Национальный исследовательский университет
Высшая школа экономики
Московский институт электроники и математики

Департамент прикладной математики
кафедра компьютерной безопасности

Домашнее задание № 3-4 по математической статистике

Дискретное распределение: *Геометрическое распределение*
Непрерывное распределение: *Нормальное II распределение*

Репозиторий с кодом: *https://github.com/danirod12/matstat-hw*

Выполнил
Федоров Д.В.

Проверил
Богданов Д.С.

Москва 2025

Содержание

3 Домашнее задание 3: Построение точечных оценок параметра распределения 2
3.1 Исходные данные . . . 2
3.2 Получение оценок методом моментов и методом максимального правдоподобия . . . 2
3.2.1 Геометрическое распределение . . . 2
3.2.2 Нормальное распределение II . . . 4
3.3 Поиск оптимальных оценок . . . 6
3.3.1 Теоретическое введение . . . 6
3.3.2 Геометрическое распределение . . . 6
3.3.3 Нормальное распределение II . . . 9
3.4 Работа с реальными данными . . . 11
3.4.1 Геометрическое распределение: данные о конверсии клиентов . . . 11
3.4.2 Нормальное распределение II: температурные данные . . . 13
3.4.3 Общий вывод . . . 14

4 Домашнее задание 4: Проверка статистических гипотез 15
4.1 Исходные данные . . . 15
4.2 Проверка гипотезы о виде распределения . . . 15
4.2.1 Теоретические основы . . . 15
4.2.2 Критерий согласия Колмогорова (Смирнова) . . . 15
4.2.3 Критерий согласия хи-квадрат (Пирсона) . . . 16
4.2.4 Сложная гипотеза: случай неизвестных параметров . . . 16
4.2.5 Геометрическое распределение . . . 17
4.2.6 Нормальное распределение II . . . 18
4.3 Проверка гипотезы об однородности выборок . . . 20
4.3.1 Теоретические основы . . . 20
4.3.2 Анализ однородности для сгенерированных выборок . . . 20
4.4 Проверка гипотез для реальных данных . . . 23
4.4.1 Геометрическое распределение: данные о конверсии клиентов . . . 23
4.4.2 Нормальное распределение II: температурные данные . . . 25
4.4.3 Общий вывод . . . 26

3 Домашнее задание 3: Построение точечных оценок параметра распределения

3.1 Исходные данные

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

3.2 Получение оценок методом моментов и методом максимального правдоподобия

3.2.1 Геометрическое распределение

Закон распределения. Случайная величина ξ имеет геометрическое распределение:

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

Параметр θ — неизвестный параметр, который требуется оценить.

Оценка методом моментов. *Теоретическое математическое ожидание:*

$$\mathbb{E}[\xi] = \frac{1}{\theta} \tag{2}$$

Вывод:

$$\begin{aligned}
\mathbb{E}[\xi] &= \sum_{x=1}^{\infty} x \cdot \theta(1-\theta)^{x-1} \\
&= \theta \sum_{x=1}^{\infty} x(1-\theta)^{x-1} \\
&= \theta \cdot \frac{d}{dq}\left[\sum_{x=1}^{\infty} q^x\right]\Bigg|_{q=1-\theta} \\
&= \theta \cdot \frac{d}{dq}\left[\frac{q}{1-q}\right]\Bigg|_{q=1-\theta} \\
&= \theta \cdot \frac{1}{(1-q)^2}\Bigg|_{q=1-\theta} = \frac{1}{\theta}
\end{aligned}$$

Приравниваем теоретический момент к выборочному:

$$\frac{1}{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

Оценка методом моментов:

$$\boxed{\hat{\theta}_{MM} = \frac{1}{\bar{X}}} \tag{4}$$

Оценка методом максимального правдоподобия. *Функция правдоподобия:*

$$L(X;\theta) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{\sum_{i=1}^n x_i - n} \tag{5}$$

Логарифм функции правдоподобия:

$$\ln L(X;\theta) = n \ln \theta + \left(\sum_{i=1}^{n} x_i - n \right) \ln(1-\theta) \tag{6}$$

Берём производную по θ и приравниваем к нулю:

$$\frac{d \ln L}{d\theta} = \frac{n}{\theta} - \frac{\sum_{i=1}^n x_i - n}{1-\theta} = 0 \tag{7}$$

Решаем уравнение:

$$\begin{aligned}
\frac{n}{\theta} &= \frac{\sum_{i=1}^n x_i - n}{1-\theta} \\
n(1-\theta) &= \theta \left(\sum_{i=1}^{n} x_i - n \right) \\
n - n\theta &= \theta \sum_{i=1}^{n} x_i - n\theta \\
n &= \theta \sum_{i=1}^{n} x_i \\
\theta &= \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}}
\end{aligned}$$

Оценка максимального правдоподобия:

$$\boxed{\hat{\theta}_{ML} = \frac{1}{\bar{X}}} \tag{8}$$

Вывод. Для геометрического распределения оценки, полученные методом моментов и методом максимального правдоподобия, **совпадают**:

$$\hat{\theta}_{MM} = \hat{\theta}_{ML} = \frac{1}{\bar{X}} \tag{9}$$

Значения оценок для сгенерированных выборок. Для выборок из пункта 2.1 со значениями $n \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$ вычислим оценку $\hat{\theta} = 1/\bar{X}$ (истинное значение $\theta = 0.4$).

