin{pmatrix} x_i \\ p_i \end{pmatrix}, i = 1,2,\ldots, n$$

Например: поставим в соответствие выпаданию орла 1, а решке –1. Можно иначе: орлу –0, решке 1. Возможны иные варианты.

$$\xi_1 = \begin{pmatrix} -1 & 1 \\ 1/2 & 1/2 \end{pmatrix} \qquad \xi_2 = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$$

Совместное распределение случайных величин

Пусть заданы две конечные случайные величины:

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \\ p_1 & p_2 & & p_m \end{pmatrix} \qquad \eta = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \\ q_1 & q_2 & & q_n \end{pmatrix}$$

Событие $\{\xi = x_i, \eta = y_j\}$ состоит в том, что одновременно случайная величина ξ принимает значение x_i, а случайная величина η – значение y_j. Назовем вероятности таких событий *совместными вероятностями* и обозначим их через p_{ij}:

$$p_{ij} = P(\xi = x_i, \eta = y_j), \quad i = 1,2,\ldots,m,\ j = 1,2,\ldots,n.$$

Таблица совместного распределения случайных величин

	x_1	x_2			x_m	
y_1	p_{11}	p_{12}			p_{1m}	q_1
y_2	p_{21}	p_{22}			p_{2m}	q_2
y_n	p_{n1}	p_{n2}			p_{nm}	q_n
	p_1	p_2			p_m	1

Две конечные случайные величины называются *независимыми,* если события $\{\xi = x_i\}$ и $\{\eta = y_j\}$ независимы при всех $i=1,2,\ldots,m$ и $j=1,2,\ldots,n$. В противном случае случайные величины *зависимы.* Для независимых случайных величин совместное распределение строится по известным распределениям величин ξ и η:

$$p_{ij} = P(\xi = x_i)P(\eta = y_j) = p_i q_j$$

$$i = 1,2,\ldots,m,\ j = 1,2,\ldots,n$$

Пусть заданы две конечные случайные величины:

$$\xi = \begin{pmatrix} x_1 & x_2 & \ldots & x_m \\ p_1 & p_2 & \ldots & p_m \end{pmatrix} \qquad \eta = \begin{pmatrix} y_1 & y_2 & \ldots & y_n \\ q_1 & q_2 & \ldots & q_n \end{pmatrix}$$

Их *суммой* называется случайная величина $\xi + \eta$, значениями которой являются всевозможные суммы

$$x_i + y_j,\ i = 1,\ldots,m,\ j = 1,\ldots,n$$

с совместными вероятностями $p_{ij} = P(\xi = x_i, \eta = y_j)$

Произведением этих случайных величин называется случайная величина $\xi\,\eta$, значениями которой являются всевозможные произведения $x_i y_j$ с теми же вероятностями p_{ij}.

Пусть заданы две конечные независимые случайные величины:

$$\xi = \eta = \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$$

Такие случайные величины называются биномиальными. Вычислим закон распределения $\xi + \eta$. Возможные значения суммы: 0- принимается с вероятностью q^2, значение 1 – принимает в двух случаях с вероятностями pq и значение 2 – с вероятностью p^2.

В результате получим таблицу

$$\xi+\eta=\begin{pmatrix} 0 & 1 & 2 \\ q^2 & 2pq & p^2 \end{pmatrix}$$

Этот результат можно обобщить на любое число слагаемых.

Теорема. Пусть $\xi_1, \xi_2, \ldots, \xi_n$ независимые бернулливые случайные величины. Тогда их сумма есть биномиальная случайная величина

$$B_{n,p}=\xi_1+\xi_2+\boxtimes+\xi_n$$

Иная трактовка: если ξ_k – число успехов в k-ом испытании, то число успехов в n испытаниях есть их сумма.

Математическое ожидание

Математическим ожиданием конечной случайной величины

$$\xi = \begin{pmatrix} x_i \\ p_i \end{pmatrix}, i = 1, \ldots, m$$

называется число

$$M\xi = x_1 p_1 + x_2 p_2 + \ldots + x_m p_m = \sum_{i=1}^{m} x_i p_i$$

Понятие математического ожидания упрощенно можно представить иначе. Пусть z_1, z_2, z_3,…, z_k - результаты некоторого испытания, описываемого случайной величиной ξ. Среднее значение случайной величины за большое число k испытаний будет

$$\frac{z_1 + z_2 + \square + z_k}{k}$$

Среди этих значений соберем все равные x_1, x_2,…. В результате получится

$$\frac{x_1 n_1 + x_2 n_2 + \square + x_m n_m}{k} =$$

$$= x_1 \frac{n_1}{k} + x_2 \frac{n_2}{k} + \square + x_m \frac{n_m}{k}$$

Но отношение n_j / k есть частота появления значения x_j. А частота при большом числе опытов близка к вероятности. В итоге получаем формулу из определения.

Например, при бросании кубика вероятность выпадения каждой грани равна 1/6. Тогда математическое ожидание числа очков равно

$$M\xi = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

Математическое ожидание обладает следующими свойствами.

1. Математическое ожидание постоянной равно ей самой:

$$Mc = c$$

2. Если случайная величина принимает только неотрицательные значения, то

$$M\xi \geq 0$$

3. Константу можно выносить *за* знак математического ожидания:

$$M\left(c\xi\right) = cM\xi$$

4. Математическое ожидание суммы (разности) случайных величин равно сумме (разности) их математических ожиданий:

$$M\left(\xi \pm \eta\right) = M\xi \pm M\eta$$

5. Для любой случайной величины справедливо равенство

$$M\left(\xi - M\xi\right) = 0$$

Операция вычитания математического ожидания из случайной величины называется *центрированием*

6. Математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий

$$M\left(\xi\eta\right) = M\xi \cdot M\eta$$

Математическое ожидание биномиальной случайной величины.

Числовое значение величины – кол-во успехов в серии испытаний. Одно испытание можно рассматривать как серию из одного испытания. Назначим успеху числовое значение 1, а неуспеху – 0. Следовательно, математическое ожидание в одном опыте равно p. Представим серию опытов как сумму отдельных испытаний. Тогда по свойству математического ожидания M=np.

Дисперсия

Дисперсией конечной случайной величины ξ называется число

$$D\xi = M\left(\xi - M\xi\right)^2$$

по определению математического ожидания, дисперсия вычисляется по следующей формуле

$$D\xi = \sum_i \left(x_i - M\xi\right)^2 p_i$$

Дисперсию иногда обозначают как $\sigma^2(\xi)$ или σ_ξ^2

$\sigma_\xi = \sqrt{D\xi}$ называется *среднеквадратичным отклонением* или *стандартным отклонением* случайной величины

Свойства дисперсии

1. *Дисперсия любой случайной величины неотрицательна* $D\xi>0$

При этом $D\xi=0$ *тогда и только тогда, когда случайная величина постоянна.*

2. *Константа выносится из-под знака дисперсии с квадратом*

$$D(c\xi) = c^2 D\xi$$

3. *Сдвиг на константу не меняет дисперсии:*

$$D(\xi + c) = D\xi$$

4. *Дисперсия суммы независимых случайных величин равна сумме их дисперсий*:

$$D(\xi + \eta) = D\xi + D\eta \qquad (\xi \text{ и } \eta \text{ независимы})$$

5. Дисперсия равна "среднему квадрата минус квадрат среднего":

$$D\xi = M\xi^2 - \left(M\xi\right)^2$$

Дисперсия биномиальной случайной величины.

Вычисление проведем по той же схеме, что и для математического ожидания. Биномиальная случайная величина есть сумма n независимых бернуллиевых величин. Но тогда используем формулу дисперсии суммы:

$$DB_{n,p} = D\xi_1 + D\xi_2 + \square + D\xi_n$$

Но $$D\xi_k = \left(p-1\right)^2 \cdot p + \left(p-0\right)^2 \cdot q = q^2 p + p^2 q =$$

$$= pq\left(p+q\right) = pq$$

и $$DB_{n,p} = npq$$

Случайная величина $$\xi^* = \frac{\xi^0}{\sqrt{D\xi}} = \frac{\xi - M\xi}{\sqrt{D\xi}}$$

называется стандартизованной (по отношению к ξ) или просто стандартизацией ξ

Стандартизованная случайная величина имеет нулевое математическое ожидание и единичную дисперсию.

Пример. Дисперсия при бросании кубика. Математическое ожидание 3.5. Считаем математическое ожидание квадрата случайной величины:

$$\frac{1}{6}1+\frac{1}{6}4+\frac{1}{6}9+\frac{1}{6}16+\frac{1}{6}25+\frac{1}{6}36=15.1667$$

$$D\xi = 15.1667 - 12.25 = 2.9167$$

Среднее квадратичное отклонение $\sigma\xi = 1.7078$

Таблица стандартизованных значений

-1.4639	-0.8783	-0.2928	0.2928	0.8783	1.4639

Задача. Проводится лотерея. Разыгрывается 50 билетов по 1 рублю. Известно, что среди билетов 1 выигрывает 30 руб., 2 – по 10 руб. Приобретено 2 билета. Вычислить математическое ожидание чистого дохода.

Коэффициент корреляции

Ковариацией двух случайных величин ξ и η (или *ковариацией между* ξ и η) называется число

$$\operatorname{cov}(\xi\eta) = M\left(\xi^0\eta^0\right) = M\left((\xi - M\xi)(\eta - M\eta)\right)$$

Из определения следуют некоторые простые свойства ковариации

1. $$\operatorname{cov}(\xi\eta) = M(\xi\eta) - M\xi \cdot M\eta$$

2. Ковариация коммутативна:

$$\operatorname{cov}(\xi\eta) = \operatorname{cov}(\eta\xi)$$

3. Ковариация суммы случайных величин

$$D(\xi + \eta) = D\xi + D\eta + 2\operatorname{cov}(\xi\eta)$$

4. Ковариация случайной величины с собой

$$\operatorname{cov}(\xi\xi) = D\xi$$

Следующее свойство важно при оценке степени зависимости двух случайных величин.

