ФИЗИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

МАТЕМАТИЧЕСКАЯ ОБРАБОТКА НАБЛЮДЕНИЙ. СЕМИНАРЫ

САЖИНА
ОЛЬГА СЕРГЕЕВНА

ФИЗФАК МГУ

Содержание

1. Семинар 1. Основы теории вероятностей

Теория вероятности оперирует такими объектами как опыт и результаты опыта. Примером опыта может служить бросание игрального кубика, а результатом этого опыта выпадание числа на его грани.

Определим вероятность события A следующим образом

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

где, m_a это число появления события A, а n общее событий.

Например, если событие A это выпадение единицы на грани кубика. Мы подбрасываем кубик 100 раз, при этом единица выпадает 34 раза. Тогда мы можем расчитать вероятность выпадения единицы

$$P(A) = \frac{34}{100} = 0.34$$

Строго говоря величина $P(A)$ называется частотой, а не вероятностью. Если мы устремим n к бесконечности мы получим вероятность, но так как мы имеем дело с конечным числом испытаний n, то величину $P(A)$ мы будем называть вероятностью.

Свойства вероятности $P(A)$

1) $0 \leq P(A) \leq 1$

Будем обозначать *невозможное события* как $\varnothing$.

2) Если $A = \varnothing$, то $P(\varnothing) = 0$.

Но если $P(A) = 0$, то A необязательно $\varnothing$. Вероятность попасть из ружья в определенную точку на стене равна нулю, но при этом попасть в нее можно.

Будем обозначать *достоверное событие* Ω.

3) $P(\Omega) = 1$

Вероятность при выстреле в стену попасть в любую точку на стене.

Противоположенное событие обозначим $\overline{A}$. Для того чтобы найти противоположенное событие, надо построить отрицание. К примеру если A хотя бы одно событие, то $\overline{A}$ ни одного события. Или если $A = \{$ровно одно$\}$, то $\overline{A} = \{$ни одного, или два, или три,...$\}$

4) $P(\overline{A}) = 1 - P(A)$

5) *Несовместные события* назовем такие которые не могут произойти одновременно в рамках одного опыта $A_1 \cdot A_2 = \varnothing$ На языке теории множеств последние утверждение можно записать через пересечение множеств $A_1 \cap A_2 = \varnothing$. Так события мы можем трактовать как множества.

Так же для несовместных событий выполняется свойство $P(A_1 + A_2) = P(A_1) + P(A_2)$. Или мы можем записать $P(A_1 + A_2)$ через объединение множеств $P(A_1 \cup A_2)$.

Для *совместных событий* $P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 \cdot A_2)$

Эти формулы легко визуализировать с помощью *диаграмм венна.*

Задача

Имеется трехтомник стихотворений, какова вероятность что при случайном расположении трехтомника на книжной полке хотя бы один том окажется на своем порядковом месте.

Решение
Всего вариантов 6, подходящих нам 4(123,132,321,213), значит

$$P(A) = \frac{4}{6}$$

1.1. Геометрическая вероятность

Рассмотрим методику вычисления вероятности когда набор значений не дискретный, а непрерывный. Как и обычно вероятность есть отношение благоприятных исходов к общему количеству, но теперь это количество измеряется не "штуками а объемами,площадями, длинами и т.п в зависимости от размерности задачи. Для задачи на плоскости формула записывается следующим образом

$$P(A) = \frac{S(A)}{S(\Omega)} \tag{2}$$

Где $S(A)$ и $S(\Omega)$ площади области A,Ω соответственно. В общем случае для задачи произвольной размерности формула записывается в виде

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}$$

где μ - сокращение от mes(мера соответствующей размерности). Обычные задачи на вероятность можно переформулировать в терминах геометрической вероятности, что позволит решить их быстрее. В этом основная ценность понятия геометрической вероятности. Рассмотрим задачи на геометрическую вероятность.

Задача Бюффона

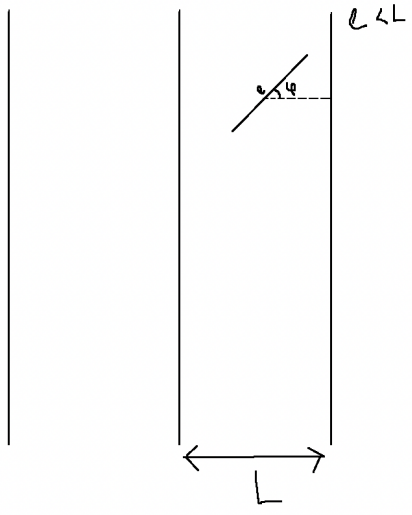

Рис. 1.1 – Задача Бюффона

Пусть у нас есть бесконечные параллельные прямые на плоскости с заданным расстоянием L между ними. На эту плоскость бросают иголку длины $l < L$, рис (1.1). Найти вероятность того что иголка пересечет хотя бы одну из этих прямых.

Идея решения

При обычном решении задачи нам нужно было бы найти количество всех случаев когда иголка пересекает прямые. Но организовать перебор всех случаев обычным образом слишком тяжело. Поэтому эту задачу нужно переформулировать в терминах геометрической вероятности. Мы не будем полностью записывать решение, лишь зададим направление в котором нужно двигаться.

Нам необходимо определить положение иголки с помощью некоторых параметров. Какими параметрами однозначно задается ее положение? Для этого достаточно двух свободных параметров таких как длина и угол поворота. Тогда в плоскости выбранных нами параметров мы можем нарисовать области для геометрической вероятности. Предлагается решить эту задачу самостоятельно, а мы решим похожую задачу.

Задача

Пусть на приемник поступает два сигнала в момент времени от 0 до T. Приемник приобретает статус "забитого". Если интервал времени между этими сигналами меньше чем τ. Найдите вероятность того что приемник забит.

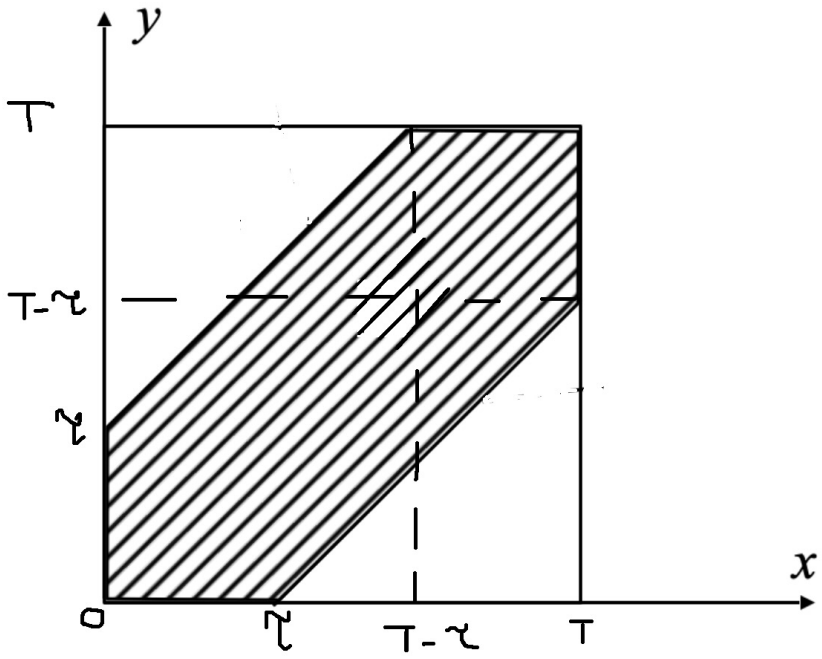

Рис. 1.2 – Задача на приемник

Решение

Зададим моменты поступления первого и второго сигнала как x и y соответственно. Область изменения x и y от 0 до T образует квадрат на плоскости x, y, рис(1.2). Точки в этом квадарате охватывают все возможные ситуации прихода двух сигналов на наш прибор. Мы можем записать условие забитости приемника как $|x - y| < \tau$. Нам остается лишь построить эти области и найти отношения их площадей.

$$P = \frac{T^2 - (T-\tau)^2}{T^2} = 1 - \left(1 - \frac{\tau}{T}\right)^2$$

Из последнего выражения легко увидеть что при $\tau = T$, приемник забит постоянно, а при $\tau = 0$ он всегда свободен. Тем самым мы понимаем что полученный нами ответ физически разумен.

