1. Предмет и основные понятия ТВ

ТВ – математическая наука изучающая закономерность в массовых однородных случаях, явлениях и процессах.

Элементарные события – это простейшие не разложимые результаты опыта. Вся совокупность элементарных событий – пространство элементарных событий.

Под опытом в ТВ понимается выполнение некоторого комплекса условий в результате которого происходят или не происходят некоторые события – факты.

Событие в ТВ – это любое конечное или счетное подмножество пространства Ω.

Три типа событий:

* Достоверные
* Случайные
* Невозможные.

События являются несовместными если они не могут происходить одновременно и наоборот.

Элементы последовательность попарно несовместны, если любые два из них попарно несовместны.

Несколько событий равновозможные, если ни одно из них не имеет объективного преимущества перед другим. События образуют полную группу если в результате опыта ничего кроме этих событий не может произойти.

Алгебра событий.

1) Суммой двух событий А + В = А∪В называется такое третье событие которое заключается в наступлении хотя бы одного из событий А или В (или).

2) Произведением двух событий А\*В = А∩В называется такое третье событие, которое заключается в наступлении двух событий одновременно (и).

3) Отрицанием события А является событие ‾А, которое заключается в ненаступлении А.

4) Если наступление события А приводит к наступлению события В и наоборот, то А=В.

Пусть множество S – это множество всех подмножеств пространства всех элементов Ω для которых выполняются следующие условия:

1. Если А∈ S, B ∈ S, то A+B = A∪B ∈ S
2. Если А∈ S, B ∈ S, то А\*В = А∩В ∈ S
3. Если А∈ S, то ‾А ∈ S.

Тогда множество S называется алгеброй событий.

При точном подходе достаточно одного из этих свойств, так как каждое из них следует из другого.

При расширении операции сложения и умножения, на случай счетного множества событий, алгебра событий называется бролевской алгеброй.

2. Определение вероятности события.

Аксиоматическое определение вероятности.

Вероятность события – это численная мера объективной возможности его появления.

Аксиомы вероятности:

* Каждому событию А ставится в соответствие неотрицательное число р, которое называется вероятностью события А. Р(А)=р ≥ 0, где А∈ S, S⊆Ω.
* Р(Ω) = 1, где Ω - истинное (достоверное) событие.

Аксиоматический подход не указывает, как конкретно находить вероятность.

Классическое определение вероятности.

Пусть событие А1,А2, …, Аn ∈ S (\*) образуют пространство элементарных событий, тогда событие из \* которое приводит к наступлению А, называют благоприятствующими исходами для А. Вероятностью А называется отношение числа исходов благоприятствующих наступлению события А, к числу всех равновозможных элементарных исходов.

|  |  |
| --- | --- |
| (А)= | m(A) |
| Рn |

Свойства вероятности:

1. 0 ≤ Р(А) ≤ 1,
2. Р (Ω) =1,
3. Р (‾Ω) = 0.

Статическое определение вероятности.

Пусть проводится серия опытов (n раз), в результате которых наступает или не наступает некоторое событие А (m раз), тогда отношение m/n, при n→∞ называются статистической вероятностью события А.

Геометрическое определение вероятности.

Геометрической вероятностью называется отношение меры области, благоприятствующей появлению события А, к мере всей области.

3. Интегральная функция распределения и ее свойства

Для непрерывной случайной величины X вероятность Р(Х= xi)→0, поэтому для НСВ удобнее использовать вероятность того, что СВ Х<хi, где хi- текущее значение переменной. Эта вероятность называется интегральной функцией распределения: P(X<xi)=F(x).

Интегральная функция является универсальным способом задания СВ (как для ДСВ, так и для НСВ).

Свойства интегральной функции распределения:

1) F(x) не убывает (если х2>x1, то F(x2)≥Р(х1));

2). F(-∞)=0;

3). F(+∞)=1;

4) вероятность попадания СВ X в интервал а<Х<b определяется по формуле

P(a≤X<b)=F(b)-F(a).

Замечание. Обычно для определённости левую границу включают в интервал, а правую нет. Вообще для НСВ верно, что

Р(а≤Х<b)= Р(а <Х≤b) =Р(а<Х < b)= Р(а≤X≤b).

4. Основные теоремы теории вероятностей

Теорема1.

Вероятность суммы двух несовместных событий А и В равна сумме их вероятностей:

Р(А+В)=Р(А)+Р(В).

Следствие1.

Если А1,А2, …, Аn - попарно несовместные события, то вероятность их суммы равна сумме вероятностей этих событий.

Следствие2.

Вероятность суммы попарно несовместных событий А1,А2, …, Аn , образующих полную группу, равна 1.

Следствие3.

События А и ‾А несовместны и образуют полную группу событий, поэтому

Р(А +‾А) = Р(А) + Р(‾А) = 1. Отсюда Р (‾А) = 1 – Р(А).

Теорема2.

Вероятность суммы двух совместных событий А и В равна сумме вероятностей этих событий без вероятности их произведения:

Р (А+В) = Р(А)+Р(В) – Р (А\*В).

Два события А и В называются независимыми, если появление одного из них не влияет на вероятность появления другого (в противном случае события зависимы).

Теорема3.

Вероятность произведения двух независимых событий равна произведению их вероятностей Р(А\*В)=Р(А)\*Р(В).

Следствие.

Вероятность произведения n независимых событий А1,А2, …, Аn равна произведению их вероятностей.

Условной вероятностью события В при условии, что событие А уже произошло, называется число Р(АВ)/Р(А)=Р(В/А)РА(В).

Теорема4.

Вероятность произведения двух зависимых событий А и В равна произведению вероятности наступления события А на условную вероятность события В при условии что событие А уже произошло:

Р(А\*В) =Р(А)\*Р(В/А).

Следствие.

Если события А и В независимы, то из теоремы 4 следует теорема 3.

Событие В не зависит от события А, если Р(В/А) = Р(В). Теорему 4 можно обобщить на n событий.

Теорема5.

Вероятность произведения n зависимых событий А1,А2, …, Аn равна произведению последовательных условных вероятностей:

Р(А1\*А2\*…\*Аn-1\*An)= P(A1)\*P(A2/A1)\*...\*P(An/A1\*A2\*...\*An-1).

Теорема6.

Вероятность наступления хотя бы одного из событий А1,А2, …, Аn равна разности между единицей и вероятностью произведении отрицаний событий А1,А2, …, Аn :

Р(А)=1-Р(‾А1\*‾А2\*…\*‾Аn)=1- P(‾A1)\*P(‾A2/‾A1)\*...\*P(‾An/‾A1\*‾A2\*...\*‾An-1).

Следствие1.

Вероятность наступления хотя бы одного из событий А1,А2, …, Аn независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий:

Р(А)=1-Р(‾А1)Р(‾А2)…Р(‾Аn).

Следствие2.

Если события А1,А2, …, Аn независимы и имеют одинаковую вероятность появиться (Р(А1)=Р(А2)=…Р(Аn)= р, Р(Аi)= 1-р=q ), то вероятность появления хотя бы одного из них равна Р(А)=1-qn .

5. Формулы полной вероятности и вероятности гипотез

Пусть событие А может наступать только одновременно с одним из попарно несовместных событий Н1, Н2, ..., Нn, образующих полную группу. Тогда вероятность события А определятся по формуле полной вероятности:

Р(А) = Р(Н1)\*P(А/Н1) + Р(Н2)\*Р(А/Н2) +...+ Р(Нn)\*Р(А/Нn), или Р(А)= Σ Р(Нi)\*Р(А/Нi),

где события Н1,Н2, ...,Нn, - гипотезы, a P(A/Hi) - условная вероятность наступления события А при наступлении i-ой гипотезы (i=1, 2,..., n).

Условная вероятность гипотезы Нi при условии того, что событие А произошло, определяется по формуле вероятности гипотез или формуле Байеса (она позволяет пересмотреть вероятности гипотез после наступления события А):

Р(Нi/А)=(P(Hi)\*P(A/Hi))/P(A).

6. Формула Бернулли

Пусть некоторый опыт повторяется в неизменных условиях n раз, причём каждый раз может либо наступить (успех), либо не наступить (неудача) некоторое событие А, где Р(А) = р - вероятность успеха, Р(А)=1-р= q - вероятность неудачи. Тогда вероятность того, что в к случаях из n произойдёт событие А вычисляется по формуле Бернулли

Pn(K) = Ckn-pk-qn-k.

Условия, приводящие к формуле Бернулли, называются схемой повторных независимых испытаний или схемой Бернулли. Так как вероятности Рn(к) для раз личных значений к представляют собой слагаемые в разложении бинома Ньютона

(p+q)n=C0n\*p0\*qn+C1n\*p1\*qn-1+…+Ckn\*pk\*qn-k+…+Cnn\*pn\*q0,

то распределение вероятностей Pn(k), где 0≤k≤n, называется биноминальным.

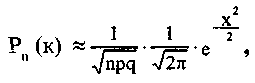
Если в каждом из независимых испытаний вероятности наступления события А разные, то вероятность наступления события А к раз в n опытах определяется как коэффициент, при к-ой степени полинома

φn(Z)=Π(qi+piZ)=anZn+an-1Zn-1+…+a1Z1+a0, где φn(Z) - производящая функция.

Невероятнейшее число наступивших событий в схеме Бернулли - ко (к0 c К) определяется из следующего неравенства: np-q≤k0≤np+p.

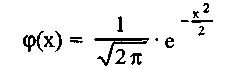
7. Локальная формула Муавра-Лапласа

Если npq>10 , то



где вероятность р отлична от 0 и 1 (р→0,5), х =(k-np)/√npq.

Для облегчения вычислений функция



представлена в виде таблицы (прил.1).

φ(х) - функция вероятности нормального распределения (рис. 6) имеет следующие свойства:

1) φ(х)-четная;

2) точки перегиба х = ± 1;

3) при х≥5, φ(х)→0, поэтому функция φ(х) представлена в виде таблицы для 0≤х≤5 (прил.1).

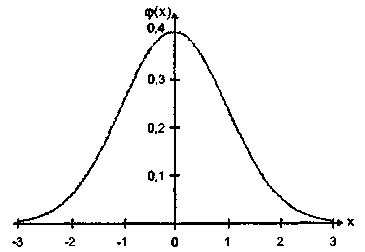


Рис. Функция вероятности нормального распределения

8. Интегральная теорема Муавра-Лапласа.

