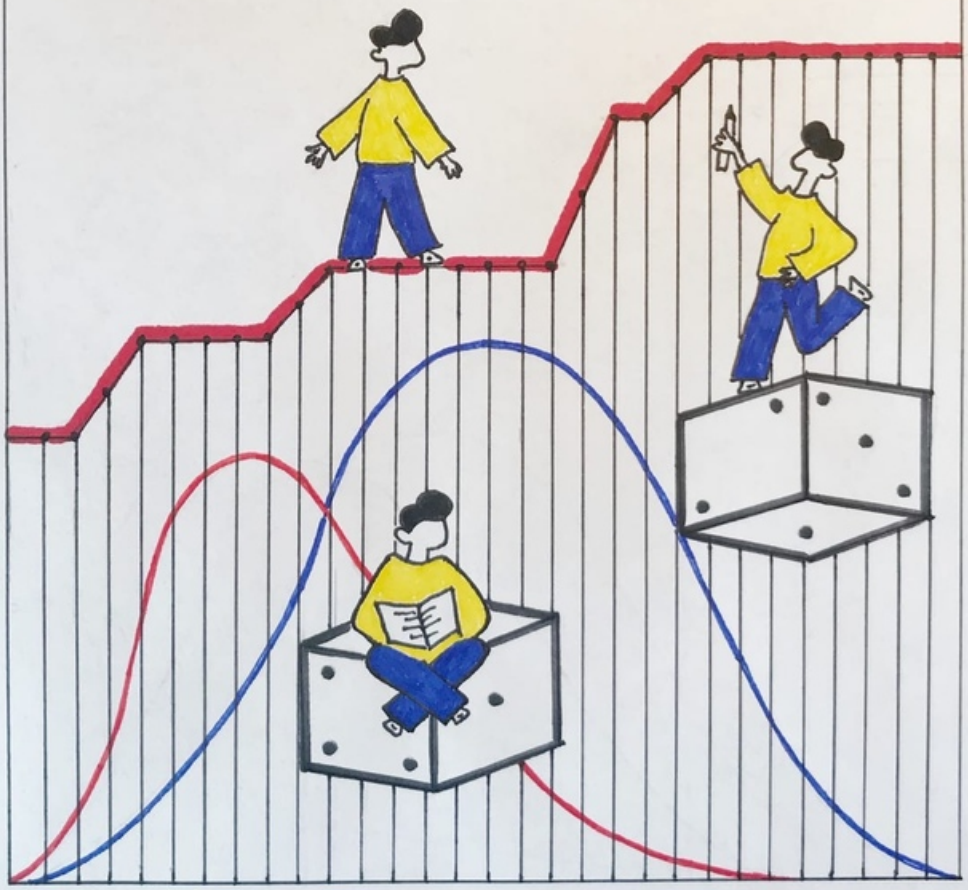
Теория вероятностей и математическая статистика

Громова Е.В., Тур А.В.



Санкт-Петербург, 2020

Посвящаем этот конспект лекций нашим мужьям и детям, мужественно выдержавшим дистанционное обучение весной и летом 2020 года. С любовью Дмитрию, Леше, Максиму, Саше, Николаю, Серёже, Мише

Конспект по курсу “Теория вероятностей и математическая статистика” был создан во время пандемии новой коронавирусной инфекции весной 2020 г.

Благодарим за помощь в компьютерной верстке конспекта Голокоза А., Портнова В.
Иллюстрации: студенты 1 курса псих.факультета СПбГУ Потёмкина Д., Митрофанова Н.

ОБНОВЛЁННАЯ ВЕРСИЯ С НАЧАЛЬНЫМИ ГЛАВАМИ ОЖИДАЕТСЯ

Содержание

1 Непрерывные случайные величины

1.1 Функция распределения случайной величины

Пусть $X : \Omega -> R^1$ — случайная величина (с.в.). *Функция распределения (ф.р.)*:

$$F(x) = P\{X < x\}.$$

Замечание 1. Иногда неравенство задается нестрого: $F(x) = P\{X \leq x\}$. В таком случае ф.р. непрерывна справа.

Замечание 2. Ф.р. $F(x) = P\{X < x\}$ всегда существует и является универсальным способом задания закона распределения случайной величины.

Замечание 3. Англ. название ф.р. $F(x) = P\{X < x\}$: distribution function or cumulative distribution function (c.d.f.). Последнее название объясняет суть ф.р., которая показывает, как "аккумулируются", т.е. накапливаются вероятности при увеличении x. В этом смысле иногда используется понятие "кумулятивной функции распределения".

Замечание 4. Для дискретных случайных величин график ф.р. $F(x) = P\{X \leq x\}$ всегда имеет ступенчатый вид (см. ниже).

Замечание 5. Классификация случайных величин осуществляется на основе свойств ф.р.: если $F(x)$ является непрерывной функцией, то с.в. X называется непрерывной, если функция $F(x)$ имеет разрывы первого рода (скачки), то с.в. X называется дискретной с.в.

Свойства функции распределения:

1. $F(x) \in [0, 1]$
2. $F(x)$ - неубывающая функция: $\forall x_2 > x_1 \;\; F(x_2) \geq F(x_1)$.

3. $P\{x_1 \leq X < x_2\} = F(x_2) - F(x_1)$.
4. Ф.р. непрерывна слева в любой точке x_0, т.е. $F(x_0 - 0) = \lim\limits_{x \to x_0 - 0} F(x) = F(x_0)$.
5. $\lim\limits_{x \to \infty} F(x) = 1$
   $\lim\limits_{x \to -\infty} F(x) = 0$

Доказательство (упрощенное):

1. $F(x)$ как вероятность принимает значения от 0 до 1.
2. $F(x_2) = p\{X < x_2\} = P\{X < x_1\} + P\{x_1 \le X < x_2\} = F(x_1) + P\{x_1 \le X < x_2\}$

   Следствие. $P\{x_1 \le X < x_2\} = F(x_2) - F(x_1)$.
3. $\lim\limits_{x\to\infty} F(x) = \lim\limits_{x\to\infty} P\{X < x\} = 1$ (достоверное событие)
4. Аналогично $\lim\limits_{x\to-\infty} F(x) = 0$ (невозможное событие)

Для (простых) дискретных случайных величин ф.р. F(x) имеет следующий вид:

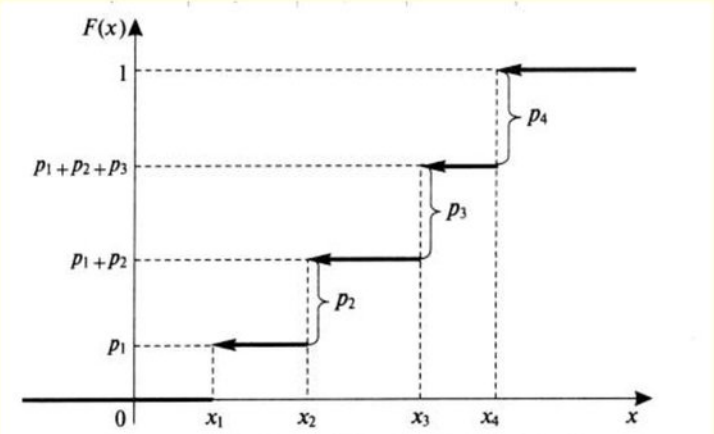
$$F(x) = \begin{cases} 0, & x \le x_1, \\ p_1, & x_1 < x \le x_2, \\ p_1 + p_2, & x_2 < x \le x_3, \\ \dots & \dots, \\ 1, & x > x_n \end{cases}. \tag{1}$$

Заметим, что величина скачка в каждой точке x_i равна соответствующей этому возможному значению случайной величины вероятности p_i. Сумма вероятностей “накапливается” и доходит до 1.

2. ***Интегральная функция распределения является неубывающей***:

$$F(x_2) > F(x_1), \text{ если } x_2 > x_1$$

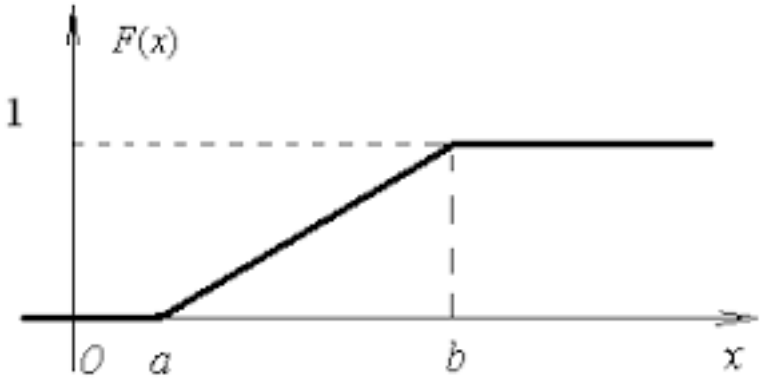
3. ***Функция распределения любой дискретной случайной величины есть разрывная ступенчатая функция***, скачки которой происходят в точках, соответствующих возможным значениям случайной величины и равны вероятностям этих значений. ***Сумма всех скачков равна*** 1.



Функция распределения дискретной случайной величины X, принимающей 4 возможных значения

11

Приведем пример графика ф.р. для непрерывной с.в.



По данному графику можно легко найти медиану распределения, первый и третий квартили, квантиль порядка p (см. далее).

1.2 Функция плотности распределения случайной величины

Случайная величина X называется *абсолютно непрерывной*, если $\exists f(x) \geq 0$:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

$f(x)$ - *плотность* распределения вероятности.

Иначе: $f(x) = F'(x)$.

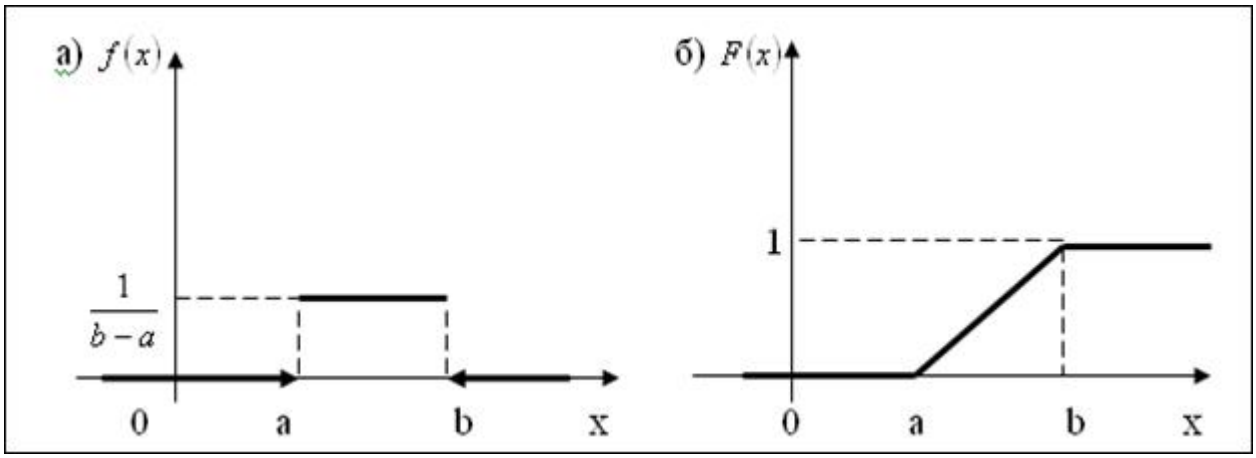
Замечание 1. Иногда требование неотрицательности $(x) \geq 0$ убирают из определения (см. свойство 1 ниже).

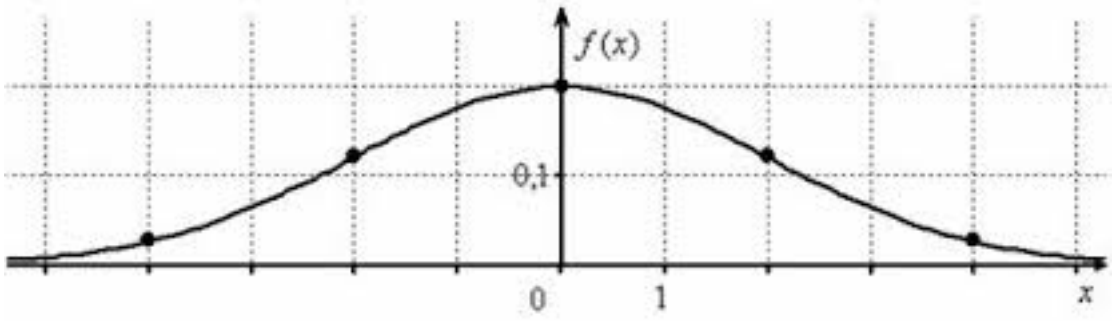
Свойства функции плотности:

1. $f(x) \geq 0$ так как $f(x) = F'(x)$ ($F(x)$ — неубывающая функция)
2. $$P\{x_1 \leq X < x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$$
3. Условие нормировки $\int_{-\infty}^{+\infty} f(x)dx = 1$

Графическое изображение функции плотности распределения с.в. называется кривой распределения.

Из условия нормировки следует, что на графике площадь под кривой всегда должна быть равной 1.





1.3 Числовые характеристики абсолютно непрерывных случайных величин

1. Математическое ожидание:

$$M(X) = \int_{-\infty}^{+\infty} x \cdot f(x)dx. \tag{2}$$

2. Дисперсия:

$$D(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 \cdot f(x)dx = \tag{3}$$
$$= \int_{-\infty}^{+\infty} x^2 \cdot f(x)dx - [M(X)]^2.$$

3. Начальный момент порядка k:

$$m_k = \int_{-\infty}^{+\infty} x^k \cdot f(x)dx. \tag{4}$$

4. Центральный момент порядка k:

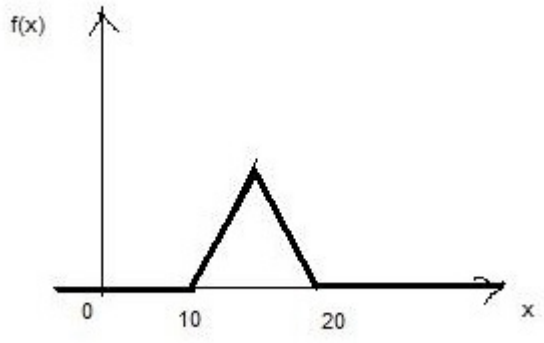
$$\nu_k = \int_{-\infty}^{+\infty} (x - M(X))^k \cdot f(x)dx. \tag{5}$$

1.4 Задачи

Задача 1

Спрос на товар фирмы есть случайная величина X (тыс. штук), плотность распределения которой имеет следующий график (закон Симпсона).

Выписать аналитическую формулу для $f(x)$. Определить вероятность того, что спрос будет меньше 10 тыс. штук; 15 тыс. штук. Поясните на графике. Найдите также математическое ожидание и дисперсию.



Решение

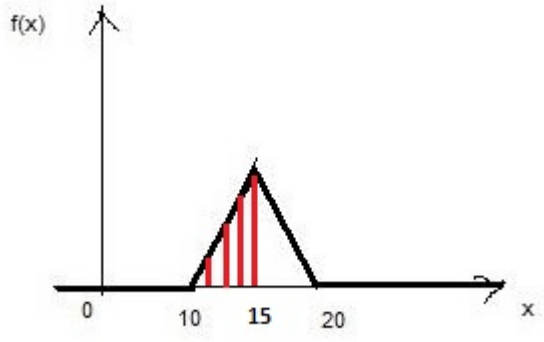
Площадь под графиком функции плотности равна 1. Значит площадь треугольника равна 1:

$$S = \frac{1}{2}h(20-10) = 1,$$

$$h = 0,2.$$

$$f(x) = \begin{cases} 0, & \text{при } x \le 10, \\ 0,04x - 0,4, & \text{при } 10 < x \le 15, \\ -0,04x + 0,8, & \text{при } 15 < x \le 20, \\ 0, & \text{при } x > 20. \end{cases}$$

Вероятность того, что спрос будет меньше 10 тыс. штук равна 0. Вероятность того, что спрос будет меньше 15 тыс. штук равна 0,5, т.к. она равна площади заштрихованного на рисунке треугольника:



Распределение является симметричным унимодальным, значит мода совпадает с

математическим ожиданием, с медианой и с абсциссой центра симметрии графика плотности распределения, т.е. $E\xi = 15$.

$$D\xi = \int\limits_{10}^{15} x^2(0,04x - 0,4)dx + \int\limits_{15}^{20} x^2(-0,04x + 0,8)dx - (E\xi)^2 = \frac{25}{6}.$$

Задача 2

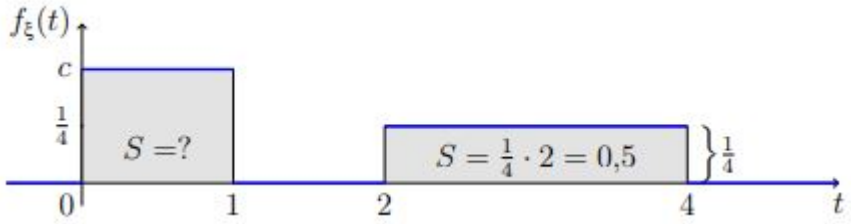
Плотность распределения случайной величины ξ известна с точностью до константы c:

$$f_\xi(x) = \begin{cases} c, & \text{при } x \in [0,1], \\ \frac{1}{4}, & \text{при } x \in [2,4], \\ 0, & \text{при } x \notin [0,1] \cup [2,4]. \end{cases}$$

Найти c и функцию распределения ξ. Нарисовать её график.

Решение:

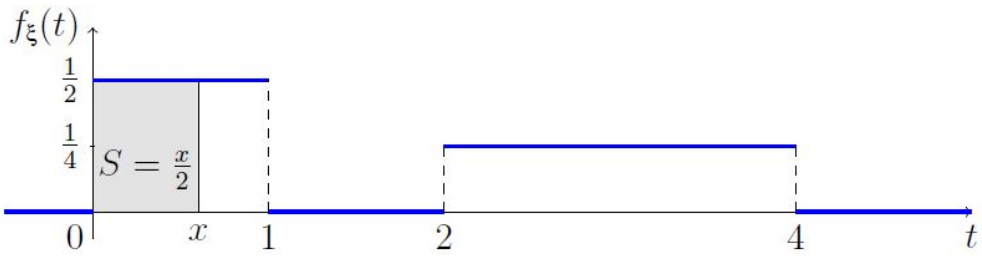
Нарисуем график плотности и найдём c из свойства нормировки: такое, чтобы площадь под графиком плотности равнялась единице.



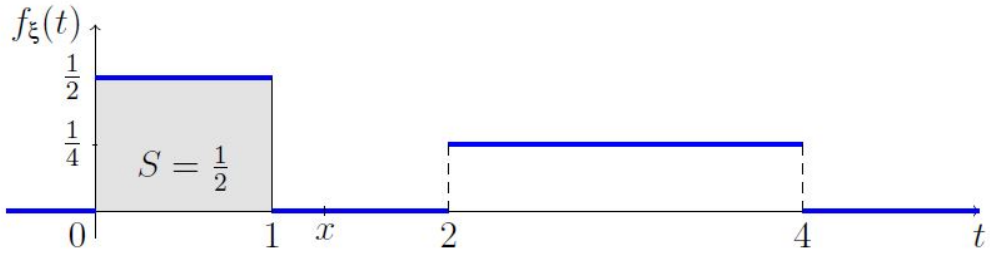
Площадь, помеченная вопросом, тоже должна равняться 0,5, откуда $c = 0,5$.

Что такое $F_\xi(x) = P(\xi < x)$? Это площадь под графиком плотности слева от точки x. Для каждого x отдельно: $x < 0$, $0 < x \leq 1$, $1 < x \leq 2$, $2 < x \leq 4$, $x > 4$, ищем функцию распределения как всю площадь под графиком плотности слева от x.

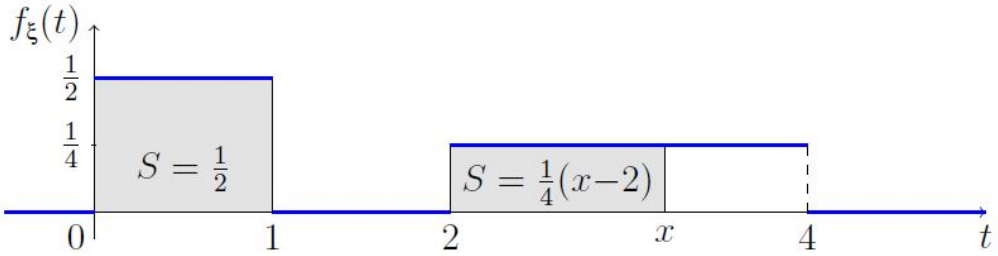
1. При $0 < x \leq 1$ площадь, выделенная на рисунке, равна $F_\xi(x) = \frac{x}{2}$



2. При $1 < x \leq 2$ площадь, выделенная на рисунке, равна $F_\xi(x) = \frac{1}{2}$



3. При $2 < x \leq 4$ площадь, выделенная на рисунке, равна $F_\xi(x) = \frac{1}{2} + \frac{x-2}{4}$



Тогда:

$$F_\xi(x) = \begin{cases} 0, & \text{при } x \leq 0, \\ \frac{x}{2}, & \text{при } 0 < x \leq 1, \\ \frac{1}{2}, & \text{при } 1 < x \leq 2, \\ \frac{x}{4}, & \text{при } 2 < x \leq 4 \\ 1, & \text{при } x > 4. \end{cases}$$

Задача 3

Плотность распределения случайной величины ξ имеет вид:

$$f_\xi(x) = \frac{a}{e^{-x} + e^x}.$$

Найти константу a и вероятность того, что случайная величина ξ примет значение меньше 1.

Решение

Найдём a, используя условие нормировки:

$$\int\limits_{-\infty}^{+\infty} \frac{a}{e^{-x} + e^x} dx = 1.$$

$$\int\limits_{-\infty}^{+\infty} \frac{a}{e^{-x} + e^x} dx = \int\limits_{-\infty}^{+\infty} \frac{a de^x}{1 + e^{2x}} = \int\limits_{0}^{+\infty} \frac{a dt}{1 + t^2} = a \cdot \operatorname{arctg} t \mid_0^{+\infty} = \frac{a\pi}{2}.$$

$$\frac{a\pi}{2} = 1 \Rightarrow a = \frac{2}{\pi}.$$

$$F_\xi(x) = \int\limits_{-\infty}^{x} \frac{2}{\pi(e^{-s} + e^s)} ds = \frac{2}{\pi} \operatorname{arctg} e^x.$$

Вероятность того, что случайная величина ξ примет значение меньше 1, равна $F_\xi(1) = \frac{2}{\pi} \operatorname{arctg} e$.

2 Основные распределения непрерывных случайных величин

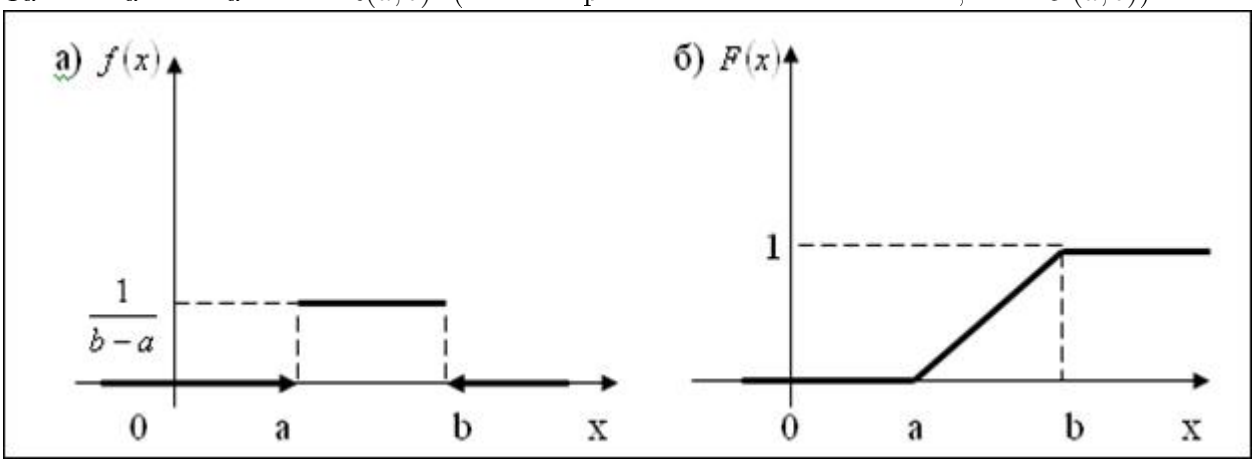
2.1 Равномерное, экспоненциальное (показательное) распределение

1. *Равномерное распределение.* (Ошибки грубых округлений, интервал ожидания в транспорте и т.д.)

С.в. X - равномерно распределённая величина, если

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b], \\ o & otherwise \end{cases}. \tag{6}$$

Записывается как $X \sim R(a,b)$. (Англ. версия - uniform distribution, $X \sim U(a,b)$)



Заметим, что площаль под графиком кривой равна 1!

$$M(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_a^b x \cdot \frac{1}{b-a}dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

$$D(X) = \frac{(b-a)^2}{12}$$

Задание: вывести формулу для ф.р. $F(x)$.

2. *Экспоненциальное (показательное) распределение.* (Время жизни простых технических объектов)

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$F(x) = \int_{-\infty}^{x} f(t)dt = \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}$

Тогда

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ o & otherwise. \end{cases} \tag{7}$$

Записывается как $X \sim Exp(\lambda)$

Замечание. Интервал времени T между двумя событиями в простейшем (Пуассоновском) потоке имеет распределение с плотностью

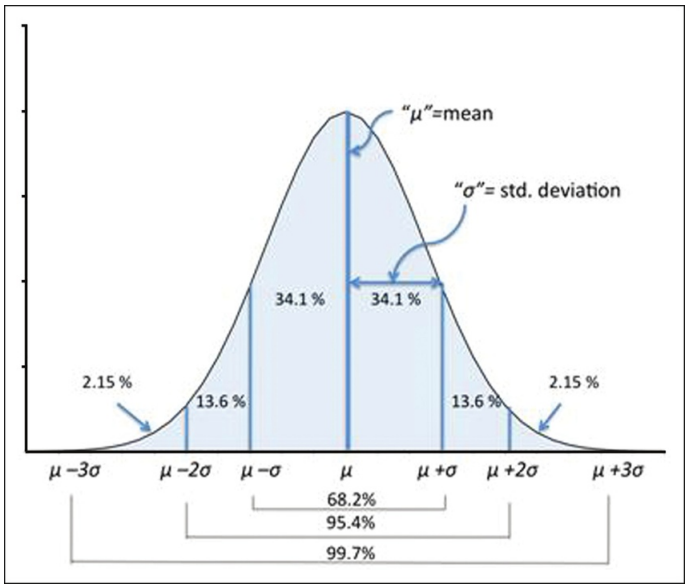
$$f(t) = \lambda \cdot e^{-\lambda \cdot t}.$$

Действительно,

$F(t) = P\{T < t\} = 1 - P_t(0) = 1 - e^{-\lambda \cdot t} \Rightarrow f(t) = \lambda \cdot e^{-\lambda \cdot t},\ t \geq 0$

3. Нормальное распределение (см. ниже)

2.2 Нормальное (гауссово) распределение



$X \sim N(a, \sigma^2)$, если:

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

$a = M(X),\ \ \sigma = \sqrt{D(X)}$

При увеличении σ увеличивается разброс возле среднего значения, кривая становится более пологой.

Стандартная нормальная величина $U \sim N(0;1),\ a = 0,\ \sigma = 1$

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

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = 1/2 + \Phi_0(x)$$

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

$\Phi_0(x)$ - *функция Лапласа*

Свойства функции Лапласа:

1. $\Phi_0(0) = 0$
2. $\Phi_0(-x) = -\Phi_0(x)$
3. $\Phi_0(+\infty) = \frac{1}{2}$
4. $\Phi_0(-\infty) = -\frac{1}{2}$

Пусть случайная величина X имеет распределение $N(a, \sigma^2)$. Тогда

$$P\{\alpha \le X \le \beta\} = \int_\alpha^\beta f(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_\alpha^\beta e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}} e^{-\frac{t^2}{2}} dt =$$

$$= \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right) = \Phi_0\left(\frac{\beta-a}{\sigma}\right) - \Phi_0\left(\frac{\alpha-a}{\sigma}\right). \quad (8)$$

$$P\{|X-a| \le l\} = P\{a-l \le X \le a+l\} = \Phi_0\left(\frac{l}{\sigma}\right) - \Phi_0\left(-\frac{l}{\sigma}\right) = 2\Phi_0\left(\frac{l}{\sigma}\right). \quad (9)$$

Таблица значений функции Лапласа прилагается!

Правило трёх сигм (3σ). Пусть случайная величина X имеет распределение: $X \sim N(a; \sigma^2)$

$$P\{|X-a| \le 3\sigma\} = 2\Phi_0\left(\frac{3\sigma}{\sigma}\right) = 2\Phi_0(3) \approx 0.997$$

Асимметрия: $As = \frac{\mu_3}{\sigma^3}$, $As > 0$ - правый хвост тяжелее левого, иначе наоборот.

Эксцесс: $Ex = \frac{\mu_4}{\sigma^4} - 3$, $Ex > 0$ - более крутое распределение, иначе наоборот. При нормальном распределении $Ex = 0$.

2.3 Задачи

Задача 1. Цены на акции распределены по нормальному закону распределения со средним значением 45 \$, средним квадратическим отклонением 7 \$. Найти вероятность того, что цена наугад выбранной акции превышает 50 \$.

Решение. Пусть X — цена акции. Тогда $X \sim N(45; 7^2)$. Имеем:

$$P\{X > 50\} = P\{50 < X < \infty\} = \Phi_0(\infty) - \Phi_0(\frac{50-45}{7}) = 1/2 - \Phi_0(0,71) = 1/2 - 0,2611 = 0,2389.$$

Здесь мы воспользовались формулой (2.2), таблицей значений функции Лапласа и тем свойством, что на бесконечности ф. Лапласа равна 1/2.

Задача 2. Рост взрослого мужчины удовлетворительно описывается нормальным распределением. Средний рост составляет 174 см, среднее квадратическое отклонение 7 см. Найти вероятность того, что рост наугад выбранного мужчины отличается от среднего не более чем на 7 см.

Замечание. Встречались ли Вам мужчины с отрицательным ростом?... На самом деле, конечно, рост лучше описывается лог-нормальным распределением, но об этом позже.

Решение. Пусть X — рост. Тогда $X \sim N(174; 7^2)$. На самом деле, условие содержит избыточную информацию: средний рост нам знать не нужно.

Имеем

$$P\{|X - a| \le 7\} = P\{a - 7 \le X \le a + 7\} = 2\Phi_0\left(\frac{7}{\sigma}\right) = 2\Phi_0(1) = 0,6826.$$

На самом деле, здесь мы использовали формулу (9) (интервал симметричен вокруг мат.ожидания, т.е. a), но поскольку по условию $l = \sigma$, мы имеем закон $k\sigma$ для $k = 1$.

Задача 3. Продолжительность жизни в некотором регионе имеет среднее квадратическое отклонение 14 лет. Чему равна средняя продолжительность жизни, если 30% жителей этого региона имеют возраст более 75 лет? Считаем, что продолжительность жизни распределена по нормальному закону.

Замечание. Конечно же, продолжительность жизни не распределена по нормальному закону, но решим задачу с таким условием.

Решение. Пусть X — продолжительность жизни. Тогда $X \sim N(a; 14^2)$. Нужно найти a. Используем условие:

$$0.3 = P\{X > 75\} = 1/2 - \Phi_0(\frac{75 - a}{14}). \Rightarrow \Phi_0(\frac{75 - a}{14}) = 0.2.$$

По таблице значений функции Лапласа находим значение, соответствующее $0,2$.

Получаем

$$\frac{75 - a}{14} = 0,524 \Rightarrow a \approx 67,66.$$

Замечание. В задачах такого типа иногда нужно “доставать” из аргумента функции Лапласа с.к.о.σ.

Задача 4. Вес пачки печенья должен равняться 200 гр. Реально вес распределен по нормальному закону со средним $a = 200$ (г) и средним квадратическим $\sigma = 4$ (г). Определите процент пачек печенья, вес которых a) меньше 192 гр. b) находится в интервале (197; 203) (г). Поясните на графике плотности.

Решение. $X \sim N(200; 4^2)$. Тогда

a)

$$P\{X < 192\} = \Phi_0(\frac{192-200}{4}) - (-1/2) = 1/2 - \Phi_0(2) \approx 0,0228.$$

Учитывая идеализацию модели (вес навряд ли может принимать отрицательные значения), получаем, что процент пачек печенья весом меньше 192 г **примерно** равен $2,28\%$.

b)

$$P\{197 < X < 203\} = P\{|X-200| < 3\} = 2\Phi_0(\frac{3}{4}) = 2\Phi_0(0.75) \approx 2 * 0,2734 = 0,5468.$$

Задача 5. В нормально распределенной совокупности 15 % значений меньше 12; 40 % значений больше 16,2. Найти a) M(X) b) $\sigma(X)$.

Решение. $X \sim N(a; \sigma^2)$. Известно

$$P\{X < 12\} = \Phi_0(\frac{12-a}{\sigma}) - (-1/2) = 1/2 + \Phi_0(\frac{12-a}{\sigma}) = 0,15; \Rightarrow \Phi_0(\frac{12-a}{)}\sigma = -0,35 \Rightarrow \Phi_0(\frac{a-12}{\sigma}) = 0$$

$$P\{X > 16,2\} = 1/2 - \Phi_0(\frac{16,2-a}{\sigma}) = 0,4; \Rightarrow \Phi_0(\frac{16,2-a}{\sigma}) = 0,1.$$

Тогда

$$\frac{a-12}{\sigma} \approx 1,04; \qquad \frac{16,2-a}{\sigma} \approx 0,26.$$

$$4,2 = 1,3\sigma \qquad \Rightarrow \sigma \approx 3,23; \qquad a \approx 15,36.$$

Задача 6. Пусть цены на акции распределены по нормальному закону со средним значением 50 \$. Какова вероятность продать акцию не менее чем за 60 \$, если известно, что вероятность продать акцию в ценовом диапазоне от 45 \$ до 55 \$ равна 0.7?

Решение. $X \sim N(50; \sigma^2)$. Известно

$$P\{45 \le X \le 55\} = P\{|X-50| \le 5\} = 2\Phi_0(\frac{5}{\sigma}) = 0,7.$$

Тогда

$$\Phi_0(\frac{5}{\sigma}) = 0,35, \Rightarrow \frac{5}{\sigma} \approx 1,04, \Rightarrow \sigma \approx 4,81.$$

Следовательно,

$$P\{X \geq 60\} = 1/2 - \Phi_0(\frac{60-50}{4,81}) \approx 1/2 - \Phi_0(2,08) \approx 0,5 - 0,4812 = 0,0188.$$

3 Виды сходимости случайных величин

Пусть случайная величина ξ и последовательность случайных величин $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ заданы на одном вероятностном прстранстве $(\Omega, \mathfrak{F}, P)$.

1. *Сходимость по вероятности.*

   Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ сходится по вероятности к случайной величине ξ,

   если $\quad \forall \varepsilon > 0 \quad P\{|\xi_n - \xi| > \varepsilon\} \underset{n\to\infty}{\to} 0.$

   Записывается это следующим образом:

   $$\xi_n \overset{p}{\to} \xi.$$

2. *Сходимость почти всюду.*

   Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ сходится почти всюду к случайной величине ξ,

   если $P\{\lim\limits_{n\to\infty} \xi_n = \xi\} = 1.$

   Записывается это следующим образом:

   $$\xi_n \overset{\text{п.в.}}{\to} \xi.$$

   Замечание 1. Сходимость почти всюду в литературе также называется сходимостью почти наверное.

   Замечание 2. Сходимость почти всюду означает сходимость всюду, за исключением множеств меры нуль.

   Утверждение 1.

   Если последовательность сходится почти всюду, то последовательность сходится и по вероятности.

Доказательство см. Буре, Парилина, Глава 9.

Н.И. Чернова, Глава 11.

3. *Сходимость по распределению (слабая сходимость).*

   Последовательность случайных величин $\{\xi_n\}_{n=1}^{\infty}$ сходится по распределению к случайной величине ξ,

   если $\forall x \in C(F)$ имеет место сходимость (в основном) соответствующих функций распределения: $F_n(x) \underset{x\to\infty}{\to} F(x)$, где $C(F)$ — область непрерывности функции распределения $F(x)$.

   Записываем это следующим образом:

$$\xi_n \xrightarrow{d} \xi.$$

   Замечание 3. Сходимость по распределению может быть определена не через сходимость функций распределения в основном, а через сходимость математических ожиданий непрерывных ограниченных функций от соответствующих случайных величин. Но вместе с тем, оба подхода являются эквивалентными.

   Замечание 4. Основным отличием слабой сходимости от остальных видов сходимости является то, что от случайных величин не требуется, чтобы они были определены на одном вероятностном пространстве, так как условия сходимости формулируются с использованием только их функций распределения.

   Утверждение. Из сходимости по вероятности следует сходимость по распределению.

Доказательство см. Буре, Парилина, Глава 9.

Н.И. Чернова, Глава 11.

Замечание 5. Существуют и другие виды сходимости случайных величин, такие как сходимость в среднем, в среднем порядка r и т.д. Мы будем использовать сходимость по вероятности и сходимость по распределению для более компактной записи некоторых важных теорем.

Имеем:

$$\xi_n \xrightarrow{\text{п.в.}} \xi \quad \Rightarrow \quad \xi_n \xrightarrow{p} \xi \quad \Rightarrow \quad \xi_n \xrightarrow{d} \xi.$$

4 Неравенство П. Л. Чебышёва

Заметим, что фамилия Чебышёв пишется через букву Ё с соответствующим ударением. В дальнейшем изложении буква "ё"заменена на "е".

4.1 Неравенство П.Л. Чебышёва

Лемма (неравенство Маркова). Пусть случайная величина $X \geq 0$ (принимает неотрицательные значения), причем $M(X) < \infty$. Тогда

$$\forall \varepsilon > 0 \qquad P\{X \geq \varepsilon\} \leq \frac{M(X)}{\varepsilon}. \tag{10}$$

Доказательство. Общая схема $X \geq 0 \Rightarrow M(X) \geq 0$; $X \geq \varepsilon \Rightarrow M(X) \geq \varepsilon \Rightarrow M(X) \geq \varepsilon P\{X \geq \varepsilon\}$. Разделим на $\epsilon > 0$ и получим требуемое неравенство (10).
Конкретизируем доказательство для случая, когда у с.в. X существует плотность $f(x)$. Напомним, что с.в. принимает значения в промежутке $[0, \infty)$. Тогда имеем

$$M(X) = \int_0^\infty x f(x)dx \geq \int_\epsilon^\infty x f(x)dx \geq \int_\epsilon^\infty \epsilon f(x)dx = \epsilon P\{X \geq \epsilon\}.$$

Отсюда следует (10). □

Замечание 1. Имеем очень простой инструмент для оценки (сверху) вероятности некоторых событий, связанных с превышением значения случайной величины некоторого порогового значения.

Например, для неотрицательной с.в. X с конечным мат.ожиданием $M(X)$ выполняется неравенство

$$P\{X \geq 2M(X)\} \leq \frac{1}{2}.$$

Также при помощи (10) может быть также получена оценка снизу (см. Задачу 1 в разделе “Задачи” данной главы).

Пусть $X \geq 0$, причем $M(X) < \infty$. Тогда

$$\forall \varepsilon > 0 \qquad P\{0 \leq X < \varepsilon\} \geq 1 - \frac{M(X)}{\varepsilon}. \tag{11}$$

Утверждение (неравенство Чебышева. Пусть с.в. X такая, что $M(X) < \infty$, $M(X^2) < \infty$. Тогда $\forall \varepsilon > 0$ выполнено следующее неравенство:

$$P\{|X - M(X)| \geq \varepsilon\} \leq \frac{D(X)}{\varepsilon^2}. \tag{12}$$

Доказательство. Отметим, что $D(X) = M(X^2) - (M(X))^2$. Следовательно, из конечности первого и второго начальных моментов следует существование (конечной) дисперсии $D(X)$.

Используем предыдущую лемму (неравенство Маркова (10)). Заметим, что вероятности следующих событий равны:

$$P\{|X - M(X)| \geq \varepsilon\} = P\{(X - M(X))^2 \geq \varepsilon^2\},\ \epsilon > 0.$$

Рассмотрим новую с.в. $Y = (X - M(X))^2$. Она принимает только неотрицательные значения. $M(Y) = M(X - M(X))^2 = M(X^2) - (M(X))^2 < \infty$. Из леммы Маркова из неравенства (10) имеем:

$$P\{|X - M(X)| \geq \varepsilon\} = P\{Y \geq \varepsilon^2\} \leq \frac{M(Y)}{\varepsilon^2} = \frac{D(X)}{\varepsilon^2}.$$

□

Очевидно, что для противоположного события неравенство Чебышева (12) может быть записано в следующем виде:

$$P\{|X - M(X)| < \varepsilon\} \geq 1 - \frac{D(X)}{\varepsilon^2}. \tag{13}$$

Замечание 2. Неравенство Чебышева используется как для дискретных, так и для непрерывных случайных величин при условии существования у них первого и второго моментов.

