1. Теорема о вероятности суммы событий

Теорема сложения вероятностей

Суммой двух случайных событий A и B называется события A + B состоящие в наступление хотя бы одного из событий A или B.

A $\underset{¿}{или}$ B: 1) только A или 2) только B или 3) A и B

A + B: 1) только A или 2) только B

Теорема сложения для 2-х несовместных событий

Если A и B – несовместны, то вероятность наступления только одного из двух несовместных событий равна сумме вероятностей этих событий

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

Следствие: эта теорема применима для любого конечного числа несовместных событий P(A+B+C) = P(A)+P(B)+P(C)

Теорема сложения для полно группы событий

Пусть события $B_1, B_2, \ldots B_n$ образуют полную группу. Сумма вероятностей событий, образующих полную группу равна 1. $P(B_1)+P(B_2)+\ldots+P(B_n)=1$

Теорема сложения для противоположных событий

P(Ā)+P(A)=1. Сумма вероятностей противоположных событий равна 1.

2. Условные вероятности. Теорема о вероятности произведения событий

Теорема умножения вероятностей

Пусть любое случайное событие называется событие A и B, состоящие в совместном наступление событий A и B

Случайное событие (с.с.) – то, что может произойти или не произойти при осуществление определенной совокупности условий S

Если никаких других ограничений кроме условия S на случайное событие не накладывается, то вероятность этого события называется *безусловной* и обозначается P(A)

Условной вероятностью события B называется вероятность этого события, вычисленную в предположении, что событие A уже произошло и обозначается $P_A(B)$

Событие B^A называется *зависимым* о события A, если вероятность события B изменяется в зависимости от того, происходит ли событие A или нет. Если не изменяется, то событие A и B – независимы

Теорема

Пусть A и B – зависимое с.с.

$P(A*B) = P(A)*P_A(B)$

Вероятность совместного наступления двух зависимых событий равна произведению вероятности одного, на условную вероятность другого, вычисленную предположением, что первое событие уже произошло

Теорема

Пусть A и B – независимое с.с.

$P_A(B)$ = P(B)

Так как вероятность события B не изменяется в зависимости от того, происходит событие A или нет

Теорема

Пусть A и B – независимое с.с.

P(A*B) = P(A)*P(B)

Вероятность совместного наступления всех независимых событий равна произведению вероятностей этих событий

Теорема

Пусть A,B,C…K, L – зависимое с.с

$P(A,B,C\ldots K, L) = P(A)*P_A(B)*P_{AB}(C)\ldots*P_{ABCK}(L)$

Вероятность совместного наступления конечного числа зависимых событий равна произведению условных вероятностей этих событий относительных произведению предшествующих каждому из них

Теорема

Пусть A,B,C…K, L – независимое с.с.

P(A,B,C…K, L) = P(A)*P(B)*P(C)…*P(L)

Вероятность наступления конечного числа независимых событий равна произведению вероятностей этих событий

3. Формула полной вероятности

Пусть событие A может произойти лишь при условии наступления одного из независимых событий $B_1, B_2, \ldots B_n$, которые образуют полную группу. В этом случае вероятность события A можно найти из теоремы

Теорема

$P(A) = P(B_1)*P_{в_1}(A)+P(¿)*P_{в_2}(¿A)+\ldots+P(B_n¿*P_{B_n}(A)$ – формула полной вероятности

Вероятность события A, которое может произойти лишь при условии наступления одного из независимых событий $B_1, B_2, \ldots B_{n'}$ которые образуют полную группу, равна сумме произведений вероятности этих событий на соответствующую условию вероятность события A

4. Формула Байеса

P(A) – вероятность события A, которое может наступить лишь при условии появления одного из несовместных событий $B_1, B_2, \ldots B_{n'}$ которые образуют полную группу.

В связи с тем, что не известно, которое из событий $B_1, B_2, \ldots B_n$ произойдет, эти события называются предположениями или гипотезами. Выясним, как изменится вероятность каждой из гипотез в связи с наступающим событием A, т.е. вычислим условие вероятности

$P_A(B_1)P_A(B_2)\ldots P_A(B_n)$

Найдем вероятность совместного наступления событий A и B_1.

Используем теорему умножения для 2-х зависимых событий

$P(A*B_1) = P(A)*P_A(B_1)$

$P(B_1*A) = P(B_1)*P_{B_1}(A)$

Т.к. в левой части обоих формул находятся вероятность одного и того же события, левые части равны, равны и правые

$P(A)*P_A(B_1) = P(B_1)*P_{B_1}(A)$

Аналогично можно получить формулы для условных вероятностей остальных гипотез

$$P_A(B_1)=\frac{P(B_1)*P_{B_1}(A)}{P(A)}$$

$$P_A(B_2)=\frac{P(B_2)*P_{B_2}(A)}{P(A)}$$

$$P_A(B_n)=\frac{P(B_n)*P_{B_n}(A)}{P(A)}$$

Эти формулы называются формулой Байеса в которых вероятность A в значении находится по формуле полной вероятности: P(A) = $P(B_1)*P_{B_1}(A)$ + $P(B_2)*P_{B_2}(A)+\ldots+P(B_n)*P_{B_n}(A)$

5. Последовательность независимых испытаний

Испытание называется независимым относительно события A, если вероятность появления этого события в каждом испытании не зависит от резервов в других испытаниях, где A – событие, появление которого интересует нас в каждом испытании.

Рассмотрим случай независимых испытаний, вероятность появления события A в каждом из которых есть величина постоянная

P(A) = p P(A) = q

Из теоремы сложения вероятности противоположных событий P(A)+P(Ā) = 1 следует, что P(Ā) = q = 1 – p

Вероятность отклонения относительной частоты от постоянной вероятности в независимых испытаниях

1. Если испытание независимо
2. Вероятность наступления события A – постоянна (в каждом испытании), то вероятность отклонения относительной частоты $\left(\frac{m}{n}\right)$ от постоянной вероятности (p) вычисляется по формуле

$$P\left(\left|\frac{m}{n}-p\right|\leq\mathcal{E}\right)=2\Phi¿$$

6. Формула Бернулли

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

$P_n(K)$ – вероятность того, что в n независимых испытаниях событие A произойдет равно K раз

В общем случае можно утверждать, что вероятность наступления события A в n независимых испытаниях:

1) не менее K раз:

$$P_n(K)+P_n(K+1)+P_n(K+2)+\ldots P_n(n)$$

2) не более K раз

$$P_n(0)+P_n(1)+\ldots+P_n(K)$$

3) более K раз

$$P_n(K+1)+P_n(K+2)+\ldots P_n(n)$$

4) менее K раз

$$P_n(0)+P_n(1)+\ldots+P_n(K-1)$$

7. Предельные теоремы Муавра – Лапласа

Локальная теорема Лапласа

$$P_{1000}(200)=C_{1000}^{200}p^{200}q^{800}$$

Если число испытаний n велико, то вычисляется вероятность $P_n(K)$ по формуле Бернулли довольно трудоемко. В этом случае можно использовать локальную теорему Лапласа: если вероятность p появления события A в каждом независимом постоянна и не равна 0 и 1,

то вероятность $P_n(K)$ того, что события А произойдет в n независимых испытаниях равно К раз приближенным (тем точнее, чем больше n)

$$P_n(K) \approx \frac{1}{\sqrt{npq}} * \varphi(x)$$

$\varphi(x)$- функция вероятности $\frac{1}{\sqrt{2\pi}} * e^{\frac{-x^2}{2}}$ эта функция находится по таб. $x = \frac{K - np}{\sqrt{npq}}$

Чтобы правильно пользоваться таблицей рассмотрим некоторые свойства $\varphi(x)$:

1. $\varphi(x)$- четная функция $\varphi(-x) = \varphi(x)$
2. $\varphi(x)$ – монотонно убывающая функция