При больших значениях n , для вычисления вероятности того, что произойдет от к1, до к2 событий по схеме



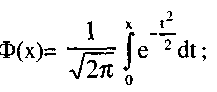
Бернулли, используется интегральная формула Муавра-Лапласа

Pn(k1≤k≤k2)=Ф(x2)- Ф(x1),

где x1=(k1-np) /(√npq), x2=(k2-np)/(√npq), Ф(x) – функция Лапласа. (рис.7)

Ф(х) имеет следующие свойства:

1. Ф(-х)= -Ф(х) - функция нечетная, поэтому достаточно изучать её для неотрицательных значений х



2. Функция Ф(х) возрастает на всей числовой оси;

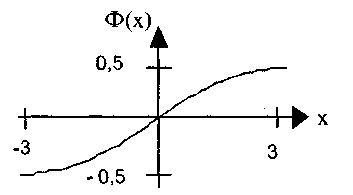
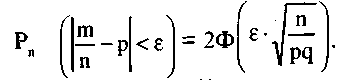


Рис. Функция Лапласа

3. При х≥5, Ф(х)→1/2 (y = 0,5 горизонтальная асимптота при х>0), поэтому функция представлена в виде таблицы Для 0≤х≤5 (прил.1).

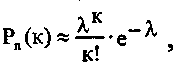
4. Вероятность отклонения относительной частоты от постоянной вероятности в независимых испытаниях не более чем на некоторое число ε>0



9. Формула Пуассона

Если npq<10 и р<0,1, то

где λ=np.



10. Случайные величины и их виды

Случайной величиной (СВ) называют такую величину, которая в результате опыта может принимать те или иные значения, причем до опыта мы не можем сказать какое именно значение она примет. (Более точно, СВ - это действительная функция, определенная на пространстве элементарных событий Q). Случайные величины обозначаются последними буквами латинского алфавита - X,Y,Z. Случайные величины могут быть трех типов: - дискретные, - непрерывные, - смешанные (дискретно-непрерывные). Дискретная случайная величина (ДСВ) может принимать конечное или бесконечное счетное число значений. Непрерывная случайная величина (НСВ) в отличие от ДСВ принимает бесконечное несчетное число значений. Например мишень имеет форму круга радиуса R. По этой мишени произвели выстрел с обязательным попаданием. Обозначим через Y расстояние от центра мишени до точки попадания, Ye [0; R]. Y - непрерывная случайная величина, так как она принимает бесконечное несчетное число значений.

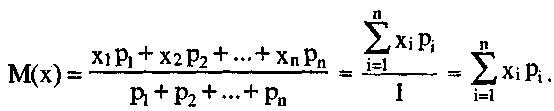
Пусть X - дискретная случайная величина, которая принимает значения х1, х2, ...,хn,... с некоторой вероятностью рi, где i = 1, 2, ..., n,... Тогда можно говорить о вероятности того, что случайная величина X приняла значение хi: рi=Р(Х=хi).

ДСВ может также представляться в виде многоугольника распределения – фигуры, состоящей из точек, соединенных отрезками. Над СВ устанавливаются операции сложения и умножения.

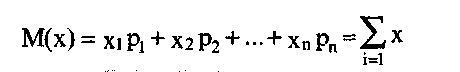
Суммой двух СВ X и Y наз-ся случайная величина, которая получается в рез-те сложения всех значений случайной величины X и всех значений СВ Y, соответствующие вероятности перемножаются. Произведением двух СВ X и Y наз-ся СВ, которая получается в рез-те перемножения всех значений СВ X и всех значений СВ Y, соответствующие вероятности перемножаются.

11. Математическое ожидание

Математическим ожиданием М(Х) ДСВ X называется среднее значение случайной величины:



Или иначе, М(Х) - это сумма парных произведений случайной величины на соответствующую вероятность:



Мода Мо(Х) распределения - это значение СВ, имеющее наиболее вероятное значение.

Медиана Ме(Х) - это значение случайной величины, которое делит таблицу распределения на две части таким образом, что вероятность попадания в одну из них равна 0,5. Медиана обычно не определяется для ДСВ.

Свойства математического ожидания:

1) М(С)=С, где С=const;

2)М(СХ) = СМ(Х);

3) M(X±Y) = М(Х) ± M(Y);

4) Если случайные величины X и Y, независимы, то M(XY) = M(X)\*M(Y).

Для биномиального распределения М(Х)=nр;

для геометрического распределения М(Х)= 1/р;

для распределения Пуассона М(Х)=λ;

для гипергеометрического распределения М(Х) = n(M/N).

12. Дисперсия ДСВ и ее свойства

Математическое ожидание квадрата отклонения СВ от ее математического ожидания:

D(X) = M(x-M(X)2) = =(х1-М(Х))2р1+(х2-М(Х))2р2+....+(xn-М(Х))2рn .(2.3.2)

Свойства дисперсии:

1) D(С) = 0, где С=соnst;

2) D(CX)=C2D(X);

3) D(X)=M(X2)-(M(X))2, где М(Х2) = х21 р1 + x22 p2 + ...+ х2n рn;

4) Если СВ X и Y независимы, то D(X±Y)=D(X) + D(Y);

5) D(OX)=D(X);

6) Для любых СВ X и Y, D(X±Y)=D(X)+D(Y)±2cov(X,Y), где cov(X,Y)=M((X-mx)(Y-m )) - ковариация случайных величин X и Y (М(Х)= mx, M(Y)= m).

Дисперсия характеризует средний квадрат отклонения ДСВ, поэтому на практике часто используют в качестве характеристики разброса среднее квадратическое отклонение σ(Х)= √D(X) , которое имеет ту же размерность, что и СВ X.

Для биноминального закона

D(X)=npq, σ(X)=√npq;

для геометрического закона D(X)= q/p2;

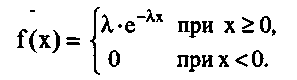
для гипергеометрического D(X)=n(M/N)(1-M/N)(N-n)/(N-1);

для распределения Пуассона D(X)=λ.

Только для распределения Пуассона M(X)=D(X)= λ.

13. Показательное распределение.

НСВ X, принимающая неотрицательные значения, имеет показательное распределение, если ее дифференциальная функция имеет вид



где Я =const, Я >0.

Интегральная функция показательного закона с параметром λ:

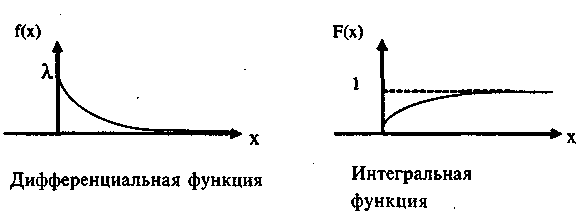
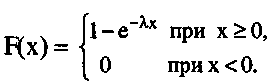


Рис. Показательный закон

Если СВ X распределена по показательному закону, то:

1. Математическое ожидание М(Х) = 1/λ ;

2. Дисперсия D(X)=1/λ2, среднее квадратическое отклонение

σ(X)=√D=1/λ.

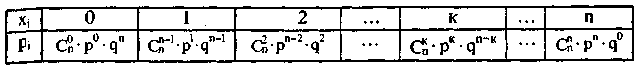
3. Вероятность попадания СВ X в заданный интервал определяется по формуле

Р(а≤х<b) = е-λа-е-λb.

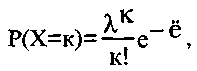
Замечание. Показательное распределение играет большую роль в теории массового обслуживания (ТМО), теории надежности. В ТМО параметр X - среднее число событий, приходящихся на единицу времени. При определенных условиях число событий, произошедших за промежуток времени т, распределено по закону Пуассона с математическим ожиданием а =λτ. Длина промежутка t, между произвольными двумя соседними событиями, подчиняется показательному закону: P(T<t)=F(t)=l-eλt.

14. Закон распределения дискретной случайной величины

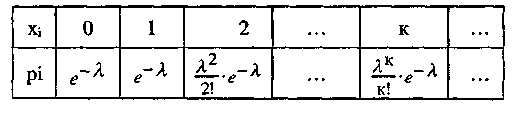
1. Биномиальный закон распределения. Случайная величина X принимает значения 0, 1, 2, 3, 4, 5,...,n, с вероятностью, определяемой по формуле Бернулли:



2. Закон распределения Пуассона. Случайная величина X принимает бесконечное счетное число значений: 0, 1, 2, 3, 4, 5,..., к,... , с вероятностью, определяющейся по формуле Пуассона:



где Х>0 - параметр распределения Пуассона.



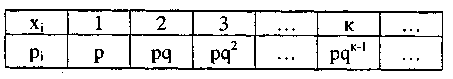
При n→∞ и р→0 биномиальный закон приближается к закону распределения Пуассона, где λ, = np.

Геометрический закон распределения. Пусть Р(А)=р - вероятность наступления события А в каждом опыте, соответственно, q=l-p - вероятность не наступления события А.

Вероятность наступления события А в к-ом опыте определяется по формуле:

P(X=k)=p-qk-1. (2.2.2.)

Случайная величина X, распределенная по геометрическому закону принимает значения 1, 2,...,к,... , с вероятностью, определяемой по формуле (2.2.2):



4. Гипергеометрический закон распределения. Пусть в урне N-шаров, из них М белых, а остальные (N - М) черные. Найдем вероятность того, что из извлеченных n шаров m белых и (n-m) черных.

N= М + (N-M); n = m + (n-m);

СmM - число способов выбора m белых шаров из М;

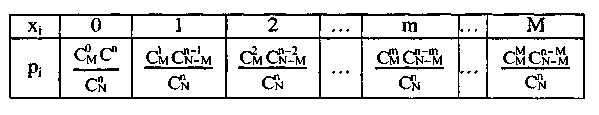
Сn-mN-M- число способов выбора (n-m) черных шаров из (N-M).

По правилу произведения, число всех возможных наборов из m белых и (n-m) черных равно СmM Сn-mN-M;

CnN- общее число способов выбора из N шаров n.

Отсюда, по формуле классического определения вероятности, P(A)= (СmM Сn-mN-M)/ CnN

Ограничения на параметры: М≤N, m≤n; m = m0, m0 +1, m0+2,..., min(M,n), где m0=max{0, n-(N-M)}. Случайная величина Х, распределенная по гипергеометрическому закону распределения (при т=0,1,2,3,...,М), имеет вид:



Гипергеометрический закон определяется тремя параметрами N, М, n. При n<0,1N этот закон стремится к биномиальному.

Замечание.

1. В теории вероятностей различают две основные схемы: выбора элементов с возвращением каждый раз обратно и выбора без возвращения, которые описываются соответственно биномиальным и гипергеометрическим законами.

2. Геометрический закон описывает схему повторения опытов (в каждом из которых может наступить или не наступить событие А: Р(А)=р, q=l-p), до первого появления события А, то есть фактически это отрицательное биномиальное распределение при m=1.