Подобные задачи зачастую встречаются в астрономии. Предположим что в некотором поле есть две пространственные или одномерные структуры и нам нужно узнать как относительно них расположена третья структура. То есть с какой вероятностью она может их пересечь, а значит наблюдение будет на одном луче зрения. Например, такой структурой может быть космическая струна. Перейдем к задаче на сложение вероятности.

Задача

Двое поочередно подбрасывают монету. Выигрывает тот у кого раньше выпадает орел. Определить вероятность выигрыша каждого игрока.

Решение

Изначально кажется что вероятности должны быть равны, чтобы узнать точно построим схему ситуаций. Пусть A и B - события, состоящие в том, что выиграет 1-й и 2-й игрок соответственно. Представим событие A в следующем виде

$$A = \text{"О"} + \text{"РРО} + \text{"РРРРО} + \ldots$$

$$P(A) = \frac{1}{2} + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^5 + \ldots = \frac{\frac{1}{2}}{1 - \left(\frac{1}{2}\right)^2} = \frac{2}{3}$$

Мы нашли вероятность выигрыша первого игрока. Если первый игрок не выиграл, то автоматически выигрывает второй игрок, значит

$$P(B) = 1 - P(A) = 1 - \frac{2}{3} = \frac{1}{3}$$

1.2. Условная вероятность

По определению условная вероятность это вероятность наступления события A при условии наступления события B и записывается следующим образом

$$P(A|B) \stackrel{\text{def}}{=} \frac{m_{AB}}{m_B} = \frac{P(AB)}{P(B)} \tag{3}$$

Где, m_{AB} количество опытов где A и B реализовались одновременно, а m_B количсетво опытов где реализовались только события B.

Событие называется *независимыми* если $P(A|B) = P(A)$. Или можно сказать по другому что два события яляются независимыми если $P(A)P(B) = P(AB)$. Такое же определение можно распространить на множество независимых событий.

Для независимых и совместных событий выполняется равенство

$$1 - P(A + B) = P(\overline{A} \cdot \overline{B})$$

Докажем это

$$1-P(A)-P(B)+P(AB) = P(\overline{A})\cdot P(\overline{B}) = (1-P(A))(1-P(B)) = 1-P(B)-P(A)+P(A)P(B)$$

В силу независимости A и B получаем верное равенство. Это же можно доказать с помощью диаграмм Венна.

Задача

В первой урне 10 белых и 5 черных шаров. Во второй 4 белых и 4 черных шара. Одна из этих урн выбирается на угад. Какова вероятность того, что шар извлеченный на угад из выбранной урны окажется белым?

Решение

1)Искомая вероятность будет складываться из двух

$$P(A) = P(A|I)P(I) + P(A|II)P(I) = \frac{10}{15} \cdot \frac{1}{2} + \frac{4}{8} \cdot \frac{1}{2} = \frac{7}{12}$$

Где $P(I), P(II)$ вероятности выбрать первую и вторую урну соответственно. Записанная нами формула называется *формулой общей вероятности.*

Теперь найдем вероятность того, что из двух урн была выбрана первая если извлеченный шар оказался белым?

Решение

Здесь нам требуется оценить вероятность условий произошедшего событие, для этого нужно использовать *формулу Байеса*, запишем ее

$$P(I|A) = \frac{P(A|I) \cdot P(I)}{P(A)} = \frac{4}{7}$$

Теперь пусть шар вынутый из наугад выбранной урны оказался белым, какова вероятность того, что второй шар из выбранной урны тоже белый?

Решение Здесь так же воспользуемся формулой для полной вероятности

$$P(B|A) = \frac{P(BA)}{P(A)} = \frac{P(AB|I)P(I) + P(AB|II)P(II)}{P(A|I)P(I) + P(A|II)P(II)} = \frac{\frac{10}{15} \cdot \frac{9}{14} \cdot \frac{1}{2} + \frac{4}{8} \cdot \frac{3}{7} \cdot \frac{1}{2}}{\frac{7}{12}} = \frac{27}{49}$$

2. Семинар 2. Элементы комбинаторики

Комбинаторика это отдельный раздел математики который позволяет вычислять количество выборов элементов одного множества из другого множества при некоторых заданных ограничениях. Например, пусть у нас есть два множества и мы хотим элементы внутреннего множества выбрать таким образом, чтобы после каждого выбора возвращать выбранный элемент обратно, тем самым перемешивая элементы множетсва. Возврат выбранного элемента и есть заданные ограничения или условия опыта. Так же можно не возвращать элементы, или возвращать без перемешивания это другие условия опыта.

Есть две принципиально разных процедуры выбора:

1) *Выборка без повторения* - при этом выборе каждый выбранный элемент икслючается из множества.

2) *Выборка с повторениями* - при этом выборе каждый выбранный элемент возвращается обратно в множество (при этом при следующем выборе есть возможность выбрать тот же самы элемент).

После того как мы осуществили выборку, то мы можем выполнить упорядочивание выбранных элементов. При этом каждая из выборок по способу упорядочивания делится на два типа:

Размещения - если нам важен порядок выбранных элементов(частным случаем являются *перестановки*)

Сочетания - если нам не важен порядок элементов (то есть элементы не упорядочиваются)

Самое сложное при решении задачи это определить к какому типу она относится.

2.1. Размещения без повторений

В начале рассмотрим размещения без повторений оно обозначется как A_n^m и по определению равно

$$A_n^m \stackrel{\text{def}}{=} n(n-1)(n-m+1)\dots(n-m-1) = \frac{n!}{(n-m)!}, \quad m \leq n \tag{4}$$

или числу размещений из n элементов по m элементам. Например, у нас есть m ячеек и мы размещаем по ним n шариков

Задача

Сколько четырехзначных чисел можно составить из пяти цифр $\{1, 2, 3, 4, 5\}$?

Решение

Для решения воспользуемся формулой (4)

$$A_5^4 = 120$$

Эту же задачу можно решить на понятийном уровне. На первое место в этом числе можно поставить одну из 5 цифр. На второе - уже одну из 4, и так далее. Остается только перемножить эти количества, и мы узнаем общее количество вариантов числа.

Задача

Белые и черные ладьи наугад ставятся на шахматную доску. Какова вероятность того, что две ладьи не будут бить друг друга?

Решение

зачастую вместо того чтобы считать количество благоприятных исходов A, легче посчитать количество противоположенных событий $\overline{A}$. Найдем количесвто ситуаций в которыха ладью бьют друг друга. Для начала посчитаем общее число способов N разместить две ладьи на шахматной доске

$$N = 64 \cdot 63 = A_{64}^2$$

Для каждого положения первой ладьи у нас есть $7 \cdot 2$ положения второй ладьи, чтобы они друг друга били.

$$n_{\overline{A}} = 64 \cdot (8-1) \cdot 2$$

Тогда вероятность того, что ладьи друг друга бьют

$$P(\overline{A}) = \frac{n_{\overline{A}}}{N} = \frac{2}{9}$$

А искомая вероятность равна

$$P(A) = \frac{7}{9}$$

2.2. Размещение с повторениями

Размещение с повторениями обозначаются $\overline{A}_n^m$ и по определению равно

$$\overline{A}_n^m \stackrel{\text{def}}{=} n^m \tag{5}$$

Вернемся к задаче о четырехзначном числе только теперь мы можем повторять цифры. Тогда на каждое из четырех мест мы можем поставить любую из пяти цифр. Поэтому число возможных размещений равно

$$n^m = 5^4$$

Задача

В поезде из 10 вагонов случайно оказались преступник и майор Томин. Какова вероятность того, что они оба едут в одном вагоне?

Решение

$$P(A) = \frac{10}{\overline{A}_{10}^2}$$

Заметим что повтор здесь в том смысле что повторяться может номер вагона для двух человек.

2.3. Перестановки с повторениями и без повторений

Перестановки без повторений это частный случай размещений без повторений A_n^m при $n = m$

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

По этой формуле решается известная задача о том каким числом способов можно разместить на полке трехтомник стихотворений(ответ 3!).