$\varphi(5) \approx 0{,}0000015$ для x>5 можно считать, что $\varphi(x) = 0$

Интегральная теорема Лапласа

В предыдущих темах была решена задача о нахождение вероятности того, что в n независимых испытаниях событие А произойдет ровно К раз. Но часто необходимо знать вероятность наступления событий неопределённое число раз , а число раз заключается в некотором интервале

$$P_{1000}(300;500) = P_{1000}(300) + P_{1000}(301) + \ldots + P_{10}$$

это решение довольно трудоемко и ответ на поставленный вопрос можно получить сразу с помощью интегральной теоремы Лапласа: вероятность того, что в n независимых испытаниях в каждом из которых вероятность появления события наступит не менее K_1 и не более K_2 раз, при p равном (0<p<1)

$$P_n(K_1 \le K \le K_2) = P_n(K_1; K_2) \approx \Phi(x'') - \Phi(x')$$

$\Phi(x)$ – функция Лапласа находящаяся по таблице

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

$$x'' = \frac{K_2 - np}{\sqrt{npq}} \quad x' = \frac{K_1 - np}{\sqrt{npq}}$$

Чтобы правильно пользоваться таблицей рассмотрим некоторые свойства функции Лапласа:

1. $\Phi(x)$ – нечетная, т.е. $\Phi(-x) = -\Phi(x)$
2. $\Phi(x)$ является монотонно возрастающей

$\Phi(x) \approx 0{,}499997$, поэтому для x>5 можно считать, что $\Phi(x) \approx 0{,}5$

8. Случайные величины

Переменные величины которые принимают различные значения в зависимости о случая, называются случайные величины

Обозначаются: заглавными латинским буквами X;Y;Z..

значение, которое принимают случайные величины в результате испытания, называют ее возможные значения

X – число очков, выпавших при подбрасывание игральной кости: x_1=1, x_2=2,x_3=3,x_4=4,x_5=5,x_6=6

Случайные величины подразделяются на 2 вида: дискретные и непрерывные

Дискретной называют случайную величину, возможное значение которой образует дискретный ряд чисел. Число этих значений может быть конечным и бесконечным.

Непрерывной называют случайную величину, возможное значение которой полностью заполняет некоторый промежуток (конечный или бесконечный) . Число всегда бесконечно

9. Закон распределения дискретной случайной величины

Для задания дискретной случайной величины недостаточно перечислить все ее возможные значения, нужно указать еще и их вероятность.

Законом распределения дискретной случайной величины называют соответствие между возможными значениями случайной величины и вероятностями их появления.

Закон распределения можно задать таблично, аналитически (в виде формулы) или графически (в виде многоугольника распределения).

Рассмотрим случайную величину X, которая принимает значения x1, x2, x3 ... xn с некоторой вероятностью pi, где i = 1.. n. Сумма вероятностей pi равна 1.

Таблица соответствия значений случайной величины и их вероятностей вида

x1 x2 x3 ... xn ... p1 p2 p3 pn

называется рядом распределения дискретной случайной величины или просто рядом распределения. Эта таблица является наиболее удобной формой задания дискретной случайной величины.

10. Числовые характеристики дискретных случайных величин

Закон распределения полностью характеризует дискретную случайную величину. Однако, когда невозможно определить закон распределения, или этого не требуется, можно ограничиться нахождением значений, называемых числовыми характеристиками случайной величины: Математическое ожидание, Дисперсия, Среднее квадратичное отклонение

Эти величины определяют некоторое среднее значение, вокруг которого группируются значения случайной величины, и степень их разбросанности вокруг этого среднего значения.

Математическое ожидание M дискретной случайной величины - это среднее значение случайной величины, равное сумме произведений всех возможных значений случайной величины на их вероятности.

11. Функция распределения вероятностей случайной величины

Определение функции распределения

Вспомним, что дискретная случайная величина может быть задана перечнем всех ее возможных значений и их вероятностей. Такой способ задания не является общим: он неприменим, например, для непрерывных случайных величин.

Действительно, рассмотрим случайную величину , возможные значения которой сплошь заполняют интервал . Можно ли составить перечень всех возможных значений ? Очевидно, что этого сделать нельзя. Этот пример указывает на целесообразность дать общий способ задания любых типов случайных величин. С этой целью и вводят функции распределения вероятностей случайной величины.

Пусть — действительное число. Вероятность события, состоящего в том, что примет значение, меньшее , т.е. вероятность события , обозначим через . Разумеется, если изменяется, то, вообще говоря, изменяется и , т.е. — функция от .Функцией распределения называют функцию , определяющую вероятность того, что случайная величина в результате испытания примет значение, меньшее , т.е.

Геометрически это равенство можно истолковать так: есть вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки .Иногда вместо термина «функция распределения» используют термин «интегральная функция».

Теперь можно дать более точное определение непрерывной случайной величины: случайную величину называют непрерывной, если ее функция распределения есть непрерывная, кусочно-дифференцируемая функция с непрерывной производной.

12. Плотность распределения вероятностей НСВ. Вероятность попадания НСВ. Свойства плотности распределения. Числовые характеристики НСВ.

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

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

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

f (x) = F′(x),

то есть является производной функции распределения.

Свойства плотности распределения.

1) f(x) ≥ 0, так как функция распределения является неубывающей.

2) , что следует из определения плотности распределения.

3) Вероятность попадания случайной величины в интервал (a, b) определяется формулой Действительно,

4) (условие нормировки). Его справедливость следует из того, что a

5) так как при

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

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

13. Числовые характеристики непрерывных случайных величин

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

Закон распределения (функция распределения и ряд распределения или плотность веро-ятности) полностью описывают поведение случайной величины. Но в ряде задач доста-точно знать некоторые числовые

характеристики исследуемой величины (например, ее среднее значение и возможное отклонение от него), чтобы ответить на поставленный во-прос. Рассмотрим основные числовые характеристики дискретных случайных величин. Математическое ожидание.
Определение 7.1. Математическим ожиданием дискретной случайной величины называ-ется сумма произведений ее возможных значений на соответствующие им вероятности: M(X) = x1p1 + x2p2 + ... + xnpn .
(7.1)Если число возможных значений случайной величины бесконечно, то , если полученный ряд сходится абсолютно.
Замечание 1. Математическое ожидание называют иногда взвешенным средним, так как оно приближенно равно среднему арифметическому наблюдаемых значений случайной величины при большом числе опытов.
Замечание 2. Из определения математического ожидания следует, что его значение не меньше наименьшего возможного значения случайной величины и не больше наибольше-го.
Замечание 3. Математическое ожидание дискретной случайной величины есть неслучай-ная (постоянная) величина. В дальнейшем увидим, что это же справедливо и для непре-рывных случайных величин.

14. Нормальное распределение.

Нормальный закон распр-я н.с.в. – закон, который хар-ся след.пл-тью распр-я .

$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-m)^2}{2\sigma^2}}$ → норм-ый закон опред-ся двумя параметрами **а** и (**жигма**)

$M[X]=\int_{-\infty}^{\infty} xf(x)dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{\frac{-(x-m)^2}{2\sigma^2}}dx$= a

$D[X]=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-m)^2 e^{\frac{-(x-m)^2}{2\sigma^2}}dx$ = Q_2

Q= D(X) под корнем = Q (X) , → параметр **а =** мат. ожид-ю , а пар. Q = среднему квадратич. О откл- ю нормально распр-ой с.в.х.

15. Генеральная совокупность и выборка.

Для исследования этого признака применяют метод сплошных наблюдений , при котором исследуют каждый объект данной соб-ти относительно изучаемого признака.
Осн-ые причины: 1) число объектов велико 2) исслед-е физич-и велико 3) исслед-е связано с большими затратами 4) исслед-е связано с ун. объектов. Если этот метод не используют, то применяют выборку.
Осн-ые способы отбора: 1) простой, случайный, повторный 2) простой, бесповторный 3) механический 4) серийный
Статистическое распределение выборки: 1) наблюдаемые значения количеств-го признака X1, X2.... X2 (варианты) 2) число наблюдений N1,N2....NK(частота этих вариантов) 3) отношение частоты к выборке(относительная частота). Выборка- число объектов выборной или ген. сов-сти . 4) Варианты расположены в порядке возрастания и образуют вариационный ряд. Хи..х1 х2 х3..хн ; ни.. н1 н2 н3..нк; ви..в1 в2 в3 ..вк. Сумма всех частот = в выборке; сумма относительных частот = 1.