Таблица 1: Оценки параметра θ для геометрического распределения

n	$\bar{X}$	$\hat{\theta}$	**Ошибка**
5	2.0400	0.4902	0.0902
10	2.1000	0.4762	0.0762
100	2.5540	0.3915	0.0085
200	2.5660	0.3897	0.0103
400	2.5350	0.3945	0.0055
600	2.5127	0.3980	0.0020
800	2.5227	0.3964	0.0036
1000	2.5116	0.3982	0.0018

3.2.2 Нормальное распределение II

Плотность распределения. Случайная величина ξ имеет нормальное распределение II:

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\theta^2}\right\}, \quad x, \mu \in \mathbb{R}, \quad \theta > 0 \tag{10}$$

Параметр $\mu = 22.5$ — известен, параметр θ — неизвестный параметр (стандартное отклонение).

Оценка методом моментов. *Теоретические моменты:*

$$\mathbb{E}[\xi] = \mu \tag{11}$$
$$\mathbb{D}[\xi] = \theta^2 \tag{12}$$

Поскольку μ известен, используем второй момент:

$$\mathbb{D}[\xi] = \mathbb{E}[(\xi - \mu)^2] = \theta^2 \tag{13}$$

Приравниваем теоретическую дисперсию к выборочной:

$$\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{14}$$

Оценка методом моментов:

$$\boxed{\hat{\theta}_{MM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}} \tag{15}$$

где $\mu = 22.5$ — известное значение.

Оценка методом максимального правдоподобия. *Функция правдоподобия:*

$$L(X;\theta) = \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\theta^2}\right\} = \frac{1}{(2\pi)^{n/2}\theta^n} \exp\left\{-\frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \tag{16}$$

Логарифм функции правдоподобия:

$$\ln L(X;\theta) = -\frac{n}{2}\ln(2\pi) - n\ln\theta - \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{17}$$

Берём производную по θ и приравниваем к нулю:

$$\frac{d\ln L}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \tag{18}$$

Решаем уравнение:

$$\frac{n}{\theta} = \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i - \mu)^2$$
$$n\theta^2 = \sum_{i=1}^{n}(x_i - \mu)^2$$
$$\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Оценка максимального правдоподобия:

$$\boxed{\hat{\theta}_{ML} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}} \tag{19}$$

Вывод. Для нормального распределения II (с известным μ) оценки, полученные методом моментов и методом максимального правдоподобия, **совпадают**:

$$\hat{\theta}_{MM} = \hat{\theta}_{ML} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - 22.5)^2} \tag{20}$$

Значения оценок для сгенерированных выборок. Для выборок из пункта 2.1 со значениями $n \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$ вычислим оценку $\hat{\theta}$ (истинное значение $\theta = 4.0$).

Таблица 2: Оценки параметра θ для нормального распределения II

n	$\hat{\theta}_{MM}$	$\hat{\theta}_{MMpravdopodob}$	**Ошибка**
5	5.3005	5.3005	1.3005
10	4.5981	4.5981	0.5981
100	4.1323	4.1323	0.1323
200	4.1279	4.1279	0.1279
400	4.0681	4.0681	0.0681
600	4.0640	4.0640	0.0640
800	4.0063	4.0063	0.0063
1000	3.9860	3.9860	0.0140

3.3 Поиск оптимальных оценок

3.3.1 Теоретическое введение

Определение 1. *Параметрическое семейство $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$ называется* ***экспоненциальным****, если плотность (или вероятность) имеет вид:*

$$f_\theta(x) = \exp\{A(\theta) \cdot B(x) + C(\theta) + D(x)\} \tag{21}$$

Теорема 1 (О полноте экспоненциальных семейств)**.** *Если $\mathcal{F}$ — экспоненциальное семейство и $A(\theta)$ содержит некоторый отрезок при изменении $\theta \in \Theta$, то статистика*

$$T(X) = \sum_{i=1}^{n} B(X_i) \tag{22}$$

является ***полной и достаточной*** *статистикой.*

Теорема 2 (Об оптимальной оценке в экспоненциальном семействе)**.** *Если параметрическое семейство регулярно и экспоненциально, то статистика*

$$T(X) = \frac{1}{n} \sum_{i=1}^{n} B(X_i) \tag{23}$$

является эффективной (оптимальной) оценкой для параметрической функции

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)} \tag{24}$$

Теорема 3. *Если существует полная достаточная статистика $T = T(X)$, то произвольная функция от нее $H(T)$ является* ***оптимальной оценкой*** *своего математического ожидания $\mathbb{E}_\theta[H(T)] = \tau(\theta)$.*

3.3.2 Геометрическое распределение

Приведение к экспоненциальному семейству. Геометрическое распределение:

$$P(\xi = x) = \theta(1-\theta)^{x-1}, \quad x \in \mathbb{N}, \quad \theta \in (0,1) \tag{25}$$

Преобразование к экспоненциальному виду:

$$\begin{aligned} P(\xi = x) &= \theta(1-\theta)^{x-1} && (26)\\ &= \exp\{\ln\theta + (x-1)\ln(1-\theta)\} && (27)\\ &= \exp\{x\ln(1-\theta) + \ln\theta - \ln(1-\theta)\} && (28)\\ &= \exp\left\{x\ln(1-\theta) + \ln\frac{\theta}{1-\theta}\right\} && (29) \end{aligned}$$

Получили экспоненциальное семейство с параметрами:

$$\begin{aligned} A(\theta) &= \ln(1-\theta) && (30)\\ B(x) &= x && (31)\\ C(\theta) &= \ln\frac{\theta}{1-\theta} = \ln\theta - \ln(1-\theta) && (32)\\ D(x) &= 0 && (33) \end{aligned}$$

Достаточная статистика. По теореме о полноте экспоненциальных семейств:

$$\boxed{T(X) = \sum_{i=1}^{n} X_i = n\bar{X}} \tag{34}$$

является **полной и достаточной** статистикой.

