Национальный исследовательский университет
Высшая школа экономики
Московский институт электроники и математики

Департамент прикладной математики
кафедра компьютерной безопасности

Домашнее задание № 1-2 по математической статистике

Дискретное распределение: *Геометрическое распределение*
Непрервное распределение: *Равномерное II распределение*

Выполнил
Федоров Д.В.

Проверил
Богданов Д.С.

Москва 2025

Содержание

Исходные данные

Для выполнения курсовой работы выбраны следующие распределения:

- **Дискретное №5**: Геометрическое распределение с параметром $\theta = 0.4$
- **Непрерывное №2**: Нормальное распределение II с параметрами $\mu = 22.5$, $\theta = 4.0$

1 Домашнее задание 1: Характеристики вероятностных распределений

1.1 Основные понятия и определения

1. **Функция распределения случайной величины** ξ

   Функция распределения случайной величины обозначается $F_\xi(x)$ — это функция, определяющая вероятность того, что случайная величина примет значение, не превосходящее x:

   $$F_\xi(x) = P(\xi \leq x)$$

   Свойства функции распределения:

   - Функция является неубывающей;
   - $\lim_{x\to-\infty} F_\xi(x) = 0$, $\lim_{x\to+\infty} F_\xi(x) = 1$;
   - Функция непрерывна справа.

2. **Математическое ожидание** случайной величины ξ

   Математическое ожидание случайной величины ξ — это интеграл Лебега от нее по мере P:

   $$\mathbb{E}[\xi] = \int\limits_\Omega \xi \, P(d\omega)$$

   Математическое ожидание определено и корректно, если $\max\{I_+, I_-\} < \infty$, где $I_+ = \int\limits_\Omega \xi^+ \, P(d\omega)$, $\quad I_- = \int\limits_\Omega \xi^- \, P(d\omega)$ — интегралы Лебега от следующих неотрицательных случайных величин: $\xi^+ = \xi \cdot I_{\{\xi \geq 0\}}$, $\quad \xi^- = -\xi \cdot I_{\{\xi < 0\}}$

   Частные случаи:

   - Если ξ — дискретная случайная величина, принимающая значения x_i при $i \geq 1$, то:

     $$\mathbb{E}[\xi] = \sum_{i\geq 1} x_i \cdot P(\xi = x_i)$$

   - Если ξ — абсолютно непрерывная случайная величина с плотностью распределения $f_\xi(x)$, и $\int\limits_\mathbb{R} |x| f_\xi(x)\, dx < \infty$, то:

     $$\mathbb{E}[\xi] = \int\limits_\mathbb{R} x f_\xi(x)\, dx$$

Итак, математическое ожидание случайной величины ξ — это среднее значение, которого можно ожидать от ξ в среднем при бесконечном числе наблюдений:

$$\mathrm{M}[\xi] = \mathbb{E}[\xi] = \begin{cases} \sum\limits_i x_i P(\xi = x_i), & \text{для дискретной} \\ \int_{-\infty}^{\infty} x f_\xi(x)\, dx, & \text{для непрерывной} \end{cases}$$

3. **Дисперсия** случайной величины ξ

   Дисперсия случайной величины ξ — это мера разброса значений ξ относительно ее математического ожидания:

$$\mathrm{D}\xi = \sigma^2(\xi) = \mathbb{E}[(\xi - \mathbb{E}[\xi])^2] = \mathbb{E}[\xi^2] - (\mathbb{E}[\xi])^2$$

4. **Квантиль распределения уровня** γ

   Квантиль распределения уровня $\gamma \in (0, 1)$ — это такое значение x_γ, для которого:

$$F_\xi(x_\gamma) \geq \gamma$$

   При строго возрастающей функции распределения:

$$F_\xi(x_\gamma) = \gamma$$

1.2 Геометрическое распределение

1.2.1 Определение распределения

Случайная величина ξ имеет геометрическое распределение с параметром $\theta \in (0, 1)$, если:

$$P(\xi = x) = \theta(1-\theta)^{x-1}, \quad x \in \mathbb{N}, \theta \in (0, 1) \tag{1}$$

Для моего варианта $\theta = 0.4$.

1.2.2 Основные характеристики

Функция распределения. Функция распределения геометрической случайной величины:

$$F(x) = P(\xi \leq x) = \sum_{k=1}^{\lfloor x \rfloor} \theta(1-\theta)^{k-1} \tag{2}$$

Для $x \geq 1$ используем формулу суммы геометрической прогрессии:

$$F(x) = \theta \sum_{k=1}^{\lfloor x \rfloor} (1-\theta)^{k-1} = \theta \cdot \frac{1 - (1-\theta)^{\lfloor x \rfloor}}{1 - (1-\theta)} \tag{3}$$

$$= \theta \cdot \frac{1 - (1-\theta)^{\lfloor x \rfloor}}{\theta} = 1 - (1-\theta)^{\lfloor x \rfloor} \tag{4}$$

Таким образом:

$$F(x) = \begin{cases} 0, & x < 1 \\ 1 - (1-\theta)^{\lfloor x \rfloor}, & x \geq 1 \end{cases} \tag{5}$$

Для $\theta = 0.4$: $F(x) = 1 - 0.6^{\lfloor x \rfloor}$ при $x \geq 1$.

Математическое ожидание. Вычислим математическое ожидание:

$$\mathbb{E}\xi = \sum_{x=1}^{\infty} x \cdot \theta(1-\theta)^{x-1} = \theta \sum_{x=1}^{\infty} x(1-\theta)^{x-1} \tag{6}$$

Используем формулу $\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2}$ при $|q| < 1$:

$$\mathbb{E}\xi = \theta \cdot \frac{1}{(1-(1-\theta))^2} = \theta \cdot \frac{1}{\theta^2} = \frac{1}{\theta} \tag{7}$$

Для $\theta = 0.4$: $\boxed{\mathbb{E}\xi = \frac{1}{0.4} = 2.5}$.

Дисперсия. Для вычисления дисперсии сначала найдем $\mathbb{E}\xi^2$:

$$\mathbb{E}\xi^2 = \sum_{x=1}^{\infty} x^2\theta(1-\theta)^{x-1} = \theta \sum_{x=1}^{\infty} x^2(1-\theta)^{x-1} \tag{8}$$

Используя формулу $\sum_{x=1}^{\infty} x^2q^{x-1} = \frac{1+q}{(1-q)^3}$:

$$\mathbb{E}\xi^2 = \theta \cdot \frac{1+(1-\theta)}{\theta^3} = \frac{2-\theta}{\theta^2} \tag{9}$$

Дисперсия:

$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \frac{2-\theta}{\theta^2} - \frac{1}{\theta^2} = \frac{1-\theta}{\theta^2} \tag{10}$$

Для $\theta = 0.4$: $\boxed{\mathbb{D}\xi = \frac{0.6}{0.16} = 3.75}$.

Квантиль уровня γ. Квантиль уровня $\gamma \in (0,1)$ определяется из условия:

$$F(x_\gamma) = \gamma \Rightarrow 1-(1-\theta)^{x_\gamma} = \gamma \tag{11}$$

Откуда:

$$x_\gamma = \left\lceil \frac{\ln(1-\gamma)}{\ln(1-\theta)} \right\rceil \tag{12}$$

где $\lceil \cdot \rceil$ — функция округления вверх.
Для $\theta = 0.4$ и $\gamma = 0.5$ (медиана):

$$x_{0.5} = \left\lceil \frac{\ln(0.5)}{\ln(0.6)} \right\rceil = \left\lceil \frac{-0.693}{-0.511} \right\rceil = \lceil 1.36 \rceil = 2 \tag{13}$$

Для $\gamma = 0.95$:

$$x_{0.95} = \left\lceil \frac{\ln(0.05)}{\ln(0.6)} \right\rceil = \left\lceil \frac{-2.996}{-0.511} \right\rceil = \lceil 5.86 \rceil = 6 \tag{14}$$

1.2.3 Интерпретация распределения

Геометрическое распределение моделирует число испытаний Бернулли до первого успеха включительно. Это одно из важнейших дискретных распределений, описывающее процессы ожидания.

Пример 1. Передача данных по ненадежному каналу.

Рассмотрим процесс отправки пакетов данных по ненадежному каналу связи. Пусть вероятность успешной доставки одного пакета равна $\theta = 0.4$. Каждая попытка отправки независима от предыдущих. Тогда случайная величина ξ — число попыток до первой успешной доставки — имеет геометрическое распределение с параметром $\theta = 0.4$.

Математическое ожидание $\mathbb{E}\xi = 2.5$ означает, что в среднем потребуется 2.5 попытки для успешной доставки пакета. С вероятностью $P(\xi = 1) = 0.4$ пакет доставлен с первой попытки, с вероятностью $P(\xi \leq 3) = 1 - 0.6^3 = 0.784$ пакет доставлен не более чем за 3 попытки.

Пример 2. Обслуживание в колл-центре.

В колл-центре вероятность того, что клиент дождется ответа оператора (не повесит трубку), составляет $\theta = 0.4$. Число звонков до первого успешно обработанного имеет геометрическое распределение. Медиана распределения равна 2 ($\gamma = 0.5$), что означает, что с вероятностью 50% первый успешный контакт произойдет не позже второго звонка.

Пример 3. Контроль качества.

При производстве изделий доля годных составляет $\theta = 0.4$. Инспектор проверяет изделия по очереди до обнаружения первого годного. Число проверенных изделий имеет геометрическое распределение. Вероятность того, что придется проверить более 6 изделий, составляет $P(\xi > 6) = 0.6^6 \approx 0.047$ или около 5%.

1.2.4 Соотношения между распределениями

1. **Поведение при $\theta \to 0$.**

   Среднее число испытаний до первого успеха:

   $$\mathbb{E}\xi = \frac{1}{\theta} \tag{15}$$

   При $\theta \to 0$ (вероятность успеха стремится к нулю):

   $$\lim_{\theta \to 0} \mathbb{E}\xi = \lim_{\theta \to 0} \frac{1}{\theta} = +\infty \tag{16}$$

   Интерпретация: Если успех очень маловероятен, то в среднем потребуется очень много попыток для его достижения. Например, при $\theta = 0.01$ среднее число попыток $\mathbb{E}\xi = 100$, при $\theta = 0.001$ уже $\mathbb{E}\xi = 1000$.