Замечание 3. Неравенство Чебышева даёт очень грубый результат, но является простым инструментом для оценок вероятностей.

Например, рассмотрим с.в. $X \sim N(a, \sigma^2)$ (распределенную по нормальному закону). Мы знаем, что для нормального распределения выполняется закон 3σ (вспомним, что $M(X) = a,\ D(X) = \sigma^2$):

$$P\{|X - M(X)| < 3\sigma\} = 2\Phi_0(\frac{3\sigma}{\sigma}) = 2\Phi_0(3) \approx 0.997.$$

Тем не менее, из неравенства Чебышева (13) имеем очень грубую оценку снизу:

$$P\{|X - M(X)| < 3\sigma\} \geq 1 - \frac{D(X)}{(3\sigma)^2} = 1 - \frac{\sigma^2)}{(3\sigma)^2} = \frac{8}{9}.$$

Задание 1. Пусть $X \sim N(a, \sigma^2)$. Выведите закон $k\sigma$. Какую оценку дает неравенство Чебышева?

Задание 2. А каким будет значение $P\{|X - M(X)| < 3\sigma\}$, если а) $X \sim R(a; b)$? b) $X \sim Exp(\lambda)$?

Замечание 4. Неравенство Чебышева имеет смысл использовать только если $\epsilon^2 > D(X)$. В противном случае верхние и нижние оценки вероятностей в неравенствах (12), (13) становятся равными 0 и 1.

4.2 Некоторые другие неравенства

Лемма (неравенство) Йенсена. Пусть задана случайная величина X,$MX < \infty$ и числовая выпуклая функция $g(x)$, тогда

$$M(g(X)) \geq g(M(X)).$$

Неравенство Коши-Буняковского-Шварца. Пусть заданы случайные величины X и Y, $MX^2 < \infty$, $MY^2 < \infty$ тогда $\exists M(X \cdot Y)$:

$$|M(X \cdot Y)| \leq \sqrt{MX^2} \cdot \sqrt{MY^2}.$$

4.3 Задачи

Задача 1. На некоторой планете значение скорости ветра в среднем равно 40 м/ с. Оценить вероятность того, что при измерении скорость ветра превысит а) 80 м/ с; b) 60 м/ с с); будет менее 60 м/ с.

Решение.

Имеем $M(X) = 40$. Из неравенства Маркова (10), (11) имеем:

a)

$$P\{X \geq 2 * 40\} = P\{X \geq 2M(X)\} \leq \frac{1}{2}.$$

b)

$$P\{X \geq 60\} \leq \frac{M(X)}{60} = \frac{40}{60} = \frac{2}{3}.$$

c)

Очевидно, что $P\{0 \leq X < 60\} + P\{X \geq 60\} = 1$. Тогда

$$1 - P\{X < 60\} = P\{X \geq 60\} \leq \frac{M(X)}{60} = \frac{40}{60} = \frac{2}{3}.$$

$$P\{0 \leq X < 60\} \geq 1 - \frac{M(X)}{60} = \frac{1}{3}.$$

Задача 2. На некоторой планете среднее значение температуры равно 100 C^0, а среднее квадратическое отклонение равно 20 C^0. Оценить вероятность того, что температура на планете отклонится от средней более чем на 40 C^0.

Решение.

Из (12) имеем

$$P\{|X - M(X)| > 40\} \leq P\{|X - M(X)| \geq 40\} \leq \frac{D(X)}{40^2} = \frac{20^2}{40^2} = \frac{1}{4}.$$

Задание 3. Вероятность вылупления цыпленка из яйца равна 0.75. В инкубатор заложено 1000 яиц. Оценить вероятность того, что вылупятся от 720 до 780 цыплят.

Решение. Вероятность успеха $p = 0.75$, число испытаний $n = 1000$. Число вылупившихся цыплят — с.в., распределенная по биномиальному закону, т.к. $S_n \sim B(1000; 0, 75)$.

Отметим, что среднее число успехов $MS_n = np = 750$, а средне квадратическое отклонение $\sigma(S_n) = \sqrt{npq} = \sqrt{750 * 0.25}$.

В данной задаче интервал [720; 780] симметричен вокруг математического ожидания 750. Получаем

$$P(|S_n - MS_n| < 30) > 1 - \frac{DS_n}{30^2} = 1 - \frac{npq}{30^2} = 0, 79.$$

Замечание. В Главе 6 данная задача будет решена при помощи интегральной теоремы Муавра-Лапласа.

Задание 4. Вероятность рождения девочки приблизительно равна 0,485. Оценить вероятность того, что число девочек среди 3000 новорожденных будет отличаться от математического ожидания этого числа по абсолютной величине менее чем на 55.

Решение. Число девочек среди n новорожденных — биномиальная случайная величина S_n ($S_n \sim B(n,p)$. В данной задаче $n = 3000,\ p = 0,485$. Тогда $MS_n = np = 3000 * 0,485 = 1455,\ DS_n = npq = 1455 * 0,515 = 749,325$.

Получаем

$$P(|S_n - MS_n| < 55) > 1 - \frac{DS_n}{55^2} = 1 - \frac{npq}{55^2} \approx 0,75.$$

Замечание. В Главе 6 данная задача будет решена при помощи интегральной теоремы Муавра-Лапласа.

5 Закон больших чисел

Закон больших чисел (ЗБЧ) — набор теорем, в которых для тех или иных условий устанавливается факт приближения средних характеристик большого числа опытов к некоторым неслучайным постоянным величинам.

Например, поведение частиц газа в замкнутом сосуде невозможно предсказать для каждой отдельной частицы. Но в среднем мы можем указать её вектор скорости и т.п. Вместо частиц в сосуде могут быть рассмотрены также цены на акции на электронных торгах, зарплаты людей определенной категории и т.д.

5.1 Закон больших чисел в форме Чебышёва

Теорема. Закон больших чисел в форме Чебышева. Для любой последовательности попарно-независимых случайных величин $X_1, X_2, \ldots, X_n, \ldots$ с конечными первыми моментами $M(X_i),\ \forall i$ и равномерно ограниченными дисперсиями $D(X_i) \le C,\ \forall i$ имеем:

$$\forall \varepsilon > 0, \qquad P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \frac{M(X_1) + \ldots + M(X_n)}{n}\right| > \varepsilon\right\} \underset{n\to\infty}{\to} 0. \tag{14}$$

Замечание 5. Заметим, что результатом теоремы является установление факта сходимости по вероятности, а именно:

$$\frac{X_1 + \ldots + X_n}{n} \quad \xrightarrow{p} \quad \frac{M(X_1) + \ldots + M(X_n)}{n} \text{ при } n \to \infty. \tag{15}$$

Последовательность случайных величин с конечными мат.ожиданиями, удовлетворяющих данному свойству, также называют "удовлетворяющими закону больших чисел".

Замечание 6. Смысл закона больших чисел: при определенных условиях и при большом числе испытаний среднее арифметическое случайных величин, являющееся случайной величиной, может быть заменено на НЕслучайную характеристику. Данная теорема также по сути связывает такие науки как теория вероятностей и математическая статистика.

Доказательство ЗБЧ в форме Чебышева. Введем следующую случайную величину:

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Очевидно, что

$$M(\bar{X}) = \frac{M(X_1) + \ldots + M(X_n)}{n},$$

$$D(\bar{X}) = D\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\sum_{i=1}^{n} D(X_i)}{n^2} \leq \frac{nC}{n^2} = \frac{C}{n}.$$

Применим неравенство Чебышева (12) для случайной величины $\bar{X}$:

$$P\{|\bar{X} - M(\bar{X})| > \varepsilon\} \leq p\{|\bar{X} - M(\bar{X})| \geq \varepsilon\} \leq \frac{D(\bar{X})}{\varepsilon^2} \leq \frac{C}{n \cdot \varepsilon^2}.$$

Возьмем предел при $n \to \infty$. Очевидно, что искомая вероятность "зажата"нулем слева и справа. Тогда, по лемме о двух милиционерах (см. курс мат.анализа), получаем выполнение (14).

□

Задание 3. Где в доказательстве теоремы использовалась независимость случайных величин?

Следующие теоремы являются **следствиями** ЗБЧ в форме Чебышева.

5.2 Закон больших чисел для независимых одинаково распределенных с.в.

Теорема. Закон больших чисел для независимых одинаково распределенных с.в.

Пусть $X_1, X_2, \ldots, X_n, \ldots$ — последовательность независимых **одинаково** распределённых случайных величин, причем $MX_i = m_x < \infty,\ DX_i \leq C\ \ \forall i$. Тогда

$$P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - m_x\right| > \varepsilon\right\} \to_{n\to\infty} 0, \qquad \forall \varepsilon > 0. \tag{16}$$

Замечание 7. Одинаково распределенные случайные величины — с.в., имеющие одинаковые распределения вероятностей. Можно рассматривать их как копии одной и той же случайной величины, причем индекс у с.в. указывает на номер испытания данной случайной величины. Очевидно, что при условии существования, моменты всех порядков у таких случайных величин совпадают (включая мат.ожидание, дисперсию и т.п.).

Доказательство. См. ЗБЧ в форме Чебышева. Очевидно, что для одинаково распределенных независимых с.в. имеем

$$\frac{M(X_1) + \ldots + M(X_n)}{n} = \frac{nm_x}{n} = m_x.$$

□

5.3 Закон больших чисел в форме Бернулли

Рассмотрим классическую формулировку схемы Бернулли: последовательность n независимых испытаний, в каждом из которых вероятность появления события A равна p.

Появление события A называем "успехом".

Теорема. Закон больших чисел в форме Бернулли. Пусть S_n — число успехов в схеме Бернулли, p — вероятность успеха, n — число испытаний. Тогда

$$\forall \varepsilon > 0 \qquad P\left\{\left|\frac{S_n}{n} - p\right| > \varepsilon\right\} \underset{n\to\infty}{\to} 0. \tag{17}$$

Доказательство. Как мы уже делали при выводе мат.ожидания и дисперсии для биномиальной с.в. S_n, представим ее в виде суммы индикаторных случайных величин:

$$S_n = I_1 + \ldots + I_n, \qquad I_i = \begin{cases} 0, & \bar{A} \quad q \\ 1, & A \quad p \end{cases}$$

Здесь q — вероятность неудачи, т.е. $q = 1 - p$. Напомним, что распределение с.в. I_j называется распределением Бернулли.

Очевидно, что последовательность с.в. I_1, $I_2, \ldots$ при $n \to \infty$ является последовательностью независимых одинаково распределенных с.в., причем $M(I_j) = 0 * q + 1 * p = p$, $D(I_j) = qp$. УМЕТЬ ЭТО ПОЛУЧАТЬ В УМЕ!

Тогда из ЗБЧ для независимых одинаково распределенных с.в. получаем:

$$\frac{I_1 + \ldots + I_n}{n} = \frac{S_n}{n} \to^p p.$$

□

Задание 4. Каким числом всегда ограничена дисперсия $D(I_j)$, $\forall j$?

Замечание 8. Из ЗБЧ в форме Бернулли следует, что частота успехов в схеме Бернулли демонстрирует свойство "устойчивости". Следовательно, при большом числе испытаний неизвестная вероятность успеха p может быть заменена на частоту успеха S_n/n. Если мы не знаем вероятность выпадения решки у монеты, то можно подбросить её, например, 1000 раз, посмотреть, сколько раз выпадет решка, и взять вместо неизвестной вероятности отношение числа появлений решки к 1000. ЗБЧ в форме Бернулли позволяет это сделать.

Замечание 9. ЗБЧ в форме Бернулли также называют одной из предельных теорем в схеме Бернулли. Данный результат был получен Я. Бернулли до появления на свет П.Л. Чебышёва при помощи формулы Дж. Стирлинга. Мы же рассматриваем ЗБЧ в форме Бернулли как следствие ЗБЧ в форме Чебышева.

Замечание 10. Законами больших чисел принято называть утверждения о том, при каких условиях последовательность случайных величин удовлетворяет закону больших чисел, т.е. обладает свойством (15). Оказывается, могут быть смягчены требования о независимости, равномерной ограниченности дисперсии и т.д. Известные законы больших чисел носят имена Маркова, Хинчина и т.д.

6 Центральная предельная теорема

Центральные предельные теоремы (ЦПТ) — класс теорем в теории вероятностей, утверждающих, что сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному.

Для доказательства одной ЦПТ нам понадобятся дополнительные сведения.

6.1 Комплексные случайные величины

Наряду с веществennозначными случайными величинами мы можем рассмотреть и комплекснозначные с.в., понимая под этим следующее.

Пусть $X : \Omega \to R^1$, $Y : \Omega \to R^1 \Rightarrow$.

Комплексная случайная величина $Z = X + iY$, где i – мнимая единица.
Модуль: $|Z| = \sqrt{X^2 + Y^2}$.

Другие формы записи: $Z = R(\cos\varphi + i\sin\varphi)$, $Z = Re^{i\varphi}$.

Фактически, с.в. Z — случайная точка на плоскости xOy.

Необходимо ввести определения мат.ожидания и дисперсии так, чтобы при $Y = 0$ сохранялись свойства введенных ранее числовых характеристик вещественнозначных случайных величин.

Определим: $M(Z) = M(X) + iM(Y)$. Это комплексное число!
Все свойства мат. ожидания $M(X)$ сохраняются для Z, в том числе
If Z_1, Z_2 — независимые с.в., то $M(Z_1Z_2) = M(Z_1)M(Z_2)$.

$D(Z) = M(|\overset{\circ}{Z}|^2) = D(X) + D(Y)$, где $\overset{\circ}{Z} = Z - M(Z)$ – центрированная случайная величина. Дисперсия для комплексной с.в. — вещественное число!

6.2 Характеристические функции

Характеристические
функции (х.ф.) введены А.М. Ляпуновым и использованы им для доказательства ЦПТ в форме Ляпунова. Кроме того, аппарат х.ф. представляет прикладную значимость сам по себе, поскольку позволяет альтернативным образом задавать закон распределения с.в., вычислять числовые характеристики (вещественнозначных) случайных величин и т.д.

6.2.1 Определение и свойства

Пусть $X : \Omega \mapsto R^1$ — (вещественнозначная) случайная величина.
Рассмотрим $Y = e^{itX}$, где i — мнимая единица, t – параметр (вещественная переменная).

Имеем, что Y — случайная точка на единичной окружности.

Характеристической функцией для случайной величины X называется функция

$$\varphi_X(t) = M(Y) = M(e^{itX})$$

Пусть X – дискретная случайная величина. Тогда:

$$\varphi_X(t) = \sum_k e^{itx_k} p_k.$$

Пусть X — абсолютно непрерывная случайная величина ($\exists f(x)$). Тогда:

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} f(x) dx.$$

Иногда индекс X опускается и пишется просто $\varphi(t)$.

Свойства характеристической функции:

1. $\varphi(t)$ равномерно непрерывна на всей прямой и

   $|\varphi(t)| \leq 1$,

   $\varphi(0) = 1$

   Доказательство (для абсолютно непрерывных с.в., для дискретных с.в. доказывается аналогично):

   $$|\varphi_X(t)| = \left| \int_{-\infty}^{+\infty} e^{itx} f(x) dx \right| \leq \int_{-\infty}^{+\infty} \left| e^{itx} f(x) dx \right| \leq \int_{-\infty}^{+\infty} |e^{itx}| f(x) dx = \int_{-\infty}^{+\infty} f(x) dx = 1$$

2. Пусть $Z = aX + b$. Тогда

   $$\varphi_Z(t) = e^{itb} \varphi_X(at) \tag{18}$$

   Доказательство:

   $$\varphi_Z(t) = \varphi_{aX+b}(t) = M(e^{it(aX+b)}) = M(e^{itaX} \cdot e^{itb}) = e^{itb} \cdot M(e^{iatX}).$$

3. Пусть случайные величины $X_1, X_2, \ldots, X_n$ независимы. Рассмотрим сумму этих случайных величин $S_n = X_1 + \ldots + X_n$. Тогда

   $$\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{X_i}(t). \tag{19}$$

   Доказательство:

   $$M(e^{it(X_1+\ldots+X_n)}) = M(e^{itX_1} \cdot \ldots \cdot e^{itX_n}) = M(e^{itX_1}) \cdot \ldots \cdot M(e^{itX_n})$$

   Это свойство красной нитью проходит через доказательство ЦПТ!

4. Пусть $\exists M(|X|^k) \Rightarrow \varphi^{(k)}(0) = i^k \cdot M(X^k)$.

Доказательство: получаем формальным дифференцированием $\varphi(t)$.

Следствие.

$$M(X^k) = i^{-k}\varphi^{(k)}(0). \tag{20}$$

Формула (20) является альтернативным способом вычисления матожидания, дисперсии и т.д. Например, для нормального распределения (см. далее) это очень удобный способ получения числовых характеристик, в отличие от применения формул (2), (3).

Из (20) имеем:

$$M(X) = \frac{\varphi'(0)}{i}. \tag{21}$$

$$D(X) = M(X^2) - (M(X))^2 = \frac{\varphi''(0)}{i^2} - \frac{(\varphi'(0)^2}{i^2} = (\varphi'(0))^2 - \varphi''(0). \tag{22}$$

Замечание*. Математическое ожидание и дисперсия легко выражаются при помощи производной от логарифма характеристической функции.

Пусть $\psi(t) = ln\varphi(t)$. Тогда $\psi'(t) = \frac{\varphi'(t)}{\varphi(t)}$, $\psi''(t) = \frac{\varphi''(t)\varphi(t) - (\varphi'(t))^2}{\varphi^2(t)}$.

Вспомним, что $\varphi(0) = 1$. Используя формулу (20), получаем:

$$\psi'(0) = \varphi'(0) = iM(X).$$

$$\psi''(0) = -D(X).$$

Тогда:

$$M(X) = \frac{1}{i}\psi'(0). \tag{23}$$

$$D(X) = -\psi''(0). \tag{24}$$

6.2.2 Примеры вычисления характеристических функций

1. Пусть $X = C$ (неслучайная величина). Тогда

$$\varphi_C(t) = \sum_k e^{itx_k} \cdot p_k = e^{itC} \cdot 1 = e^{itC}.$$

2. Пусть X_i - индикаторная случайная величина в схеме Бернулли. X_i имеет только два значения 1 и 0 с соответствующими вероятностями $p,\ q = 1 - p$. Говорят также, что с.в. X_i имеет распределение Бернулли. Тогда

$$\varphi_{X_i}(t) = \sum_k e^{itx_k} \cdot p_k = e^{it\cdot 0} \cdot q + e^{it\cdot 1} \cdot p = q + pe^{it}.$$

3. Пусть $S_n \sim B(n,p)$ (S_n — число успехов в схеме Бернулли, состоящей из n испытаний, иначе говоря, биномиальная случайная величина). Тогда по свойству (19)

$$S_n = X_1 + \ldots + X_n \Rightarrow \varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{X_i}(t) = (q + p \cdot e^{it})^n.$$

4. $X \sim R(a;b) \quad \varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} \cdot f(x)dx = \int_a^b e^{itx} \cdot \frac{1}{b-a}dx = \frac{e^{itb}-e^{ita}}{it(b-a)}$

5. $X \sim N(a,\sigma^2) \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{(x-a)^2}{2\sigma^2}}$

$$\varphi_X(t) = \int_{-\infty}^{+\infty} e^{itx} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma} e^{ita} \int_{-\infty}^{+\infty} e^{ity - \frac{y^2}{2\sigma^2}} dy$$

$ity - \frac{y^2}{2\sigma^2} = (\frac{y - it\sigma}{\sqrt{2}\sigma})^2 - \frac{t^2\sigma^2}{2}$

$$\varphi_X(t) = e^{ita - \frac{t^2\sigma^2}{2}} \cdot \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{y-itb}{\sqrt{2}\sigma})^2} dy = e^{ita - \frac{t^2\sigma^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{z^2}{2}} dz = e^{ita - \frac{t^2\sigma^2}{2}}.$$

Имеем

$$\varphi_X(t) = e^{ita - \frac{t^2\sigma^2}{2}}. \tag{25}$$

Заметим, что для стандартной нормальной с.в. $a = 0,\ \sigma = 1$. Тогда из (25) получаем

$$\varphi_X(t) = e^{\frac{-t^2}{2}}. \tag{26}$$

Задание. Получите х.ф. для случайной величины $X \sim P(\lambda)$ (распределенной по закону Пуассона).

6.2.3 Примеры вычисления числовых характеристик с помощью характеристических функций

Воспользуемся выражениями для х.ф., полученными в предыдущем разделе, а также формулами (21), (22).

1. $X = C$. Тогда

$$(e^{itC})' = iCe^{itC});\ \varphi'_C(0) = iCe^0 = iC;\ \Rightarrow M(C) = \frac{iC}{i} = C.$$

$$(e^{itC})'' = i^2C^2e^{itC});\ \varphi''_C(0) = i^2C^2e^0 = -C^2;\ \Rightarrow D(C) = -(-C^2) + (iC)^2 = 0.$$

Тут мы сложным образом доказали очень простой факт.

2. X_i — индикаторная с.в.

$\varphi(t) = q + pe^{it}$

$\varphi'(t) = ipe^{it}$

$M(X_i) = \frac{\varphi'(0)}{i} = \frac{ipe^0}{i} = p$

$\varphi''(t) = (ipe^{it})' = i^2pe^{it} = -pe^{it}$

$D(X_i) = p - p^2 = p(1-p) = pq$

3. $S_n \sim B(n,p)$ (биномиальное распределение)

$\varphi_{S_n}(t) = (q + pe^{it})^n$

$\varphi'_{S_n}(t) = n(q + pe^{it})^{n-1} \cdot ipe^{it}, \varphi'_{S_n}(0) = in(q+p)^{n-1}p = inp$

$MS_n = \frac{\varphi'_{S_n}(0)}{i} = np$

$DS_n = npq$

4. $X \sim N(a;\sigma^2)$ (нормальное распределение)

$\varphi(t) = e^{ita - \frac{t^2\sigma^2}{2}}$

$\varphi'(t) = e^{ita - \frac{t^a\sigma^2}{2}}(ia - t\sigma^2),\ \varphi'(0) = ia$

$MX = \frac{\varphi'(0)}{i} = \frac{ia}{i} = a$

$D(X) = \sigma^2$

А здесь мы очень просто доказали непростую формулу!

Задание. 1) Получите выражения для дисперсий для биномиального и нормального распределений (п.3, 4 выше) при помощи аппарата х.ф.

2) При помощи х.ф. докажите, что для распределения Пуассона $M(X) = \lambda;\ D(X) = \lambda$.

6.2.4 Теоремы о единственности и непрерывности

Теорема 1 (о единственности). Пусть $F(x), G(x)$ – функции распределения,

соответствующие характеристической функции $\varphi(t)$. Тогда

$$F(x) \equiv G(x).$$

(см. теорему Леви)

Теорема 2 (о непрерывности характеристической функции). Пусть дана последовательность $\{\varphi_n\}_{n=1}^{\infty}$ такая, что $\varphi_n(t) \underset{n\to\infty}{\to} \varphi(t)$. Тогда

$$F_n(x) \underset{n\to\infty}{\to} F(x)$$

$\forall x \in \mathbb{C}(F), \varphi(t)$ — характеристическая функция.

Следствие.

Сумма нормально распределенных с.в. Докажите, что если $X_1 \sim N(a_1; \sigma_1^2)$, $X_2 \sim N(a_2; \sigma_2^2)$, они независимы, то $X = X_1 + X_2 \sim N(a_1 + a_2; \sigma_1^2 + \sigma_2^2)$.

Доказательство. $\varphi_{X_1}(t) = e^{ita_1 - \frac{t^2\sigma_1^2}{2}}$, $\varphi_{X_2}(t) = e^{ita_2 - \frac{t^2\sigma_2^2}{2}}$.

Тогда из независимости получаем: $\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) = e^{it(a_1+a_2) - \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}}$.

Это соответствует нормально распределенной с.в. $N(a_1 + a_2; \sigma_1^2 + \sigma_2^2)$. Обратите внимание, что $\sigma(X_1 + X_2) = \sqrt{\sigma(X_1)^2 + \sigma(X_2)^2}$.

Это свойство легко обобщается на сумму n независимых с.в., распределенных по норм. закону. Если они все одинаково распределены, то их сумма X подчиняется нормальному закону и

$$M(X) = M(X_1 + \ldots + X_n) = na,$$

$$\sigma(X) = \sigma(X_1 + \ldots + X_n) = \sigma\sqrt{n}.$$

Данное свойство нам понадобится на статистике.

Задание. Сумма случайных величин, распределенных по закону Пуассона. Докажите, что если $X_1 \sim P(\lambda_1),\ X_2 \sim P(\lambda_2)$, они независимы, то $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$.

6.3 Центральная предельная теорема для независимых, одинаково распределённых случайных величи.

Центральная предельная теорема Пусть имеется последовательность $\{\xi_n\}_{n=1}^{\infty}$ независимых, одинаково распределённых случайных величин, $M\xi_i = a,\ D\xi_i = \sigma^2$.

Рассмотрим:

- частичные суммы $S_n = \sum_{i=1}^{n} \xi_i$
- нормированные частичные суммы

  $\bar{S_n} = \frac{S_n - MS_n}{\sqrt{DS_n}} = \frac{S_n - an}{\sigma\sqrt{n}}$, $M\bar{S_n} = 0, D\bar{S_n} = 1$.

  Тогда

$$\bar{S_n} \xrightarrow{d} U \sim N(0,1),$$

  т.е.

$$F_{\bar{S_n}}(x) = P\{\bar{S_n} < x\} \underset{n\to\infty}{\to} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt.$$

Пояснение.

1) ЦПТ утверждает, что если у нас есть последовательность независимых, одинаково распределенных с.в., у которых существует конечное матожидание и дисперсия, то их частичная (нормированная) сумма в пределе имеет **нормальное распределение**, независимо от того, какое распределение было изначально.

2) Получаем, что последовательность функций распределения частичных нормированных сходится в основном к функции распределения $\Phi(x)$, соответствующей стандартной нормальной случайной величине. Значения этой функции (как таковой или на основе функции Лапласа, $\Phi(x) = 1/2 + \Phi_0(x)$) можно брать из таблицы.

Замечания.

Замечание 1: Говорят, что $\bar{S_n}$ асимптотически нормальна.

Замечание 2: Очевидно, что если $\bar{S_n} \xrightarrow{d} U \sim N(0,1),$ то $S_n \xrightarrow{d} \xi \sim N(na, n\sigma^2)$

Идея доказательства.

1. Для независимых одинаково распределенных случайных величин: $\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{\xi_i}(t) = (\varphi_{\xi_i}(t))^n$
2. $\varphi_{\bar{S_n}}(t) \underset{n\to\infty}{\to} e^{-t^2/2}$, что соответствует характеристической функции стандартной нормальной случайной величины. По теореме о непрерывности имеем, что тогда и последовательность соответствующих функций распределения сходится в основном к ф.р. стандартной норм. с.в.

Доказательство. Рассмотрим $\xi_i' = \xi_i - a$:

$$S_n' = \sum_{i=1}^{n} \xi_i' = \sum_{i=1}^{n} \xi_i - na$$

Отцентрируем (отнимем матожидание) и отнормируем (поделим на среднее квадратическое отклонение). Тогда

$$\bar{S}'_n = \frac{S'_n - MS'_n}{\sqrt{D\bar{S}'_n}} = \frac{\sum_{i=1}^n \xi_i - na}{\sigma\sqrt{n}} = \bar{S}_n$$

Нормированные частичные суммы у этих последовательностей совпадают, поэтому далее, не умаляя общности будем рассматривать

$\xi_i : M\xi_i = 0,\ D\xi_i = \sigma^2$. (штрих опустим)

Тогда $\bar{S}_n = \frac{S_n}{\sigma\sqrt{n}}$.

Построим характеристическую функцию: $\varphi_{\bar{S}_n}(t) = \prod_{i=1}^n \varphi_{\xi_i}(t) = (\varphi_{\xi_i}(t))^n$ Вспомним свойство характеристической функции:

$$\varphi_{aX+b}(t) = e^{itb}\varphi_X(at) \Rightarrow \varphi_{\bar{S}_n}(t) = \varphi_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

Разложим $\varphi_{\xi_i}(t)$ в ряд Тейлора в окрестности нуля с остаточным членом в форме Пеано:

$$\varphi_{\xi_i}(t) = \varphi_{\xi_i}(0) + \varphi'_{\xi_i}(0)\cdot t + \frac{\varphi''_{\xi_i}(0)\cdot t^2}{2} + O(t^2)$$

$\varphi_{\xi_i}(0) = 1$

$\varphi^{(k)}(0) = i^k \cdot M(X^k)$

$\varphi'_{\xi_i}(0) = i \cdot M\xi_i = 0$

$\varphi''(0) = i^2 \cdot M\xi_i^2 = -1 \cdot \sigma^2 = -\sigma^2$

$\varphi_{\xi_i}(t) = 1 - \frac{\sigma^2}{2} + O(t^2)$

$\varphi_{\xi_i}(\frac{t}{\sigma\sqrt{n}}) = 1 - \frac{\sigma^2}{2}\cdot\frac{t^2}{\sigma^2\cdot n} + O(t^2) = 1 - \frac{t^2}{2n} + O(t^2)$

$\varphi_{\bar{S}_n}(t) = \left(\varphi(\frac{t}{\sigma\sqrt{n}})\right)^n = \left(1 - \frac{t^2}{2n} + O(t^2)\right)^n$

$\lim\limits_{n\to\infty} \varphi_{\bar{S}_n}(t) = \lim\limits_{n\to\infty}\left(1 - \frac{t^2}{2n}\right)^n = e^{-\frac{t^2}{2}}$

□

Задание. Докажите утверждение Замечания 2.

6.4 Интегральная и локальная теорема Муавра-Лапласа

Рассмотрим схему Бернулли:

$$P\{S_n = k\} = C_n^k \cdot p^k \cdot q^{n-k}, \quad k = 0, \ldots, n$$

При больших n биномиальное распределение с.в. S_n можно заменять на нормальное: $S_n \sim B(n,p) \to \ S_n \sim N(np; npq)$.

Основанием для этого является то, что число успехов в схеме Бернулли представимо в виде суммы $S_n = X_1 + \ldots + X_n, \quad MX_i = p, DX_i = pq, q = 1-p \Rightarrow$ см. Центральную предельную теорему.

$$\bar{S_n} = \frac{S_n - np}{\sqrt{npq}} \xrightarrow{d} U \sim N(0,1)$$

Теорема (интегральная теорема Муавра-Лапласа). Пусть $S_n \sim B(n,p)$. Тогда при $n \to \infty$

$$P\{k_1 \le S_n \le k_2\} \approx \Phi_0\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{k_1 - np}{\sqrt{npq}}\right). \tag{27}$$

Доказательство: Число успехов в схеме Бернулли представимо в виде суммы индикаторных с.в., $S_n = \sum_{i=1}^{n} X_i$. Все предпосылки ЦПТ для независимых, одинаково распределенных с.в. выполнены. Следовательно, S_n сходится по распределению к нормальной с.в. с математическим ожиданием np и дисперсией npq (либо, что то же, $\frac{S_n - np}{\sqrt{npq}}$ сходится к стандартной нормальной с.в. $N(0;1)$).

Тогда для S_n выполнены формулы для нахождения попадания с.в. в интервал (2.2). Имеем:

$$P\{k_1 \le S < k_2\} = P\{\frac{k_1 - MS_n}{\sqrt{DS_n}} \le \frac{S_n - MS_n}{\sqrt{DS_n}} < \frac{k_2 - MS_n}{\sqrt{DS_n}}\} = \Phi_0\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{k_1 - np}{\sqrt{npq}}\right)$$

Теорема (локальная теорема Муавра-Лапласа). При $n \to \infty$

$$P\{S_n = k\} \approx \frac{1}{\sqrt{2\pi npq}} e^{-x_k^2/2}, \quad x_k = \frac{k - np}{\sqrt{npq}}$$

Доказательство следует из центральной предельной теоремы.

6.5 Задачи

Задание 1. Вероятность вылупления цыпленка из яйца равна 0.75. В инкубатор заложено 1000 яиц. Найти вероятность того, что вылупятся от 720 до 780 цыплят.

Решение. Вероятность успеха $p = 0.75$, число испытаний $n = 1000$. Число вылупившихся цыплят — с.в., распределенная по биномиальному закону, т.к. $S_n \sim B(1000; 0,75)$.

Отметим, что среднее число успехов $MS_n = np = 750$, а средне квадратическое отклонение $\sigma(S_n) = \sqrt{npq} = \sqrt{750 * 0.25}$.

Согласно ЦПТ (интегральной теореме Муавра–Лапласа) имеем:

$$P\{S_n \in [720; 780]\} = \Phi_0(\frac{780-750}{13.7}) - \Phi_0(\frac{720-750}{13.7}) = 2\Phi_0(\frac{30}{13.7}) = 2\Phi_0(2.19) = 0.9708.$$

Замечание. В данной задаче интервал [720; 780] симметричен вокруг математического ожидания 750. Для нормального распределения можно было использовать формулу (9).

Кроме того, посмотрим, что дает неравенство Чебышёва (13) (не пользуясь асимптотической нормальностью).

Получаем

$$P(|S_n - MS_n| < 30) > 1 - \frac{DS_n}{30^2} = 1 - \frac{npq}{30^2} = 0,79.$$

Мы видим, что нижняя оценка по неравенству Чебышёва верная, но дает достаточно грубый результат.

Задание 2. Вероятность рождения девочки приблизительно равна 0,485. Найти вероятность того, что число девочек среди 3000 новорожденных будет отличаться от математического ожидания этого числа по абсолютной величине менее чем на 55.

Решение. Число девочек среди n новорожденных — биномиальная случайная величина S_n ($S_n \sim B(n,p)$. В данной задаче $n = 3000,\ p = 0,485$. Тогда $MS_n = np = 3000 * 0,485 = 1455,\ DS_n = npq = 1455 * 0,515 = 749,325$.

Согласно ЦПТ (интегральной теореме Муавра–Лапласа) имеем:

$$P\{|S_n - MS_n| < 55\} \approx 2\Phi_0(\frac{55}{27.37}) \approx 2\Phi_0(2) = 2 * 0.4772 = 0,9544.$$

Кроме того, посмотрим, что дает неравенство Чебышёва (13) (не пользуясь асимптотической нормальностью).

Получаем

$$P(|S_n - MS_n| < 55) > 1 - \frac{DS_n}{55^2} = 1 - \frac{npq}{55^2} \approx 0,75.$$

Мы видим, что нижняя оценка по неравенству Чебышёва верная, но дает достаточно грубый результат.

Задание 3. В 1980 г. на президентских выборах 34.9 млн жителей США проголосовали за кандидата от Демократической партии (Картера) и 43.2 млн жителей США проголосовали за кандидата от Республиканской партии (Рейгана). Не принимая во внимание другие партии, найдите вероятность того, что в случайной выборке из 1500 человек большинство набрал кандидат от Республиканской партии?

Решение. Доля избирателей, голосующих за республиканца, равна: $\frac{43.2}{34.9+43.2} = 0.553$. Пусть X — число избирателей в выборке из 1500 человек, голосующих за республиканца. Очевидно, это биномиальная с.в., $n = 1500,\ \ p = 0.553$. Имеем:

$$MX = np = 1500 * 0,553 = 829.5; \qquad DX = npq = 1500 * 0,553 * 0,447 = 370,7865;$$

$$\sigma(X) = \sqrt{DX} = 19.26.$$

Тогда вероятность того, что $X > 750$ согласно интегральной теореме Муавра-Лапласа равна:

$$P\{X > 750\} = \Phi_0(\infty) - \Phi_0(\frac{750 - 829.5}{19.26}) = 1/2 + \Phi_0(4.12) \approx 1/2 + 0,49996 \approx 0.9999.$$

Задание 4. Сколько опытов с бросанием монеты нужно произвести, чтобы с вероятностью 0.92 можно было ожидать отклонение частоты выпадения герба от теоретической вероятности 0.5 по абсолютной величине, меньше чем на 0.01?

Решение. Имеем схему Бернулии, $p = 0,5,\ \ n =?$. Из условий задачи известно:

$$P\{|\frac{S_n}{n} - 0.5| < 0,01\} = P\{|S_n - 0.5n| < 0.01n\} = 0.92.$$

Очевидно, что

$$MS_n = n0.5, \qquad DS_n = n * 0.5^2.$$

По интегральной теореме М.-Л. имеем:

$$P\{|S_n - 0.5n| < 0.01n\} = 2\Phi_0(\frac{0.01n}{0,5\sqrt{n}}) = 0.92.$$

По таблице значений функции Лапласа находим:

$$\Phi_0\left(\frac{0.01n}{0,5\sqrt{n}}\right) = 0.92/2 = 0.46, \Rightarrow \qquad \frac{0.01n}{0.5\sqrt{n}} \approx 1.75,$$

$$\Rightarrow 0.02\sqrt{n} = 1.75, \Rightarrow \sqrt{n} = 87.5, \Rightarrow n \approx 7657.$$

Задание 5. Небольшой город ежедневно посещают 100 туристов, которые днем идут обедать. Каждый из них выбирает для обеда один из двух городских ресторанов с равными вероятностями и независимо друг от друга. Владелец одного из ресторанов желает, чтобы с вероятностью приблизительно 0,99 все пришедшие в его ресторан туристы могли там одновременно пообедать. Сколько мест должно для этого быть в его ресторане?

Решение. Имеем схему Бернулии, $p = 0,5$, $n = 100$.

Обозначим через m – количество необходимых мест в ресторане. Тогда:

$$P\{X \leq m\} = 0,99.$$

$np = 50$, $\sqrt{npq} = 5$

По интегральной теореме М.-Л. имеем:

$$P\{X \leq m\} = \Phi_0\left(\frac{m-50}{5}\right) + 0,5 = 0,99.$$

$$\Phi_0\left(\frac{m-50}{5}\right) = -0,5 + 0,99 = 0,49$$

По таблице значений функции Лапласа находим:

$$\frac{m-50}{5} = 2,34.$$

Значит,

$m = 61,7$.

Таким образом, в ресторане должно быть 62 места.

7 Моделирование случайных величин

В практике создания и использования имитационных моделей весьма часто приходится сталкиваться с необходимостью моделирования случайных величин различных типов, а именно, необходимо уметь “разыгрывать” случайную величину X, т.е. получать набор ее реализаций, соответсвующих распределению.

Часто методы разыгрывания с.в. опираются на использование (псевдо) случайных чисел, полученных при помощи таблиц случайных чисел, генераторов (псевдо)случайных

чисел. Под таблицей случайных чисел в данном разделе будем понимать набор реализаций с.в. R, имеющей равномерное распределение на отрезке $[0,1]$ ($R_i \sim R(0,1)$).

Алгоритмы получения случайных числе при помощи генератора случайных чисел изучаются в разделах современной информатики и криптографии.

7.1 Приближённое разыгрывание нормально распределённых случайных величин

7.1.1 Стандартные нормальные величины

Пусть $X \sim N(0;1)$.

Необходимо получить k независимых реализаций с.в. X.

Рассмотрим $R_i \sim R(0,1)$, последовательность независимых одинаково распределенных с.в., имеющих равномерное распределение на $[0,1]$ ($a=0;\ b=1$). Тогда

$MR_i = \frac{a+b}{2} = \frac{1}{2},\ DR_i = \frac{(b-a)^2}{12} = \frac{1}{12}$.

Согласно центральной предельной теореме рассмотрим последовательность частичных сумм:

$$\bar{S}_n = \frac{\sum\limits_{i=1}^{n} R_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \xrightarrow{d} U \sim N(0;1).$$

Пусть $n = 12$. Тогда $\bar{S}_n \approx \sum\limits_{i=1}^{12} R_i - 6$.

Очевидно, что $n = 12$ не является большим числом, но очень примерно получившуюся $\bar{S}_n$ можно считать стандартной нормальной случайной величиной.

Тогда алгоритм получения реализаций $X \sim N(0,1)$ может быть следующим:

1. выбираем 12 значений r_i (реализаций с.в. R_i) из таблицы случайных чисел

2. Вычисляем $x_1 = \sum\limits_{i=1}^{12} R_i - 6$.

Данную процедуру повторяем еще $k-1$ раз для получения значений $x_2,\ x_3,\ \ldots, x_k$.

Каждый раз выбираются новые 12 значений из таблицы случайных чисел!

7.1.2 Нормальные величины с произвольными a, σ

Теперь рассмотрим $Y \sim N(a, \sigma^2)$.