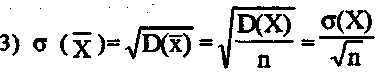
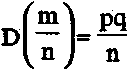
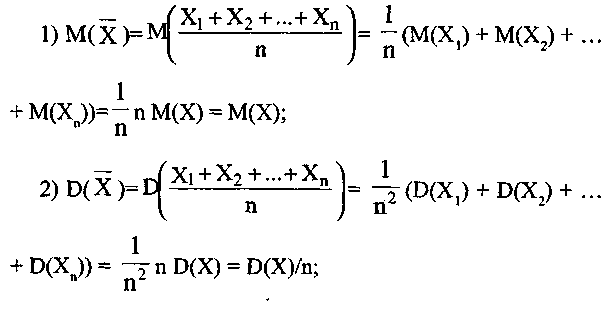
16. Одинаково распределённые, взаимонезависимые дискретные случайные величины

СВ называют одинаково распределенными, если они имеют одинаковые законы распределения. Поэтому у них совпадают числовые характеристики: математическое ожидание, дисперсия, среднее квадратическое отклонение.

Пусть X1, Х2,..., Хn одинаково распределенные, взаимонезависимые ДСВ, тогда:

M(X1) = М(Х2) = ... = М(Хn) = М(Х), D(X1) = D(X2) = ...= D(Xn)=D(X).

Рассмотрим характеристики их средней арифметической X = (X1+X2+…+Xn)/n:



-стандартное отклонение СВ X.

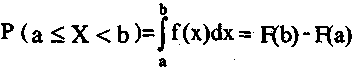
Дисперсия относительной частоты (m/n) появления события А в n независимых испытаниях (в каждом из которых событие А появляется с вероятностью равной р, и не появляется с вероятностью q= 1-р; m-число появлений события А в серии из n испытаний), определяется по формуле

15. Дифференциальная функция распределения и ее свойства

СВ X непрерывна, если ее интегральная функция непрерывна на всей числовой оси. СВ X непрерывна и имеет дифференциальную функцию, если ее интегральная функция непрерывна и дифференцируема всюду, за исключением конечного числа точек на любом конечном промежутке.

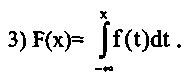
Дифференциальной функцией (функцией плотности вероятности) СВ X называется производная ее функции распределения: f(x)=F'(x).

С помощью дифференциальной функции можно получить формулу вероятности попадания СВ X в заданный интервал:



Свойства дифференциальной функции:

). f(x)≥0;



16. Числовые характеристики непрерывных случайных величин

1. Математическое ожидание НСВ X определяется по формуле

М(Х)= ∫xf(x)dx. (2.7.1)

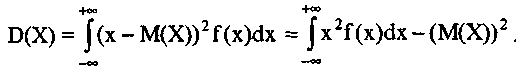
Если НСВ X определена на интервале (а; b), то

М(Х)= ∫xf(x)dx.

2) Мода НСВ X будет определяться как максимум ее дифференциальной функции: Мо(Х) = max f (x).

3) Медиана определяется как значение случайной величины, которое делит площадь под дифференциальной функцией на две равные части. Me(X): P(x<Me(X))=P(x>Me(X))=1/2.

4). Дисперсия НВС



Все св-ва дисперсии и мат-го ожидания, установленные для ДСВ, сохраняется для НСВ.

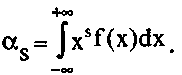
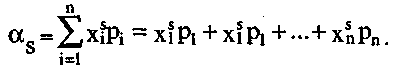
Замечание. Если распределение симметрично, то его мода, медиана и математическое ожидание совпадают.

5). Моменты случайных величин.

Кроме характеристик положения и рассеяния существует ряд других числовых характеристик распределения, например, моменты.

Начальным моментом порядка s называется математическое ожидание степени s CB X: αs=M(Xs).

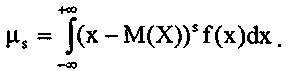
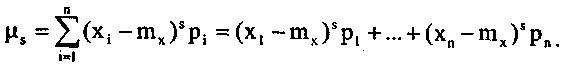
Для ДСВ:



При s=l:α1, = M(X) = mx, то есть, первый начальный момент - это математическое ожидание СВ.

Отклонение СВ от ее математического ожидания называется центрированной СВ X: X = Х-mх.

Центральным моментом порядка s СВ X называется математическое ожидание степени s, соответствующей центрированной СВ: μs=μ(Xs)=M((x-M(X))s).



При вычислении центральных моментов пользуются формулами связи между центральными и начальными моментами:

μ1=0,

μ2=α2-m2x,

μ3=α3-3mxα2+2m3x,

μ4=α4-4mxα3+6m2xα2-3m4x.

Обычно рассматривают первые четыре центральных момента:

1). μ1=M(x-mx)=0 – мат-ое ожидание центрированной СВ равно нулю;

2). μ2= M(x-mx)2=D(X) – второй центральный момент – это дисперсия;

3). μ3= M(x-mx)3- третий центральный момент может служить для характеристики асимметрии, обычно рассматривают безразмерный коэффициент асимметрии

Sk=μ3/σ3.

4). Четвёртый центральный момент

μ4=M(x-mx)4,

может служить для характеристики “крутости” или островершинности распределения, описывающиеся с помощью эксцесса:

Ex=(μ4/σ4)-3.

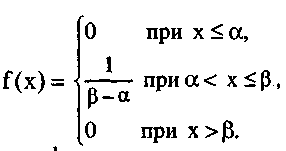
Основным моментом порядка s называется нормированный центральный момент порядка s:

rs= μs/σs, то есть Sk=r3, Ex=r4-3

17. Равномерный закон распределения

СВ X распределена по равномерному (прямоугольному) закону, если все значения СВ лежат внутри некоторого интервала и все они равновероятны (точнее обладают одной плотностью вероятности).Например, если весы имеют точность 1г и полученное значение округляется до ближайшего целого числа k, то точный вес можно считать равномерно распределенной СВ на интервале (k-0,5; k+0,5).

Дифференциальная функция равномерного закона на интервале (α,β) (рис. 11):



Интегральная функция равномерного закона на интервале (α,β) (рис. 11):

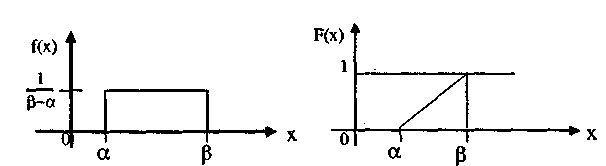
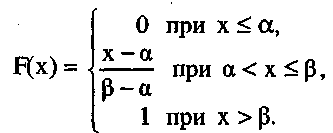
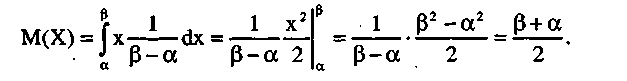


Рис. Дифференциальная функция

2). Интегральная функция.

Основные числовые характеристики равномерного закона:

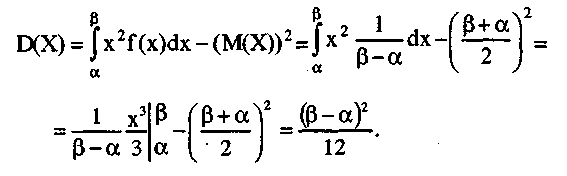
1. Математическое ожидание



М(Х) совпадает, в силу симметрии распределения, с медианой.

Моды равномерное распределение не имеет.

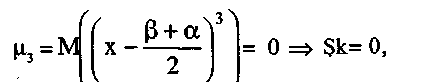
Дисперсия



Отсюда, среднее квадратическое отклонение

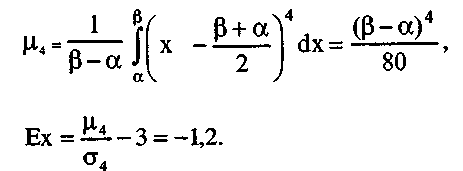


Третий центральный момент



поэтому распределение симметрично относительно М(Х).

Четвёртый центральный момент



Вероятность попадания СВ в заданный интервал (а;b). Пусть СВ X распределена по равномерному закону,

18. Закон больших чисел

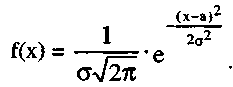
Под законом больших чисел в теории вероятностей понимают совокупность теорем, в которых утверждается, что существует связь между средним арифметическим достаточно большого числа случайных величин и средним арифметическим их математических ожиданий.

В1927 г. Гейзенберг открыл принцип неопределенности, который утверждает, что измерительное познание ограничено. Неопределенность является неотъемлемой частью нашей жизни, однако, при большом числе однотипных опытов можно установить определенные закономерности.

19. Нормальный закон распределения

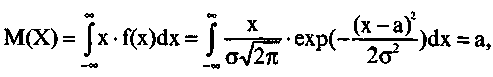
Нормальный закон распределения играет исключительную роль в теории вероятностей. Это наиболее часто встречающийся закон распределения, главной особенностью которого является то, что он является предельным законом, к которому, при определённых условиях, приближаются другие законы распределения.

Дифференциальная функция нормального закона имеет вид



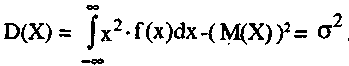
Числовые характеристики нормального закона:

1. Математическое ожидание характеризует центр распределения



где ex=exp(x);

2. Дисперсия характеризует форму распределения



Свойства дифференциальной функции нормального закона:

1. Область определения: Df = R;

2. Ось ОХ - горизонтальная асимптота;

3. х = а±σ - две точки перегиба;

4. Максимум в точке с координатами (а; 1/(σ√2π);

5. График симметричен относительно прямой х=а;

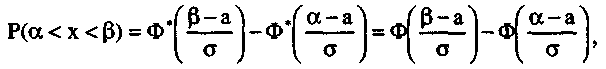
6. Моменты:

μ1=μ3=…=μ2k+1=…=0,

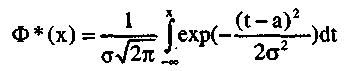
μ2=σ2, μ4=3σ4,

Sk=μ3/σ3=0, Ex=μ4/σ4-3=0

7. Вероятность попадания нормально распределенной случайной величины в заданный интервал определяется, по свойству интегральной функции



где



интегральная функция нормального закона (рис.14); Ф(х)- функция Лапласа.

Свойства интегральной функции нормального закона:

1. Ф\* (-∞)=0;

2. Ф\*(+)=1;

3. Ф\*(x)=1/2+Ф(x);

4. Ф\*(-x)=1-Ф\*(x).

Вероятность заданного отклонения. Правило трех сигм.

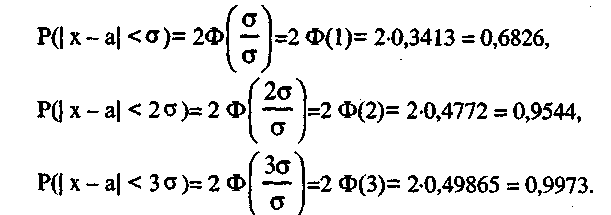
Найдем вероятность того, что случайная величина X, распределённая по нормальному закону, отклонится от математического ожидания М(Х)=а не более чем на величину ε>0.