2. **Свойство отсутствия памяти.**

   Геометрическое распределение обладает уникальным свойством:

   $$P(\xi > m + n \mid \xi > m) = P(\xi > n), \quad \forall m, n \in \mathbb{N} \tag{17}$$

   Доказательство:

   По определению условной вероятности:

   $$P(\xi > m + n \mid \xi > m) = \frac{P(\xi > m + n, \xi > m)}{P(\xi > m)} = \frac{P(\xi > m + n)}{P(\xi > m)} \tag{18}$$

Так как событие $\{\xi > m+n\} \subset \{\xi > m\}$, то $P(\xi > m+n, \xi > m) = P(\xi > m+n)$.

Вычислим вероятности:

$$P(\xi > k) = \sum_{j=k+1}^{\infty} \theta(1-\theta)^{j-1} = \theta(1-\theta)^k \sum_{j=0}^{\infty}(1-\theta)^j \tag{19}$$

$$= \theta(1-\theta)^k \cdot \frac{1}{1-(1-\theta)} = \theta(1-\theta)^k \cdot \frac{1}{\theta} = (1-\theta)^k \tag{20}$$

Тогда:

$$P(\xi > m+n \mid \xi > m) = \frac{(1-\theta)^{m+n}}{(1-\theta)^m} = (1-\theta)^n = P(\xi > n) \tag{21}$$

Интерпретация: Если уже произошло m неудачных попыток, вероятность того, что потребуется еще более n попыток, такая же, как если бы мы начинали с начала. Прошлые неудачи не влияют на будущее.

Пример: Бросаем монету до первого выпадения орла. Если уже выпало 10 решек подряд, вероятность того, что потребуется еще хотя бы 5 бросков, равна $P(\xi > 5) = 0.5^5$, независимо от предыстории.

3. **Связь с отрицательным биномиальным распределением.**

Пусть $\xi_1, \xi_2, \ldots, \xi_k$ — независимые случайные величины, каждая с геометрическим распределением $\mathrm{Geom}(\theta)$. Тогда их сумма:

$$S_k = \sum_{i=1}^{k} \xi_i \tag{22}$$

имеет отрицательное биномиальное распределение $\mathrm{NB}(k, \theta)$ с параметрами k (число успехов) и θ (вероятность успеха).

Обоснование:

Случайная величина S_k представляет общее число испытаний до достижения k-го успеха. Функция вероятности:

$$P(S_k = n) = \binom{n-1}{k-1}\theta^k(1-\theta)^{n-k}, \quad n = k, k+1, k+2, \ldots \tag{23}$$

Интерпретация: Для k-го успеха нужно, чтобы на $(n-1)$-м испытании был $(k-1)$-й успех, а на n-м испытании был k-й успех.

Частный случай: При $k = 1$ отрицательное биномиальное распределение совпадает с геометрическим.

Математическое ожидание и дисперсия суммы:

$$\mathbb{E}S_k = k \cdot \mathbb{E}\xi_1 = \frac{k}{\theta} \tag{24}$$

$$\mathbb{D}S_k = k \cdot \mathbb{D}\xi_1 = \frac{k(1-\theta)}{\theta^2} \tag{25}$$

4. **Дискретный аналог экспоненциального распределения.**

Геометрическое распределение является дискретным аналогом экспоненциального распределения в следующем смысле:

- Оба описывают время до первого события в последовательности независимых испытаний
- Оба обладают свойством отсутствия памяти
- При дискретизации времени экспоненциальное распределение переходит в геометрическое

Формальная связь:

Пусть $\eta \sim \text{Exp}(\lambda)$ — экспоненциально распределенная случайная величина с плотностью $f(t) = \lambda e^{-\lambda t}$, $t > 0$.

Рассмотрим дискретизацию времени с шагом Δt. Вероятность события в малом интервале $[0, \Delta t]$ равна:

$$p = P(\eta \leq \Delta t) = 1 - e^{-\lambda \Delta t} \approx \lambda \Delta t \quad \text{при малых } \Delta t \tag{26}$$

Число интервалов до первого события имеет геометрическое распределение с параметром $\theta = p$.

При $\Delta t \to 0$ и $n \to \infty$ так, что $n\Delta t \to t$ (фиксированное время):

$$P(\xi > n) = (1 - \theta)^n = (1 - \lambda \Delta t)^n \to e^{-\lambda t} = P(\eta > t) \tag{27}$$

Свойство отсутствия памяти для экспоненциального распределения:

$$P(\eta > s + t \mid \eta > s) = P(\eta > t) \tag{28}$$

Это свойство объединяет оба распределения и делает их естественными моделями для процессов без последействия.

1.2.5 Моделирование геометрического распределения

Пусть имеется генератор равномерно распределенных на $[0, 1]$ случайных величин $U \sim \mathcal{U}[0, 1]$.

Метод 1: метод испытаний (прямое моделирование).

Этот метод непосредственно реализует определение геометрического распределения:

1. Установить счетчик $k = 1$
2. Генерировать $U_k \sim \mathcal{U}[0, 1]$
3. Если $U_k \leq \theta$, то вернуть $\xi = k$ (первый успех произошел на k-м шаге)
4. Иначе увеличить k на 1 и перейти к шагу 2

Псевдокод:

```
function generate_geometric_trials(theta):
    k = 1
    while true:
        U = random_uniform(0, 1)
        if U <= theta:
            return k
        k = k + 1
```

Метод 2: обратное преобразование.
Из определения квантиля:

$$F(\xi) = U \Rightarrow 1 - (1-\theta)^{\xi} = U \Rightarrow \xi = \left\lceil \frac{\ln(1-U)}{\ln(1-\theta)} \right\rceil \tag{29}$$

Алгоритм:

1. Сгенерировать $U \sim \mathcal{U}[0,1]$
2. Вычислить $\xi = \left\lceil \frac{\ln(1-U)}{\ln(1-\theta)} \right\rceil = \left\lceil \frac{\ln(1-U)}{\ln(0.6)} \right\rceil$
3. Значение ξ имеет геометрическое распределение с параметром $\theta = 0.4$

Псевдокод:

```
function generate_geometric(theta):
    U = random_uniform(0, 1)
    X = ceil(log(1 - U) / log(1 - theta))
    return X
```

Оба метода дают одинаковое распределение. Метод обратного преобразования более эффективен вычислительно, так как требует только один вызов генератора случайных чисел.

1.3 Нормальное распределение II

1.3.1 Определение распределения

Случайная величина ξ имеет нормальное распределение с параметрами $\mu \in \mathbb{R}$ и $\theta > 0$, если ее плотность:

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\theta^2}\right), \quad x \in \mathbb{R} \tag{30}$$

где μ — параметр сдвига (известен), θ — неизвестный параметр масштаба (среднеквадратическое отклонение).
Для нашего варианта: $\mu = 22.5$, $\theta = 4.0$.
Обозначение: $\xi \sim \mathcal{N}(\mu, \theta^2) = \mathcal{N}(22.5, 16)$.

1.3.2 Основные характеристики

Функция распределения. Функция распределения нормального распределения:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\theta^2}\right) dt \tag{31}$$

Делая замену $z = \frac{t-\mu}{\theta}$:

$$F(x) = \int_{-\infty}^{\frac{x-\mu}{\theta}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi\left(\frac{x-\mu}{\theta}\right) \tag{32}$$

где $\Phi(z)$ — функция распределения стандартного нормального распределения $\mathcal{N}(0,1)$.
Для наших параметров:

$$F(x) = \Phi\left(\frac{x-22.5}{4}\right) \tag{33}$$

Математическое ожидание. Для нормального распределения $\mathcal{N}(\mu, \theta^2)$:

$$\mathbb{E}\xi = \mu \tag{34}$$

Это следует из симметрии плотности относительно точки μ.
Для нашего варианта: $\boxed{\mathbb{E}\xi = 22.5}$.

Дисперсия. Для нормального распределения:

$$\mathbb{D}\xi = \theta^2 \tag{35}$$

Это можно показать через вычисление интеграла:

$$\begin{aligned}
D_\xi &= \mathrm{E}[(X-\mu)^2] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\theta^2}\right) dx \\
&\xrightarrow[z=\frac{x-\mu}{\theta}]{x=\mu+\theta z,\ dx=\theta\,dz} \theta^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz \\
&= \theta^2\,\mathrm{E}[Z^2], \quad Z \sim N(0,1) \\
&= \theta^2 \cdot 1 \\
&= \theta^2.
\end{aligned}$$

Для нашего варианта: $\boxed{\mathbb{D}\xi = 16}$.
Среднеквадратическое отклонение: $\sigma = \theta = 4.0$.

Квантиль уровня γ. Квантиль уровня γ определяется из уравнения:

$$F(x_\gamma) = \gamma \Rightarrow \Phi\left(\frac{x_\gamma - \mu}{\theta}\right) = \gamma \tag{36}$$

Откуда:

$$x_\gamma = \mu + \theta \cdot \Phi^{-1}(\gamma) = 22.5 + 4 \cdot z_\gamma \tag{37}$$

где $z_\gamma = \Phi^{-1}(\gamma)$ — квантиль стандартного нормального распределения. Найти примерное значение Φ^{-1} можно по таблицам: Z-таблицы правая, левая, отрицательная

Примеры:

- Медиана ($\gamma = 0.5$): $x_{0.5} = 22.5 + 4 \cdot 0 = 22.5$ (так как $\Phi^{-1}(0.5) = 0$)
- Квантиль 0.95: $x_{0.95} \approx 22.5 + 4 \cdot 1.645 \approx 29.08$
- Квантиль 0.05: $x_{0.05} \approx 22.5 + 4 \cdot (-1.645) \approx 15.92$

Интервал $[15.92, 29.08]$ содержит 90% вероятностной массы распределения.