5. Если случайные величины ξ *и* η *независимы, то их ковариация равна нулю.*

Для независимых величин ξ и η их центрированные величины также независимы.

Поэтому $$\operatorname{cov}(\xi\eta) = M\xi^0 \cdot M\eta^0 = 0$$

Ковариация стандартизованных величин называется *коэффициентом, корреляции* между случайными величинами ξ и η

$$r_{\xi\eta} = \frac{\operatorname{cov}(\xi\eta)}{\sqrt{D\xi}\sqrt{D\eta}} = \frac{M\left[(\xi - M\xi)(\eta - M\eta)\right]}{\sqrt{D\xi}\sqrt{D\eta}}$$

Предполагается, что случайные величины ξ и η имеют ненулевые дисперсии

свойства коэффициента корреляции:

1. $$r_{\xi\eta} = M\left(\xi^* \eta^*\right)$$

2. Коэффициенты корреляции между ξ и η и между их стандартизациями совпадают

$$r_{\xi\eta} = r_{\xi^*\eta^*}$$

3. Коэффициент корреляции всегда по модулю меньше 1

$$\left|r_{\xi\eta}\right| \le 1$$

4. Если ξ и η независимы, то $r_{\xi\eta} = 0$

5. *Коэффициент корреляции равен* +1 или -1 *тогда и только тогда, когда случайные величины линейно зависимы:*

$$\left|r_{\xi\eta}\right| = 1 \Leftrightarrow \eta = a\xi + b$$

Вычислительная формула $$r_{\xi\eta} = \frac{M(\xi\eta) - M\xi \cdot M\eta}{\sqrt{D\xi \cdot D\eta}}$$

Примеры. Даны таблицы распределения. Найти коэффициенты корреляции

P	0.2	0.2	0.2	0.2	0.2
X	-2	-1	0	1	2
Y	1	0	-1	-2	-3

P	0.2	0.2	0.2	0.2	0.2
X	-2	-1	0	1	2
Y	1	2.5	4	5.5	6

P	0.2	0.2	0.2	0.2	0.2
X	-2	-1	0	1	2
Y	4	1	0	1	4

P	0.2	0.2	0.2	0.2	0.2
X	-2	-1	0	1	2
Y	1	0	1	2	1

1) Mx=0, Dx=2, MY=1, DY=2 MXY=-2, R=-1

2) Mx=0, Dx=2, MY=3.8, DY=3.46 MXY=2.6, R=0.9884

3) Mx=0, Dx=2, MY=2, DY=2.8 MXY=0, R=0

4) Mx=0, Dx=2, MY=1, DY=0.4 MXY=0.4, R=0.4472

Функция распределения

Функция действительной переменной

$$F_\xi(x) = P(\xi < x)$$

называется *функцией распределения* случайной величины ξ .

Свойства функции распределения

1. $$P(\xi \geq x) = 1 - F_\xi(x)$$

2. $$P(a \leq \xi < b) = F_\xi(b) - F_\xi(a)$$

3. При любом x выполняется неравенство.

$$0 \leq F_\xi(x) \leq 1$$

Это справедливо, поскольку функция распределения есть вероятность

4. Функция распределения есть неубывающая функция.

5. При $x \to -\infty$ событие стремится к невозможному и вероятность соответственно, стремится к нулю. При $x \to +\infty$ событие становится достоверным

6. Функция распределения непрерывна слева, то есть

$$\lim_{x \to x_0 + 0} F_\xi(x) = F_\xi(x_0)$$

Случайная величина ξ называется *непрерывной случайной величиной,* если существует функция $f_\xi(x)$ такая, что

$$P(\xi \in [a,b]) = \int_a^b f_\xi(x)dx$$

Функция $f_\xi(x)$ называется плотностью *вероятности* или *плотностью распределения* случайной величины ξ

7. *Для любой непрерывной случайной величины*

$$P(a \le \xi \le b) = P(a \le \xi < b) = P(a < \xi \le b) =$$

$$= P(a < \xi < b) = \int_a^b f_\xi(x)dx$$

8. *Функция распределения непрерывной случайной величин имеет вид*

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(x)dx$$

Свойства плотности функции распределения

1. *Функция* $f_\xi(x)$ *неотрицательна при всех x*

2. Условие нормировки. *Справедливо равенство*

$$\int_{-\infty}^{+\infty} f_\xi(x)dx = 1$$

3. *В точках непрерывности плотность вероятности равна производной функции распределения:*

$$F'_\xi(x) = f_\xi(x)$$

Математическим ожиданием непрерывной случайной величины называется число

$$M\xi = \int_{-\infty}^{+\infty} x\, f_\xi(x)dx \qquad \sum_{i=1}^{n} z_i f(z_i)\Delta x_i$$

(если соответствующий интеграл существует).

Дисперсия вычисляется через интеграл:

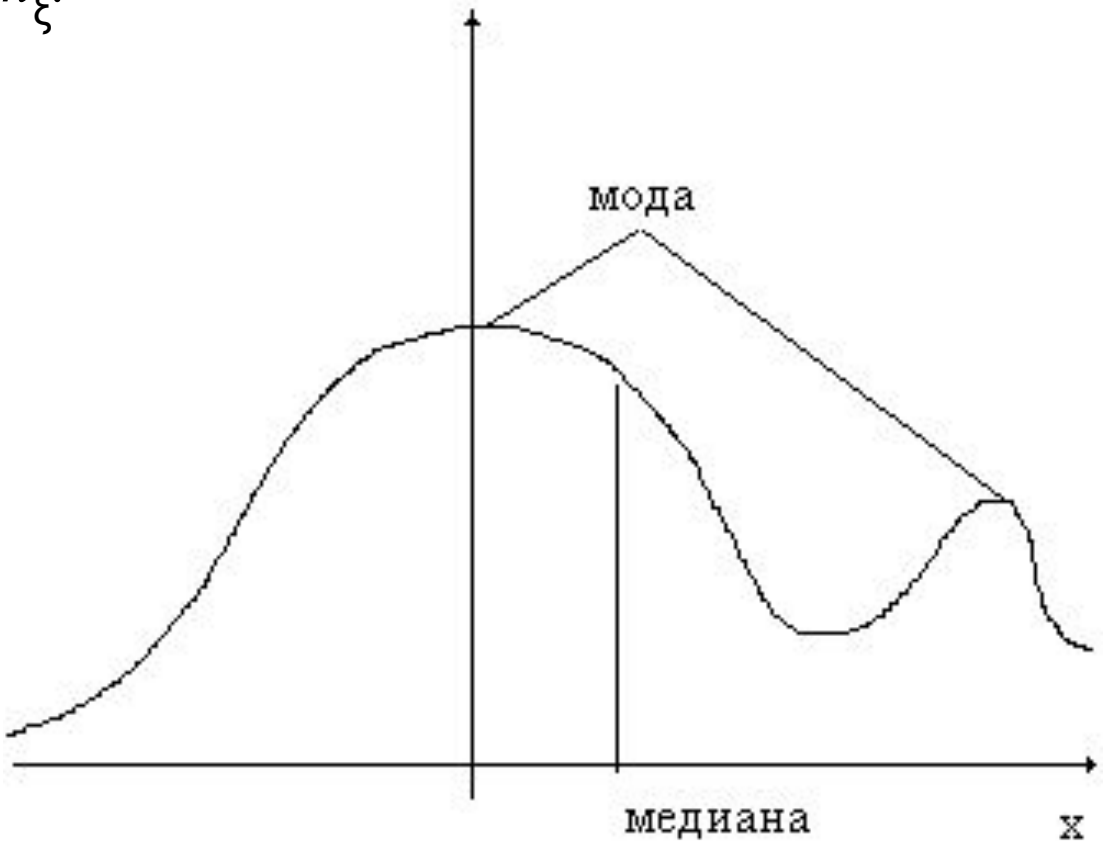
$$D\xi = \int_{-\infty}^{+\infty} \left(x - M\xi\right)^2 f_\xi(x)dx$$

(если интеграл существует).

Некоторые определения

- *Квантилью* случайной величины ξ порядка p называется число x_p такое, что вероятность события $\{\xi<x\}$ равна p.
- *Модой* распределения случайной величины ξ называется точка локального максимума плотности распределения
- *Медианой* называется квантиль $x_{0.5}$ порядка 0.5 (50-процентная квантиль) распределения m_ξ.