Перестановки с повторениями обозначаются буквой P и равны

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_k!}$$

где аргументы $n_1, n_2, \ldots, n_k$ обозначают количество повторов. Например, для слова это количество одинаковых букв.

Задача

Пусть есть слово "АВИАЦИЯ". Сколько разных семибуквенных слов можно составить из букв этого слова?

Решение

В комбинаторике словом называют любую комбинацию букв. Всего у нас 7 букв значит $n = 7$, при этом буква "А"и буква "И"встречаются два раза, а все остальные по одной значит $n_1 = n_2 = 2, n_3 = n_4 = n_5 = 1$, тогда

$$P(2, 2, 1, 1, 1) = \frac{7!}{2!2!} = 1260$$

Если бы все буквы были разные, то количество семибуквенных слов оказалось бы больше (7!), ведь тогда у нас были бы перестановки без повторений.

С помощью формулы для перестановок можно решить следующий пример

$$(a_1 + a_2 + \ldots + a_k)^n = \sum P(n_1, n_2, \ldots, n_k) a_1^{n_1} \cdot a_2^{n_2} \cdot \ldots \cdot a_k^{n_k}$$

Где суммирование ведется по все n_i таким что $n_1 + n_2 + \ldots + n_k = n$

2.4. Сочетания без повторений

Сочетания без повторений C_n^m это число способов выбрать m элементов из n элементов при этом $m \leq n$ и по определению равно

$$C_n^m \stackrel{\text{def}}{=} \frac{n!}{(n-m)!m!} = \frac{A_n^m}{m!} \tag{6}$$

Из формулы 6 следует что количество сочетаний без повторений всегда меньше или равно чем количества размещений без повторений

$$C_n^m \leq A_n^m$$

Формула для сочетаний встречается в **биноме Ньютона**

$$(a+b)^n = \sum_{m=0}^{n} C_n^m a^{n-m} b^m \tag{7}$$

Рассмотрим его свойства

1)

$$C_n^m = C_n^{n-m}$$

2)

$$C_{n+1}^{m+1} = C_n^{m+1} + C_n^m$$

Проведем идею доказательства этого свойства. Распишем вторую часть равенства

$$\frac{n!}{(m+1)!(n-m-1)} + \frac{n!}{n!(n-m)!} = \frac{n!(n-m)+n!(m+1)}{(m+1)!(n-m)!} = \frac{(n+1)!}{(m+1)!(n-m)!} = C_{n+1}^{m+1}$$

3) Формула Стирлинга

$$n! \xrightarrow[n\to\infty]{} \sqrt{2\pi n} \cdot n^n \cdot e^{-n}$$

Рассмотрим пример на сочетания без повторений из статистической физики.

Пример

Пусть m частиц, размещают в n ячеек, причем $m < n$. Кажая ячейка может содержать только одну частицу. Тогда по статистики Ферми-Дирака для тождественных частиц типа электронов, протонов, нейтронов и т.д. тогда число равно вероятных состояний будет выражаться как C_n^m.

2.5. Сочетания с повторениями

Сочетания с повторениями это количество наборов в которых каждый элемент может участвовать несколько раз

$$C_n^m \overset{\text{def}}{=} \frac{(n+m-1)!}{(m-1)!n!} \tag{8}$$

2.6. Симметрия в теории вероятности

Задача

Есть множество чисел от 1 до 100, из этого множества последовательно без возвращений выбирают два числа. Какова вероятность что второе число больше чем первое?

Решение

Из соображения симметрии количество пар, в которых первое число больше второго, равно количеству пар, в которых первое число меньше второго, отсюда следует что вероятность равна $\frac{1}{2}$.

3. Семинар 3. Нормальное распределение

Задача

Пусть деталь размера D изготовляется с ошибками, поэтому D распределено по нормальному закону

$$D \sim N(\mu, \sigma^2)$$

Где $\mu = 40$мм - математическое ожидание , $\sigma = 0,05$мм - среднеквадратическое отклонение. В результате контроля бракованными признаются все детали у которых $D < a = 39.85$мм или $D > b = 40.05$мм. Определить вероятность того, что наугад выбранная деталь будет признана бракованной и определить процент бракованных деталей.

Решение

$$P(A) = 1 - P(\overline{A}) = 1 - \Phi\left(\frac{40.05 - 40.0}{0.05}\right) + \Phi\left(\frac{39.85 - 40.0}{0.05}\right) = 1 - \Phi(1) + \Phi(-3) =$$

$$= 1 - \Phi(1) + (1 - \Phi(3)) = 1 - 0.8413 + 1 - 0.9987 = 0.16$$

Значит искомая вероятность будет равна 0.16, а процент забракованных деталей равен 16%.

3.1. Дискретная случайная величина

Задача

Пусть X - дискретная случайная величина, а $h(X) = \cos X$ функция этой случайной величины. Постройте ряд распределения функции $h(X)$ если ряд распределения X задан таблицей 9

x_i	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$
p_i	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

(9)

Решение

Для начала проверяем что сумма всех p_i равно 1, после этого переходим к вычислениям

$$P_1^H = P(h(0) = \cos 0 = 1) = \frac{1}{6}$$

$$P_2^H = P\left(h\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}\right) = \frac{2}{6}$$

$$P_3^H = P\left(h\left(\frac{\pi}{2}\right) = 0\right) = \frac{3}{6}$$

Теперь мы можем составить таблицу

h_i	0	$\frac{\sqrt{2}}{2}$	1
p_i^H	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Более сложным случай будет когда есть несколько значений x, а функция принимает одно и то же значение.

Задача

Пусть теперь значения x_i определяются таблицей 10

x_i	$-\frac{\pi}{2}$	$-\frac{\pi}{4}$	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$
p_i	$\frac{1}{35}$	$\frac{3}{35}$	$\frac{6}{35}$	$\frac{10}{35}$	$\frac{15}{35}$

(10)

Решение

Если случайная величина X принимает несколько значений $x_\nu, x_{\nu+1} \dots, x_{k-1}x_k$ при которых $h(X) = h(x_k) = h_0$, то

$$P(h(x) = h_0) = \sum_{j=\nu}^{k} P(X = x_j)$$

Например, в точке $\frac{\pi}{2}$ и $-\frac{\pi}{2}$ функция косинуса равна 0, поэтому в таблице для p_1^H значения $\frac{1}{35}$ и $\frac{15}{35}$ складываются, запишем таблицу для функции

h_i	0	$\frac{\sqrt{2}}{2}$	1
p_i^H	$\frac{16}{35}$	$\frac{13}{35}$	$\frac{6}{35}$

3.2. Непрерывная случайная величина

Решим задачу для непрерывной случайной величины, которая задается не таблицей, а функцией или плотностью распределения

Задача

Пусть X непрерывная случайная величина для которой задана плотность распределения в дифференцальной форме записи

$$f(x)dx = P(x \le X \le x + dx)$$

Ставится задача найти плотность распределения $g(h)$ функции $h(x) = h$, такую что

$$g(h)dh = P(h \le H \le h + dh)$$

Решение

Для простоты будем считать что $h(x)$ это однозначная функция. Тогда по аналогии с дискретным случаем, можно найти малый интервал значений $h(x)$, соответствующий заданному малому интервалу значений X с известной вероятность $f(x)dx$

$$dx = \left|\frac{dx(h)}{dh}\right| dh$$

где $dx(h)$ обратная функция, тогда

$$f(x)dx = f(x(h))\left|\frac{dx(h)}{dh}\right|dh$$

Для дискретной величины в однозначном случае вероятность для X такая же, как и для функции от X взятой в одной и той же точке(т.е. $x_i = p_i \Rightarrow h(x_i) \to p_i$), то же можно написать и для плотностей

$$g(h) = f(x(h))\left|\frac{dx(h)}{dh}\right|$$

Пример
Пусть $h(x) = \cos x$. Распределение вероятности для X

$$f(x)dx = a + bx$$

где

$$0 \le x \le \frac{\pi}{2}$$

Найдем плотность $g(h)$

$$g(h)dh = f[x(h)]\left|\frac{dx(h)}{dh}\right|dh = [a + b\arccos h] \cdot \frac{dh}{\sqrt{1-h^2}}, \quad 0 \le h \le 1$$

Окончательно

$$g(h) = [a + b\arccos h] \cdot \frac{1}{\sqrt{1-h^2}}, \quad 0 \le h \le 1$$

Задача
Вероятность обнаружить звезду в объеме dv равна kdx. Для каждой звезды найдется другая звезда - ее ближайший сосед. Найти функцию распределения $F(X)$ расстоянии до ближайшего соседа, а также среднее расстояние $M[X]$ до ближайшего соседа, а так же дисперсию расстояний $D[X]$.