Р(|х-а|<ε)= Р(-ε< х-а<+ε) = Р(а-ε<х< а+е) =Ф\*((a+ε-a)/σ)-Ф\*((a-ε-a)/σ)=Ф\*(ε/σ)-(1-Ф\* (ε/σ))=2Ф\* (ε/σ)-1.

Или, используя функцию Лапласа:

P(|X-a|<ε)=2Ф(ε/σ).

Найдём вероятность того, что нормально распределённая СВ X отклонится от M(X)=a на σ, 2σ, 3σ:



Отсюда следует правило Зσ. если случайная величина X имеет нормальное распределение, то отклонение этой случайной величины от ее математического ожидания по абсолютной величине не превышает утроенное среднее квадратическое отклонение (Зσ).

20. Многомерные случайные величины

В практических задачах приходится сталкиваться со случаями, когда результат описывается двумя и более случайными величинами, образующими систему случайных величин (случайный вектор). Например, точка попадания снаряда имеет две координаты: х и у, которые можно принять за систему случайных величин, определенных на одном и том же пространстве элементарных событий Ω.

Закон распределения дискретной двумерной случайной величины можно представить в виде таблицы, характеризующей собой совокупность всех значений случайных величин и соответствующих вероятностей:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | x1 | x2 | … | xn | Σ P(yj) |
| y1 | P(x1,y1) | P(x2,y2) | … | P(xn,y1) | P(y1) |
| y2 | P(x1,y2) | P(x2,y2) | … | P(xn,y2) | P(y2) |
| … | … | … | … | … | … |
| ym | P(x1,ym) | P(x2,ym) | … | P(xn,ym) | P(ym) |
| Σ Pxi | P(x1) | P(x2) | … | P(xn) | 1 |

В общем случае двумерная случайная величина задается в виде интегральной функции, которая означает вероятность попадания двумерной случайной величины в квадрант левее и ниже точки с координатами (х, y):

F(x, у) = Р(Х<х, Y<y).

21. Свойства интегральной функции:

1. F - не убывает и непрерывна слева по каждому аргументу.

2. F(-∞, у)= F(x,-∞)= F(-∞, -∞)= 0.

3. F(+∞, у)= F2(y) - функция распределения случайной величины Y. F(x,+∞)= F1,(x) - функция распределения случайной величины X.

4. F(+∞,+∞)=l.

Вероятность попадания двумерной случайной величины в прямоугольник определяется исходя из определения интегральной функции двумерной случайной величины:

Р((х, у) c D) = F(β,δ) - F(α,β) - F(β,γ) + F(α,γ).

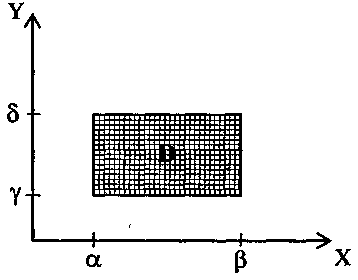


Рис. Вероятность попадания точки (х, у) в прямоугольник D

Случайные величины X, Y независимы, если

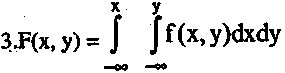
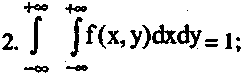
F(x, у) = = F1(x)\* F2(y).

Дифференциальная функция системы двух непрерывных случайных величин определяется как вторая смешанная производная функции распределения:

f(x,y)=(∂2F(x,y))/∂x∂y=F″xy(x,y).

Свойства дифференциальной функции:

l.f(x,y)>0;



Геометрически свойство 2 означает, что объем тела, ограниченного поверхностью f (x, у) и плоскостью XOY, равен 1.

Если случайные величины X и Y независимы, то

f(x,y) = f1(x) f2(y), где f1(x)=F’1(x),f2(y)=F’2(y).

# В противном случае

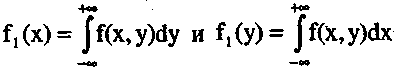
# f ( x , у ) = f1( x ) f ( y / x )

# или f ( x, y) = f2( y ) f (x / y ),

# где f(y/x)=f(x,y)/f1(x) - условная дифференциальная функция CB Y при заданном значении

# X = x, f(y/x)=f(x,y)/f2(x) - условная дифференциальная функция СВ X при заданном значении Y= у;

- дифференциальные функции отдельных случайных величин X и Y, входящих в систему.



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

Начальным моментом порядка s,h системы двух случайных величин X, Y называется математическое ожидание произведения степени s случайной величины X и степени h случайной величины Y:

αs,h =M(XsYh)

Центральным, моментом порядка s, h системы СВ (X, Y) называется математическое ожидание произведения степеней s, h соответствующих центрированных случайных величин:

μs,h =M(XSYh), где X =X-М(X),

Y=Y-М(Y)

-центрированные случайные величины X и Y.

Основным моментом порядка s, h системы СВ (X,Y) называется нормированный центральный момент порядка s, h:

Начальные моменты α1.0, α0,1

α1.0=M(X1Y0)=M(X); α0.1=M(X0Y1)=M(Y).

Вторые центральные моменты:

μ2,0=M(X2Y0)=M(x-M(X))2=D(X)

- характеризует рассеяние случайных величин в направлении оси ОХ.

μ2,0 = M(X0Y2) = M(y-M(Y))2 = D(Y)

- характеризует рассеяние случайных величин в направлении оси OY.

Особую роль в качестве характеристики совместной вариации случайных величин X и Y играет второй смешанный центральный момент, который называется корреляционным моментом - K(X,Y) или ковариацией –

cov(X,Y): μ1,1=K(X,Y)=cov(X,Y)=M(X1Y1)=M(XY)-M(X)M(Y).

Корреляционный момент является мерой связи случайных величин.

Если случайные величины X и Y независимы, то математическое ожидание равно произведению их математических ожиданий:

М (XY)= М (X) М (Y), отсюда cov(X,Y)=0

## Если ковариация случайных величин не равна нулю, то говорят, что случайные величины коррелированны. Ковариация может принимать значения на всей числовой оси, поэтому в качестве меры связи используют основной момент порядка s=1, h=1 ,который называют коэффициентом корреляции:



Свойства коэффициента корреляции:

1. -1<rху<1.

2. Если r = +1, то случайные величины линейно зависимы;

3. Если rху = 0, то случайные величины некоррелированны, что не означает их независимости вообще.

Замечание. Если случайные величины X и Y подчиняются нормальному закону распределения, то некоррелированность СВ X и Y означает их независимость.

23. Функции случайных величин

Закон распределения функции случайных величин.

Пусть имеется непрерывная случайная величина X с функцией плотности вероятности f(x). Другая случайная величина Y связана со случайной величиной X функциональной зависимостью: Y=φ(X). Случайная точка (X, Y) может находиться только на кривой у=φ(х).

Дифференциальная функция случайной величины Y определяется при условии, что φ(х) - монотонна на интервале (а,b), тогда для функции φ(х) существует обратная функция: φ-1= Ψ, x= Ψ(x).

Обычно, числовая прямая разбивается на n промежутков монотонности и обратная функция находится на каждом из них, поэтому g(y) -дифференциальная функция СВ Y определяется по формуле



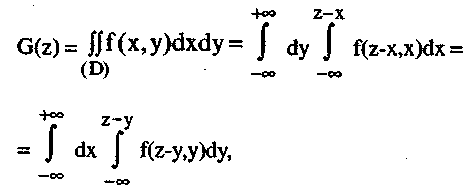
Замечание.

Математическое ожидание и дисперсию СВ Y - функции случайной величины X(Y=φ(x)), имеющей дифференциальную функцию f(x), можно определить по формулам:



24. Композиция законов распределения

В приложениях часто рассматривается вопрос о распределении суммы нескольких случайных величин. Например, пусть Z=X+Y, тогда G(z) -интегральную функцию СВ Z можно определить по формуле



где: f(х,у)-дифференциальная функция системы случайных величин (X,Y);

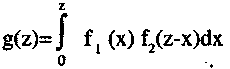
область D - полуплоскость, ограниченная сверху прямой y= z-x.

Отсюда

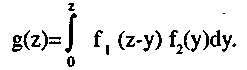
g(z) = G'(z) = ∫f(x, z - x)dx.

Если Х и Y независимы, то говорят о композиции законов распределения случайных величин и дифференциальная функция СВ Z определяется как g(z)=f1 (x) f2(z-x)dx, где f ,(х) и f2(y) дифференциальные функции СВ X и Y соответственно.

Если возможные значения аргументов неотрицательны, то дифференциальную функцию СВ Z определяют по формуле



Или



25. Понятие и виды статистических гипотез.

Статистическая гипотеза – всякое высказывание о генеральной совокупности, проверяемое по выборке. Статистические гипотезы делятся на: 1. параметрические – гипотезы, сформулированные относительно параметров (среднего значения, дисперсии и т.д.) распределения известного вида; 2. непараметрические – гипотезы, сформулированные относительно вида распределения (например, определение по выборке в степени нормальности генеральной совокупности). Процесс использования выборки для проверки гипотезы называется статистическим доказательством. Основную выдвигаемую гипотезу называют нулевой Н0. Наряду с нулевой гипотезой рассматривают ей альтернативную Н1.

26. Выборочный метод

В реальных условиях обычно бывает трудно или экономически нецелесообразно, а иногда и невозможно, исследовать всю совокупность, характеризующую изучаемый признак (генеральную совокупность). Поэтому на практике широко применяется выборочное наблюдение, когда обрабатывается часть генеральной совокупности (выборочная совокупность). Свойства (закон распределения и его параметры) генеральной совокупности неизвестны, поэтому возникает задача их оценки по выборке. Для получения хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была репрезентативной (представительной). Репрезентативность, в силу закона больших чисел, достигается случайностью отбора.

Различают 5 основных типов выборок. 1).Собственно-случайная: а) повторная (элементы после выбора возвращаются обратно); б) бесповторная (выбранные элементы не возвращаются).

2). Типическая - генеральная совокупность предварительно разбивается на группы типических элементов, и выборка осуществляется из каждой. Следует различать: а) равномерные выборки (при равенстве объемов исходных групп в генеральной совокупности выбирается одинаковое количество элементов из каждой); б) пропорциональные (численность выборок формируют пропорционально численностям или средним квадратическим отклонениям групп генеральной совокупности); в) комбинированные (численность выборок пропорциональна и средним квадратическим отклонениям, и численностям групп генеральной совокупности).

3) механическая отбор элементов проводится через определенный интервал.

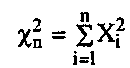
4).Серийная - отбор проводится не по одному элементу, а сериями для проведения сплошного обследования.

5). Комбинированная - используются различные комбинации вышеуказанных методов, например, типическая выборка сочетается с механической и собственно случайной.

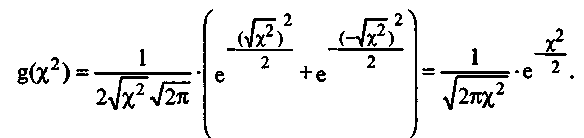
После осуществления выборки возникает задача оценки числовых характеристик генеральной совокупности по элементам выборочной совокупности. Различают точечные и интервальные оценки.