На рисунке показано полимодальное распределение



Биномиальное распределение

$$P\left(\left\{\xi\left(\omega\right)=x_k\right\}\right)=P_n\left(k\right)=C_n^k p^k q^{n-k},$$

где $q=1-p$

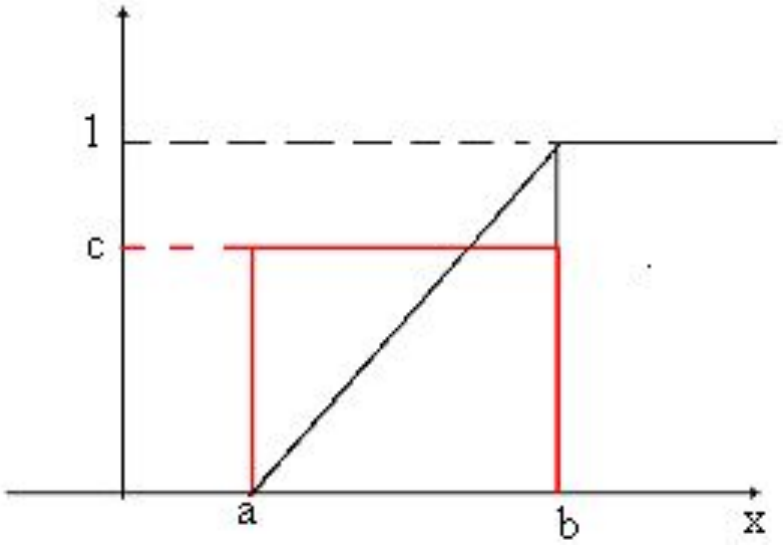
Распределение Пуассона. Получается как предельное при очень большом числе испытаний маловероятных событий.

$$P_n(k)=\lambda^k e^{-\lambda}\big/k! \qquad \lambda=n\cdot p$$

Равномерное распределение

График равномерного на отрезке (a,b) распределения представлен на рисунке. Значение C определяется из условия нормировки.

$$c = \frac{1}{b - a}$$



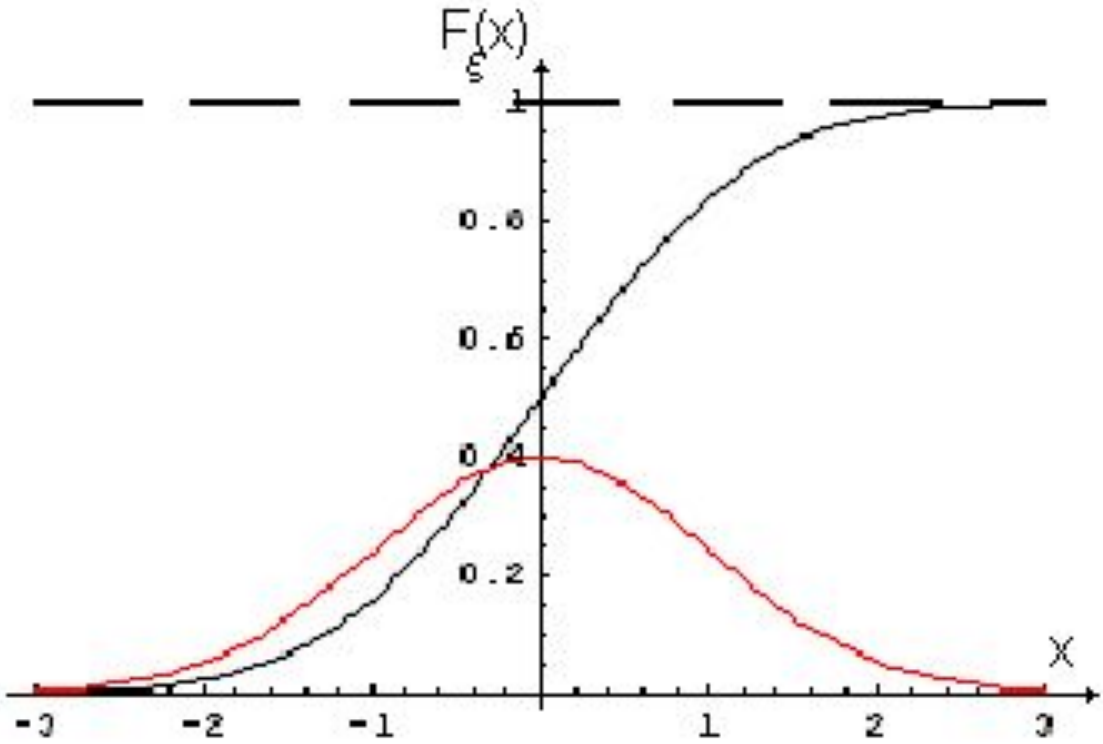
Распределение Гаусса

Говорят, что случайная величина ξ , распределена по *нормальному закону* (имеет *нормальное распределение)* с параметрами m и σ, (σ>0) если она имеет плотность распределения

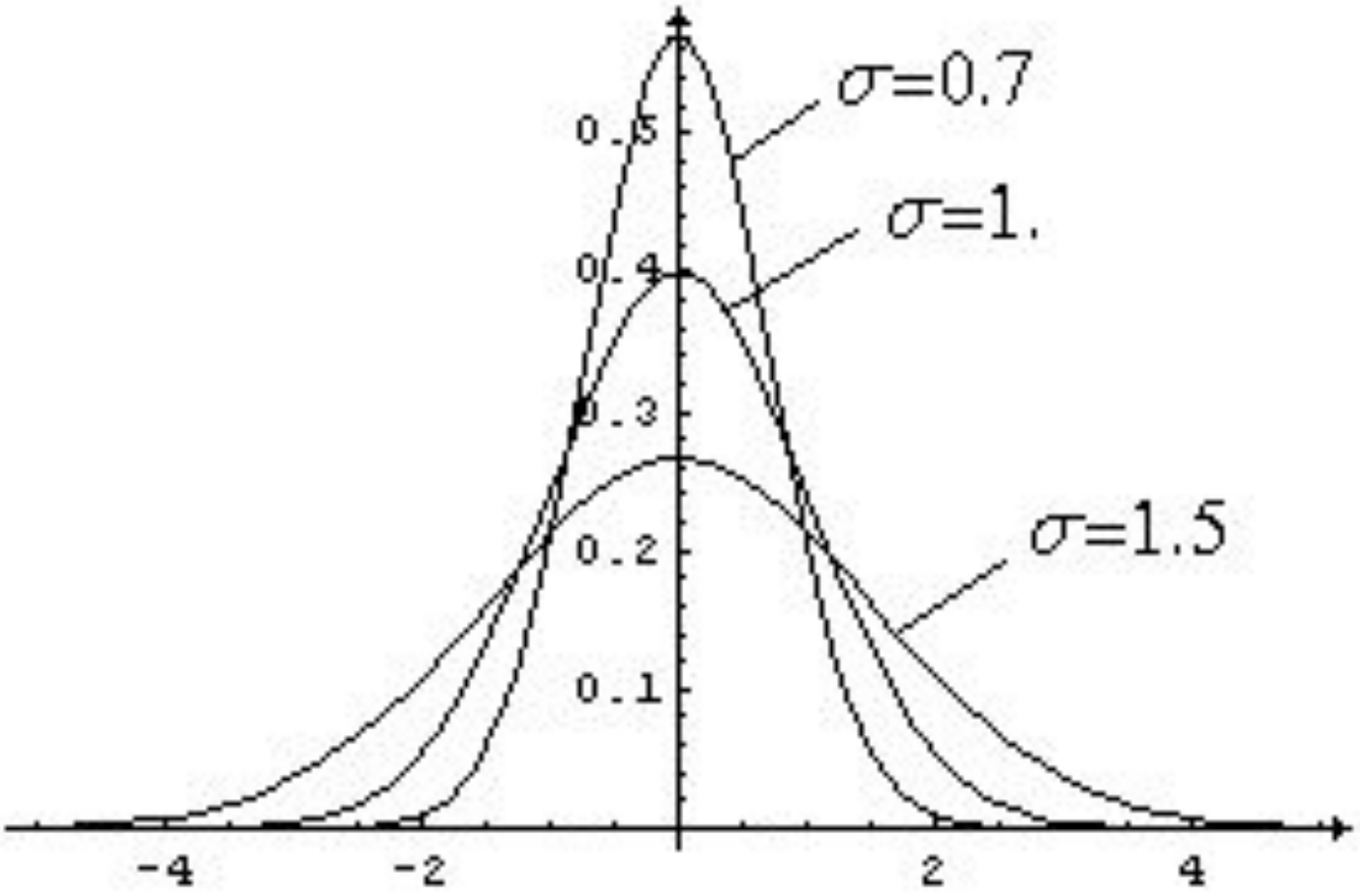
$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

$$F_\xi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

На рисунке представлены графики стандартного (при m=0 и s=1) нормального распределения Гаусса (черный) и его плотности (красный)



Графики плотности нормального распределения при различных значениях дисперсии



Свойства нормального распределения

- график симметричен относительно прямой $x=m$;
- функция достигает максимума в точке $x=m$;
- график приближается к нулю при возрастании $|x|$

$$\lim_{x\to-\infty} f_\xi(x) = \lim_{x\to+\infty} f_\xi(x) = 0$$

Нормальное распределение обозначают $N(m,\sigma)$. Нормальное распределение с параметрами $m=0$, $\sigma=1$ называется *стандартным* нормальным распределением и задается плотностью

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

Функция распределения стандартной нормальной случайной величины обозначается через $\Phi(x)$

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

Пусть $\xi \sim N(m,\sigma)$. Тогда квантиль x_p случайной величины ξ связана с квантилью стандартного нормального распределения следующим соотношением:

$$x_p = m + \sigma z_p$$

Законы больших чисел
Теорема Бернулли

Пусть μ_n *- число успехов в n испытаниях Бернулли, p - вероятность успеха в единичном испытании. Тогда относительная частота успеха сходится по вероятности к вероятности p. Другими словами, для любого* $\varepsilon > 0$ *выполняется предельное соотношение*

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

Центральная предельная теорема Ляпунова

Пусть случайные величины $X_1, X_2, \ldots, X_n$ *независимы, одинаково распределены с математическим ожиданием M и конечной дисперсией* σ^2. *Тогда справедливо предельное соотношение*

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \ldots + X_n - nM}{\sigma\sqrt{n}} < x \right) = \Phi(x)$$

Теорема Чебышева

Если $X_1, X_2, \ldots, X_n, \ldots$ - попарно независимые случайные величины, причем дисперсии их равномерно ограничены, то как бы мало ни было $\varepsilon > 0$, вероятность неравенства

$$\left| \frac{X_1 + X_2 + \ldots + X_n}{n} - \frac{MX_1 + MX_2 + \ldots + MX_n}{n} \right| < \varepsilon$$

Сколь угодно близка к единице, если n достаточно велико.