Параметрическая функция $\tau(\theta)$. Вычислим производные:

$$A'(\theta) = \frac{d}{d\theta} \ln(1-\theta) = -\frac{1}{1-\theta} \tag{35}$$

$$C'(\theta) = \frac{d}{d\theta} \left[\ln\theta - \ln(1-\theta)\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)} \tag{36}$$

По формуле для экспоненциального семейства:

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)} = -\frac{\frac{1}{\theta(1-\theta)}}{-\frac{1}{1-\theta}} = \frac{1}{\theta} \tag{37}$$

Заметим, что $\tau(\theta) = \frac{1}{\theta} = \mathbb{E}[\xi]$ — математическое ожидание геометрического распределения.

Оптимальная оценка для $\tau(\theta) = \frac{1}{\theta}$. По теореме об оптимальной оценке в экспоненциальном семействе, статистика

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

является эффективной (оптимальной) оценкой для $\tau(\theta) = \frac{1}{\theta}$.

Оптимальная оценка:

$$\boxed{\hat{\tau}_{opt} = \bar{X}} \tag{39}$$

Проверка несмещённости:

$$\mathbb{E}_\theta[\bar{X}] = \mathbb{E}_\theta[X_1] = \frac{1}{\theta} = \tau(\theta) \tag{40}$$

Об оценке для θ. Параметрическая функция $\tau(\theta) = \frac{1}{\theta}$ взаимно однозначна, поэтому:

$$\theta = \frac{1}{\tau(\theta)} \tag{41}$$

Используя инвариантность оценки максимального правдоподобия, получаем оценку для θ:

$$\boxed{\hat{\theta} = \frac{1}{\bar{X}}} \tag{42}$$

Важно: Эта оценка является функцией от достаточной статистики, но **не является оптимальной** оценкой для θ в смысле минимальной дисперсии среди несмещённых оценок (поскольку она смещённая).

Действительно:

$$\mathbb{E}_\theta\left[\frac{1}{\bar{X}}\right] \neq \theta \tag{43}$$

Это связано с тем, что $g(x) = 1/x$ — нелинейная функция, и математическое ожидание обратной величины не равно обратной величине математического ожидания.

Значения оценок для сгенерированных выборок. Для выборок объёмов $n \in \{5, 10, 100, 200,$ вычислим оптимальные оценки (истинное значение $\theta = 0.4$, $\tau(\theta) = 2.5$).

Таблица 3: Оптимальные оценки для геометрического распределения

n	$\hat{\tau}_{opt} = \bar{X}$	$\hat{\theta} = 1/\bar{X}$
5	2.0400	0.4902
10	2.1000	0.4762
100	2.5540	0.3915
200	2.5660	0.3897
400	2.5350	0.3945
600	2.5127	0.3980
800	2.5227	0.3964
1000	2.5116	0.3982
Истинные значения: $\tau(\theta) = 2.5$, $\theta = 0.4$		

3.3.3 Нормальное распределение II

Приведение к экспоненциальному семейству. Нормальное распределение II с известным μ и неизвестным θ:

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\theta^2}\right\}, \quad x, \mu \in \mathbb{R}, \quad \theta > 0 \tag{44}$$

Преобразование к экспоненциальному виду:

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\theta^2}\right\} \tag{45}$$

$$= \exp\left\{-\frac{(x-\mu)^2}{2\theta^2} - \ln\theta - \frac{1}{2}\ln(2\pi)\right\} \tag{46}$$

$$= \exp\left\{-\frac{1}{2\theta^2} \cdot (x-\mu)^2 - \ln\theta - \frac{1}{2}\ln(2\pi)\right\} \tag{47}$$

Получили экспоненциальное семейство с параметрами:

$$A(\theta) = -\frac{1}{2\theta^2} \tag{48}$$

$$B(x) = (x-\mu)^2 \tag{49}$$

$$C(\theta) = -\ln\theta - \frac{1}{2}\ln(2\pi) \tag{50}$$

$$D(x) = 0 \tag{51}$$

Достаточная статистика. По теореме о полноте экспоненциальных семейств:

$$\boxed{T(X) = \sum_{i=1}^{n} (X_i - \mu)^2} \tag{52}$$

является **полной и достаточной** статистикой.

Параметрическая функция $\tau(\theta)$. Вычислим производные:

$$A'(\theta) = \frac{d}{d\theta}\left(-\frac{1}{2\theta^2}\right) = \frac{1}{\theta^3} \tag{53}$$

$$C'(\theta) = \frac{d}{d\theta}\left[-\ln\theta - \frac{1}{2}\ln(2\pi)\right] = -\frac{1}{\theta} \tag{54}$$

По формуле для экспоненциального семейства:

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)} = -\frac{-\frac{1}{\theta}}{\frac{1}{\theta^3}} = \theta^2 \tag{55}$$

Заметим, что $\tau(\theta) = \theta^2 = \mathbb{D}[\xi]$ — дисперсия нормального распределения.

Оптимальная оценка для $\tau(\theta) = \theta^2$. По теореме об оптимальной оценке в экспоненциальном семействе:

$$\frac{1}{n}\sum_{i=1}^{n} B(X_i) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 \tag{56}$$

является эффективной (оптимальной) оценкой для $\tau(\theta) = \theta^2$.