27. Специальные законы распределения

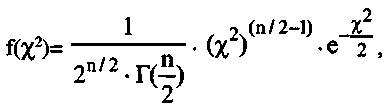
1. х2 -распределение Пирсона. Пусть X1, X2, ...,Хn одинаково распределенные по нормальному закону случайные величины, являющиеся взаимно-независимыми, для которых математическое ожидание равно нулю, а среднеквадратическое отклонение 1, тогда сумма квадратов этих случайных величин носит название случайной величины х2 - xu-квадрат с v=n степенями свободы:



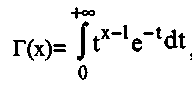
При v=l (учитывая дифференциальная функция:



Дифференциальная функция распределения χ2 с v=n степенями свободы задается формулой



где Г(х) - гамма, функция Эйлера.



при R+; если n Z, то Г(n+ 1)=n!

С возрастанием числа степеней свободы v = n, распределение χ2 медленно приближается к нормальному закону распределения. На практике используют обычно не плотность вероятности, а квантили распределения.

Квантилью χ2n распределения, отвечающей заданному уровню значимости α (альфа) – χ2α,ν , называется такое значение χ2= χ2α,ν, при котором вероятность того, что χ2 превысит значение χ2α,ν, равна α:

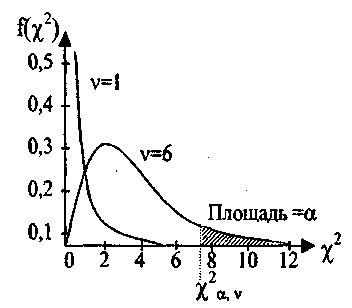
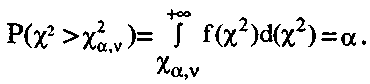


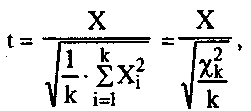
Рис. Дифференциальная функция распределения χ 2 с ν степенями свободы.

С геометрической точки зрения нахождение квантили заключается в выборе такого значения Х2= 5Ca v при котором площадь криволинейной трапеции ограниченной дифференциальной функцией была бы равна а. Значения квантилей затабулированы. При n>30 распределение практически не отличается от нормального.

Замечание. Квантиль СВ X порядка a - это такое значение СВ X, что F(xa) = а, где F(x)=P(X<x). Например, медиана – это квантиль x0.5.

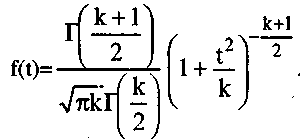
2. t- распределение Стъюдента. Это распределение имеет важное значение при статистических вычислениях, связанных с нормальным законом, распределения, где a - неизвестный параметр распределения и подлежит определению из опытных данных, например, при статистической обработке наблюдений с неизвестной точностью.

Пусть X, X,, X2,...,Xk независимые нормально распределённые случайные величины с нулевыми математическими ожиданиями и одинаковыми дисперсиями. Безразмерная величина



называется дробью Стьюдента.

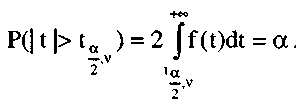
Ее распределение не зависит от а в силу ее безразмерности. Дифференциальная функция t-распределения с v=k степенями свободы имеет вид



t - распределение Стьюдента быстрее, чем х2 стремится к нормальному.

На практике используют квантили распределения в зависимости от числа степеней свободы и уровня значимости α.

С геометрической точки зрения нахождение квантилей (для двусторонней области) заключается в выборе такого значения t, при котором суммарная площадь криволинейной трапеции была бы равна α, в силу симметрии распределения:



F-распределение Фишера-Снедекора.

Пусть Х1, X2, ...,Xm и Y1, Y2, ...,Yn одинаково распределенные по нормальному закону случайные величины, являющиеся взаимно-независимыми, для которых математическое ожидание равно нулю, а среднеквадратическое отклонение равно единице.

Рассмотрим дробь Фишера F(m,n)=(χ2m/m)/(χ2n/n), она имеет F - распределение с v1= m - числом степеней свободы числителя, и v2=n - числом степеней свободы знаменателя ((m, n) степенями свободы), которое называется распределением Фишера-Снедекора. Обычно используют квантили распределения в зависимости от числа степеней свободы (m, n) и уровня значимости а:

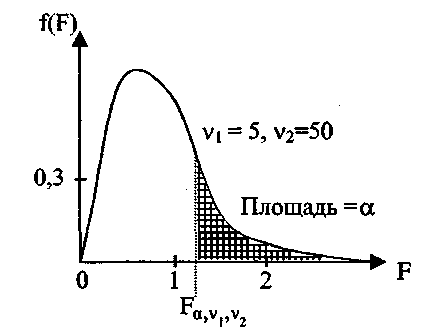
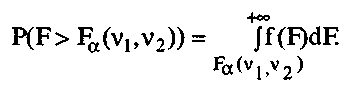
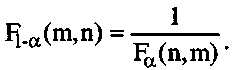


Рис. Дифференциальная функция F распределения Фишера -Снедекора с v1=5, v2=50 степенями свободы

Для квантилей распределения Фишера-Снедекора геометрический смысл аналогичен другим распределениям (рис.23). Имеет место равенство



Распределения χ2 - Пирсона, t - Стьюдента, F -Фишера-Снедекора нашли широкое применение в математической статистике, в частности при проверке статистических гипотез и в дисперсионном анализе.

28. Особенности статистического анализа количественных и качественных показателей

Методы шкалирования при обработке качественных признаков.

Основной задачей статистического анализа является оценка связи признаков м/у собой. Необходимо измерить признаки, в гуманитарных исследованиях более сложны, т.к. они касаются измерения не только количественных, но и качественных признаков.

Суть статистических методов – анализ чисел как таковых, а не истинных значений некоторого признака.

Если количественные показатели можно, то для качественных показателей можно экспертным путем оценить степень сходства или различия м/у парами объектов.

Объекты отражают в некотором многомерном пространстве, где каждая точка – это объект, а координаты – признаки.

Для этого используют методы многомерного шкалирования.

- матрица парных расстояний (количественный признак)

- матрица парных отклонений (качественный признак)

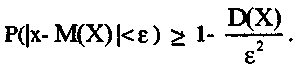
По матрицам изучается степень сходства и различия.

29. Неравенство Чебышева

Рассмотрим закон больших чисел в форме Чебышева.

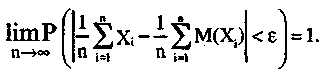
Лемма Чебышева (Маркова). Если случайная величина X принимает только неотрицательные значения и имеет математическое ожидание М(Х), то для любого α>0 имеет место неравенство: P(X≥α)≤(M(X))/α.

Неравенство Чебышева. Если случайная величина X имеет математическое ожидание М(Х) и дисперсию D(X), то для любого ε>0 имеет место неравенство:



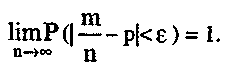
Неравенство Чебышева является в теории вероятностей общим фактом и позволяет оценить нижнюю границу вероятности.

Теорема. Закон больших чисел Чебышева. Пусть Х1, Х2, .. .,Хn - последовательность попарно независимых случайных величин, имеющих конечные математические ожидания и дисперсии, ограниченные сверху постоянной С = const (D(Xi)≤C(i=l, 2,...,n)). Тогда для любого ε>0,



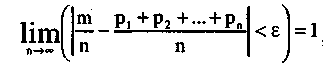
Теорема показывает, что среднее арифметическое большого числа случайных величин с вероятностью сколь угодно близкой к 1 будет мало отклоняться от среднего арифметического математических ожиданий.

Следствие 1. Если вероятность наступления события А в каждом из n независимых испытаний равна р, m - число наступлений события А в серии из n независимых испытаний, то, каково бы ни было число е > 0, имеет место предел:



Таким образом устанавливается связь между относительной частотой появления события А и постоянной вероятностью р в серии из n независимых испытаний.

Следствие 2. Теорема Пуассона. Если в последовательности независимых испытаний вероятность появления события А в к-ом испытании равна р, то

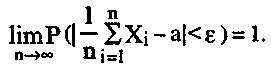


где m - число появлений события А в серии из n испытаний.

Следствие 3. Теорема Бернулли. Если X1, Х2,.. .,Хn - последовательность независимых случайных величин таких, что

М(Х1) = М(Х2)=...= М(Хn) = а, D(Х1)< С, D(X2) < С,.. .,D(Xn)< С, где С = const

то, каково бы ни было постоянное число ε>0, имеет место предел:



Этот частный случай закона больших чисел позволяет обосновать правило средней арифметической.

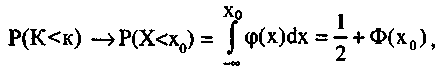
Законы больших чисел не позволяют уменьшить неопределённость в каждом конкретном случае, они утверждают лишь о существовании закономерности при достаточно большом числе опытов. Например, если при подбрасывании монеты 10 раз появился герб, то это не означает, что в 11 раз появится цифра.

30. Центральная предельная теорема

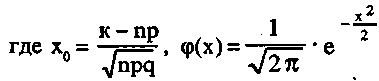
В теории вероятностей и математической статистике большое значение имеет центральная предельная теорема Ляпунова, в которой утверждается, что если сложить большое число случайных величин, имеющих один или различные законы распределения, то случайная величина, являющаяся результатом суммы, при некоторых условиях, будет иметь нормальный закон распределения.

Примером центральной предельной теоремы (для последовательности независимых случайных величин) является интегральная теорема Муавра-Лапласа.

Теорема 1. Пусть производится n независимых опытов в каждом из которых вероятность наступления события А равна р (не наступления q=l-p, p≠0, р≠1). Если К - число появлений события А в серии из n испытаний, то при достаточно больших n СВ К можно считать нормально распределенной (М(К)=nр, σ(К)=√D(K)= √npq).

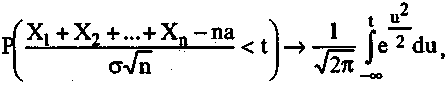


,Ф(x0) – функция Лапласа.



В более общем случае верна следующая теорема.

Теорема 2. Если случайные величины X1, Х2... Хn независимы, одинаково распределены и имеют конечную дисперсию, то при n→∞:



где М(Х)=а, σ2=D(Х); U - нормально распределенная случайная величина, M(U)=0,D(U)=1.

31. Определение вариационных рядов. Графическое изображение вариационных рядов

В реальных социально-экономических системах нельзя проводить эксперименты, поэтому данные обычно представляют собой пассивные наблюдения за происходящим процессом, например: курс валюты на бирже в течение месяца, урожайность пшеницы в хозяйстве за 30 лет, производительность труда рабочих за смену и т.д. Результаты наблюдений, в общем случае, ряд чисел, расположенных в беспорядке, который для изучения необходимо упорядочить (проранжировать).