Статистика

- Генеральной совокупностью называется вся совокупность исследуемых объектов
- Выборочной совокупностью или просто выборкой называют совокупность случайно отобранных из генеральной совокупности объектов
- Объемом совокупности называют число объектов этой совокупности

Способы формирования выборочной совокупности

- Повторный – после измерений объект возвращают в генеральную совокупность
- Бесповторный – после измерений объект в генеральную совокупность не возвращается

Выборка должна быть репрезентативной - представительной. Для этого объекты из генеральной совокупности должны отбираться случайно.

- **Простой случайный отбор** – объекты извлекают по одному из всей генеральной совокупности
- **Типический отбор -** объекты отбирают не из всей генеральной совокупности, а из каждой ее «типической части»
- **Механический отбор** – генеральную совокупность делят механически на несколько групп и из каждой группы отбирают один объект
- **Серийный отбор** – объекты из генеральной совокупности отбирают не по одному, а сериями, которые подвергают сплошному обследованию.

На практике, как правило, используется смешанная схема.

Выборка и ее обработка

- Упорядочивание. Элементы выборки $x_1, x_2, \ldots, x_n$ располагаются в порядке возрастания.
- Частотный анализ. Пусть выборка содержит k различных значений $z_1, z_2, \ldots, z_k$, причем z_i встречается n_i (i=1,2,…,k) Число n_i называют частотой элемента z_i,

$$\sum_{i=1}^{k} n_i = n$$

- Совокупность пар (z_i, n_i) называют статистическим рядом выборки. Часто его представляют в виде таблицы – в первой строке z_i, во второй n_i.
- Величина $v_i = n_i / n$ называется относительной частотой
- Накопленная частота значения z_i равна $n_1+n_2+\ldots+n_i$.
- Относительная накопленная частота $v_1+v_2+\ldots+v_i$

- Группировка**.** При большом объеме выборки ее элементы объединяют в группы, представляя результаты опытов в виде *группированного статистического ряда.* Для этого интервал, содержащий все значения выборки, разбивается на k интервалов. Для выборки большого объема число интервалов определяется по формуле Стерджесса $k = 1 + 3.322 \ln(n)$

- Удобнее всего разбивать на равные интервалы. При этом считается, что правая граница интервала принадлежит следующему интервалу. Последний интервал включает правую границу. После этого подсчитываются частоты – количество n_i элементов выборки, попавших в *i*-й интервал. Получающийся статистический ряд в первой строке содержит середины интервалов группировки z_i, а во второй строке -частоты n_i, попадания в соответствующий интервал. Наряду с частотами подсчитываются относительные частоты v_i, накопленные частоты и накопленные относительные частоты. Результаты обычно сводятся в *таблицу частот группированной выборки,* а процесс формирования такой таблицы называется *частотной табуляцией* выборки.

Пример

Дана выборка

0,0473 0,0543 0,0561 0,0989 0,1107 0,1112 0,1204
0,1647 0,2030 0,2138 0,2147 0,2463 0,2725 0,2734
0,3029 0,3222 0,3389 0,3841 0,3909 0,4037 0,4071
0,4173 0,4238 0,4308 0,4451 0,5382 0,5454 0,5472
0,6124 0,6320 0,6417 0,6776 0,6908 0,7399 0,7715
0,7853 0,8038 0,8174 0,8201 0,8287 0,8693 0,8704
0,8704 0,8718 0,8965 0,9025 0,9130 0,9366 0,9629

Она содержит 49 чисел в отрезке [0,1]. Все числа различны.

Проведем группировку. Разобьем отрезок на 10 полуинтервалов **[0,0.1),[0.1,0.2),…[0.8,0.9),[0.9,1.0]**. Подсчитаем, сколько элементов выборки попало в каждый интервал и получим статистический ряд

0.05	**0.15**	**0.25**	**0.35**	**0.45**	**0.55**	**0.65**	**0.75**	**0.85**	**0.95**
4	**4**	**6**	**5**	**6**	**3**	**5**	**3**	**9**	**4**

Обработку этого примера продолжим в дальнейшем.

Эмпирическая функция распределения

Каждой выборке $\{x_1, x_2, ..., x_n\}$ можно поставить в соответствие конечную случайную величину, принимающую эти значения с равными вероятностями 1/n

$$\xi_n = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1/n & 1/n & \dots & 1/n \end{pmatrix}$$

Это распределение называется *выборочным, или эмпирическим, распределением.* Как и для любой конечной случайной величины, для эмпирической случайной величины можно построить ступенчатую функцию распределения; она называется *выборочной функцией распределения.* Кроме того, можно вычислить все числовые характеристики выборочной случайной величины ξ_n- математическое ожидание, дисперсию, СКО, медиану и т.д.

Все эти величины снабжаются определением "выборочный": *выборочное математическое ожидание* (его обычно называют *выборочным средним), выборочная дисперсия, выборочная медиана* и т.д. Например, выборочное среднее (его обозначают через $\bar{x}$) есть не что иное как среднее арифметическое значений выборки

$$M\xi_n = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Соответственно выборочная дисперсия s^2 равна

$$D\xi_n = s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Оценки параметров распределения
Точечные оценки

Будем предполагать, что имеется выборка $\{x_1, x_2, ...x_n\}$ из генеральной совокупности с функцией распределения. $F_\xi(x)$ Для удобства опустим индекс ξ в обозначении функции распределения. Пусть функция распределения на самом деле зависит от неизвестного параметра θ : $P(x_k < x) = F(x, \vartheta)$

Одна из главных задач математической статистики - оценить значение параметра ϑ, имея в распоряжении только выборку. Например, нам известно, что генеральная совокупность распределена по биномиальному закону при 10 испытаниях. Неизвестным параметром в этом случае является вероятность p успеха в единичном испытании. Иногда требуется оценить несколько параметров. Например, требуется оценить математическое ожидание *m* и дисперсию σ^2 нормально распределенной генеральной совокупности; у равномерного распределения - границы отрезка [a, b] и т.д.

Оценка ϑ_n^* является случайной величиной

$$\vartheta_n^* = \vartheta_n^*(x_1, x_2, \ldots x_n)$$

Индекс *n* в обозначении оценки напоминает, что она получена по выборке объема n, «звездочка» показывает, что это не истинное значение параметра, а его оценка. Произвольную функцию от выборки называют еще *статистикой.*

Оценка ϑ_n^* является случайной величиной

Оценка называется несмещенной, если при любом объеме выборки n ее математическое ожидание совпадает с истинным значением параметра $M\vartheta_n^* = \vartheta$

Разность $M\vartheta_n^* - \vartheta$ называется смещением оценки ϑ_n^*. Несмещенная оценка имеет нулевое смещение.

Оценка называется *состоятельной,* если при увеличении объема выборки вероятность того, что оценка мало отличается от истинного значения, приближается к единице.

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P(|\vartheta_n^* - \vartheta| < \varepsilon) = 1$$

Если ϑ_n^**- несмещенная оценка параметра и ее дисперсия стремится к нулю при* $n \to \infty \ (D\vartheta_n^* \to 0)$ *, то данная оценка является состоятельной .*

Качество оценки характеризуют *средним квадратом ошибки*

$$\delta = M(\vartheta_n^* - \vartheta)^2$$

Для несмещенных оценок ($M\vartheta_n^* = \vartheta$) этот показатель равен дисперсии оценки. Если ϑ_{n1}^* и ϑ_{n2}^* две несмещенные оценки параметра ϑ и $D\vartheta_{n1}^* < D\vartheta_{n1}^*$, то говорят, что первая оценка *эффективнее* второй.

Несмещенная оценка называется *наиболее эффективной* (или просто *эффективной),* если она имеет минимальную дисперсию среди всех несмещенных оценок данного параметра.

Теорема Бернулли. *Пусть* μ_n *- число успехов в n испытаниях Бернулли, p - вероятность успеха в единичном испытании. Тогда относительная частота успеха сходится по вероятности к вероятности p:*

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

Или в терминах статистики: относительная частота есть состоятельная оценка вероятности.

Оценка является также и несмещенной

$$M\left(\frac{\mu_n}{n}\right) = \frac{1}{n} M\mu_n = \frac{1}{n} MB_{p,n} = \frac{1}{n} np = p$$

ОЦЕНКА ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Пусть в нашем распоряжении имеется выборка $\{x_1, x_2, ...x_n\}$ из генеральной совокупности с функцией распределения *F(x)*. Функция распределения $F_n^*(x)$ эмпирической случайной величины

$$\xi_n = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ 1/n & 1/n & & 1/n \end{pmatrix}$$

есть вероятность события $\{\xi_n < x\}$:

Пусть среди значений выборки имеется $\mu_n(x)$ чисел, меньших данного числа *x*. Тогда, очевидно,

$$F_n^*(x) = \frac{\mu_n(x)}{n}$$

Покажем, что выборочная функция распределения $F_n^*(x)$ есть оценка функции распределения генеральной совокупности. Зададимся числом ***x***; и применим схему Бернулли. Будем считать успехом событие, состоящее в том, что выборочное значение меньше ***x***.

Поскольку каждое значение из выборки есть случайная величина с функцией распределения, то вероятность успеха равна $p=F(x)$. Число успехов равно $\mu_n(x)$, а относительная частота успеха равна $\mu_n(x)/n$ и совпадает с выборочной функцией распределения.