Оптимальная оценка:

$$\boxed{\hat{\tau}_{opt} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2} \tag{57}$$

где $\mu = 22.5$ — известное значение.

Проверка несмещённости:

$$\mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \mathbb{E}_\theta[(X_1 - \mu)^2] = \mathbb{D}[\xi] = \theta^2 = \tau(\theta) \tag{58}$$

Об оценке для θ. Параметрическая функция $\tau(\theta) = \theta^2$ не является взаимно однозначной (два значения θ и $-\theta$ дают одно значение τ). Однако, поскольку $\theta > 0$ по определению, можно рассмотреть:

$$\theta = \sqrt{\tau(\theta)} \tag{59}$$

Естественная оценка для θ:

$$\hat{\theta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2} \tag{60}$$

Важно: Эта оценка **не является оптимальной** для θ в смысле минимальной дисперсии среди несмещённых оценок, хотя и является функцией от достаточной статистики.

Причина: функция $g(x) = \sqrt{x}$ нелинейна, и:

$$\mathbb{E}_\theta\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}\right] \neq \theta \tag{61}$$

Оценка является **асимптотически несмещённой** и **состоятельной**, но смещена для конечных выборок.

Значения оценок для сгенерированных выборок. Для выборок объёмов $n \in \{5, 10, 100, 200,$ вычислим оптимальные оценки (истинные значения: $\theta = 4.0$, $\tau(\theta) = 16.0$, $\mu = 22.5$).

Таблица 4: Оптимальные оценки для нормального распределения II

n	$\hat{\tau}_{opt} = \frac{1}{n}\sum(X_i - \mu)^2$	$\hat{\theta} = \sqrt{\hat{\tau}_{opt}}$
5	28.0952	5.3005
10	21.1426	4.5981
100	17.0758	4.1323
200	17.0399	4.1279
400	16.5498	4.0681
600	16.5158	4.0640
800	16.0506	4.0063
1000	15.8879	3.9860
Истинные значения: $\tau(\theta) = 16.0$, $\theta = 4.0$		

3.4 Работа с реальными данными

В данном разделе проводится анализ реальных данных, соответствующих интерпретациям распределений из первого домашнего задания. Используются открытые датасеты из репозиториев UCI Machine Learning Repository и FiveThirtyEight.

3.4.1 Геометрическое распределение: данные о конверсии клиентов

Источник данных. Использован датасет UCI Online Retail Dataset[1], содержащий 541 909 транзакций интернет-магазина за период 01.12.2010 – 09.12.2011. Датасет загружен через библиотеку `ucimlrepo`.

Интерпретация и подбор параметров. Рассматривается модель геометрического распределения: для каждого клиента подсчитывается количество покупок до первой «крупной» покупки, превышающей заданный порог. Вероятность успеха θ зависит от выбора порога.

Для подбора порога, обеспечивающего $\hat{\theta} \approx 0.4$, проведён анализ зависимости оценки от порога:

Таблица 5: Зависимость оценки $\hat{\theta}$ от порога крупной покупки

Порог, £	n	$\bar{X}$	$\hat{\theta}$
5	4336	1.538	0.6504
8	**4327**	**2.379**	**0.4203**
10	4308	3.319	0.3013
12	4290	4.326	0.2312
15	4237	5.934	0.1685

Выбран порог **£8**, при котором $\hat{\theta} = 0.42$ наиболее близко к теоретическому значению $\theta = 0.4$.

Таблица 6: Описательная статистика для данных о конверсии (порог £8)

Характеристика	**Значение**
Объём выборки n	500
Минимум	1
Максимум	49
Медиана	1.0

Описательная статистика.

Выборочные моменты.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 2.1380 \tag{62}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 18.1990 \tag{63}$$

[1] https://archive.ics.uci.edu/dataset/352/online+retail

Оценки параметра θ. Оценка методом моментов и методом максимального правдоподобия:

$$\hat{\theta}_{MM} = \hat{\theta}_{ML} = \frac{1}{\bar{X}} = \frac{1}{2.1380} = 0.4677 \tag{64}$$

Оптимальная оценка. Оптимальная оценка для параметрической функции $\tau(\theta) = 1/\theta$:

$$\hat{\tau}_{opt} = \bar{X} = 2.1380 \tag{65}$$

Таблица 7: Сравнение выборочных и теоретических характеристик (геометрическое)

Характеристика	**Теор.** ($\theta = 0.4$)	**Выборочное**	**Разница**
$\mathbb{E}[\xi] = 1/\theta$	2.5000	2.1380	0.3620
$\mathbb{D}[\xi] = (1-\theta)/\theta^2$	3.7500	18.1990	14.4490
θ	0.4000	0.4677	0.0677

Сравнение с теоретическими значениями.

Вывод. Оценка параметра $\hat{\theta} = 0.47$ находится в разумной близости к теоретическому значению $\theta = 0.4$ (отклонение около 17%). Выборочное среднее $\bar{X} = 2.14$ также близко к теоретическому значению $1/\theta = 2.5$.

Завышенная выборочная дисперсия ($S^2 = 18.2$ вместо теоретических 3.75) объясняется наличием выбросов в реальных данных — некоторые клиенты совершали до 49 покупок до первой крупной. Это типично для реальных данных и отражает неидеальность модели. Тем не менее, геометрическое распределение остаётся адекватной моделью для описания процесса конверсии клиентов.