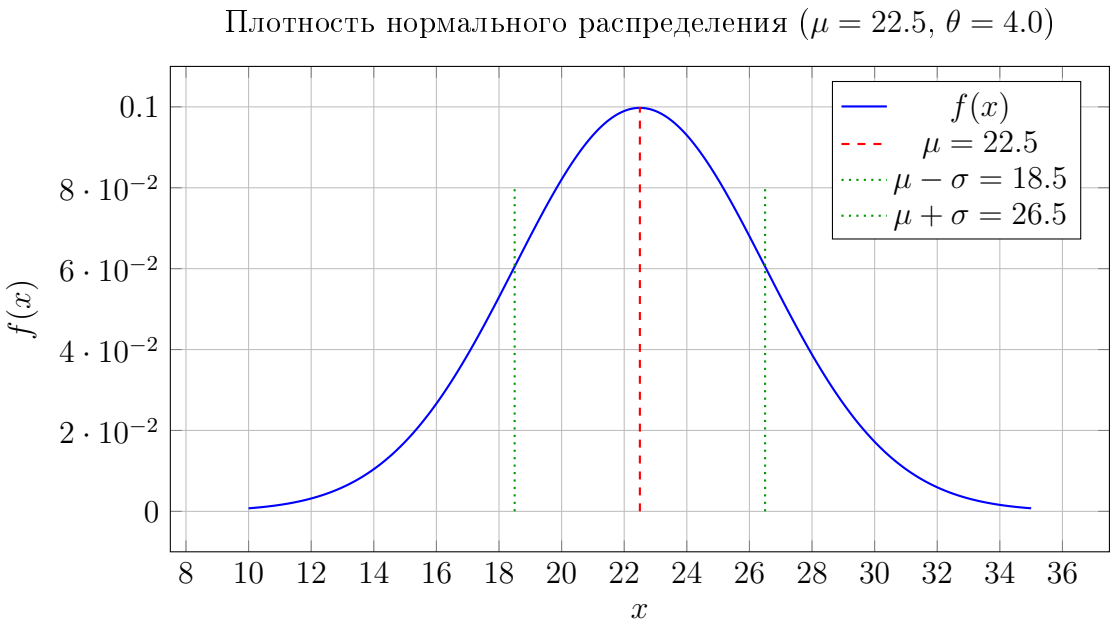

Рис. 1: Плотность нормального распределения с $\mu = 22.5$, $\theta = 4.0$

Правило трех сигм.

$$X \sim N(\mu, \theta^2), \qquad Z = \frac{X - \mu}{\theta} \sim N(0, 1)$$

$$P(\mu - k\theta < X < \mu + k\theta) = P(-k < Z < k) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1.$$

Для k=1,2,3:

$$\Phi(1) = 0.84134 \quad \Rightarrow \quad 2\Phi(1) - 1 = 0.68268 \approx 0.683$$
$$\Phi(2) = 0.97724 \quad \Rightarrow \quad 2\Phi(2) - 1 = 0.95449 \approx 0.954$$
$$\Phi(3) = 0.99865 \quad \Rightarrow \quad 2\Phi(3) - 1 = 0.99730 \approx 0.997$$
$$k = 1: \ [\mu - \theta, \mu + \theta] = [22.5 - 4, \ 22.5 + 4] = [18.5, \ 26.5]$$
$$k = 2: \ [14.5, \ 30.5] \quad (\approx 95.4\%)$$
$$k = 3: \ [10.5, \ 34.5] \quad (\approx 99.73\%)$$

Для нормального распределения выполняются следующие соотношения:

- $P(\mu - \theta < \xi < \mu + \theta) \approx 0.683$ — около 68% значений в интервале $[18.5, 26.5]$
- $P(\mu - 2\theta < \xi < \mu + 2\theta) \approx 0.954$ — около 95% значений в интервале $[14.5, 30.5]$
- $P(\mu - 3\theta < \xi < \mu + 3\theta) \approx 0.997$ — около 99.7% значений в интервале $[10.5, 34.5]$

1.3.3 Интерпретация распределения

Нормальное распределение — одно из важнейших распределений в теории вероятностей и статистике. Его центральная роль обусловлена центральной предельной теоремой: сумма большого числа независимых случайных величин приближенно нормальна.

Пример 1. Погрешности измерений.

Пусть измерительный прибор имеет систематическую погрешность $\mu = 22.5$ единиц и случайную компоненту погрешности со среднеквадратическим отклонением $\theta = 4.0$ единиц. Если истинное значение измеряемой величины равно 0, то результат измерения будет иметь распределение $\mathcal{N}(22.5, 16)$.

Систематическая погрешность означает, что прибор постоянно завышает показания на 22.5 единиц, а случайная компонента приводит к дополнительному разбросу измерений со стандартным отклонением 4.0 единиц.

Для отдельного измерения X вероятность попадания в различные интервалы вычисляется следующим образом:

- С вероятностью 68% результат попадет в интервал $[\mu-\theta, \mu+\theta] = [18.5, 26.5]$ (правило одной сигмы)
- С вероятностью 95% результат попадет в интервал $[\mu - 2\theta, \mu + 2\theta] = [14.5, 30.5]$ (правило двух сигм)
- С вероятностью 99.7% результат попадет в интервал $[\mu - 3\theta, \mu + 3\theta] = [10.5, 34.5]$ (правило трех сигм)

Для уменьшения случайной погрешности можно провести n независимых измерений $X_1, X_2, \ldots, X_n$ и вычислить их среднее:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{38}$$

Так как измерения независимы и одинаково распределены по закону $\mathcal{N}(22.5, 16)$, то среднее имеет распределение:

$$\bar{X} \sim \mathcal{N}\left(22.5, \frac{16}{n}\right) \tag{39}$$

То есть математическое ожидание остается равным $\mu = 22.5$ (систематическая ошибка не уменьшается), но дисперсия уменьшается в n раз, а стандартное отклонение уменьшается в $\sqrt{n}$ раз:

$$\sigma_{\bar{X}} = \frac{\theta}{\sqrt{n}} = \frac{4.0}{\sqrt{n}} \tag{40}$$

Числовые примеры:

- При $n = 4$ измерениях: $\bar{X} \sim \mathcal{N}(22.5, 4)$, стандартное отклонение $\sigma_{\bar{X}} = 2.0$, 95% интервал: $[18.58, 26.42]$
- При $n = 16$ измерениях: $\bar{X} \sim \mathcal{N}(22.5, 1)$, стандартное отклонение $\sigma_{\bar{X}} = 1.0$, 95% интервал: $[20.54, 24.46]$
- При $n = 100$ измерениях: $\bar{X} \sim \mathcal{N}(22.5, 0.16)$, стандартное отклонение $\sigma_{\bar{X}} = 0.4$, 95% интервал: $[21.72, 23.28]$

Таким образом, усреднение измерений позволяет существенно повысить точность, уменьшая влияние случайной погрешности, однако систематическая погрешность при этом не устраняется и требует калибровки прибора.

Пример 2. Результаты тестирования.

Баллы студентов на экзамене часто распределены нормально. Если средний балл $\mu = 22.5$ из 30 возможных, а стандартное отклонение $\theta = 4.0$, то:

- 68% студентов получают балл в диапазоне $[18.5, 26.5]$
- 95% студентов получают балл в диапазоне $[14.5, 30.5]$
- Балл выше 26.5 получают примерно 16% студентов

- Балл ниже 18.5 получают примерно 16% студентов

Пример 3. Время реакции.
Время реакции человека на визуальный стимул может быть описано нормальным распределением. Если среднее время реакции $\mu = 22.5$ сотых секунды (225 мс) со стандартным отклонением $\theta = 4.0$ сотых секунды (40 мс), то нормальное распределение хорошо описывает вариацию времени реакции в повторных испытаниях.

1.3.4 Соотношения между распределениями

1. **Стандартизация.**

   Если $\xi \sim \mathcal{N}(\mu, \theta^2)$, то стандартизованная величина:

   $$Z = \frac{\xi - \mu}{\theta} \sim \mathcal{N}(0, 1) \tag{41}$$

   имеет стандартное нормальное распределение с параметрами 0 (математическое ожидание) и 1 (дисперсия).

   Доказательство:

   Вычислим математическое ожидание:

   $$\mathbb{E}Z = \mathbb{E}\left[\frac{\xi - \mu}{\theta}\right] = \frac{1}{\theta}(\mathbb{E}\xi - \mu) = \frac{1}{\theta}(\mu - \mu) = 0 \tag{42}$$

   Вычислим дисперсию:

   $$\mathbb{D}Z = \mathbb{D}\left[\frac{\xi - \mu}{\theta}\right] = \frac{1}{\theta^2}\mathbb{D}\xi = \frac{\theta^2}{\theta^2} = 1 \tag{43}$$

   Для нормального распределения линейное преобразование сохраняет нормальность, следовательно $Z \sim \mathcal{N}(0, 1)$.

   Применение: Стандартизация позволяет использовать табличные значения функции распределения стандартного нормального распределения $\Phi(z)$ для вычисления вероятностей:

   $$P(\xi \leq x) = P\left(\frac{\xi - \mu}{\theta} \leq \frac{x - \mu}{\theta}\right) = \Phi\left(\frac{x - \mu}{\theta}\right) \tag{44}$$

   Пример: Для $\xi \sim \mathcal{N}(22.5, 16)$ вероятность $P(\xi \leq 26.5)$:

   $$P(\xi \leq 26.5) = \Phi\left(\frac{26.5 - 22.5}{4}\right) = \Phi(1) \approx 0.8413 \tag{45}$$

2. **Линейное преобразование.**

   Если $\xi \sim \mathcal{N}(\mu, \theta^2)$, то для любых констант $a \neq 0$ и b случайная величина:

   $$\eta = a\xi + b \sim \mathcal{N}(a\mu + b, a^2\theta^2) \tag{46}$$

   Доказательство:

   Математическое ожидание:

   $$\mathbb{E}\eta = \mathbb{E}[a\xi + b] = a\mathbb{E}\xi + b = a\mu + b \tag{47}$$

Дисперсия:

$$\mathbb{D}\eta = \mathbb{D}[a\xi + b] = a^2\mathbb{D}\xi = a^2\theta^2 \tag{48}$$

Нормальность сохраняется при линейных преобразованиях, что следует из свойств характеристической функции нормального распределения.

Обратное утверждение: Любое нормальное распределение $\mathcal{N}(\mu, \sigma^2)$ может быть получено из стандартного линейным преобразованием:

$$\xi = \mu + \sigma Z, \quad Z \sim \mathcal{N}(0,1) \Rightarrow \xi \sim \mathcal{N}(\mu, \sigma^2) \tag{49}$$

Пример: Если температура в градусах Цельсия $T_C \sim \mathcal{N}(22.5, 16)$, то в градусах Фаренгейта:

$$T_F = 1.8 \cdot T_C + 32 \sim \mathcal{N}(1.8 \cdot 22.5 + 32, 1.8^2 \cdot 16) = \mathcal{N}(72.5, 51.84) \tag{50}$$

3. **Сумма независимых нормальных величин.**

Если $\xi_1 \sim \mathcal{N}(\mu_1, \theta_1^2)$ и $\xi_2 \sim \mathcal{N}(\mu_2, \theta_2^2)$ независимы, то их сумма также нормальна:

$$\xi_1 + \xi_2 \sim \mathcal{N}(\mu_1 + \mu_2, \theta_1^2 + \theta_2^2) \tag{51}$$

Обобщение: Для линейной комбинации n независимых нормальных величин:

$$\sum_{i=1}^{n} a_i\xi_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\theta_i^2\right) \tag{52}$$

где $\xi_i \sim \mathcal{N}(\mu_i, \theta_i^2)$ независимы.