16. Вариационный ряд.

Вариационный ряд - последовательность всех элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются.
По этому ряду уже можно сделать несколько выводов. Например, средний элемент вариационного ряда (***медиана***) может быть оценкой наиболее вероятного результата измерения. Первый и последний элемент вариационного ряда (т.е. минимальный и максимальный элемент выборки) показывают **разброс элементов выборки**. Иногда если первый или последний элемент сильно отличаются от остальных элементов выборки, то их исключают из результатов измерений, считая, что эти значения получены в результате какого-то грубого сбоя, например, техники.

17. Графическое изображение вариационных рядов, полигон и гистограмма.

Графическое изображение вариац. рядов: 1. Полигонная частота- линия отрезка, которой соединяют точки с координатами (х1,н1) (х2,н2) (хн,нк). Точки соединяются с координатами (х1,в1) (х2,в2) (хн,вн).
2. Для непрерывного распр-я колличеств-ти признака X, используют гистограмму частот или относит. частот. Для гистограммы относит. частот высоты прямоугол = ви : альфа.

18. Эмпирическая ф-ция распределения.

m_x- число наблюдений, при которых наблюдалось значение признака, меньшее ***x;*** *n*- общее число наблюдений (объем выборки). Ясно, что относительная частота события X < *x* равна. m_x/n. Если *x*изменяется, то изменяется и относительная частота, т. е. относительная частота есть функция от *x.* Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической. *Эмпирической функцией распределения (функцией распределения выборки) называют функцию определяющую для каждого значения x относительную частоту события X < x, т.е.*

$\widehat{F}(x)=\frac{m_x}{n}$ Из теоремы Бернулли следует, что относительная частота события X < *x,* т. е. эмпирическая функция стремится по вероятности к вероятности *F (x)* этого события. Отсюда следует целесообразность использования эмпирической функции распределения выборки для приближенного представления теоретической (интегральной) функции распределения генеральной совокупности.
Эмпирическая функция обладает всеми свойствами *F(x):*
1) ее значения принадлежат отрезку [0, 1]; 2) неубывающая; 3) если x_i -наименьшая варианта, то

$\widehat{F}(x)=0 \; при \; x\le x_1$, если x_k- наибольшая варианта, то

$\widehat{F}(x)$=1 при $x>x_k$

Итак, эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

19. Выборочная средняя, её свойства.

Вы́борочное (эмпири́ческое) сре́днее — это приближение теоретического среднего распределения, основанное на выборке из него.

Определение: Пусть $X_1, \dots, X_n$ — выборка из распределения вероятности, определённая на некотором вероятностном пространстве (Ω,F,P). Тогда её выборочным средним называется случайная величина. $\overline{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$

Свойства выборочного среднего :

Пусть $\widehat{F}(x)$ — выборочная функция распределения данной выборки.

Тогда для любого фиксированного ω $\in \Omega$ функция $\widehat{F}(\omega, x)$ является (неслучайной) функцией дискретного распределения.

Тогда математическое ожидание этого распределения равно $\overline{X}(\omega)$

Выборочное среднее — несмещённая оценка теоретического среднего:

$E[\overline{X}]=E[X_i], i=1, \dots, n$

Выборочное среднее — сильно состоятельная оценка теоретического среднего:

$\overline{X}\to E[X_i]$ почти наверное при $n\to\infty$.

Выборочное среднее — асимптотически нормальная оценка.

Пусть дисперсия случайных величин X_i конечна и ненулевая, то

есть $D[X_i]=\sigma^2<\infty, \sigma^2\neq 0, i=1, \dots, n$

Тогда $\sqrt{n}$¿ по распределению при $n\to\infty$,

где $N(0,\sigma^2)$ — нормальное распределение со средним 0 и дисперсией σ^2.

Выборочное среднее из нормальной выборки — эффективная оценка её среднего

20. Выборочная дисперсия, её свойства.

Выборочная дисперсия в математической статистике — это оценка теоретической дисперсии распределения на основе выборки. Различают выборочную дисперсию и несмещённую, или исправленную, выборочные дисперсии.
Определения
Пусть $X_1, \dots, X_n, \dots$ — выборка из распределения вероятности. Тогда
Выборочная дисперсия — это случайная величина

$S_n^2\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-$¿¿,

где символ $\bar{X}$ обозначает выборочное среднее.
Несмещённая (исправленная) дисперсия — это случайная величина

$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$

Замечание
Очевидно,

$S^2=\frac{1}{n-1}S_n^2.$

Свойства выборочных дисперсий
Выборочная дисперсия является
теоретической дисперсией выборочного распределения. Более точно,
пусть $\widehat{F}(x)$ — выборочная функция распределения данной выборки.

Тогда для любого фиксированного ω $\in \Omega$ функция $\widehat{F}(\omega, x)$ является (неслучайной) функцией дискретного распределения.

Дисперсия этого распределения равна $S_n^2(\omega)$.

Обе выборочные дисперсии являются состоятельными оценками теоретической дисперсии. Если ,

$D[X_i] = \sigma^2 < \infty$, *то есть*

$S_n^2 \xrightarrow{P} \sigma^2$ и $S^2 \xrightarrow{P} \sigma^2$,

где $\xrightarrow{P}$ обозначает сходимость по вероятности.

Выборочная дисперсия является смещённой оценкой теоретической дисперсии, а исправленная выборочная дисперсия несмещённой:

$$E\left[S_n^2\right] = \frac{n-1}{n}\sigma^2, \text{и} E\left[S^2\right] = \sigma^2$$

Выборочная дисперсия нормального распределения имеет распределение хи-квадрат.

Пусть $X_i \sim \mathrm{N}(\mu, \sigma^2),\ i = 1, 2, \ldots$. Тогда

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

21. Статистические оценки: несмещенные, эффективные, состоятельные

Состоятельной называют такую точечную статистическую оценку, которая при n стрем к бесконечн стремится по вероятности к оцениваемому параметру. В частности, если дисперсия несмещенной оценки при n стр к беск стремится к нулю, то такая оценка оказывается и состоятельной.

Рассмотрим оценку θ_n числового параметра θ, определенную при *n* = 1, 2, ... Оценка θ_nназывается *состоятельной*, если она сходится по вероятности к значению оцениваемого параметра θ при безграничном возрастании объема выборки. Выразим сказанное более подробно.

Статистика θ_n является состоятельной оценкой параметра θ тогда и только тогда, когда для любого положительного числа ε справедливо предельное соотношение

$\lim\limits_{n\to\infty} P\{|\theta_n - \theta| > \varepsilon\} = 0.$

Пример 3. Из закона больших чисел следует, что $\theta_n = \overline{X}$ является состоятельной оценкой θ = *M(X)* (в приведенной выше теореме Чебышёва предполагалось существование дисперсии *D(X);* однако, как доказал А.Я. Хинчин [6], достаточно выполнения более слабого условия – существования математического ожидания *M(X)*).

Пример 4. Все указанные выше оценки параметров нормального распределения являются состоятельными.

Вообще, все (за редчайшими исключениями) оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются состоятельными.

Пример 5. Так, согласно теореме В.И. Гливенко, эмпирическая функция распределения$F_n(x)$ является состоятельной оценкой функции распределения результатов наблюдений*F(x)*

*Несмещенной называют такую точечную статистическую оценку Q*математическое ожидание которой равно оцениваемому параметру: M(Q*)=Q*

Второе важное свойство оценок – *несмещенность*. Несмещенная оценка θ_n – это оценка параметра θ, математическое ожидание которой равно значению оцениваемого параметра: $M(\theta_n) = \theta$.

Пример 6. Из приведенных выше результатов следует, что $\overline{X}$ и S_0^2 являются несмещенными оценками параметров *m* и σ^2 нормального распределения. Поскольку M($\widetilde{X}$) = M(*m***) = *m*, то выборочная медиана $\widetilde{X}$ и полусумма крайних членов вариационного ряда *m*** - также несмещенные оценки математического ожидания *m*нормального распределения. Однако

$M(s^2) \neq \sigma^2, \quad M[(\sigma^2)^{**}] \neq \sigma^2,$

поэтому оценки s^2 и (σ^2)** не являются состоятельными оценками дисперсии σ^2нормального распределения.