3.4.2 Нормальное распределение II: температурные данные

Источник данных. Использован датасет FiveThirtyEight US Weather History[2], содержащий ежедневные метеорологические данные для городов США за 2014–2015 годы. Выбран город Хьюстон (код KHOU) как город с умеренно-тёплым климатом.

Интерпретация и фильтрация данных. Для получения данных с температурой, близкой к теоретическому значению $\mu = 22.5°C$, отфильтрованы весенние месяцы (март–май), когда средняя температура в Хьюстоне составляет примерно 20–25°C.

Температура переведена из шкалы Фаренгейта в Цельсий:

$$T_C = (T_F - 32) \times \frac{5}{9} \tag{66}$$

Таблица 8: Описательная статистика для температурных данных (Хьюстон, весна)

Характеристика	**Значение**
Объём выборки n	92
Минимум	6.67°C
Максимум	27.78°C
Медиана	22.78°C

Описательная статистика.

Выборочные моменты.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 21.5882 \tag{67}$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 22.4408 \tag{68}$$

Оценки параметра θ (при известном $\mu = 22.5$). Оценка методом моментов и методом максимального правдоподобия:

$$\hat{\theta}_{MM} = \hat{\theta}_{ML} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2} = \sqrt{\frac{1}{92}\sum_{i=1}^{92}(X_i - 22.5)^2} = 4.8241 \tag{69}$$

Оптимальная оценка. Оптимальная оценка для параметрической функции $\tau(\theta) = \theta^2$:

$$\hat{\tau}_{opt} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = 23.2723 \tag{70}$$

Оценка для θ:

$$\hat{\theta} = \sqrt{\hat{\tau}_{opt}} = 4.8241 \tag{71}$$

Сравнение с теоретическими значениями.

[2]https://github.com/fivethirtyeight/data/tree/master/us-weather-history

Таблица 9: Сравнение выборочных и теоретических характеристик (нормальное)

Характеристика	**Теор.** ($\mu = 22.5$, $\theta = 4.0$)	**Выборочное**	**Разница**
$\mathbb{E}[\xi] = \mu$	22.5000	21.5882	0.9118
$\tau(\theta) = \theta^2$	16.0000	23.2723	7.2723
θ	4.0000	4.8241	0.8241

Вывод. Температурные данные демонстрируют хорошее согласие с теоретической моделью:

- Выборочное среднее $\bar{X} = 21.59$°C близко к теоретическому значению $\mu = 22.5$°C (разница менее 1°C, что составляет около 4%).
- Оценка стандартного отклонения $\hat{\theta} = 4.82$ близка к теоретическому значению $\theta = 4.0$ (отклонение около 20%).

Несколько большее значение $\hat{\theta}$ объясняется климатическими особенностями весеннего периода, когда температурные колебания более выражены из-за смены сезонов. Тем не менее, нормальное распределение является адекватной моделью для описания температурных данных.

3.4.3 Общий вывод

Проведённый анализ **реальных данных** из открытых датасетов (UCI Online Retail и FiveThirtyEight Weather) подтверждает практическую применимость методов оценивания параметров:

1. **Геометрическое распределение**: данные о конверсии клиентов интернет-магазина (UCI Online Retail Dataset) хорошо описываются геометрическим распределением. При пороге крупной покупки £8 оценка $\hat{\theta} = 0.47$ близка к теоретическому значению $\theta = 0.4$.
2. **Нормальное распределение II**: температурные данные города Хьюстон (FiveThirtyEight) за весенний период согласуются с нормальным распределением. Выборочное среднее $\bar{X} = 21.6$°C и оценка $\hat{\theta} = 4.8$ близки к теоретическим параметрам $\mu = 22.5$, $\theta = 4.0$.
3. Методы моментов и максимального правдоподобия дают совпадающие оценки, что подтверждает теоретические результаты.
4. Наблюдаемые расхождения между выборочными и теоретическими характеристиками (10–20%) являются типичными для реальных данных и объясняются конечным объёмом выборок и неидеальностью моделей.

Результаты демонстрируют, что теоретические модели распределений успешно применимы для анализа реальных явлений из области электронной коммерции и метеорологии.

4 Домашнее задание 4: Проверка статистических гипотез

4.1 Исходные данные

Для проверки статистических гипотез используются выборки объёмов $n \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$, сгенерированные из:

- **Геометрического распределения** с параметром $\theta = 0.4$
- **Нормального распределения II** с параметрами $\mu = 22.5$, $\theta = 4.0$

4.2 Проверка гипотезы о виде распределения

4.2.1 Теоретические основы

Постановка задачи. Задача проверки гипотезы о виде распределения состоит в следующем: по данным выборки $X_1, \ldots, X_n$ требуется проверить гипотезу

$$H_0 : F(x) = F_0(x), \quad x \in \mathbb{R} \tag{72}$$

против альтернативы

$$H_1 : F(x) \neq F_0(x) \text{ хотя бы для некоторого } x \in \mathbb{R} \tag{73}$$

Здесь различают два случая:

1. **Простая гипотеза**: функция распределения $F_0(x)$ полностью известна (включая все параметры)
2. **Сложная гипотеза**: вид функции $F_0(x)$ известен, но параметры неизвестны

4.2.2 Критерий согласия Колмогорова (Смирнова)

Определение статистики. Статистика критерия Колмогорова определяется как

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \tag{74}$$

где $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$ — эмпирическая функция распределения, $F_0(x)$ — предполагаемая функция распределения.