Доказательство (через характеристические функции):

Характеристическая функция $\xi \sim \mathcal{N}(\mu, \theta^2)$:

$$\varphi_\xi(t) = \exp\left(i\mu t - \frac{\theta^2 t^2}{2}\right) \tag{53}$$

Для суммы независимых величин характеристические функции перемножаются:

$$\varphi_{\xi_1+\xi_2}(t) = \varphi_{\xi_1}(t) \cdot \varphi_{\xi_2}(t) \tag{54}$$

$$= \exp\left(i\mu_1 t - \frac{\theta_1^2 t^2}{2}\right) \cdot \exp\left(i\mu_2 t - \frac{\theta_2^2 t^2}{2}\right) \tag{55}$$

$$= \exp\left(i(\mu_1 + \mu_2)t - \frac{(\theta_1^2 + \theta_2^2)t^2}{2}\right) \tag{56}$$

Это характеристическая функция $\mathcal{N}(\mu_1 + \mu_2, \theta_1^2 + \theta_2^2)$.

Следствие: Для n независимых одинаково распределенных величин $\xi_i \sim \mathcal{N}(\mu, \theta^2)$:

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i \sim \mathcal{N}\left(\mu, \frac{\theta^2}{n}\right) \tag{57}$$

Важное замечание: Это свойство уникально для нормального распределения — только сумма нормальных величин остается нормальной.

4. **Центральная предельная теорема (ЦПТ).**

   Пусть $\xi_1, \xi_2, \ldots, \xi_n$ — независимые одинаково распределенные случайные величины с $\mathbb{E}\xi_i = \mu$ и $\mathbb{D}\xi_i = \sigma^2 < \infty$. Тогда:

$$\frac{\sum_{i=1}^{n} \xi_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{\xi}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{при } n \to \infty \tag{58}$$

   где $\xrightarrow{d}$ обозначает сходимость по распределению.

   Эквивалентная формулировка:

$$\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{при } n \to \infty \tag{59}$$

   или

$$\sum_{i=1}^{n} \xi_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2) \quad \text{при } n \to \infty \tag{60}$$

   Содержательный смысл: Сумма (или среднее) большого числа независимых случайных величин имеет приближенно нормальное распределение, независимо от исходного распределения слагаемых (при выполнении условий теоремы).

   Условия применимости (условия Линдеберга):

   - Слагаемые независимы
   - Дисперсии конечны
   - Выполнено условие Линдеберга: вклад каждого слагаемого в общую дисперсию пренебрежимо мал

   Практическое правило: Для большинства распределений хорошее приближение достигается при $n \geq 30$. Для симметричных распределений достаточно $n \geq 10$.

   Пример: Сумма показаний 100 независимых игральных костей приближенно нормальна:

$$S_{100} = \sum_{i=1}^{100} X_i \approx \mathcal{N}\left(100 \cdot 3.5, 100 \cdot \frac{35}{12}\right) = \mathcal{N}(350, 291.67) \tag{61}$$

   где $\mathbb{E}X_i = 3.5$, $\mathbb{D}X_i = 35/12 \approx 2.917$.

   Значение ЦПТ: Объясняет, почему нормальное распределение встречается повсеместно в природе и технике — большинство наблюдаемых величин являются результатом суммирования многих независимых случайных факторов.

1.3.5 Моделирование нормального распределения

Для моделирования нормально распределенной случайной величины $\xi \sim \mathcal{N}(\mu, \theta^2)$ существует несколько методов.

Метод Бокса-Мюллера.

Это точный и эффективный метод, основанный на преобразовании двух независимых равномерных величин. Wiki: Метод Бокса-Мюллера

Пусть $U_1, U_2 \sim \mathcal{U}[0,1]$ — независимые равномерно распределенные случайные величины. Тогда:

$$Z_1 = \sqrt{-2\ln U_1}\cos(2\pi U_2) \tag{62}$$

$$Z_2 = \sqrt{-2\ln U_1}\sin(2\pi U_2) \tag{63}$$

являются независимыми стандартными нормальными величинами $Z_1, Z_2 \sim \mathcal{N}(0,1)$. Для получения $\xi \sim \mathcal{N}(\mu, \theta^2)$ используем линейное преобразование:

$$\xi = \mu + \theta \cdot Z_1 = 22.5 + 4.0 \cdot Z_1 \tag{64}$$

Алгоритм:

1. Сгенерировать две независимые величины $U_1, U_2 \sim \mathcal{U}[0,1]$
2. Вычислить $Z_1 = \sqrt{-2 \ln U_1} \cos(2\pi U_2)$
3. Вычислить $\xi = 22.5 + 4.0 \cdot Z_1$
4. Значение ξ имеет распределение $\mathcal{N}(22.5, 16)$

Псевдокод:

```
function generate_normal(mu, theta):
    U1 = random_uniform(0, 1)
    U2 = random_uniform(0, 1)
    Z = sqrt(-2 * log(U1)) * cos(2 * pi * U2)
    X = mu + theta * Z
    return X
```

Метод Бокса-Мюллера предпочтительный, так как дает точное распределение и он в целом более эффективен. Например, есть метод через ЦПТ с приближением или усовершенствованные методы Бокса-Мюллера.

2 Домашнее задание 2. Основные понятия математической статистики

2.1 Генерация выборок

Для каждого из распределений сгенерировано по 5 массивов из 50000 значений и сделаны выбороки следующих объемов:

$$n \in \{5, 10, 100, 200, 400, 600, 800, 1000, 1001, 50000\} \tag{65}$$

Выборки 1001 и 50000 используются для некоторых вычислений, чтобы показать конкретную зависимость. Например, для статистики Смирнова сравнение выборок 1000 и 1001 дает ожидаемо маленькое D_{mn}, а для ЭФР и гистограммы значение 50000 подтверждает еще сильнее закон больших чисел. По заданию требуются лишь выборки $\{5, 10, 100, 200, 400, 600, 800, 1000\}$, соответственно основная часть расчетов выполнена только для них.

Генерация производилась с использованием описанных выше алгоритмов:

- Для геометрического распределения — метод обратного преобразования
- Для нормального распределения — метод Бокса-Мюллера

Для генерации используются сиды **{500, 501, 502, 503, 504}** для генератора *np.random.uniform* для каждой из 5 выборок соответственно.

2.2 Эмпирическая функция распределения

2.2.1 Определение и свойства

Для выборки $X = (X_1, \ldots, X_n)$ эмпирическая функция распределения (ЭФР) определяется как:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \tag{66}$$

Свойства ЭФР:

- $\hat{F}_n(x)$ принимает значения $\{0, 1/n, 2/n, \ldots, 1\}$
- $\hat{F}_n(x)$ является ступенчатой функцией со скачками в точках наблюдений
- Для каждого фиксированного x величина $\hat{F}_n(x)$ является несмещенной оценкой $F(x)$: $\mathbb{E}\hat{F}_n(x) = F(x)$
- $\hat{F}_n(x)$ состоятельно сходится к $F(x)$: $\hat{F}_n(x) \xrightarrow{P} F(x)$ при $n \to \infty$

2.2.2 Построение графиков ЭФР

Покажем два вида графиков эмпирической функции распределения. Сначала для одного сида построим график с разными объемами выборки. Потом для каждого объема выборки n построим график с разными сидами. На каждый график наносим теоретическую функцию распределения.

2.2.3 График для геометрического распределения (выборки из одного сгенерированного массива seed=500):