Оценки, для которых соотношение $M(\theta_n) = \theta$ неверно, называются смещенными. При этом разность между математическим ожиданием оценки θ_n и оцениваемым параметром θ, т.е. $M(\theta_n) - \theta$, называется смещением оценки.

Пример 7. Для оценки s^2, как следует из сказанного выше, смещение равно

$M(s^2) - \sigma^2 = -\sigma^2/n$.

Смещение оценки s^2 стремится к 0 при $n \to \infty$.

Оценка, для которой смещение стремится к 0, когда объем выборки стремится к бесконечности, называется *асимптотически несмещенной*. В примере 7 показано, что оценка s^2 является асимптотически несмещенной.

Практически все оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются либо несмещенными, либо асимптотически несмещенными. Для несмещенных оценок показателем точности оценки служит дисперсия – чем дисперсия меньше, тем оценка лучше. Для смещенных оценок показателем точности служит математическое ожидание квадрата оценки $M(\theta_n - \theta)^2$. Как следует из основных свойств математического ожидания и дисперсии,

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] = D(\theta_n) + (M(\theta_n) - \theta)^2, \quad (3)$$

т.е. математическое ожидание квадрата ошибки складывается из дисперсии оценки и квадрата ее смещения.

Для подавляющего большинства оценок параметров, используемых в вероятностно-статистических методах принятия решений, дисперсия имеет порядок 1/*n*, а смещение – не более чем 1/*n*, где *n* – объем выборки. Для таких оценок при больших *n* второе слагаемое в правой части (3) пренебрежимо мало по сравнению с первым, и для них справедливо приближенное равенство

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] \approx D(\theta_n) \approx \frac{c}{n}, \quad c = c(\theta_n, \theta). \quad (4)$$

где *c* – число, определяемое методом вычисления оценок θ_n и истинным значением оцениваемого параметра θ.

Эффективной называют такую точечную статистическую оценку, которая при фиксированном n имеет наименьшую дисперсию.

С дисперсией оценки связано третье важное свойство метода оценивания –*эффективность*. Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра.

Доказано [11], что $\overline{X}$ и S_0^2 являются эффективными оценками параметров *m* и σ^2нормального распределения. В то же время для выборочной медианы $\widetilde{X}$ справедливо предельное соотношение

$\lim\limits_{n\to\infty} \frac{D(\widetilde{x})}{D(\widetilde{x})} = \frac{2}{\pi} \approx 0{,}637.$

Другими словами, эффективность выборочной медианы, т.е. отношение дисперсии эффективной оценки $\overline{X}$ параметра *m* к дисперсии несмещенной оценки $\widetilde{X}$ этого параметра при больших n близка к 0,637. Именно из-за сравнительно низкой эффективности выборочной медианы в качестве оценки математического ожидания нормального распределения обычно используют выборочное среднее арифметическое.

Понятие эффективности вводится для несмещенных оценок, для которых $M(\theta_n) = \theta$ для всех возможных значений параметра θ. Если не требовать несмещенности, то можно указать оценки, при некоторых θ имеющие меньшую дисперсию и средний квадрат ошибки, чем эффективные.

Пример 8. Рассмотрим «оценку» математического ожидания $m_1 \equiv 0$. Тогда $D(m_1) = 0$, т.е. всегда меньше дисперсии $D(\overline{X})$ эффективной оценки $\overline{X}$. Математическое ожидание среднего квадрата ошибки $d_n(m_1) = m^2$, т.е. при $|m| < \sigma/\sqrt{n}$ имеем $d_n(m_1) < d_n(\overline{X})$. Ясно, однако, что статистику $m_1 \equiv 0$ бессмысленно рассматривать в качестве оценки математического ожидания *m*.

Пример 9. Более интересный пример рассмотрен американским математиком Дж. Ходжесом:

$$T_n = \begin{cases} \bar{x}, & |\bar{x}| > n^{-1/4}, \\ 0{,}5\bar{x}, & |\bar{x}| \le n^{-1/4}. \end{cases}$$

Ясно, что T_n – состоятельная, асимптотически несмещенная оценка математического ожидания *m*, при этом, как нетрудно вычислить,

$$\lim_{n\to\infty} n d_n(T_n) = \begin{cases} \sigma^2, & m \neq 0, \\ \dfrac{\sigma^2}{4}, & m = 0. \end{cases}$$

Последняя формула показывает, что при $m \neq 0$ оценка T_n не хуже $\overline{X}$ (при сравнении по среднему квадрату ошибки d_n), а при *m* = 0 – в четыре раза лучше.

Подавляющее большинство оценок θ_n, используемых в вероятностно-статистических методах, являются асимптотически нормальными, т.е. для них справедливы предельные соотношения:

$$\lim_{n\to\infty} P\left\{\frac{\theta_n - M(\theta_n)}{\sqrt{D(\theta_n)}} < x\right\} = \Phi(x)$$

для любого *x*, где *Φ(x)* – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Это означает, что для больших объемов выборок (практически - несколько десятков или сотен наблюдений) распределения оценок полностью описываются их математическими ожиданиями и дисперсиями, а качество оценок – значениями средних квадратов ошибок $d_n(\theta_n)$.

22. Точечные и интервальные оценки.

Точечной оценкой неизвестного параметра называют число (точку на числовой оси), которое приблизительно равно оцениваемому параметру и может заменить его с достаточной степенью точности в статистических расчетах.

Для того чтобы точечные статистические оценки обеспечивали “хорошие” приближения неизвестных параметров, они должны быть несмещенными, состоятельными и эффективными.

Определение :Пусть $X_1, \ldots, X_n, \ldots$ — случайная выборка из распределения, зависящего от параметра $\theta \in \Theta$. Тогда статистику $\hat{\theta}(X_1, \ldots, X_n)$,

принимающую значения в Θ, называют точечной оценкой параметра θЗамечание

Формально статистика $\hat{\theta}$ может не иметь ничего общего с интересующим нас значением параметра θ. Её полезность для получения практически приемлемых оценок вытекает из дополнительных свойств, которыми она обладает или не обладает.
Свойства точечных оценок

Оценка $\hat{\theta} = \hat{\theta}(X)$ называется несмещённой, если её математическое ожидание равно оцениваемому параметру генеральной совокупности:
$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta, \quad \forall\theta \in \Theta$,

где $\mathbb{E}_\theta$ обозначает математическое ожидание в предположении, что θ — истинное значение параметра (распределения выборки X).

Оценка $\hat{\theta}$ называется эффективной, если она обладает минимальной дисперсией среди всех возможных несмещенных точечных оценок.

Оценка $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ называется состоятельной, если она по вероятности с увеличением объема выборки n стремится к параметру генеральной совокупности: $\forall\theta \in \Theta$,
$\hat{\theta}_n \to \theta$ по вероятности при $n \to \infty$.

Оценка $\hat{\theta}_n$ называется сильно состоятельной, если $\forall\theta \in \Theta$,
$\hat{\theta}_n \to \theta$ почти наверное при $n \to \infty$.

Надо отметить, что проверить на опыте сходимость «почти наверное» не представляется возможным, поэтому с точки зрения прикладной статистики имеет смысл говорить только о сходимости по вероятности.
Интервальной называют оценку, которая определяется двумя числами – концами отрезка.
Интервальные оценки – характеризуют не единственно возможную ситуацию, а их множественность. Этот вид экспертных оценок широко распространен. Одним из определяющих свойств интервальной оценки является то, что на множестве задано бинарное отношение МЕЖДУ.
Определение
Пусть ϑ - неизвестный параметр генеральной совокупности. По сделанной выборке по определенным правилам находятся числа θ_1 и θ_2, такие чтобы выполнялось неравенство:
$P\{\theta_1 < \theta < \theta_2\} = 1-\alpha$.