Решение
Случайную величину, расстояние от звезды до ее ближайшего соседа, обозначим за X, рис(3.1). Тогда вероятность того, что сосед находится ближе расстояния x равно, по определению, функции распределения, $F(x) = P(X < x)$. Вероятность того, что ближайший сосед находится не ближе x равно $1 - F(x)$. Вероятность того, что сосед находится на расстоянии между x и $x + dx$ определяется как $f(x)dx$, и равна произведению $1 - F(x)$ на вероятность того, что между сферами с радиусами x и xdx имеется звезда, таким образом

$$f(x)dx = [1 - F(x)] \cdot k \cdot 4\pi \cdot x^2 dx$$

Продифференцируем это выражение по x и разделим на $\cdot k \cdot 4\pi \cdot x^2$, учтя, что $F'(x) = f(x)$

$$\frac{1}{k \cdot 4\pi}\frac{2}{x^3}f(x) + \frac{1}{k \cdot 4\pi x^2}f'(x) = -f(x)$$

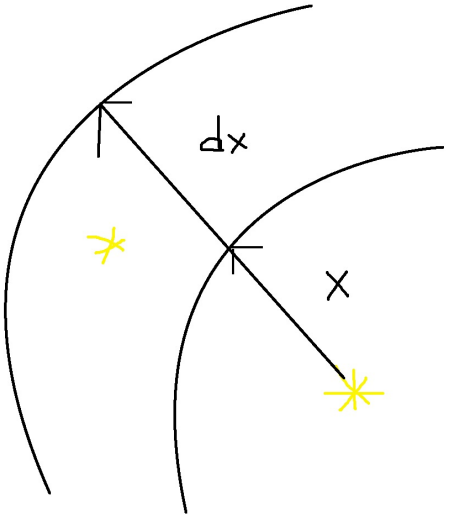

Рис. 3.1 – Расстояние между соседними звездами

Разделим на $f(x)$ и умножим на $k \cdot 4\pi \cdot x^2$, чтобы получить *логарифмическую производную*

$$\frac{f'(x)}{f(x)} = \frac{2}{x} - 4\pi \cdot k \cdot x^2$$

После интегрирования получаем

$$f(x) = cx^2 exp\left\{-\frac{4}{3}\pi \cdot k \cdot x^3\right\}$$

Произвольная константа c определяется из условия равенства интеграла плотности распределения единице на всей числовой прямой

$$\int_{\mathbb{R}} f(x)dx = \frac{c}{3}\frac{1}{\frac{4}{3}\pi k}\int_0^\infty exp\left\{-\frac{4}{3}\pi \cdot k \cdot x^3\right\} d\left(x^3\frac{4}{3}\pi k\right) = 1$$

Здесь мы так же учли что расстояние не может быть отрицательным и интеграл по всей числовой прямой превратился в интеграл по полупрямой. Окончательно находим

$$f(x) = 4\pi \cdot kx^2 exp\left\{-\frac{4}{3}\pi \cdot k \cdot x^3\right\}$$

Среднее расстояние до ближайшего соседа

$$\overline{x} = \int_0^\infty x \cdot f(x)dx = \left(\frac{3}{4\pi \cdot k}\right)^{\frac{1}{3}} \Gamma\left(\frac{4}{3}\right) \approx 0.554 \cdot k^{-1/3}$$

Дисперсия расстояния до ближайшего соседа

$$D[x] = \int_0^\infty (x - \overline{x})^2 f(x)dx \approx 0.0405 \cdot k^{-2/3}$$

Среднеквадратичное отклонение

$$\sigma = 0.201 \cdot k^{-1/3}$$

3.3. Логарифмически-нормальное распределение

Пример Случайная величина $Y = \log X$ распределена по нормальному закону $Y \sim N(0,1)$ тогда $X \sim logN(0,1)$. Получим плотность нормального распределения X

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

4. Семинар 4. Задачи на теорию множеств и комбинаторику

Задача

Доказать дистрибутивность умножения относительно сложения

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

Решение

Задачу можно решить с помощью диаграмм Венна, но мы решим это с помощью другого метода. Нужно доказать что элемент принадлежащий левой части равенства, так же принадлежит и второй части равенства

$$\omega \in (A+B)C \Rightarrow X(\omega) = \begin{cases} \omega \in A+B \\ \omega \in C \end{cases} = \left[\begin{array}{l} \begin{cases} \omega \in A \\ \omega \in C \end{cases} \\ \\ \begin{cases} \omega \in B \\ \omega \in C \end{cases} \end{array} \right. = \left[\begin{array}{l} \omega \in AC \\ \omega \in BC \end{array} \right. \Rightarrow \omega \in AC+BC$$

Вторым шагом нужно доказать обратное включение $\omega \in AB+BC \Rightarrow \omega \in (A+B)C$

Задача

Доказать

$$A - B = A \cdot \overline{B}$$

Решение

$$\begin{cases} \omega \in A \\ \omega \notin B \end{cases} \Leftrightarrow \begin{cases} \omega \in A \\ \omega \in \overline{B} \end{cases} \Leftrightarrow \omega \in A \cdot \overline{B}$$

Правила Моргана

1)

$$\overline{A+B} = \overline{A} \cdot \overline{B}$$

2)

$$\overline{AB} = \overline{A} + \overline{B}$$

Приведем пример, которй наглядно показывает отличие алгебры множеств, от алгебры чисел

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

или в другой форме записи

$$(A \cup B)/B$$

Задача

Пусть заданы два несовместных событий A и B. Доказать что AC и BC также несовместны

Решение

Необходимо доказать, что

$$(AC)(BC) = 0$$

Так же рассмотрим элемент ω который принадлежит левой части. По условию ω должен принадлежать A, C и BC одновременно, так как A и B не имеют общих элементов, тогда заключаем что $(AC)(BC)$ пустое множество.

Задача

Нужно составить пятизначное число используя цифры $\overline{0,9}$, так чтобы оно читалось одинаково слева направо и справа налево

Решение

На первое место в числе можно поставить 9 цифр, а на остальные 10. Поэтому общее число исходов равно $9 \cdot 10^4$. Далее посмотрим какие исходы нас устраивают. На первое число мы можем поставить цифру 9 способами, при этом мы определяем какое число будет стоять на последнем месте. На второе мы можем выбрать число 10 способами при этом мы также ставим цифру на четвертое место. И в середине числа мы можем выбрать число также 10 способами, значит

$$P(A) = \frac{9 \cdot 10 \cdot 10}{9 \cdot 10^4}$$

Задача

Нужно составить пятизначное число используя цифры $\overline{0,9}$, так чтобы оно было кратно пяти

Решение

Общее число исходов остается таким же как и в предыдущей задаче. На первое место мы ставим число 9 способами, на второе, третье и четвертое 10 способами, а на последнее только двумя, значит

$$P(A) = \frac{9 \cdot 10^3 \cdot 2}{9 \cdot 10^4}$$

Задача

Две команды играют n партий. Нужно найти вероятность того, что первая команда выиграет ровно m раз

Решение

$$P = \frac{C_n^m}{2^n}$$

Задача

В библиотеке имеются книги по 16-ти разделам науки. Сколько есть способов выбрать себе 4 книги?

Решение Для начала посмотрим какие слова указывают нам на тип задачи. Из формулировки задачи мы можем понять что из каждого раздела мы можем брать столько книг сколько захотим, допустим мы можем взять все 4 книги только по физике или только по биологии. Это указание на то, что в задаче есть повторения. Возьмем 3 книги по физики и 1 по биологии и обозначим это множество книг как $E_1 = \{e_1, e_1, e_1, e_2\}$. Так же составим множество, где возьмем сначала одну книгу по физике, потом одну по биологии и затем еще 2 книги по физике и обозначим это множество как $E_2 = \{e_1, e_2, e_1, e_1\}$. Множество E_1 будет равноправно множеству E_2, то есть порядок элементов нам неважен, также как нам неважно какая книга будет

лежать сверху, главное какие книги мы взяли. Это указания на сочетания. Итак, мы имеем дело сочетания с повторениями.