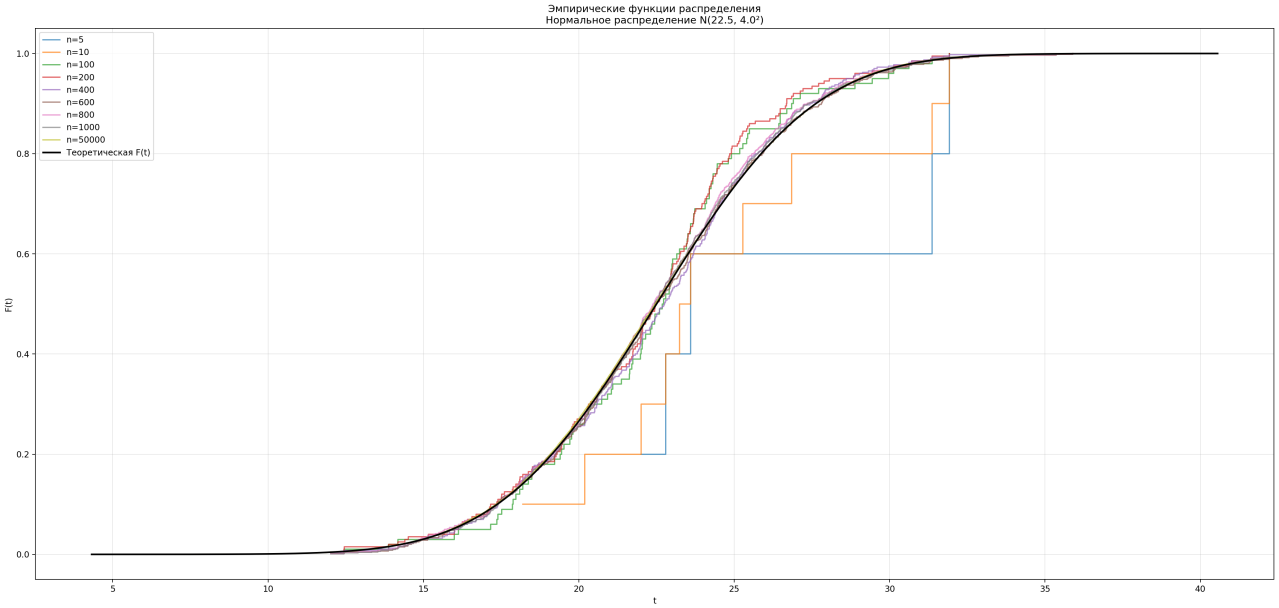

Рис. 2: Эмпирическая функция распределения для геометрического распределения

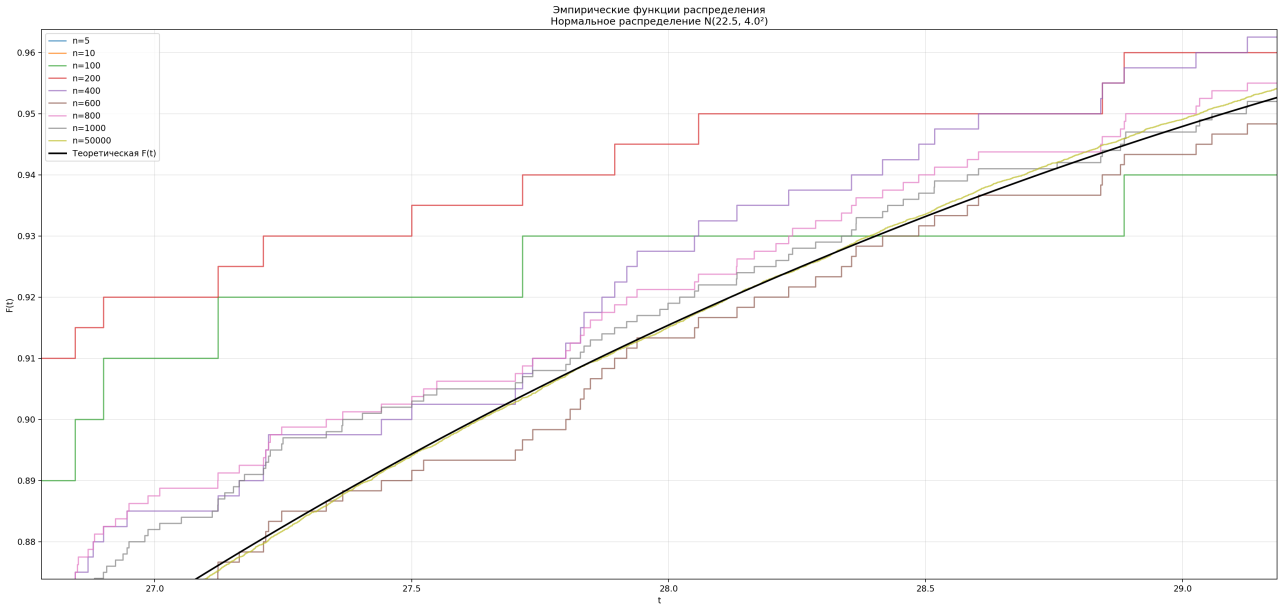

Рис. 3: Увеличенный фрагмент ЭФР геометрического распределения

2.2.4 График для нормального распределения (выборки из одного сгенерированного массива seed=500):

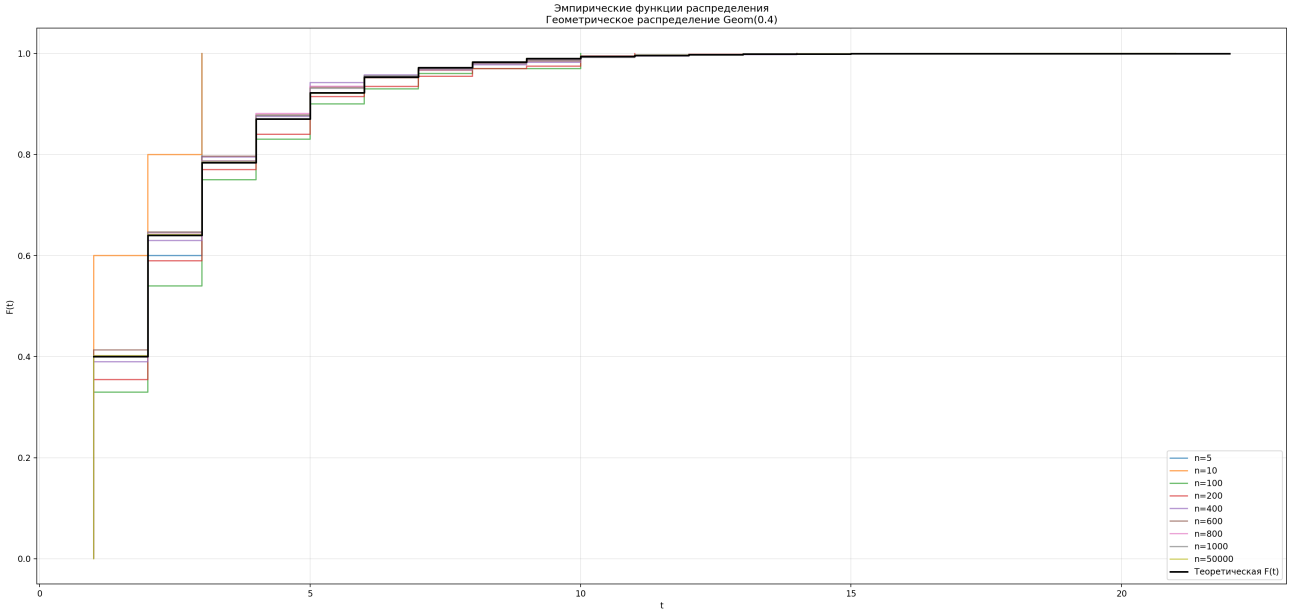

Рис. 4: Эмпирическая функция распределения для нормального распределения

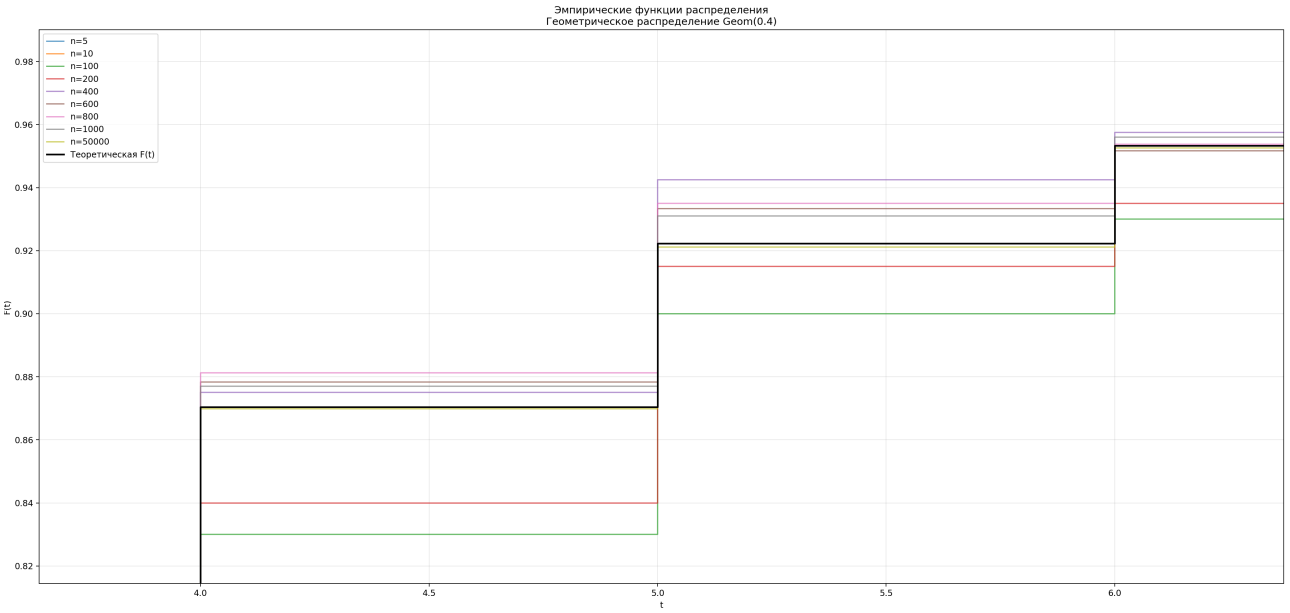

Рис. 5: Увеличенный фрагмент ЭФР нормального распределения

2.2.5 График для геометрического распределения (Разные сиды):