Интервал (θ_1, θ_2) является доверительным интервалом для параметра ϑ, а число $1-\alpha$ - *доверительной вероятностью* или *надежностью* сделанной оценки. Обычно надежность задается заранее, причем выбираются числа близкие к 1 (0.95, 0.99 или 0.999).
Примеры интервальных оценок
Пример 1. Доверительное оценивание по вариационному ряду.
Пусть задана выборка $X^n = (x_1, \cdots, x_n)$ некоторой случайной величины X. Построим вариационный ряд выборки $x^{(1)} < \cdots < x^{(n)}$:

Очевидно, что вероятность попасть в любой из $(n+1)$- го интервалов значений случайной величины X одинакова и равна $\frac{1}{n+1}$. Тогда вероятность того, что случайная величина X приняла значение из интервала $(x^{(k)}, x^{(l)})$, где $l > k$ будет равна:
$P_{X^n, x}\left\{x \in (x^{(k)}, x^{(l)})\right\} = \frac{l-k}{n+1}$.

Вопрос: чему должен быть равен размер выборки n чтобы вероятность попасть в интервал $(\min_i(x_i), \max_i(x_i))$ составила 95%.
Подставляя значение для доверительной вероятности в формулу выше, получим:
$0.95 = P_{X^n, x}\left\{x \in (x^{(1)}, x^{(n)})\right\} = \frac{n-1}{n+1}$,

откуда $n = 39$.
Таким образом, при достаточном для заданной доверительной вероятности числе измерений случайной величины X по набору ее порядковых статистикможет быть оценен диапазон принимаемых ею значений.
Пример 2. Доверительный интервал для медианы.
Пусть задана выборка $X^n = (x_1, \cdots, x_n)$ некоторой случайной величины X
При $n > 50$ доверительный интервал для медианы $\tilde{x}$ определяется порядковыми статистиками
$x_k \le \tilde{x} \le x_{n-k+1}$,

где
$k = \frac{1}{2}(n - 1.64\sqrt{n} - 1)$ при $\alpha = 0.1$;
$k = \frac{1}{2}(n - 1.96\sqrt{n} - 1)$ при $\alpha = 0.05$;
$k = \frac{1}{2}(n - 2.58\sqrt{n} - 1)$ при $\alpha = 0.01$.

Для значений $n \le 50$ номера порядковых статистик, заключающих в себе медиану, при $\alpha = 0.05$ и $\alpha = 0.01$ приведены в таблице 1, взятой из [3].
Пример 3. Доверительный интервал для математического ожидания.
Пусть задана выборка $X^n = (x_1, \cdots, x_n)$ некоторой случайной величины X, арактеристики которой (дисперсия D и математическое ожидание M) неизвестны. Эти параметры оценим так:
$M^* = \frac{\sum_{i=1}^{n} x_i}{n}$
$D^* = \frac{\sum_{i=1}^{n} (x_i - M^*)^2}{n-1}$ - несмещенная оценка дисперсии.
Величину $\sqrt{D^*}$ называют оценкой среднего квадратического отклонения. Воспользуемся тем, что величина M^* представляет собой сумму n независимых случайных величин, и, согласно центральной предельной теореме, при достаточно большом n ее закон близок к нормальному. Поэтому будем считать, что величина M^* распределена по нормальному закону. Характеристики этого закона - математическое ожидание и дисперсия - равны соответственно M (настоящее МО случайной величины X) и $\frac{D}{n}$.
Найдем такую величину δ, для которой $P(|M^* - M| < \delta) = \alpha$.
Перепишем это в эквивалентном виде $P(\frac{|M^* - M|}{\sqrt{D/n}} < \frac{\delta}{\sqrt{D/n}}) = \alpha$ и скажем, что случайная величина перед знаком неравенства есть модуль от стандартной нормальной. Получаем, что $2\Phi(\frac{\delta}{\sqrt{D/n}}) - 1 = \alpha$,
и $\delta = \sqrt{\frac{D}{n}} * \Phi^{-1}\left(\frac{\alpha+1}{2}\right)$. В случае неизвестной дисперсии ее можно заменить на оценку D^*.
Например, выбирая $\alpha = 0.05$, получаем коэффициент $\Phi^{-1}\left(\frac{\alpha+1}{2}\right) = 1.96$
Окончательно: с вероятностью α можно сказать, что $M \in \left(M^* - \Phi^{-1}\left(\frac{\alpha+1}{2}\right)\frac{D^*}{\sqrt{n}}, M^* + \Phi^{-1}\left(\frac{\alpha+1}{2}\right)\frac{D^*}{\sqrt{n}}\right)$

23. Точность и надежность оценки, доверительный интервал.

Точность оценка характеризуется положительным числом δ, которое характеризует величину расхождения между оценками выборки и генеральной совокупности:
$\left|\theta - \theta^{\dot{c}}\right| < \delta, \delta > 0$

Надежностью (доверительной вероятностью) оценки 0 по 0* называют вероятность у, с которой осуществляется неравенство $\left|\theta - \theta^{\dot{c}}\right| < \delta$
$P\left[\theta^{\dot{c}} - \delta < \theta < \theta^{\dot{c}} + \delta\right] = \gamma$

В качестве параметров надежности наиболее часто используют величины, близкие к единице: 0,95; 0,99 и 0,999.

Доверительным называют интервал $\left(\theta^{\dot{c}} - \delta, \theta^{\dot{c}} + \delta\right)$, который покрывает
неизвестный параметр с заданной надежностью у.

24. Интегральная оценка неизвестного математического ожидания нормально распределенной генеральной совокупности.

Существуют два основных метода построения доверительных интервалов: байесовский метод и метод доверительных интервалов, предложенный Нейманом. Применяя метод построения доверительных интервалов, основанный на формуле Байеса, исходят из предположения, что оцениваемый параметр сам случаен. Предполагается также, что известно априорное распределение параметра. Этот метод часто неприменим, так как оцениваемая величина на практике является просто неизвестной постоянной, а не случайной величиной. Кроме того, ее распределение бывает также неизвестным. От этих недостатков свободен метод доверительных интервалов. Рассмотрим примеры построения доверительных интервалов в ряде случаев.

2.1. Доверительный интервал для математического ожидания при известной дисперсии
Пусть по выборке достаточно большого объема, $n > 30$, и при заданной доверительной вероятности γ необходимо определить доверительный интервал для математического ожидания $M[X] = m$, в качестве оценки которого используется среднее арифметическое (среднее выборочное) $\overline{X}_B = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot n_i$.
Закон распределения оценки математического ожидания близок к нормальному (распределение суммы независимых случайных величин с конечной дисперсией асимптотически нормально). Если потребовать

абсолютную надежность оценки математического ожидания, то границы доверительного интервала будут бесконечными $(-\infty,+\infty)$. Выбор любых более узких границ связан с риском ошибки, вероятность которой определяется уровнем значимости $1-\gamma$, где значения γ выбираются достаточно близкими к единице, например, 0,9, 0,95, 0,98, 0,99. Величину γ называют надежностью или доверительной вероятностью. Интерес представляет максимальная точность оценки, т.е. наименьшее значение интервала. Для симметричных функций минимальный интервал тоже будет симметричным относительно оценки $\overline{X}_B$. В этом случае выражение для доверительной вероятности имеет вид $P(|\overline{X}_B - m| < \delta) = \gamma$, где δ – абсолютная погрешность оценивания.

Нормальный закон $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}$ полностью определяется двумя параметрами – математическим ожиданием $m = a$ и дисперсией σ^2. Величина $\overline{X}_B$ является несмещенной, состоятельной и эффективной оценкой математического ожидания, поэтому ее значение принимаем за значение математического ожидания в качестве точечной оценки.