Изобразим в общем виде схему решения нашей задачи на рисунке 4.1. Вертикальные стенки обозначают границы разделов, а круги книги которые мы выбираем. Мы можем переставлять все стенки, кроме крайних $(n+1)-2$. Рисунок 4.1 а) иллюстрирует ситуацию когда мы взяли 2 книги из первого раздела, и по одной из последних двух. Рисунок 4.1 б) иллюстрирует ситуацию когда все книги были взяты из первого раздела. Мы должны выбрать m штук элементов из общего числа элементов, которые мы можем переставлять $n-1+m$, общая формула решения задачи

$$C_{n-1+m}^{m}$$

Формула в числах для нашей задачи

$$C_{16-1+4}^{4} = C_{19}^{4} = C_{19}^{15}$$

а) б)

Рис. 4.1 – Иллюстрация к задаче

Задача
Сколько существует способов распределить 5 шаров по двум лункам?
Решения
Повторяться у нас может номер лунки, так как все шары могут попасть в одну лунку. При этом нам важно какой именно шар попал в лунку, нам важен порядок, поэтому мы имеем дело с размещениями с повторениями.

$$\overline{A_n^m} = 2^5$$

Задача
Какова вероятность что сумма трех наудачу взятых отрезков длинна которых не превосходит l будет больше чем l?
Решение
Нам необходимо найти вероятность того, что $a+b+c > l$. Решим задачу изобразив область вероятностей на плоскости (рис(4.2)). Событие $a+b+c > l$ соответствует

многограннику, отсекаемому от куба плоскостью $a+b+c=l$. Для задания плоскости необходимы три точки, возьмем их как $a=l, b=l, c=l$. Область, которая дополняет многогранник, является тетраэдром, поэтому

$$P(A)=1-P(a+b+c<l)=1-\frac{V_{\text{тетр}}}{V_{\text{куб}}}=1-\frac{\frac{1}{3}\cdot l\cdot\frac{l^2}{2}}{l^2}=1-\frac{1}{6}$$

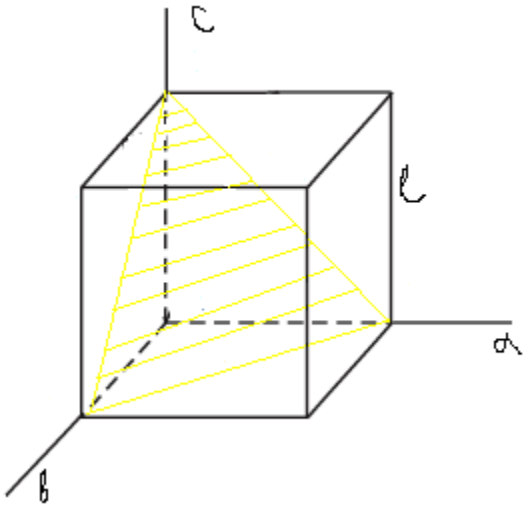

Рис. 4.2 – Иллюстрация к задаче

Задача Вероятность попасть в самолет 2/5, вероятность сбить самолет 1/10. Найдите вероятность того, что при попадании самолет будет сбит.

Решение

Нам необходимо найти $P(B|A)$, вероятность того, что мы сбили B самолет при условии, что мы в него попали A.

$$P(B|A)=\frac{P(AB)}{P(A)}=\frac{P(B)}{P(A)}=\frac{1}{4}$$

Задача

Пусть $A\subset B$, доказать что $P(A)\leq P(B)$

Решение

$$A\subset B\Rightarrow A\cdot B=A$$

$$B=B(A+\overline{A})=BA+B\overline{A}=A+B\overline{A}$$

Из этого следует что

$$P(B)=P(A+B\overline{A})=P(A)+P(A\overline{B})\geq A$$

5. Семинар 5. Построение оценок

5.1. t- распределения Стьюдента

$$z = \frac{\mu^* - \overline{x}\sqrt{n}}{\sigma}$$

Распределение Стюдента и использование таблицы распределения Стьюдента этого распределения разберем на следующей задаче

Задача

Задана выборка в виде вариационного ряда - измерения роста 10-ти человек

$$160, 160, 167, 170, 173, 176, 178, 178$$

Нужно проверить, действительно ли средний рост большой группы людей равен $\mu^* = 167$ см.

Пусть случайная величина X - это рост. Пусть эта случайная величина распределена по нормальному закону со средним $\mu^* = 167$ и неизвестной дисперсией $(\sigma)^2$

$$X \sim N(\mu^*, (\sigma^*)^2)$$

Вычислим характеристики выборки. Выборочное среднее есть среднее арифметическое всех элементов выборки : $\overline{x} = 172.4$. Это точечная оценка среднего. Теперь нужно вычислить интервальную оценку среднего при неизвестной дисперсии, чтобы определить попадает ли в этот интервал ожидаемая величина $\mu^* = 167$. Введем величину

$$z = \frac{\mu^* - \overline{x} \cdot \sqrt{n}}{\sigma}$$

распределенную по стандартному нормальному закону $N(0, 1)$.

Если бы дисперсия σ^2 была известна, можно было бы воспользоваться таблицами стандартного нормального распределения и проверить, является ли величина z значимо больше нуля. Но поскольку величина дисперсии не известна, надо сначала ее оценить при помощи выборочной дисперсии s^2

$$(\sigma^*)^2 = s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} = \sum_{i=1}^{1} 0 \frac{x_i - 172.4}{29 = 62.9}$$

т.е $s = 7.93$ Оценка среднеквадратичного отклонения для величины $\overline{x}$ есть

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{62.9}{10}} = 2.51$$

По анологии с z введем величину t

$$t = \frac{\mu^* - \overline{x}\sqrt{n}}{s}$$

Величина t служит критерием проверки, и нам необходимо вычислить ее распреедление для $\mu^* = 167$.

Если переписать выражение для t в виде

$$t = \frac{\frac{\mu^* - \overline{x}}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{s^2}{\sigma^2}}}$$

То числитель оказывается распределен по нормальному закону $\sim N(0,1)$,а квадратный корень из знаменателя имеет также известное распередление, называемое $\chi^2_\gamma(k)$(*"хи-квадрат"*) с $k = n-1$ степенями свободы. Более подробно

$$u = \frac{s^2}{\sigma^2} = \sum_{j=1}^{n-1} y_i^2 \sim \frac{\chi^2_{n-1}}{n-1}$$

Где $Y = \{y_1, y_2, \ldots, y_{n-1}\}$ - распределенная по стандартному нормальному закону случайная величина с независимыми компонентами.

Таким образом, величина t есть функция двух случайных величин, чье распределение известно, а значит, может быть вычислена по известным правилам. Величина t имеет распределение, называемое распределением Стьюдента с числом степеней свободы $k = n-1$. Из таблицы распределения Стьюдента видно, что при большом количестве степеней свободы $k \to \infty$, распределение Стьюдента стремится к нормальному распределению

$$t_{\infty,\gamma} = T^{-1}\left(\infty, \frac{1+\gamma}{2}\right) = u_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

В нашем примере доверительная вероятность или надежность оценки есть

$$P\left(|\mu^* - \overline{x}| < \tilde{\varepsilon}\right) = P\left(|\mu^* - \overline{x}|\right) < \frac{t_{n-1,\gamma} \cdot s}{\sqrt{n}} = P\left(\frac{|\mu^* - \overline{x}|}{s/\sqrt{n}} < t_{n-1,\gamma}\right) = \gamma$$

Подставляя найденные величины

$$\mu^* = 167.0,\ \ \overline{x} = 172.4,\ \ s/\sqrt{n} = 2.51, n = 10$$

Получаем

$$P\left(\frac{|167.0 - 172.4|}{2.51} < t_{9,\gamma}\right) =$$

$$= P(172.4 - 2.51 \cdot t_{9,\gamma} < 167.0 < 172.4 + 2.51 \cdot t_{9,\gamma}) = \gamma$$

Осталось выбрать надежность γ, вычислить с помощью таблицы распределения Стьюдента $t_{9,\gamma}$ и проверить выполнение неравенства.