Теорема Колмогорова (простая гипотеза). Если $F_0(x)$ — непрерывная функция распределения, то при верности гипотезы H_0 распределение статистики $\sqrt{n}D_n$ сходится к распределению Колмогорова:

$$\lim_{n \to \infty} P(\sqrt{n}D_n \leq t) = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2} \tag{75}$$

Критерий: **отвергаем** H_0 при уровне значимости α, если

$$D_n > k_{1-\alpha}/\sqrt{n} \tag{76}$$

где $k_{1-\alpha}$ — квантиль уровня $1 - \alpha$ распределения Колмогорова.

Поправка Большева. Для повышения точности приближения при конечных выборках используется модифицированная статистика:

$$S = (6nD_n + 1)\sqrt{\frac{1}{6\sqrt{n}}} \tag{77}$$

которая сходится к распределению Колмогорова быстрее.

4.2.3 Критерий согласия хи-квадрат (Пирсона)

Определение статистики. Процедура критерия хи-квадрат состоит из следующих этапов:

1. **Группировка данных**: область значений разбивается на k интервалов:

$$(a_0, a_1], (a_1, a_2], \ldots, (a_{k-1}, a_k) \tag{78}$$

2. **Подсчёт частот**: для каждого интервала j подсчитывается число наблюдений ν_j, попавших в этот интервал.
3. **Вычисление ожидаемых частот**:

$$np_j = n \cdot P_0(X \in (a_{j-1}, a_j]) = n \cdot (F_0(a_j) - F_0(a_{j-1})) \tag{79}$$

4. **Статистика Пирсона**:

$$\chi^2 = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j} \tag{80}$$

Теорема Пирсона (простая гипотеза). При верности гипотезы H_0, при условии, что $np_j \geq 5$ для всех j, статистика χ^2 асимптотически следует распределению $\chi^2(k-1)$ с $k-1$ степенями свободы:

$$\chi^2 \xrightarrow{d} \chi^2(k-1) \quad \text{при } n \to \infty \tag{81}$$

Критерий: **отвергаем** H_0 при уровне значимости α, если

$$\chi^2 > \chi^2_{1-\alpha}(k-1) \tag{82}$$

Правило Старджесса для выбора числа интервалов. Рекомендуемое число интервалов:

$$k = 1 + \log_2 n \approx 1 + 3.322 \lg n \tag{83}$$

4.2.4 Сложная гипотеза: случай неизвестных параметров

Процедура проверки. При проверке сложной гипотезы выполняются следующие шаги:

1. **Оценивание параметров**: по выборке вычисляется оценка $\hat{\theta}$ неизвестного параметра θ, обычно методом максимального правдоподобия.

2. **Построение гипотетического распределения**: конструируется функция $F_{\hat{\theta}}(x)$ с подставленной оценкой.

3. **Вычисление статистики**: вычисляется значение выбранной статистики критерия (Колмогорова или хи-квадрат) с использованием $F_{\hat{\theta}}(x)$ вместо $F_0(x)$.

4. **Учёт потери степеней свободы**: при использовании критерия хи-квадрат число степеней свободы уменьшается на m (число оценённых параметров):

$$\chi^2 \xrightarrow{d} \chi^2(k-1-m) \tag{84}$$

5. **Определение критической области**: для критерия Колмогорова используются специальные таблицы (если они доступны для данного распределения) или применяется асимптотический подход.

Практический подход при отсутствии таблиц. Если для рассматриваемого распределения не известны точные предельные распределения статистики при сложной гипотезе, используется следующий подход:

- Разделить выборку на две части: первая (объёма n_1) служит для оценивания параметров, вторая (объёма n_2) — для проверки гипотезы.
- По первой части вычислить $\hat{\theta}$.
- По второй части вычислить статистику критерия, используя $F_{\hat{\theta}}(x)$ как известную функцию.
- Применить стандартные процедуры проверки простой гипотезы.

4.2.5 Геометрическое распределение

Проверка простой гипотезы ($\theta = 0.4$ известен). Функция распределения геометрического распределения:

$$F_0(x) = 1 - (1-\theta)^{\lfloor x \rfloor} = 1 - 0.6^{\lfloor x \rfloor}, \quad x \geq 1 \tag{85}$$

Критерий Колмогорова:
Вычисляем

$$D_n = \max_i \left\{ \max \left\{ \frac{i}{n} - F_0(X_{(i)}), F_0(X_{(i)}) - \frac{i-1}{n} \right\} \right\} \tag{86}$$

где $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ — порядковые статистики.
Статистика с поправкой Большева:

$$S = (6nD_n + 1)/(6\sqrt{n}) \tag{87}$$

Критическое значение на уровне $\alpha = 0.05$: $k_{0.95} \approx 1.36$ (из таблиц распределения Колмогорова).

Критерий хи-квадрат:
Число интервалов: $k = 1 + \log_2 n$ (с учётом того, что геометрическое распределение дискретно).
Интервалы: $\{1\}, \{2\}, \{3\}, \ldots, \{k-1\}, \{k, k+1, \ldots\}$
Вероятности:

$$p_j = P(\xi = j) = 0.4 \cdot 0.6^{j-1}, \quad j = 1, 2, \ldots, k-1 \tag{88}$$

$$p_k = P(\xi \geq k) = 0.6^{k-1} \tag{89}$$

Статистика Пирсона:

$$\chi^2 = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j} \tag{90}$$

Число степеней свободы: $\nu = k - 1$.

Проверка сложной гипотезы (θ неизвестен). 1. Оцениваем параметр: $\hat{\theta} = 1/\bar{X}$ (ММП-оценка).