Будем полагать, что дисперсия σ^2 известна, тогда выборочное среднее X_B – нормально распределенная случайная величина с параметрами $\left(a, \frac{\sigma}{\sqrt{n}}\right)$. Для такой случайной величины вероятность попадания на симметричный относительно математического ожидания интервал выражается через функцию

$$P(|\overline{X}_B - a| < \delta) = 2\Phi\left(\frac{\delta}{\frac{\sigma}{\sqrt{n}}}\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t) = \gamma$$

Лапласа, где $t = \left(\frac{\delta\sqrt{n}}{\sigma}\right)$. При заданной надежности γ, уравнение $2\Phi(t) = \gamma$ можно решить приближенно с помощью таблицы значений функции Лапласа (см. приложение, таблица 1). Если точного значения $\frac{\gamma}{2}$ в списке значений нет, то надо найти два ближайших к нему значения, одно большее, а другое меньшее, чем $\frac{\gamma}{2}$, и найти их среднее арифметическое. Известное значение параметра t позволяет записать абсолютную погрешность $\delta = \frac{t\sigma}{\sqrt{n}}$. Теперь можно указать симметричный интервал $P\left(|\overline{X}_B - a| < \frac{t\sigma}{\sqrt{n}}\right) = P\left(\overline{X}_B - \frac{t\sigma}{\sqrt{n}} < a < \overline{X}_B + \frac{t\sigma}{\sqrt{n}}\right) = \gamma$. Полученное соотношение означает, что доверительный интервал $\left(\overline{X}_B - \frac{t\sigma}{\sqrt{n}}; \overline{X}_B + \frac{t\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a (математическое ожидание) с вероятностью (надежностью) $p = \gamma$, а точность оценки $\delta = \frac{t\sigma}{\sqrt{n}}$.

При фиксированном объеме выборки из оценки $\delta = \frac{t\sigma}{\sqrt{n}}$ следует, что чем больше доверительная вероятность γ, тем шире границы доверительного интервала (тем больше ошибка в оценке математического ожидания). Чтобы снизить ошибку в оценке значения, можно увеличить объем выборки. При этом, чтобы снизить относительную погрешность на порядок, необходимо увеличить объем выборки на два порядка.

25. Метод моментов для точечной оценки параметров распределения.

Метод моментов оценивания параметров распределения генеральной совокупности состоит в том, на основании выборки *x1, x2, ..., xn* вычисляются выборочные моменты (начальные или центральные). Полученные значения приравниваются соответствующим теоретическим моментам. Количество моментов должно ровняться числу оцениваемых параметров. Затем решают полученную систему уравнений относительно этих параметров.

Рассмотрим случай, когда метод моментов используется для нахождения оценки одного параметра. Положим, что плотность распределения *f(x;a)* случайной величины *X* зависит только от одного параметра*a*, и необходимо найти оценку параметра *a*. Для нахождения оценки одного параметра достаточно иметь одно уравнение относительно этого параметра, используя, например, на основании выборки *x1*, *x2*, ... , *,xn*первый начальный момент

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

Приравняем его значение первому теоретическому моменту

$$\frac{1}{n}\sum_{j=1}^{n} x_j = \int_{-\infty}^{\infty} x f(x;a)dx = \varphi(a),$$

рассматривая правую часть равенства как функцию от *a*. Решая это уравнение относительно неизвестного параметра *a*, получаем точечную оценку $\overline{a}$, которая теперь является функцией от вариант выборки, то есть

$$\overline{a} = \varphi^{-1}(x_1, x_2, ..., x_n).$$

Пример. Пусть *X* – непрерывная случайная величина подчинена показательному (экспоненциальному) закону, плотность распределения которого зависит от одного неизвестного параметра λ:

$$f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0.$$

Используя полученные экспериментальные данные *x1*, *x2*, ... , *xn*, получить оценку параметра λ.

Решение. На основании выборки *x1*, *x2*, ... , *xn* находим первый выборочный момент и приравниваем его первому моменту случайной величины *X*, подчиненной показательному закону:

$$\frac{1}{n}\sum_{j=1}^{n} x_j = \int_{-\infty}^{\infty} x f(x;\lambda)dx = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \frac{1}{\lambda}.$$

Отсюда получаем оценку параметра λ:

$$\overline{\lambda} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} x_j}.$$

Если функция плотности распределения случайной величины *X* зависит от двух параметров, например $f(x; a_1, a_2)$, то для отыскания оценок параметров a_1, a_2 необходимо иметь уже два уравнения относительно этих параметров. Для этого можно воспользоваться, например, первым начальным моментом (математическим ожиданием) и вторым центральным (дисперсией).

Примеры.

1. По выборке *x1*, *x2*, ... , *xn* методом моментов найти точечные оценки параметров *mx* и σ_x^2 нормального распределения:

$$f(x; m_x, \sigma_x^2) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}.$$

Решение. Так как первый начальный момент нормального распределения равен параметру *mx.*, а второй центральный момент равен параметру σ_x^2,

то

$$\overline{m}_x = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \overline{\sigma}_x^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \overline{m}_x\right)^2.$$

2. По выборке *x1*, *x2*, ... , *xn* методом моментов найти точечные оценки параметров *a1*, *a2* равномерного распределения на интервале [*a1*, *a2*]:

$$f(x;a_1,a_2) = \begin{cases} \frac{1}{a_2 - a_1}, x \in [a_1, a_2], \\ 0, x \notin [a_1, a_2]. \end{cases}$$

Решение. Используя выборку *x1*, *x2*, ... , *xn*, находим выборочные первый начальный и второй центральные моменты:

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \overline{\sigma}_x^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2, \quad \overline{\sigma}_x = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \overline{x}\right)^2} \quad (8.2)$$

Для равномерного распределения имеем теоретические моменты

$$m_x = \frac{a_1 + a_2}{2}, \quad \sigma_x^2 = \frac{(a_2 - a_1)^2}{12}.$$

Приравняем теоретические моменты выборочным и получаем систему двух уравнений с двумя неизвестными для нахождения оценок параметров *a1*, *a2*:

$$\begin{cases} \frac{\overline{a}_2 + \overline{a}_1}{2} = \overline{x}, \\ \frac{(\overline{a}_2 - \overline{a}_1)^2}{12} = \overline{\sigma}_x^2. \end{cases}$$

Решая эту систему, получаем в окончательном виде

$$\overline{a}_1 = \overline{x} - \sqrt{3}\overline{\sigma}_x, \quad \overline{a}_2 = \overline{x} + \sqrt{3}\overline{\sigma}_x$$

где величины $\overline{x}, \overline{\sigma}_x$ определены соотношениями (2).

26. Функциональная, статистическая и корреляционная зависимости.

Пусть у нас имеются n серии значений двух параметров X и Y: $(x_1;y_1)$, $(x_2;y_2)$,...,$(x_n;y_n)$. Подразумевается, что у одного и того же объекта измерены два параметра. Нам надо выяснить есть ли значимая связь между этими параметрами.

Как известно, случайные величины X и Y могут быть либо зависимыми, либо независимыми. Существуют следующие формы зависимости – функциональная и статистическая. В математике функциональной зависимостью переменной Y от переменной X называют зависимость вида y=f(x), где каждому допустимому значению X ставится в соответствие по определенному правилу единственно возможное значение Y.

Однако, если X и Y случайные величины, то между ними может существовать зависимость иного рода, называемая статистической. Дело в том, что на формирование значений случайных величин X и Y оказывают влияние различные факторы. Под воздействием этих факторов и формируются конкретные значения X и Y. Допустим, что на X и У влияют одни те же факторы, например Z_1, Z_2, Z_3, тогда X и Y находятся в полном соответствии друг с другом и связаны функционально. Предположим теперь, что на X воздействуют факторы Z_1, Z_2, Z_3, а на только Y и Z_1, Z_2. Обе величины и X и Y являются случайными, но так как имеются общие факторы Z_1 и Z_2, оказывающие влияние и на X и на Y, то

значения X и Y обязательно будут взаимосвязаны. И связь это уже не будет функциональной: фактор Z_3, влияющий лишь на одну из случайных величин, разрушает прямую (функциональную) зависимость между значениями X и Y, принимаемыми в одном и том же испытании. Связь носит вероятностный случайный характер, в численном выражении меняясь, от испытания к испытанию, но эта связь определенно присутствует и называется статистической. При этом каждому значению X может соответствовать не одно значение Y, как при функциональной зависимости, а целое множество значений.

ОПРЕДЕЛЕНИЕ. Зависимость случайных величин называют статистической, если изменения одной из них приводит к изменению закона распределения другой.

