



Финансовый университет
при Правительстве Российской Федерации

Дистанционное обучение



Финансовый университет
при Правительстве Российской Федерации

Шевелёв

Александр Юрьевич
доцент, кандидат физико-
математических наук.



Финансовый университет
при Правительстве Российской Федерации

Математика



Финансовый университет
при Правительстве Российской Федерации

Тема №13.
Элементы математической
статистики



Математическая статистика

Предметом математической статистики является изучение совокупности однородных объектов относительно некоторого количественного или качественного признака, характеризующего эти объекты по результатам наблюдений.



Математическая статистика

Наблюдения могут заключаться либо в измерении какого-нибудь параметра исследуемого объекта, либо в регистрации у него того или иного признака. В общем случае измеряемых параметров или регистрируемых признаков может быть несколько. При этом наблюдения могут производиться как над самими объектами, так и над их моделями.

Математическая статистика



К числу наиболее часто встречающихся задач математической статистики относятся:

1. Определение по результатам независимых наблюдений частоты наступления случайного события и оценка на этой основе его вероятности;
2. Оценка законов распределения случайных величин по результатам наблюдений;
3. Определение неизвестных значений числовых характеристик случайных величин, оценка их точности и надёжности;

Математическая статистика



4. Проверка статистических гипотез о виде закона распределения или его числовых характеристиках;
5. Оценка степени взаимосвязи между несколькими характеристиками исследуемых объектов (корреляция).



Математическая статистика

В практике статистических наблюдений различают два вида: *сплошное*, когда изучаются все объекты и *выборочное*, когда изучается часть объектов (*выборочный метод*).

Генеральной совокупностью называют множество всех объектов над которыми необходимо произвести наблюдение.

Выборочной совокупностью (выборкой) называется та часть генеральной совокупности, которая отобрана для непосредственного изучения.



Число объектов в совокупности называется её объёмом. N – объём генеральной совокупности, n – объём выборки.

Суть выборочного метода в том, чтобы по выборке можно было бы делать выводы о тех же свойствах генеральной совокупности.





Математическая статистика

Чтобы по выборке можно было уверенно судить об изучаемой случайной величине выборка должна быть *собственно-случайной*: любой объект генеральной совокупности может быть с одинаковой вероятностью отобран в выборку.



Математическая статистика

При этом возможны два способа образования выборки: повторная и бесповторная.

Повторной называют выборку, при которой случайно отобранный и обследованный объект возвращается в генеральную совокупность и после этого снова может быть отобран в выборку.

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



Математическая статистика

Накопленные в процессе исследования или эксперимента данные сначала подвергают сортировке: *ранжируют* (упорядочение в порядке возрастания или убывания), затем *группируют* (в каждой группе возможные значения случайной величины одинаковы).

Различные возможные значения случайной величины, соответствующие отдельной группе сгруппированного ряда наблюдаемых данных называются *вариантами*.

Численность отдельной группы сгруппированного ряда наблюдаемых данных называется *частотой* варианта.

Отношение частоты данного варианта к объёму совокупности называется *долей* (относительной частотой) варианта.





Математическая статистика

Частоты и доли вариантов обобщённо называются весами.

Сумма частот равна объёму совокупности, а сумма долей равна единице.

Ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им весами называется дискретным *вариационным рядом*.

Обычно представляется в виде таблицы.



Математическая статистика



X	x_1	x_2	\dots	x_m	Σ
n_i	n_1	n_2	\dots	n_m	n

X	x_1	x_2	\dots	x_m	Σ
w_i	w_1	w_2	\dots	w_m	1



Математическая статистика

Если изучаемая случайная величина является непрерывной, то строится интервальный вариационный ряд.

Длины интервалов называются *интервальными разностями*. В нашем случае для удобства расчётов будем брать ряды с одинаковыми интервальными разностями и затем заменять интервальный ряд дискретным, в котором в качестве варианта принимается середина интервала.



Математическая статистика

Для наглядности интервальный вариационный ряд можно изобразить в прямоугольной системе координат в виде *гистограммы*, которая представляет собой ступенчатую фигуру, состоящую из прямоугольников, основаниями которых на оси абсцисс являются интервалы значений признака, а высоты равны соответствующим им частотам или долям (на оси ординат).