Любая случайная величина Y может быть стандартизирована следующим образом:

сначала центрируем ее: $Y - M(Y)$ (тогда $M(Y - M(Y)) = M(Y) - M(Y) = 0$);

затем нормализуем: $\frac{Y-M(Y)}{\sqrt{D(Y)}}$ (тогда $D(\frac{Y-M(Y)}{\sqrt{D(Y)}}) = \frac{1}{D(Y)} D(Y - M(Y)) = \frac{D(Y)}{D(Y)} = 1$.)

Тогда если $Y \sim N(a, \sigma^2)$, то $X = \frac{Y-a}{\sigma} \sim N(0, 1)$.

Как разыгрывать $X \sim N(0, 1)$ см. выше.

Тогда получаем $x_i = \frac{y_i - a}{\sigma}$. Отсюда выражаем $y_i = a + \sigma x_i$.

7.2 Моделирование непрерывных случайных величин

Пусть X — непрерывная случайная величина. Предположим, что $F(x)$ монотонно возрастает на промежутке $[x_1, x_2]$. Моделируем выборку объёма $k : (x_1^*, \ldots, x_k^*)$.

Выберем k значений из таблицы случайных чисел:

$$r_j = F(x),\ x_j = F^{-1}(r_j),\ j = 1 \ldots, k.$$

Пример. Пусть $X \sim R(2; 5)$.

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{x-2}{3} & 2 \le x < 5 \\ 1 & x \ge 5. \end{cases}$$

$\frac{x-2}{3} = r_j \Rightarrow x_1 = 2 + 3r_1, x_2 = 2 + 3r_2, \ldots$

Задание. Разыграйте пять значений с.в., распределенной по экспоненциальному закону (с каким-либо конкретным параметром λ). Отметьте полученные значения на графике плотности распределения.

7.3 Моделирование дискретных случайных величин

Пусть X — дискретная с.в. с конечным числом возможных значений:

x_1	x_2	$\ldots$	x_n
p_1	p_2	$\ldots$	p_n

Как получить k реализаций с.в. X?

Поставим в соответствие каждому возможному значению с.в. X один (и только один) из интервалов из $[0; 1]$ следующим образом.

Разобьем интервал $[0;1]$ на n интервалов длины $p_1,\ p_2, \ldots,\ p_n$ с соответствующими границами интервалов $z_0,\ z_1, \ldots, z_n$.

$\Delta_1 : |\Delta_1| = p_1$

$\Delta_2 : |\Delta_2| = p_2$

$\ldots$

$\Delta_n : |\Delta_n| = p_n$

Пусть $R \sim R(0;1)$, плотность: $f(x) = 1\ \forall x \in [0,1]$, $f(x) = 0\ \ \forall x \notin [0,1]$.

Тогда

$$P\{R \in \Delta_i\} = \int_{z_{i-1}}^{z_i} f(x)dx = z_i - z_{i-1} = |\Delta_i| = p_i.$$

Из таблицы случайных чисел выбираем $r_1, r_2, \ldots, r_k$. Тогда вероятность события $P\{r_1 \in \Delta_i\} = p_i = P\{X = x_i\}$ и т.д.

$r_1 \in \Delta_i \Rightarrow X = x_i \Rightarrow x_1^*$

$r_2 \in \Delta_j \Rightarrow X = x_j \Rightarrow x_2^*$

$\ldots$

Задание. Разыграйте десять значений с.в., распределенной по следующему закону:

1	3	10	12
0.4	0.1	0.25	0.25

Как получившийся результат соотносится с $M(X)$? О чем нам говорит закон больших чисел применительно к данной задаче?

7.4 Вычисление определенных интегралов. Простейший метод Монте-Карло

Название метода происходит от названия коммуны в княжестве Монако, широко известного своими многочисленными казино, поскольку именно рулетка является одним из самых широко известных генераторов случайных чисел. Пусть нужно вычислить $\int_a^b g(x)dx$.

Рассмотрим случайную величину $u \sim R(a;b)$. Тогда плотность распределения с.в. u имеет вид:

$$f(x) = 1/(b-a)\ \forall x \in [a,b], \quad f(x) = 0 \ \ \forall x \notin [a,b].$$

Имеем:

$$M(g(u)) = \int_a^b g(x)f(x)dx = 1/(b-a)\int_a^b g(x)dx.$$

Отсюда получаем:

$$\int_a^b g(x)dx = (b-a)M(g(u)).$$

Математическое ожидание $M(g(u))$ можно заменить на выборочное среднее, т.е. смоделировать N значений с.в. u и взять их среднее арифметическое.

Итак, берем N значений с.в., равномерно распределённой на $[a,b]$ (можно взять N случайных чисел), для каждой точки u_i вычисляем $g(u_i)$. Затем вычисляем выборочное среднее: $\frac{1}{N}\sum_{i=1}^{N} g(u_i)$. В итоге получаем оценку интеграла:

$$\int_a^b g(x)dx \approx \frac{(b-a)}{N}\sum_{i=1}^{N} g(u_i).$$

Чем больше N, тем точнее полученная оценка.

Этот метод имеет и геометрическую интерпретацию. Он очень похож на описанный выше детерминистический метод, с той разницей, что вместо равномерного разделения области интегрирования на маленькие интервалы и суммирования площадей получившихся «столбиков» мы забрасываем область интегрирования случайными точками, на каждой из которых строим такой же «столбик», определяя его ширину как $\frac{b-a}{N}$, и суммируем их площади.

Задание. Методом Монте-Карло возьмите интеграл $\int\limits_0^3 x^2 dx$. Проиллюстрируйте на графике.

8 Многомерные случайные величины

8.1 Многомерные дискретные с.в.

Рассмотрим дискретные случайные величины $X : \Omega \to R^1$, $Y : \Omega \to R^1$. Дискретная случайная величина X называется *простой случайной величиной*, если у неё существует

конечный набор возможных значений $x_1, \ldots, x_n$ $(p_1, \ldots, p_n)$.

Будем говорить, что (X, Y) — *случайный вектор (двумерная случайная величина)*, если закон распределения имеет следующий вид:

X \ Y	y_1	$\ldots$	y_j	$\ldots$	y_m
x_1			$\vdots$		
$\vdots$			$\vdots$		
x_i	$\ldots$	$\ldots$	p_{ij}		
$\vdots$					
x_n					

где $p_{ij} = P\{X = x_i, Y = y_j\}$.

"Отдельно" рассмотренные распределения случайных величин X, Y случайного вектора (X, Y) называются *маргинальными*.

X	x_1	$\ldots$	x_n
P	$\sum_{j=1}^{m} p_{1j}$	$\ldots$	$\sum_{j=1}^{m} p_{nj}$

Y	y_1	$\ldots$	y_m
P	$\sum_{i=1}^{n} p_{i1}$	$\ldots$	$\sum_{i=1}^{n} p_{im}$

Заметим, что для нахождения маргинальных законов суммируются вероятности по строке и столбцу соответственно.

Случайные величины X, Y *независимы*, если

$$p_{ij} = P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\} = p_i \cdot p_j \quad \forall i = 1, \ldots, n \quad \forall j = 1, \ldots, m.$$

Совместная (n-мерная) функция распределения $X_1, \ldots, X_n$:

$$F(x_1, \ldots, x_n) = P\{X_1 < x_1, \ldots, X_n < x_n\}.$$

При независимости $X_1, \ldots, X_n$ верно:

$$F(x_1 \ldots x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n),$$

где $F_i(x_i)$ — маргинальная функция распределения случайной величины X_i.

Свойства $F(x_1, \ldots, x_n)$:

1. $F(x_1, \ldots, x_n) \in [0; 1]$
2. Неубывающая функция своих аргументов

   $\forall x_i'' > x_i' \;\; F(x_1 \ldots x_i'' \ldots x_n) \geq F(x_1 \ldots x_i' \ldots x_n) \;\; \forall i = 1, \ldots, n$
3. Непрерывность слева
4. (X, Y):

   $\lim\limits_{y \to \infty} F(x, y) = F_1(x)$

   $\lim\limits_{x \to \infty} F(x, y) = F_2(y)$

   $\lim\limits_{y \to -\infty} F(x, y) = 0$

   $\lim\limits_{\substack{x \to \infty \\ y \to -\infty}} F(x, y) = 0$

В общем случае случайные величины X, Y *независимы*, если $F(x, y) = F_1(x) \cdot F_2(y)$.

Если существует плотность, то для независимости имеем:

$f(x, y) = f_1(x) \cdot f_2(y)$

Для дискретных величин имеем следующее определение независимости:

$p_{ij} = p_i \cdot p_j, \forall i, j$

8.2 Корреляция случайных величин

Пусть имеются две случайные величины X и Y. Введём следующее определение.

Ковариация случайных величин X, Y:

$$cov\langle X, Y \rangle = K_{X,Y} = M((X - M(X)) \cdot (Y - M(Y))) = M(X \cdot Y) - M(X) \cdot M(Y)$$

Коэффициент корреляции:

$$r_{X,Y} = \frac{K_{X,Y}}{\sigma_x \cdot \sigma_y}$$

Свойства *корреляции*:

1. Если X, Y независимые случайные величины, то $r_{X,Y} = 0$

   Для независимых с.в. $M[X \cdot Y] = M(X)M(Y)$

   $K_{XY} = M[X \cdot Y] - M(X) \cdot M(Y) = 0$

Если X, Y независимые случайные величины, то $K_{XY} = 0 \Rightarrow r_{XY} = 0$, но обратное неверно

$r_{XY} = 0 \Rightarrow X, Y$ *некоррелируемые* случайные величины

2. $|r_{XY}| \leq 1$, причём $|r_{XY}| = 1 \Leftrightarrow Y = a \cdot X + b$ (величины связаны линейно)

Доказательство.

$r_{X,Y} = \frac{M(XY)-M(X)M(Y)}{\sigma(X)\sigma(Y)} = \frac{M[X(a\cdot X+b)]-MX\cdot M(aX+b)}{\sigma_x\cdot\sigma_y} = \frac{a\cdot M(X^2)+b\cdot M(XX)-a\cdot aM^2(X)-b\cdot M(X)}{|a|\cdot D(X)} = \frac{a(MX^2-M^2(X))}{|a|\cdot D(X)} = \frac{a\cdot D(X)}{|a|\cdot D(X)} = sign(a) \Rightarrow |r_{X,Y}| = 1$

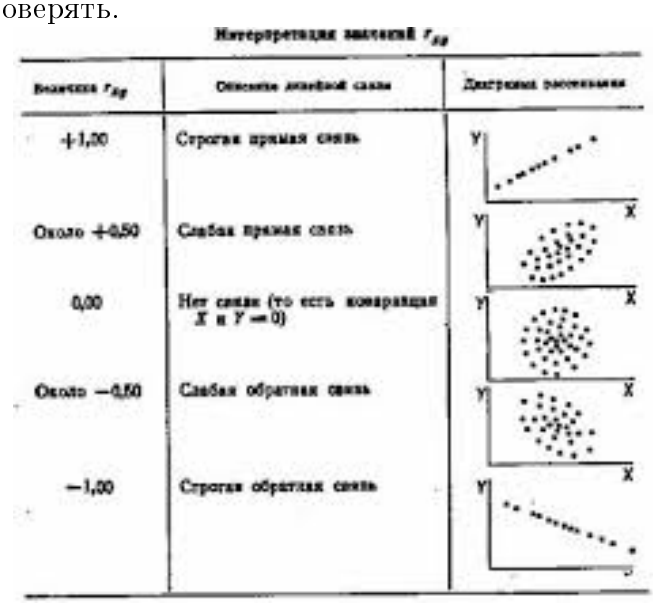
Дисперсия:

$$D(X \pm Y) = D(X) + D(Y) \pm 2K_{X,Y}. \tag{28}$$

Задание. Докажите формулу (28).

Замечание. 1. Коэффициент корреляции показывает степень *линейной* зависимости случайных величин.

2. На выборочных данных значимость коэффициента корреляции необходимо проверять.



8.3 Задачи

Задача 1. Из коробки, в которой 4 красных, 2 синих и 3 зеленых карандаша, наудачу извлекли 3 карандаша. Пусть X — число красных, Y — число синих карандашей среди отобранных. Найти совместное распределение X и Y, законы распределения X, Y.

Решение. Совместное распределение получаем по комбинаторным правилам.

X \ Y	0	1	2
0	1/84	1/14	1/28
1	1/7	2/7	1/21
2	3/14	1/7	0
3	1/21	0	0

Маргинальные распределения имеют вид:

X	0	1	2	3
P	5/42	10/21	5/14	1/21

Y	0	1	2
P	5/12	0,5	1/12

Задача 2. Два контракта случайным образом распределяются между тремя фирмами A, B, C. Обозначим через Y_1, Y_2 число контрактов, полученных фирмами A, B. Вычислите коэффициент корреляции с.в. Y_1, Y_2. Объясните полученный знак.

Совместное распределение получаем по комбинаторным правилам.

X \ Y	0	1	2
0	1/9	2/9	1/9
1	2/9	2/9	0
2	1/9	0	0

Очевидно, $M(Y_1Y_2) = 1 * 1 * 2/9 = 2/9$ (остальные нули в сумме).

Маргинальное распределение Y_i, $i = 1, 2$.

Y_i	0	1	2
P	4/9	4/9	1/9

Имеем $M(Y_i) = 2/3$, $D(Y_i) = 4/9$.

Тогда

$$cov(Y_1; Y_2) = M(Y_1Y_2) - M(Y_1)M(Y_2) = 2/9 - (2/3)^2 = -2/9.$$

Отсюда

$$r(Y_1; Y_2) = \frac{-2/9}{\sqrt{4/9^2}} = -1/2.$$

Корреляция отрицательная, т.к. "угол наклона прямой — тупой" (хотя здесь и плохо подходит линейная взаимосвязь). Увеличение числа контрактов одной фирмы приводит к уменьшению числа контрактов другой.

9 Математическая статистика, основные разделы

Математическая статистика — раздел математики, разрабатывающий методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений.

Пусть

$$X : \Omega \to R^1.$$

Функция распределения с.в. X часто неизвестна. Далее иногда будем называть с.в. X *генеральной совокупностью*, хотя так же называется и набор элементов одной природы, из которого отбирается некоторое подмножество (выборка). Изучим поведение случайной величины (с.в.) X на основе эмпирических данных.

Пусть $(x_1^*, x_2^*, \ldots, x_n^*)$ – наблюдаемые значения с.в. X. Будем говорить, что имеется *выборка* объёма n (n **независимых** реализаций с.в. X).

Иногда будем использовать обозначение $(X_1, X_2, \ldots, X_n)$ для выборки до ее фактического получения. В данном случае будем говорить, что имеется n копий случайной величины X.

Важным свойством выборки является ее *репрезентативность*, т.е. случайность отбора элементов выборки. Например, при изучении среднего роста студентов университета была сделана выборка из 30 студентов. Если все отобранные студенты оказались членами баскетбольной команды, то представление о среднем росте студентов всего университета будет искаженным.

Возможна следующая интерпретация при помощи урновой схемы. Имеется генеральная совокупность элементов X_1, X_2,...,X_N. Из этой генеральной совокупности n раз осуществляется *выбор с возвращением* и формируется выборка $x_1^*, \ldots, x_n^*$. Тогда

вероятность попадания каждого элемента генеральной совокупности в выборку одинакова и равна $1/N$.

Под репрезентативностью выборки часто понимается именно равновероятность попадания элементов генеральной совокупности в выборку.

Методы получения выборок (серийная, районированная и пр.) описаны в теории планирования эксперимента.

Будут изучены следующие разделы мат. статистики:

1. Дескриптивная (описательные) статистика
2. Точечные и интервальные оценки неизвестных параметров распределения с.в. X
3. Проверка гипотез о виде и параметрах распределения с.в. X
4. Анализ взаимосвязи между случайными величинами на основе выборочных данных. Корреляционный и регрессионный анализ

10 Дескриптивная (описательная) статистика

Пусть имеется выборка $(x_1^*, x_2^*, \ldots, x_n^*)$. Упорядочим ее значения в порядке невозрастания:
Вариационный ряд: $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, $\quad x_{(i)}$ — i-я варианта.

Если выборка рассмотрена до фактического появления, то будем говорить, что $X_{(1)} = min\{X_1, \ldots, X_n\}$ — первая порядковая статистика, ..., $X_{(n)} = max\{X_1, \ldots, X_n\}$ — n-ая порядковая статистика.

Под статистикой здесь и далее понимается функция выборки $f(X_1, \ldots, X_n)$.

10.1 Полигон частот

Пусть с.в. X — дискретная с.в.

Пример 1. Изучается распределения месяцев рождения студентов. Пусть 20 студентов группы сообщили свой месяц рождения. Имеем выборку объема 20:

7 3 2 5 8 2 11 10 9 6 11 6 5 12 1 6 3 11 7 3

Вариационный ряд:

1 2 2 3 3 3 5 5 6 6 6 7 7 8 9 10 11 11 11 12

Выделим различные элементы вариационного ряда (выборки) x_i, $i = 1, \ldots, k$ и вычислим количество их повторений n_i, $i = 1, \ldots, k$. Будем называть n_i *эмпирической или наблюдаемой частотой*. Очевидно, что

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

Разделим n_i на n. Будем называть n_i/n относительной эмпирической частотой. По закону больших чисел $n_i/n \approx p_i = P\{X = x_i\}$ в том случае, если X имеет дискретное (неизвестое нам) распределение вероятностей.

Статистическое распределение выборки:

x_1	x_2	$\ldots$	x_k	
n_1	n_2	$\ldots$	n_k	$\sum\limits_{i=1}^{k} n_i = n$

Статистическое распределение относительных частот (эмпирическое распределение):

x_1	x_2	$\ldots$	x_k	
n_1/n	n_2/n	$\ldots$	n_k/n	$\sum\limits_{i=1}^{k} n_i/n = 1$

Пример. Продолжим рассматривать пример с месяцами рождения студентов. Из вариационного ряда легко получаем статистическое распределение:

1	2	3	5	6	7	8	9	10	11	12	
1	2	3	2	3	2	1	1	1	3	1	$\sum = 20$

Иногда возникает вопрос, нужно ли записывать в таблицу стат. распределения значение с.в. X, которое не было реализовано в выборке. Например, мы знаем, что месяц рождения может быть любым из целых чисел $1, \ldots, 12$, однако в данной выборке месяц 4 реализован не был.

В таблицу стат. распред. не будем вписывать 4 и соответствующий 0, поскольку потенциально мы не знаем, каким распределением обладает с.в. X.

Эмпирическое распределение (относительных частот):

1	2	3	5	6	7	8	9	10	11	12	
1/20	2/20	3/20	2/20	3/20	2/20	1/20	1/20	1/20	3/20	1/20	$\sum = 1$

Графическое изображение статистического ряда называется *полигоном частот*. По закону больших чисел полигон относительных частот в пределе дает многоугольник распределения для дискретных случайных величин.

Пример. Для рассматриваемого примера имеем следующий полигон частот:

Полигон относительных частот отличается только масштабом.

Полигон имеет смысл строить для дискретных с.в., поскольку для непрерывных с.в. значения выборки, как правило, являются различными и график статистического ряда не является информативным.

10.2 Гистограмма распределения

Пусть X — непрерывная случайная величина, тогда целесообразно произвести группировку данных (разбиение рассматриваемого интервала от x_{min} до x_{max} на m

интервалов).

Интервальное разбиение: определяется интервал $[z_0; z_m]$ ($z_0 = x_{min}$; $z_m = x_{max}$), в пределах которого варьируются значения выборки, затем данный интервал делится на m частичные интервалы, и по каждому интервалу j подсчитываются частоты a_j — количество вариант, которые в него попали. В дальнейшем иногда будем писать вместо a_i прежнее обозначение эмпирической частоты n_j.

Очевидно, что

$$\sum_{j=1}^{m} a_j = n,$$

где n — объем выборки.

Часто (но не обязательно) интервалы выбираются равной длины.

Графическое изображение группированного ряда с равными интервалами называется *гистограммой* частот. В пределе (при больших n) гистограмма позволяет судить о поведении плотности распределения с.в. X.

Если интервалы не равны, то по оси OY откладывают $\frac{a_j}{|\delta z_j|}$.

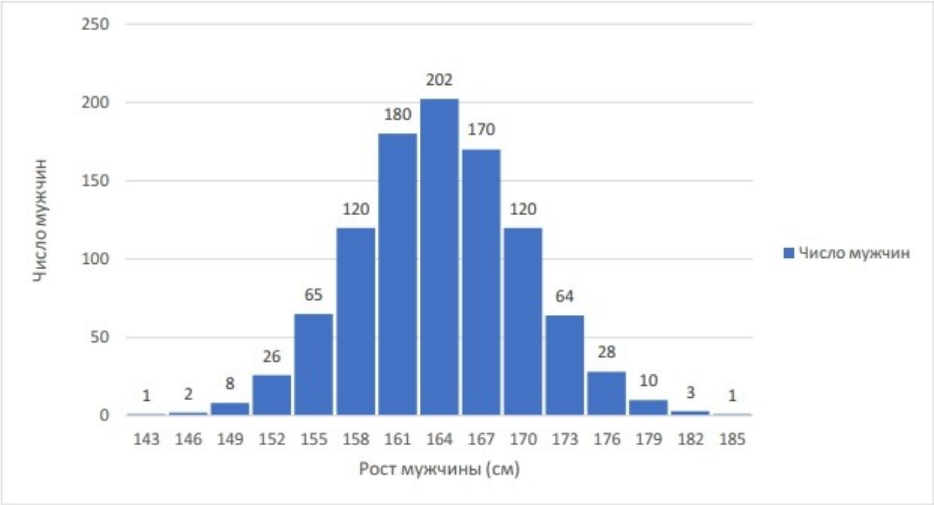
Часто группированный ряд получен специалистами в предметной области и предоставляется статистикам в готовом виде.

Пример 2. В 1889-1890 гг. был измерен рост 1000 взрослых мужчин (рабочих московских фабрик). Результаты измерений представлены в таблице.

Рост (см) $(z_{i-1}; z_i]$	[143; 146]	(146; 149]	(149; 152]	(152; 155]	(155; 158]
Число мужчин a_i	1	2	8	26	65
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]
Число мужчин a_i	120	180	202	170	120
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	(185; 188]
Число мужчин a_i	64	28	10	3	1

Замечание. Обратите внимание на вид интервалов (за исключением первого и последнего): $(z_{i-1}; z_i]$. Это означает, что значение выборки, попавшее на границу интервалов учитывается В ЛЕВОМ интервале.

Построим гистограмму распределения.



Очевидно, что для данного эмпирического распределения имеет смысл рассматривать гипотезу о нормальности распределения генеральной совокупности.

Из гистограммы частот может быть получен полигон частот, например, путем соединения середин интервалов. Тем не менее, точность при этом, очевидно, теряется.

Оптимальное число интервалов разбиения определяется, как правило, эмпирической **формулой Стерджесса**:

$$m = [1 + \log_2 n] \approx [1 + 3.322 \cdot \log_{10} n],$$

либо может быть задано из условия задачи.

Квадратные скобки означают целую часть числа.

В примере 2 (рост мужчин) $n = 1000$. Тогда оптимальное число интервалов разбиения $m = 10$. (Тем не менее, данные сгруппированы в 8 интервалах по условию задачи).

10.3 Выборочные характеристики

Пусть изучается поведение случайной величины X. Истинные значения ее числовых характеристик (математическое ожидание, дисперсия и т.д.) нам неизвестны. В дальнейшем будем называть их *теоретическим* математическим ожиданием $M(X)$, теоретической дисперсией $D(X)$ и т.д. По наблюдаемым значениям случайной величины X могут быть получены *выборочные* числовые характеристики.

Напомним, что

$$\frac{n_i}{n} \approx p_i.$$

(Чем больше объем выборки n, тем ближе эти значения).

Для дискретных случайных величин математическое ожидание вычислялось по следующей формуле:

$$M(X) = \sum_i x_i p_i.$$

Подставив вместо p_i выражение $\frac{n_i}{n}$ приходим к понятию выборочного математического ожидания:

$$\sum_i x_i \frac{n_i}{n} = \frac{1}{n} \sum_i x_i n_i.$$

Выборочное математическое ожидание (выборочное среднее):

$$\bar{x} = \mu^* = \frac{1}{n} \sum_{i=1}^{k} x_i n_i = \frac{1}{n} \sum_{i=1}^{n} x_i^*.$$

Выборочная дисперсия:

$$\sigma^{2^*} = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^{k} x_i^2 n_i - (\bar{x})^2.$$

Замечание. ВАЖНО! В литературе часто встречается обозначение s^2 для выборочной дисперсии. Тем не менее, в дальнейшем мы иногда будем использовать так называемую "исправленную" выборочную дисперсию (unbiased sample variance):

$$s^2 = \frac{n}{n-1} \sigma^{2^*} = \frac{1}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i.$$

В литературе исправленная выборочная дисперсия иногда обозначается как $\tilde{s}^2$ и т.д. Во избежание путаницы будем придерживаться введенных в данном параграфе обозначений.

Знак $*$ всегда будет обозначать принадлежность характеристики к выборочной.

Стандартное отклонение:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}.$$

Пример 1. Продолжим рассматривать пример про месяцы рождения. Получаем

$$\bar{x} = 6,4; \qquad \sigma^{2^*} \approx 12,45; \qquad s^2 \approx 11,83; \qquad s \approx 3,44.$$

Если имеется группировочный ряд, то выборочные математическое ожидание и дисперсия вычисляются аналогичным образом. Принято вместо значений x_i брать середины интервалов разбиения:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} t_j a_j$$

$$\sigma^{2^*} = \frac{1}{n}\sum_{j=1}^{m}(t_j - \bar{X})^2 a_j = \frac{1}{n}\sum_{j=1}^{m} t_j^2 a_j - (\bar{X})^2$$

где $t_j = \frac{z_{j-1}+z_j}{2}$, $z_j = \begin{cases} x_{min} & j = 0 \\ x_{max} & j = r \end{cases}$

Начальный выборочный момент порядка s : $\mu_s^* = \frac{1}{n}\sum_{i=1}^{k} x_i^s n_i$

Центральный момент порядка s: $\nu_s^* = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^s n_i$

Выборочные моменты более высоких порядков могут пригодиться для вычисления асимметрии и эксцесса распределения (для сравнения статистического распределения с нормальным).

Пример 2. Вычислим выборочное мат.ожидание и дисперсию для примера с ростом мужчин.

Получаем:

$$\bar{x} = 165,533;\ \sigma^{2^*} = 36,565911,\ s^2 = 36,60251351,\ s \approx 6,05.$$

При больших объемах выборок n разница между исправленной и неисправленной выборочной дисперсией стирается.

Мода: x_m — наиболее часто встречающееся значение в выборке. Очевидно, что статистическое распределение может иметь не одну моду. Соответственно, различают “унимодальные”, “бимодальные”, “мультимодальные” распределения.

Пример 1. Статистическое распределение месяцев рождения имеет три моды (3, 6, 11).

Выборочный коэффициент корреляции:

$$r^*_{X,Y} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma^*_X \cdot \sigma^*_Y}.$$

10.4 Выборочная функция распределения

$F(x) = P\{X < x\}$ - *теоретическая функция распределения.* Она обычно неизвестна либо известна с точностью до параметров распределения.

$x_1^*, \ldots, x_n^*$ — выборка; $x_{(1)}, \ldots, x_{(n)}$ — вариационный ряд.

Введём *эмпирическую функцию распределения*:

$$F_n^*(x) = \begin{cases} 0, & x \le x_{(1)} \\ \frac{n_x}{n}, & x_{(1)} < x \le x_{(n)}, \\ 1, & x > x_{(n)}, \end{cases}$$

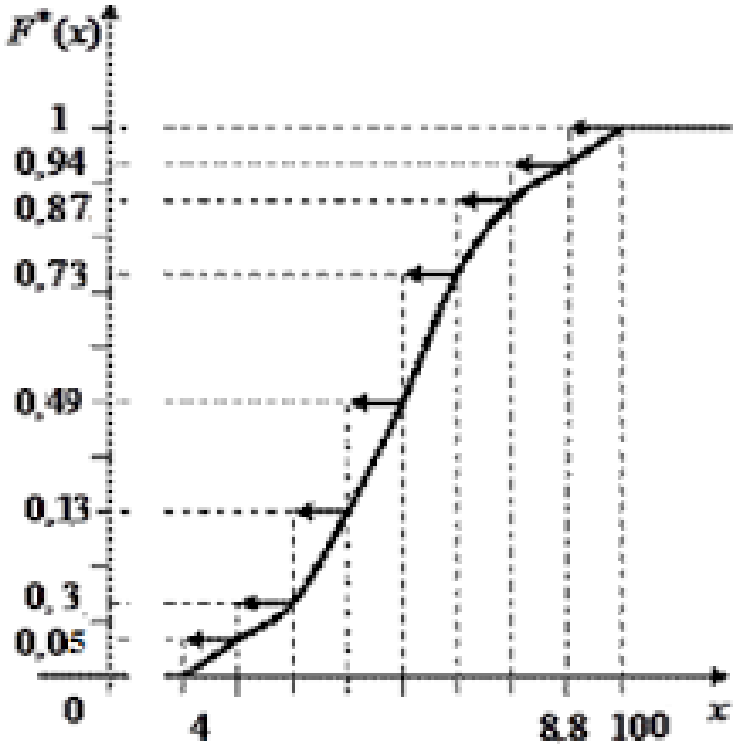
где n_x — число вариант в вариационном ряду строго левее x $(< x)$.

Свойства эмпирической функции распределения:

1. $F_n^*(x) \in [0; 1]$.
2. $F_n^*(x)$ — неубывающая функция.
3. $\lim\limits_{x \to x_0 - 0} F_n^*(x) = F_n^*(x_0)$ (непрерывность слева).
4. $\lim\limits_{x \to +\infty} F_n^*(x) = 1$.
   $\lim\limits_{x \to -\infty} F_n^*(x) = 0$.

График эмпирической функции распределения имеет ступенчатый вид (как для дискретных с.в. в теории вероятностей) и показывает, как накапливаются частоты в статистическом распределении. Если соединить накопленные частоты на графике эмпирической ф.р., то получится *кумулята распределения.*

Огива строится аналогично кумуляте с той лишь разницей, что накопленные частоты помещают на оси абсцисс, а значения признака — на оси ординат.



Интуитивно понятно, что при больших n эмпирическая функция распределения ведет себя как теоретическая (неизвестная нам). Следующая теорема подтверждает это и позволяет нам в дальнейшем заменять неизвестные теоретические числовые характеристики на выборочные.

Теорема Гливенко-Кантелли.

$$P\{sup_x|F(x) - F_n^*(x)| \underset{n\to\infty}{\to} 0\} = 1$$

10.5 Квантиль, квартиль, медиана. Диаграмма “ящик с усами”

Пусть X имеет функцию распределения $F(x)$. Тогда по значению аргумента x можно найти значение функции $F(x)$.

Иногда нужно решить противоположную задачу: по значению функции распределения $F(x) = p$ найти соответствующее значение аргумента $x_p = F^{-1}(p)$. Тогда x_p называется квантилью порядка p.

Квантиль порядка $p = 1/2$ называется *медианой* распределения. Медиана — значение, которое делит распределение пополам: ровно половина значений находится левее и ровно

половина значений находится правее.

Квартили: первый квартиль $Q_1 = x_{0,25}$ — квантиль порядка $p = 1/4$; второй квартиль $Q_2 = x_{0,5}$ — медиана ($p = 1/2$); третий квартиль $Q_3 = x_{0,75}$.

Межквартильный размах:

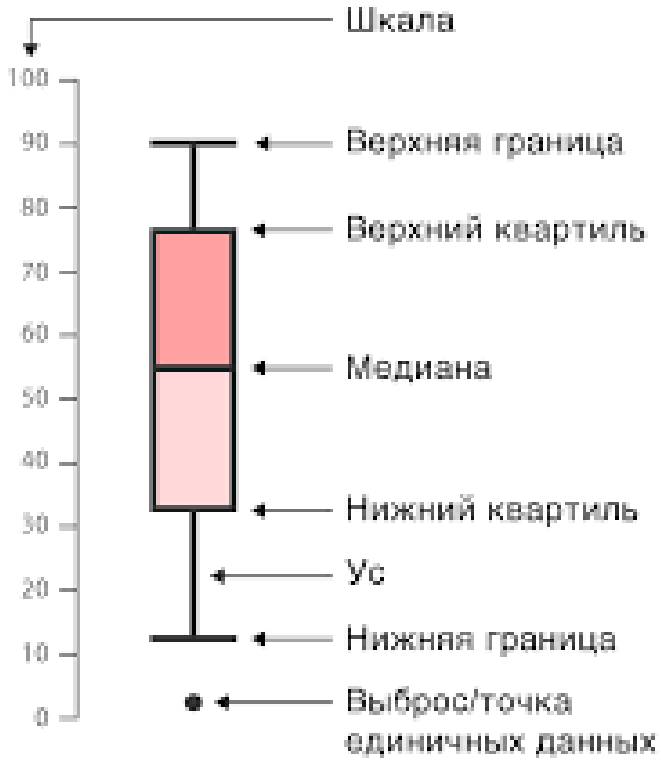
$$Q_3 - Q_1.$$

Децилъ: квантили $x_{0,1}, \ldots, x_{0,9}$.

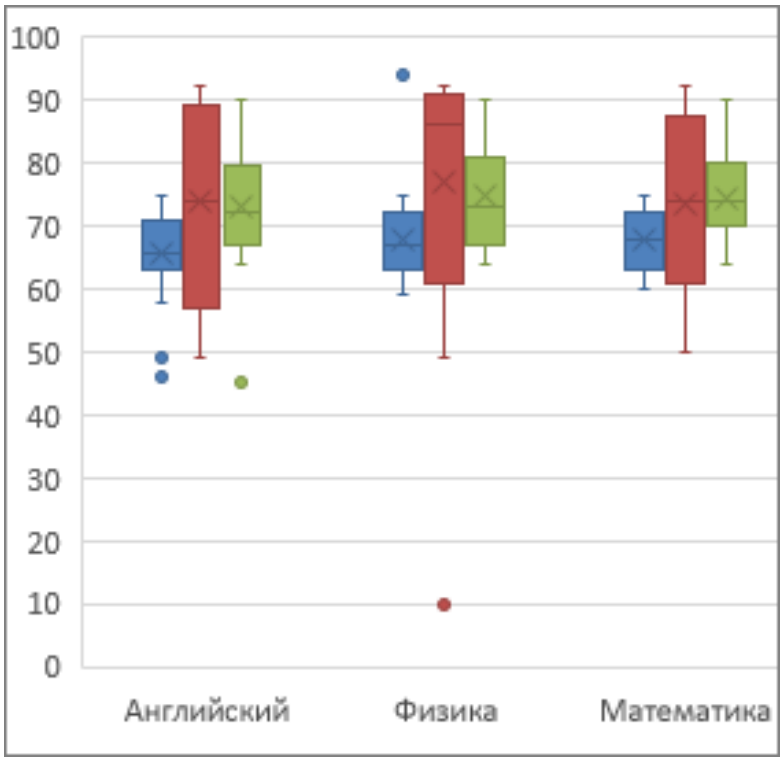
Персентилъ: квантили $x_{0,01}, \ldots, x_{0,99}$.

По выборке данные характеристики также могут быть посчитаны.

Вариативность признака может быть изображена при помощи диаграммы “ящик с усами” (boxplot diagram), один из вариантов которой представлен ниже.



С помощью диаграммы ящик с усами можно визуально сравнивать распределения нескольких выборок.



10.6 Задачи

Задача 1.

В таблице приведены значения обменных курсов двух валют (реал и дублон).

	Реал	Дублон
Январь	0.15	0.03
Февраль	0.18	0.05
Март	0.12	0.06
Апрель	0.14	0.08
Май	0.19	0.07
Июнь	0.19	0.05
Июль	0.18	0.07
Август	0.13	0.04
Сентябрь	0.11	0.05
Октябрь	0.10	0.05
Ноябрь	0.12	0.06
Декабрь	0.17	0.08

Найдите медиану, выборочное среднее, выборочную дисперсию, размах, межквартильный размах каждого распределения. Сравните эти распределения.

Решение

Упорядочим выборки по возрастанию:

X= (0,1; 0,11; 0,12; 0,12; 0,13; 0,14; 0,15; 0,17; 0,18; 0,18; 0,19; 0,19)

Y=(0,03; 0,04; 0,05; 0,05; 0,05; 0,05; 0,06; 0,06; 0,07; 0,07; 0,08; 0,08)

1. Найдем выборочные медианы по формуле:

$$me = \begin{cases} \frac{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}}{2}, & \text{если } n \text{ четное,} \\ x_{(\frac{n+1}{2})}, & \text{если } n \text{ нечетное.} \end{cases}$$

В нашем случае $n_1 = n_2 = 12$, значит:

$$me_1 = \frac{0,14 + 0,15}{2} = 0,145,$$

$$me_2 = \frac{0,05 + 0,06}{2} = 0,055.$$

2. Найдём выборочное среднее и выборочную дисперсию для каждой выборки:

$$\overline{x} = \frac{\sum\limits_{i=1}^{12} x_i}{12} = 0,1483,$$

$$\overline{y} = \frac{\sum\limits_{i=1}^{12} y_i}{12} = 0,0575,$$

$$\sigma_1^{2*} = \frac{\sum\limits_{i=1}^{12} x_i^2}{12} - \overline{x}^2 = 0,000981,$$

$$\sigma_2^{2*} = \frac{\sum\limits_{i=1}^{12} y_i^2}{12} - \overline{y}^2 = 0,000219.$$

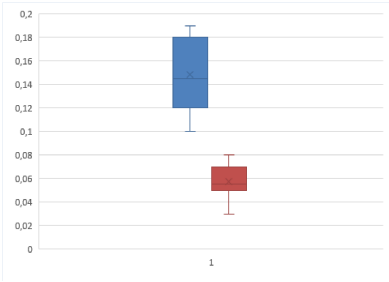
3. Размах для первой выборки $R_1 = X_{(12)} - X_{(1)} = 0,19 - 0,1 = 0,09$, для второй - $R_2 = Y_{(12)} - Y_{(1)} = 0,08 - 0,03 = 0,05$.

4. Найдём межквартильные размахи. Выборочную квартиль порядка α будем искать как элемент выборки с номером $[n\alpha] + 1$. Q_1 – квартиль порядка 0,25, элемент выборки с номером $[\frac{n}{4}] + 1 = 4$, Q_3 – квартиль порядка 0,75, элемент выборки с номером $[\frac{3n}{4}] + 1 = 10$.

Для первой выборки: $Q_1 = 0,12$, $Q_3 = 0,18$, $Q_3 - Q_1 = 0,06$.

Для второй выборки: $Q_1 = 0,05$, $Q_3 = 0,07$, $Q_3 - Q_1 = 0,02$.

Для сравнения двух выборок построим диаграмму "ящик с усами".



Видно, что значительной разницы между распределениями нет. С точностью до масштаба вариабельность примерно одинакова. У обоих распределений медиана находится посередине между наибольшим и наименьшим значениями выборки, что говорит нам о симметричности распределения. Однако другие характеристики распределения указывают на их несимметричность: верхний ус у обоих распределений короче нижнего, что может говорить в пользу скошенности влево.

Задача 2.

В таблице приведены кумулятивные (накопленные) частоты распределения дохода в США в 1950 и 1959 гг.

Доход, $	1950	1959
$<$ 1000	0.13	0.07
1000–1999	0.28	0.19
2000–2999	0.46	0.29
3000-3999	0.64	0.40
4000–4999	0.79	0.52
5000–7499	0.93	0.78

Найдите приближенно медиану и межквартильный размах каждого распределения. На основании полученных результатов сравните распределения дохода в 1950 и 1959 гг.

Решение

Для определения медианы нужно найти интервал, с кумулятивной частотой близкой к 0,5. Для 1950 г. мы выберем интервалы 2000–2999 и 3000–3999 и найдём среднее значение между их серединами, это и будет приближенное значение медианы:

$$me_1 \approx \frac{2500 + 3500}{2} = 3000.$$

Для 1959 г. мы выберем интервалы 3000–3999 и 4000–4999:

$$me_2 \approx \frac{3500 + 4500}{2} = 4000.$$

Медианный доход вырос на 33%. Это говорит о том, что существенно возросла доля состоятельных людей.

Для определения Q_1 и Q_3 необходимо найти интервалы, соответствующие кумулятивным частотам 0,25 и 0,75.

Для 1950 г.: $Q_1 \approx 1500$, $Q_3 \approx 4500$, $Q_3 - Q_1 = 3000$.

Для 1959 г.: $Q_1 \approx 2500$, $Q_3 \approx 6000$, $Q_3 - Q_1 = 3500$.

Межквартильный размах вырос, это позволяет говорить о том, что распределение дохода в 1959 г. более «размазано», более вариабельно, чем в 1959 г.

Задача 3.

Управляющий службы доставки корреспонденции намеревается обновить парк грузовиков. При погрузке пакетов следует учитывать два вида ограничений: вес (в фунтах) и объём (в кубических футах) каждого пакета.