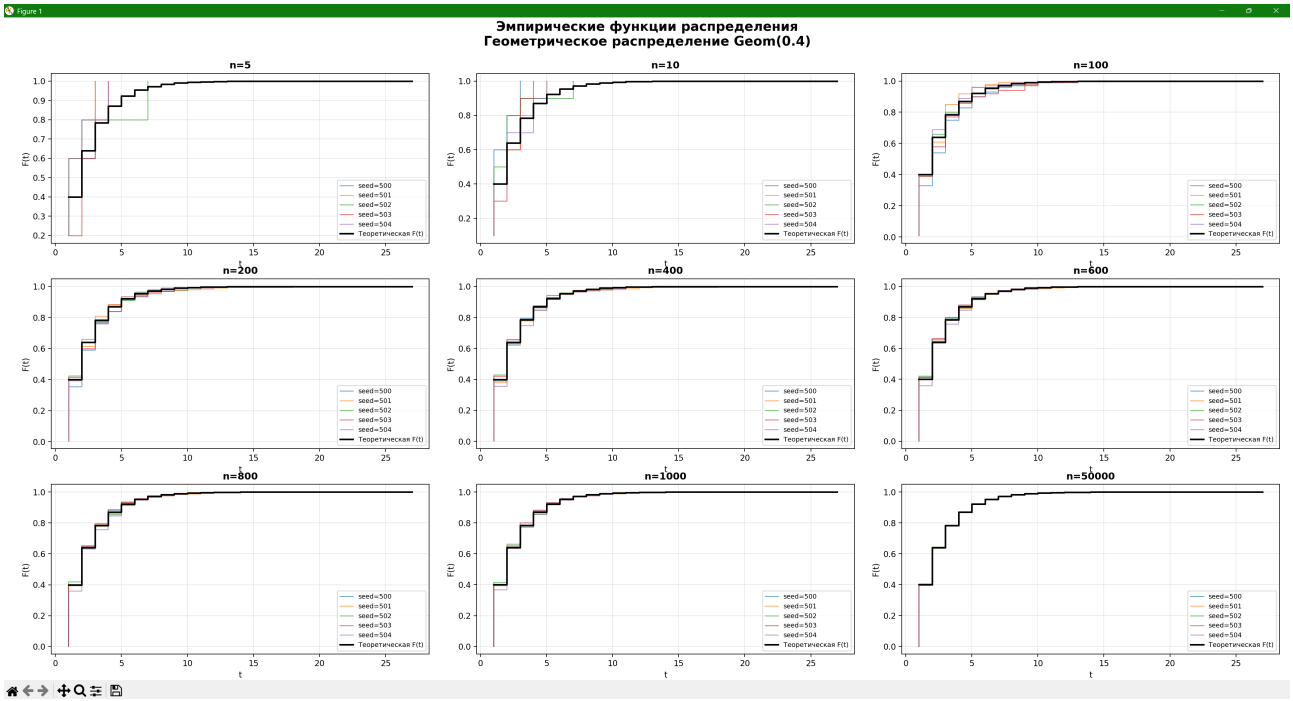

Рис. 6: Эмпирическая функция распределения для геометрического распределения

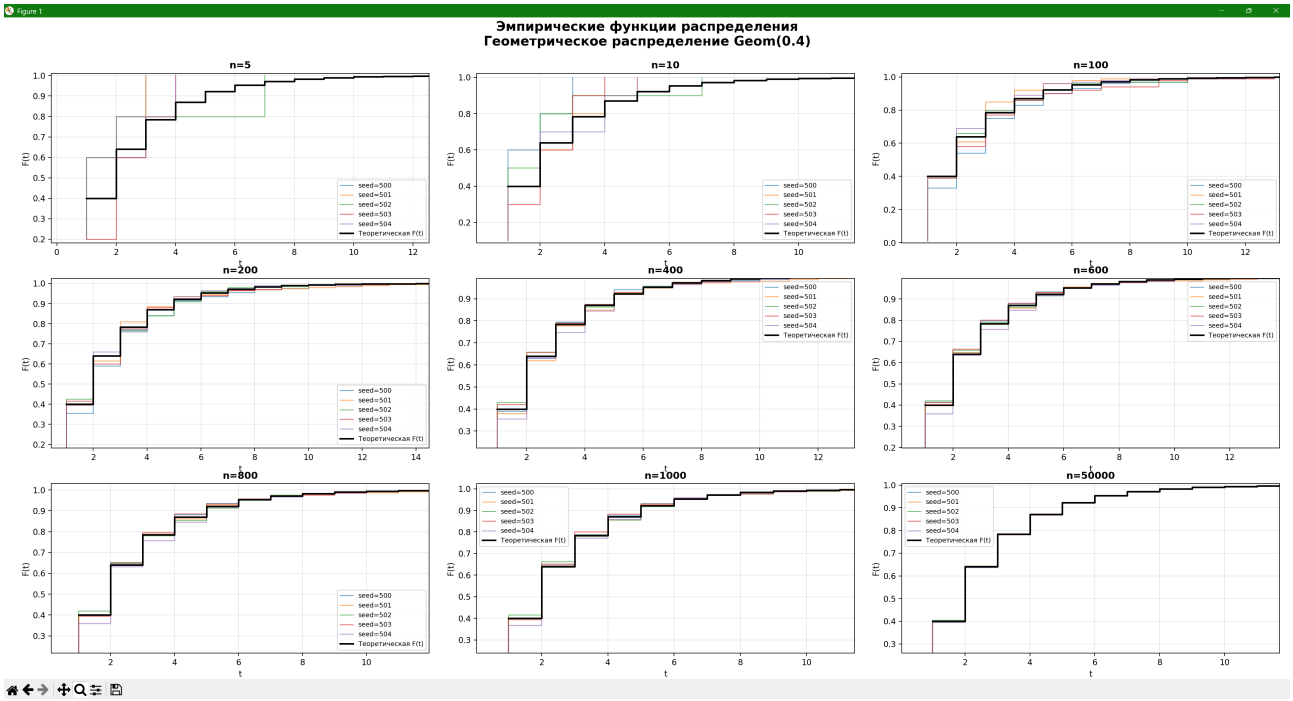

Рис. 7: Увеличенный фрагмент ЭФР геометрического распределения

2.2.6 График для нормального распределения (Разные сиды):

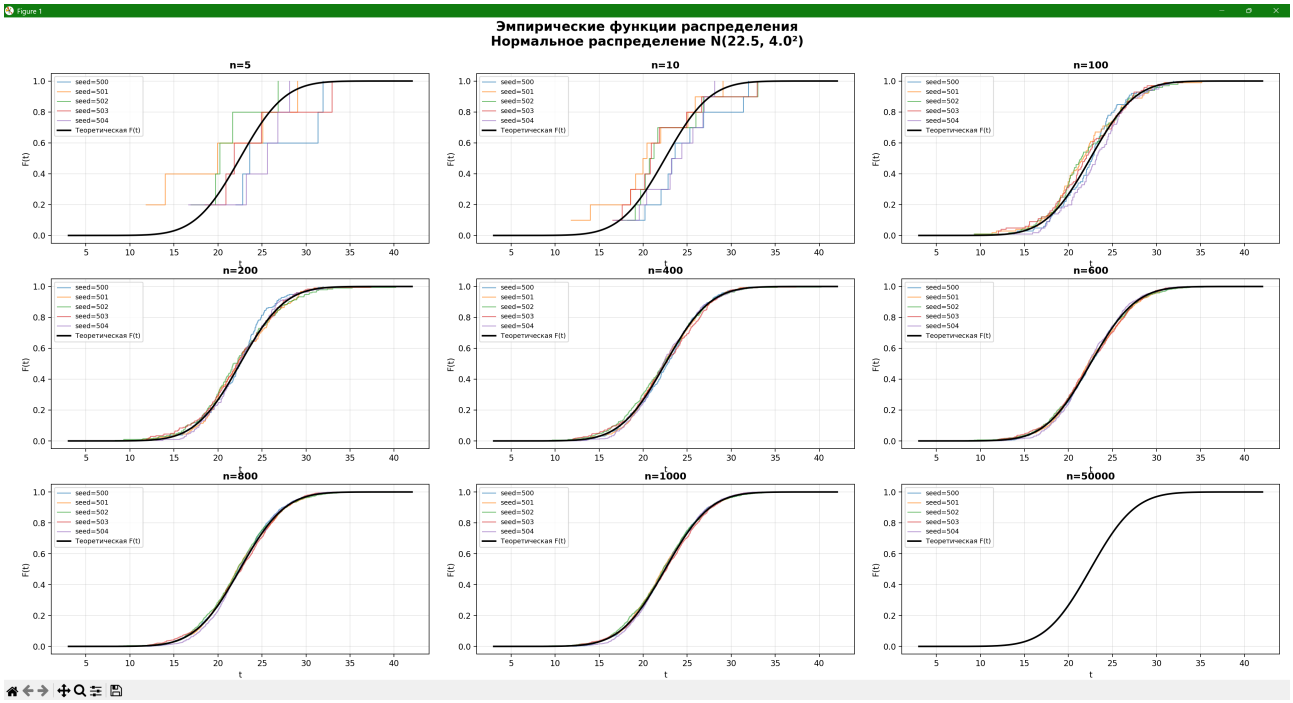

Рис. 8: Эмпирическая функция распределения для нормального распределения

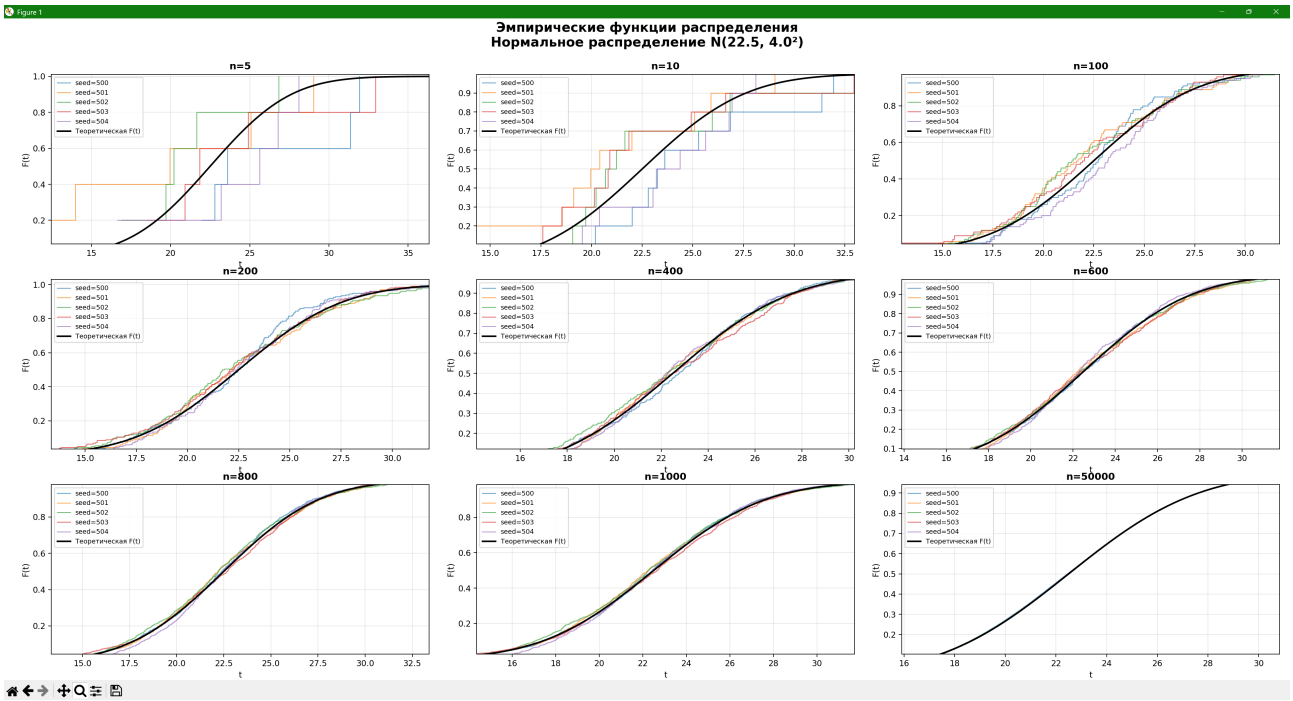

Рис. 9: Увеличенный фрагмент ЭФР нормального распределения

- При малых n (5, 10) эмпирические функции сильно отличаются друг от друга и заметно отклоняются от теоретической
- При средних n (100, 200) уже видно хорошее приближение, а различия между графиками уменьшаются

- При больших n (600, 800, 1000) эмпирические функции практически совпадают с теоретической
- При еще большем n (50000) даже при детальном рассмотрении (сильном приближении) эмпирическая функция очень близка к теоретической

2.2.7 Статистика Смирнова $D_{m,n}$

Для каждой пары выборок объемов n и m вычисляется статистика Смирнова:

$$D_{m,n} = \sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - \hat{F}_m(x) \right|$$

Эта статистика используется для проверки гипотезы об однородности двух выборок. Значение $D_{m,n}$ показывает максимальное расхождение между эмпирическими функциями распределения, нормированное на размеры выборок. Чем меньше значение $D_{m,n}$, тем более схожи эмпирические распределения выборок. При сравнении выборок близких размеров из одного и того же распределения ожидается получение относительно малых значений статистики, что указывает на согласованность эмпирических функций распределения.