Математическая статистика

Полигоном частот или относительных частот называется ломаная линия, соединяющая точки с координатами $(x_i; n_i)$ или $(x_i; w_i)$.



Математическая статистика

Основными числовыми характеристиками вариационных рядов являются средняя арифметическая и дисперсия вариационного ряда.

Средней арифметической вариационного ряда называется сумма произведений всех вариантов ряда на соответствующие им частоты, делённая на объём.

Дисперсией вариационного ряда называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической.



Математическая статистика

По определению вести расчёты средней арифметической и дисперсии вариационного ряда бывает сложно. Можно пользоваться следующими формулами:



Математическая статистика

$$\bar{x} = \frac{k}{n} \cdot \sum_{i=1}^m \frac{x_i - c}{k} \cdot n_i + c$$

$$\sigma_x^2 = \frac{k^2}{n} \cdot \sum_{i=1}^m \left(\frac{x_i - c}{k} \right)^2 \cdot n_i - (\bar{x} - c)^2$$

σ^2 – генеральная дисперсия

s^2 – выборочная дисперсия



Математическая статистика

Известно, что для описания случайной величины достаточно знать её числовые характеристики (параметры). Например, математическое ожидание, дисперсию, с.к.о. Поэтому встаёт задача определения этих характеристик генеральной совокупности по тем же параметрам выборки.

Поскольку объём выборки мал, по сравнению с объёмом генеральной совокупности, то по выборке можно лишь оценить значения параметров генеральной совокупности.

Выборочная числовая характеристика t , используемая в качестве приближённого значения неизвестной числовой характеристики генеральной совокупности t , называется её *точечной статистической оценкой*.





Математическая статистика

Средние арифметические, дисперсии, а также с.к.о. распределения признака в генеральной и выборочной совокупностях называются *генеральной средней, выборочной средней, генеральной дисперсией, выборочной дисперсией, генеральным с.к.о., выборочным с.к.о.*



Математическая статистика

Выборочная средняя и выборочная доля являются точечными оценками генеральной средней и генеральной доли. Но точечных оценок не достаточно, следует выяснить степень рассеивания их относительно истинных параметров, т.е. дисперсию.

Интервальной оценкой параметра t называется числовой интервал $(a; b)$, который с заданной доверительной вероятностью γ «накрывает» неизвестное значение параметра t .

В этом случае интервал $(a; b)$ называется *доверительным интервалом*, а вероятность γ — *доверительной вероятностью*.



Доверительной вероятностью (надёжностью) γ называется вероятность того, что оценка x отклонится от оцениваемого параметра t по абсолютной величине не более, чем на положительное число Δ .

$$P(|x - t| \leq \Delta) = \gamma$$



Наибольшее отклонение Δ выборочной числовой характеристики от соответствующей ей генеральной характеристики, которое возможно с заданной доверительной вероятностью γ называется *предельной ошибкой выборки*.





Математическая статистика

$$\gamma = \Phi\left(\frac{\Delta}{\sigma_x}\right)$$

Φ - Функция Лапласа, значения которой находятся в таблице.

x - выборочная средняя или доля,

σ_x - соответствующее ей с.к.о.

Среднее квадратическое отклонение σ_x оценки x параметра t собственно случайной выборки называется *средней квадратической ошибкой* выборки.

Из последней формулы следует, что при заданной доверительной вероятности предельная ошибка выборки равна u -кратной величине средней квадратической ошибки, т.е. $\Delta = u \cdot \bar{\sigma}_x$ (u – аргумент функции Лапласа).





Математическая статистика

Формулы для средних квадратических ошибок имеют вид:

Оцениваемый параметр	Повторная выборка	Бесповторная выборка
Генеральная средняя \bar{x}	$\sigma_{\bar{x}} = \sqrt{\frac{s^2}{n}}$	$\sigma'_{\bar{x}} = \sqrt{\frac{s^2}{n} \left(1 - \frac{n}{N}\right)}$
Генеральная доля w	$\sigma_w = \sqrt{\frac{w(1-w)}{n}}$	$\sigma'_w = \sqrt{\frac{w(1-w)}{n} \left(1 - \frac{n}{N}\right)}$



Математическая статистика

При интервальном оценивании решаются следующие задачи:

1. Определение доверительного интервала при заданной доверительной вероятности и фиксированном объёме выборки;
2. Определение доверительной вероятности при заданном доверительном интервале и фиксированном объёме выборки;
3. Определение необходимого объёма выборки для достижения заданной точности и надёжности исследований.