Операция, заключенная в расположении значений признака по не убыванию, называется ранжированием опытных данных.

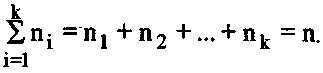
После операции ранжирования опытные данные можно сгруппировать так, чтобы в каждой группе признак принимал одно и то же значение, которое называется вариантом (хi). Число элементов в каждой группе называется частотой варианта (ni).

Размахом выборки называется число

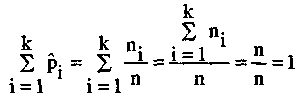
W=xmax- х min,

где xmax - наибольший вариант, х min - наименьший вариант.

Сумма всех частот равна определенному числу n, которое называется объемом совокупности:



Отношение частоты данного варианта к объему совокупности называется относительной частотой (pi) или частостью этого варианта: pi=ni/n.



Последовательность вариант, расположенных в возрастающем порядке, называется вариационным рядом (вариация - изменение).

Вариационные ряды бывают дискретными и непрерывными. Дискретным вариационным рядом называется ранжированная последовательность вариант с соответствующими частотами и (или) частостями.

Построение дискретного вариационного ряда нецелесообразно, если число значений признака велико или признак является непрерывным, то есть может принимать любые значения в пределах некоторого интервала. В этом случае следует построить интервальный вариационный ряд. Для построения такого ряда промежуток изменения признака разбивается на ряд отдельных интервалов и подсчитывается количество значений величины в каждом из них.

Будем считать, что отдельные (частичные) интервалы имеют одну и ту же длину. Число интервалов (k), в случае нормально распределённой совокупности, можно определить по формуле Стерджесса k = l +3,3221g n. или приближённо: k [6;12]. Длина частичного интервала определяется по формуле



Графическое изображение вариационных рядов.

Вариационные ряды изображают графически с помощью полигона и гистограммы.

Полигон частот - это ломаная, отрезки которой соединяют точки (x1; n1), (x2;n2),... (хk;nk).

Полигон относительных частот - это ломаная, отрезки которой соединяют точки



Гистограммой частот называется фигура, состоящая из прямоугольников с основанием h и высотами ni. Для гистограммы относительных частот в качестве высоты рассматривают ni/n. Гистограмма относительных частот является аналогом дифференциальной функции случайной величины.

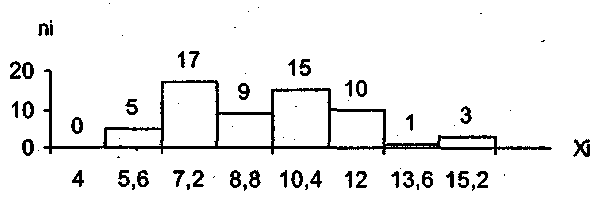


Рис. Гистограмма частот

32. Общие модели статистического анализа

Характеристика методов многомерного анализа, (компонентный анализ, факторный анализ, кластер-анализ(классификация без обучения). Дискриминантный анализ (классификация с обучением. Канонические корреляции. Множественный ковариационный анализ).

Реальные процессы зависят от параметров, их характеристик, поэтому возникает необходимость в применении мер, методов статистического анализа.

Методы МСА следует рассматривать, как логическое продолжение методов ТВ и МС. Принципиальное различие состоит в учете более 3-х факторов.

Методы МСА базируются на представлении информации в многомерном пространстве и позволяют определить латентные зак-ти, сущ-ие объективно.

Методы:

- моделирования и первичной обработки данных

- анализа и построения зависимости

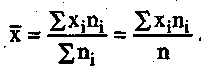
- классификация и снижение зависимости размерности

33. Средняя арифметическая ряда

Вариационные ряды позволяют получить первое представление об изучаемом распределении. Далее необходимо исследовать числовые характеристики распределения (аналогичные характеристикам распределения теории вероятностей): характеристики положения (средняя арифметическая, мода, медиана); характеристики рассеяния (дисперсия, среднее квадратическое отклонение, коэффициент вариации); характеристики меры скошенности (коэффициент асимметрии) и островершинности (эксцесс) распределения.

Средней арифметической (х) дискретного вариационного ряда называется отношение суммы произведений вариантов на соответствующие частоты к объему совокупности:

(3.2.1)



Модой (М\*(Х)) дискретного вариационного ряда называется вариант, имеющий наибольшую частоту.

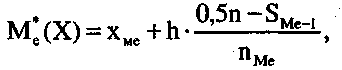
Медианой (М\*(Х)) дискретного вариационного ряда называется вариант, делящий ряд на две равные части. Если дискретный вариационный ряд имеет 2n членов: x1, x2, ..., xn, xn+1, ... x2n, то

Ме\*(Х)=(xn+xn+1)/2.

Если дискретный вариационный ряд имеет 2n+1 членов: x1, x2, ..., xn-1, xn, xn+1, ... x2n+1, то

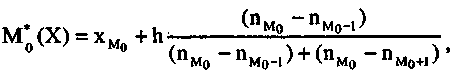
М\*e(Х)=xn+1.

Для интервальных вариационных рядов (с равными интервалами для медианы и моды) имеют место формулы: а) медианы



где хМе - начало медианного интервала, h - длина частичного интервала, n - объем совокупности, SMe-i, - накопленная частота интервала, предшествующего медианному, nМе -частота медианного интервала;

б) моды



где хМо- начало модального интервала, h -длина частичного интервала, nмо - частота модального интервала, nМо-1 -частота предмодального интервала, nМо+1 - частота послемодального интервала;

в) средней арифметической, совпадающей с формулой (3.2.1) для дискретного вариационного ряда, причем в качестве вариант хi принимаются середины соответствующих интервалов (интервалы могут иметь как одинаковую, так и разную длину).

Мода и медиана используются в качестве характеристики среднего положения в случае, если границы ряда нечеткие или если ряд не симметричен.

34. Проблема размерностей в многомерных методах исследования

Метод МСА базируется на представлении данных в многомерном признаковом пространстве размерностью, равной числу признака. При этом исследователь часто сталкивается с понятием размерности.

В общем случае изучается n-мерное эвклидово пространство. При n>3 все задачи решаются только логически и алгебраически (n>>m) (m=2-3). Для этого обычно стараются снизить размерность изучаемого пространства без видимых потерь информации.

Основные предпосылки перехода к производству меньшей размерности.

1. дублирование информации
2. ненормативность признаков
3. возможность агрегирования (простого или взвешенного суммирования)

Основной минус МСА: статистические методы оценивания и сравнения основываются только на многомерном нормальном законе раск-ния.

35. Введение в Excel

Табулирование – вычисление значений функций, при известных значениях аргумента.

БД – это фактически любой набор данных. Создание баз данных упрощает обработку данных и их анализ.

Группировка – разбиение на группы, удовлетворяющие определенным критериям

Можно для облегчения работы с данными использовать Пакет анализа содержащий 13 категорий функций:

Финансовые (51 функция)

Дата и время (19 функций)

Математические (60)

Пользовательские (11-при сложных вычислениях)

Логические (6)

Статистические (самая объемная - 78)

Ссылки и массивы

Информационные и тд.

36. Современные пакеты прикладных программ МС исследования. Пакет статистика. Стандарт качества ISO 9000. Система SEWS применение многомерных статистических методов в социально экономических исследованиях

За 200 лет математиками, экономистами, психологами был создан аппарат принятия решений, которых называется МС, а позже прикладной С или анализом данных

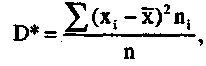
Широкому внедрению методов анализа данных в 60-70гг. способствовало появление компонентов, причем если до середины 80г. Эти методы рассматривались, как инструмент научных исследований, то теперь основными показателями стали компоненты организации и тд.

Пакет statistika версия stat 5.5 русскоязычная поддержка всех архитектур документация 3000с.

1. иногда слишком поверхностны
2. неудобный редактор отсчета
3. высокая стоимость

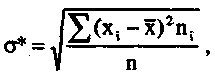
37. Дисперсия дискретного ряда

Дисперсия дискретного ряда распределения:



характеризует средний квадрат отклонения х от х---,

Среднее квадратическое отклонение дискретного ряда распределения:



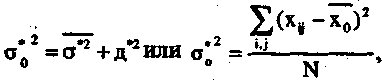
выражается в тех же единицах, что и хi.

Коэффициент вариации:



характеризует относительное значение среднего квадратического отклонения и обычно служит для сравнения колеблемости несоизмеримых показателей.

Если объединяются несколько распределений в одно, то общая дисперсия σ0\*2 нового распределения равна средней арифметической из дисперсий объединяемых распределений, сложенной с дисперсией частных средних относительно общей средней нового распределения:

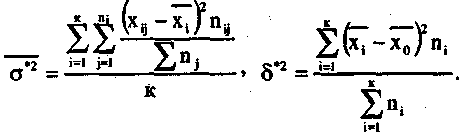


где x0-- - средняя ариф-кая нового распределения, xi-- - средняя ариф-кая i–го частного распределения (I=1,…,k).

n - объем i-гo частного распределения, хij - j-й член i-го частного распределения (j=l,..., ni; i=l,2,..., к), δ\*2 -

межгрупповая дисперсия, --σ\*2 - внутригрупповая дисперсия, N=∑ni - объем нового распределения.

Значения --σ\*2 и δ\*2 определяются по формулам

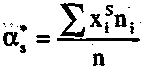


Дисперсия имеет важное свойство, заключающееся в том, что

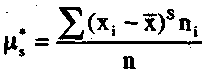
D\*=(∑(xi-d)2ni)/k принимает наименьшее значение при d=--x.

38. Моменты для вариационных рядов в математической статистике находятся по формулам, аналогичным формулам (2.7.6), (2.7.7)>(2.7.11), (2.10.3):

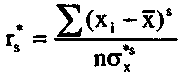
- начальный момент s–го порядка,



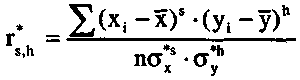
- центральный момент s–го порядка.



- основной момент s-гo порядка



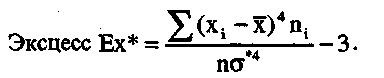
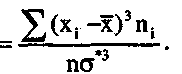
- основной момент порядка s, h.



Соотношения между начальными и центральными моментами в математической статистике соответствуют формулам (2.7.8).

Коэффициент асимметрии

Sk\*=



39. Проверка адекватности модели регрессии

После построения уровня регрессии возникает вопрос о качестве решения.

Пусть при исследовании n пар наблюдений (хi, уi) получено уравнение регрессии У на Х.

‾yi = a + bxi

Рассмотрим тождество:

yi - ‾yi = yi - ‾yi – (‾yi -‾yi)

Если переписать это уравнение в виде

(yi-‾y) = (‾yi-‾y) + (yi-‾y)

возвести обе части в квадрат и просуммировать по i, то получим

Σ(yi-‾y)2 =Σ (‾yi-‾y)2 + Σ(yi-‾y)2 (\*)

Уравнение (\*) является основополагающим в дисперсионном анализе.