ОПРЕДЕЛЕНИЕ. Если изменение одной из случайных величин влечет изменение среднего другой случайной величины, то статистическую зависимость называют корреляционной. Сами случайные величины, связанные корррреляционной зависимостью, оказываются коррелированными.

Примерами корррреляционной зависимости являются: зависимость массы от роста:

- каждому значению роста (X) соответствует множество значений массы (Y), причем, несмотря на общую тенденцию, справедливую для средних, большему значению роста соответствует и большее значение массы – в отдельных наблюдениях субъект с большим ростом может иметь и меньшую массу.
- зависимость заболеваемости от воздействия внешних факторов, например, запыленности, уровня радиации, солнечной активности и т.д.
- количество (X) вводимого объекту препарата и его концентрация в крови (Y).
- между показателями уровня жизни населения и процентом смертности;
- между количеством пропущенных студентами лекций и оценкой на экзамене.

Именно корреляционные зависимости наиболее часто встречаются в природе в силу взаимовлияния и тесного переплетения огромного множества самых различных факторов, определяющих значения изучаемых показателей.

Корреляционную зависимость Y от X можно описать с помощью уравнения вида:

$y_x=f(x)$ (1)

где y_x - условное среднее величины Y, соответствующее значению x величины X, а f(x) некоторая функция. Уравнение (1) называется выборочным уравнением регрессии Y на X. Функцию f(x) называют выборочной регрессией Y на X, а ее график – выборочной линией регрессии Y на X.

Совершенно аналогично выборочным уравнением регрессии X на Y является уравнение: $x_y=\phi(y)$

В зависимости от вида уравнения регрессии и формы соответствующей линии регрессии определяют форму корреляционнной зависимости между рассматриваемыми величинами – линейной, квадратической, показательной, экспоненциальной.

Важнейшим является вопрос выбора вида функции регрессии f(x) [или ф(y)], например линейная или нелинейная (показательная, логарифимическая и т.д.)

На практике вид функции регрессии можно определить, построив на координатной плоскости множество точек, соответствующих всем имеющимся парам наблюдений (x;y).

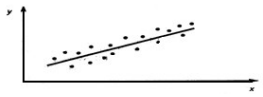

Рис. 1. Линейная регрессия значима. Модель Y=a+bX.

Рис. 2. Линейная регрессия незначима. Модель $Y=\overline{Y}$

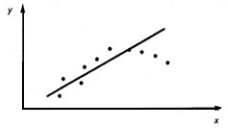

Рис. 3. Линейная регрессия значима. Нелинейная модель ($y=ax^2+bx+c$)

Например, на рис.1. видна тенденция роста значений Y с ростом X, при этом средние значения Y располагается визуально на прямой. Имеет смысл использовать линейную модель (вид зависимости Y от X принято называть моделью) зависимости Y от X. На рис.2. средние значения Y не зависят от x, следовательно линейная регрессия незначима (функция регрессии постоянна и равна $\overline{Y}$). На рис. 3. прослеживается тенденция нелинейности модели.

27. Условные средние выборочные уравнения регрессии.

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

Рассмотрим сначала зависимость Y от одной случайной (или неслучайной)

величины X. Две случайные величины могут быть связаны либо функциональ-

ной зависимостью, либо зависимостью другого рода, называемой статистиче-

ской, либо быть независимыми.

Строгая функциональная зависимость реализуется редко, так как обе вели-

чины или одна из них подвержены еще действию случайных факторов, причем

среди них могут быть и общие для обеих величин (под общими здесь подразу-

меваются такие факторы, которые воздействуют и на Y и на X). В этом случае

возникает статистическая зависимость.

Например, если Y зависит от случайных факторов Z1, Z2, V1, V2, а X зависит от случайных факторов Z1, Z2, U1, U2, то между Y и X имеется ста-

тистическая зависимость, так как среди случайных факторов есть общие: Z1 и

Z2.

Определение 71. Статистической называют зависимость, при которой из-

менение одной из величин влечет изменение распределения другой.

В частности, статистическая зависимость проявляется в том, что при из-

менении одной из величин изменяется среднее значение другой; в этом случае

статистическую зависимость называют корреляционной.

Рассмотрим пример случайной величины Y , которая не связана с величи-

ной X функционально, а связана корреляционно. Пусть Y - урожай зерна, X

- количество удобрений. С одинаковых по площади участков земли при рав-

ных количествах внесенных удобрений снимают различный урожай, т. е. Y не

является функцией от X.

Это объясняется влиянием случайных факторов (осадки, температура возду-

ха и др.). Вместе с тем, как показывает опыт, средний урожай является функци-

ей от количества удобрений, т. е. Y связан с X корреляционной зависимостью.

15.1.2. Условные средние

В качестве оценок условных математических ожиданий принимают условные

средние, которые находят по данным наблюдений (по выборке).

Определение 72. Условным средним yx называется среднее арифметическое

наблюдавшихся значений Y, соответствующих X = x.

Пример.

Если при x1 = 2 величина Y приняла значения y1 = 5, y2 = 6, y3 = 10, то условное среднее yx1 =5 + 6 + 10/3= 7

Аналогично определяется условное среднее xy.15.1. Теория корреляций 119

Определение 73. Условным средним xy называется среднее арифметическое

наблюдавшихся значений X, соответствующих Y = y.

15.1.3. Выборочные уравнения регрессии

При изучении условных вероятностей мы вводили понятие условного математи-

ческого ожидания

Условным математическим ожиданием дискретной случайной величины

Y при X = x (x - определенное возможное значение X) называется произведе-

ние возможных значений Y на их условные вероятности:

$M[Y|X = x] = Xm$

$j=1$

$yjp(yj|x)$.

Для непрерывных величин

$M[Y |X = x] = Z \infty$

$-\infty$

$y\psi(y|x)dy$,

где $\psi(y|x)$ - условная плотность случайной величины Y при X = x.

Условное математическое ожидание $M[Y|X = x]$ есть функция от x:

$M[Y|X = x] = M[Y|x] = f(x)$,

которая называется функцией регрессии Y на X.

Аналогично определяются условное математическое ожидание случайной ве-

личины X и функция регрессии X на Y :

$M[X|Y = y] = M[X|y] = \phi(y)$.

Условное математическое ожидание $M[Y |x]$ является функцией от x, следова-

тельно, его оценка, т. е. условное среднее yx

, также функция от x; обозначив

эту функцию через f
•
(x), получим уравнение
yx = f
•
(x).
Определение 74.
Выборочным уравнением регрессии Y на X называется уравнение
yx = f
•
(x).
Выборочной регрессией Y на X называется функция f
•
(x).
Выборочной линией регрессии Y на X называется график функции f
•
(x).
Аналогично уравнение
xy = ϕ
•
(y).
называется выборочным уравнением регрессии X на Y; функция ϕ
•
(y) называ-
ется выборочной регрессией X на Y , а ее график - выборочной линией регрессии
X на Y.

28. Построение линейных моделей:

1)по не сгруппированным выборочным данным.
2) по сгруппированным выборочным даннымю
Для того, чтобы найти объясненную часть, т. е. величину математического ожидания Mx(У), требуется нахождение условных распределений случайной величины Y. На практике это почти никогда не возможно.
В большинстве случаях при решении задач по эконометрике применяется стандартная процедура **сглаживания экспериментальных данных**. Эта процедура состоит из двух этапов:
1) определяется параметрическое семейство, к которому принадлежит искомая функция Mx(У) (определяемая как функция от значений объясняющих переменных X). Это может быть линейная функция, показательная функция и т.д.;
2) находятся оценки параметров этой функции с помощью одного из методов мат. статистики.
Формально никаких способов выбора параметрического семейства нет. Однако в большинстве случаев модели в задачах
предмета эконометрика выбираются линейными.
Кроме очевидного **преимущества линейной модели** — ее относительной простоты, — для этого выбора имеются, как минимум, две существенные причины.
Первая причина: если случайная величина (X, У) имеет совместное **нормальное распределение**, то уравнения регрессии линейные.
В других случаях сами величины Y или X могут не иметь нормального распределения, но некоторые функции от них распределены нормально. Например, известно, что логарифм доходов на душу населения — нормально распределенная случайная величина. В большинстве случаев гипотеза о нормальном распределении принимается, когда нет явного ей противоречия, и, как показывает практика, подобная предпосылка бывает вполне разумной.
Вторая причина, по которой линейная регрессионная модель оказывается предпочтительнее других, является меньший риск значительной ошибки прогноза.