Предположим, что в выборке, содержащей 200 пакетов, средний вес пакетов равен 26 фунтов, стандартное отклонение веса - 3,9 футов, средний объём пакета – 8,8 кубических футов, а стандартное отклонение объёма – 2,2 кубических фута. Как сравнить разброс веса и объёма пакетов?

Решение

Поскольку единицы измерения веса и объёма отличаются друг от друга, управляющий должен сравнить относительный разброс этих величин. Можно использовать для этого коэффициент вариации, равный

$$v = \frac{s}{\overline{x}}.$$

Коэффициент вариации веса равен $\frac{3,9}{26} = 0,15$, а коэффициент вариации объёма – $\frac{2,2}{8,8} = 0,25$. Таким образом, относительный разброс объёма пакетов намного больше относительного разброса их веса.

Задача 4.

Системный администратор, руководящий работой корпоративной сети, подсчитывает количество сбоев сервера, происходящих за день. В следующей выборке приведены данные его наблюдений за последние две недели $X_{[14]} = (1; 3; 0; 3; 26; 2; 7; 4; 0; 2; 3; 3; 6; 3)$. Вычислить моду, выборочное среднее и межквартильный размах для данной выборки.

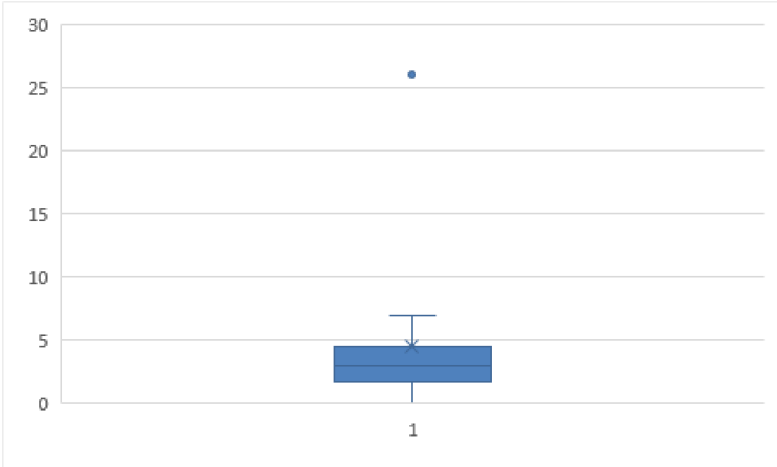
Решение

Упорядочим значения выборки по возрастанию

0 0 1 2 2 3 3 3 3 3 4 6 7 26

У данной выборки мода равна 3, медиана также равна 3, $\overline{x} = 4,5$, $Q_1 = 2$, $Q_3 = 4$, $Q_3 - Q_1 = 2$, $R = X_{14} - X_1 = 26$.

Для полученных данных построим диаграмму "ящик с усами".



На диаграмме отдельная точка соотвествует значению выборки 26. Эта точка является выбросом.

Как мы видим, при наличии выброса, медиана и мода больше подходят для оценки среднего значения, поскольку не учитывают значения выбросов. Также видно, что межквартильный размах, учитывающий только 50% выборки и не учитывающий выбросы, лучше характеризует вариативность выборки, чем размах.

Задача 5.

Постройте эмпирическую функцию распределения для выборки

$X = (0; 2; 1; 2; 6; 3; 1; 4; 6; 1; 4; 6; 6; 2; 6; 6; 7; 9; 9; 2; 6)$.

Решение:

Статистическое распределение выборки:

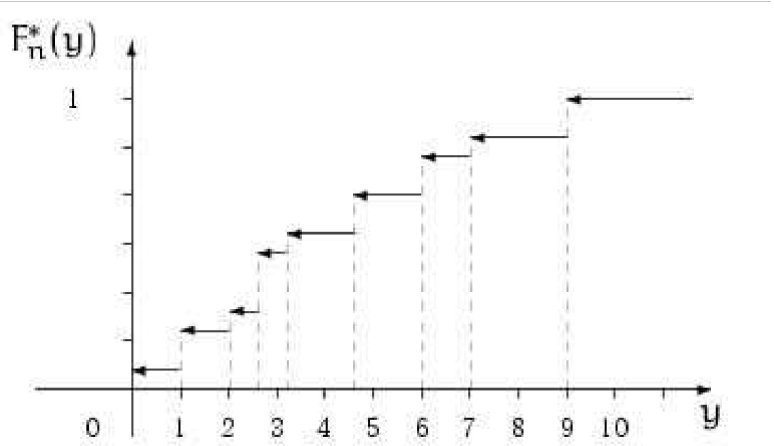
x_i	0	1	2	3	4	6	7	9
n_i	1	3	4	1	2	7	1	2

Объём выборки – $n = 21$. Вычислим эмпирическую функцию распределения: $F_n^*(x) =$

$$\begin{cases} 0, & \text{при } x \le 0, \\ \frac{1}{21}, & \text{при } 0 < x \le 1, \\ \frac{1+3}{21}, & \text{при } 1 < x \le 2, \\ \frac{1+3+4}{21}, & \text{при } 2 < x \le 3, \\ \frac{1+3+4+1}{21}, & \text{при } 3 < x \le 4, \\ \frac{1+3+4+1+2}{21}, & \text{при } 4 < x \le 6, \\ \frac{1+3+4+1+2+7}{21}, & \text{при } 6 < x \le 7, \\ \frac{1+3+4+1+2+7+1}{21}, & \text{при } 7 < x \le 9, \\ \frac{1+3+4+1+2+7+1+2}{21}, & \text{при } x > 9. \end{cases}$$

Эмпирическая функция распределения имеет вид:

$$F_n^*(x) = \begin{cases} 0, & \text{при } x \le 0, \\ \frac{1}{21}, & \text{при } 0 < x \le 1, \\ \frac{4}{21}, & \text{при } 1 < x \le 2, \\ \frac{8}{21}, & \text{при } 2 < x \le 3, \\ \frac{9}{21}, & \text{при } 3 < x \le 4, \\ \frac{11}{21}, & \text{при } 4 < x \le 6, \\ \frac{18}{21}, & \text{при } 6 < x \le 7, \\ \frac{19}{21}, & \text{при } 7 < x \le 9, \\ 1, & \text{при } x > 9. \end{cases}$$



11 Точечные оценки для неизвестных параметров распределения

11.1 Свойства точечных оценок

Пусть имеется генеральная совокупность X. Предположим, что функция распределения случайной величины X известна с точностью до параметров распределения. Обозначим её как $F(x, \vartheta)$.

Рассмотрим выборку $X_1, \ldots, X_n$ до её фактического получения. Рассмотрим функцию выборки $f(X_1, \ldots, X_n)$. Функция от выборки также называется *статистикой*.

Пусть распределение с.в. X известно с точностью до параметра ϑ, $\vartheta \in \Theta \subset R^m$. Например, известно, что рост человека распределен по нормальному закону, но неизвестны a, σ; частички золота в песке распределены по закону Пуассона, но неизвестен параметр λ и т.д.

Точечная оценка $\hat{\vartheta}$ неизвестного параметра ϑ —статистика $\hat{\vartheta} = f(X_1, \ldots, X_n)$, которую с той или иной надёжностью можно взять за истинное значение ϑ.

Замечание. Заметим, что $\hat{\vartheta} = f(X_1, \ldots, X_n)$ — случайная величина! На разных выборках дает различные значения!

11.1.1 Несмещенные оценки

Определение. Если

$$M\hat{\vartheta} = \vartheta,$$

то $\hat{\vartheta}$ называется *несмещённой оценкой* неизвестного параметра ϑ.

Примеры.

1. Биномиальное распределение, пусть неизвестный параметр p. В качестве точечной оценки возьмем $\hat{\vartheta} = \frac{S_n}{n} \Rightarrow M\left(\frac{S_n}{n}\right) = \frac{M_n}{n} = \frac{np}{n} = p$
   (S_n — число успехов, т.е. число появлений события A в выборке объема n. Тогда $\frac{S_n}{n}$ — относительная частота появления события A в выборке.)
2. Дана случайная величина X, $\vartheta = M(X)-?$
   $\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

$M(\hat{\vartheta}) = M(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} MX_i = M(X) = \vartheta$

3. $\vartheta = D(X) = \sigma^2 - ?$

   $\hat{\vartheta} = \sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$

   $M(\sigma^{2*}) = \frac{1}{n}\sum X_i^2 - (\frac{\sum X_i}{n})^2 = \frac{1}{n}\sum MX_i^2 - \frac{1}{n^2}M(\sum X_i)^2 = MX^2 - \frac{1}{n^2}M(\sum X_i)^2 = MX^2 - \frac{1}{n^2}(n\sigma^2 + n^2(MX)^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \Rightarrow \sigma^{2*}$ не является несмещённой оценкой для $\sigma^2 \Rightarrow$ рассмотрим $S^2 = \frac{n}{n-1}\sigma^{2*} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ — исправленная выборочная дисперсия (unbiased sample variance).

Задание 1. Покажите, что S^2 является несмещенной оценкой для неизвестной дисперсии σ^2. ($\hat{\vartheta} = S^2 \Rightarrow M(S^2) = \sigma^2$ несмещённая оценка.)

Замечание. При больших n разница между выборочной и исправленной выборочной дисперсией стирается.

В литературе под S^2 часто понимается σ^{2*}.

В англояз. литературе: σ^{2*} — biased sample variance; S^2 — unbiased sample variance.

Определение. Если

$$M\hat{\vartheta} \to_{n\to\infty} \vartheta,$$

то $\hat{\vartheta}$ называется *асимптотически несмещённой оценкой* неизвестного параметра ϑ.

11.1.2 Эффективные оценки

Пусть ϑ, т.ч. $\vartheta \in \Theta \subset R^1$.

Определение. $\hat{\vartheta}$ называется *эффективной точечной оценкой*, если

1. $\hat{\vartheta}$ — несмещённая оценка;
2. $\hat{\vartheta} = \min\limits_{\tilde{\vartheta}\in K} D(\tilde{\vartheta})$, где K — класс несмещенных оценок неизвестного параметра ϑ.

Неравенства Рао-Крамера (частный случай). Пусть $\hat{\vartheta}$ - несмещённая оценка. Тогда

$$D(\hat{\vartheta}) \geq \left\{ nM\left[\left(\frac{\partial}{\partial\vartheta} \ln f(X,\vartheta) \right)^2 \right] \right\}^{-1},$$

если X — абсолютно-непрерывная случайная величина;

$$D(\hat{\vartheta}) \geq \left\{ nM\left[\left(\frac{\partial}{\partial \vartheta} \ln p_X(\vartheta) \right)^2 \right] \right\}^{-1},$$

если X — дискретная случайная величина.

Замечание. На практике эффективность оценок часто проверяется с помощью неравенства Рао-Крамера. Если дисперсия оценки $D(\hat{\vartheta})$ в точности совпадает с выражением в правой части неравенства, значит, это и есть минимальное значение дисперсии на классе несмещенных оценок.

Пример. Пусть $\xi \sim N(a, \sigma^2)$. Доказать, что оценка $\widehat{a} = \bar{X}$ является эффективной оценкой неизвестного параметра a.

Решение.

Воспользуемся неравенством Рао-Крамера:

$$\frac{\partial \ln f_\xi(X,a)}{\partial a} = \frac{\partial \ln(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-a)^2}{2\sigma^2}})}{\partial a} = \frac{(X-a)^2}{\sigma^2}$$

Тогда

$$M(\frac{\partial \ln f_\xi(X,a)}{\partial a})^2 = \frac{M(X-a)^2}{\sigma^4} = \frac{DX}{\sigma^4} = \frac{1}{\sigma^2}.$$

Тогда правая часть неравенства Рао-Крамера равна $I_n(a) = nI_1(a) = \frac{\sigma^2}{n}$.

Найдём дисперсию оценки $\overline{X}$

$$D\overline{X} = \frac{\sum\limits_{i=1}^{n} DX_i}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Сравнивая левую и правую часть неравенства Рао-Крамера получаем равенство:

$$D\overline{X} = \frac{DX_1}{n} = \frac{\sigma^2}{n},$$

эффективность доказана.

Задание 2. Пусть $\xi \sim P(\lambda)$. Доказать, что оценка $\widehat{\lambda} = \bar{X}$ является эффективной оценкой неизвестного параметра λ.

11.1.3 Состоятельные оценки

Определение. Оценка называется *состоятельной*, если $\forall \epsilon > 0$ $P\{|\vartheta - \hat{\vartheta}| > \epsilon\} - > 0$ if $n - > \infty$, т.е.

$$\hat{\vartheta} \xrightarrow{p} \vartheta.$$

Теорема. Пусть $\hat{\vartheta}$ — несмещённая оценка $\Rightarrow \hat{\vartheta}$ — состоятельная $\Leftrightarrow D(\hat{\vartheta}) \to_{n\to\infty} 0$

Доказательство основано на использовании неравенства Чебышева:

$$P\{|X - M(X)| > \varepsilon\} \leq \frac{D(X)}{\varepsilon^2}.$$

Здесь имеем $M(\hat{\vartheta}) = \vartheta$. Тогда

$$P\{|\hat{\vartheta} - \vartheta| > \varepsilon\} = P\{|\hat{\vartheta} - M(\hat{\vartheta})| > \varepsilon\} \leq \frac{D(\hat{\vartheta})}{\varepsilon^2}.$$

Очевидно, что если $D(\hat{\vartheta}) \to_{n\to\infty} 0$, то по лемме о двух милиционерах (поскольку $P\{ldots\}$ ограничена снизу нулем), получаем состоятельность оценки.

В обратную сторону утверждение доказывается аналогично.

Пример.

Пусть выборка извлекается из генеральной совокупности, подчиняющейся закону распределения Пуассона $P(\lambda), \lambda > 0$. Исследуем несмещенность и состоятельность оценки $\widehat{\lambda} = \overline{X}$ параметра λ.

Решение. Для распределения Пуассона справедлива формула $M(X) = \lambda$. Ранее было показано, что выборочное среднее является несмещенной оценкой математического ожидания (для любого распределения). Значит несмещенность оценки $\widehat{\lambda}$ доказана, т.к. $M(\widehat{\lambda}) = M(X) = \lambda$.

Известно, что для распределения Пуассона $D(X) = \lambda$. Тогда дисперсия оценки $D\widehat{\lambda} = D\overline{X} = \frac{nD(X)}{n^2} = \frac{D(X)}{n} = \frac{\lambda}{n} \xrightarrow[n\to\infty]{} 0$. Состоятельность доказана.

Задание 3. Пусть $\xi \sim B(n, p)$. Доказать, что оценка $\widehat{p} = \frac{S_n}{n}$ является состоятельной оценкой неизвестного параметра p.

Наилучшая точечная оценка — несмещённая, эффективная, состоятельная оценка.

Задание 4. Приведите пример наилучшей оценки.

11.2 Основные методы построения точечных оценок

11.2.1 Метод моментов.

Метод моментов (Карл Пирсон).

Пусть ϑ — неизвестный параметр распределения. Отметим, что ϑ может быть вектором! (Неизвестно несколько параметров, как, например, для нормального, равномерного, биномиального распределений).

Суть метода (строгое обоснование см. лекции Черновой Н.И.): для получения точечной оценки неизвестного параметра ϑ нужно приравнять теоретические моменты порядка k к эмпирическим моментам порядкам k:

$$\mu_k = \mu_k*, \tag{29}$$

где k — размерность вектора ϑ.

Таким образом, нужно брать столько моментов, сколько неизвестных параметров!

Для однопараметрических распределений (Пуассона, геометрическое, экспоненциальное и т.д.) имеем:

$$M(X) = \bar{X}$$

Для двухпараметрических распределений (нормальное, равномерное, биномиальное и т.д.) получаем:

$$M(X) = \bar{X},\ D(X) = \sigma^{2*}.$$

Пример 1. Имеется выборка $x_1^*, \dots x_n^*$. Известно, что с.в. X распределена по закону Пуассона. Оценить неизвестный параметр λ методом моментов.

Решение. Известно, что для р. Пуассона $M(X) = \lambda$. Отсюда

$$M(X) = \lambda = \bar{X},$$

решение этого уравнения $\hat{\lambda} = \bar{X}$. В прикладных задачах даны реальные выборочные значения.

Пример 2. Имеется выборка $x_1^*, \dots x_n^*$. Известно, что с.в. X распределена по экспоненциальному закону. Оценить неизвестный параметр λ методом моментов.

Решение. Известно, что для экспоненциального распределения $M(X) = \frac{1}{\lambda}$. Отсюда

$$M(X) = \frac{1}{\lambda} = \bar{X},$$

решение этого уравнения $\hat{\lambda} = \frac{1}{\bar{X}}$. В прикладных задачах даны реальные выборочные значения.

Задание 5. Получить выражения для оценок неизвестных параметров нормального, равномерного и биномиального распределения методом моментов.

11.2.2 Метод наибольшего правдоподобия

Метод наибольшего правдоподобия (Р. Фишер)

Оценку, полученную данным методом, будем называть МНП-оценкой. Рассмотрим применение метода для дискретных и абсолютно непрерывных случайных величин.

Пусть X — дискретная случайная величина, неизвестный параметр $\vartheta \in R^1$, закон распределения
$P_\vartheta\{X = x\} = P_\vartheta(x)$. Пусть дана выборка $x_1^*, \ldots, x_n^*$. Запишем статистическое распределение выборки: различные значения x_i и соответствующие им частоты n_i, $i = 1, \ldots, k$.

x_1	x_2	$\ldots$	x_k
n_1	n_2	$\ldots$	n_k

Значение x_1 "выпадает" с вероятностью $P_\vartheta\{X = x_1\}$ (известной с точностью до параметра ϑ), причем n_1 раз, значит, вероятность этого события $P_\vartheta\{X = x_1\})\text{x}\ldots\text{x}P_\vartheta\{X = x_1\}) = (P_\vartheta\{X = x_1\})^{n_1}$. Аналогично для следующего значения x_2 и т.д.

Тогда запишем ВЕРОЯТНОСТЬ НАБЛЮДАТЬ НАБЛЮДАЕМОЕ!

$$L(x_1, x_2, \ldots, x_k; \vartheta) = (P_\vartheta\{X = x_1\})^{n_1}(P_\vartheta\{X = x_2\})^{n_2}\text{x}\ldots\text{x}(P_\vartheta(X = x_k))^{n_k}. \qquad (30)$$

L – *функция правдоподобия* – вероятность "наблюдать наблюдаемое".

Метод наибольшего правдоподобия основан на максимизации функции правдоподобия! Выражение для параметра, обеспечивающее максимум ф. правдоподобия, и является решением:

Если $\vartheta|\max L \Rightarrow \hat{\vartheta}$ — МНП-оценка.

Уравнение правдоподобия:

$$\frac{dL}{d\vartheta} = 0$$

Иногда заменяют $L \to \ln L$ - это *логарифмическая* функция правдоподобия (она достигает максимум при тех же значениях аргумента? поскольку логарифм – это монотонное преобразование).

Пример.

Пусть $X \sim \Pi(\lambda)$(имеем выборку из распределения Пуассона), параметр $\vartheta = \lambda$, известно статистическое распределение выборки $x_i \sim n_i$.

$$P_\vartheta(x) = P_\vartheta\{X = x\} = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

$$L = \left[\frac{\lambda^{x_1} e^{-\lambda}}{x_1!}\right]^{n_1} \cdot \ldots \cdot \left[\frac{\lambda^{x_k} e^{-\lambda}}{x_k!}\right]^{n_k} = \frac{\lambda^{\sum_{i=1}^n x_i n_i} \cdot e^{-\lambda n}}{\prod_{i=1}^k (x_i!)^{n_i}} \qquad (n = \sum_{i=1}^k n_i)$$

$\frac{\partial \ln L}{\partial \lambda} = \sum_i x_i n_i \frac{1}{\lambda} - n = 0 \Rightarrow \hat{\lambda} = \frac{\sum_i x_i n_i}{n} = \bar{X}$.

Мы получили, что методом наибольшего правдоподобия оценка для неизвестного параметра λ распределения Пуассона такая же. (Но это не всегда так).

Пусть X — абсолютно непрерывная случайная величина, т.е. существует плотность: $f(x, \vartheta) = f_\vartheta(x)$.

Поскольку плотность является аналогом вероятности для дискр.с.в., запишем функцию правдоподобия в следующем виде:

$$L(x_1, \ldots, x_n, \vartheta) = f_\vartheta(x_1) \cdot \ldots \cdot f_\vartheta(x_n).$$

Имеем уравнение правдоподобия:

$$\frac{\partial \ln L}{\partial \vartheta} = 0.$$

Далее действуем по описанному выше алгоритму.

Пример.

$X \sim Exp(\lambda)$ (экспоненциальный закон распределения)

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

$x_1, \ldots, x_n$ — выборка

$L = \lambda e^{-\lambda x_1} \cdot \ldots \cdot \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$

Возьмем логарифмическую функцию правдоподобия: $\ln L = n \cdot \ln \lambda - \lambda \sum_{i=n}^n x_i$

Берем производную и приравниваем к нулю: $\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}$

Получился такой же ответ, как и методом моментов.

Замечание. Пусть $\vartheta \in R^2$ (два параметра). Тогда $L(x_1, \ldots, x_n, \vartheta_1, \vartheta_2)$.

Запишем частные производные и приравняем их к нулю $\frac{\partial \ln L}{\partial \vartheta_1} = 0$,
$\frac{\partial \ln L}{\partial \vartheta_2} = 0$.

Замечание. Вообще говоря, полученные экстремальные точки необходимо исследовать далее (действительно ли в них достигается максимум, а не минимум и т.п.) Изучите знак второй производной функции правдоподобия для одномерного случая, а также условие максимизации L для случая многомерного параметра ϑ.

Задание 6.

a) Найти неизвестный параметр p геометрического распределения,

b) Найти неизвестный параметр p биномиального распределения (считая, что n известно) методом моментов и методом наибольшего правдоподобия.

Задание 7. В течение Второй мировой войны на южную часть Лондона упало 535 снарядов. Территория южного Лондона была разделена на 576 участков площадью 0.25 км^2. В следующей таблице приведены числа участков n_k, на каждый из которых упало k снарядов:

Число снарядов k	0	1	2	3	4	5	
Число участков n_k	229	211	93	35	7	1	$\sum\limits_{n_k} = n = 576$

Методом моментов и методом наибольшего правдоподобия оцените среднее число снарядов, упавших на участок 0.25 км^2, если принять, что с.в. X — число снарядов, упавших на один участок, распределена по закону Пуассона.

12 Основные распределение математической статистики

12.1 Распределения Гаусса, Пирсона, Стьюдента, Фишера

1. *Нормальное (гауссово) распределение* $X \sim N(a, \sigma^2)$

$$M(X) = a;\ D(X) = \sigma^2.$$

Плотность:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Функция распределения:

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt.$$

Стандартное нормальное распределение ($a = 0$, $\sigma = 1$). Тогда плотность $f(x)$ обозначается как $\phi(x)$,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Функция распределения $F(x)$ обозначается как $\Phi(x)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt = \Phi_0(x) + \frac{1}{2}, \tag{31}$$

где $\Phi_0(x)$— функция Лапласа, $\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{\frac{-t^2}{2}} dt$.

2. *Распределение* χ^2 *(Пирсона)*

Случайная величина $\chi^2 \sim \chi^2(n)$ (имеет распределение χ^2 с n степенями свободы), если она может быть представлена в следующем виде:

$$\chi^2 = \xi_1^2 + \ldots + \xi_n^2, \quad \xi_i \sim N(0,1). \tag{32}$$

(ξ_i, $i = 1, \ldots, n$ имеют (одинаковое) стандартное нормальное распределение, независимые с.в.

Плотность распределения:

$$f(y) = \begin{cases} 0 & y \leq 0 \\ \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{\frac{-y}{2}} y^{\frac{n}{2}-1} & y > 0 \end{cases}$$

Очевидно, что поскольку n — целое и является единственным параметром распределения, значения для плотности могут быть рассчитаны и заданы в виде таблицы.

Имеем, что $M(\chi^2) = n$; $D(\chi^2) = 2n$. Тогда по центральной предельной теореме, χ^2 можно считать примерно нормальной при $n \geq 30$, $\chi^2 \sim N(n, 2n)$.

3. *Распределение Стьюдента (t-распределение)*

Случайная величина T имеет распределение Стьюдента с n степенями свободы ($T \sim t(n)$), если она может быть представлена в следующем виде:

$$T = \frac{\eta}{\sqrt{\chi^2/n}}, \quad \eta \sim N(0,1), \quad \chi^2 \sim \chi^2(n), \quad \Rightarrow T \sim t(n). \tag{33}$$

Плотность распределения:

$$f(y) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} (1 + \frac{y^2}{n})^{-\frac{n+1}{2}}.$$

График плотности распределения Стьюдента напоминает форму колокола, как у стандартного нормального распределения, но он ниже и шире. При достаточно больших n распределение Стьюдента асимптотически сходится к стандартному нормальному: $T \sim N(0;1)$.

В настоящее время достаточно актуально в исследованиях с так называемыми “тяжелыми хвостами”.

4. *Распределение Фишера*

Случайная величина $F_{n,m}$ имеет распределение Фишера с с n, m степенями свободы, если она может быть представлена в следующем виде:

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}, \quad \chi_n^2 \sim \chi^2(n), \quad \chi_m^2 \sim \chi^2(m), \tag{34}$$

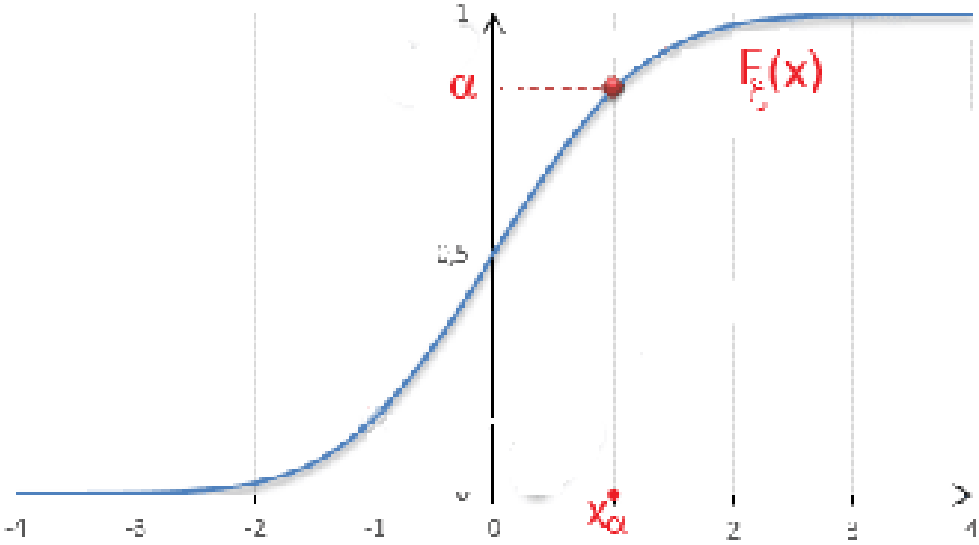
где χ_n^2, χ_m^2 — независимые случайные величины, распределенные по χ^2 с соответствующими степенями свободы.

12.2 Вычисление квантилей для некоторых распределений

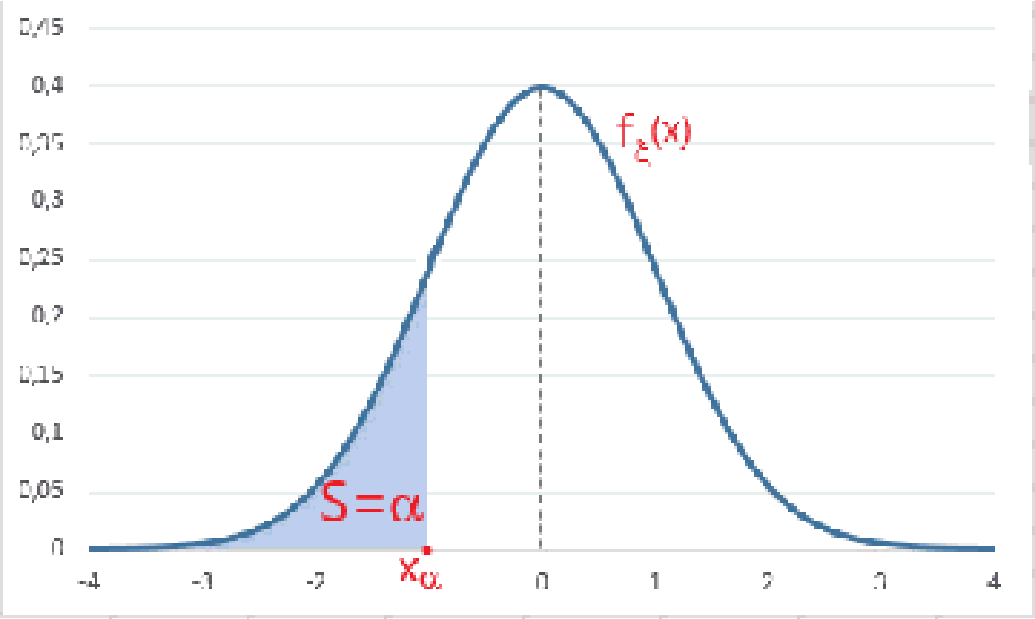
Пусть случайная величина ξ имеет функцию распределения $F_\xi(x)$, $X \in R$. Квантиль уровня α распределения $F_\xi(x)$ есть точка x_α, определяемая из условия

$$F_\xi(x_\alpha) = \alpha.$$

(или, что то же самое, $P\{\xi \le x_\alpha\} = \alpha$)



Посмотрим геометрический смысл квантиля на графике плотности распределения. Площадь под графиком плотности левее точки x_α равна α (это есть вероятность того, что сл.в. принимает значения меньшие x_α):


0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-4
-3
-2
0
1
2
3
4
$f_\xi(x)$
$S=\alpha$
x_α

Пример 1

Найдём квантиль уровня 0,05 равномерного на отрезке $[0, 4]$ распределения.

Функция распределения имеет вид:

$$F_\xi(x) = \begin{cases} 0, & \text{при } x < 0, \\ \frac{x}{4}, & \text{при } 0 \le x < 4, \\ 1, & \text{при } x \ge 4. \end{cases}$$

Квантиль уровня 0,05 будем искать как корень уравнения:

$$F_\xi(x) = 0,05.$$

$$\frac{x}{4} = 0,05 \Rightarrow x = 0,2.$$

Значит, $x_{0,05} = 0,2$ – квантиль уровня 0,05 данного распределения.

Квантиль стандартного нормального распределения уровня α будем обозначать через: $\Phi^{-1}(\alpha)$ или u_α.

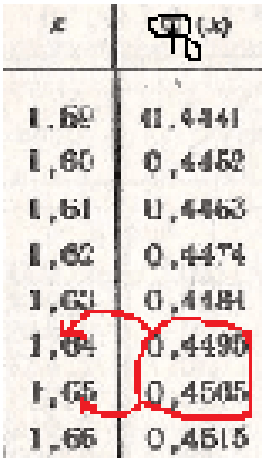
Пример 2 Найти квантиль стандартного нормального распределения уровня 0,05.

Для нахождения квантилей стандартного нормального распределения будем пользоваться таблицей значений функции Лапласа.

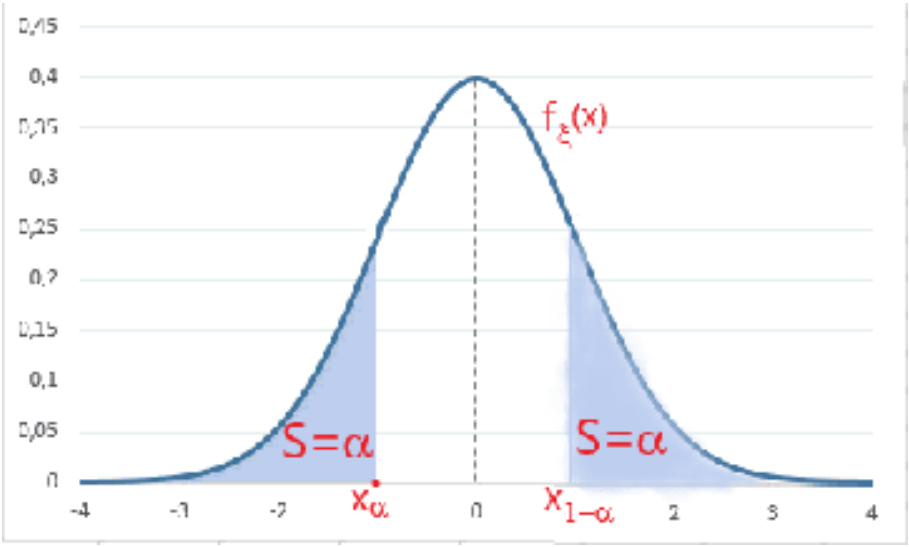
$$\Phi(u_{0,05}) = \frac{1}{2} + \Phi_0(u_{0,05}) = 0,05,$$

$$\Phi_0(u_{0,05}) = -0,45 \Rightarrow \Phi_0(-u_{0,05}) = 0,45$$

Находим в таблице значение, близкое к 0,45 и соответствующее ему значение x (см. рисунок) Получаем: $-u_{0,05} = 1,645$ (у нас значение между 1,64 и 1,65). Значит, $u_{0,05} = -1,645$ – квантиль уровня 0,05 стандартного нормального распределения.



Заметим, что стандартное нормальное распределение симметрично, значит $u_\alpha = -u_{1-\alpha}$. На картинке ниже это проиллюстрировано:



Обозначение: $\chi^2_\alpha(k)$ – квантиль уровня α распределения χ^2 с k степенями свободы.

Обозначение: $t^1_k(\alpha)$ – квантиль уровня α распределения Стьюдента с k степенями свободы.

Обозначение: $F(\alpha, k_1, k_2)$ – квантиль уровня α распределения Фишера с k_1 и k_2 степенями свободы.

13 Доверительные интервалы для неизвестных параметров распределения

13.1 Интервальное оценивание. Доверительный интервал

Пусть ϑ, $\vartheta \in R^1$ — неизвестный параметр распределения. (Например, вероятность успеха p для биномиального распределения, неизвестное мат.ожидание a для нормального распределения и т.д.)

Пусть задано некоторое число $0 < \gamma < 1$, называемое *коэффициентом доверия (надёжности)*. Если указаны две статистики (функции выборки) $\vartheta'(x_1, \ldots, x_n)$ и

$\vartheta''(x_1, \ldots, x_n)$, такие что

$$P\{\vartheta' < \vartheta < \vartheta''\} = \gamma,$$

то $(\vartheta', \vartheta'')$ называется *доверительным интервалом* для неизвестного параметра ϑ с коэффициентом доверия γ.

Как правило, коэффициент доверия выбирается близким к 1, например, $\gamma = 0,95; 0,97; 0,99$.

Отметим, что статистики $\vartheta'(x_1, \ldots, x_n)$ и $\vartheta''(x_1, \ldots, x_n)$ — случайные величины, и на различных выборках принимают различные значения!

Имеем, что с наперед заданной вероятностью γ неизвестный нам параметр ϑ будет находиться в указанном доверительном интервале $(\vartheta', \vartheta'')$. Такой способ оценивания неизвестных параметров называется *интервальным* оцениванием (до этого мы находили одно число по конкретной выборке, которым заменяли неизвестный параметр — точечное оценивание).

Вопрос. Почему нельзя взять $\gamma = 1$?

Если $\gamma = 0,95$, то говорят, что построен *95 %-ный* (читается "девяносто пяти процентный") доверительный интервал и т.п. В большинстве экономических, социологических и пр. приложений используется именно 95%-ный доверительный интервал.

Пусть $\hat{\vartheta}$ — точечная оценка неизвестного параметра ϑ (найденная по методу моментов, наибольшего правдоподобия или др.). Часто доверительный интервал ищется в следующем симметричном виде:

$$\vartheta' = \hat{\vartheta} - \delta; \;\; \vartheta'' = \hat{\vartheta} + \delta, \tag{35}$$

т.е.

$$P\{\hat{\vartheta} - \delta \leq \vartheta \leq \hat{\vartheta} + \delta\} = \gamma.$$

Тогда δ называется *точностью* оценки. Для построения доверительного интервала как раз и необходимо найти δ.

Очевидно, что при увеличении γ доверительный интервал становится шире, т.е. δ увеличивается! (Что на самом деле приводит к уменьшению точности в обычном

понимании, но в приведенных выше терминах точность увеличивается).

δ также называется *предельной ошибкой оценки.*

Иногда требуется найти минимальный объем выборки n, при котором при фиксированном коэффициенте доверия γ будет обеспечиваться некоторая фиксированная точность δ.

Как правило, доверительные интервалы ищут именно в описанном выше виде. Такие интервалы называют также двусторонними доверительными интервалами. Могут быть также рассмотрены односторонние доверительные интервалы, в которых одна из границ "обрезается" исходя из естественных требований конкретной задачи. Имеем:

$(\vartheta', \vartheta'')$– *двусторонний доверительный интервал.*

$(-\infty, \vartheta'')$ $((0, \vartheta''))$ – *нижний доверительный интервал.*

(ϑ'', ∞) $((\vartheta'', 1)$ – *верхний доверительный интервал.*

13.1.1 Построение доверительных интервалов в случае асимптотически нормальных оценок

Пусть ϑ — неизвестный параметр распределения с.в. X. Предположим, что получена точечная оценка $\hat{\vartheta}$, которая асимптотически нормальна, т.е.

$$\hat{\vartheta} \underset{n\to\infty}{\to} \vartheta^* \sim N(M(\hat{\vartheta}), D(\hat{\vartheta})).$$

Пусть $\hat{\vartheta}$ — также несмещенная оценка, т.е.

$$M(\hat{\vartheta}) = \vartheta.$$

Будем искать двусторонний доверительный интервал в симметричном относительно точечной оценки виде:

$$P\{\hat{\vartheta} - \delta \le \vartheta \le \hat{\vartheta} + \delta\} = \gamma. \tag{36}$$

Необходимо найти δ.

Перепишем (36) в виде

$$P\{|\hat{\vartheta} - \vartheta| < \delta\} = \gamma. \tag{37}$$

Из свойства несмещенности получаем из (37):

$$P\{|\hat{\vartheta} - M(\hat{\vartheta})| < \delta\} = \gamma. \tag{38}$$

Воспользуемся асимптотической нормальностью $\hat{\vartheta}$. Тогда для левой части формулы (38) при больших n можно использовать формулу (9) для вероятности попадания нормально распределенной с.в. в определенный интервал (который здесь также симметричен относительно матожидания):

$$P\{|\hat{\vartheta} - M(\hat{\vartheta})| < \delta\} = 2\Phi_0(\frac{\delta}{\sigma(\hat{\vartheta})}) = 2\Phi(\frac{\delta}{\sigma(\hat{\vartheta})}) - 1 = \gamma. \tag{39}$$

Из (39) получаем:

$$\Phi_0(\frac{\delta}{\sigma(\hat{\vartheta})}) = \gamma/2 \Rightarrow \frac{\delta}{\sigma(\hat{\vartheta})} = \Phi_0^{-1}(\gamma/2) \Rightarrow \delta = \Phi_0^{-1}(\gamma/2)\sigma(\hat{\vartheta}). \tag{40}$$

Таким образом, для асимптотически нормальных несмещенных оценок требуется найти

$$\sigma(\hat{\vartheta}) = \sqrt{D(\hat{\vartheta})}$$

и подставить в формулу

$$\delta = \Phi_0^{-1}(\gamma/2)\sigma(\hat{\vartheta}). \tag{41}$$

Если $\sigma(\hat{\vartheta})$ неизвестно, то вместо него берется состоятельная оценка $\bar{\sigma}(\hat{\vartheta})$, посчитанная по выборочным значениям.

13.2 Построение приближенных доверительных интервалов для неизвестного параметра p биномиального распределения

Задачу оценивания неизвестного параметра p (вероятность успеха) в схеме Бернулли также называют “оценкой для генеральной доли”.

Пусть имеется генеральная совокупность объема N. Требуется оценить долю p элементов этой совокупности, обладающих некоторым качественным признаком A. Предполагается, что p не слишком мало.

Производится выборка объема n. Рассмотрим случай, когда объем выборки достаточно велик ($n \geq 100$). В качестве точечной оценки для неизвестного параметра p возьмем частоту появления события A в выборке объема n, т.е. $\hat{p} = \frac{S_n}{n}$, где S_n — число появлений события A в выборке.

Тогда справедливо следующее утверждение.

Утверждение. С вероятностью γ неизвестная доля p находится в интервале:

$$(\hat{p} - \delta; \hat{p} + \delta), \tag{42}$$

где $\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, т.е.

$$P\{p \in (\hat{p} - \delta; \hat{p} + \delta)\} = \gamma.$$

Доказательство.

Пусть $n \to \infty$. Тогда согласно центральной предельной теореме с.в. $S_n \sim B(n;p)$ $(MS_n = np,\ DS_n = npq\ )$ может быть заменена на нормальную с.в. $S_n \sim N(np, npq)$.