Для сумм обычно вводятся названия:

Σyi2 – нескорректированная сумма квадратов У-ков;

- коррекция на среднее суммы квадратов У-ков.

-сумма квадратов отношений относительно среднего наблюдений.

Σ (‾yi-‾y)2- сумма квадратов относительно регрессии.

Σ(yi-‾yi)2 – сумма квадратов обусловленная регрессией.

40. Интервальные оценки. Доверительная вероятность, доверительный интервал

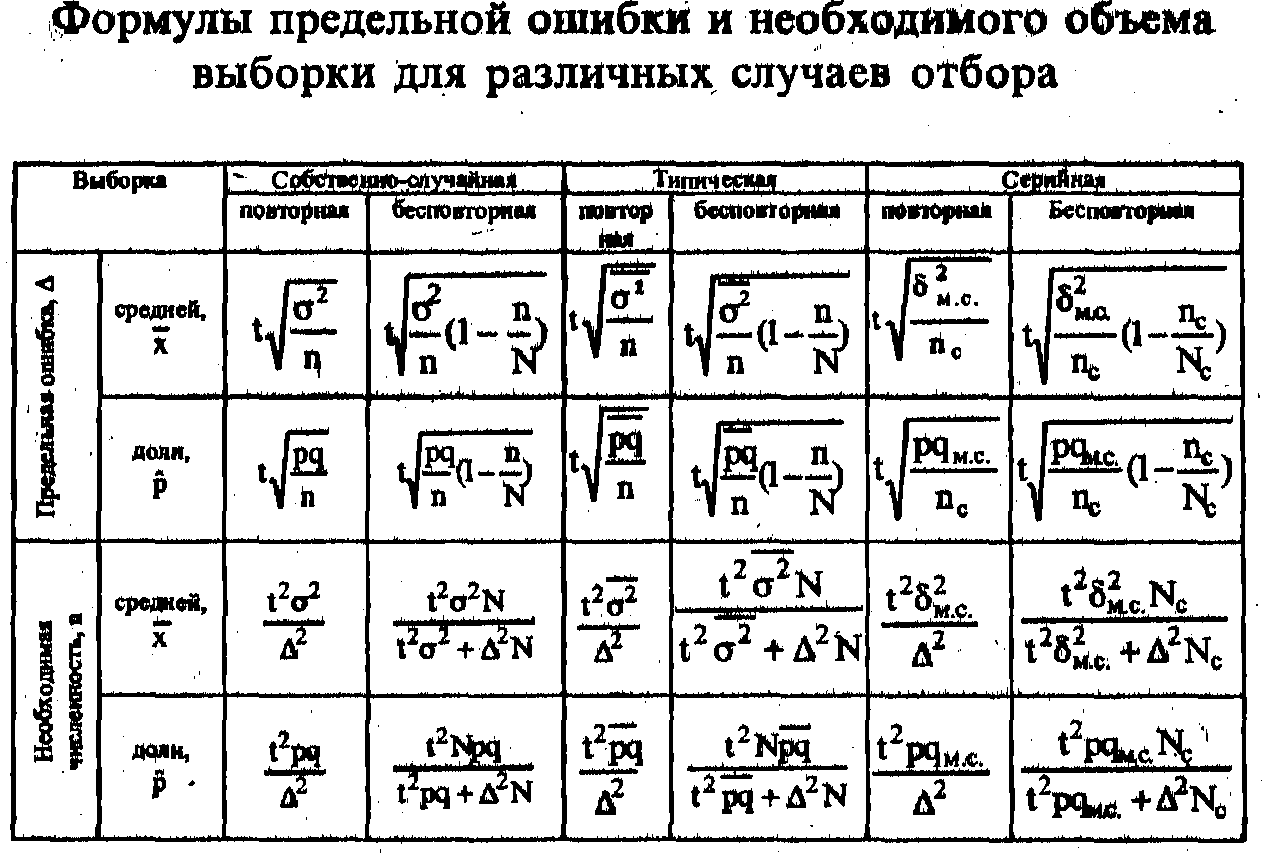
Интервальной называют оценку, которая определяется 2 числами – границами интервала. Она позволяет ответить на вопрос: внутри какого интервала и с какой вероятностью находится неизвестное значение оцениваемого параметра генеральной совокупности. Пусть θ точечная оценка параметра θ. Чем меньше разность θ - θ , тем точнее и лучше оценка. Обычно говорят о доверительной вероятности p = 1-α, с которой θ будет находиться в интервале θ-Δ < θ < θ+Δ, где: Δ (Δ > 0) – предельная ошибка выборки, которая может быть либо задана наперед, либо вычислена; α - риск или уровень значимости (вероятность того, что неравенство будет неверным). В качестве 1-α принимают значения 0,90;0,95;0,99;0,999. Доверительная вероятность показывает, что в (1-α) 100% случаев оценка будет накрываться указанным интервалом. Для построения доверительного интервала параметра а – математического ожидания нормального распределения, составляют выборочную характеристику (статистику), функционально зависимую от наблюдений и связанную с а, например, для повторного отбора:

Статистика u распределена по нормальному закону распределения с математическим ожиданием а = 0 и средним квадратическим отклонением σ = 1. Отсюда

P(|u|<u α/2)= 1-σ или 2Ф(uα/2)=1-σ,

где Ф-функция Лапласа, uα/2 – квантиль нормального закона распределения, соответствующая уровню значимости α.

Определение доверительного интервала для средней и доли при случайном обороте. Определение доверительного интервала для средней и доли при типическом обороте;. Определение необходимой численности выборки. Распространение данных выборки на генеральную совокупность).



Где:

1) t— квантиль распределения соответствующая уровню значимости:



а) при n 30 t=- квантиль нормального закона распре деления,



б) при n<30t - квантиль распределения Стьюдента с v=n-1 степенями свободы для двусторонней области;

2) - выборочная дисперсия:



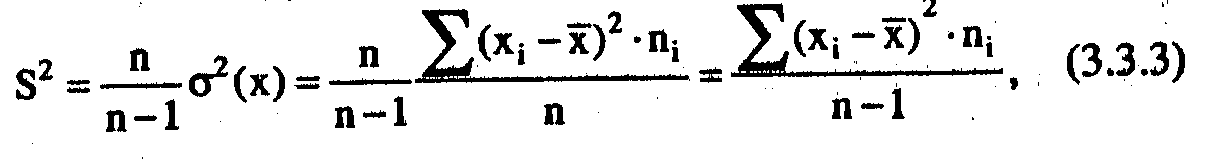
а) при n30 можно считать, что



б) при n<30 вместо берут исправленную выборочную дисперсию



S2 ()



далее везде рассматривается исправленная выборочная дисперсия S2;

З) рq — дисперсия относительной частоты в схеме повторных независимых испытаний;

4) N — объем генеральной совокупности;

5) n — объем выборки;

6) — средняя арифметическая групповых дисперсий (внутригрупповая дисперсия);



7) — средняя арифметическая дисперсий групповых долей,



8) — межсерийная дисперсия,



9) pqм.с. — межсерийная дисперсия доли;

10) Nc — число серий в генеральной совокупности;

11) nc — число отобранных серий (объем выборки);

12) — предельная ошибка выборки.



41. Статистические критерии проверки гипотез, уровень значимости и мощность критерия. Выбор м/у гипотезами Н0 и Н1 может сопровождаться ошибками 2 родов. Ошибка первого рода α означает вероятность принятия Н1, если верна гипотеза

Н0: α=Р(Н1/Н0)

Ошибка второго рода β означает вероятность принятия Н0 если верна гипотеза

Н1: β=Р(Н0/Н1)

Существует правильное решение двух видов

Р(Н0/Н0) = 1-α и Р(Н1/Н1)=1-β.

Правило, по которому принимается решение о том, что верна или неверна гипотеза Н0 называется критерием, где:

α=Р(Н1/Н0)

уровень значимости критерия;

М= Р(Н1/Н1)=1-β

мощность критерия. Статистический критерий К – случайная величина, с помощью которой принимают решение о принятии или отклонении Н0.

42. Концепция Data Mining

Data Mining переводится как "добыча" или "раскопка данных". Нередко рядом с Data Mining встречаются слова "обнаружение знаний в базах данных" (knowledge discovery in databases) и "интеллектуальный анализ данных". Их можно считать синонимами Data Mining. Возникновение всех указанных терминов связано с новым витком в развитии средств и методов обработки данных. Традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, откровенно спасовала перед лицом возникших проблем. Главная причина — концепция усреднения по выборке, приводящая к операциям над фиктивными величинами (типа средней температуры пациентов по больнице, средней высоты дома на улице, состоящей из дворцов и лачуг и т.п.). Методы математической статистики оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для “грубого” разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical processing, OLAP). В основу современной технологии Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих фрагменты многоаспектных взаимоотношений в данных. Эти шаблоны представляют собой закономерности, свойственные подвыборкам данных, которые могут быть компактно выражены в понятной человеку форме. Поиск шаблонов производится методами, не ограниченными рамками априорных предположений о структуре выборке и виде распределений значений анализируемых показателей.

43. Понятие корреляционной зависимости

При изучении случайных величин в общем случае необходимо рассматривать стохастическую зависимость, когда каждому значению СВ Х может соответствовать одно и более значений СВ Y, причем до опыта нельзя предсказать возможное соответствие. В случае стохастической связи изменение CВY, вследствие изменения СВ Х, можно разбить на 2 компоненты: 1. функциональную, связанную с зависимостью Y от Х, 2. случайную, связанную со случайным характером самих СВ Х и Y. Соотношение м/у функциональной и случайной компонентой определяет силу связи. Отсутствие первой компоненты указывает на независимость СВ Х и Y, отсутствие второй компоненты показывает, что м/у CВ X и Y существует функциональная связь.

Важным частным случаем стохастической зависимость является корреляционная. Корреляционная зависимость м/у переменными величинами – это та функциональная зависимость, которая существует м/у значениями одной из них и групповыми средними другой. (Корреляционные зависимости Y на Х и Х на Y обычно не совпадают). Корреляционная связь чаще всего характеризуется выборочным коэффициентом корреляции r, который характеризует степень линейной функциональной зависимости м/у CB X и Y. Для двух СВ Х и Y коэффициент корреляции имеет => св-ва:

1. -1≤r≤1;
2. если r=+ 1, то м/у СВ Х и Y существует функциональная линейная зависимость;
3. если r=0, то СВ Х и Y некоррелированны, что не означает независимости вообще;
4. если Х и Y образуют систему нормально распределенных СВ, то из их некоррелированности => их независимость.

Коэффициенты корреляции Y на Х и Х на Y совпадают.