2. Строим функцию распределения: $F_{\hat{\theta}}(x) = 1 - (1-\hat{\theta})^{\lfloor x \rfloor}$.

3. Вычисляем статистики критериев Колмогорова и хи-квадрат с $F_{\hat{\theta}}(x)$.

4. Для хи-квадрата число степеней свободы: $\nu = k - 1 - 1 = k - 2$ (вычтено одно: число оценённых параметров равно 1).

Таблица 10: Проверка гипотез для геометрического распределения ($\theta = 0.4$)

n	D_n	S	**Выв. K**	χ^2	$\chi^2_{0.95}$	**Выв.** χ^2
5	0.4000	2.0333	принимается	0.8444	5.9915	принимается
10	0.4000	2.8520	принимается	1.7093	7.8147	принимается
100	0.4000	8.9517	отвергается	5.7274	14.0671	принимается
200	0.4000	12.6544	отвергается	7.4788	15.5073	принимается
400	0.4000	17.8923	отвергается	23.4995	16.9190	отвергается
600	0.4000	21.9119	отвергается	18.0282	18.3070	принимается
800	0.4000	25.3009	отвергается	19.2401	18.3070	отвергается
1000	0.4000	28.2866	отвергается	28.4846	18.3070	отвергается
Выв.: вывод (H0 отвергается/принимается)						

Таблица результатов проверки гипотез для геометрического распределения.

Таблица 11: Проверка сложной гипотезы для геометрического распределения

n	$\hat{\theta}$	D_n	**Выв. K**	χ^2	$\chi^2_{0.95}(k-2)$	**Выв.** χ^2
5	0.4902	0.4902	принимается	0.8444	3.8415	принимается
10	0.4762	0.4762	отвергается	1.7093	5.9915	принимается
100	0.3915	0.3915	отвергается	5.7274	12.5916	принимается
200	0.3897	0.3897	отвергается	7.4788	14.0671	принимается
400	0.3945	0.3945	отвергается	23.4995	15.5073	отвергается
600	0.3980	0.3980	отвергается	18.0282	16.9190	отвергается
800	0.3964	0.3964	отвергается	19.2401	16.9190	отвергается
1000	0.3982	0.3982	отвергается	28.4846	16.9190	отвергается

Таблица результатов для сложной гипотезы (геометрическое).

4.2.6 Нормальное распределение II

Проверка простой гипотезы ($\mu = 22.5$, $\theta = 4.0$ известны). Функция распределения:

$$F_0(x) = \Phi\left(\frac{x - 22.5}{4.0}\right) \tag{91}$$

где $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$ — функция распределения стандартного нормального закона.

Критерий Колмогорова:

Вычисляется по стандартной формуле с использованием $F_0(x)$.

Критерий хи-квадрат:

Число интервалов: $k = 1 + \log_2 n$.

Интервалы выбираются так, чтобы в каждый попало примерно одинаковое число наблюдений (равновероятные интервалы).

Ожидаемые частоты: $np_j = n/k$ для равновероятного разбиения.

Проверка сложной гипотезы (μ и θ неизвестны). 1. Оцениваем параметры:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{92}$$

$$\hat{\theta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{93}$$

Важно: используется смещённая оценка дисперсии (с делением на n, а не на $n-1$) для согласованности с теорией.

2. Строим стандартизованные данные:

$$Z_i = \frac{X_i - \hat{\mu}}{\hat{\theta}} \tag{94}$$

3. Проверяем гипотезу о том, что Z_i имеют стандартное нормальное распределение $N(0,1)$.

4. Для хи-квадрата число степеней свободы: $\nu = k - 1 - 2$ (вычтено 2: два оценённых параметра).

Таблица 12: Проверка гипотез для нормального распределения II ($\mu = 22.5$, $\theta = 4.0$)

n	D_n	S	**Выв. K**	χ^2	$\chi^2_{0.95}$	**Выв.** χ^2
5	0.1375	0.7208	принимается	0.1200	7.8147	принимается
10	0.0957	0.7005	принимается	0.0000	9.4877	принимается
100	0.0381	0.8593	принимается	0.0320	14.0671	принимается
200	0.0339	1.0769	принимается	0.0080	15.5073	принимается
400	0.0146	0.6545	принимается	0.0000	16.9190	принимается
600	0.0176	0.9694	принимается	0.0080	18.3070	принимается
800	0.0145	0.9181	принимается	0.0070	18.3070	принимается
1000	0.0117	0.8315	принимается	0.0060	18.3070	принимается

Таблица результатов проверки гипотез для нормального распределения II.

Таблица 13: Проверка сложной гипотезы для нормального распределения II

n	$\hat{\mu}$	$\hat{\theta}$	D_n	**Выв. K**	χ^2	**Выв.** χ^2
5	23.0532	5.2716	0.0630	принимается	0.1200	принимается
10	22.6372	4.5961	0.0943	принимается	0.0000	принимается
100	22.3163	4.1282	0.0222	принимается	0.0320	принимается
200	22.3370	4.1247	0.0199	принимается	0.0080	принимается
400	22.4632	4.0680	0.0139	принимается	0.0000	принимается
600	22.5284	4.0639	0.0202	принимается	0.0080	принимается
800	22.4776	4.0063	0.0122	принимается	0.0070	принимается
1000	22.4798	3.9859	0.0097	принимается	0.0060	принимается

Таблица результатов для сложной гипотезы (нормальное).