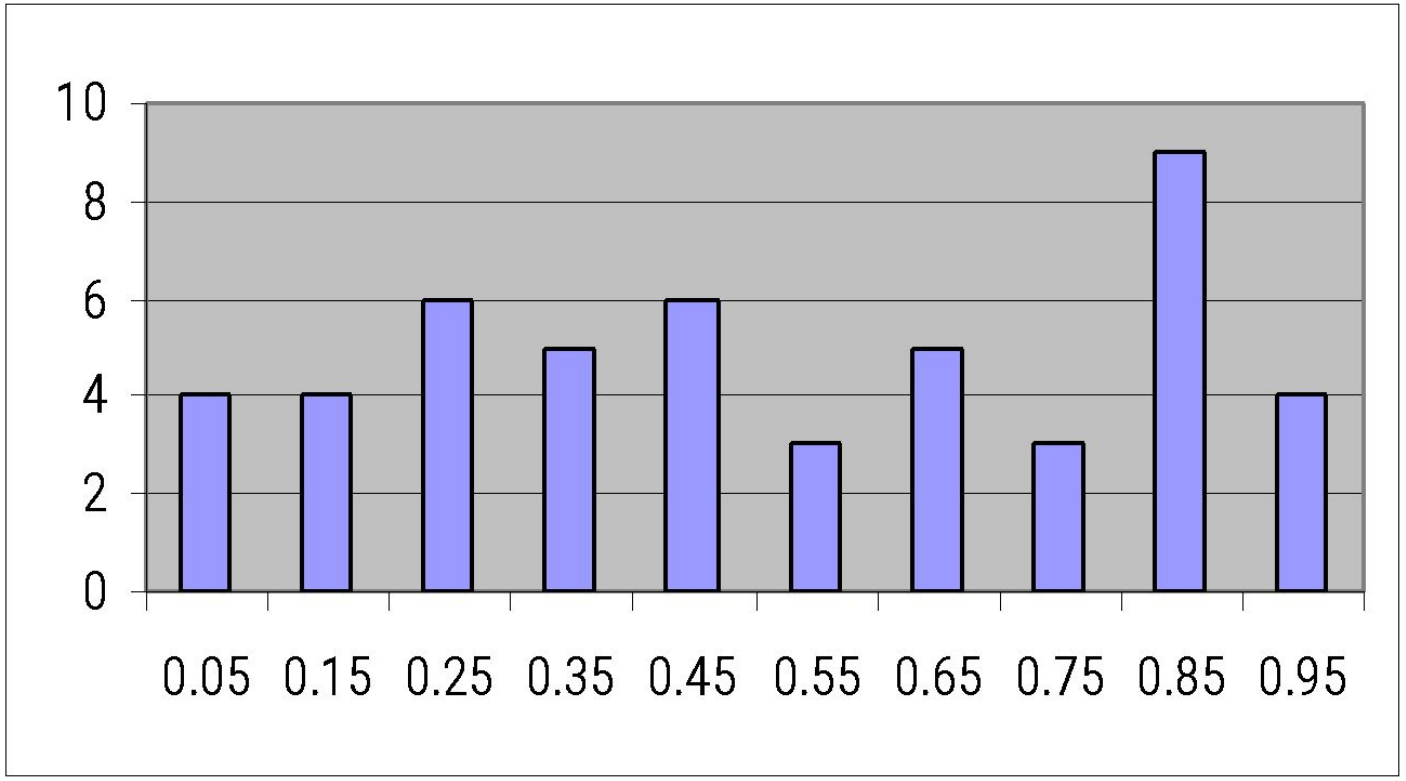
Следовательно, выборочная функция распределения представляет собой относительную частоту успеха, а функция распределения генеральной совокупности - вероятность успеха. Из предыдущего нам известно, что относительная частота есть несмещенная состоятельная оценка вероятности. Значит, выборочная функция распределения действительно является несмещенной, состоятельной и эффективной оценкой функции распределения:

$$MF_n^*(x) = F(x)$$

$$\lim_{n\to\infty} P\left(\left|F_n^*(x) - F(x)\right| < \varepsilon\right) = 1$$

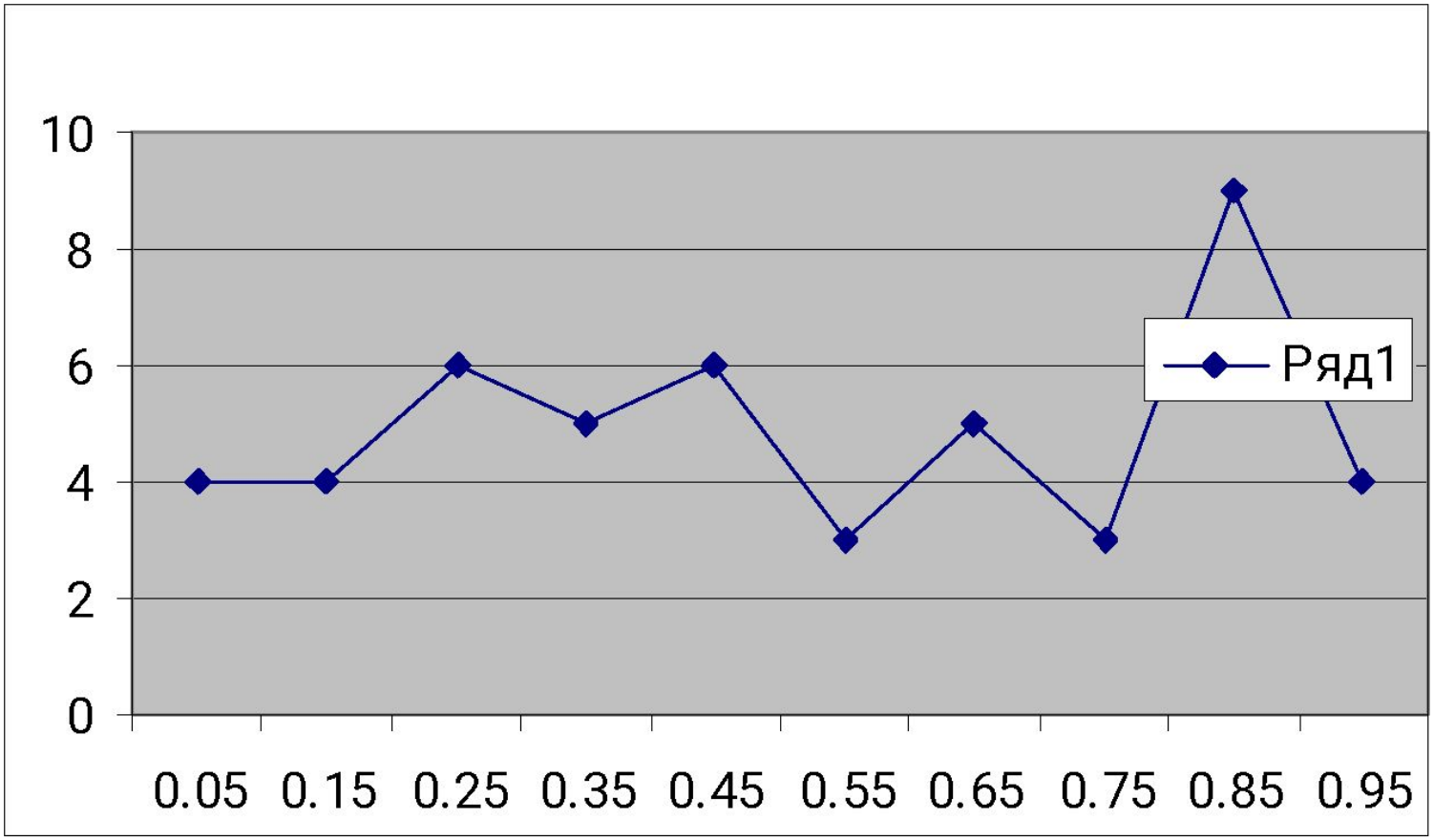
Гистограмма

Для оценки плотности распределения генеральной совокупности используется специальный график - *гистограмма*



На рисунке представлена гистограмма, построенная по примеру, рассмотренному ранее.

Полигон



Если соединить отрезками середины верхних сторон прямоугольников гистограммы, получится еще одно графическое представление для плотности распределения – *полигон*. На рисунке представлен полигон, построенный на основе примера.

Точечная оценка математического ожидания

Выборочное среднее

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

дает несмещенную и состоятельную оценку математического ожидания

$$M = Mx_1 = Mx_2 = \ldots = Mx_n = M\xi$$

Найдем математическое ожидание оценки : $\overline{x}$

$$M\overline{x} = M\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n}\left(Mx_1 + Mx_2 + \ldots + Mx_n\right) = \frac{nM}{n} = M$$

Для проверки состоятельности этой оценки найдем ее дисперсию, обозначив дисперсию генеральной совокупности через σ^2

$$D\overline{x} = D\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n^2}\left(Dx_1 + Dx_2 + \ldots + Dx_n\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Точечная оценка дисперсии

Оценкой дисперсии является выборочная дисперсия

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Вычислим математическое ожидание выборочной дисперсии. Для этого преобразуем выражение для s^2 (через M обозначено математическое ожидание генеральной совокупности):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - M\right) - \left(\bar{x} - M\right)\right)^2 =$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}\left(x_i - M\right)^2 - 2\left(\bar{x} - M\right)\sum_{i=1}^{n}\left(x_i - M\right) + n\left(\bar{x} - M\right)^2\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - M\right)^2 - \frac{2}{n}\left(\bar{x} - M\right)\sum_{i=1}^{n}\left(x_i - M\right) + \left(\bar{x} - M\right)^2$$

Рассмотрим сумму из второго слагаемого в квадратных скобках:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - M\right) = \frac{x_1 + x_2 + \ldots + x_n - nM}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} - M = \bar{x} - M$$

в итоге получаем

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - M\right)^2 - \left(\bar{x} - M\right)^2$$

Тогда математическое ожидание выборочной дисперсии будет

$$Ms^2 = \frac{1}{n}\sum_{i=1}^{n}M\left(x_i - M\right)^2 - M\left(\bar{x} - M\right)^2$$

В итоге

$$Ms^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Если домножить выборочную дисперсию s^2 на дробь $\frac{n}{n-1}$ то получится несмещенная оценка

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Приведенное выражение дает состоятельную несмещенную оценку дисперсии генеральной совокупности

Для вычисления выборочной дисперсии можно вывести более удобную формулу

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2$$

Пример

$$\bar{x} = \frac{25,35}{49} = 0.51735$$

$$\overline{x^2} = \frac{17,1225}{49} = 0,34944$$

Z_i	n_i	Z_i^2	$Z_i n_i$	$Z_i^2 n_i$
0.05	4	0.0025	0,2	0,01
0.15	4	0.0225	0,6	0,09
0.25	6	0.0625	1,5	0,375
0.35	5	0,1225	1,75	0,6125
0.45	6	0,2025	2,7	1,215
0.55	3	0,3025	1,65	0,9075
0.65	5	0,4225	3,25	2,1125
0.75	3	0,5625	2,25	1,6875
0.85	9	0,7225	7,65	6,5025
0.95	4	0,9025	3,8	3,61
	49		25,35	17,1225

s^2=0,34944-0,51735^2= 0,081791

S^2=49 s^2 /48=0,083495

Выборочные мода, медиана, квантили

Выборочные мода, медиана и квантиль легко определяются по упорядоченной, но не сгруппированной выборке.