Корреляция используется для количественной оценки взаимосвязи двух наборов данных с помощью коэффициента корреляции. Коэффициент корреляции выборки представляет собой ковариацию двух наборов данных, деленную на произведение их стандартных отклонений.

44. Критерий согласия

Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т. е. при помощи специально подобранной случайной величины — критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: χ2 («хи квадрат») К. Пирсона, Колмогорова, Смирнова и др.

Ограничимся описанием применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений, в этом состоит его достоинство). С этой целью будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные в предположении нормального распределения) частоты. Обычно эмпирические и теоретические частоты различаются.

Случайно ли расхождение частот? Возможно, что расхождение случайно и объясняется малым числом; наблюдений, либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены, исходя из неверной гипотезы о нормальном распределении генеральной совокупности. Критерий Пирсона отвечает на поставленный выше вопрос. Правда, как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает, на принятом уровне значимости, ее согласие или несогласие с данными наблюдений.

Итак, пусть по выборке объема п получено эмпирическое распределение:

варианты xl, x1, x2 ... xs,

эмп. частоты ni n1 п2 ... ns.

Допустим, что в предположении нормального распределения генеральной совокупности, вычислены теоретические частоты п. При уровне значимости α, требуется проверить нулевую гипотезу; генеральная совокупность распределена нормально.

В качестве критерия проверки нулевой гипотезы примем случайную величину

(\*)



Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения. Ясно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия (\*) и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределений.

Заметим, что возведением в квадрат разностей частот устраняют возможность взаимного погашения положительных и отрицательных разностей. Делением на n’i достигают уменьшения каждого из слагаемых; в противном случае сумма была бы настолько велика, что приводила бы к отклонению нулевой гипотезы даже и тогда, когда она справедлива. Разумеется, приведенные соображения не являются обоснованием выбранного критерия, а лишь пояснением.

Доказано, что при n→∞ закон распределения случайной величины (\*), независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ2 с k степенями свободы. Поэтому случайная величина (\*) обозначена через χ2, а сам критерий называют критерием согласия «хи квадрат».

Число степеней свободы находят по равенству

k=s-1-r

где s — число групп выборки; r — число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение — нормальное, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение) поэтому r=2 и число степеней свободы

k=s-1-r=s-1-2-s-3.

Если, например, предполагают, что генеральная совокупность распределена по закону Пуассона, то оценивают один параметр X, поэтому r=1 и k=s-2.

Поскольку односторонний критерий более «жестко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α:



Т.о., правосторонняя критическая область определяется неравенством



а область принятия нулевой гипотезы — неравенством



Обозначим значение критерия, вычисленное по данным наблюдений, через χ2набл и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы, при заданном уровне значимости, проверить нулевую гипотезу H0: генеральная совокупность распределена нормально, надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия

(\*\*)



и по таблице критических точек распределения χ2, по заданному уровню значимости α, и числу степеней свободы k=s-3, найти критическую точку χ2 (α; k).

Если χ2набл<χ2кр – нет оснований отвергнуть нулевую гипотезу.

Если χ2набл >χ2кр — нулевую гипотезу отвергают.

Замечание 1. Объем выборки должен быть достаточно велик, во всяком случае не менее 50. Каждая группа должна содержать не менее 5—8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Замечание 2. Поскольку возможны ошибки первого и второго рода, в особенности, если согласование теоретических и эмпирических частот «слишком хорошее», следует проявлять осторожность.

Замечание 3. В целях контроля вычислений, формулу (\*\*) преобразуют к виду



45. Понятие и модели дисперсионного анализа

Дисперсионный анализ позволяет ответить на вопрос о наличии существенного влияния некоторых факторов на изменчивость фактора, значения которого могут быть получены в результате опыта. При проверке статистических гипотез предполагается случайность вариации изучаемых факторов. В дисперсионном анализе один или несколько факторов изменяются заданным образом, причем, эти изменения могут влиять на результаты наблюдений. Исследование такого влияния и является целью дисперсионного анализа.

Идея дисперсионного анализа заключается в том, что основная дисперсия разлагается в сумму составляющих ее дисперсий, каждое слагаемое которой соответствует действию определенного источника изменчивости. Например, в двухфакторном анализе мы получим разложение вида:

σС2=σА2+σВ2+σАВ2+δZ’2,

где

σС2 –общая дисперсия изучаемого признака С

σА2 –доля дисперсии, вызванная влиянием фактора А

σВ2 - доля дисперсии, вызванная влиянием фактора В

σАВ2 - доля дисперсии, вызванная взаимодействием факторов А и В

δZ’2 –доля дисперсии, вызванная неучтенными случайными причинами (случайная дисперсия).

В дисперсионном анализе рассматривается гипотеза: Н0 – ни один из рассматриваемых факторов не оказывает влияния на изменчивость признака. Значимость каждой из оценок дисперсии проверяется по величине ее отношения к оценке случайной дисперсии и сравнивается с соответствующим критическим значением, при уровне значимости α, с помощью таблиц критических значений F – распределения Фишера-Снедекора. Гипотеза Н0 относительно того или иного источника изменчивости отвергается, если Fрасч.> Fкр.

В дисперсионном анализе рассматриваются эксперименты трех видов:

А) эксперименты, в которых все факторы имеют систематические (фиксированные) уровни;

Б) эксперименты, в которых все факторы имеют случайные уровни;

В) эксперименты, в которых есть факторы, имеющие случайные уровни, а так же факторы, имеющие фиксированные уровни.

Все три случая соответствует трем моделям, которые рассматриваются в дисперсионном анализе.

Однофакторный дисперсионный анализ.

Рассмотрим единичный фактор, который принимает р различных уровней, и предположим, что на каждом уровне сделано n наблюдений, что дает N = np наблюдений. (все факторы имеют фиксированные уровни)

Пусть результаты представлены в виде Хij (i=1,2...,p; j=1,2...,n).

Предполагается, что доля каждого уровня n наблюдений имеется средняя, которая равна сумме общей средней и ее вариации обусловленной выбранным уровнем:

Xij = μ + Ai + εij,

где μ - общая средняя;

Ai – эффект, обусловленный i-м уровнем фактора;

εij – вариация результатов внутри отдельного уровня фактора. С помощью члена εij принимаются в расчет все неконтролируемые факторы.

Пусть наблюдения на фиксированном уровне фактора нормально распределены относительно среднего значения μ + Ai с общей дисперсией σ2.

Тогда (точка вместо индекса обозначает усреднение соответствующих наблюдений по этому индексу):

Xij – X.. = (Xi. – X..) + (Xij – Xi.).

Иначе первую формулу можно записать: S = S1 + S2. Величина S1 вычисляется по отклонениям р средних от общей средней X.. , поэтому S1 имеет (р-1) степеней свободы. Величина S2 вычисляется по отклонениям N наблюдений от р выборочных средних и, следовательно, имеет N – р = np – p = p(n - 1) степеней свободы. S имеет (N -1) степеней свободы.

Если гипотеза о том, что влияние всех уровней одинаково, справедлива, то обе величины М1 и М2 будут несмещенными оценками σ2. Значит, гипотезу можно проверить, вычислив отношение (М1/М2) и сравнив его с Fкр. с ν1= (р-1) и ν2= (N – р) степенями свободы.

Если Fрасч.> Fкр. , то гипотеза о незначимом влиянии фактора А на результат наблюдений не принимается.

Многофакторный дисперсионный анализ. Дисперсионный анализ в Excel.

Дисперсионный анализ позволяет ответить на вопрос о наличии существенного влияния некоторых факторов на изменчивость фактора, значение которого могут быть получены в результате опыта. При проверке статистических гипотез предполагается случайность вариации изучаемых факторов. В дисперсионном анализе один или несколько факторов изменяются заданным образом, причем, эти изменения могут влиять на результаты наблюдений. Исследование такого влияния и является целью дисперсионного анализа. Идея дисперсионного анализа заключается в том, что основная дисперсия разлагается на сумму составляющих ее дисперсий, каждое слагаемое которой соответствует действию определенного источника изменчивости. Например, в двухфакторном анализе мы получим разложение вида:

σC2=σA2 + σB2 + σAB2 + δZ2

σC2 – общая дисперсия изучаемого признака С;

σA2 – доля дисперсии, вызванная влиянием фактора А;

σB2 – доля дисперсии, вызванная влиянием фактора В;

σAB2 – доля дисперсии, вызванная взаимодействием факторов А и В;

δZ2 – доля дисперсии, вызванная неучтенными случайными причинами (случайная дисперсия);

В дисперсионном анализе рассматривается гипотеза Н0 – и один из рассматриваемых факторов не оказывает влияния на изменчивость признака. Значимость каждой из оценок дисперсии проверяется по величине ее отношения к оценке случайной дисперсии и сравнивается с соответствующим критическим значением, при уровне значимости α, с помощью таблиц критических значений F-распределения Фишера-Снедекора. Гипотеза Н0 относительно того или иного источника изменчивости отвергается, если Fрасч>Fкр. В дисперсионном анализе рассматриваются эксперименты 3 видов:

1. эксперименты, в которых все факторы имеют систематические (фиксированные) уровни;

2. эксперименты, в которых все факторы имею случайные уровни;

3. эксперименты, в которых есть факторы, имеющие случайные уровни, а так же факторы, имеющие случайные уровни.

Двухфакторный дисперсионный анализ с повторениями представляет собой более сложный вариант однофакторного анализа, включающего более чем одну выборку для каждой группы данных. Двухфакторный дисперсионный анализ позволяет статистически обосновать существенность влияния факторных признаков А и В взаимодействия факторов (А и В) на результативный фактор F.

Двухфакторный дисперсионный анализ без повторений позволяет оценить существенность воздействия факторов А и В на результирующий фактор без учета воздействия взаимодействии факторов А и В.

46. Оценка методом наименьших квадратов коэффициентов регрессии

Регрессионный анализ – один из основных методов современной мат статистики. Корреляционный анализ позволяет установить существует или не существует зависимость м/у парами наблюдений, то регрессионный анализ дает целый арсенал методов построения соответствующих зависимостей. Классическим методом оценивания коэффициентов уравнения регрессии является метод наименьших квадратов (МНК).

На основании известных n пар наблюдений (xi, yi) делается предположение о виде зависимости, например:

y=a+bx,

где y – зависимая (результативная) переменная, х – независимая (факторная) переменная.

Пусть переменная x задается точно (без ошибок), тогда отклонение наблюдений yi от зависимости y=a+bx является случайным и параметры a и b можно найти из условия минимизации суммы квадратов ошибок

εi=yi–a–bxi

S= Σεi2→ min,

S= Σ( yi–a–bxi)2→ min,

Эта система носит название системы нормальных уравнений Гаусса, т.к. получена из условия минимизации суммы квадратов отклонении, в предположении, что xi – фиксированы, т.е. отклонения перпендикулярны оси ОХ.