Кроме того, мы знаем, что

$$M(S_n/n) = \frac{np}{n} = p,$$

т.е. оценка $\hat{p} = \frac{S_n}{n}$ является несмещенной оценкой для неизветного p.

Таким образом, $\frac{S_n}{n}$ — асимптотически нормальная, несмещенная оценка параметра p.

Воспользуемся формулой (41). Для этого предварительно вычислим

$$D(\hat{p}) = \frac{D(S_n)}{n^2} = \frac{pq}{n} = \frac{p(1-p)}{n}. \Rightarrow \sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

Поскольку значение p нам неизвестно, возьмем вместо него оценку $\hat{p}$.

Утверждение доказано. $\square$

Напомним, здесь $\Phi_0^{-1}(\gamma/2)$— значение, которое мы находим по таблице значений функции Лапласа. В отличие от задач по теории вероятности, где по значению аргумента (левый столбец) находилось значение функции (правый столбец), здесь, наоборот, по значению функции, равному $\gamma/2$, находится значение аргумента!

Например, если $\gamma = 0,9$, то $\Phi_0^{-1}(\gamma/2) = \Phi_0^{-1}(0,45) = 1,65$.

Если $\gamma = 0,95$, то $\Phi_0^{-1}(\gamma/2) = \Phi_0^{-1}(0,475) = 1,96$.

Напомним, что функция распределения стандартной нормальной случайной величины имеет вид

$$\Phi(x) = \Phi_0(x) + 1/2.$$

Тогда если

$$\Phi_0(x) = \alpha, \Rightarrow \Phi(x) = \alpha + 1/2, \Rightarrow \Phi_0^{-1}(\alpha) = \Phi^{-1}(\alpha + 1/2).$$

Пример 1. При проверке лекарственного препарата у 23 животных из 400 наблюдались побочные эффекты. Найти 95%-ный доверительный интервал для доли животных, дающих побочный эффект.

Решение. Событие A — появление побочного эффекта. Коэффициент доверия $\gamma = 0,95$. Объем выборки $n = 400$, число появлений события A в выборке $S_n = 23$. Тогда $\hat{p} = 23/400$; $\Phi_0^{-1}(\gamma/2) = \Phi_0^{-1}(0,475) = 1,96$. Подставляя полученные значения в (48), имеем: $0,0335 < p < 0,0815$.

Это означает, что с вероятностью 0.95 доля животных находится в указанных пределах, т.е. от 3 до 8 % животных имеют побочный эффект.

13.3 Построение доверительных интервалов для неизвестных параметров нормального распределения

Пусть $X \sim N(a, \sigma^2)$.

I. Рассмотрим случай, когда требуется найти доверительный интервал для a при *известном* σ.

Итак, неизвестный параметр $a = M(X)$, но откуда-то (из предварительной информации) известно σ.

Замечание. Либо σ неизвестно, но объем выборки n достаточно большой. Тогда в качестве σ можно использовать выборочную характеристику s (стандартное отклонение).

В качестве точечной оценки a возьмем выборочное среднее $\bar{x} = \frac{x_1+\ldots+x_n}{n}$, т.е. $\hat{a} = \bar{x}$.

Тогда справедливо следующее утверждение.

Утверждение. С вероятностью γ неизвестное математическое ожидание a находится в интервале:

$$(\bar{x} - \delta; \bar{x} + \delta), \tag{43}$$

где $\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\sigma^2}{n}}$, т.е.

$$P\{a \in (\bar{x} - \delta; \bar{x} + \delta)\} = \gamma.$$

Доказательство.

Заметим, что если $X \sim N(a, \sigma^2)$, то

$$M(\bar{X}) = a; \qquad D(\bar{X}) = \frac{\sigma^2}{n}; \qquad \sigma(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}.$$

причем $\bar{X}$ распределена нормально (доказательство см. аппарат характеристических функций).

Таким образом, точечная оценка $\bar{X}$ является несмещенной (асимптотически) нормальной с.в. Воспользуемся формулой (41).

Тогда

$$\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\sigma^2}{n}}.$$

□

Пример 2. Средняя жирность молока у коров некоторого региона неизвестна. 100 коров обследуются на жирность молока. По данным обследования, средняя жирность составила 3,64 (%) при известной дисперсии $2,56$. С вероятностью 0.95 определите предельные значения для средней жирности молока. Считаем, что жирность молока у коров распределена по нормальному закону.

Решение. $n = 100,\ \gamma = 0,95$. Неизвестный параметр $M(X) = a$. Известно: $\sigma = \sqrt{2.56} = 1,6;\ \ \bar{x} = 3,64$. Тогда доверительный интервал для неизвестного a при известном σ:

$$(3,64 - 1.96 * \frac{1,6}{10}; 3,64 + 1.96 * \frac{1,6}{10}) \Rightarrow a \in (3,3264; 3,6776).$$

Средняя жирность молока (%) с вероятностью 0.95 находится в этом интервале.

Примечание. Это именно то, что вы иногда можете увидеть на пакетах с молоком!

Вопрос. А как изменится доверительный интервал, если задать $\gamma = 0.9545$?

Пример 3. Рекламное агентство, обслуживающее крупную радиостанцию, хочет оценить среднее количество времени, которое аудитория радиостанции тратит на прослушивание радио ежедневно. Из прошлых исследований среднее квадратическое отклонение оценивается как 45 минут. Считаем, что время на прослушивание распределено по нормальному закону. Какой размер выборки необходим, если агентство хочет быть на 90 % уверенным в том, что оно будет правильным с точностью до 5 минут?

Решение. Необходимо оценить минимальный объем выборки n, такой что

$$P\{a \in (\bar{x} - \delta; \bar{x} + \delta)\} = \gamma;\ \ \delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\sigma^2}{n}}.$$

Здесь нам уже дано δ, по которому нужно найти n.

Имеем: $\sigma = 45,\ \delta = 5;\ \gamma = 0,9$. Тогда $\Phi_0^{-1}(\gamma/2) = \Phi_0^{-1}(0,45) = 1,65$.

Тогда $5 = 1,65 * 45/\sqrt{n}. \Rightarrow \sqrt{n} = 14,85 \Rightarrow n \approx 220$.

Замечание. Не всегда в задачах сразу дано выборочное среднее $\bar{x}$. Если дана выборка, посчитайте его сами.

II. Рассмотрим теперь случай построения интервала для a при *неизвестном* σ.

Предварительно сформулируем без доказательства следующую теорему.

Теорема (Фишера). Пусть выборка $\{x_1^*, \ldots, x_n^*\}$ получена из генеральной совокупности, подчиняющейся нормальному закону распределения $N(a, \sigma^2)$. Тогда справедливы следующие распределения:

1. статистика $\frac{\bar{x}-a}{\sigma}\sqrt{n}$ подчиняется стандартному нормальному распределению;
2. статистика $\frac{\bar{x}-a}{\sigma^*}\sqrt{n-1} = \frac{\bar{x}-a}{s}\sqrt{n}$ подчиняется распределению Стьюдента с $n-1$ степенями свободы;
3. статистика $\sum\limits_{i=1}^{n}\left(\frac{x_i-a}{\sigma}\right)^2 = \frac{n(\overline{x^2}-2a\bar{x}+a^2)}{\sigma^2}$ подчиняется χ^2 распределению с n степенями свободы;
4. статистика $\sum\limits_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 = \frac{n\sigma^{*2}}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$ подчиняется χ^2 распределению с $n-1$ степенью свободы.

Доказательство данных утверждений может быть произведено самостоятельно.

Замечание. Вывод доверительного интервала для неизвестного a при известном σ может быть получен из п.1. приведенной теоремы. Воспользуемся п.2 для вывода доверительного интервала для случая неизвестного σ.

Вычислим по выборке выборочное среднее $\bar{x}$ и выборочную дисперсию σ^{2*} (здесь имеется в виду выборочная дисперсия до ее "исправления", т.е. смещенная оценка $\sigma^{2*} = 1/n\sum(\ldots)$).

Будем искать доверительный интервал в следующем виде:

$$P\{\bar{x} - \delta_1 < a < \bar{x} + \delta_1\} = \gamma,$$

где

$$\delta_1 = \delta\sqrt{\sigma^{2*}}.$$

Тогда

$$P\{\bar{x} - \delta_1 < a < \bar{x} + \delta_1\} = P\{-\delta \le \frac{a-\bar{x}}{\sqrt{\sigma^{2*}}} \le \delta\} = \gamma.$$

Воспользуемся теоремой Фишера, п.2. Для этого умножим на $\sqrt{n-1}$ соответствующее выражение. Имеем

$$P\left\{-\delta\sqrt{n-1} < \frac{\bar{X}-a}{\sqrt{\sigma^{2*}}}\delta\sqrt{n-1} < \sqrt{n-1}\right\} = \gamma.$$

Обозначим как

$$T = \frac{\bar{X} - a}{\sqrt{\sigma^{2*}}}\delta\sqrt{n-1},$$

имея в виду, что T распределена по закону Стьюдента (t– распределение). Напомним, что график плотности t– распределения имеет симметричный относительно нуля вид (похож на кривую стандартной нормальной с.в. с более тяжелыми хвостами). Тогда:

$$P\{-\delta\sqrt{n-1} \le T \le \delta\sqrt{n-1}\} = \gamma$$

$$P\{T \le \delta\sqrt{n-1}\} - P\{T \le -\delta\sqrt{n-1}\} = \gamma$$

$$P\{T \le \delta\sqrt{n-1}\} = \frac{1+\gamma}{2}$$

$$F_T(\delta\sqrt{n-1}) = \frac{1+\gamma}{2},$$

где

$$\delta\sqrt{n-1} = t_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)$$

— квантиль распределения Стьюдента порядка $\frac{1+\gamma}{2}$ с числом степеней свободы $n-1$.

Это табличное значение! Берем его из таблицы значений распределения Стьюдента.

Имеем:

$$\delta = t_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{1}{\sqrt{n-1}}. \tag{44}$$

Фактически, мы доказали утверждение.

Утверждение. С вероятностью γ неизвестное математическое ожидание a при неизвестном σ находится в интервале:

$$(\bar{x} - \delta_1; \bar{x} + \delta_1), \tag{45}$$

где $\delta_1 = t_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{\sigma^{2*}}{n-1}}$, т.е.

$$P\{a \in (\bar{x} - \delta_1; \bar{x} + \delta_1)\} = \gamma.$$

Замечание. Иногда вместо $\frac{\sigma^{2*}}{n-1}$ используют $\frac{s}{n}$, имея в виду под s исправленную выборочную дисперсию.

Пример 4. Суточный расход авиационного топлива (т) по данным 10 дней составил: 220, 200, 240, 190, 160, 260, 210, 200, 170, 150 т. Для коэффициента доверия 0.9 построить

доверительный интервал для среднего расхода авиационного топлива в сутки, считая, что суточный расход имеет нормальное распределение.

Решение. Имеем $\gamma = 0,9$. Нужно построить доверительный интервал для a при неизвестном σ. Воспользуемся формулой (49).

Вычислим по выборке мат.ожидание и дисперсию:

$$\bar{x} = 200;\ \sigma^{2*} = 1080$$

Учитывая, что $n = 10$, $\frac{1+\gamma}{2} = 0.95$, по таблице значений распределения Стьюдента находим $t_9^{-1}(0.95) = 1,833$. Тогда доверительный интервал имеет вид:

$$(200 - 1,833\sqrt{\frac{1080}{10-1}}; 200 + 1,833\sqrt{\frac{1080}{10-1}}) = (179,92; 220,08).$$

III. Доверительный интервал для неизвестной дисперсии σ^2 .

Посчитаем выборочную дисперсию σ^{2*}. Ищем доверительный интервал в следующем виде:

$$P\{\vartheta' \le \sigma^2 \le \vartheta''\} = \gamma,$$

$\vartheta' = \delta_1\sigma^{2*}, \quad \vartheta'' = \delta_2\sigma^{2*}$.

Воспользуемся п.4 теоремы Фишера. Тогда

$$P\{\delta_1\sigma^{2*} \le \sigma^2 \le \delta_2\sigma^{2*}\} = P\{\frac{n}{\delta_2} < \frac{n\sigma^{2*}}{\sigma^2} < \frac{n}{\delta_1}\} = P\{\frac{n}{\delta_2} < \chi^2 < \frac{n}{\delta_1}\},$$

где

$$\chi^2 = \frac{n\sigma^{2*}}{\sigma^2},$$

поскольку имеет распределение Пирсона χ^2 с $n-1$ степенью свободы.

Проблема в том, что распределение Пирсона не является симметричным и из выражения

$$P\{\frac{n}{\delta_2} < \chi^2 < \frac{n}{\delta_1}\} = \gamma$$

нельзя однозначно определить оба значения $\delta_1,\ \delta_2$.

Положим, что вероятности выхода за пределы интервала одинаковы, тогда

$$P\{\chi^2 < \frac{n}{\delta_2}\} = \frac{1-\gamma}{2},$$

$$P\{\chi^2 > \frac{n}{\delta_1}\} = \frac{1-\gamma}{2}, \Rightarrow P\{\chi^2 \le \frac{n}{\delta_1}\} = \frac{1+\gamma}{2}.$$

Используя понятие квантилей распределения Пирсона, получаем:

$$\delta_1 = \frac{n}{\chi^2_{\frac{1+\gamma}{2}}(n-1)},\ \delta_2 = \frac{n}{\chi^2_{\frac{1-\gamma}{2}}(n-1)},$$

где $\chi^2_{\frac{1+\gamma}{2}}(n-1),\ \chi^2_{\frac{1-\gamma}{2}}(n-1)$ — квантили распределения хи-квадрат соответствующих порядков с числом степеней свободы $n-1$. Это табличные значения!

Пример 5. Шесть стограммовых баночек кофе были открыты, их содержимое тщательно взвешено. Получены следующие значения: 105, 99, 94, 102, 91, 104. Найдите 90 %-ный доверительный интервал для среднего квадратического отклонения σ, предполагая, что вес распределен по нормальному закону.

Решение. Посчитаем выборочную дисперсию σ^{2*}. Ищем доверительный интервал в следующем виде:

$$P\{\vartheta' \le \sigma^2 \le \vartheta''\} = \gamma,$$

$\vartheta' = \delta_1\sigma^{2*},\ \ \vartheta'' = \delta_2\sigma^{2*}$.

Имеем :

$$\delta_1 = \frac{n}{\chi^2_{\frac{1+\gamma}{2}}(n-1)},\ \delta_2 = \frac{n}{\chi^2_{\frac{1-\gamma}{2}}(n-1)},$$

где $\chi^2_{\frac{1+\gamma}{2}}(n-1),\ \chi^2_{\frac{1-\gamma}{2}}(n-1)$ — квантили распределения хи-квадрат соответствующих порядков с числом степеней свободы $n-1$. Это табличные значения!

В данной задаче

$$\sigma^2 \in (\frac{158.83}{11.07}; \frac{158.83}{1.07}) = (14.347; 138.66).$$

Тогда

$$\sigma \in (\sqrt{14.347}; \sqrt{138.66}) = (3.79; 11.77).$$

13.4 Задачи

Задача 1. Пусть p — доля спортсменов, которые получали травмы на тренировках в течение последнего года. Из 330 спортсменов, участников опроса, 167 спортсменов указали, что они получали такие травмы. Постройте 90%-ный доверительный интервал для доли спортсменов p, получивших травмы.

Решение.

$n = 330,\ \gamma = 0,9,\ S_n = 167,$ тогда $\hat{p} = 167/330 = 0,506$. Квантиль $\Phi_0^{-1}(\gamma/2) = 1,65$.

С вероятностью $\gamma = 0,9$ неизвестная доля p находится в интервале:

$$(\hat{p} - \delta; \hat{p} + \delta), \tag{46}$$

где

$$\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0,045.$$

Доверительный интервал:

$$(167/330 - 0,045; 167/330 + 0,045) = (0,46; 0,55).$$

Задача 2. Фирма-производитель зубной пасты занимала 23 % рынка. Для того чтобы увеличить свою долю рынка, фирма наняла рекламную компанию. По истечению времени эта компания сообщила, что после ее усилий из выборки 1000 покупателей 28 % выбрали пасту этой фирмы. Свидетельствует ли это о том, что реклама была эффективной?

Указание. Постройте доверительный интервал для доли рынка для уровня доверия 0.95. Если нижняя граница превосходит 0.23, то реклама была эффективной.

Решение. $n = 1000,\ \gamma = 0,95,\ \hat{p} = 0,28$. Квантиль $\Phi_0^{-1}(\gamma/2) = 1,96$.

С вероятностью $\gamma = 0,95$ неизвестная доля p находится в интервале:

$$(\hat{p} - \delta; \hat{p} + \delta), \tag{47}$$

где

$$\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0,0278.$$

Доверительный интервал:

$$(0,28 - 0,0278; 0,28 + 0,0278) \approx (0,252; 0,308).$$

Нижняя граница находится правее 0.23. Значит, реклама была эффективной.

Задание 3. По данным департамента здравоохранения, только 25 % пациентов, пользующихся тростью, используют трость правильной длины. Какого размера надо взять выборку, чтобы оценить долю пациентов, пользующихся тростью правильной длины, с ошибкой не более 0.04 и уровнем доверия 0.95? Решение. $n = ?,\ \gamma = 0,95,\ \hat{p} = 0,25.$

Квантиль $\Phi_0^{-1}(\gamma/2) = 1,96, \delta < 0,04$

$$\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 0,04.$$

Тогда

$$1,96\sqrt{\frac{0,25 * 0,75}{n}} = 1,96\sqrt{\frac{3}{16n}} < 0,04. \Rightarrow n > 450.$$

Ответ. $n \geq 451$.

Задача 4. Из генеральной совокупности супружеских пар была сделана случайная выборка размера 400. Пусть

- в выборке 20 пар, в которых жена выше мужа, и 380 пар, в которых муж выше жены;
- рост женатого мужчины имеет нормальное распределение со средним 70 дюймов и средним квадратическим отклонением 3 дюйма;
- рост замужней женщины имеет нормальное распределение со средним 65 дюймов и средним квадратическим отклонением 2.5 дюйма.

Постройте доверительные интервалы с коэффициентом доверия $\gamma = 0,95$ для

a) доли пар, в которых жена выше мужа;

b) среднего роста мужчины;

c) среднего роста женщины.

Решение. $n = 400,\ \gamma = 0,95,$ Квантиль $\Phi_0^{-1}(\gamma/2) = 1,96.$

a)

$$\hat{p} = 20/400 = 0,05.$$

С вероятностью $\gamma = 0,95$ неизвестная доля p находится в интервале:

$$(\hat{p} - \delta; \hat{p} + \delta), \tag{48}$$

где

$$\delta = \Phi_0^{-1}(\gamma/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0,0214.$$

Доверительный интервал:

$$(0,05 - 0,0214; 0,05 + 0,0214) \approx (0,0286; 0,0714).$$

С вероятностью 0.95 от 2.86 до 7.14 процентов пар обладают заданным в условии свойством.

b) Необходимо построить доверительный интервал для неизвестного a при известном σ.

$n = 400,\ \gamma = 0,95$. Неизвестный параметр $M(X) = a$. Известно: $\sigma = 3;\ \bar{x} = 70$. Тогда доверительный интервал для неизвестного a при известном σ:

$$(70 - 1.96 * \frac{3}{20}; 70 + 1.96 * \frac{3}{20}) \Rightarrow a \in (69,7; 70,3).$$

Средний рост мужчины с с вероятностью 0.95 находится в этом интервале.

c) Аналогично b).

Средний рост женщины с с вероятностью 0.95 находится в интервале:

$$(64,76; 65,25)$$

Задача 5. На контрольных испытаниях $n = 15$ ламп была определена средняя продолжительность горения лампы, $\bar{X} = 3000$ ч. Считая, что срок службы лампы распределен нормально с $\sigma = 16$ ч., определить уровень доверия доверительного интервала для генеральной средней, если известно, что точность оценки средней равна 10 ч.

Решение

Точность задаётся формулой $\delta = \Phi_0^{-1}(\frac{\gamma}{2})\frac{\sigma}{\sqrt{n}}$.

Откуда $\Phi_0^{-1}(\frac{\gamma}{2}) = 2,42$.

Значит $\frac{\gamma}{2} = 0,4922$.

Тогда уровень доверия: $\gamma = 0,9844$.

Задача 6. Разработан новый тест IQ. Для определения среднего времени его выполнения студентами случайным образом выбран 61 студент из различных вузов России. Результаты приведены в таблице:

Время выполнения (мин)	[25; 30)	[30; 35)	[35; 40)	[40; 45)	[45; 50)	[50; 55]
Число студентов	1	4	16	26	10	4

Известно, что время выполнения теста распределено нормально. С вероятностью 0.9 оцените среднее время выполнения теста.

Решение.

$$(\bar{x} - \delta; \bar{x} + \delta), \tag{49}$$

где $\delta = t_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\sqrt{\frac{\sigma^{2*}}{n-1}}$

Получаем:

$$\bar{x} = 41,76 \qquad \sigma^{2*} \approx 5,19^2, \qquad t_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right) = 1,671.$$

Окончательно

$$a \in (40,64; 42,88).$$

Задача 7. При анализе точности фасовочного автомата было проведено $n = 24$ независимых контрольных взвешивания пятисотграммовых пачек кофе. Известно, что фасовочный автомат отрегулирован без смещения, так что его ошибка подчиняется $(0; \sigma^2)$-нормальному закону распределения, однако значение параметра σ^2 неизвестно. По результатам контрольных взвешиваний была рассчитана выборочная дисперсия $\sigma^{2*} = 0,64$ (г^2). Оценить точность работы фасовочного автомата, т.е. построить интервальную оценку для его среднеквадратической ошибки σ с уровнем доверия $0,95$.

Решение

По условию $n = 24$, $\sigma^{2*} = 0,64$, $\gamma = 0,95$.

Найдём квантили распределения χ^2 с 23 степенями свободы уровней $\frac{1+\gamma}{2} = 0,975$ и $\frac{1-\gamma}{2} = 0,025$. Пользуемся таблицей квантилей распределения χ^2: $\chi^2_{0,025;23} = 11,689$, $\chi^2_{0,975;23} = 38,076$.

Найдём границы интервала:

$$\frac{n\sigma^{2*}}{\chi^2_{1-\frac{\varepsilon}{2},n-1}} = \frac{24 \cdot 0,64}{38,076} = 0,403,$$

$$\frac{n\sigma^{2*}}{\chi^2_{\frac{\varepsilon}{2},n-1}} = \frac{24 \cdot 0,64}{11,689} = 1,314.$$

С вероятностью $0,95$ σ принадлежит интервалу $(0,634,\ 1,146)$.

14 Проверка статистических гипотез

Статистическая гипотеза - любое предположение о виде или о параметрах распределения случайной величины X.

Простая гипотеза – гипотеза, которая однозначно определяет распределение случайной величины X, если она верна. В остальных случаях гипотеза называется *сложной*.

Сначала формулируется *нулевая* (основная) гипотеза $H_0 : F(x) = F_0(x)$. Затем выдвигается *альтернативная* (конкурирующая) гипотеза $H_1 : F(x) \neq F_0(x)$. Если основная гипотеза отвергается, то автоматически принимается альтернативная.

Замечание. В качестве нулевой гипотезы H_0 выбирается та гипотеза, неверное отклонение от которой приводит к более опасным последствиям.

Например, H_0: человек здоров, H_1: человек болен. Если гипотеза H_0 верна (человек здоров), но по ошибке эту гипотезу отвергли, то принята гипотеза H_1 — человек болен. Таким образом, имеем ложноположительный результат: здорового человека признали больным. О вероятности совершить такую ошибку см. ниже.

Пусть H_0: человек болен, H_1: человек здоров. Если гипотеза H_0 верна (человек болен), но по ошибке эту гипотезу отвергли, то принята гипотеза H_1 — человек здоров. Таким образом, имеем ложнотрицательный результат: больного человека признали здоровым.

В данной задаче лучше выбирать второй вариант формулировки H_0, H_1.

Если преподаватель проверяет гипотезу “студент знает 20 из 25 вопросов” на основе заданных трех вопросов, то H_0, H_1 выбираются исходя из того, что преподаватель считает страшнее: поставить хорошую оценку тому, что не знает, или плохую оценку студенту, который знает 20 из 25.

Очевидно, что при проверке ответов студента на экзамене, может оказаться, что студент выучил 20 из 25 вопросов, но ему попались 3 вопроса, которые он не знает. Либо, наоборот, он мог не выучить 20 из 25 (т.е. он знает меньше необходимого минимума), но ему попались билеты, которые он знает.

При принятии решения мы можем совершить два вида ошибок.

2 вида ошибок:

I рода – отвергнуть нулевую гипотезу при её истинности;

II рода — принять нулевую гипотезу при её ложности.

Введем следующие обозначения.

$\alpha = P\{\bar{H}_0|H_0\}$ – *уровень значимости* – вероятность совершить ошибку I рода.

$\beta = P\{H_0|\bar{H}_0\}$ – вероятность совершить ошибку II рода.

$(1-\beta)$ – *мощность критерия.*

Уровень значимости α фиксируется, а мощность критерия максимизируется $(1-\beta)$.

В задаче про принятие решения по поводу пациента (болен или здоров) имеем: H_0: человек болен, H_1: человек здоров. Если гипотеза H_0 верна (человек болен), но по ошибке эту гипотезу отвергли, то принята гипотеза H_1 — человек здоров. Таким образом, имеем ложнотрицательный результат: больного человека признали здоровым. Это и есть ошибка первого рода. Её фиксируют заранее, делают очень маленькой: $\alpha = 0,01; 0,05, \ldots$. Вероятность признать человека больным, при условии, что он здоров - – это вероятность ложноположительного результата. Он не так опасен, и метод проверки гипотезы (согласно *методологии Неймана-Пирсона*) выбирается таким, чтобы вероятность совершить ошибку II рода была минимальной.

Как же математически проверить справедливость гипотезы H_0?

Выбирается *вспомогательная* случайная величина $R = R(x_1^*, \ldots, x_n^*)$ — *критерий проверки гипотезы, то же — статистика критерия* H_0.

Замечание. Статистика критерия R выбирается исходя из *методологии Неймана–Пирсона.* Метод основан на *лемме Неймана-Пирсона.* Необходимо выписать функции правдоподобия для случая выполнения гипотез H_0, H_1 и найти их отношение (отношение правдоподобия). Отношение правдоподобия – это отношение вероятности осуществления альтернативной гипотезы, деленное на вероятность осуществления нулевой гипотезы. Если результаты теста очень ожидаемые, если нулевая гипотеза верна по сравнению с альтернативной, отношение правдоподобия должно быть небольшим. К сожалению, мы пропускаем эту часть курса.

Область возможных значений R разбивается на две области:

$\mathbb{R}_{\text{ОПГ}}$ — *область принятия гипотезы.*

$\mathbb{R}_{\text{КР}}$ — *критическая* область.

В критической области $\mathbb{R}_{\text{КР}}$ нулевая гипотеза *критикуется*, т.е. отвергается.

Итак, если

$R_{\text{набл}} \in \mathbb{R}_{\text{ОПГ}} \Rightarrow H_0$ (гипотеза принимается).

$R_{\text{набл}} \in \mathbb{R}_{\text{КР}} \Rightarrow \bar{H}_0$ (гипотеза отвергается). $R_{\text{набл}}$ — это наблюдаемое значение критерия,

т.е. значение случайной величины R на конкретной выборке $x_1^*, \ldots, x_n^*$. Далее индекс "набл" будем опускать.

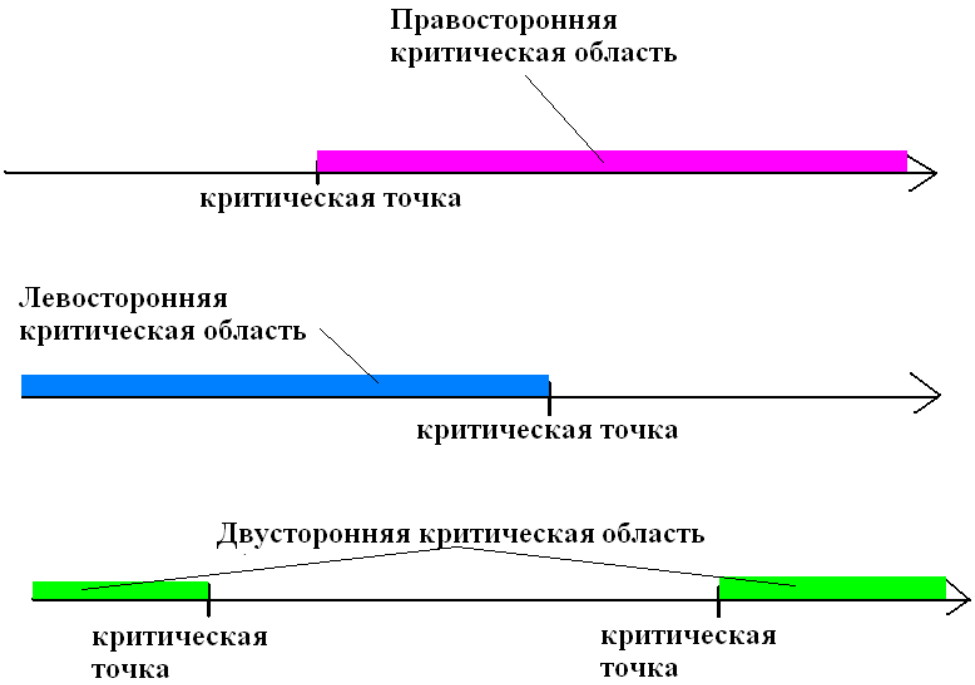
Вспомним, что такое вероятности совершить ошибку I и II рода. Сопоставим их с только что предложенным алгоритмом проверки.

$\alpha = P\{\bar{H}_0|H_0\} = P\{R \in \mathbb{R}_{\text{КР}}|H_0\}$ – *уровень значимости* – вероятность отвергнуть нулевую гипотезу при её истинности (совершить ошибку I рода).

$\beta = P\{H_0|\bar{H}_0\}$ – вероятность принять нулевую гипотезу при её ложности.

$(1-\beta)$ – *мощность критерия* (совершить ошибку II рода).

3 возможных типа расположения критической области:



1. *Правосторонняя:* критическая область правее области принятия гипотезы.

   Критическая точка $R_{\text{КР}}$ — значение на границе критической области и области принятия гипотезы. Если $R \geq R_{\text{КР}} \Rightarrow \bar{H}_0$. (нулевая гипотеза отвергается)

2. *Левосторонняя:* критическая область левее области принятия гипотезы.

   Если $R \leq R_{\text{КР}} \Rightarrow \bar{H}_0$.

3. *Двусторонняя:* критическая область и слева, и справа.

   Если $\left[\begin{array}{l} R \leq R_{\text{КР}_1} \\ R \geq R_{\text{КР}_2} \end{array}\right. \Rightarrow \bar{H}_0$.

Для двусторонней критической области имеем две критические точки $R_{\text{КР}_1}$, $R_{\text{КР}_2}$.

Уровень значимости α (вероятность совершить ошибку I рода) для разных видов критических областей:

1. Для правосторонней $\alpha = P\{R \geq R_{\text{КР}}|H_0\}$
2. Для левосторонней $\alpha = P\{R \leq R_{\text{КР}}|H_0\}$
3. Для двусторонней $1 - \alpha = P\{R_{\text{КР}_1} < R < \mathbb{R}_{\text{КР}_2}\}$ $(\mathbb{R}_{\text{КР}_1}, R_{\text{КР}_2})$ — доверительный интервал для R с коэффициентом доверия $1 - \alpha$.

14.1 Общий алгоритм проверки статистических гипотез

Общий алгоритм проверки статистических гипотез:

1. Формулируются H_0 и H_1
2. Задаётся уровень значимости α
3. Выбирается критерий R
4. Задаются $\mathbb{R}_{\text{ОПГ}}$ и $\mathbb{R}_{\text{КР}}$
5. Вычисляется $R_{\text{набл}}$ по выборке
6. Если $\left[\begin{array}{l} R_{\text{набл}} \in \mathbb{R}_{\text{ОПГ}} \Rightarrow H_0 \\ R_{\text{набл}} \in \mathbb{R}_{\text{КР}} \Rightarrow \bar{H_0} \end{array}\right.$ совершаем ошибку I рода с вероятностью α

Критерий согласия – $H_0 : F(x) = F_0(x); H_1 : F(x) \neq F(x)$ – проверка соответствия (согласия) статистического распределения теоретическому (с известной функцией распределения). *Критерий значимости* — проверка гипотезы о значении параметров распределения с.в.

14.2 Сравнение эмпирических и теоретических частот

14.2.1 Дискретные случайные величины

Пример 1. Полицейское исследование 840 ограблений показало следующее распределение числа ограблений по дням недели.

Можно ли утверждать, что ограбления происходят с равной вероятностью каждый день?

День недели k	1	2	3	4	5	6	7
Число ограблений n_k	105	121	129	115	110	134	126

Запишем имеющееся статистическое распределение выборки в следующем виде.

День недели k	1	2	3	4	5	6	7	
Эмпирическая частота n_k	105	121	129	115	110	134	126	$\sum n_k = n = 840$
Эмпир. относит. частота n_k/n	105/840	121/840	129/840	115/840	110/840	134/840	126/840	$\sum n_k/n = 1$

Пусть X — число ограблений в день. Поскольку в неделе 7 дней, то если ограбления происходят с равной вероятностью каждый день, имеем *теоретические вероятности*:

$$P\{X = 1\} = P\{X = 2\} = \ldots = P\{X = 7\} = 1/7.$$

Поскольку при справедливости нашего предположения о равной вероятности преступлений по дня недели по закону больших чисел имеем

$$n_k/n \approx p_k = 1/7,$$

получаем, что *теоретические частоты* могут быть вычислены по формуле

$$n'_k = np_k$$

Здесь

$$n'_k = 840 * 1/7 = 120.$$

Теоретические вероятности частоты иногда также называют ожидаемыми.

Имеем

Сведем данные с эмпирическими и теоретическими частотами в одну таблицу.

Рассогласование теоретических и эмпирических частот можно представить графически. Аналогично можно поступить с вероятностями и относительными частотами. Обычно выбирается один из видов графиков, т.к. они отличаются только масштабом.

День недели k	1	2	3	4	5	6	7	
Теоретическая вероятность p_k	1/7	1/7	1/7	1/7	1/7	1/7	1/7	$\sum p_k = 1$
Теоретическая частота n'_k	120	120	120	120	120	120	120	$\sum n'_k = n = 840$

День недели k	1	2	3	4	5	6	7	
Эмпирическая частота n_k	105	121	129	115	110	134	126	$\sum n_k = n = 840$
Теоретическая частота n'_k	120	120	120	120	120	120	120	$\sum n'_k = n = 840$

Такая процедура также называется *выравниванием* эмпирических и теоретических частот.

Далее при помощи специального *критерия* мы проверим утверждение о том, что вероятности в исходном распределении равны.

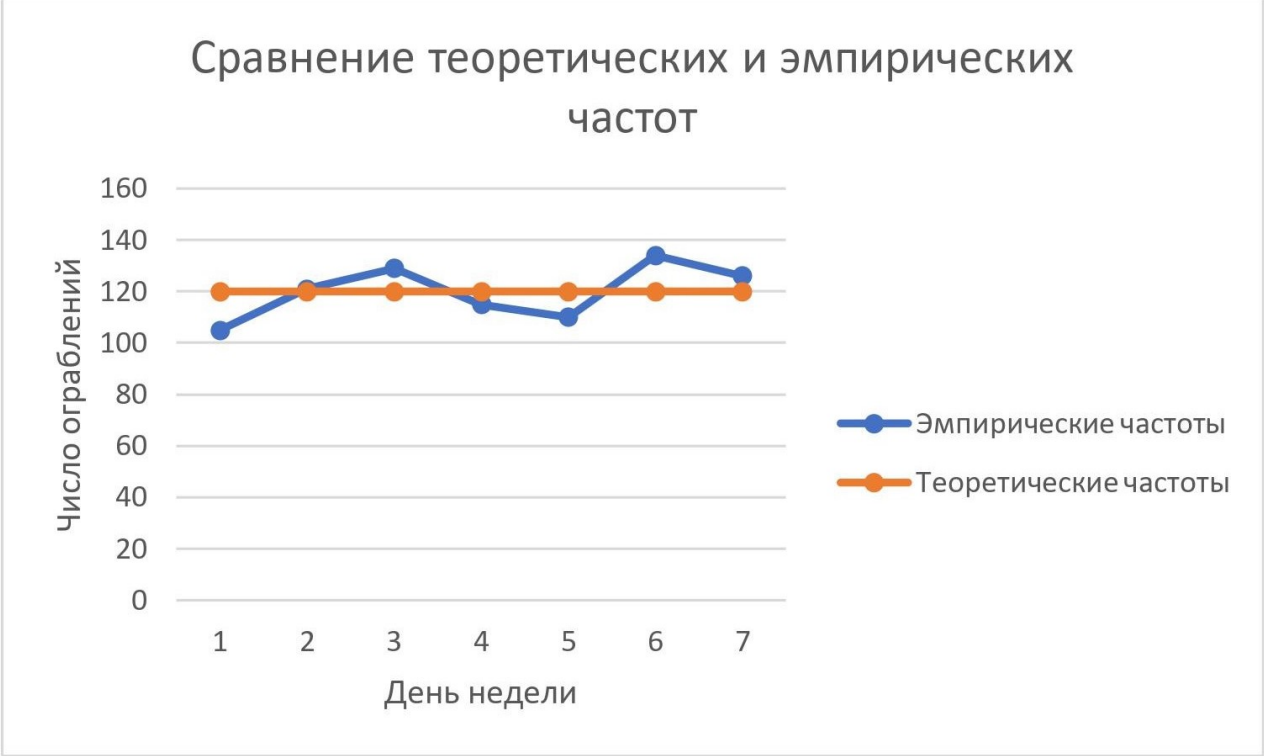
Пример 2. Рассмотрим задачу из § **??**. В течение Второй мировой войны на южную часть Лондона упало 535 снарядов. Территория южного Лондона была разделена на 576 участков площадью 0.25 км^2. В следующей таблице приведены числа участков n_k, на каждый из которых упало k снарядов:

Число снарядов k	0	1	2	3	4	5	
Число участков n_k	229	211	93	35	7	1	$\sum\limits_{n_k} = n = 576$

Проверим предположени о том, что с.в. X — число снарядов, упавших на один участок, — распределена по закону Пуассона.

Предварительно произведем следующий наглядный анализ. Получим выражения для *теоретических* (иногда также они называются ожидаемыми) частот, т.е. тех частот, которые должны были бы быть, если бы распределение соответствовало закону Пуассона.

Сначала запишем таблицу для *эмпирических* (наблюдаемых) частот из условия задачи в следующем виде:



Теоретические вероятности находим по формуле Пуассона:

$$p_k = P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}. \tag{50}$$

В формуле (50) в качестве λ возьмем точечную оценку $\hat{\lambda}$, посчитанную по выборке методом моментов, методом наибольшего правдоподобия или др. методом.

Ранее для данной задачи было получено: $\hat{\lambda} = \bar{x} \approx 0,93$.