Математическая статистика

Формулы расчёта объёма выборки имеют вид:

Оцениваемый параметр	Повторная выборка	Бесповторная выборка
Генеральная средняя \bar{X}	$n = \frac{u^2 \cdot s^2}{\Delta^2}$	$n' = \frac{n \cdot N}{n + N}$
Генеральная доля w	$n' = \frac{u^2 \cdot w \cdot (1 - w)}{\Delta^2}$	$n' = \frac{Nu^2 w(1 - w)}{u^2 w(1 - w) + N\Delta^2}$



Математическая статистика

При оценке генеральной доли в отсутствии предварительных сведений о значениях дисперсии и доли нет, то формула для объёма повторной выборки имеет следующий вид:

$$n = \frac{u^2}{4 \cdot \Delta^2} \cdot$$



Математическая статистика

В науке и на практике часто ставится задача нахождения неизвестного закона распределения признака, являющегося случайной величиной. С этой целью производится эксперимент, в результате которого получают эмпирическое распределение случайной величины в виде вариационного ряда. Далее на основе анализа опытных данных по отношению к известным теоретическим распределениям делают предположение о том, какое распределение лучше других отражает опытное.



Математическая статистика

Т.е. выдвигается *статистическая гипотеза* (предположение о виде или параметрах неизвестного закона распределения).
Необходимо выяснить, справедлива ли она (степень её согласованности с имеющимся эмпирическим вариационным рядом).



Математическая статистика

Если на основании теоретических предпосылок и анализа опытных данных приходим к выводу, что изучаемый признак распределён по *нормальному закону*, то нахождение нормального закона этого признака сводится к определению средней арифметической и дисперсии опытного распределения признака.



Математическая статистика

Затем вычисляют теоретические частоты, соответствующие опытным частотам по формуле:

$$n_i^0 = \frac{n \cdot k}{s} \cdot f\left(\frac{x_i - \bar{x}}{s}\right)$$

k - интервальная разность

f - функция Гаусса (значения в таблице)



Математическая статистика

После этого выясняется степень согласованности данных эксперимента и статистической гипотезы. Для ответа на этот вопрос существуют критерии согласия, одним из которых является критерий Пирсона. В нём за меру расхождения эмпирического ряда с гипотезой принимают величину χ^2 , которая вычисляется по формуле:

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - n_i^0)^2}{n_i^0},$$

n_i — эмпирическая частота.



Математическая статистика

Полученное значение χ^2 сравниваем с критическим (табличным). Для критического значения определяются число степеней свободы, которое на 3 единицы меньше, чем число интервалов и уровень значимости, который в наших гипотезах принимается равным 0,05. Если полученное значение χ^2 больше критического, то гипотеза о нормальном распределении опытных данных отвергается, а если полученное меньше критического, то не отвергается.

Задача



Пример 1. Для исследования количества рабочих часов, выработанных одним работником на фирме в течение декады из тысячи сотрудников по схеме собственно-случайной выборки отобрано 200 человек. Получены следующие данные:

Число часов	До 51	51-54	54-57	57-60	60-63	63-66	66-69	69-72	72-75	75-78	78-81	81-84	Свыше 81
Число работников	6	10	12	15	17	20	22	21	18	15	18	16	10

Задача



Найти доверительную вероятность того, что среднее количество рабочих часов всех сотрудников отклонится от выборочной средней на более, чем на полчаса.

Найти границы, в которых с вероятностью 0,9876 заключено среднее количество рабочих часов для всех сотрудников.

Определить минимальный объём выборки, по которой с вероятностью 0,9876 можно было утверждать, что среднее количество часов, полученное по выборке, отличалось от генеральной средней не более, чем на 1,725 часа.