Согласно стандартным рекомендациям, зададимся доверительной вероятностью $\gamma = 0.9$. Табличное значение $t_{9,0.9} = T^{-1}(9,(1+0,9)/2) = T^{-1}(9,0.95) = 1.833$, т.е с вероятностью 0.9. должно быть $|t_{9,0.9}| \leq 1.8331$. Тогда предполагаемое среднее значение μ^* должно с вероятностью 0.9 лежать в интервале $\{167.8, 177.0\}$. Но в нашем случае это не так. Можно прийти к такому же выводу, сравнив табличное значение $t_{n-1,\gamma}$ с вычисленной статистикой t:

$$\frac{|167.0 - 172.4|}{2.51} = 2.15 > 1.833$$

Следовательно, идея принять средний рост 167 см для данной выборки оказалась неудачной.

Можно было подобрать и другую доверительную вероятность.

Однако надо иметь в виду, что чем больше доверительная вероятность, тем меньше уровень значимости и, следовательно, тем менее точен результат. Так, к примеру, для уровня значимости 0.05% доверительный интервал станет большим, в нашем случае он станет $\{160.4, 184.4\}$ и, хотя он покроет значение 167.0, никакой практической ценности иметь уже не будет.

5.2. Односторонняя оценка

Задача

Старые автомобили имеют расход топлива 10 литров на 100 километров. Для уменьшения расхода двигатели автомобилей изменили. При проверке 25 новых машин оказалось что их средний расход топлива составляет $\overline{x} = 9.3$ литра. Проверить действительно ли уменьшился расход топлива в новых автомобилях если также известно, что они распределены по нормальному закону с дисперсией $\sigma^2 = 4$ литра^2.

Решение

Построим статистику для нормального стандартного распределения

$$u = \frac{\overline{x} - 10}{\sqrt{4/25}} \sim N(0, 1)$$

Зададим процентную точку $\alpha = 5\%$.(Так как мы берем только половину интервала)

Нам нужно сравнить величину U с квантилью нормального распределения $u_{0.90} = -1,645$

$$u = \frac{9.3 - 10}{2/5} = -1.75 < -1.645$$

Это означает что оценка $\overline{x}$ является хорошей, и расход топлива действительно уменьшился.

Найдем *критическую область* $\hat{\overline{x}}$ для данной задачи, которая определит верхнюю допустимую границу для $\overline{x}$ при которой еще можно говорить что расход топлива уменьшается

$$\frac{\hat{\overline{x}} - 10}{2/5} = -1.645$$

$$\hat{\overline{x}} = 9.342$$

То есть при $\overline{x} < 9.342$ можно считать что расход топлива уменьшается.

5.3. Понятие ошибки 1-ого и 2-ого рода

Эти ошибки возникают в задачах на односторонние интервалы. Они возникают когда статистические предположения оказываются ошибочными

Ошибка первого рода обозначается $I(\alpha)$ и конретно для задачи про автомобили означает вероятность того, что старый автомобиль будет ошибочно принят за новый. Здесь $\alpha = 5\%$. Для того чтобы посмотреть как устроена эта ошибка, изменим

критическая область предыдущей задачи $\hat{\overline{x}} < 9.44$ Найдем α, то есть вероятность того, что $\overline{x}$ попало в критическую область, при условии, что старый автомобиль с расходом 10 литров ошибочно посчитали за новый т.к его значение по средним характеристикам попала в критическую область

$$\alpha = \Phi\left(\frac{9.44 - 10}{\sqrt{4/25}}\right) = \Phi(-1.4) = 8\%$$

Ошибка второго рода обозначается $I(\beta)$ и для задачи про автомобили означает вероятность того, что новый автомобиль будет ошибочно принят за старый. Еще раз изменим условия задачи заменив 10 литров на 9. Посмотрим какой будет ошибка β при условии что автомобиль классифицируется как старый(то есть его среднее не попадает в критическую область) при этом имея расход 9 литров.

$$\beta = 1 - \Phi\left(\frac{9.44 - 9}{\sqrt{4/25}}\right) = 1 - \Phi(1.1) \approx 0.136$$

При таких условиях задачи получается что примерно $13,6\%$ автомобилей имеющие расход 9 литров классифицируются как имеющие больший расход топлива.

Изобразим графически ошибку первого и второго рода рис(5.1).

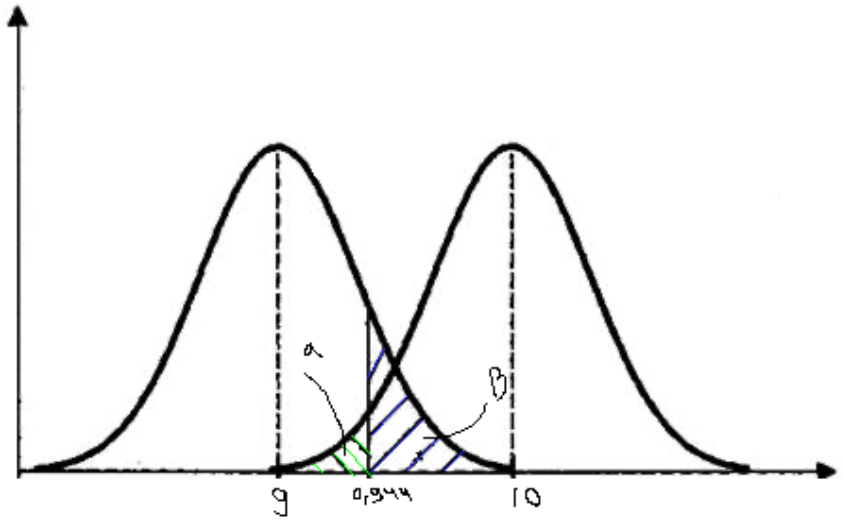

Рис. 5.1 – Иллюстрация к задаче

Если задано α, то β можно уменьшить если объем выборки увеличить. Если задано и α и β, то можно определить минимальный объем выборки который нужен.

Пример

Остановимся на том же примере с машинами. Пусть расход топлива старых автомобилей равен 10, а новых 9. При этом $\sigma^2 = 4$ литра, $\alpha = 0.01$, а β должно быть меньше 0.1. Узнаем, сколько машин n требуется взять чтобы эти условия выполнялись.

$$\begin{cases} \Phi\left(\frac{\hat{\overline{x}}-10}{\sqrt{4/n}}\right) = 0.01 \\ 1-\Phi\left(\frac{\hat{\overline{x}}-9}{\sqrt{4/n}}\right) \leq 0.1 \end{cases}$$

Решим уравнения для этих функций

$$\begin{cases} \frac{\hat{\overline{x}}-10}{2}\sqrt{n} = -2.326 \\ \frac{\hat{\overline{x}}-9}{2}\sqrt{n} \geq 1.282 \end{cases}$$

Отсюда получаем $n \geq 53$

Задача

Имеется две выборки

$$I - \quad 14.2 \quad 10.1 \quad 14.7 \quad 13.7 \quad 14.0$$

$$II - \quad 14.0 \quad 14.5 \quad 13.7 \quad 12.7 \quad 14.1$$

Каждая из них выбрана из нормальной генеральной совокупности. Процентная точка $\alpha = 10\%$ для доверительного интервала 0.9. Одинаковы ли эти выборки?

Решение

Для того чтобы выборки были одинаковы необходимо равенство дисперсий и средних, начнем с первого

$$\frac{s_I^2}{s_{II}^2} = \frac{\sigma_I^{*2}}{\sigma_{II}^*} \approx \frac{3.37}{0.46} \approx 7.33$$

Нужно сравнить полученное число со статистикой Фишера

$$F_{1-\frac{\alpha}{2}}(n_I - 1, n_{II} - 1) = F_{0.95}(4,4) = 6.39$$

Получаем, что

$$7.33 > 6.39$$

Расчетная статистика оказалась больше чем табличная, значит предположение о равенстве дисперсий неверно.