- Медиана – середина вариационного ряда. Справа и слева располагается одинаковое число значений выборки.
- Мода– наиболее часто встречающееся значение выборки.
- Квантиль – левее должно располагаться кол-во значений, соответствующее индексу квантили. Например, для квантили $x_{0.8}$

Левее должно располагаться 80% значений выборки.

В нашем примере: мода=0.85, медиана= 0,4451, $x_{0..8}$= 0,8287 – левее должно располагаться 49*0.8=39.2

39 значений выборки.

Интервальные оценки

Интервальная оценка – некоторый интервал [a,b]. По заданной выборке мы должны найти $a(x_1,x_2,\ldots,x_n)$ и $b(x_1,x_2,\ldots,x_n)$ такие, чтобы накрывали неизвестное значение параметра ϑ с заданной вероятностью γ – уровнем значимости. Уровень значимости выбирается в зависимости от необходимой точности решения задачи. Обычно 0.9 – 0.99. Считается 0.9 – средняя точность, 0.99 – высокая, 0.999 – очень высокая.

Часто доверительный интервал строится симметричным относительно точечной оценки.

В дальнейшем будем предполагать, что выборка $\{x_1,x_2,\ldots,x_n\}$ получена из нормально распределенной генеральной совокупности: $x_i \sim N(m,\sigma)$ и при различных условиях требуется найти доверительные интервалы для параметров *m* и σ^2.

Доверительный интервал математического ожидания

Случай 1. Считаем, что известна дисперсия генеральной совокупности σ^2.

$$M\bar{x} = m, \quad D\bar{x} = \frac{\sigma^2}{n}$$

Если все x_i распределены по нормальному закону, то выборочное среднее тоже имеет нормальное распределение и $\bar{x} \sim N(m, \sigma/\sqrt{n})$
После стандартизации

$$U = \frac{\bar{x} - m}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (1)$$

Строим симметричный относительно выборочного среднего интервал

$$P\left(\left|m - \bar{x}\right| < \Delta\right) = \gamma$$

Случ
ай 1.

Мы должны найти такое число Δ, что вероятность попадания разности $m - \bar{x}$ в отрезок (-Δ, Δ) равна заданному числу γ. Разделим обе части неравенства (2) на дисперсию выборочного среднего $\sigma/\sqrt{n}$. В результате получим

$$P\left(\frac{|m-\bar{x}|}{\sigma/\sqrt{n}} < \frac{\Delta}{\sigma/\sqrt{n}}\right) = \gamma$$

Обозначим для краткости $\frac{\Delta}{\sigma/\sqrt{n}} = \delta$

Статистика U (1) должна попадать в интервал (-δ,δ) с вероятностью γ. Вероятность попадания случайной величины в интервал равна

$$\int_{-\delta}^{\delta} \varphi(x)dx = \gamma$$

Вспомнив свойства нормального распределения, получим

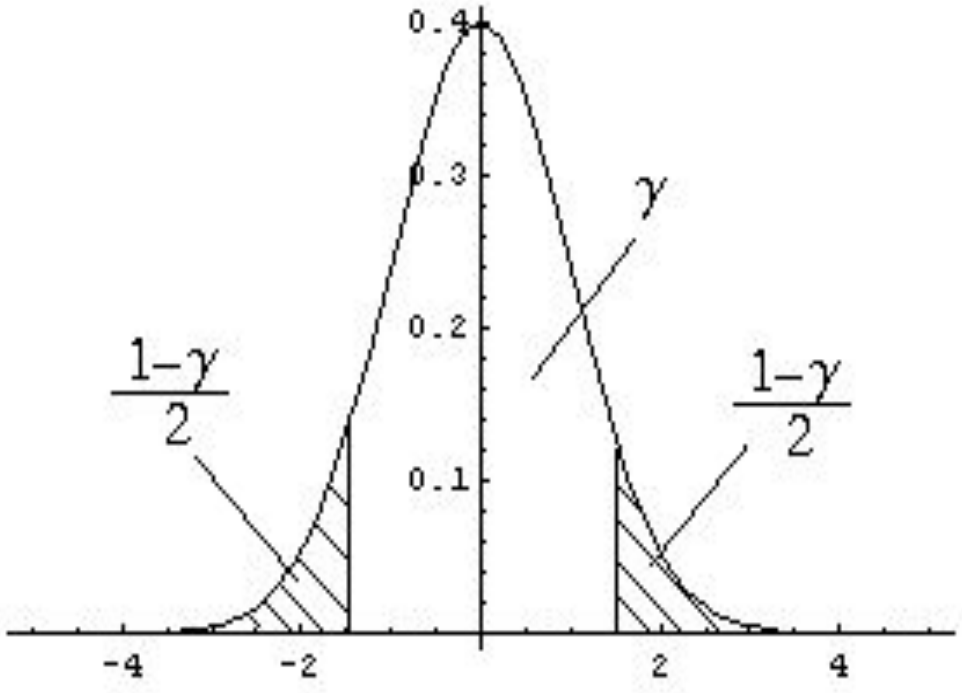
$$F(x) = (1+\gamma)/2 \qquad (3)$$

где $F(x)$ функция нормального распределения

Напомним, корень уравнения (3) называется квантилем распределения с индексом (1+γ)/2. Следовательно, $\delta = u_{(1+\gamma)/2}$ и

$$\Delta = \frac{u_{(1+\gamma)/2}\sigma}{\sqrt{n}} \qquad \bar{x} - \frac{u_{(1+\gamma)/2}\sigma}{\sqrt{n}} < m < \bar{x} + \frac{u_{(1+\gamma)/2}\sigma}{\sqrt{n}}$$

Геометрически. Площадь под графиком плотности распределения равна вероятности попадания в отрезок. Следовательно, нужно построить симметричный отрезок, такой, что площадь над ним равна заданному числу γ. Общая площадь хвостов 1-γ. Площадь одного (1-γ)/2.



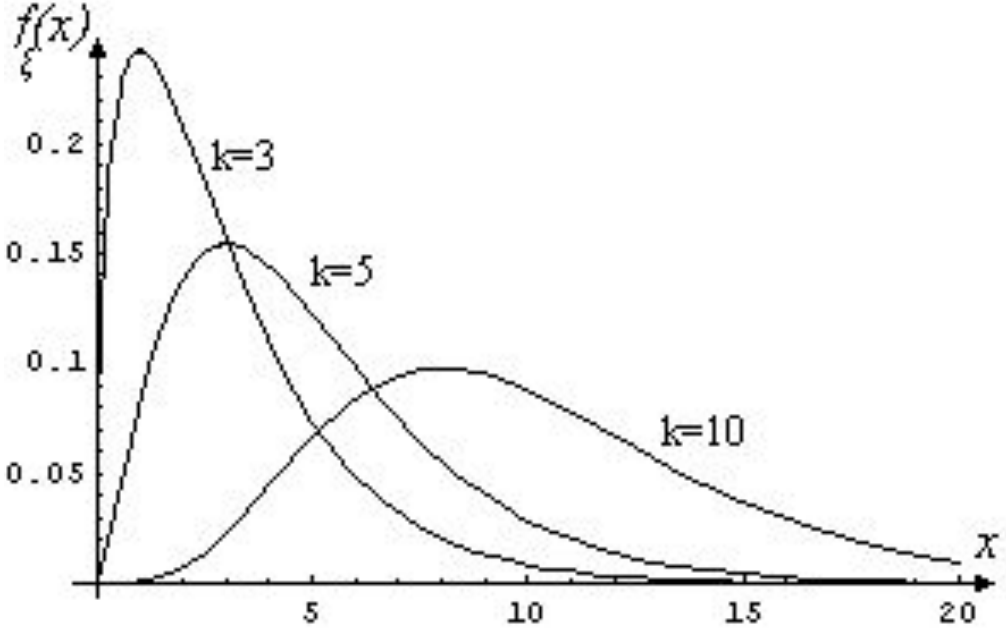
Распределение χ^2

Пусть $\xi_1,\xi_2, \ldots,\xi_k$ независимые случайные величины, распределенные по стандартному нормальному закону

$\xi_1,\xi_2, \ldots,\xi_k \sim N(0,1)$. Говорят, что сумма квадратов этих величин распределена по закону χ^2 с k степенями свободы.

Обозначают $\chi^2 \sim \xi_1,\xi_2, \ldots,\xi_k$. Запись $\xi \sim \chi^2(k)$ означает, что случайная величина ξ распределена по закону $\chi^2(k)$ с k степенями свободы.

На рисунке представлены графики распределения $\chi^2(k)$ с различным числом степеней свободы.



Свойства распределения χ^2 .