Задача



Рассмотреть повторную и бесповторную выборки.

Проверить гипотезу о том, что количество рабочих часов, выработанных рабочим в течение декады распределено по нормальному закону.

Решение: сначала вычислим выборочную среднюю и выборочную дисперсию, для этого составим вспомогательную таблицу:

Задача



Кол. часов	до 51	51-54	54-57	57-60	60-63	63-66	66-69	69-72	72-75	75-78	78-81	81-84	св. 84	Итого
Середина интервала x_i	49,5	52,5	55,5	58,5	61,5	64,5	67,5	70,5	73,5	76,5	79,5	82,5	85,5	
Кол. сотрудников n_i	6	10	12	15	17	20	22	21	18	15	18	16	10	200
$\frac{x_i - c}{k}$	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	
$\frac{x_i - c}{k} \cdot n_i$	-36	-50	-48	-45	-34	-20	0	21	36	45	72	80	60	81
$\left(\frac{x_i - c}{k}\right)^2 \cdot n_i$	216	250	192	135	68	20	0	21	72	135	288	400	360	2157

Задача



$$c = 67,5; \quad k = 3$$

$$\bar{x} = \frac{81}{200} \cdot 3 + 67,5 = 68,715$$

$$s^2 = \frac{2157}{200} \cdot 3^2 - (68,715 - 67,5)^2 \approx 95,59$$

$$s \approx 9,78$$

Задача



Найдём средние квадратические ошибки:

$$\sigma_{\bar{x}} = \frac{9,78}{\sqrt{200}} \approx 0,69 \quad \text{п.в.}$$

$$\sigma_{\bar{x}}' = \sqrt{\frac{95,59}{200} \cdot \left(1 - \frac{200}{1000}\right)} \approx 0,2 \quad \text{б.п.в.}$$

Задача



Подставим их в формулу доверительной вероятности:

$$P(|X - 68,715| \leq 0,5) \approx \Phi\left(\frac{0,5}{0,69}\right) \approx \Phi(0,72) \approx 0,5285 \text{ п.в.}$$

$$P(|X - 68,715| \leq 0,5) \approx \Phi\left(\frac{0,5}{0,2}\right) = \Phi(2,5) \approx 0,9876 \text{ б.п.в.}$$

Задача



Для нахождения доверительного интервала нужно найти предельную ошибку выборки. Используем найденные ранее значения средних квадратических ошибок.

$$\Phi(u) = 0,9876 \Rightarrow u = 2,5$$

$$\Delta_{\bar{x}} = 2,5 \cdot 0,69 = 1,725$$

$$\Delta'_{\bar{x}} = 2,5 \cdot 0,2 = 0,5$$

$$68,715 \pm 1,725 \Rightarrow (66,99; 70,44) \quad \text{п.в.}$$

$$68,715 \pm 0,5 \Rightarrow (68,215; 69,215) \quad \text{б.п.в.}$$

Задача



Найдём минимальный объём выборки.

$$\Phi(u) = 0,9876 \Rightarrow u = 2,5 \quad (\text{табл.})$$

$$n = \left(\frac{2,5}{1,725} \right)^2 \cdot 95,59 \approx 200 \quad \text{н.в.}$$

$$n' = \frac{200 \cdot 1000}{200 + 1000} = \frac{1000}{1 + 5} \approx 167 \quad \text{б.н.в.}$$

Задача



Для нахождения теоретических частот
составим вспомогательную таблицу

Задача



x_i	49,5	52,5	55,5	58,5	61,5	64,5	67,5	70,5	73,5	76,5	79,5	82,5	85,5	Ито- го
n_i	6	10	12	15	17	20	22	21	18	15	18	16	10	200
$x_i - \bar{x}$	-19,2	-16,2	-13,2	-10,2	-7,2	-4,2	-1,2	1,8	4,8	7,8	10,8	13,8	16,8	
$\frac{x_i - \bar{x}}{s}$	-1,96	-1,66	-1,35	-1,04	-0,74	-0,43	-0,12	0,18	0,49	0,8	1,1	1,41	1,72	
$f\left(\frac{x_i - \bar{x}}{s}\right)$	0,0584	0,1006	0,1604	0,2323	0,3034	0,3637	0,3961	0,3925	0,3538	0,2897	0,2179	0,1476	0,0909	
n_i^0	4	6	10	14	19	22	24	24	22	18	13	9	6	191