Из этого одного уже следует что выборки не равны. Тем не менее все равно сравним математическое ожидание.

$$\overline{x}_I = 13.32, \quad \overline{x}_{II} = 13.80$$

$$\frac{|\overline{x}_I - \overline{x}_{II}|}{\sqrt{\frac{s_I^2}{n_I} + \frac{s_{II}^2}{n_{II}}}} = 0.55$$

Сравним полученную величину с распределением Стьюдента $t_{0.95}$, осталось понять число степеней свободы

$$\frac{\left(\frac{\sigma_I^2}{n_I}+\frac{\sigma_{II}^2}{n_{II}}\right)^2}{\frac{\left(\frac{\sigma_I^2}{n_I}\right)}{n_I+1}+\frac{\left(\frac{\sigma_I^2}{n_{II}}\right)}{n_{II}+1}}-2\approx 6$$

Получаем

$$t_{0.95}(6)=1.943$$

$$u_{0.95}=1.960$$

В итоге

$$0.55<1.960 \qquad 0.55<1.943$$

Мы видим что в данной задаче мы могли обойтись только нормальным распределением, но так бывает не всегда. Например, если вместо величины 0.55 мы бы получили 1.950, то, во-первых, распределение Стьюдента оказалось бы более точным. Так же это указывало бы на то, что необходимы более точные статистические исследования и данные изначально могли бы быть не из нормальной генеральной совокупности. Окончательно делаем вывод, что средние равны, но тем не менее, выборки разные.

6. Семинар 6. Регрессия

Линейная регрессия ищется в виде

$$y = a + bx$$

Это система условных уравнений, по ней ищется система нормальных уравнений как производная суммы не вязок по параметру α и по параметру β

$$\begin{cases} \frac{\partial S}{\partial \alpha} = 0 \\ \frac{\partial S}{\partial \beta} = 0 \end{cases}$$

Выпишем нормальную систему

$$\begin{cases} n\alpha + b\sum x_i = \sum y_i \\ a\sum x_i + \beta \sum x_i^2 = \sum y_i x_i \end{cases} \tag{11}$$

Из системы (11) можно в общем виде через x_i, y_i выписать коэффициенты α и β

$$\alpha = \frac{\sum y_i - b\sum x_i}{n}$$
$$\beta = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Для проверки вычислений можно воспользоваться равенством

$$\overline{y} = a + b\overline{x}$$

Задача
Задана таблица чисел

x_i	1.2	2.4	2.8	4.2	5.9	6.8	8.1	9.2	10.1	11.0
y_i	7	12	17	24	29	38	46	45	54	68

(12)

Постройте доверительные интервалы для параметров линейной регрессии

Решение
Вычислим необходимые величины

$$\sum x_i = 61.7 \qquad \sum x_i y_i = 2695.1$$
$$\left(\sum x_i\right)^2 = 3806.89 \qquad \beta^* = 5.6189$$
$$\sum x_i^2 = 489.99 \qquad \alpha^* = -0.669$$
$$\sum y_i = 340 \qquad \overline{x} = 6.17$$

β тангенс угла наклона прямой которая может быть проведена по данным таблицы 12. Построим доверительный интервал для β с помощью распределения Стьюдента

$$\beta = \beta^* \pm T^{-1}\left(n-2, \frac{1+\gamma}{2}\right) s_\beta$$

Число параметров в квантильи Стьюдента равно разности количества точек и количества параметров регрессионной модели. γ примем равной 0.95.

$$s_\beta = \frac{s}{s_x\sqrt{n-1}}$$

Величина s^2 это нормированная сумма квадратов невязок

$$s^2 = \frac{1}{n-2}\sum (y_i - a^* - b^* x_i)^2$$

s_x^2 это дисперсия x

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2 = 11.8112$$

Величина s_β имеет такую структуру, что при делении на разность точного β и приближенного β^* найденного по выборке дает величину которая имеет распределение Стьюдента

$$\frac{\beta - \beta^*}{s_\beta} \sim t_{n-2,\gamma}$$

Так как нам изначальна не заданно ни дисперсия, ни математическое ожидание, чтобы сравнить как сильно отличается β от β^*, нам необходимо сравнить их с квантилью Стюдента. Продолжим вычислять величины

$$S_x = 3.4367$$

$$S^2 = \frac{1}{n-2}\sum (y_i - a^* - b^* x_i)^2 = 13.4755$$

$$s_\beta = 0.356$$

$$t_{8,0.975} = 2.306$$

Если линейное приближение правильно, то β должен быть отличен от нуля

$$|\beta - \beta^*| = 0.356 \cdot 2.306 = 0.821$$

$$4.8 \le \beta \le 6.462$$

Далее примем $\beta = 5$. Построим доверительный интервал для α

$$\alpha = \alpha^* T^{-1}(n-2)\left(n-2, \frac{1+\gamma}{2}\right) s_\alpha$$

$$\frac{\alpha - \alpha^*}{s_a} \sim t$$

$$s_a = s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} = 2.485$$

$$-6.399 \leq \alpha \leq 5.062$$

Отсюда следует что мы можем принять α равным нулю, получаем

$$y = 5x$$

6.1. Коэффициент детерминации

Коэффициент детерминации R^2 показывает связь меру вклада регрессии в общее отклонение от среднего или, другими словами, определяет корреляцию y и его регрессии

$$R^2 = \frac{\sum(\alpha^* + \beta^* x - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

Коэффициент детерминации для задачи предыдущей равен

$$R^2 = 0.968$$

Это означает что 96.8 процентов отклонения всех данных от общего значения описывается уравнением линейной регрессии. Коэффициент детерминации можно связать с распределением Фишера

$$R^2 = \frac{F}{F + (n-2)}$$

Где F величина имеющая распределение фишера

$$F = \frac{\left(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\right)(n-2)}{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right] - \left(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\right)^2}$$

Величина F сравнивается с табличным значением $F(1, n-2) = (t_{n-2})^2$ и если

$$F > 4F(1, n-2)$$

Значит линейная регрессия хорошо описывает отклонение от среднего

6.2. Анализ остатков

Остатком называют величину

$$e_i = y_i - \alpha^* - \beta^* x_i$$

При этом

$$\sum e_i = 0$$

Остатки можно представить в виде суммы случайной и систематической компоненты

$$e_i = q_i + B_i$$

$$q_i = (y_i - \alpha^* - \beta^* x_i) - (M[Y] - M[\alpha^* + \beta^* x])$$

$$B_i = M[Y] - M[\alpha^* + \beta^* x]$$

Если B_i равно 0, то регресссионная модель называется корректной. Это значит что мы так аппроксимировали данные, что среднее по данным и среднее по модели оказались одинаковыми. Для корректной модели остатки являются наблюдаемыми ошибками. Они должны быть распределены по нормальному закону и должны обладать нулевой дисперсией и нулевым средним. Если B_i не равно нулю тогда нужно изучать ошибки в следующих направлениях

1) Исследование e_i на нормальное распределение.
2) Исследование на постоянство дисперсии для e_i.

6.3. Нелинейная регрессия

Пример

Задана зависимость температуры от времени

t,мин	5	10	15	20	25
T, °C	59.3	59.8	60.1	64.9	70.2

Будем считать что $T = \alpha + \beta t + t^2$

Для упрощения расчетов введем новые переменные

$$x = \frac{t - 15}{5}, \qquad y = 10(T - 60)$$

Нужно минимизировать

$$(y_i - \beta_0 - \beta_1 x - \beta_2 x^2)$$

По параметрам $\beta_0, \beta_1, \beta_2$. Все отличие от линейной регрессии заключается в том, что вместо двух уравнений теперь будет три. Выпишем систему для трех уравнений

$$\begin{cases} \beta_0 n + \beta_1 \sum x_i + \beta_2 \sum x_i^2 = \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 = \sum y_i x_i \\ \beta_0 \sum x_i^2 + \beta_1 \sum x_i^3 + \beta_2 \sum x_i^4 = \sum y_i x_i^2 \end{cases}$$

Подставим все суммы которые имеются в уравнениях

$$\begin{cases} 5\beta_0 + 10\beta_2 = 143 \\ 10\beta_1 = 269 \\ 10\beta_0 + 34\beta_2 = 427 \end{cases}$$

Отсюда получаем

$$\begin{cases} \beta_0 = 8.457 \\ \beta_1 = 26.9 \\ \beta_2 = 10.07 \end{cases}$$

Вернемся к старым переменным

$$(T-60)\cdot 10 = 8.457 + 26.9\cdot\frac{t-15}{5} + 10.07\left(\frac{t-15}{5}\right)^2$$

$$T = 61.84 - 0.67t + 0.04t^2$$

7. Семинар 7. Соответствие выборки конкретному распределению, критерии

Задача

Исследуется работа некоторых приборов в количестве $n = 757$. У них есть сбои, которые выражаются таким параметром как число отказов

число отказов, k	количество случаев в которых наблюдались отказы
0	427
1	235
2	72
3	21
4	1
5	1
≥ 6	0

Имеет ли распределение числа отказов распределение Пуассона?