Вычисление:

1. Объединить обе выборки и упорядочить все значения
2. В каждой точке скачка вычислить значения обеих ЭФР
3. Найти максимум модуля разности
4. Умножить на нормировочный коэффициент $\sqrt{\frac{nm}{n+m}}$

Ниже приведены таблицы статистики $D_{m,n}$, значения усреднены по 5 генерациям.

Таблица 1: Статистика $D_{m,n}$ для нормального распределения

	5	10	100	200	400	600	800	1000	1001
5	0.000	0.402	0.685	0.749	0.776	0.800	0.794	0.799	0.799
10	0.402	0.000	0.573	0.738	0.804	0.850	0.844	0.847	0.848
100	0.685	0.573	0.000	0.523	0.863	0.938	0.893	0.929	0.932
200	0.749	0.738	0.523	0.000	0.554	0.674	0.686	0.710	0.710
400	0.776	0.804	0.863	0.554	0.000	0.377	0.502	0.526	0.526
600	0.800	0.850	0.938	0.674	0.377	0.000	0.444	0.493	0.497
800	0.794	0.844	0.893	0.686	0.502	0.444	0.000	0.288	0.292
1000	0.799	0.847	0.929	0.710	0.526	0.493	0.288	0.000	0.015
1001	0.799	0.848	0.932	0.710	0.526	0.497	0.292	0.015	0.000

Таблица 2: Статистика $D_{m,n}$ для геометрического распределения

	5	10	100	200	400	600	800	1000	1001
5	0.000	0.292	0.458	0.455	0.472	0.470	0.464	0.464	0.464
10	0.292	0.000	0.416	0.407	0.395	0.388	0.397	0.391	0.392
100	0.458	0.416	0.000	0.302	0.604	0.636	0.610	0.610	0.606
200	0.455	0.407	0.302	0.000	0.456	0.527	0.455	0.480	0.473
400	0.472	0.395	0.604	0.456	0.000	0.253	0.278	0.350	0.352
600	0.470	0.388	0.636	0.527	0.253	0.000	0.187	0.225	0.231
800	0.464	0.397	0.610	0.455	0.278	0.187	0.000	0.163	0.162
1000	0.464	0.391	0.610	0.480	0.350	0.225	0.163	0.000	0.017
1001	0.464	0.392	0.606	0.473	0.352	0.231	0.162	0.017	0.000

2.3 Гистограмма и полигон частот

Для повышения надежности результатов построение гистограмм и полигонов частот выполнено на основе **усредненных по 5 независимым выборкам характеристик**. Усреднение позволяет снизить влияние отдельных выборок и получить более устойчивые оценки распределения.

2.3.1 Для дискретного распределения

Полигон частот. Для дискретного распределения строится полигон частот — ломаная, соединяющая точки $(x_i, \bar{\nu}(x_i))$, где:

- x_i — возможные значения, встретившиеся хотя бы в одной из 5 выборок
- $\bar{\nu}(x_i)$ — усредненная относительная частота

Процедура усреднения:
Для каждого значения x_i, встретившегося хотя бы в одной из 5 выборок, вычисляется усредненная относительная частота:

$$\bar{\nu}(x_i) = \frac{1}{5} \sum_{k=1}^{5} \frac{\nu_k(x_i)}{n} \tag{67}$$

где $\nu_k(x_i)$ — частота появления значения x_i в k-й выборке размера n. Если значение x_i не встретилось в некоторой выборке, его вклад в сумму равен нулю (т.е. $\nu_k(x_i) = 0$).

Сравнение с функцией вероятности. На том же графике изображается теоретическая функция вероятности:

$$P(\xi = x) = 0.4 \cdot 0.6^{x-1}, \quad x \in \mathbb{N} \tag{68}$$

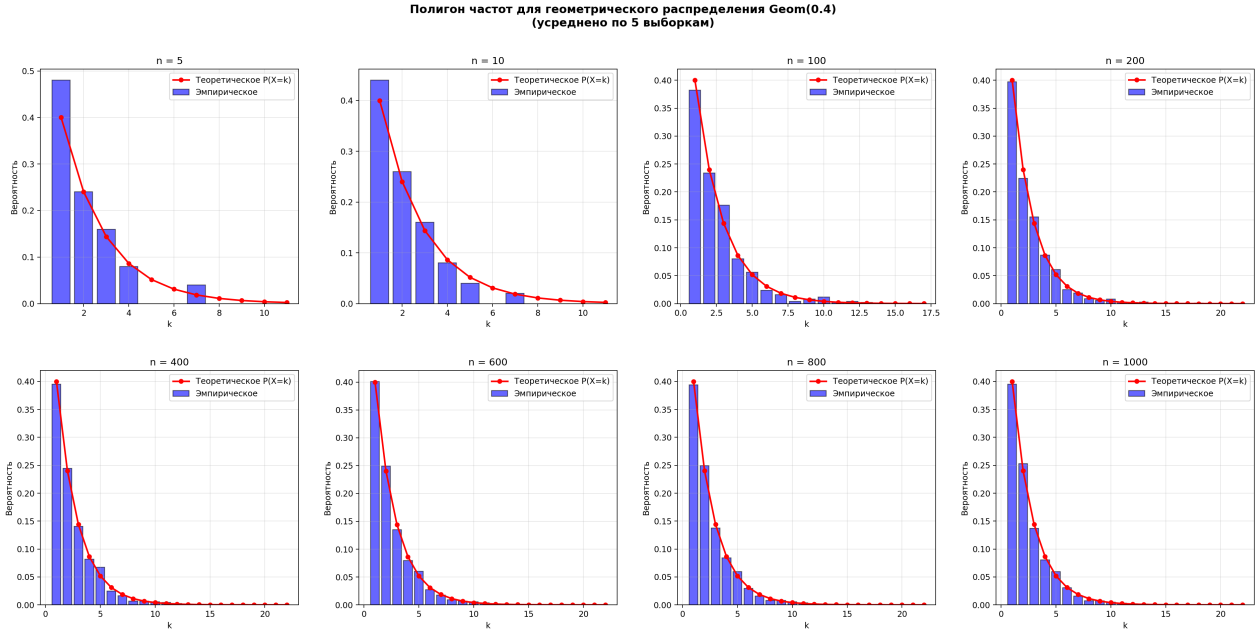

Рис. 10: Усредненный полигон частот геометрического распределения

Наблюдения:

- При $n = 5, 10$ усредненный полигон частот все еще демонстрирует заметные отклонения от теоретического распределения из-за малого размера выборок
- При $n = 100$ уже видно хорошее соответствие форме распределения
- При $n = 1000$ усредненные относительные частоты очень близки к теоретическим вероятностям, что подтверждает закон больших чисел

2.3.2 Для непрерывного распределения

Гистограмма. Для непрерывного распределения строится усредненная гистограмма с использованием следующей процедуры:

1. **Определение единой сетки интервалов.** На основе объединения всех 5 выборок размера n определяется диапазон $[\min(X), \max(X)]$, который разбивается на $k = 20$ интервалов равной ширины:
$$\text{bins} = \{a_0, a_1, \ldots, a_k\}, \quad a_{j+1} - a_j = \Delta_j \tag{69}$$
2. **Вычисление плотностей для каждой выборки.** Для каждого интервала $[a_j, a_{j+1})$ и каждой выборки k вычисляется высота столбца:
$$h_{k,j} = \frac{\nu_{k,j}}{n \cdot \Delta_j} \tag{70}$$
где $\nu_{k,j}$ — частота попадания значений из k-й выборки в j-й интервал.
3. **Усреднение высот.** Для каждого интервала j вычисляется усредненная высота столбца:
$$\bar{h}_j = \frac{1}{5} \sum_{k=1}^{5} h_{k,j} = \frac{1}{5} \sum_{k=1}^{5} \frac{\nu_{k,j}}{n \cdot \Delta_j} \tag{71}$$
4. **Построение гистограммы.** Строятся прямоугольники с основанием Δ_j и высотой $\bar{h}_j$.

Высота выбирается так, чтобы площадь прямоугольника равнялась усредненной относительной частоте: $\bar{h}_j \cdot \Delta_j = \frac{1}{5} \sum_{k=1}^{5} \frac{\nu_{k,j}}{n}$.

Сравнение с плотностью. На том же графике изображается теоретическая плотность:
$$f(x) = \frac{1}{4\sqrt{2\pi}} \exp\left(-\frac{(x - 22.5)^2}{32}\right) \tag{72}$$

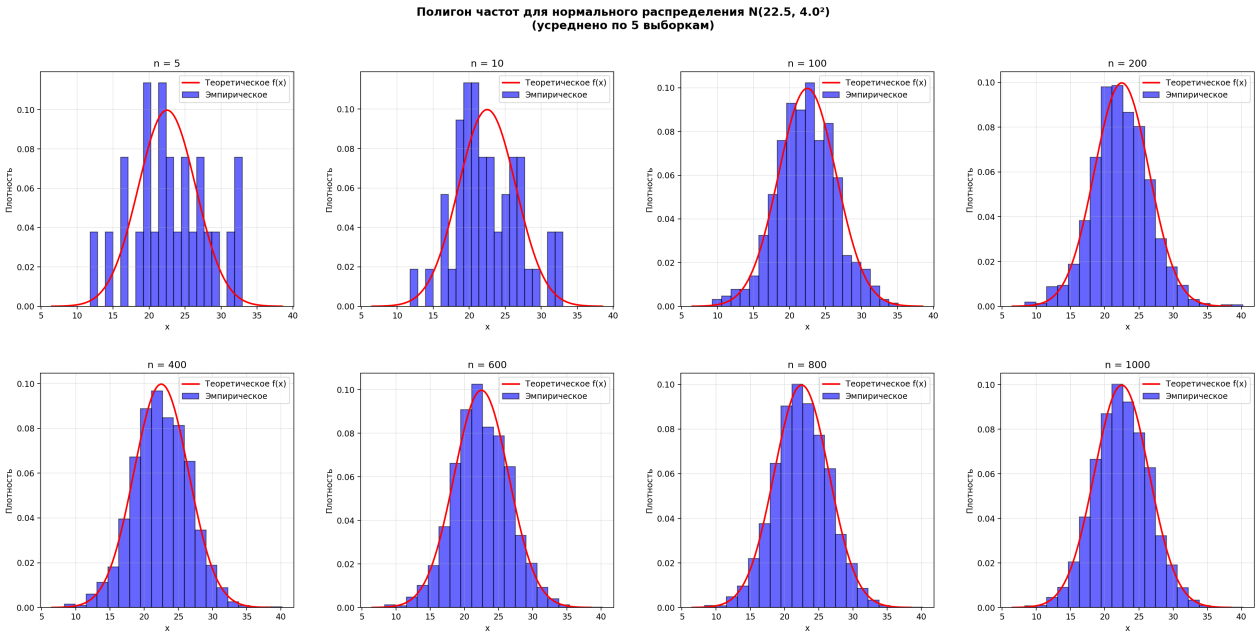

Рис. 11: Усредненная гистограмма нормального распределения

Наблюдения:

- При малых размерах выборок ($n = 5, 10$) усредненная гистограмма имеет неровную форму и существенно отличается от теоретической плотности
- При увеличении n до 100-200 форма гистограммы начинает приближаться к колоколообразной кривой нормального распределения
- При $n = 800, 1000$ усредненная гистограмма практически совпадает с теоретической плотностью

2.3.3 Графики иллюстрируют теорему математического анализа

Полученные графики иллюстрируют **закон больших чисел**: усредненная относительная частота сходится по вероятности к истинной вероятности при увеличении объема выборки.