Замечание. Закон больших чисел говорит нам, что относительные эмпирические частоты n_k/n должны быть примерно равны теоретическим вероятностям p_k, если наше предположение о виде распределения было верным:

$$n_k/n \approx p_k.$$

Умножим вероятности, полученные по формуле (50) на объем выборки n. Получаем так называемые *теоретические частоты*:

$$n'_k = p_k * n. \tag{51}$$

Сведем все полученные результаты (с округление) в таблицу для теоретических частот и вероятностей:

Наблюдаемое значение k случайной величины X	0	1	2	3	4	5	
Наблюдаемые частоты n_k	229	211	93	35	7	1	$\sum\limits_k n_k = n = 576$
Наблюдаемые относительные частоты n_k/n	$\frac{229}{576} \approx 0,3976$	$\frac{211}{576} \approx 0,3663$	$\frac{93}{576} \approx 0,1615$	$\frac{35}{576} \approx 0,0608$	$\frac{7}{576} \approx 0,01215$	$\frac{1}{576} \approx 0,0017$	$\sum\limits_k n_k/n = 1$

Значение случайной величины $X = k$	0	1	2	3	4	5	$[\mathbf{6};\infty)$	
Теоретические вероятности p_k	0,3950	0,3669	0,17039	0,05276	0,01225	0,0028	**0,00040**	$\sum\limits_k p_k = 1$
Теоретические частоты $n'_k = p_k * n$	227,5314	211,3356	98,1463	30,3867	7,0560	1,3107	**0,2333**	$\sum\limits_k n'_k = n = 576$

Заметим, что в таблице появилась еще одна колонка для значений случайной величины X больших или равных 6. Это необходимо сделать, поскольку *теоретически* с.в.X может принимать любые целочисленные значения ≥ 0, а не только $k = 0, \ldots, 5$. Если посчитать сумму теоретических вероятностей $p_0 + \ldots + p_5$, то она не будет равна 1, а чуть меньше. Вот этот "остаток" мы и запишем для теоретических вероятностей $X \in [6;\infty)$. Таким образом, имеем

$$p_6 = 1 - \sum_{k=0}^{k} p_k.$$

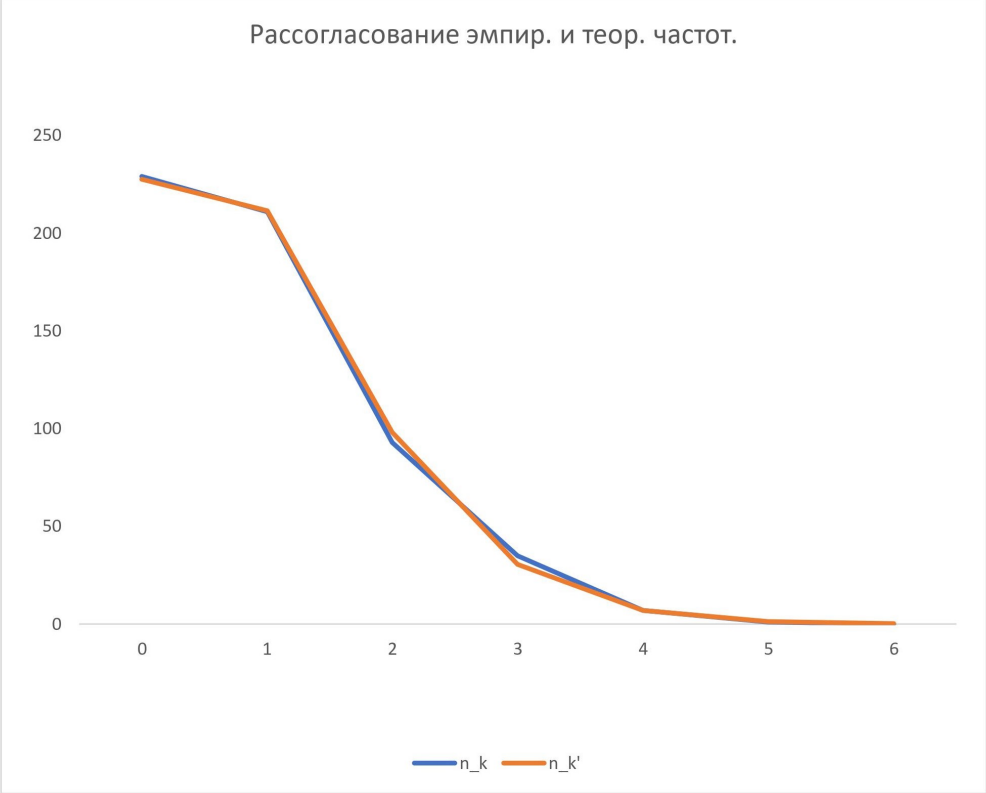
Сведем полученные результаты для наглядной демонстрации сравнения эмпирических и теоретических частот и вероятностей.

Значение случайной величины $X = k$	0	1	2	3	4	5	$[\mathbf{6}; \infty)$	
Наблюдаемые частоты n_k	229	211	93	35	7	1	**0**	$\sum\limits_k n_k = n = 576$
Теоретические частоты $n'_k = p_k * n$	227,5314	211,3356	98,1463	30,3867	7,0560	1,3107	**0,2333**	$\sum\limits_k n'_k = n = 576$

Значение случайной величины $X = k$	0	1	2	3	4	5	$[\mathbf{6}; \infty)$	
Наблюдаемые относительные частоты n_k/n	$\frac{229}{576} \approx 0,3976$	$\frac{211}{576} \approx 0,3663$	$\frac{93}{576} \approx 0,1615$	$\frac{35}{576} \approx 0,0608$	$\frac{7}{576} \approx 0,01215$	$\frac{1}{576} \approx 0,0017$	0	$\sum\limits_k n_k/n = 1$
Теоретические вероятности p_k	0,3950	0,3669	0,17039	0,05276	0,01225	0,0028	**0,0004**	$\sum\limits_k p_k = 1$

Рассогласование теоретических и эмпирических частот можно представить графически. Аналогично можно поступить с вероятностями и относительными частотами. Обычно выбирается один из видов графиков, т.к. они отличаются только масштабом.

Такая процедура также называется *выравниванием* эмпирических и теоретических частот.



Мы видим, что эмпирические и теоретические частоты очень близки (*согласованы*), и, скорее всего, наше предположение о том, что с.в. имеет распределение Пуассона, верное. Как проверить это при помощи *критерия согласия* см. далее.

14.2.2 Непрерывные случайные величины

Рассмотрим другой пример (см.), в котором случайная величина X имеет непрерывное распределение. Это значит, что выборка должна быть представлена в виде группировочного ряда с указанием интервалов $[z_0, z_1]$, $(z_1, z_2]$, $\dots (z_{r-1}, z_r]$, в которые попадает случайная величина.

При подсчете теоретических вероятностей попадания случайной величины в интервал $(z_i; z_{i+1}]$ воспользуемся формулой, использующей *теоретическую* функцию распределения (т.е. ту, которая была бы, если бы наше предположение о виде распределения было верным):

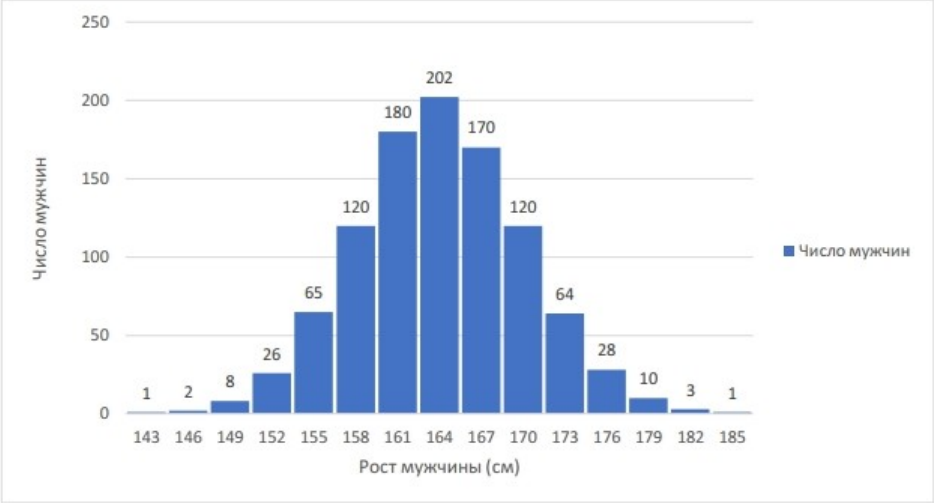
$$P\{X \in (z_{i-1}; z_i]\} = F(z_i) - F(z_{i-1}), \qquad F(\infty) = 1; \;\; F(-\infty) = 0. \tag{52}$$

Замечание. Напомню, что для непрерывных с.в. в формуле (52) с включением и исключением границ интервалов можно поступать достаточно небрежно.

Пример 3. В 1889-1890 гг. был измерен рост 1000 взрослых мужчин (рабочих московских фабрик). Результаты измерений представлены в таблице.

Рост (см) $(z_{i-1}; z_i]$	[143; 146]	(146; 149]	(149; 152]	(152; 155]	(155; 158]
Число мужчин a_i	1	2	8	26	65
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]
Число мужчин a_i	120	180	202	170	120
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	(185; 188]
Число мужчин a_i	64	28	10	3	1

Построим гистограмму распределения.



По ней сразу можно предположить, что рост мужчин распределен по нормальному закону распределения.

Вычислим эмпирические относительные частоты. Для этого разделим эмпирические частоты на объем выборки $n = 1000$.

Данные приведем в таблице относительных частот.

Рост (см) $(z_{i-1}; z_i]$	[143; 146]	(146; 149]	(149; 152]	(152; 155]	(155; 158]	
Эмпирические частоты a_i	1	2	8	26	65	
Эмпирические относительные частоты a_i/n	0,001	0,002	0,008	0,026	0,065	
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]	
Эмпирические частоты a_i	120	180	202	170	120	
Эмпирические относительные частоты a_i/n	0,12	0,18	0,202	0,17	0,12	
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	(185; 188]	
Эмпирические частоты a_i	64	28	10	3	1	$\sum_i a_i = n = 1000$
Эмпирические относительные частоты a_i/n	60,064	0,028	0,01	0,003	0,001	$\sum_i a_i/n = 1$

Вычислим теоретические вероятности $p_i = P\{X \in (z_{i-1}; z_i]$, предполагая, что $X \sim N(a, \sigma^2)$.

Для случая нормального распределния с.в. X формула (52) приобретает вид:

$$P\{X \in (z_{i-1}; z_i]\} = \Phi_0(\frac{z_i - a}{\sigma}) - \Phi_0(\frac{z_{i-1} - a}{\sigma}), \qquad \Phi_0(\infty) = 1/2; \;\; \Phi_0(-\infty) = -1/2, \quad (53)$$

где $\Phi_0()$ — функция Лапласа ($\Phi_0(-x) = -\Phi_0(x)$).

Заметим, что для того, чтобы воспользоваться формулой (53), необходимо знать математическое ожидание a и среднее квадратическое отклонение σ.

Поскольку они нам неизвестны, будем использовать их оценки, полученные по выборке каким-либо методом.

Согласно методу моментов

$$\hat{a} = \bar{x}; \qquad \hat{\sigma}^2 = \sigma^{2*}.$$

Замечание. Для малых выборок лучше считать “исправленную” выборочную дисперсию s^2 (unbiased sample variance).

Необходимо посчитать выборочное среднее $\bar{x}$ и выборочную дисперсию σ^{2*} по группировочному ряду. Для этого возьмем середины интервалов $(z_{i-1}; z_i]$ и вычислим искомые величины.

Имеем

$$\bar{x} = 165,533;\ \sigma^{2*} = 36,565911.$$

Будем использовать округленные значения. Тогда

$$\hat{a} = 165,53;\ \hat{\sigma}^2 = 36,57,\ \hat{\sigma} = \sqrt{36,57} = 6,05.$$

Подставляя полученные значения оценок для a, σ в формулу (53) для каждого интервала значений получаем теоретические вероятности.

Понятно, что с.в. X теоретически принимает значения от минус бесконечности до бесконечности. Тогда изменим вид первого и последнего интервалов: вместо 143 напишем минус бесконечность, вместо 188 напишем плюс бесконечность.

Очевидно, что в первом и последнем интервалах теоретические вероятности должны вычисляться по следующим формулам:

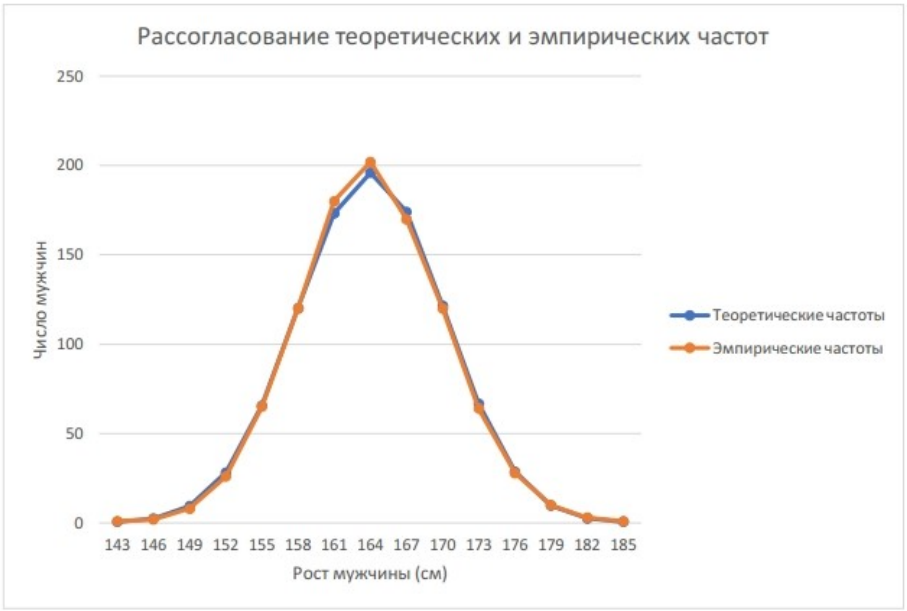
$$P\{X \in (-\infty; z_1]\} = \Phi_0(\frac{z_1 - a}{\sigma}) + 1/2, \tag{54}$$
$$P\{X \in (z_{r-1}; +\infty)\} = 1/2 - \Phi_0(\frac{z_{r-1} - a}{\sigma}).$$

Как и для дискретных величин, теоретические частоты вычисляются как $a'_i = p_i * n$, $n = 1000$ в данной задаче (объем выборки).

Получаем следующие результаты (с округлением).

Рост (см) $(z_{i-1}; z_i]$	$(-\infty; \mathbf{146}]$	(146; 149]	(149; 152]	(152; 155]	(155; 158]	
Теоретические вероятности p_i	0,0006	0,0025	0,00948	0,0281	0,06566	
Теоретические частоты $a'_i = p_i * n$	0,6185	2,5091	9,4836	28,1550	65,6624	
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]	
Теоретические вероятности p_i	0,1203	0,1732	0,1959	0,1741	0,1216	
Теоретические частоты $a'_i = p_i * n$	120,3098	173,1966	195,9075	174,1178	121,5931	
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	$(\mathbf{185}; +\infty)$	
Теоретические вероятности p_i	0,0667	0,02876	0,0097	0,0026	0,0006	$\sum p_i = 1$
Теоретические частоты $a'_i = p_i * n$	66,7160	28,7590	9,7386	2,5903	0,6425	$\sum a'_i = 1000$

Составим таблицу, наглядно демонстрирующую расхождение эмпирических и теоретических частот. Изобразим графически рассогласование частот.



Из графика мы видим, что скорее всего гипотезу о нормальности распределения с.в. X

Рост (см) $(z_{i-1}; z_i]$	$(-\infty; \mathbf{146}]$	(146; 149]	(149; 152]	(152; 155]	(155; 158]	
Эмпирические частоты a_i	1	2	8	26	65	
Теоретические частоты $a'_i = p_i * n$	0,6185	2,5091	9,4836	28,1550	65,6624	
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]	
Эмпирические частоты a_i	120	180	202	170	120	
Теоретические частоты $a'_i = p_i * n$	120,3098	173,1966	195,9075	174,1178	121,5931	
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	$(\mathbf{185}; +\infty)$	
Эмпирические частоты a_i	64	28	10	3	1	$\sum_i a_i = 1000$
Теоретические частоты $a'_i = p_i * n$	66,7160	28,7590	9,7386	2,5903	0,6425	$\sum a'_i = 1000$

можно принять.

14.3 Критерий согласия Пирсона (χ^2 критерий).

Критерий согласия К. Пирсона был предложен для сравнения того, насколько согласованы эмпирические (наблюдаемые) частоты/ относительные частоты и теоретические ("идеальные", в предположении истинности нулевой гипотезы) частоты / вероятности.

Кроме того, данный критерий может быть использован для проверки независимости двух признаков, а также для проверки однородности двух выборок.

14.3.1 Критерий согласия Пирсона для проверки гипотезы о виде распределения

Предположим, получены графики, показывающие рассогласование эмпирических и теоретических частот. Как проверить то, что нулевая гипотеза о том, что случайная величина X имеет конкретное распределение $F_0(x)$, может быть принята?

Имеем:

$$H_0 : F(x) = F_0(x)$$

$$H_1 : F(x) \neq F_0(x)$$

Вся область возможных значений X разбивается на r непересекающихся множеств (интервальное разбиение):

$$(z_0; z_1), (z_1; x_2), \ldots, (z_{r-2}; z_{r-1}), (z_{r-1}; z_r)$$

$z_0 = -\infty$

$z_r = +\infty$

$\Delta_i = z_i - z_{i-1}$

Сделаем выборку $x_1^*, \ldots, x_n^*$. Считаем, сколько значений попало в каждый интервал. Получаем группировочный ряд.

$\sum_{i=1}^{r} a_i = n$

Теоретическая частота:

$$a_i' = np_i,$$

Таблица 1: *Интервальное разбиение выборки*

Δ_1	Δ_2	...	Δ_r
a_1	a_2	...	a_r

где

$$p_i = p\{X \in \Delta_i | H_0\}$$

— теоретическая вероятность.

$$p_i = p\{X \in [z_{i-1}; z_i] \mid F_0(x)\} = F_0(z_i) - F_0(z_{i-1})$$

a_i – эмпирические частоты (то, что "в реальности")

a'_i — теоретические частоты (то, что "в теории")

Эти значения нужно сравнить.

Теорема Пирсона. Пусть гипотеза H_0 верна, тогда

$$R = \sum_{i=1}^{n} \frac{(a_i - a'_i)^2}{a'_i} \sim \chi^2(r - l - 1)$$

где $\chi^2(r-l-1)$ — χ^2 распределение с $(r-l-1)$ степенями свободы,

r — число интервалов разбиения,

l — число **неизвестных** параметров, определяемых для $F_0(x)$.

Замечание. Если проверяем гипотезу о нормальности распределения, то число степеней свободы $= (r-3)$. Для распределения Пуассона имеем число степеней свободы $= (r-2)$.

Если проверяется гипотеза о том, что вероятности значений в генеральной совокупности равны, то имеет место *простая* гипотеза, число неизвестных параметров равно 0 и при справедливости гипотезы $R \sim \chi^2(r-1)$, где r— число различных значений генеральной совокупности.

Критическая область имеем правосторонний вид!

$$R_{\text{КР}} = F^{-1}_{\chi^2}(1-\alpha) = \chi^2_{1-\alpha}(r-l-1).$$

Это значение берется из таблицы критических точек χ^2. Число степеней свободы $l-r-1$, уровень значимости (вероятность совершить ошибку I рода) α.

Если $\left[\begin{array}{l} R < R_{\text{КР}} \Rightarrow H_0 \\ R \geq R_{\text{КР}} \Rightarrow \bar{H}_0 \end{array}\right.$

Если H_0 принимаем, то $F(x) = F_0(x)$.

Замечание. Пусть X — дискретная случайная величина, тогда вместо интервалов используются отдельные изолированные значения x_i:

$$p_i = P\{x \in \Delta_i\} = p\{X = x_i\}.$$

Пример 4. Рассмотрим задачу из примера 1 §14.2.1 (о равной вероятности для ограблений по дням недели).

Имеем:

День недели k	1	2	3	4	5	6	7	
Эмпирическая частота n_k	105	121	129	115	110	134	126	$\sum n_k = n = 840$
Теоретическая частота n'_k	120	120	120	120	120	120	120	$\sum n'_k = n = 840$

Проверяется простая гипотеза

$$H_0 : p_1 = p_2 = \ldots = p_7 = 1/7$$

Проверим гипотезу H_0 на уровне значимости $\alpha = 0,1$.

Вычислим наблюдаемое значение критерия

$$R = \sum_{i=1}^{n} \frac{(n_i - n'_i)^2}{n'_i} \approx 5,53.$$

Тогда число степеней свободы для нахождения критической точки в таблице распределения χ^2 равно $7 - 1 = 6$. Напомним, что $\alpha = 0,1$.

Имеем из таблицы:

$$R_{\text{КР}} = 10,64.$$

Таким образом, наблюдаемое значение $R = 5,53$ находится левее $R_{\text{КР}} = 10,64$, значит наблюдаемое значение попало в область принятия нулевой гипотезы.

Нет основания отвергнуть гипотезу о том, что преступления совершаются с равной вероятностью в разные дни недели.

Пример 5. Рассмотрим задачу из примера 1 §14.2.1 (о снарядах в Лондоне во время Второй мировой войны). Проверим на уровне значимости $\alpha = 0,05$ гипотезу о том, что число разорвавшихся снарядов имеет распределение Пуассона.

Имеем таблицу, в которой сопоставлены эмпирические и теоретические частоты.

Значение случайной величины $X = k$	0	1	2	3	4	5	$[\mathbf{6}; \infty)$	
Наблюдаемые частоты n_k	229	211	93	35	7	1	**0**	$\sum\limits_k n_k = n = 576$
Теоретические частоты $n'_k = p_k * n$	227,5314	211,3356	98,1463	30,3867	7,0560	1,3107	**0,2333**	$\sum\limits_k n'_k = n = 576$

Вычислим наблюдаемое значение критерия

$$R = \sum_{i=1}^{n} \frac{(n_i - n'_i)^2}{n'_i} \approx 1,2877.$$

Для распределения Пуассона имеем один параметр (λ), а число интервалов разбиения здесь равно 7 (сколько различных значений).

Тогда число степеней свободы для нахождения критической точки в таблице распределения χ^2 равно $7 - 1 - 1 = 5$. Напомним, что $\alpha = 0,05$.

Имеем из таблицы:

$$R_{\text{КР}} = 11,07$$

Таким образом, наблюдаемое значение $R = 1,2877$ находится левее $R_{\text{КР}} = 11,07$, значит наблюдаемое значение попало в область принятия нулевой гипотезы.

Гипотеза о том, что с.в. X имеет распределение Пуассона принимается.

Пример 5. Рассмотрим задачу из примера 3 §14.2.2 (о росте мужчин в конце 19 в.). Проверим на уровне значимости $\alpha = 0,05$ гипотезу о том, что рост имеет нормальное распределение.

Имеем

Рост (см) $(z_{i-1}; z_i]$	$(-\infty; \mathbf{146}]$	(146; 149]	(149; 152]	(152; 155]	(155; 158]	
Эмпирические частоты a_i	1	2	8	26	65	
Теоретические частоты $a'_i = p_i * n$	0,6185	2,5091	9,4836	28,1550	65,6624	
Рост (см) $(z_{i-1}; z_i]$	(158; 161]	(161; 164]	(164; 167]	(167; 170]	(170; 173]	
Эмпирические частоты a_i	120	180	202	170	120	
Теоретические частоты $a'_i = p_i * n$	120,3098	173,1966	195,9075	174,1178	121,5931	
Рост (см) $(z_{i-1}; z_i]$	(173; 176]	(176; 179]	(179; 182]	(182; 185]	$(\mathbf{185}; +\infty)$	
Эмпирические частоты a_i	64	28	10	3	1	$\sum_i a_i = 1000$
Теоретические частоты $a'_i = p_i * n$	66,7160	28,7590	9,7386	2,5903	0,6425	$\sum a'_i = 1000$

Тогда

$$R = \sum_{i=1}^{n} \frac{(a_i - a'_i)^2}{a'_i} \approx 1,7194.$$

Для нормального распределения имеем два параметра (a, σ), а число интервалов разбиения здесь равно 15.

Тогда число степеней свободы для нахождения критической точки в таблице распределения χ^2 равно $15 - 2 - 1 = 12$. Напомним, что $\alpha = 0,05$.

Имеем из таблицы:

$$R_{\text{КР}} = 21,03$$

Таким образом, наблюдаемое значение $R = 1,7194$ находится левее $R_{\text{КР}} = 21,03$, значит наблюдаемое значение попало в область принятия нулевой гипотезы.

Гипотеза о том, что с.в. X имеет распределение Гаусса принимается.

Замечания.

- Если в интервал попало меньше трех значений выборки, то такой интервал лучше объединить с соседним.
- Критерий согласия Пирсона очень чувствителен к выбору интервального разбиения.
- Если наблюдаемое значение критерия попало ровно на границу области принятия гипотезы и критической области (либо очень близко), это нехорошо. Нужно проверить нулевую гипотезу с другим уровнем значимости α.
- Не все дискретные с.в. распределены по закону Пуассона и не все непрерывные с.в. распределены по нормальному закону. Существуют и другие виды распределения. К ним применим подход, описанный в χ^2 критерии.
- Существуют другие критерии согласия, например, Колмогорова-Смирнова.

14.3.2 Критерий согласия Пирсона для проверки гипотезы о независимости признаков. Таблицы сопряженности

Пусть имеются два признака (A и B), измеряемые в номинальных шкалах. Номинальная шкала — категориальная (т.е. качественная, а не количественная) шкала измерения, где каждое значение определяет отдельную категорию, в которую попадают значения переменной (каждая категория "отличается"от других, но это отличие не может быть количественно измерено).

Критерий согласия Пирсона может быть использован для проверки гипотезы о независимости признаков A и B.

Пусть признак A имеет r градаций, признак B имеет s градаций. Запишем *таблицу сопряженности*, где n_{ij} — количество элементов в выборке объема n, которые обладают одновременно и признаком A, и признаком B.

A \ B	B_1	B_2	...	B_s	$\sum$ по строке
A_1	n_{11}	n_{12}	...	n_{1s}	m_1
A_2	n_{21}	n_{22}	...	n_{2s}	m_2
...	n_{k1}	n_{k2}	...	n_{ks}	m_k
A_r	n_{r1}	n_{r2}	...	n_{rs}	m_r
$\sum$ по столбцу	n_1	n_2	...	n_s	$n = \sum_{i=1}^{r} m_i = \sum_{j=1}^{s} n_j$

Разделим на n эмпирические частоты. Получим таблицу относительных эмпирических частот.

A \ B	B_1	B_2	...	B_s	$\sum$ по строке
A_1	n_{11}/n	n_{12}/n	...	n_{1s}/n	m_1/n
A_2	n_{21}/n	n_{22}/n	...	n_{2s}/n	m_2/n
...	n_{k1}/n	n_{k2}/n	...	n_{ks}/n	m_k/n
A_r	n_{r1}/n	n_{r2}/n	...	n_{rs}/n	m_r/n
$\sum$ по столбцу	n_1/n	n_2/n	...	n_s/n	$1 = \sum_{i=1}^{r} m_i/n = \sum_{j=1}^{s} n_j/n$

Проверим гипотезу о том, что признаки (случайные величины) A и B независимы. Тогда вероятность $p_{ij} = P(A = A_i; B = B_j)$ должна быть равна произведению вероятностей $p_i p_j = P(A = A_i)P(B = B_j)$.

Имеем:

$$H_0 : p_{ij} = p_i p_j;$$

$$H_1 : p_{ij} \neq p_i p_j.$$

Из закона больших чисел имеем, что

$$n_{ij}/n \approx p_{ij},$$

$$m_i/n \approx p_i,$$

$$n_j/n \approx p_j.$$

Тогда нам необходимо проверить предположение о том, что

$$n_{ij}/n = \frac{m_i}{n}\frac{n_j}{n},$$

или (что то же)

$$n_{ij} = \frac{m_i n_j}{n},$$

Для уменьшения громоздкости обозначим n_{ij} — эмпирические (или наблюдаемые (observed)) частоты как O_{ij}.

Теоретические (или ожидаемые (expected)) частоты обозначим как E_{ij}. Это частоты, которые должны быть при справедливости гипотезы H_0, т.е.

$$E_{ij} = \frac{m_i n_j}{n}.$$

Вычислим теоретические частоты E_{ij} для каждой ячейки таблицы сопряженности. Запишем получившийся результат в скобках.

B / A	B_1	B_2	...	B_s	$\sum$ по строке
A_1	$O_{11}(E_{11})$	$O_{12}(E_{12})$	...	$O_{1s}(E_{1s})$	m_1
A_2	$O_{21}(E_{21})$	$O_{22}(E_{22})$	...	$O_{2s}(E_{2s})$	m_2
...	$O_{k1}(E_{k1})$	$O_{k2}(E_{k2})$	...	$O_{ks}(E_{ks})$	m_k
A_r	$O_{r1}(E_{r1})$	$O_{r2}(E_{r2})$	...	$O_{rs}(E_{rs})$	m_r
$\sum$ по столбцу	n_1	n_2	...	n_s	$n = \sum_{i=1}^{r} m_i = \sum_{j=1}^{s} n_j$

Для проверки гипотезы H_0 используем вспомогательную случайную величину:

$$R = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{55}$$

При выполнении гипотезы H_0 случайная величина R имеет χ^2 распределение с $(r-1)(s-1)$ степенью свободы.

Критическая

область является правосторонней. Значение критичекой точки вычисляется по таблице критических точек χ^2,

$$R_{\text{КР}} = \chi^2_{1-\alpha}(r-1)(s-1).$$

Замечание. Обратите внимание, что из таблицы, прикрепленной к конспекту лекций, нужно брать пересечение строк, соответствующих числу степеней свободы и α (не нужно отнимать $1-\alpha$).

Если наблюдаемое значение критерия $R < R_{\text{КР}}$, то нулевая гипотеза принимается. Признаки независимы.

Если наблюдаемое значение критерия $R \geq R_{\text{КР}}$, то гипотеза о независимости признаков отвергается. При этом совершается ошибка первого рода с вероятностью α.

Пример (А.И. Кибзун и др.) По переписи населения Швеции 1936 г. из совокупности всех супружеских пар была получена выборка в 25 263 пары, вступивших в брак в течение 1931- 1936 гг. В следующей таблице приведено распределение годовых доходов (в тыс. крон) и количество детей у супружеских пар. На уровне значимости 0.05 проверьте, являются ли зависимыми количество детей в семье и уровень годового дохода.

число детей \ годовой доход	0–1	1–2	2–3	> 3	$\sum$
0	2161	3577	2184	1636	9558
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	225	419	96	38	778
≥ 4	39	98	31	14	182
$\sum$	6116	10928	51173	3046	25263

Решение. Проверим гипотезу о том, что данные признаки независимы.

$n = 25263$.

Для каждой ячейки таблицы сопряженности вычислим теоретические (ожидаемые) частоты по формуле

$$E_{ij} = \frac{m_i n_j}{n},$$

где m_i — сумма элементов по строке i, n_j— сумма элементов по столбцу j.

число детей \ годовой доход	0–1	1–2	2–3	> 3
0	2161 (2313,93)	3577 (4134,5)	2184 (1957,15)	1636 (1152,42)
1	2755 (2689,66)	5081 (4805,85)	2222(2274,95)	1052 (1339, 55)
2	936 (880,01)	1753 (1572,39)	640 (744,32)	306 (438,28)
3	225 (188,35)	419 (336,54)	96 (159,31)	38 (93,80)
≥ 4	39(44,06)	98 (78,73)	31 (37,27)	14 (21,94)

Для каждой ячейки вычислим величину

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

а затем просуммируем их.

Получили

$$R = 568,5663.$$

$$R_{\text{КР}} = \chi^2_{1-\alpha}(r-1)(s-1) = \chi^2_{1-0.95}(3*4) = 21.02.$$

Поскольку наблюдаемое значение критерия находится правее критической точки, гипотезу о независимости признаков отвергаем. Доход семьи и количество детей согласно представленным данным связаны.

14.3.3 Критерий согласия Пирсона для проверки гипотезы об однородности выборок

Пусть даны две независимые выборки $x_1^*, \ldots, x_n^*$, $y_1^*, \ldots, y_m^*$ из генеральной совокупности X, Y соответственно. Пусть F_X, F_Y — функции распределения (неизвестные нам), соответствующие первой и второй с.в. Требуется проверить, что

$$H_0: \qquad F_X = F_Y,$$

$$H_1: \qquad F_X \neq F_Y,$$

т.е. выборки взяты из одной и той же генеральной совокупности.

Пусть n, m — объемы первой и второй выборок соответственно. Пусть r — число интервалов разбиения для первой и второй выборки (одинаковое).

Пусть n_i, $i = 1, \ldots, r$ — эмпирические частоты для первой выборки. Добавим дополнительный индекс 1, чтобы подчеркнуть, что это частоты первой выборки. Будем писать n_{i1}, $i = 1, \ldots, r$ — эмпирические частоты для первой выборки.

Аналогично n_{i2}, $i = 1, \ldots, r$ — эмпирические частоты для второй выборки.

Тогда критерий R для проверки гипотезы H_0 принимает вид:

$$R = n \cdot m \sum_{j=1}^{r} \frac{1}{n_{j1} + n_{j2}} \left(\frac{n_{j1}}{n} - \frac{n_{j2}}{m} \right)^2, \tag{56}$$

причем в случае выполнении гипотезы об однородности выборок R имеет распределение χ^2 с $r-1$ степенью свободы.

Критическая область имеет вид: $[R_{\text{КР}};\infty)$, где

$$R_{\text{КР}} = \chi^2_{1-\alpha}(r-1).$$

Если наблюдаемое значение критерия $R < R_{\text{КР}}$, то нулевая гипотеза принимается. Выборки однородны.

Если наблюдаемое значение критерия $R \geq R_{\text{КР}}$, то гипотеза об однородности признаков отвергается. При этом совершается ошибка первого рода с вероятностью α.

Замечание. Существует много других критерией проверки гипотезы об однородности выборок. Наиболее популярные из них — ранговые критерии Вилкоксона, Манна-Уитни.

Пример (А.И. Кибзун и др.). В таблице приведены данные о распределении доходов (в тыс. крон) всех промышленных рабочих и служащих Швеции в 1930 г. для возрастных групп 40-50 лет и 50-60 лет. Проверить на уровне значимости $\alpha = 0.05$ гипотезу H_0 о том, что доходы для обеих возрастных групп распределены одинаково.

возраст / доходы	40-50 лет	50-60 лет
$0-1$	7831	7558
$1-2$	26740	20685
$2-3$	35572	24186
$3-4$	20009	12280
$4-6$	11527	6776
>6	6919	4222
$\sum$	108598	75707

Решение.

Имеем $r = 6$ (6 групп), $n = 108598$, $m = 75707$. Числа n_{i1}, n_{i2} приведены во втором и третьем столбце таблицы соответственно.

Ищем критическую точку распределения χ^2 с 5 степенями свободы для $\alpha = 0.05$. Получаем $R_{\text{КР}} = 11.07$.

Вычисляем наблюдаемое значение критерия по формуле (56). Получаем $R = 840,62$.

Поскольку $840.62 > 11.07$, нулевая гипотеза отвергается. Выборки не являются

однородными. Доходы для разных возрастных групп не распределены одинаково.

14.4 Критерий согласия Колмогорова

Критерий согласия Колмогорова основан на сравнении эмпирической и гипотетической функций распределения и используется только в тех случаях, когда преполагаемое распределение непрерывно.

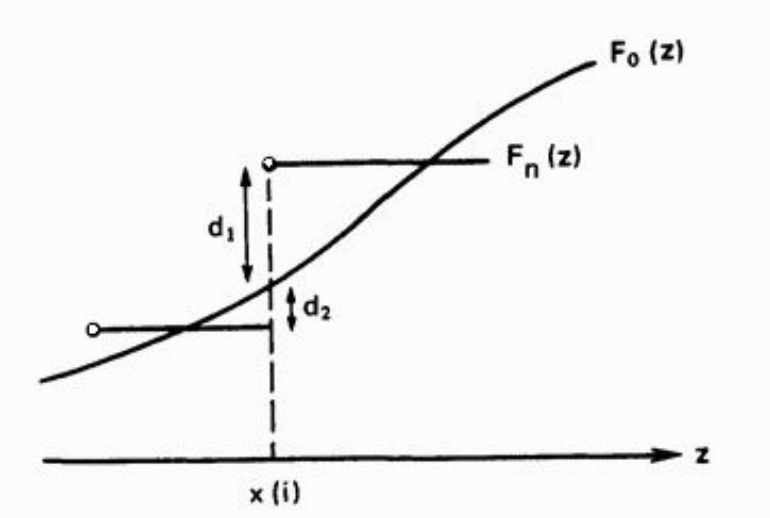
Алгоритм критерия согласия Колмогорова. Простая гипотеза

1. Выдвигаем нулевую гипотезу $H_0 : F_\xi(\cdot) = F_0(\cdot)$, при этом предполагается, что $F_0(x)$ полностью известна и представляет собой непрерывную функцию. Альтернативная гипотеза $H_1 : F_\xi(\cdot) \neq F_0(\cdot)$.
2. Задаем уровень значимости критерия α.
3. Вычисляем значение статистики критерия. Для этого:
   (a) По выборке $X_{[n]}$ строим эмпирическую функцию распределения $F_n^*(x)$.
   $F_n^*(x) = \frac{\nu(x)}{n}$, где $\nu(x)$ – число элементов выборки, меньших x. Если построен точечный вариационный ряд $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(k)}$, то функция $F_n^*(x)$ определяется по формуле:
   $$F_n^*(x) = \begin{cases} 0, & \text{при } x \leq y_{(1)}, \\ \frac{n_1}{n}, & \text{при } y_{(1)} < x \leq y_{(2)}, \\ \frac{n_1+n_2}{n}, & \text{при } y_{(2)} < x \leq y_{(3)}, \\ \frac{n_1+n_2+n_3}{n}, & \text{при } y_{(3)} < x \leq y_{(4)}, \\ \ldots & \\ 1, & \text{при } x > y_{(k)}. \end{cases}$$
   (b) Определяем статистику критерия по формуле
   $$D_n^* = \sup_{|x|<\infty} |F_n^*(x) - F_0(x)|.$$
   Для практического вычисления этой статистики полезна формула, эквивалентная предыдущей:
   $$D_n^* = \max_{1 \leqslant i \leqslant k} \left\{ |F_n^*(y_{(i)}) - F_0(y_{(i)})|, |F_0(y_{(i)}) - F_n^*(y_{(i+0)}| \right\}. \quad (57)$$

На графике ниже это сравнение расстояний d_2 и d_1 в каждой точке $y_{(i)}$.

В случае, когда все элементы выборки различны(а в идеале так и должно быть для предполагаемого непрерывного распределения), эта формула имеет вид:

$$D_n^* = \max_{1 \leqslant i \leqslant n} \left\{ |\frac{i}{n} - F_0(x_{(i)})|, |F_0(x_{(i)}) - \frac{i-1}{n}| \right\}. \tag{58}$$



При условии справедливости гипотезы H_0 и при $n \to \infty$ случайная величина $\sqrt{n}D_n^*$ асимптотически подчиняется распределению Колмогорова с функцией распределения

$$K(x) = \lim_{n \to \infty} \mathrm{P}\left\{\sqrt{n}D_n^* \leqslant x\right\} = 1 + 2\sum_{m=1}^{\infty}(-1)^m e^{-2m^2x^2}. \tag{59}$$

4. Критическая область будет иметь вид: $(k_{1-\alpha}; \infty)$, где $k_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения Колмогорова с функцией распределения (59).

Алгоритм критерия согласия Колмогорова. Сложная гипотеза

Пусть имеется выборка $X_{[n]} = \{x_1, \ldots, x_n\}$ из генеральной совокупности ξ с функцией распределения $F_\xi(x, \theta)$, где θ — неизвестный параметр, который может быть вектором. Используя метод максимального правдоподобия, найдем оценку $\hat{\theta}$ неизвестного параметра θ.

Тогда для проверки сложных гипотез согласия $H_0 : F_\xi(\cdot) = F_0(\cdot, \theta)$ используются модифицированные статистики. Модифицированная статистика Колмогорова имеет вид:

$$\hat{D}_n^* = \sup_x |F_n(x) - F_0(x, \hat{\theta})|.$$

Модифицированная статистика $\hat{D}_n^*$ не обладает свойством «свободы от распределения выборки», поэтому для каждого параметрического семейства распределений нужны свои таблицы.

Начиная с $n = 5$ можно использовать исправленные формы модифицированных статистик $\hat{D}_n^*$. В таблицах 1–2 приведены квантили уровня $1 - \alpha$ для $\alpha = 0.15, 0.10, 0.05, 0.025$ и 0.01, используемые при проверке гипотез о нормальном и экспоненциальном распределении генеральной совокупности, когда все параметры этих распределений оцениваются по выборке.

Табл. 2. Квантили уровня $1 - \alpha$ для проверки нормальности

Модифицированная форма	0.15	0.10	0.05	0.025	0.01
$\hat{D}_n^*\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)$	0.775	0.819	0.895	0.955	1.035

Табл. 3. Квантили уровня $1 - \alpha$ для проверки экспоненциальности

Модифицированная форма	0.15	0.10	0.05	0.025	0.01
$\left(\hat{D}_n^* - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$	0.926	0.990	1.094	1.190	1.308

Алгоритм критерия согласия Колмогорова в случае сложной гипотезы о нормальности распределения генеральной совокупности

1. Выдвигаем нулевую гипотезу $H_0 : F_\xi(\cdot) = F_0(\cdot, \theta)$. Сформулируем альтернативную гипотезу $H_1 : F_\xi(\cdot) \neq F_0(\cdot, \theta)$.
2. Задаем уровень значимости критерия α.
3. Находим оценки $\hat{\theta} = (\bar{x}, s^2)$ неизвестных параметров распределения $\theta = (a, \sigma^2)$.
4. Вычисляем значение исправленной формы статистики:
$$\hat{D}_n^*\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right).$$
5. Находим критическую область — интервал $(d_{1-\alpha}; \infty)$. Квантиль $d_{1-\alpha}$ можно найти из таблицы 1.
6. Если численное значение статистики $\hat{D}_n^*$ попадет в интервал $(d_{1-\alpha}; \infty)$, то нулевая гипотеза H_0 отвергается, в противном случае нет оснований отвергнуть нулевую гипотезу при уровне значимости приближенно равном α.

Алгоритм критерия согласия Колмогорова в случае сложной гипотезы об экспоненциальности распределения генеральной совокупности такой же, как и в случае проверки сложной гипотезы о нормальности распределения, но исправленная форма модифицированной статистики Колмогорова будет следующей: $\left(\hat{D}_n^* - \frac{0.2}{n}\right)\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right)$, и квантили этой статистики можно найти в таблице 2.