Решение

$$P_k = P(x = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$\overline{\lambda} = \lambda^* = \frac{0 \cdot 427 + 1 \cdot 235 + \ldots + 5 \cdot 1}{757} = \frac{451}{757} \approx 0.6$$

Мы ожидаем, что если распределение подчиняется закону Пуассона $\lambda = 0.6$. Продолжим таблицу для P_k и $n \cdot P_k$

число отказов, k	количество случаев в которых наблюдались отказы	$p_k = \frac{0.6^k}{k!} \cdot e^{-0.6}$	n$\cdot P_k$
0	427	0.549	416
1	235	0.329	249
2	72	0.099	75
3	21	0.019	15
4	1	0.0029	2
5	1	0.00036	0
≥ 6	0	0.00004	0

Для некоторых k величина $n \cdot p_k$ меньше чем 5 и для них мы не можем использовать критерий хи-квадрат, поэтому объединим интервалы для $k = 3, 4, 5$ и ≥ 6 в один

число отказов, k	$n \cdot p_k$	$\frac{(n_k - n \cdot p_k)^2}{n \cdot p_k}$
0	416	0.291
1	249	0.787
2	75	0.120
≥ 3	17	2.118

Сумма элементов третьего столбца равна 3.316 это Расчетная характеристика хи-квадрата, сравним это значение с табличным

$$\chi^2(4-1-1) = \chi(2) = 9.21, \qquad \gamma = 0.99$$

Расчетное значение оказалось меньше табличного, значит данные точки имеют распределение Пуассона.

Пример на нормальное распределение

Есть 55 точек которые необходимо проверить на соответствие нормальному распределению. Задан самый большой $x_{54} = 23.8$ и самый маленький элемент $x_o = 10.1$

Для использования критерия хи-квадрат нужно определить на какое количество интервалов нам следует разделить наши данные для этого нужно найти размах выборки

$$J_n(x) = 23.8 - 10.1 = 13.7$$

Длина интервала группировки

$$\frac{13.7}{7} \approx 2$$

Составим таблицу

$\triangle_k$	n_k	p_k
$(-\infty, 12]$	2	1.26
[12,14]	4	3.89
[14,16]	8	9.18
[16,18]	12	13.95
[18,20]	15	13.63
[20,22]	11	8.56
$[22, +\infty)$	3	4.52
	$\sum = 55$	$\sum 1 \cdot 55$

Для того что бы сравнивать наше распределение не с усеченным нормальным распределением, а с обычным нормальным, мы включили $-\infty$ и $+\infty$

$$p_k = \Phi\left(\frac{\beta_k - \overline{x}}{s}\right) - \Phi\left(\frac{\alpha_k - \overline{x}}{s}\right), \qquad k = \overline{1,7}$$

Для некоторых k величина $n \cdot p_k$ меньше чем 5, поэтому объединим первые и последние два интервала и постром статистику хи-квадрат

$\triangle_k$	n_k	p_k	$\frac{(n_k - n\cdot p_k)^2}{n\cdot p_k}$
$(-\infty, 14]$	6	5.14	0.14
[14,16]	8	9.18	0.15
[16,18]	12	13.95	0.27
[18,20]	15	13.63	0.14
$[20, +\infty)$	14	13.09	0.06
	$\sum = 55$	$\sum 1 \cdot 55$	$\sum = 0.77$

при этом

$$\chi^2(5-2-1) = \chi^2(2) = 4.61, \qquad \gamma = 0.9$$

Где 5 - число интервалов, 2 - число параметров(среднее и дисперсия)

Окончательно, сравнивая полученную нами вероятность и табличную мы видим что точки распределены по нормальному закону.

7.1. Критерий Колмогорова

Критерий Колмогорова необходим для сопоставления полученных данных с заданным распределением, то есть может являться альтернативой критерию хи-квадрат. В качестве меры сравнения выбирается расстояние

$$D_n = \underset{|x|<\infty}{max}|F_n^*(x) - F(x)|$$

Где $F_n^*(x)$ - эмпирическая функция распределения, а $F(x)$ - теоретическая функция распределения

А.Н. Колмогоровым было доказано, что закон распределения величины

$$\Lambda = D_n\sqrt{n}, \quad n \to \infty$$

определяется функцией

$$F(\lambda) = \sum_{k=-\infty}^{\infty}(-1)^k exp\{-2k^2\lambda^2\}$$

Критерий Колмогорова используется при больших выборках.

7.2. Непараметрические критерий

Непараметрические критерий используется когда распределение генеральной совокупность либо неизвестно, либо к нему тяжело подобрать аналитическую функцию распределения. Эти методы не нуждаются в знание распределения. Например, в методе наименьших квадратов знать закон распределенияне было не нужно было, в то время как это знание необходимо для метода максимального правдоподобия. Основная идея использования непараметрических критериев это изучение структуры однородной выборки составленной из двух выборок, которые мы хотим сравнить.

7.2.1. Критерий знаков

Скорость автомобиля измеряется двумя приборами. С помощью непараметрических критериев необходимо проверить завышена ли оценка скорости для одного из приборов.

Пример

v_1	70	85	63	54	65	80	75	95	52	55
v_2	72	86	62	55	63	80	78	90	53	57

Построим разность $v_1 - v_2$ и посчитаем количество выражений с плюсами и минусами Если приборы меряют одинаково то, вероятность появления плюсов и минусов также должна быть одинакова. Для того чтобы проверить, что количество плюсов и минусов равно используется следующая статистика. Обозначим за r количество плюсов, l - общее число разностей без нулевого элемента, $\alpha = 1 - \gamma$ уровень значимости. Построим статистику

$$\frac{l-r}{1+r} = 1.5$$

и сравним ее с распределением Фишера

$$F_{1-\frac{\alpha}{2}}(2(r+1)2(l-r)) = F_{0,9}(8,12) = 2.24$$

В нашем случае $r = 3$, $l = 9$.

Получаем, что ни один из приборов не завышает оценку скорости, а отклонения в показателях носят чисто случайный характер.

7.3. Критерий Уилкоксона-Манна-Уитни

Пример

Необходимо сравнить две выборки разного объема

$$\begin{aligned} I \ - \quad & 39, 50, 61, 67, 40, 40, 54 \\ II - \quad & 60, 53, 42, 41, 40, 54, 63, 69 \end{aligned}$$

Выпишем объединенную выборку таким, образом чтобы ее элементы были построены в виде вариационного ряда. Будем отмечать элементы взятые из первой выборки чертой сверху

Объединенная выборка	$\overline{39}$	$\overline{40}$	$\overline{40}$	40	41	42	$\overline{50}$	53	$\overline{54}$	54	60	$\overline{61}$	63	$\overline{67}$	69
Ранг	1	3	3	3	5	6	7	8	9.5	9.5	11	12	13	14	15

Ранг элементов равен среднему арифметическому их порядковых номеров в объединенной выборке, например для элемента 40 он равен $\frac{2+3+4}{3}$. Для статистики нужно посчитать сумму рангов для элементов первой и второй выборки

$$\sum_I R = 1 + 3 + 3 + 7 + 9.5 + 12 + 14 = 49.5$$
$$\sum_{II} R = 70.5$$

$$\omega_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$\omega_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$\omega_1 + \omega_2 = n_1 n_2$$

$$\omega = min(\omega_1, \omega_2) = \frac{\omega - \frac{1}{2}n_1 n_2}{\sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}} \sim N(0, 1)$$

Окончательно, эту величину нужно сравнить с квантилью нормального распределения u_γ.

ФИЗИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in

ЛЕКЦИИ УЧЕНЫХ МГУ