Для дискретного случая:

$$\bar{\nu}(x_i) \xrightarrow{P} P(\xi = x_i) \quad \text{при } n \to \infty \tag{73}$$

Для непрерывного случая усредненная гистограмма сходится к плотности распределения:

$$\bar{h}_j \xrightarrow{P} \frac{1}{\Delta_j} \int_{a_j}^{a_{j+1}} f(x)\,dx \approx f(\xi_j) \quad \text{при } n \to \infty,\ \Delta_j \to 0 \tag{74}$$

где $\xi_j \in [a_j, a_{j+1})$ — некоторая точка в интервале.

Усреднение по 5 независимым выборкам снижает дисперсию оценок, что обеспечивает более гладкие и устойчивые графики, особенно при малых размерах выборок.

2.4 Выборочные моменты

Для повышения надежности оценок все вычисления выборочных моментов производятся на основе **усреднения по 5 независимым выборкам**. Такой подход позволяет снизить влияние случайной вариабельности и получить более устойчивые оценки математического ожидания и дисперсии.

2.4.1 Определения и вычисление

Для каждой выборки $X = (X_1, \ldots, X_n)$ вычислены следующие характеристики:
Выборочное среднее:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{75}$$

Выборочная дисперсия:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{76}$$

Процедура усреднения:
Для каждого размера выборки n выполняется следующая процедура:

1. Для каждой из 5 независимых выборок (с различными seed = 500, 501, 502, 503, 504) вычисляются $\bar{X}_k$ и S_k^2, где $k = 1, \ldots, 5$
2. Вычисляются усредненные значения:

$$\overline{\bar{X}} = \frac{1}{5} \sum_{k=1}^{5} \bar{X}_k \tag{77}$$

$$\overline{S^2} = \frac{1}{5} \sum_{k=1}^{5} S_k^2 \tag{78}$$

Далее в таблицах и на графиках приводятся именно усредненные значения $\overline{\bar{X}}$ и $\overline{S^2}$.

Таблица 3: Усредненные выборочные моменты для нормального распределения $N(22.5, 16)$

n	$\hat{X}$	$\mathbf{S^2}$
5	23.053	22.617
10	22.637	19.352
100	22.316	16.855
200	22.337	16.989
400	22.463	16.535
600	22.528	16.513
800	22.478	16.040
1000	22.480	15.878

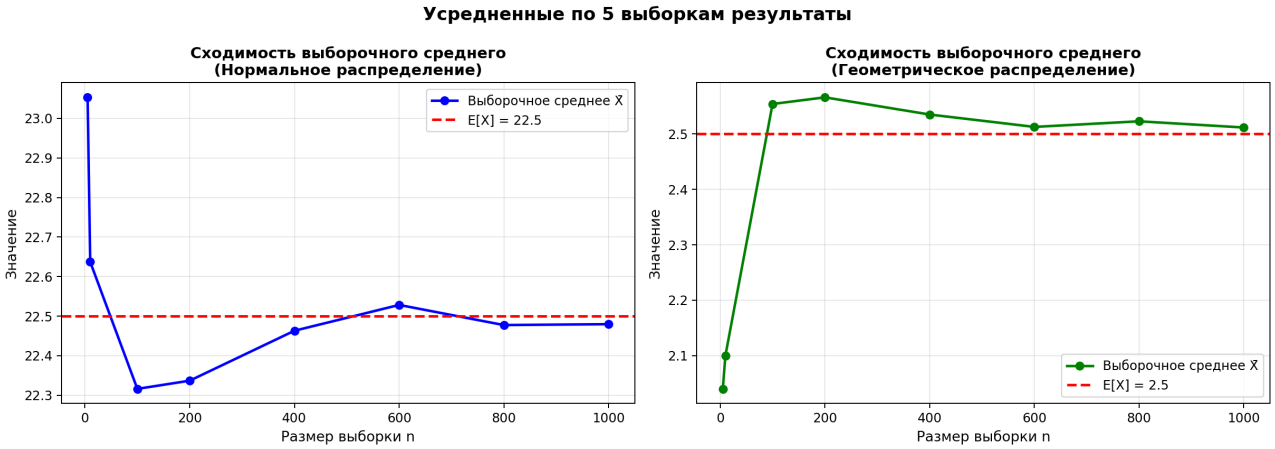

Рис. 12: Сходимость усредненного выборочного среднего

Таблица 4: Усредненные выборочные моменты для геометрического распределения Geom(0.4)

n	$\hat{X}$	S^2
5	2.040	1.856
10	2.100	1.702
100	2.554	3.927
200	2.566	4.190
400	2.535	4.105
600	2.513	4.082
800	2.523	3.975
1000	2.512	3.952

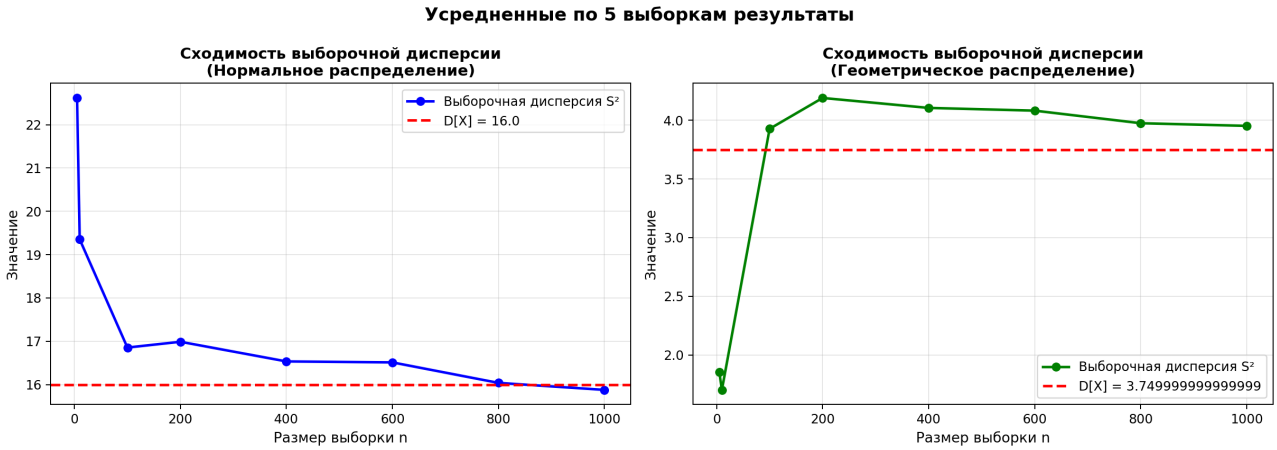

Рис. 13: Сходимость усредненной выборочной дисперсии

2.5 Сравнение с истинными значениями

Для нормального распределения $N(22.5, 16)$**:**

- Истинное математическое ожидание: $E[X] = \mu = 22.5$
- Истинная дисперсия: $\mathrm{D}[X] = \sigma^2 = 16$

Для геометрического распределения Geom(0.4):

- Истинное математическое ожидание: $E[X] = \frac{1}{\theta} = \frac{1}{0.4} = 2.5$
- Истинная дисперсия: $\mathrm{Var}[X] = \frac{1-\theta}{\theta^2} = \frac{0.6}{0.16} = 3.75$

Из графиков и таблиц видно, что при увеличении размера выборки n усредненные выборочные оценки $\overline{\bar{X}}$ и $\overline{S^2}$ сходятся к истинным значениям, что подтверждает свойство состоятельности оценок и иллюстрирует закон больших чисел.

Для малых выборок ($n = 5, 10$) наблюдаются существенные отклонения от истинных значений, однако усреднение по 5 реализациям уже снижает случайную вариабельность. При $n \geq 100$ усредненные оценки становятся достаточно точными и стабильными.

2.5.1 Свойства выборочных оценок

1. Несмещенность. Выборочное среднее является несмещенной оценкой математического ожидания:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = E[\xi] \tag{79}$$

Выборочная дисперсия S^2 является смещенной оценкой дисперсии:

$$E[S^2] = \frac{n-1}{n}\mathrm{D}[\xi] \tag{80}$$

Несмещенная оценка дисперсии:

$$S^2_{\text{несм}} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{81}$$

2. Состоятельность. Обе оценки состоятельны (сходятся по вероятности к истинным значениям при $n \to \infty$):

$$\bar{X} \xrightarrow{P} E[\xi], \quad S^2 \xrightarrow{P} \mathrm{D}[\xi] \tag{82}$$

Это следует из закона больших чисел. Данные из таблиц подтверждают состоятельность: при увеличении n усредненные оценки приближаются к теоретическим значениям.

3. Эффект усреднения. Усреднение оценок по 5 независимым выборкам дополнительно улучшает точность. Если $\bar{X}_k$ — выборочное среднее из k-й выборки, то дисперсия усредненной оценки:

$$\mathrm{D}\left[\overline{\bar{X}}\right] = \mathrm{D}\left[\frac{1}{5}\sum_{k=1}^{5}\bar{X}_k\right] = \frac{1}{25}\sum_{k=1}^{5}\mathrm{D}[\bar{X}_k] = \frac{1}{5}\cdot\frac{\sigma^2}{n} = \frac{\sigma^2}{5n} \tag{83}$$

Таким образом, усреднение по 5 выборкам дополнительно снижает дисперсию оценки в 5 раз по сравнению с одиночной выборкой того же размера, что особенно важно при малых n.