Задача 1

Пассажир, приходящий в случайные моменты времени на автобусную остановку, в течение пяти поездок фиксировал время ожидания автобуса: 5,1; 3,7; 1,2; 9,2; 4,8 (мин.) Проверить гипотезу о том, что время ожидания равномерно распределено на отрезке [0; 10] на уровне значимости $\alpha = 0,05$.

Решение

Рассматриваемая гипотеза простая. Все параметры распределения заданы.

Эмпирическая функция распределения имеет вид:

$$F_5^*(x) = \begin{cases} 0, & \text{при } x \le 1,2, \\ \frac{1}{5}, & \text{при } 1,2 < x \le 3,7, \\ \frac{2}{5}, & \text{при } 3,7 < x \le 4,8, \\ \frac{3}{5}, & \text{при } 4,8 < x \le 5,1, \\ \frac{4}{5}, & \text{при } 5,1 < x \le 9,2, \\ 1, & \text{при } x > 9,2. \end{cases}$$

Предполагаемое распределение равномерное, поэтому $F_0(x) = \frac{x}{10}$, $0 \le x \le 10$.

$$D_n^* = \max_{1 \leqslant i \leqslant 5} \left\{ \tfrac{1}{5} - \tfrac{1,2}{10}, \tfrac{1,2}{10} - 0, \tfrac{2}{5} - \tfrac{3,7}{10}, \tfrac{3,7}{10} - \tfrac{1}{5}, \tfrac{3}{5} - \tfrac{4,8}{10}, \tfrac{4,8}{10} - \tfrac{2}{5}, \right.$$
$$\left. \tfrac{4}{n} - \tfrac{5,1}{10}, \tfrac{5,1}{10} - \tfrac{3}{5}, 1 - \tfrac{9,2}{10}, \tfrac{9,2}{10} - \tfrac{4}{5} \right\} = 0,29,$$

$$\sqrt{n} D_n^* = 0,65.$$

Критическая область будет иметь вид: $(k_{0,95}; \infty)$, где $k_{0,95} = 1,36$. Значение статистики не попадает в критическую область, значит нет основания отвергнуть гипотезу.

Задача 2

Урожайность (ц/га) зерновых культур в России в 1992-2001гг. представлена таблицей.

Год	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Урож.	18	17,1	15,3	13,1	14,9	17,8	12,9	14,4	15,6	19,4

При уровне значимости $\alpha = 0,1$, используя критерий Колмогорова, проверить гипотезу о нормальном распределении урожайности.

Решение:

В данном сдучае нулевая гипотеза H_0: $F_\xi(x) = N(a;\sigma^2)$ является сложной, параметры a, σ^2 неизвестены. Альтернативная гипотеза $H_1:\ F_\xi(x) \neq N(a;\sigma^2)$.

Эмпирическая функция распределения имеет вид:

$$F_n^*(x) = \begin{cases} 0, & \text{при } x \leq 12,9, \\ 0,1, & \text{при } 12,9 < x \leq 13,1, \\ 0,2, & \text{при } 13,1 < x \leq 14,4, \\ 0,3, & \text{при } 14,4 < x \leq 14,9, \\ 0,4, & \text{при } 14,9 < x \leq 15,3, \\ 0,5, & \text{при } 15,3 < x \leq 15,6, \\ 0,6, & \text{при } 15,6 < x \leq 17,1, \\ 0,7, & \text{при } 17,1 < x \leq 17,8, \\ 0,8, & \text{при } 17,8 < x \leq 18, \\ 0,9, & \text{при } 18 < x \leq 19,4, \\ 1, & \text{при } x > 19,4. \end{cases}.$$

Оценим неизвестные параметры предполагаемого распредеения по выборке: $\hat{a} = \bar{x} = 15,85$, $\hat{\sigma}^2 = \sigma^{2*} = 4,2225$.

Необходимо найти $F_0(x_i)$ – значения предполагаемой функции распредления в точках x_i. Можно сделать это, используя таблицу значений функци Лапласа: $F_0(x_i) = \frac{1}{2} + \Phi_0(\frac{x_i - \hat{a}}{\hat{\sigma}})$.

$F_0(12,9) = 0,08$, $F_0(13,1) = 0,09$, $F_0(14,4) = 0,24$, $F_0(14,9) = 0,32$, $F_0(15,3) = 0,39$, $F_0(15,6) = 0,45$, $F_0(17,1) = 0,73$, $F_0(17,8) = 0,83$, $F_0(18) = 0,85$, $F_0(19,4) = 0,96$.

Тогда

$$\hat{D}_n^* = 0,15,$$

$$\hat{D}_n^* \left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right) = 0,51.$$

Критическая область будет иметь вид: $(d_{0,95}; \infty)$, где $d_{0,95} = 0,819$ (из табл. 1). Значение статистики 0,51 не попало в критическую область, нет оснований отвергнуть нулевую гипотезу.

14.4.1 Критерий однородности Колмогорова—Смирнова

Пусть имеются две выборки $X_{[n]} = \{x_1, \ldots, x_n\}$ и $Y_{[m]} = \{y_1, \ldots, y_m\}$ из генеральных совокупностей ξ и η соответственно. Элементы $x_1, \ldots, x_n$ образуют вариационный ряд $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$, а элементы $y_1, \ldots, y_m$ образуют вариационный ряд $y_{(1)} \leqslant \ldots \leqslant y_{(m)}$. Объемы выборок могут быть различны, но, не нарушая общности, предположим, что $m \leqslant n$. Функции распределения этих генеральных совокупностей равны $F(x)$ и $G(x)$ соответственно. Функции распределения $F(x)$ и $G(x)$ непрерывны.

1. Выдвигаем нулевую гипотезу $H_0 : F(\cdot) = G(\cdot)$ (проверяет гипотезу о равенстве функций распределения двух генеральных совокупностей ξ и η, из которых извлечены выборки).

   Альтернативная гипотеза H_1: $\sup\limits_{|x|<\infty} |F(x) - G(x)| > 0$.
2. Задаем уровень значимости критерия α.
3. Вычисляем значение статистики критерия следующим образом.
   (a) По выборкам $X_{[n]}$ и $Y_{[m]}$ строим эмпирические функции распределения $F_n^*(x)$ и $G_m^*(x)$
   (b) Статистика критерия имеет вид:

$$\sqrt{\frac{mn}{m+n}} D_{m,n} \tag{60}$$

   где

$$D_{m,n} = \sup_{|x|<\infty} |G_m^*(x) - F_n^*(x)|. \tag{61}$$

   На практике значение статистики $D_{m,n}$ вычисляют по формуле

$$D_{m,n} = \max_{1 \leqslant r \leqslant m} \left[|G_m^*(y_{(r)}) - F_n^*(y_{(r)})|, |F_n^*(y_{(r)}) - G_m^*(y_{(r)} + 0)| \right]. \tag{62}$$

   Или, что то же самое

$$D_{m,n} = \max_{1 \leqslant s \leqslant n} \left[|F_n^*(x_{(s)}) - G_m^*(x_{(s)})|, |G_m^*(x_{(s)}) - F_n^*(x_{(s)} + 0)| \right]. \tag{63}$$

   Если в каждой из выборок элементы различны (что бывает очень часто, т.к. критерий применим лишь для непрерывных распределений), то эта формула принимает вид:

$$D_{m,n} = \max_{1 \leqslant r \leqslant m} \left[|\frac{r}{m} - F_n^*(y_{(r)})|, |F_n^*(y_{(r)}) - \frac{r-1}{m}| \right]. \tag{64}$$

Или, что то же самое

$$D_{m,n} = \max_{1 \leqslant s \leqslant n} \left[|\frac{s}{n} - G_m^*(x_{(s)})|, |G_m^*(x_{(s)}) - \frac{s-1}{n}| \right]. \tag{65}$$

При справедливости нулевой гипотезы и неограниченном увеличении объемов выборок статистика

$$\sqrt{\frac{mn}{m+n}} D_{m,n} \tag{66}$$

асимптотически подчиняется распределению Колмогорова.

4. Находим критическую область — интервал $(k_{1-\alpha}; \infty)$, где $k_{1-\alpha}$ — квантиль уровня $1 - \alpha$ распределения Колмогорова (таблицу значений функции распрееделения Колмогорова можно найти в файле для прошлой пары).
5. Если численное значение статистики критерия (60) попадет в интервал $(k_{1-\alpha}; \infty)$, то нулевая гипотеза H_0 отвергается, в противном случае нет оснований ее отвергнуть при уровне значимости приближенно равном α.

Задача 4

Проверить гипотезу об однородности двух выборок

X:	3,49	3,5	3,53	3,62	3,79	3,8	3,81	3,99	4,01	4,05
Y:	3,7	3,81	3,83	3,85	3,9	4,1	4,38	4,66	4,96	

Решение

1. $H_0 : F(\cdot) = G(\cdot)$

   H_1: $\sup\limits_{|x|<\infty} |F(x) - G(x)| > 0.$
2. Задаем уровень значимости критерия α.
3. Строим эмпирические функции распределения $F_n^*(x)$ (красным цветом на графике) и $G_m^*(x)$ (синим цветом на графике)

$$F_n^*(x) = \begin{cases} 0, & \text{при } x \le 3,49, \\ 0,1, & \text{при } 3,49 < x \le 3,5, \\ 0,2, & \text{при } 3,5 < x \le 3,53, \\ 0,3, & \text{при } 3,53 < x \le 3,62, \\ 0,4, & \text{при } 3,62 < x \le 3,79, \\ 0,5, & \text{при } 3,79 < x \le 3,8, \\ 0,6, & \text{при } 3,8 < x \le 3,81, \\ 0,7, & \text{при } 3,81 < x \le 3,99, \\ 0,8, & \text{при } 3,99 < x \le 4,01, \\ 0,9, & \text{при } 4,01 < x \le 4,05, \\ 1, & \text{при } x > 4,05. \end{cases},$$

$$G_m^*(x) = \begin{cases} 0, & \text{при } x \le 3,7, \\ 0,11, & \text{при } 3,7 < x \le 3,81, \\ 0,22, & \text{при } 3,81 < x \le 3,83, \\ 0,33, & \text{при } 3,83 < x \le 3,85, \\ 0,44, & \text{при } 3,85 < x \le 3,9, \\ 0,56, & \text{при } 3,9 < x \le 4,1, \\ 0,67, & \text{при } 4,1 < x \le 4,38, \\ 0,78, & \text{при } 4,38 < x \le 4,66, \\ 0,89, & \text{при } 4,66 < x \le 4,96, \\ 1, & \text{при } x > 4,96. \end{cases}.$$

Найдём

$$D_{m,n} = \sup_{|x|<\infty} |G_m^*(x) - F_n^*(x)|. \tag{67}$$

Для этого находим $F_n^*(y_{(r)})$ (зелёные точки на графике):

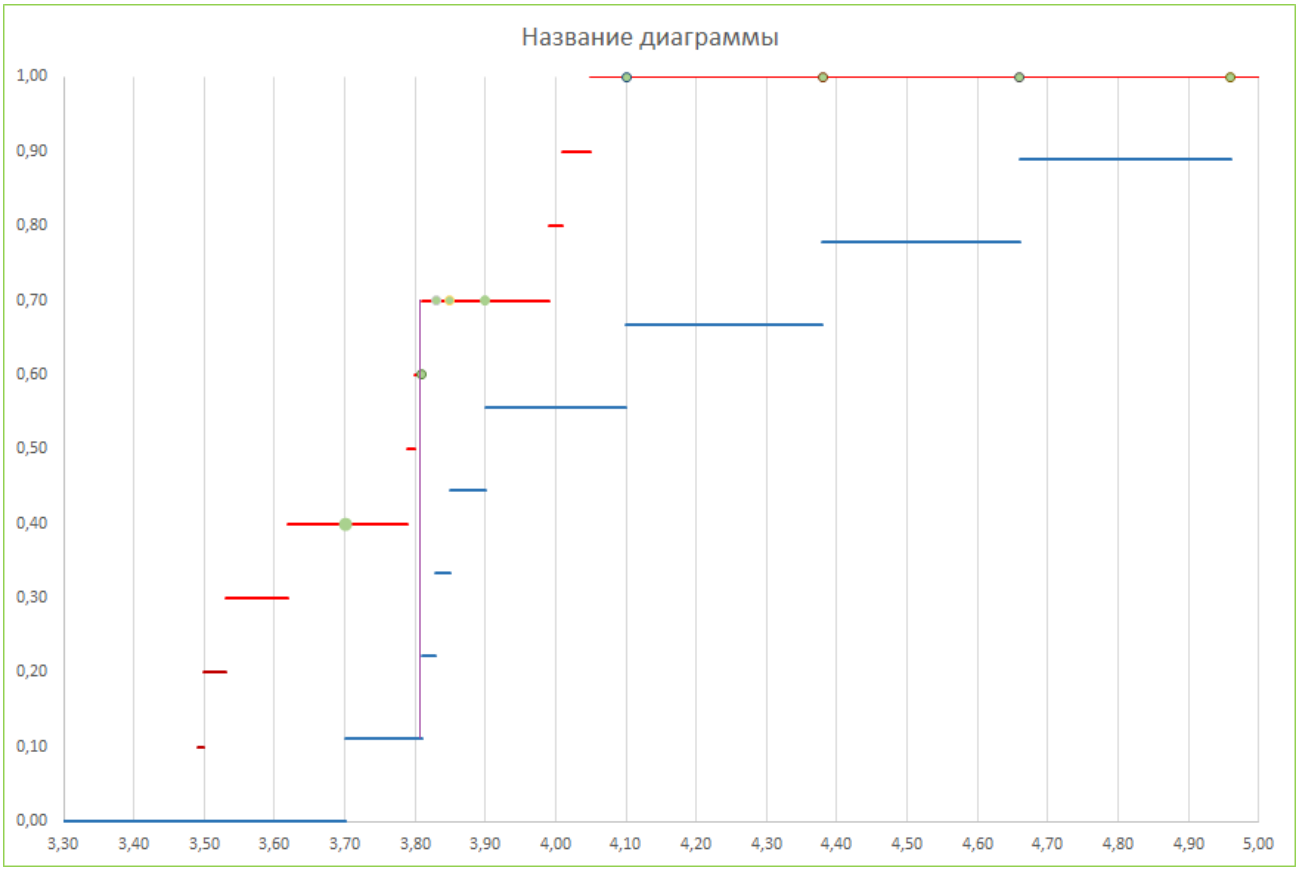
$F_n^*(3,7) = 0,4$, $F_n^*(3,81) = 0,6$

$F_n^*(3,83) = F_n^*(3,85) = F_n^*(3,9) = 0,7$

$F_n^*(4,1) = F_n^*(4,38) = F_n^*(4,66) = F_n^*(4,96) = 1.$

Найдём значение статистики $D_{m,n}$:

$D_{m,n} = \max\limits_{1\le r\le m} \big[|G_m^*(y_{(r)}) - F_n^*(y_{(r)})|, |F_n^*(y_{(r)}) - G_m^*(y_{(r)}+0)|, F_n^*(3,81+0) - G_n^*(3,81)\big]$ (т.к. значение 3,81 повторяется в двух выборках, то $G_n^*(3,81)$ нужно сравнить не только с $F_n^*(3,81)$, но и с $F_n^*(3,81+0)$, см. рисунок).



$D_{m,n} = 0,59$ (длина фиолетового отрезка на графике, $F_n^*(3,81+0) - G_n^*(3,81)$.

Тогда значение статистики критерия

$$\sqrt{\frac{mn}{m+n}} D_{m,n} = 1,28. \tag{68}$$

4. Находим критическую область — интервал $(k_{1-\alpha}; \infty)$, где $k_{0,95} = 1,36$
5. Значение статистики критерия не попало в критическую область, значит нет оснований отвергнуть нулевую гипотезу при уровне значимости приближенно равном $0,05$.

14.5 Задачи

Не забудьте графически изобразить рассогласование теоретических и эмпирических частот!

Задача 1. Пакетик арахиса, покрытого шоколадом, содержит 224 конфеты разных цветов: коричневые, оранжевые, желтые и зеленые. Проверьте гипотезу о том, что автомат, наполняющий эти пакетики конфетами, не отдает предпочтения каким-либо цветам, т.е.

гипотезу

$$p_1 = p_2 = p_3 = p_4 = 1/4.$$

Результаты подсчетов конфет в случайно выбранном пакетике приведены в таблице.

Цвет конфеты	коричневый	оранжевый	желтый	зеленый
Число конфет n_k	42	64	53	65

Уровень значимости α можно выбрать самостоятельно.

Задача 2. Результаты 60 подбрасываний кубика приведены в таблице.

Выпавшее число k	1	2	3	4	5	6
Число раз n_k	0	10	10	10	15	15

С вероятностью 0.05 проверьте предположение о том, что кубик правильный.

Задача 3. Ермолаев О.Ю. Задача 8.3 (об опросе респондентов, стр.131.

Психолог решает задачу: будет ли удовлетворенность работой на данном предприятии распределена равномерно по следующим градациям:

1 — Работой вполне доволен;

2 — Скорее доволен, чем не доволен;

3 — Трудно сказать, не знаю, безразлично;

4 — Скорее не доволен, чем доволен;

5 — Совершенно недоволен работой.

Был проведен опрос 65 респондентов. Результаты опроса представлены в виде таблицы.

Альтернатива k	1	2	3	4	5
Число респондентов n_k	8	22	14	9	12

Замечание. Здесь под равномерным распределением понимается то, что альтернативы имеют равные вероятности.

Задача 4. Гмурман В.Е. Задача 667 (задача Борткевича о кавалеристах). На основании 200 донесений, полученных в течение двадцати лет о количестве кавалеристов прусской

армии, которые погибли в результате гибели под ними коня, было получено следующее эмпирическое распределение:

Количество погибших k	0	1	2	3	4
Число донесений n_k	109	65	22	3	1

На уровне значимости $\alpha = 0,05$ проверить гипотезу о том, что число кавалеристов, погибших в результате гибели коня под ними, распределено по закону Пуассона.

Задача 5. (Гмурман В.Е.) Используя критерий Пирсона, при уровне значимости 0.05 проверить, согласуется ли гипотеза о нормальном распределении с.в. X с эмпирическим распределением:

Номер интервала i	1	2	3	4	5	6	7	8
Границы интервала i	[6; 16]	(16; 26]	(26; 36]	(36; 46]	(46; 56]	(56; 66]	(66; 76]	(76; 86]
Частота n_i	8	7	16	35	15	8	6	5

Задача 6. Компании, сменившие генерального директора, были классифицированы как “провалившиеся” в случае становления банкротом в течение следующих 3 месяцев и “не провалившиеся” в случае успешного развития. Эти же компании были разделены в зависимости от того, был ли новый генеральный директор из числа сотрудников компании или нет. Полученные результаты приведены в таблице. Протестируйте гипотезу о независимости признаков.

успешность компании / новый руководитель	Провал	Успех
из компании	21	14
не из компании	39	11

Задача 7. Основываясь на интервью семейных пар, подавших на развод, социальный работник собрал данные по продолжительности знакомства до свадьбы и

продолжительностью брака. Подтверждают ли данные связь между стабильностью брака и продолжительностью знакомства до брака? (Протестируйте гипотезу о независимости признаков).

продолжительность брака / знакомство до свадьбы	≤ 4 лет	> 4 лет
до 1/2 года	11	8
$0.5 - 1.5$ года	28	24
более 1.5 лет	21	19

Задача 8. Опрос 110 студентов был нацелен на поиск наиболее предпочитаемых занятий в свободное от учебы время. На выбор давалось 4 вариант: бег, статистика, пиво, поп-музыка. Связаны ли между собой выбор занятий во время досуга и курс студента? (Протестируйте гипотезу о независимости признаков).

любимое занятие / курс	бег	статистика	пиво	поп-музыка
1 курс	12	3	10	18
2 курс	11	9	10	10
3 курс	11	9	2	5

Задача 9. Опрос 2000 студентов в Москве и Париже показал следующее распределение по специальностям

специальность / город	математика	инженер	химия	экономика	другие	Всего
Москва	95	300	160	250	320	1125
Париж	75	200	100	230	270	875

На уровне значимости $\alpha = 0.05$ проверьте данные о том, что студенты в Москве и Париже распределены по специальностям одинаково.

Задача 10 (Ермолаев О.Ю., 8.9, стр. 148. Психолог сравнивает два эмпирических распределения, в каждом из которых было исследовано 200 человек по тесту интеллекта. Вопрос: различаются ли между собой эта два распределения?

уровень интеллекта \ частоты	выборка 1	выборка 2
60	1	1
70	5	3
80	17	7
90	45	22
100	70	88
110	51	69
120	10	7
130	1	2
140	0	1

Задача 11. В январе 2019 года диагноз «внебольничная пневмония» получили 1 576 человек, в январе 2020 года —1 731. Февраль 2020 года оказался даже более благополучным, чем 2019-го — 1 777 человек против 1 830. Зато в марте начался очевидный всплеск: в 2019 году 1 778 человек, в 2020-м — 2 716.

Проверьте выборки 2019 и 2020 года на однородность. Уровень значимости выберите сами.

14.6 Некоторые критерии значимости

14.6.1 Проверка гипотезы о генеральной доле (вероятности)

Предположим, что необходимо проверить гипотезу о вероятности p в схеме Бернулли. Напомним, что p — один из параметров биномиально распределенной случайной величины S_n — числа успехов в схеме Бернулли из n независимых испытаний.

Выдвигается гипотеза о том, что p равно конкретному значению p_0.

Рассмотрим три варианта для альтернативной гипотезы.

I. Двусторонняя критическая область.

$H_0 : p = p_0$

$H_1 : p \neq p_0$ Пусть $R = S_n$ — число успехов в n независимых испытаниях. Если нулевая гипотеза верна, то $MS_n = np_0$.

Имеет место *двусторонняя критическая* область.

$$R_{cr_1} = np_0 - \delta,$$

$$R_{cr_2} = np_0 + \delta,$$

где

$$\delta = \Phi_0^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{np_0(1-p_0)}.$$

Если $S_n \in [np_0 - \delta; np_o + \delta] \Rightarrow H_0$ не отвергается, иначе $\bar{H}_0$ (нулевая гипотеза отвергается, при этом совершается ошибка 1 рода с вероятностью α).

Напомним, значение $\Phi_0^{-1}\left(\frac{1-\alpha}{2}\right)$ берется из таблицы значений функции Лапласа (это значение для аргумента, соответствующего значению функции $(\frac{1-\alpha}{2})$.

II. Правосторонняя критическая область.

$H_0 : p = p_0$

$H_1 : p > p_0$

Имеет место *правосторонняя критическая* область. Это легко понять из интуитивных соображений: чем больше истинное значение p предполагаемого p_0, тем лучше это согласуется с альтернативной гипотезой.

$R_{\text{КР}} = np_0 + \Phi_0^{-1}(1/2 - \alpha)\sqrt{np_0(1-p_0)}$,

$\mathbb{R}_{\text{КР}} = [R_{\text{КР}}; n]$

Если $0 \leq S_n \leq np_0 + \Phi_0^{-1}(1/2 - \alpha)\sqrt{np_0(1-p_0)} \Rightarrow H_0$ не отвергается, иначе $\bar{H}_0$ (нулевая гипотеза отвергается, при этом совершается ошибка 1 рода с вероятностью α).

III. Левосторонняя критическая область

$H_0 : p = p_0$

$H_1 : p < p_0$

Имеет место *левосторонняя критическая* область. Это легко понять из интуитивных соображений: чем меньше истинное значение p предполагаемого p_0, тем лучше это согласуется с альтернативной гипотезой.

$R_{\text{КР}} = np_0 - \Phi_0^{-1}(1/2 - \alpha)\sqrt{np_0(1-p_0)}$,

$\mathbb{R}_{\text{КР}} = [0; R_{\text{КР}}]$

Если $n \geq S_n \geq np_0 - \Phi_0^{-1}(1/2-\alpha)\sqrt{np_0(1-p_0)} \Rightarrow H_0$ не отвергается, иначе $\bar{H}_0$ (нулевая гипотеза отвергается, при этом совершается ошибка 1 рода с вероятностью α).

Пример 1. Производитель противогриппозной вакцины утверждает, что в 87% случаев его продукция эффективна. Проверить $p \neq p_0$ при $\alpha = 0.05$, если известно, что из 400 привитых пациентов 45 заболели.

Решение.

$H_0 : p = 0.87, \quad H_1 : p \neq 0.87$

$S_n = 355, \quad n = 400, \quad np_0 = 400 * 0,87 = 348$

(Если нулевая гипотеза верна, то в среднем должны не заболевать 348 чел.)

Область принятия гипотезы:

$$\left[np_0 - \Phi_0^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{np_0(1-p_0)}; \quad np_0 + \Phi_0^{-1}\left(\frac{1-\alpha}{2}\right)\sqrt{np_0(1-p_0)}\right]$$

$$\sqrt{np_0(1-p_0)} = 6.8, \quad \Phi_0^{-1}(\frac{1-\alpha}{2}) = 1,96.$$

Область принятия гипотезы: $[335; 361] \Rightarrow 355 \in [335; 361] \Rightarrow H_0$ нет основания отвергнуть.

Пример 2. Владелец агенства по продаже недвижимости утверждает, что 40% обращавшихся в агенство (зарегистрированных) клиентов совершают сделки. Для проверки из регистрационной книги случайным образом выбраны 100 клиентов. Оказалось, что 34 из них совершили сделки. Можно ли при уровне значимости $\alpha = 0,05$ считать, что владелец завысил результаты работы агенства?

Решение.

$H_0 : p = 0.4, \quad H_1 : p < 0.4$

$S_n = 34, \quad n = 100, \quad np_0 = 100 * 0,4 = 40$ (Если нулевая гипотеза верна, то в среднем должны заключать сделки 40 чел.)

$$\sqrt{np_0(1-p_0)} = 24, \qquad \Phi_0^{-1}(1/2-\alpha) = \Phi_0^{-1}(0,45) = 1,65.$$

Область принятия гипотезы (для левосторонней альтернативы!) $n \geq S_n \geq np_0 - \Phi_0^{-1}(1/2-\alpha)\sqrt{np_0(1-p_0)}$, т.е. $[15,92; 100]$. Поскольку наблюдаемое значение S_n равно 24, имеем $24 \in [15,92; 100]$. Следовательно, нет оснований отвергнуть нулевую гипотезу H_0. Нет оснований утверждать, что владелец завысил результаты работы агенства.

14.6.2 Проверка гипотезы о равенстве вероятностей

Пусть имеются две генеральные совокупности. Требуется сравнить генеральные доли $p_1,\ p_2$ какого-то признака в этих совокупностях на основе выборочных данных. Пусть первая выборка имеет объем $n_1,$ вторая выборка имеет объем n_2. Пусть $S_{n1},\ S_{n2}$ — число появлений признака в первой и второй выборке соответственно. Тогда

$$\hat{p}_1 = \frac{S_{n1}}{n_1}; \qquad \hat{p}_2 = \frac{S_{n2}}{n_2}; \qquad \hat{p} = \frac{S_{n1} + S_{n2}}{n_1 + n_2}.$$

Здесь $\hat{p}$ — оценка вероятности появления признака в объединенных совокупностях.

В качестве вспомогательной случайной величины R возьмем так называемую z-статистику (на самом деле, это нормированная с.в., которую можно было использовать и в предыдущем пункте):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}.$$

Если нулевая гипотеза верна, то статистика z распределена по стандартному нормальному закону $N(0,1)$.

I. Двусторонняя критическая область.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$$R_{cr_1} = -\Phi_0^{-1}(\frac{1-\alpha}{2}),\ \ R_{cr_2} = \Phi_0^{-1}(\frac{1-\alpha}{2}).$$

Если наблюдаемое значение $z \in [R_{cr_1}; R_{cr_2}]$, тогда H_0 принимается. В противном случае гипотеза о равенстве вероятностей отвергается.

II. Правосторонняя критическая область.

$H_0 : p_1 = p_2$

$H_1 : p_1 > p_2$

$$R_{cr} = \Phi_0^{-1}(1/2 - \alpha).$$

Если наблюдаемое значение $z \leq R_{cr}$, тогда H_0 принимается. В противном случае гипотеза о равенстве вероятностей отвергается.

III. Левосторонняя критическая область.

$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$

$$R_{cr} = -\Phi_0^{-1}(1/2 - \alpha).$$

Если наблюдаемое значение $z \geq R_{cr}$, тогда H_0 принимается. В противном случае гипотеза о равенстве вероятностей отвергается.

Пример 3. В одном роддоме в течение лета родилось 1000 младенцев, а вдругом — 500. Известно, что в первом доме из родившихся детей 515 – мальчики, а во втором их 240. Протестируйте на уровне значимости $0,05$ гипотезу о том, что доля младенцев мужского пола одинакова.

Решение.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

Имеем

$$n_1 = 1000;\ S_{n1} = 515; \qquad n_2 = 500;\ S_{n2} = 240.$$

Тогда

$$\hat{p}_1 = 515/1000 = 0,515; \qquad \hat{p}_2 = 240/500 = 0,48; \qquad \hat{p} = (515 + 240)/(1000 + 500) \approx 0,503.$$

Получаем

$$z = \frac{0,515 - 0,48}{\sqrt{0,5030,497(1/1000 + 1/500)}} \approx 1,2774.$$

$$\Phi_0^{-1}(\frac{1-\alpha}{2}) = \Phi_0^{-1}(0,475) = 1,96.$$

$$R_{cr_1} = -1,96,\ \ R_{cr_2} = 1,96.$$

Поскольку $1,2774 \in [-1,96; 1,96]$, нет оснований отвергнуть нулевую гипотезу.

14.7 Задачи

Задача 1. В одном роддоме в течение лета родилось 1000 младенцев, а вдругом — 500. Известно, что в первом доме из родившихся детей 515 – мальчики, а во втором их 240. Предполагая, что вероятность рождения мальчика равна 0.5, проверьте гипотезу о том,

что в первом роддоме доля мальчиков равна 0.5. Проверьте ту же гипотезу для второго роддома. Какую гипотезу стоит выбрать в качестве альтернативной?

Задача 2. Исследователи в области лечения онкологических заболеваний часто сообщают количество пациентов, выживших за определенный период времени после проведенного курса лечения. Пусть известно, что среди пациентов, которые прошли стандартный курс лечения, 30 % прожили более 5 лет после проведенного курса лечения. Курс лечения по новой методике применялся к 100 пациентам, из которых 38 прожили более 5 лет.

a) Сформулируйте нулевую и альтернативную гипотезы для проверки утверждения, что новая методика лечения более эффективна, чем стандартная.

b) Проверьте гипотезу на уровне значимости $\alpha = 0,05$ и сформулируйте ваше заключение.

Решение.

$$S_n = 38, \qquad n = 100, \qquad \alpha = 0,05.$$

$H_0 : p = 0,3$

$H_1 : p > 0,3$

Имеет место *правосторонняя критическая* область.

$R_{\text{КР}} = np_0 + \Phi_0^{-1}(1/2 - \alpha)\sqrt{np_0(1-p_0)} \approx 37,56,$

$\mathbb{R}_{\text{КР}} = [37,56; 100]$

Поскольку $38 > 37,56$, нулевая гипотеза отвергается. Новая методика более эффективна.

Задача 3. В Бостоне в 1968 г. доктор Спок (знаменитый врач-педиатр) предстал перед судом за активность в протестах против Вьетнамской войны. Среди выбранных судьей 700 кандидатов в присяжные оказалось только 15 % женщин. В то же время в городе среди возможных кандидатов в присяжные было 29 % женщин. Пусть p — вероятность выбора женщины в члены жюри присяжных. Какую гипотезу надо проверить, чтобы оценить непредвзятость судьи? Может H_0 быть отвергнута на уровне значимости 0.05?

Решение.

$$S_n = 700 * 0,15 = 105, \qquad n = 100, \qquad \alpha = 0,05.$$

$H_0 : p = 0,29$

$H_1 : p < 0,29$

Имеет место *левосторонняя критическая* область.

$R_{\text{КР}} = np_0 - \Phi_0^{-1}(1/2 - \alpha)\sqrt{np_0(1-p_0)} \approx 183,19,$

$\mathbb{R}_{\text{КР}} = [0; 183,19]$

Поскольку $105 < 183,19$, нулевая гипотеза отвергается в пользу гипотезы о гендерной дискриминации.

Задача 4. Исследование 10 000 транспортных происшествий, в которых участвовали автомобили с ремнями безопасности, дало следующие результаты.

использование ремня / серьезные травмы	да	нет	всего
да	3	119	122
нет	829	9040	9878
всего	832	9168	10 000

Указывают ли данные на пользу ремней безопасности?

Решение. Обозначим через p_1 долю происшествий с серьезными травмами, когда использовались ремни безопасности, через p_2 долю происшествий с серьезными травмами, когда не использовались ремни безопасности.

Тогда

$$n_1 = 832;\ S_{n1} = 3; \qquad n_2 = 9168;\ S_{n2} = 119.$$

$$\hat{p}_1 = \frac{S_{n1}}{n_1} \approx 0,003; \qquad \hat{p}_2 = \frac{S_{n2}}{n_2} \approx 0,0013; \qquad \hat{p}\frac{S_{n1} + S_{n2}}{n_1 + n_2} = 0,0122.$$

Здесь $\hat{p}$ — оценка вероятности появления признака в объединенных совокупностях.

В качестве вспомогательной случайной величины R возьмем так называемую z-статистику:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \approx -2,53.$$

Левосторонняя критическая область.

$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$

$$R_{cr} = -\Phi_0^{-1}(1/2 - \alpha) = -1,65.$$

Если наблюдаемое значение $z < R_{cr}$, тогда H_0 отвергается. Использование ремней безопасности делает поездку более безопасной.

Задача 5. В выборке 1200 жителей Дании 480 положительно относятся к автодилерам. В независимой выборке 1000 жителей Франции 440 положительно относятся к автодилерам. Проверьте на 1% уровне значимости нулевую гипотезу о равенстве пропорций в генеральных совокупностях.

Решение.

Имеем

$$n_1 = 1200;\ S_{n1} = 480; \qquad n_2 = 1000;\ S_{n2} = 440.$$

$$\hat{p}_1 = \frac{S_{n1}}{n_1} = 0,4; \qquad \hat{p}_2 = \frac{S_{n2}}{n_2} = 0,44; \qquad \hat{p}\frac{S_{n1} + S_{n2}}{n_1 + n_2} = 0,4182.$$

Здесь $\hat{p}$ — оценка вероятности появления признака в объединенных совокупностях.

В качестве вспомогательной случайной величины R возьмем так называемую z-статистику:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = -1,8939.$$

Левосторонняя критическая область.

$H_0 : p_1 = p_2$

$H_1 : p_1 < p_2$

$$R_{cr} = -\Phi_0^{-1}(1/2 - \alpha) = -2,326.$$

Если наблюдаемое значение $z > R_{cr}$, тогда H_0 не отвергается. Гипотеза о равенстве пропорций не отвергается.

Задача 6. Майкл Джордан забросил 35 из 40 мячей во время финальной серии. Означает ли это, что процент попаданий для него меньше 90? (Уровень значимости задайте сами)

Задача 7. В 1980 г. институтом Гэллапа было произведено исследование по изучению мнения американцев относительно того, являются ли текущие меры по обеспечению атомных станций достаточными. Из 420 респондентов в возрасте от 18 до 30 лет 24 % ответили "да". Из 510 респондентов в возрасте от 30 до 50 лет 34 % ответили "да". Проверьте нулевую гипотезу о том, что возраст не имеет значения на уровне значимости 0.05.

14.8 Проверка гипотез о параметрах нормальных совокупностей

14.8.1 Проверка гипотезы о математическом ожидании нормальной совокупности

Здесь и далее предполагаем, что имеется выборка из нормально распределенной генеральной совокупности. Тем не менее, при больших n критерии проверки применимы и для других видов распределения.

Пусть X — случайная величина, распределенная по нормальному закону $N(a, \sigma^2)$. Предполагаем, что один или оба параметра неизвестны.

I. Пусть σ (среднее квадратическое отклонение) известно. Как и в случае проверки гипотезы относительно вероятности, рассмотрим три случая.

Двусторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a \neq a_0$

Вычисляем

$$R = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}. \tag{69}$$

Замечание. Мы знаем (см. Теорему Фишера и выводы до нее), что в случае справедливости нулевой гипотезыR (69) имеет стандартное нормальное распределение $N(0, 1)$.

Значение $R_{cr1} = -\Phi_0^{-1}(\frac{1-\alpha}{2})$, $R_{cr2} = \Phi_0^{-1}(\frac{1-\alpha}{2})$. Если $R \in [-R_{cr1}; R_{cr2}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Правосторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a > a_0$

Вычисляем

$$R = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}.$$

Значение $R_{cr} = \Phi_0^{-1}(1/2 - \alpha)$, Если $R \in [-\infty; R_{cr}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Левосторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a < a_0$

Вычисляем

$$R = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma}.$$

Значение $R_{cr} = -\Phi_0^{-1}(1/2 - \alpha)$, Если $R \in [R_{cr}; \infty]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Пример 4

Директор швейной фабрики желает определить, соотвествует ли ткань, произведенная на новом станке, заданным техническим требованиям. В частности ткань должна иметь прочность 70 фунтов на квадратный дюйм при стандартном отклонении 3,5 дюйма. Анализ выборки, состоящей из 49 отрезков ткани, показал, что средняя прочность ткани рана 69,1. Есть ли основания утверждать, что новый станок не соответствует техническим требованиям (уровень значимости равен 0,05, распределения считать нормальным)?

Решение. Действуем по алгоритму проверки статистических гипотез.

1. $H_0 : a = 70$,

   $H_1 : a \neq 70$.
2. Уровень значимости равен $\alpha = 0,05$.
3. По формуле (69): $R = \frac{\bar{x} - a_0}{\sigma}\sqrt{n} = \frac{69,1-70}{3,5} \cdot 7 = -1,8$.
4. Ищем двустороннюю критическую область. $\Phi_0^{-1}(\frac{1-\alpha}{2}) = 1,96$. Значит, критическая область имеет вид: $(-\infty, -1,96) \cup (1,96, +\infty)$.
5. Делаем вывод. Значение статистики -1,8 не попало в критическую область, значит нет оснований отвергнуть нулевую гипотезу и нет оснований утверждать, что новый

станок не соответствует техническим требованиям.

Пример 5

Компания, производящая сыр, желает проверить качество поставляемого молока. В частности, её интересует, не подмешивает ли производитель воду в молоко. Как известно, добавление воды снижает температуру замерзания молока, которая равна $-0,545^0$ С. Стандартное отклонение температуры замерзания молока равно $0,08^0$ С. Из партии молока случайным образом выбраны 25 бидонов. Выборочная средняя температура замерзания равна $-0,550^0$ С. Проверить гипотезу об обмане (уровень значимости 0,01, распределение считаем нормальным).

Решение

Действуем по алгоритму проверки статистических гипотез.

1. $H_0 : a = -0,545$,
   $H_1 : a < -0,545$ (Мы знаем, что если H_0 не верна, то температура уменьшится).
2. Уровень значимости равен $\alpha = 0,01$.
3. По формуле (69): $R = \frac{\bar{x}-a_0}{\sigma}\sqrt{n} = \frac{-0,55+0,545}{0,08} \cdot 5 = -3,125$.
4. Ищем левостороннюю критическую область. $-\Phi_0^{-1}(\frac{1}{2} - \alpha) = -2,33$. Значит, критическая область имеет вид: $(-\infty, -2,33)$.
5. Делаем вывод. Значение статистики -3,125 попало в критическую область, значит нулевую гипотезу следует отвергнуть. А компания должна провести расследование данного факта.

II. Пусть σ (среднее квадратическое отклонение) неизвестно. Опять рассмотрим три случая.

В качестве критерия каждый раз будем использовать

$$R = \frac{(\bar{x} - a_0)\sqrt{n-1}}{\sqrt{\sigma^{2*}}} = \frac{(\bar{x} - a_0)\sqrt{n}}{s}, \tag{70}$$

где σ^{2*}— выборочная дисперсия, s^2— “исправленная” выборочная дисперсия (см. формулы (??), (??)).

$$\sigma^{2^*} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - (\bar{x})^2.$$

$$s^2 = \frac{n}{n-1}\sigma^{2^*} = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i.$$

Замечание. Мы знаем (см. Теорему Фишера и выводы до нее), что в случае справедливости нулевой гипотезы статистика R (70) имеет распределение Стьюдента с (n-1) степенью свободы.

Опять рассмотрим три случая.

Двусторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a \neq a_0$

Значение $R_{cr1} = -t^{-1}_{n-1}(1-\alpha/2)$ — квантиль распределения Стьюдента порядка $1-\alpha/2$ для числа степеней свободы $n-1$. $R_{cr2} = t^{-1}_{n-1}(1-\alpha/2)$. Если $R \in [-R_{cr1}; R_{cr2}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Правосторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a > a_0$

Значение $R_{cr} = t^{-1}_{n-1}(1-\alpha)$, Если $R \in [-\infty; R_{cr}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Левосторонняя критическая область.