Задача



Рассчитаем значение критерия Пирсона:

$$\begin{aligned}\chi^2 &= \frac{(6-4)^2}{4} + \frac{(10-6)^2}{6} + \frac{(12-10)^2}{10} + \frac{(15-14)^2}{14} + \\ &+ \frac{(17-19)^2}{19} + \frac{(20-22)^2}{22} + \frac{(22-24)^2}{24} + \frac{(21-24)^2}{24} + \\ &+ \frac{(18-22)^2}{22} + \frac{(15-18)^2}{18} + \frac{(18-13)^2}{13} + \frac{(16-9)^2}{9} + \\ &+ \frac{(10-6)^2}{6} = 1 + 2,6667 + 0,4 + 0,0714 + 0,2105 + \\ &+ 0,1818 + 0,1667 + 0,375 + 0,7273 + 0,5 + 1,9231 + \\ &+ 5,4444 + 2,6667 = 16,3336\end{aligned}$$

Задача



Найдём по таблице критическое значение критерия Пирсона (число степеней свободы $k=10$, уровень значимости принимается равным $0,05$).

$$\chi_{кр}^2(0,05; 10) = 18,3$$

$$16,3336 < 18,3$$

Это позволяет утверждать, что при уровне значимости $0,05$ опытные данные не противоречат гипотезе о нормальном законе распределения (или опытные данные согласуются с выдвинутой гипотезой).

Задача



Пример 2. Проверяется партия из 5000 консервов. Проверили 10%, среди проверенных оказалось 12% просроченных. Найти доверительную вероятность того, что процент годных консервов во всей партии отличается от процента годных в выборке не более, чем на 3% по абсолютной величине.

Найти границы в которых с вероятностью 0,95 заключён процент годных консервов во всей партии.

Задача



Каким должен быть минимальный объём выборки по которой можно было бы утверждать, что отклонение доли годных консервов не превысит 2,8% по абсолютной величине (рассмотреть повторную и бесповторную выборки).

Задача



Решение:

Дано: $N = 5000$; $n = 500$; $w = 0,88$.

$$\sigma_w = \sqrt{\frac{0,88 \cdot 0,12}{500}} \approx 0,0145$$

$$\sigma_w' = \sqrt{\frac{0,88 \cdot 0,12}{500} \cdot \left(1 - \frac{500}{5000}\right)} \approx 0,0138$$

$$P(|w - 0,88| \leq 0,03) \approx \Phi\left(\frac{0,03}{0,0145}\right) \approx \Phi(2,06) \approx 0,9606 \text{ п.в.}$$

$$P(|w - 0,88| \leq 0,03) \approx \Phi\left(\frac{0,03}{0,0138}\right) \approx \Phi(2,17) \approx 0,97 \text{ б.п.в.}$$

Задача



Для нахождения доверительного интервала найдём предельные ошибки выборки, используя найденные значения средних квадратических ошибок.

$$\Phi(u) = 0,95 \Rightarrow u = 2,5$$

$$\Delta_w = 2,5 \cdot 0,01145 \approx 0,028$$

$$\Delta'_w = 2,5 \cdot 0,01138 \approx 0,027$$

$$88\% \pm 2,8\% \Rightarrow (85,2\%; 90,8\%) \quad \text{н.в.}$$

$$88\% \pm 2,7\% \Rightarrow (85,3\%; 90,7\%) \quad \text{б.н.в.}$$

Задача



Найдём минимальный объём выборки:

$$\Phi(u) = 0,9 \Rightarrow u = 1,65$$

$$n = \frac{(1,65)^2 \cdot 0,88 \cdot 0,12}{(0,028)^2} \approx 367$$

$$n' = \frac{5000 \cdot (1,65)^2 \cdot 0,88 \cdot 0,12}{5000 \cdot (0,027)^2 + (1,65)^2 \cdot 0,88 \cdot 0,12} \approx 366$$



Финансовый университет
при Правительстве Российской Федерации

Конец лекции