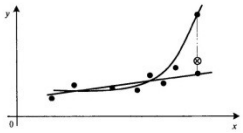

Рисунок показывает два выбора функции регрессии — линейной и квадратичной. Как видно, имеющееся множество экспериментальных данных (точек) парабола сглаживает, пожалуй, даже лучше, чем прямая. Однако парабола быстро удаляется от корреляционного поля и для добавленного наблюдения теоретическое значение может очень значительно отличаться от эмпирического.
Можно определить точный математический смысл этому утверждению: ожидаемое значение ошибки прогноза, т.е. математическое ожидание квадрата отклонения наблюдаемых значений от сглаженных (или теоретических) оказывается меньше в том случае, если выбрано линейное уравнение регрессии.

29. Выборочный коэффициент корреляции,методика его вычисления

Понятие корреляции является одним из основных понятий теории вероятностей и математической статистики, оно было введено Гальтоном и Пирсоном.
Закон природы или общественного развития может быть представлен описанием совокупности взаимосвязей. Если эти зависимости стохастичны, а анализ осуществляется по выборке из генеральной совокупности, то данная область исследования относится к задачам стохастического исследования зависимостей, которые включают в себя корреляционный, регрессионный, дисперсионный и ковариационный анализы. В данном разделе рассмотрена теснота статистической связи между анализируемыми переменными, т.е. задачи корреляционного анализа.
В качестве измерителей степени тесноты парных связей между количественными переменными используются коэффициент корреляции (или то же самое "коэффициент корреляции Пирсона") и корреляционное отношение.
Пусть при проведении некоторого опыта наблюдаются две случайные величины X и Y, причем одно и то же значение x встречается n_x раз, $y - n_y$ раз, одна и та же пара чисел (x,y) наблюдается n_{xy} раз. Все данные записываются в виде таблицы, которую называют корреляционной.
Выборочная ковариация k(X,Y) величин X и Y определяется формулой

$$k(X,Y) = \frac{1}{n}\sum (x_i - x^*)(y_i - y^*)n_{xy},$$

где $n = \sum n_{xy}$, а $\mathbf{x}^*$, y^* - выборочные средние величин X и Y. При небольшом количестве экспериментальных данных k(X,Y) удобно находить как полный вес ковариационного графа:

Выборочный коэффициент корреляции находится по формуле

$$r(X,Y) = \frac{k(X,Y)}{\sigma_x^* \cdot \sigma_y^*} = \frac{\sum n_{xy} xy - x^* y^*}{n\sigma_x^* \cdot \sigma_y^*},$$

где σ_x^*, σ_y^* - выборочные средние квадратические отклонения величин X и Y.
Выборочный коэффициент корреляции $r(X,Y)$ показывает тесноту линейной связи между X и Y: чем ближе $|r(X,Y)|$ к единице, тем сильнее линейная связь между X и Y.

30. Статистические гипотезы, ошибки 1-го и 2-го рода, уровень значимости.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют статистической. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т. е. могут быть допущены ошибки двух родов.
Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза.
Ошибка второго рода состоит в том, что будет принята неправильная гипотеза.
Вероятность совершить ошибку первого рода принято обозначать через α; ее называют уровнем значимости. Наиболее часто уровень значимости принимают равным 0.05 или 0.01. Если, например, принят уровень значимости, равный 0.05, то это означает, что в пяти случаях из ста мы рискуем допустить ошибку первого рода (отвергнуть правильную гипотезу).

Пусть дана выборка $\mathbf{X} = (X_1, \ldots, X_n)^{\mathsf{T}}$ из неизвестного совместного распределения $\mathbb{P}^X$, и поставлена бинарная задача проверки статистических гипотез:
H_0
H_1,
где H_0 — нулевая гипотеза, а H_1 — альтернативная гипотеза.
Предположим, что задан статистический критерий
$f : \mathbb{R}^n \to \{H_0, H_1\}$,

31. Статистический критерий проверки нулевой гипотезы.

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближенное распределение которой известно. Эту величину обозначают через U или Z, если она распределена нормально, F или v^2 - по закону Фишера-Снедекора, T – по закону Стьюдента, c^2 – по закону «хи квадрат» и т. д. Все эти случайные величины обозначим через K.
Статистическим критерием (или просто *критерием*) называют случайную величину K, которая служит для проверки нулевой гипотезы. Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин, и таким образом получают частное (наблюдаемое) значение критерия.
Наблюдаемым значением $K_{набл}$ назначают значение критерия, вычисленное по выборкам.

32. Критическая область. Область принятия гипотезы, критические точки.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества, одно из которых содержит значения критерия, при которых нулевая гипотеза отвергается, а другое – при которых она принимается.
Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значений критерия, при которых гипотезу принимают.
Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если области принятия гипотезы – гипотезу принимают.
Так как критерий K – одномерная случайная величина, то все ее возможные значения принадлежат некоторому интервалу и, соответственно, должны существовать точки, разделяющие критическую область и область принятия гипотезы. Такие точки называются критическими точками.
Различают одностороннюю (правостороннюю и левостороннюю) и двустороннюю критические области.
Правосторонней называют критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр}$ – положительное число.
Левосторонней называют критическую область, определяемую неравенством $K < k_{кр}$, где $k_{кр}$ – отрицательное число.
Двусторонней называют критическую область, определяемую неравенствами $K < k_1$, $K > k_2$, где $k_2 > k_1$. В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами $K < -k_{кр}$, $K > k_{кр}$ или равносильным неравенством $|K| > k_{кр}$. Различия между вариантами критических областей иллюстрирует следующий рисунок.

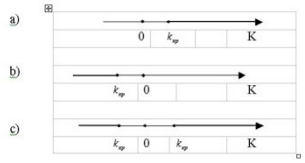

Рис. 1. Различные варианты критических областей a) правосторонняя, b) левосторонняя, c) двусторонняя
Резюмируя, сформулируем этапы проверки статистической гипотезы:
Формулируется нулевая гипотеза H_0; Определяется критерий K, по значениям которого можно будет принять или отвергнуть H_0 и выбирается уровень значимости α; По уровню значимости определяется критическая область; По выборке вычисляется значение критерия K, определяется, принадлежит ли оно критической области и на основании этого принимается H_0 или H_1.

33. Сравнение 2-х дисперсий нормально распределенных генеральных совокупностей.

Пусть имеются две выборки объемов n1
и n2, извлеченные из нормально
распределенных генеральных совокупностей X и Y. Требуется по
исправленным выборочным дисперсиям Sx2 и Sy2
проверить нулевую
гипотезу о равенстве генеральных дисперсий рассматриваемых
генеральных
совокупностей:Но: D (X) = D (Y).
Критерием служит случайная величина = - 2
2
м
б
s
s
F отношение большей
исправленной дисперсии к меньшей, которая при условии
справедливости
нулевой гипотезы имеет распределение Фишера - Снедекора со
степенями
свободы k
1
= n1 – 1 и k
2
= n2 – 1. Критическая область зависит от вида
конкурирующей гипотезы:
1) если H1: D (X) > D (Y), то критическая область правосторонняя:
((, ,)) . p F > Fкр α k
1
k
2
=α
Критическая точка (, ,)
1 2
F k k
кр α находится по таблице критических точек
распределения Фишера - Снедекора. Если = < кр -
м
б
набл F
s
s
F 2
2
нулевая гипотеза
принимается, в противном случае – отвергается.
2) При конкурирующей гипотезе H1: D (X) ≠ D (Y) критическая область двусторонняя: .
2
, ( )
2
( )
1 2
α α
p F < F = p F > F =
При этом достаточно найти
, ,).
2
(
2 1 2
F F k k
кр
α
=
Тогда, если = < кр -
м
б
набл F
s
s
F 2
2
нет оснований отвергнуть
нулевую гипотезу, если Fнабл > Fкр - нулевую гипотезу отвергают.