$H_0 : a = a_0$

$H_1 : a < a_0$

Значение $R_{cr} = -t^{-1}_{n-1}(1-\alpha)$, Если $R \in [R_{cr}; \infty]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Пример 6

В компании проводится аудиторская проверка. Для проверки аудитор извлекает из информационной системы выборку накладыных, заполненных в течение последнего месяца. Средняя сумма накладных за 5 лет равна 120 долл. Аудитор должен оценить, изменилась ли сумма накладных. Было извлечено 12 накладных, их средняя сумма составила 112,85 долл., а исправленное выборочное отклонение равно 20,85 долл.(уровень значимости равен 0,05, распределения считать нормальным).

Решение

Действуем по алгоритму проверки статистических гипотез, дисперсия генеральной совокупности не дана, значит это II случай.

1. $H_0 : a = 120$,

   $H_1 : a \neq 120$.
2. Уровень значимости равен $\alpha = 0,05$.
3. По формуле (2): $R = \frac{\bar{x}-a_0}{s}\sqrt{n} = \frac{112,85-120}{20,85} \cdot \sqrt{12} = -1,19$.
4. Ищем двустороннюю критическую область. $t^{-1}_{n-1}(1-\frac{\alpha}{2}) = 2,201$. Значит, критическая область имеет вид: $(-\infty, -2,201) \cup (2,201, +\infty)$.
5. Делаем вывод. Значение статистики -1,19 не попало в критическую область, значит нет оснований отвергнуть нулевую гипотезу. Аудитор не имеет право заключить, что средняя сумма накладных значительно отличается от 120.

14.8.2 Проверка гипотезы о равенстве математических ожиданий

I. Предположим, имеются две случайные величины X, Y, которые независимы и нормально распределены с неизвестными математическими ожиданиями a_x, a_y и *известными* дисперсиями σ_x^2, σ_y^2.

Производятся независимые повторные выборки объемами n_x, n_y из указанных совокупностей X, Y. На основании выборочных средних $\bar{x}$, $\bar{y}$ необходимо сделать вывод о равенстве неизвестных математических ожиданий a_x, a_y.

Описанные ниже методы применяются для проверки *однородности* двух выборок.

Вычисляем

$$R = \frac{(\bar{x} - \bar{y})}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}.$$

Двусторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x \neq a_y$

Значение $R_{cr1} = -\Phi_0^{-1}(\frac{1-\alpha}{2})$, $R_{cr2} = \Phi_0^{-1}(\frac{1-\alpha}{2})$. Если $R \in [-R_{cr1}; R_{cr2}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Правосторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x > a_y$

Значение $R_{cr} = \Phi_0^{-1}(1/2 - \alpha)$, Если $R \in [-\infty; R_{cr}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Левосторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x < a_y$

Значение $R_{cr} = -\Phi_0^{-1}(1/2 - \alpha)$, Если $R \in [R_{cr}; \infty]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Пример

1. *Руководство фирмы сравнивает результаты работы двух своих подразделений, для чего была сделана выборка в количестве 50 счетов из первого отделения и 60 счетов из второго. Результаты проверки следующие: в первом подразделении средний размер счета равен* $\bar{x} = 110$ *(фунтов стерлингов), во втором средний счет равен* $\bar{y} = 100$. *На основании прошлых проверок можно считать, что* $\sigma_x = 25$, $\sigma_y = 20$ *(фунтов). Можно ли на уровне значимости 0.05 считать, что средний размер счета в первом подразделении больше чем во втором?*

Решение.

Правосторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x > a_y$

Вычисляем

$$R = \frac{(\bar{x} - \bar{y})}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}} = \frac{(110 - 100)}{\sqrt{25^2/50 + 20^2/60}} = 2.28.$$

Значение $R_{cr} = \Phi_0^{-1}(1/2 - \alpha) = 1,65$. Поскольку $R \notin [-\infty; R_{cr}]$, то нулевая гипотеза отвергается. На уровне значимости 0.05 считать, что средний размер счета в первом подразделении больше чем во втором.

II. Предположим, имеются две случайные величины X, Y, которые независимы и нормально распределены с неизвестными математическими ожиданиями a_x, a_y и *неизвестными* дисперсиями σ_x^2, σ_y^2 (но известно, что $\sigma_x^2 = \sigma_y^2$).

Замечание. Сначала следует проверить гипотезу о равенстве дисперсий $\sigma_x^2 = \sigma_y^2$! См. следующий раздел.

Вычислим по выборке исправленные дисперсии $s_x^2,\ s_y^2$ (см. формулу (**??**)).

Рассмотрим

$$R = \frac{\bar{x} - \bar{y}}{\sqrt{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}.$$

Опять рассмотрим три случая.

Двусторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x \neq a_y$

Значение $R_{cr1} = -t^{-1}_{n_x+n_y-2}(1 - \alpha/2)$ — квантиль распределения Стьюдента порядка $1 - \alpha/2$ для числа степеней свободы $n_x + n_y - 2$. $R_{cr2} = t^{-1}_{n_x+n_y-2}(1 - \alpha/2)$. Если $R \in [-R_{cr1}; R_{cr2}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Правосторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x > a_y$

Значение $R_{cr} = t^{-1}_{n_x+n_y-2}(1 - \alpha)$, Если $R \in [-\infty; R_{cr}]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Левосторонняя критическая область.

$H_0 : a_x = a_y$

$H_1 : a_x < a_y$

Значение $R_{cr} = -t^{-1}_{n_x+n_y-2}(1 - \alpha)$, Если $R \in [R_{cr}; \infty]$, то нулевая гипотеза принимается. В противном случае она отвергается.

Пример 6. Пусть в условиях примера 1 о средних счетах в двух подразделениях σ_x, σ_y неизвестны, но известно, что они равны. Пусть посчитаны стандартные отклонения по выборкам:

$$s_x = 25,\ s_y = 20.$$

Можно ли на уровне значимости 0.05 утверждать, что средний счет в первом отделении больше, чем во втором?

Решение.

Непосредственными вычислениями получаем

$$R = \frac{\bar{x} - \bar{y}}{\sqrt{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}} = 6.475.$$

Значение $R_{cr} = t^{-1}_{50+60-2}(1 - 0,05) = 1,66$. Поскольку $R \notin [-\infty; R_{cr}]$, то нулевая гипотеза отвергается. Можно на уровне значимости 0.05 утверждать, что средний счет в первом отделении больше, чем во втором.

III. Пусть генеральные совокупности X, Y распределены нормально, причем их дисперсии неизвестны. Из этих совокупностей извлечены **зависимые** выборки одинакового объема n. Зависимость означает, что как правило мы имеем пары наблюдений (x_i, y_i), связанные между собой.

Введем следующие обозначения:

$$d_i = x_i - y_i; \qquad \bar{d} = \sum \frac{d_i}{n}; \qquad s_d = \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2/n}{n - 1}}.$$

Критерий

$$R = \frac{\bar{d}\sqrt{n}}{s_d}.$$

Двусторонняя критическая область.

$H_0 : M(X) = M(Y)$

$H_1 : M(X) \neq M(Y)$

Значение $R_{cr1} = -t^{-1}_{n-1}(1 - \alpha/2)$ — квантиль распределения Стьюдента порядка $1 - \alpha/2$ для числа степеней свободы $n - 1$. $R_{cr2} = t^{-1}_{n-1}(1 - \alpha/2)$. Если $R \in [R_{cr1}; R_{cr2}]$, т.е. $|R| < R_{cr2}$, то нулевая гипотеза принимается. В противном случае она отвергается.

Пример 7. Двумя приборами в одном и том же порядке измерены шесть деталей и получены следующие результаты измерений (в сотых долях миллиметра):

$$x_1 = 2 \qquad y_1 = 10;$$

$$x_2 = 3 \qquad y_2 = 3;$$

$$x_3 = 5 \qquad y_3 = 6;$$

$$x_4 = 6 \qquad y_4 = 1;$$

$$x_5 = 8 \qquad y_5 = 7;$$

$$x_6 = 10 \qquad y_6 = 4.$$

При уровне значимости $0,05$ установить, значимо или незначимо отличаются результаты измерений, в предположении, что они распределены нормально.

Решение.

Имеем:

$$\begin{array}{llll} x_1 = 2 & y_1 = 10; & d_1 = -8; & d_1^2 = 64; \\ x_2 = 3 & y_2 = 3; & d_2 = 0; & d_2^2 = 0; \\ x_3 = 5 & y_3 = 6; & d_3 = -1; & d_3^2 = 1; \\ x_4 = 6 & y_4 = 1; & d_4 = 5; & d_4^2 = 25; \\ x_5 = 8 & y_5 = 7; & d_5 = 1; & d_5^2 = 1; \\ x_6 = 10 & y_6 = 4; & d_6 = 6; & d_6^2 = 36; \\ n = 6 & n = 6 & \sum d_i = 3 & \sum d_i^2 = 127; \end{array}$$

$$\bar{d} = \sum d_i/6 = 3/6 = 0,5$$

$$(\sum d_i)^2 = 9$$

Тогда

$$s_d = \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2/n}{n-1}} = \sqrt{\frac{127 - 9/6}{6-1}} = \sqrt{25,1}.$$

Наблюдаемое значение критерия:

$$R = \frac{\bar{d}\sqrt{n}}{s_d} = \frac{\bar{0},5\sqrt{6}}{25,1} = 0,24.$$

Значение $R_{cr2} = t_{6-1}^{-1}(1 - 0,05/2) = 2,57$.

Поскольку $0,24 < 2,57$, нет оснований отвергнуть нулевую гипотезу. Другими словами, средние результаты измерений различаются незначительно.

14.8.3 Проверка гипотезы о дисперсии нормальной совокупности

Пусть имеется генеральная совокупность X, распределенная по нормальному закону. Предположим, что имеется выборка $x_1^*, \ldots, x_n^*$ объема n. Вычислим *исправленную* выборочную дисперсию s^2.

На основании выборочных данных проверим, равно ли значение дисперсии $D(X) = \sigma^2$ генеральной совокупности некоторому гипотетическому значению σ_0^2.

В качестве вспомогательной случайной величины будем использовать

$$R = \frac{(n-1)s^2}{\sigma_0^2}.$$

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 > \sigma_0^2$

Тогда

$$R_{cr} = \chi^2_{\alpha}(n-1).$$

Критическое значение выбирается из таблицы критических точек χ^2 распределения.

Если $R < R_{cr}$, то нет оснований отвергнуть нулевую дисперсию. Если $R > R_{cr}$, нулевая гипотеза отвергается.

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 < \sigma_0^2$

Тогда

$$R_{cr} = \chi^2_{1-\alpha}(n-1).$$

Критическое значение выбирается из таблицы критических точек χ^2 распределения.

Если $R > R_{cr}$, то нет оснований отвергнуть нулевую дисперсию. Если $R < R_{cr}$, нулевая гипотеза отвергается.

$H_0 : \sigma^2 = \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2$

Тогда

$$R_{cr1} = \chi^2_{1-\alpha/2}(n-1), \qquad R_{cr2} = \chi^2_{\alpha/2}(n-1).$$

Критические значения выбирается из таблицы критических точек χ^2 распределения.

Если $R \in (R_{cr1}; R_{cr2})$, то нет оснований отвергнуть нулевую дисперсию. В противном случае нулевая гипотеза отвергается.

Пример 8. Качество работы фасовочного автомата определяется средним квадратическим отклонением σ веса упаковок. По стандарту это отклонение не должно превышать $\sqrt{8}$ гр. ($\sigma^2 \leq 8$). Считается, что вес упаковок распределен по нормальному закону. Для 25 случайным образом выбранных упаковок оказалось, что $s^2 = 10$. Можно ли утверждать при уровне значимости 0.05, что $\sigma^2 > 8$.

Решение.

$H_0 : \sigma^2 = 8$

$H_1 : \sigma^2 > 8$

$$R = \frac{10 \cdot 24}{8} = 30, \qquad R_{cr} = \chi^2_{0,05}(24) = 36,42.$$

Имеем $R < R_{cr}$, следовательно, нет основания отвергнуть нулевую гипотезу. Стандарт не превышен.

14.8.4 Проверка гипотезы о равенстве дисперсий

Напомним, что вопрос о равенстве дисперсий двух генеральных совокупностей уже возникал при проверке гипотезы о равенстве математических ожиданий в разделе 14.8.2. Кроме того, дисперсия является одной из характеристик точности измерений. Следовательно, вопрос сравнения дисперсий (точности измерений) является актуальным.

Пусть имеются две независимые выборки объемами n_1, n_2 соответственно из генеральных совокупностей X, Y. Предположим, что найдены *исправленные* выборочные дисперсии s^2_X, s^2_Y соответственно.

Проверим, равны ли дисперсии генеральных совокупностей X, Y.

$H_0 : D(X) = D(Y)$

$H_1 : D(X) > D(Y)$

Сначала выясним, какая из исправленных выборочных дисперсий большая (обозначим индексом “Б”), а какая меньшая (обозначим индексом “М”). Соответствующие им объемы выборок также отметим индексами. Например, если $s^2_X > s^2_Y$, то $s^2_X = s^2_{\text{Б}}$, $n_1 = n_{\text{Б}}$.
Если $s^2_Y > s^2_X$, то $s^2_Y = s^2_{\text{Б}}$, $n_2 = n_{\text{Б}}$.

В качестве вспомогательной случайной величины выбирается

$$F = \frac{s^2_{\text{Б}}}{s^2_{\text{М}}}.$$

Значение $F_{cr} = F(\alpha; n_{\text{Б}} - 1; n_{\text{М}} - 1)$ выбирается из таблицы критических точек распределения Фишера– Снедекора. Порядок степеней свободы важен!

Если $F < F_{cr}$ — нет оснований отвергнуть нулевую гипотезу.

Если $F > F_{cr}$ — нулевую гипотезу отвергают.

Если проверяется

$H_0 : D(X) = D(Y)$

$H_1 : D(X) \neq D(Y)$, то

значение $F_{cr} = F(\alpha/2; n_{\text{Б}} - 1; n_{\text{М}} - 1)$ выбирается из таблицы критических точек распределения Фишера– Снедекора. Порядок степеней свободы важен! Здесь уровень $\alpha/2$!

Если $F < F_{cr}$ — нет оснований отвергнуть нулевую гипотезу.

Если $F > F_{cr}$ — нулевую гипотезу отвергают.

Замечание. Обратите внимание, что в обоих случаях имеет место правосторонняя критическая область!

Пример 9. Для сравнения точности работы двух фасовочных аппаратов различных фирм сделаны две независимые выборки объемами $n_1 = 25,\ n_2 = 30$. Оказалось, что $s_1^2 = 3,85,\ s_2^2 = 3,24$. Можно ли при уровне значимости 0.05 считать, что фасовочный аппарат второй фирмы более точен?

Решение. Аппарат более точен — это означает, что дисперсия при фасовке для него меньше.

$H_0 : D(X) = D(Y)$

$H_1 : D(X) > D(Y)$

Имеем $s_1^2 > s_2^2$, то

$$3,85 = s_{\text{Б}}^2,\ \ n_{\text{Б}} = 25, \qquad 3,24 = s_{\text{М}}^2,\ \ n_{\text{М}} = 30.$$

$$F = 3,85/3,24 = 1,19, \qquad F_{cr} = F(0,05; 24; 29) = 1,90.$$

Значит, $F < F_{cr}$, нет оснований отвергнуть нулевую гипотезу. В условиях задачи нет оснований усомниться в том, что аппараты имеют одинаковую точность.

14.9 Задачи

Задача 1

Средняя жирность молока у коров некоторого региона неизвестна. 100 коров обследуются на жирность молока. По данным обследования, средняя жирность составила 3,64 (%) при известной дисперсии $2,56$. Считаем, что жирность молока у коров распределена по нормальному закону. Проверить гипотезу о том, что средняя жирность молока у коров в данном регионе равна 3,5 % против альтернативы, что жирность больше.

Решение.

$n = 100,$ пусть $\alpha = 0,05$. Неизвестный параметр $M(X) = a$. Известно: $\sigma = \sqrt{2.56} = 1,6$; $\bar{x} = 3,64$.

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : a = 3,5$

   $H_1 : a > 3,5$

2. Задаем уровень значимости критерия $\alpha = 0,05$.

3. Вычисляем значение статистики

$$R = \frac{(3,64 - 3,5)\sqrt{100}}{1,6} = 0,875.$$

4. Находим критическую область. Значение $R_{cr} = \Phi_0^{-1}(1/2 - 0,05) = \Phi_0^{-1}(0,45) = 1,65$. Критическая область: $(1,65, \infty)$.

5. Наблюдаемое значение статистики R не попало в критическую область: $0,875 < 1,65$. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Задача 2

Суточный расход авиационного топлива (т) по данным 10 дней составил: 220, 200, 240, 190, 160, 260, 210, 200, 170, 150 т. Считая, что суточный расход имеет нормальное распределение, на уровне значимости 0.1 проверить гипотезу о том, что средний расход топлива равен 200.

Решение.

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : a = 200$

   $H_1 : a \neq 200$

2. Задаем уровень значимости критерия $\alpha = 0,1$.

3. Вычисляем значение статистики.

   Вычислим по выборке мат.ожидание и дисперсию:

$$\bar{x} = 200; \ \sigma^{2*} = 1080.$$

   Учитывая, что $n = 10$,

$$R = \frac{(200 - 200)\sqrt{9}}{\sqrt{1080}} = 0.$$

4. Находим критическую область. По таблице квантилей распределения Стьюдента находим $t_9^{-1}(0.95) = 1,833$. Критическая область: $(-\infty, -1,833) \cup (1,833, +\infty)$.

5. Наблюдаемое значение статистики R не попало в критическую область, $0 \in [-1,833; 1,833]$. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Задача 3

Качество работы фасовочного автомата определяется средним квадратическим отклонением σ веса упаковок. По стандарту это отклонение не должно превышать $\sqrt{8}$ гр. ($\sigma^2 \leq 8$). Считается, что вес упаковок распределен по нормальному закону. Для 25 случайным образом выбранных упаковок оказалось, что $s^2 = 10$. Можно ли утверждать при уровне значимости 0.05, что $\sigma^2 > 8$.

Решение.

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : \sigma^2 = 8$

   $H_1 : \sigma^2 > 8$

2. Задаем уровень значимости критерия $\alpha = 0,05$.

3. Вычисляем значение статистики.

$$R = \frac{10 \cdot 24}{8} = 30.$$

4. Находим критическую область. $R_{cr} = \chi^2_{0,95}(24) = 36,42$. Критическая область: $(36,42, +\infty)$.

5. Имеем $R < R_{cr}$. Наблюдаемое значение статистики R не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Задача 4

Пусть $X_1, ..., X_n$ – выборка из нормального распределения со средним a и единичной дисперсией. Для проверки основной гипотезы $a = 0$ против альтернативы $a = 1$ используется следующий критерий: основная гипотеза принимается, если $X_{(n)} < 3$, и отвергается в противном случае. Найти вероятности ошибок первого и второго рода ($X_{(n)}$ – наибольший элемент выборки).

Решение

$H_0 : a = 0$,

$H_1 : a = 1$.

$$\alpha = P_{H_0}\{H_0 \text{ отвергаем}\} = P_{H_0}\{X_{(n)} \geq 3\} = 1 - P_{H_0}\{X_{(n)} < 3\} =$$

$$= 1 - (P_{H_0}\{X_1 < 3\})^n = 1 - (0,5 + \Phi_0(3))^n = 1 - (0,9987)^n.$$

$$\beta = P_{H_1}\{H_0 \text{ принимаем}\} = P_{H_1}\{X_{(n)} < 3\} = (P_{H_1}\{X_1 < 3\})^n =$$

$$= (P_{H_1}\{X_1 - 1 < 2\})^n = (0,5 + \Phi_0(2))^n = (0,9772)^n.$$

Задача 5

Тестируются два расфасовочных автомата, выпускающих 100-граммовые баночки кофе. Несколько случайным образом отобранных баночек были открыты и их содержимое тщательно взвешено. Для I автомата для 8 банок выборочное среднее оказалось равным 98.71 г, а стандатное отклонение 2.38 г. Для II автомата для 12 банок получено выборочное среднее 101.87 г, стандартное отклонение 4.32 г. Для уровня значимости 0.05 проверьте следующие гипотезы:

a) Среднее значение веса баночек кофе для первого автомата совпадает со средним значением для второго. (Альтернатива $H_1 : M(X) < M(Y)$ – для первого автомата средний вес меньше).

b) Средний вес кофе для второго автомата равен 100 г. (Альтернатива $H_1 : a \neq 100$).

c) Точность аппаратов одинакова $H_0 : \sigma_X = \sigma_Y$ при альтернативе $H_1 : \sigma_X < \sigma_Y$.

d) Точность аппаратов одинакова $H_0 : \sigma_X = \sigma_Y$ при альтернативе $H_1 : \sigma_X \neq \sigma_Y$.

С какого пункта следует начать решение этой задачи?

Решение

Предположим нормальное распределение количества кофе в банке. Начнём с критериев о дисперсиях.

c)

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : \sigma_X = \sigma_Y$

   $H_1 : \sigma_Y > \sigma_X$
2. Задаем уровень значимости критерия $\alpha = 0,05$.
3. Вычисляем значение статистики.

$$F = \frac{s_2^2}{s_1^2} = 3,29.$$

4. Находим критическую область. $F^{-1}_{11,7}(0,95) = 3,6$ – квантиль распределения Фишера уровня 0,95 с 11 и 7 степенями свободы (через таблицу критических точек это $F(0,05,11,7)$). Критическая область: $(3,6,+\infty)$.
5. Имеем $F < 3,6$. Наблюдаемое значение статистики F не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

d)

Точность аппаратов одинакова $H_0 : \sigma_X = \sigma_Y$ при альтернативе $H_1 : \sigma_X \neq \sigma_Y$.

Здесь меняется только критическая область: $(-\infty; 0,266) \cup (4,7; +\infty)$. Наблюдаемое значение статистики F не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

a)

1. Считаем, что $\sigma_X = \sigma_Y$. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : M(X) = M(Y)$

   $H_1 : M(X) < M(Y)$
2. Задаем уровень значимости критерия $\alpha = 0,05$.
3. Вычисляем значение статистики.

$$R = \frac{98,71 - 101,87}{\sqrt{7 \cdot 2,38^2 + 11 \cdot 4,32^2}} \sqrt{\frac{12 \cdot 8(12+8-2)}{12+8}} = -1,88.$$

4. Находим критическую область. $t^{-1}_{8+12-2}(0,05) = -1,734$ – квантиль распределения Стьюдента уровня 0,05 с 18 степенями свободы. Критическая область: $(-\infty; -1,734)$.
5. Имеем $R < -1,734$. Наблюдаемое значение статистики R попало в критическую область. Значит, отвергаем нулевую гипотезу в пользу альтернативной при уровне значимости α.

b)

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : M(Y) = 100,$

   $H_1 : M(Y) \neq 100.$
2. Задаем уровень значимости критерия $\alpha = 0,05$.

3. Вычисляем значение статистики.

$$R = \frac{101,87 - 100}{4,32}\sqrt{12} = 1,5.$$

4. Находим критическую область. $t^{-1}_{12-1}(0,975) = 2,2$ – квантиль распределения Стьюдента уровня 0,975 с 11 степенями свободы. Критическая область: $(-\infty; -2,2) \cup (2,2; \infty)$.

5. Наблюдаемое значение статистики R не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Задача 6

Оценки по мидтерму по статистике и по финальному экзамену 10 случайно выбранных студентов курса представлены в таблице:

мидтерм	77	33	13	73	62	92	17	87	58	51
экзамен	80	20	27	65	64	69	30	74	46	59

На уровне значимости 0.1 проверьте гипотезу о равенстве средних значений оценок по мидтерму и экзамену. Найдите 90 % доверительный интервал для средних значений оценок.

Решение

Дисперсии неизвестны, выборки зависимые.

1. $H_0 : M(X) = M(Y)$

   $H_1 : M(X) \neq M(Y)$

2. Задаем уровень значимости критерия $\alpha = 0,05$.

3. Вычисляем значение статистики.

   $\bar{d} = 2,9,\ s_d = 12,62$

$$R = \frac{\bar{d}\sqrt{n}}{s_d} = 0,73$$

4. Находим критическую область. $t^{-1}_{9}(0,95) = 1,833$ – квантиль распределения Стьюдента уровня 0,95 с 9 степенями свободы. Критическая область: $(-\infty; -1,833) \cup (1,833; \infty)$.

5. Наблюдаемое значение статистики R не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Найдём доверительный интервал для средней оценки по мидтерму (X):

$\bar{X} = 56,3$, $\sigma_X^{2*} = 695,01$.

$\delta_1 = t^{-1}(\frac{1+\gamma}{2})\sqrt{\frac{\sigma_X^{2*}}{n-1}} = 1,833\sqrt{\frac{695,01}{9}} = 16,1$.

Доверительный интервал для средней оценки за мидтерм: $(40,2; 72,41)$.

Найдём доверительный интервал для средней оценки по экзамену (Y):

$\bar{Y} = 53,4$, $\sigma_Y^{2*} = 406,84$.

$\delta_1 = t^{-1}(\frac{1+\gamma}{2})\sqrt{\frac{\sigma_Y^{2*}}{n-1}} = 1,833\sqrt{\frac{406,84}{9}} = 12,32$.

Доверительный интервал для средней оценки за экзамен: $(41,1; 65,72)$.

Задача 7

Традиционный способ консервирования овощей состоит в кипячении в большом объеме воды. Предложенный новый метод парового консервирования (ПК), который, как предполагается, сохраняет больше витаминов и минералов. Десять партий бобов с разных ферм использованы для сравнения ПК и традиционного метода. Половина каждой партии была подвержена ПК, а другая — традиционному методу. Считаем, что дисперсия содержания витаминов одинакова для двух методов. Содержание витаминов на кг консервированных бобов представлено в таблице:

ПК	35	48	65	33	61	54	49	37	58	65
традиционный	33	40	55	41	62	54	40	35	59	56

Проверьте гипотезу о том, что ПК не дает преимущества по сравнению с традиционным с альтернативой, что ПК лучше на 2% уровне значимости.

1. Известно, что $\sigma_X = \sigma_Y$. Выдвигаем нулевую гипотезу и альтернативную
   $H_0 : M(X) = M(Y)$
   $H_1 : M(X) > M(Y)$
2. Задаем уровень значимости критерия $\alpha = 0,02$.
3. Вычисляем значение статистики. $\bar{X} = 50,5$, $\bar{Y} = 47,5$, $s_x^2 = 148,5$, $s_y^2 = 114,94$

$$R = \frac{50,5 - 47,5}{\sqrt{9 \cdot 148,5 + 9 \cdot 114,94}} \sqrt{\frac{10 \cdot 10(10 + 10 - 2)}{10 + 10}} = 0,58.$$

4. Находим критическую область. $t^{-1}_{10+10-2}(0,98) = 2,21$ – квантиль распределения Стьюдента уровня 0,98 с 18 степенями свободы. Критическая область: $(2,21; \infty)$.
5. Имеем $R < 2,21$. Наблюдаемое значение статистики R не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

Задача 8

Фирма А выпускает на рынок новую машину по упаковке стограммового кофе. Фирма утверждает, что новая машина точнее старой, точность которой равна $\sigma = 5$ г. Независимый эксперт для проверки этого утверждения произвел испытания, результаты которых приведены в таблице:

Модель А, вес г	102	103	97	99	100	97

a) Проверьте на 5 % уровне значимости справедливость утверждения фирмы

b) Найдите 95 % -ный доверительный интервал для точности σ новой машины;

c) Фирма В предложила свою модель упаковочной машины. По утверждению этой фирмы, её модель более точная, чем модель фирмы А. Эксперт также произвел испытания модели В и получил следующие результаты:

Модель В, вес г	100	96	97	98	101	99	99	97

Показывют ли эти результаты справедливость утверждения фирмы В?

Решение

a)

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : \sigma = 5$

   $H_1 : \sigma < 5.$
2. Задаем уровень значимости критерия $\alpha = 0,05.$
3. Вычисляем значение статистики. $\sigma^{2*} = 5,22$

$$R = \frac{n\sigma^{2*}}{25} = 1,25.$$

4. Находим критическую область. $\chi^2_{0,05}(5) = 1,145$ – квантиль распределения χ^2 уровня 0,05 с 5 степенями свободы. Критическая область: $(-\infty; 1,145)$.
5. Имеем $R > 1,145$. Наблюдаемое значение статистики R не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

b)

$$\frac{n\sigma^{2*}}{\chi^2_{0,975}(5)} < \sigma^2 < \frac{n\sigma^{2*}}{\chi^2_{0,025}(5)},$$

$$\frac{31,3}{12,8333} < \sigma^2 < \frac{31,3}{0,831}$$

$$2,44 < \sigma^2 < 37,7,$$

$$1,56 < \sigma < 6,14.$$

С вероятностью 0,95 σ лежит в интервале $(1,56; 6,14)$.

c)

1. Выдвигаем нулевую гипотезу и альтернативную

   $H_0 : \sigma_A = \sigma_B$

   $H_1 : \sigma_A > \sigma_B$
2. Задаем уровень значимости критерия $\alpha = 0,05$.
3. Вычисляем значение статистики: $s_A^2 = 6,27$, $s_B^2 = 2,84$,

$$F = \frac{s_A^2}{s_B^2} = 2,21.$$

4. Находим критическую область. $F^{-1}_{5,7}(0,95) = 3,97$ – квантиль распределения Фишера уровня 0,95 с 5 и 7 степенями свободы (через таблицу критических точек это $F(0,05, 5, 7)$). Критическая область: $(3,97, +\infty)$.
5. Имеем $F < 3,97$. Наблюдаемое значение статистики F не попало в критическую область. Значит, нет оснований отвергнуть нулевую гипотезу при уровне значимости α.

15 Парная линейная регрессия

Предположим, что требуется не только оценить степень линейной зависимости между случайными величинами X, Y по набору эмпирических данных, но и построить эту самую линейную зависимость.

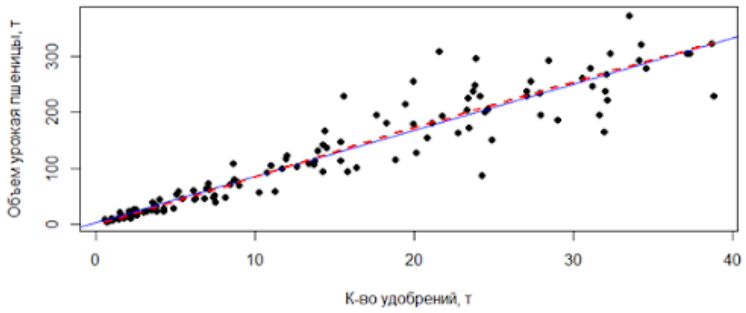
Будем называть X *фактором* или *объясняющей* переменной, а случайную величину Y — *откликом* или *объясняемой* переменной.

Предположим, что имеется парная выборка значений $(x_1, y_1), \ldots, (x_n, y_n)$. Например, x_i — затраты на рекламу в неделю, y_i — объем продаж за эту неделю. Другой пример (см. далее) — количество удобрений x_i и объем урожая y_i.

Таблица 4: *Количество удобрений и объем урожая*

X – *Количество удобрений*	X_1	X_2	$\ldots$	X_n
Y – *Объем урожая*	Y_1	Y_2	$\ldots$	Y_n

Изобразим набор эмпирических данных на графике. Получим набор точек (x_i, y_i), который называется *регрессионным облаком* или *диаграммой рассеяния*. Пусть в примере с количеством удобрений x_i и объемом урожая y_i диаграмма рассеяния имеет следующий вид:



Очевидно, что для представленного набора данных прослеживается *линейный тренд*, а именно, реальные значения y_i в некотором смысле не сильно отклоняются от идеальных значений $\hat{y}_i$, лежащих на некоторой прямой:

$$\hat{y}_i = \alpha + \beta x_i. \tag{71}$$

Колебания реальных данных y_i вокруг "идеальных" $\hat{y}_i$ можно объяснить присутствием случайной составляющей:

$$Y = \alpha + \beta X + \varepsilon, \tag{72}$$

где $\varepsilon \sim N(0, \sigma^2)$. Условия, накладываемые на "белый шум" (случайную составляющую ε) называются условиями Гаусса – Маркова.

Зависимость (71) называется теоретической парной линейной регрессией, а зависимость (72) — просто парной линейной регрессией. Прямая называется *прямой парной линейной регрессии*.

Замечание 1. Название "парная" соответствует "паре" переменных X, Y. Различают также *множественную* регрессию, для которой отклик Y соответствует множеству факторов $X_1, \ldots, X_n$.

Замечание 2. Название "линейная" соответствует линейной зависимости между X, Y. Различают также нелинейные формы трендов, такие например как параболический, гиперболический, показательный, степенной и др. Общий вид регрессионной модели:

$$Y = f(X_1, X_2, \ldots, X_n) + \varepsilon.$$

Замечание 3. Название *регрессия* является историческим. Проводилось исследование взаимосвязи роста высоких отцов и их сыновей. Оказалось, что сыновья высоких отцов в среднем **ниже** своих отцов. Зависимость роста сына Y от роста отца X назвали регрессией, имея в виду **регресс роста**. Никакой другой смысловой нагрузки регресс в данном случае не имеет.

Если мы вычислим *выборочный* коэффициент корреляции для приведенного на графике примера, мы получим значение, достаточно близкое к 1. Это также дает основания строить зависимость между Y и X в линейном виде.

Замечание 4. Если выборочный коэффициент корреляции близок к минус 1, это указывает на близкую к линейной зависимость, в которой угол наклона прямой регрессии является тупым. На практике это означает то, что при росте значений X значение Y уменьшается, но связь является линейной.

Вспомним, как считается *теоретический* коэффициент корреляции:

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \cdot \sigma_Y},$$

где $cov(X, Y)$ — ковариация величин X и Y, вычисленная по формуле:

$$cov(X, Y) = M[(X - M(X))(Y - M(Y))] = M(XY) - M(X)M(Y).$$

По выборке мы можем вычислить оценку этой величины, а именно, *выборочный коэффициент корреляции*:

$$r^*_{X,Y} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma^*_X \cdot \sigma^*_Y},$$

где $\bar{x}, \bar{y}$ — выборочные средние для $X,\ Y$ соответственно, $\sigma^*_X,\ \sigma^*_Y$ — выборочные средние квадратические отклонения и

$$\overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

Методом наименьших квадратов найдем оценки для неизвестных $\alpha,\ \beta$ в уравнении регрессии (71):

$$\alpha = \overline{y} - \beta\overline{X}$$

$$\beta = r^*_{X,Y}\frac{\sigma^*_Y}{\sigma^*_X}$$

β называется также *коэффициентом линейной регрессии.*

Пример. В таблице даны оценки 10 студентов за промежуточный мидтерм и финальный экзамен по статистике.

Таблица 5: *Количество удобрений и объем урожая*

X (мидтерм)	70	74	80	84	80	67	70	64	74	82
Y (экзамен)	87	79	88	98	96	73	83	79	91	94

Оцените линейную регрессию финальной оценки на промежуточную оценку. Каким предположительно будет результат на экзамене для студента с результатом 75 баллов по мидтерму?

Решение.

Имеем:

$$\bar{x} = \frac{1}{10}\sum x_i = 74,5, \qquad \bar{y} = \frac{1}{10}\sum y_i = 86,8; \qquad \overline{x \cdot y} = \frac{1}{10}\sum x_i y_i = 6508,7.$$

$$\sigma^2_X = 41,45, \qquad \sigma^2_Y = 60,76, \qquad \sigma_X = 6,44, \qquad \sigma_Y = 7,79.$$

$$r^*_{XY} \approx 0,84.$$

Тогда

$$\beta = 0,84 * 7,79/6,44 \approx 1,016.$$

Отсюда

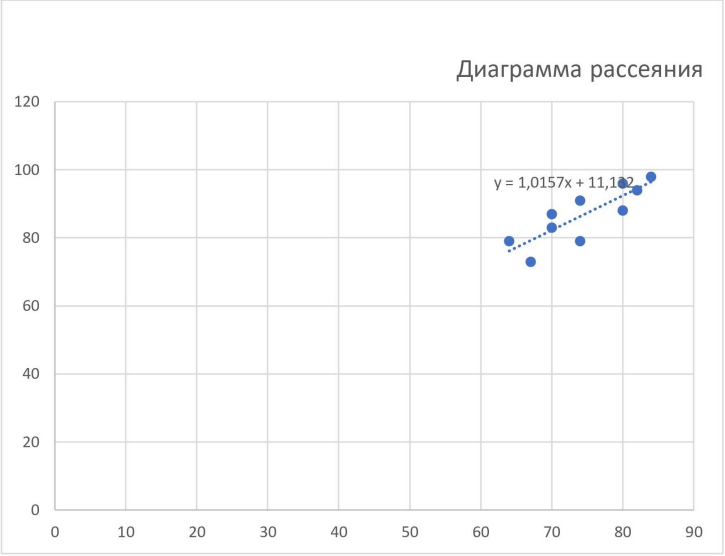
$$\alpha = \bar{y} - \beta\bar{x} = 86,8 - 1,016 \cdot 74,5 = 11,13.$$

Имеем зависимость:

$$\hat{Y} = 11,13 + 1,016X.$$

В данное уравнение можно подставить $X = 75$ и получить прогнозное $\hat{Y} = 87,33$.

График см. ниже.



Адекватность модели может быть оценена при помощи средней ошибки аппроксимации $\bar{A}$ и коэффициента детерминации R^2:

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \bar{y}_i}{|y_i|}100\%,$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}.$$

$R^2 \in [0,1]$, и чем ближе это значение к 1, тем лучше построенная модель аппроксимирует эмпирические данные.

Значение средней ошибки детерминации $\bar{A}$ до 12 % является хорошим результатом для реальных прикладных задач.

Кроме того, может быть проверена *значимость* коэффициентов регрессии и коэффициента корреляции.

15.1 Проверка значимости коэффициента корреляции

Предположим, вычислен выборочный коэффициент корреляции $r^*_{X,Y}$ и он оказался не равным нулю.

Нужно проверить на основании выборочного коэффициента корреляции предположение о коэффициенте корреляции генеральной совокупности.

$H_0 : r_{X,Y} = 0$ $H_1 : r_{X,Y} \neq 0$

Если нулевая гипотеза отвергается, то принимается альтернативная гипотеза и коэффициент корреляции оказывается *значимым*. В противном случае коэффициент корреляции равен нулю, он *незначим*.

Выберем вспомогательную случайную величину (статистику критерия):

$$t = \frac{r^*_{X,Y}}{\sqrt{1 - {r^*_{X,Y}}^2}}\sqrt{n-2},$$

где n — объем выборки.

Двусторонняя критическая область: $R_{cr1} = -t^{-1}_{n-2}(1-\alpha/2)$ — квантиль распределения Стьюдента порядка $1-\alpha/2$ для числа степеней свободы $n-2$, $R_{cr2} = t^{-1}_{n-2}(1-\alpha/2)$.

Если $|t| \geq R_{cr2}$ $(t \notin (R_{cr1}; R_{cr2})) \Rightarrow \overline{H_0} \Rightarrow$ коэффициент корреляции значим.

Иначе $H_0 \Rightarrow$ коэффициент незначим (равен нулю и случайные величины не коррелируют).

Пример. Проверим значимость коэффициента корреляции для предыдущего примера (с баллами за мидтерм и экзамен).

Решение.

Пусть $\alpha = 0,05$.

$$t \approx 4,36.$$

$$R_{cr2} = t^{-1}_8(0,975) = 2,3.$$

Тогда $t > R_{cr2}$, нулевая гипотеза отвергается. Коэффициент корреляции значим.

15.2 Задачи

Задача 1. Для каждого из 8 сортов марочного вина известны Y — число покупок в расчете на одного покупателя в год, X — покупательский рейтинг вина. Данные приведены в таблице:

Таблица 6: *Рейтинг вина и число покупок*

X (рейтинг вина)	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
Y (число покупок)	24	21	22	22	18	13	9	6

Найдите выборочный коэффициент корреляции, проверьте его значимость.

Постройте регрессию Y на X. Можно ли спрогнозировать число покупок для вина, имеющего рейтинг 2?

Задача 2. Врачей интересует зависимость времени выздоровления больного от дозы лекарства. В таблице приведены результаты наблюдения за 5 пациентами, примерно совпадающими по своим характеристикам (доза лекарства в граммах, время выздоровления в днях): Найдите выборочный коэффициент корреляции, проверьте его

Таблица 7: *Доза лекарства и время выздоровления*

X (доза лекарства)	1.2	1.0	1.5	1.2	1.4
Y (время выздоровления)	25	40	10	27	16

значимость.

Поясните знак коэффициента корреляции.

Постройте регрессию Y на X. Можно ли спрогнозировать время выздоровления больного, если он примет 1.6 грамм лекарства?