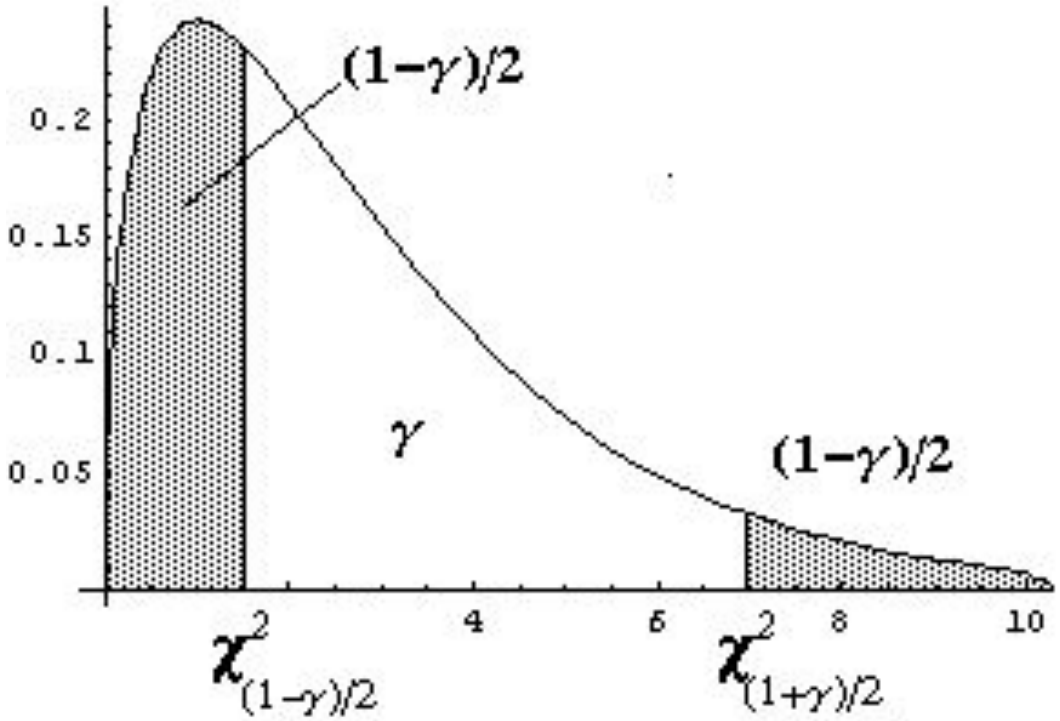
- *Случайная величина имеет нулевую плотность распределения при* $x \leq 0$.
- *При большом числе степеней свободы k распределение* $\chi^2(k)$ *близко к нормальному.*
- *Математическое ожидание случайной величины, распределенной по закону k степенями свободы, равно k:* $M\,\chi^2(k)=k$

Доверительный интервал для дисперсии

Теорема. Случайная величина S^2/σ^2 распределена по закону

$$\frac{S^2}{\sigma^2} \sim \frac{\chi^2(n-1)}{n-1}$$

Отрезок доверительного интервала выберем так, чтобы площади под графиком правее и левее были равны, т.е. равны вероятности попадания справа и слева.



Из рисунка видно, что положение отрезка определяется квантилями $\chi^2_{(1-\gamma)/2}$ и $\chi^2_{(1+\gamma)/2}$. На основании теоремы получим

$$\frac{\chi^2_{(1-\gamma)/2}(n-1)}{n-1} < \frac{S^2}{\sigma^2} < \frac{\chi^2_{(1+\gamma)/2}(n-1)}{n-1}$$

После преобразования неравенства найдем интервал для σ^2

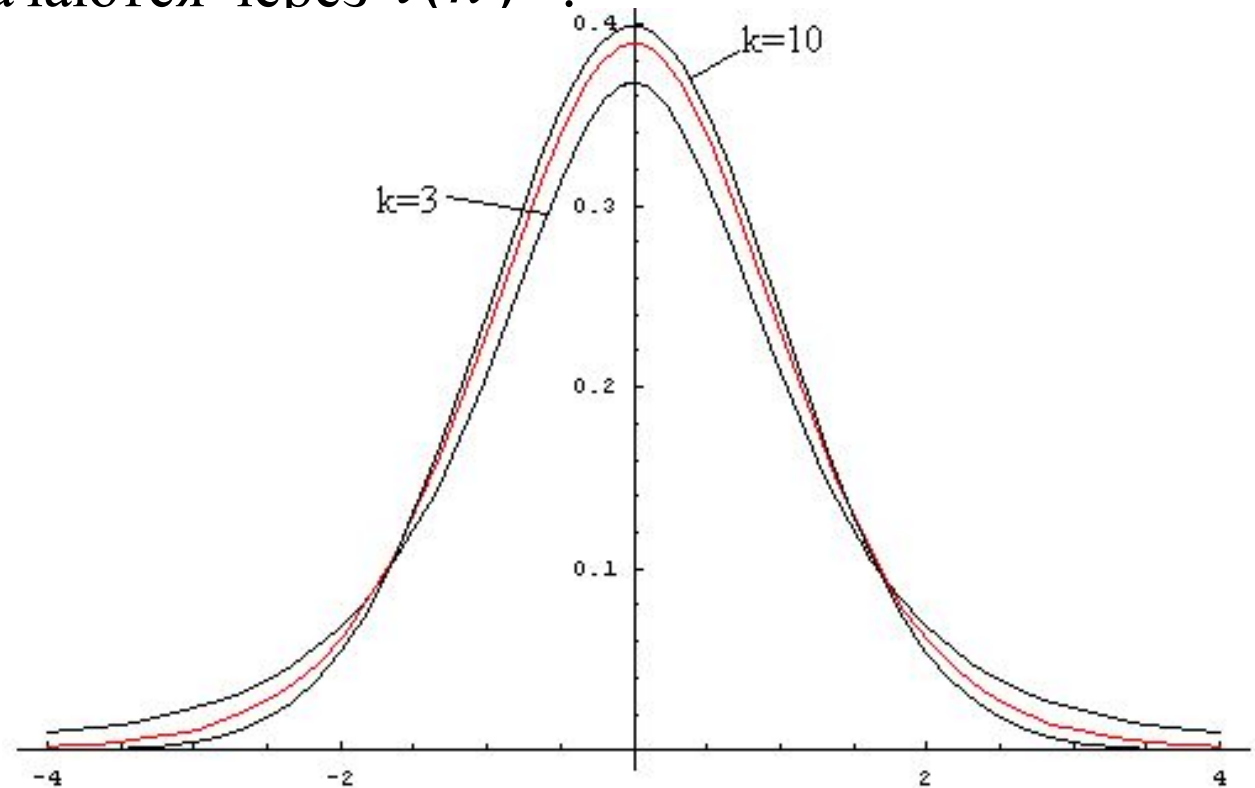
$$\frac{S^2(n-1)}{\chi^2_{(1+\gamma)/2}(n-1)} < \sigma^2 < \frac{S^2(n-1)}{\chi^2_{(1-\gamma)/2}(n-1)}$$

Распределение Стьюдента

Пусть случайная величина ξ распределена по стандартному нормальному закону: $\xi \sim N(0,1)$. Разделим ξ на корень из $\chi^2(k)/k$, то есть из случайной величины, распределенной по закону $\chi^2(k)$ с k степенями свободы, деленной на k). Полученная случайная величина имеет распределение Стьюдента с k степенями свободы. Данная случайная величина и соответствующий закон распределения обозначаются через $t(k)$:

$$t(k) = \frac{\xi}{\sqrt{\chi^2(k)/k}}$$

На рисунке красным выделено нормальное распределение, черным – распределение Стьюдента.



Свойства распределения Стьюдента

- *Распределение Стьюдента симметрично, причем Mt(k) = 0.*
- *При больших k распределение Стьюдента близко к стандартному нормальному распределению N(0,*1).

Доверительный интервал математического ожидания. Случай 2.

Случайная величина U распределена по нормальному закону

$$\frac{\bar{x}-m}{\sigma/\sqrt{n}} \sim N(0,1)$$

Разделим обе части на $\frac{S}{\sigma} \sim \sqrt{\frac{\chi^2(n-1)}{n-1}}$. σ сократится, а в правой части появится распределение Стьюдента t(n-1).

Следовательно, случайная величина

$$\frac{\bar{x}-m}{S/\sqrt{n}} \sim t(n-1)$$

распределена по закону Стьюдента, а доверительный интервал математического ожидания примет вид (τ_α - квантиль распределения Стьюдента, $\alpha = (1+\gamma)/2$)

$$\left(\bar{x} - \tau_\alpha \frac{S}{\sqrt{n}}, \bar{x} + \tau_\alpha \frac{S}{\sqrt{n}} \right)$$

Пример

Вычислим доверительные интервалы для нашей выборки.

Интервал для математического ожидания. Случай 1. Будем считать, что несмещенная оценка дисперсии – точное значение.

Выберем уровень значимости $\gamma = 0.95$. По таблице найдем квантиль стандартного распределения $u_{0.975} = 1.96$. Подставим в формулу

$$u_{(1+\gamma)/2} \frac{\sigma}{\sqrt{n}}$$

m=0.51735, σ=0,288955, n=49. После вычислений получим 0,0809074.

Интервал будет 0.51735- 0,0809074<m< 0.51735+ 0,0809074

0,4364426<m<0,5982574.

Пример. Интервал для дисперсии

S^2=0,083495

$$\frac{S^2(n-1)}{\chi^2_{(1+\gamma)/2}(n-1)} < \sigma^2 < \frac{S^2(n-1)}{\chi^2_{(1-\gamma)/2}(n-1)}$$

Находим квантили распределения $\chi^2_{(1+\gamma)/2}$ и $\chi^2_{(1-\gamma)/2}$.

$\chi^2_{0.975} = 71.4$ $\chi^2_{0.025} = 42.85$

Находим интервал 0,056131<σ^2<0,09353

Интервал для математического ожидания. Случай 2.

Используем распределение Стьюдента. Формула та же, что и раньше, но вместо квантиля нормального распределения используется квантиль распределения Стьюдента. $t(48)_{0.975} = 2.0105$

После вычислений получим

0.51735- 0,082992 <m< 0.51735+ 0,082992

0,434358<m<0,600342

Основы теории проверки статистических гипотез

Статистической гипотезой называется предположение относительно параметров или вида распределения наблюдаемой случайной величины .

Гипотеза называется *простой,* если она однозначно определяет распределение генеральной совокупности. В противном случае гипотеза называется *сложной.*

1. **Гипотезы о параметрах распределения.** Эти гипотезы представляют собой предположение о значении некоторых параметров распределения генеральной совокупности.

2. **Гипотезы о виде распределения.** Эти гипотезы более общего характера выдвигаются в условиях недостаточной информации о генеральной совокупности.

Проверяемая гипотеза называется *нулевой гипотезой и* обычно обозначается H_0. Наряду с H_0 рассматривают *альтернативную* (конкурирующую) гипотезу H_1.

Например: выдвигается гипотеза о значении математического ожидания $H_0 : m=a$

Возможные альтернативные

$H_1: m \neq a$,

$m>a$,

$m<a$,

$m=b$, $b \neq a$

Гипотеза $m=a$, $\sigma^2= b$ – сложная гипотеза.

Можно выдвигать и другие гипотезы.

Общая схема проверки гипотез

Формирование решающего правила опирается на ту же идею, которая используется при построении доверительных интервалов. Ищется случайная величина (так называемая *статистика критерия),* удовлетворяющая двум основным требованиям:

1) ее значение можно посчитать, используя только выборку;

2) ее распределение известно в предположении, что нулевая гипотеза верна.

После того, как такая статистика выбрана, на числовой оси выделяется область, попадание в которую для этой случайной величины маловероятно (*критическая область).* Малая вероятность задается числом α *(уровнем значимости).* Основной принцип проверки гипотез состоит в следующем. *Маловероятное событие считается невозможным. Событие с большой вероятностью считается достоверным.*

Построение решающего правила на основе критерия значимости можно разбить на следующие основные шаги.

1. Сформировать нулевую H_0 и альтернативную H_1 гипотезы.

2. Назначить уровень значимости α .

3. Выбрать статистику Z критерия для проверки гипотезы H_0 .

4. Найти плотность распределения статистики $f_z(x)=f_z(x|H_0)$ критерия **в предположении, что гипотеза H_0 верна.**

5. Определить на числовой оси критическую область V_c из условия $P(Z \in V_c | H_0)= \alpha$ (условная вероятность того, что Z попадает в область V_c, при условии, что гипотеза H_0 верна). Область $R\backslash V_c$ в этом случае называется *областью принятия решения.* Условия, задающие критическую область, называются просто *критерием.*

6. По выборке вычислить выборочное значение Z_s статистики критерия.

7. Принять решение:

• если $Z_s \in V_c$, гипотеза H_0 отклоняется (то есть принимается гипотеза H_1*):*

• если $Z_s \in R \backslash V_c$, гипотеза H_0 не отклоняется.

Принятое решение носит вероятностный, случайный характер. Поэтому обычно применяют более осторожные формулировки. Вместо того чтобы сказать “гипотеза отклоняется, говорят: "данные эксперимента не подтверждают гипотезу *“,* “гипотеза не согласуется с экспериментом”

Значение уровня значимости ***не определяет критическую*** *область однозначно.*

Пример: проверка гипотезы о математическом ожидании

$H_0: \quad m = a$ основная гипотеза

$H_1: \quad m \neq a$ альтернативная гипотеза

Считаем, что дисперсия σ^2 известна. В качестве статистики выбираем величину

$$\frac{a - \bar{x}}{\sigma / \sqrt{n}}$$

Известно, что эта величина распределена по стандартному нормальному закону. Тогда, если гипотеза верна, она должна попадать в интервал

$$\left| \frac{a - \bar{x}}{\sigma / \sqrt{n}} \right| < u_{1-\alpha/2}$$

Но тогда получается, что $\bar{x}$ должно попадать в интервал

$$\left(a - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, a + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Пример 2. Иной вариант альтернативной гипотезы

$H_0: \quad m = a$ основная гипотеза

$H_1: \quad m < a$ альтернативная гипотеза

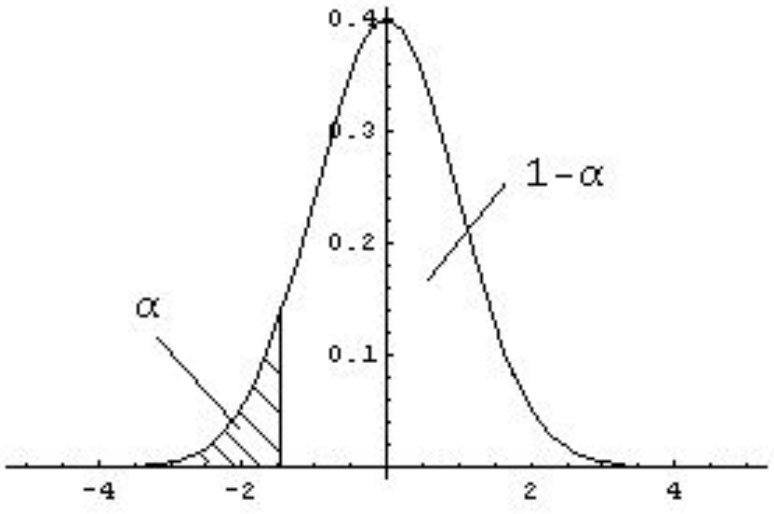
Считаем, что дисперсия σ^2 известна. В качестве статистики выбираем величину

$$\frac{a - \bar{x}}{\sigma / \sqrt{n}}$$

которая имеет нормальное стандартное распределение см. рисунок.
Критическая область находится слева.

$$\bar{x} < a + \frac{\sigma u_{1-\alpha}}{\sqrt{n}}$$

Если дисперсия неизвестна, то используется распределение Стьюдента.



Ошибки при проверке статистических гипотез

Принятие решения на основе статистического критерия носит случайный характер. Возможны следующие ситуации.

1. **Гипотеза верна H_0, и она не отвергается.**
2. **Гипотеза H_0 верна, но она** отвергается. В этом случае говорят, что допущена *ошибка I рода.* Поскольку нулевая гипотеза верна, статистика *Z* действительно имеет то распределение, на основании которого принималось решение. Тем не менее выборочное значение статистики попало в критическую область. Вероятность этого события по определению равна уровню значимости α. *Вероятность ошибки I рода равна уровню значимости критерия.* (это риск производителя)
3. **Гипотеза H_0 неверна, и она отвергается.**
4. **Гипотеза H_0 неверна, но она не отвергается.** Тогда говорят, что допущена *ошибка II рода.* (это риск потребителя)

В этой ситуации выборочное значение попало в область принятия решения, тогда как гипотеза на самом деле неверна. Если распределение статистики Z известно и в предположении, что верна альтернативная гипотеза H_1, то можно посчитать вероятность ошибки II рода: это условная вероятность того, что Z попадает в область $R\backslash V_c$ при условии, что верна гипотеза H_1. Вероятность ошибки II рода обычно обозначают через β

$$\beta = P\left(Z \in R \setminus V_C \mid H_1\right)$$

Для оценки вероятности ошибки второго рода нужно знать функцию распределения в предположении, что альтернативная гипотеза верна.

Проверка гипотезы о функции распределения

Пусть $\{x_1, x_2, \ldots, x_n\}$ - выборка наблюдений некоторой случайной величины ξ.

Гипотеза: H_0 *:* генеральная совокупность имеет функцию распределения *F(x)*

против альтернативы H_1, что функция распределения не такова.

За меру расхождения примем величину . $\delta = \sum_{i=1}^{k} \frac{(n_i - p_i n)^2}{np_i}$

Теорема (Пирсона). *Пусть m параметров функции распределения F(x) оцениваются по выборке. Тогда при* $n \to \infty$ *распределение меры расхождения* δ *стремится к распределению* χ^2 *с k-m-1 степенями свободы*

$$\sum_{i=1}^{k} \frac{(n_i - p_i n)^2}{np_i} \sim \chi^2 (k - m - 1)$$

Понятие о факторном анализе

Пусть результаты наблюдений составляют k независимых выборок (групп), полученных из k нормально распределенных генеральных совокупностей, которые имеют, вообще говоря, различные средние $m_1,m_2,\ldots,m_k$. Каждая группа содержит n_j значений, $j=1,2,\ldots,k$. Общее число наблюдений равно n:

$n_1+n_2+\ldots n_k=n$

Проверяется гипотеза о равенстве средних во всех k выборках:

H_0: $m_1=m_2=\ldots=m_k$

Нулевая гипотеза является сложной: предполагается лишь, что математические ожидания совпадают. Альтернативная гипотеза состоит в том, что хотя бы две выборки имеют различные средние.

Обозначим через x_{ij} i-й элемент j-й выборки, $i=1,2,\ldots,n_j$, $j=1,2,\ldots,k$.

Групповое среднее : $\bar{x}_j$ $\bar{x}_j = \sum_{i=1}^{n_j} x_{ij}$

Общее среднее $\bar{x} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}$

Основное тождество дисперсионного анализа

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2$$

Общая сумма квадратов отклонений от среднего есть сумма квадратов между группами плюс сумма квадратов внутри групп

$$Q_1 = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x}\right)^2 \qquad Q_2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2$$

Пример

Даны две выборки {1.5, 2.5, 2., 1.7, 2.25} и {2., 1.8, 2.2, 2.5, 1.7, 1.6}

Выборочное среднее для первой 1.99, для второй 1.96667.

Значимо ли различие?

Оценки дисперсии S_1=0.163, S_2=0.114667

Генеральное среднее 1.97727, генеральная дисперсия 0.122682