4.3 Проверка гипотезы об однородности выборок

4.3.1 Теоретические основы

Постановка задачи. Имеются две независимые выборки $X_1^{(1)}, \ldots, X_{n_1}^{(1)}$ и $X_1^{(2)}, \ldots, X_{n_2}^{(2)}$ с эмпирическими функциями распределения $F_{n_1}(x)$ и $F_{n_2}(x)$ соответственно.

Гипотеза однородности:

$$H_0 : F_1(x) = F_2(x) \quad \text{для всех } x \in \mathbb{R} \tag{95}$$

против альтернативы

$$H_1 : F_1(x) \neq F_2(x) \quad \text{хотя бы для некоторого } x \in \mathbb{R} \tag{96}$$

Статистика Смирнова. Статистика Смирнова определяется как

$$D_{n_1,n_2} = \sup_{x \in \mathbb{R}} |F_{n_1}(x) - F_{n_2}(x)| \tag{97}$$

При верности гипотезы H_0 распределение статистики

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot D_{n_1,n_2} \tag{98}$$

асимптотически совпадает с распределением Колмогорова.

4.3.2 Анализ однородности для сгенерированных выборок

Метод вычисления. Для каждой пары выборок объёмов n_i и n_j (где $i, j \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$, $i < j$) вычисляется:

1. Эмпирические функции распределения $F_{n_i}(x)$ и $F_{n_j}(x)$.
2. Статистика Смирнова:

$$D_{n_i,n_j} = \max_k \left| F_{n_i}(X_{(k)}) - F_{n_j}(X_{(k)}) \right| \tag{99}$$

3. Нормированная статистика:

$$\sqrt{\frac{n_i n_j}{n_i + n_j}} \cdot D_{n_i,n_j} \tag{100}$$

4. Сравнение с критическим значением $k_{0.95} \approx 1.36$ на уровне значимости $\alpha = 0.05$.

Таблица значений D_{n_1,n_2} для геометрического распределения.

Таблица значений D_{n_1,n_2} для нормального распределения II.

Вывод. При увеличении объёма выборок значения D_{n_1,n_2} должны монотонно убывать, что отражает сходимость эмпирических функций распределения к истинной функции распределения. Все выборки должны быть однородны друг другу в асимптотическом смысле, поскольку они сгенерированы из одного распределения.

Таблица 14: Статистика Смирнова для геометрического распределения

n_1	n_2	D_{n_1,n_2}	**Вывод однородности**
5	10	0.0400	однородны
5	100	0.1040	однородны
5	200	0.1040	однородны
5	400	0.1000	однородны
5	600	0.0953	однородны
5	800	0.0988	однородны
5	1000	0.0954	однородны
10	100	0.0840	однородны
10	200	0.0840	однородны
10	400	0.0800	однородны
10	600	0.0753	однородны
10	800	0.0788	однородны
10	1000	0.0754	однородны
100	200	0.0160	однородны
100	400	0.0235	однородны
100	600	0.0340	однородны
100	800	0.0275	однородны
100	1000	0.0318	однородны
200	400	0.0185	однородны
200	600	0.0290	однородны
200	800	0.0225	однородны
200	1000	0.0268	однородны
400	600	0.0105	однородны
400	800	0.0043	однородны
400	1000	0.0083	однородны
600	800	0.0070	однородны
600	1000	0.0058	однородны
800	1000	0.0043	однородны

Таблица 15: Статистика Смирнова для нормального распределения II

n_1	n_2	D_{n_1,n_2}	**Вывод однородности**
5	10	0.1000	однородны
5	100	0.1520	однородны
5	200	0.1520	однородны
5	400	0.1385	однородны
5	600	0.1313	однородны
5	800	0.1370	однородны
5	1000	0.1376	однородны
10	100	0.0760	однородны
10	200	0.0800	однородны
10	400	0.0845	однородны
10	600	0.0893	однородны
10	800	0.0890	однородны
10	1000	0.0898	однородны
100	200	0.0180	однородны
100	400	0.0325	однородны
100	600	0.0333	однородны
100	800	0.0352	однородны
100	1000	0.0338	однородны
200	400	0.0275	однородны
200	600	0.0257	однородны
200	800	0.0243	однородны
200	1000	0.0246	однородны
400	600	0.0108	однородны
400	800	0.0127	однородны
400	1000	0.0132	однородны
600	800	0.0123	однородны
600	1000	0.0118	однородны
800	1000	0.0050	однородны

4.4 Проверка гипотез для реальных данных

В данном разделе проводится проверка статистических гипотез о виде распределения для реальных данных, проанализированных в разделе 3.3 (дз), в данной работе раздел 3.4. Используются критерий Колмогорова-Смирнова и критерий хи-квадрат (χ^2) Пирсона. Уровень значимости: $\alpha = 0.05$.

4.4.1 Геометрическое распределение: данные о конверсии клиентов

Источник данных. Используются данные UCI Online Retail Dataset (раздел 3.3): число покупок клиента до первой покупки, превышающей порог £8. Объём выборки $n = 500$, минимальное значение 1, максимальное 49.

Проверка простой гипотезы. Гипотеза:

$$H_0 : \xi \sim \text{Geom}(\theta = 0.4) \quad \text{против} \quad H_1 : \xi \not\sim \text{Geom}(\theta = 0.4) \tag{101}$$

Критерий Колмогорова-Смирнова.
Статистика критерия:

$$D_n = \sup_x |F_n(x) - F_0(x)| \tag{102}$$

где $F_n(x)$ — эмпирическая функция распределения, $F_0(x) = 1 - (1-\theta)^{\lfloor x \rfloor}$ — теоретическая функция распределения геометрического распределения.

Таблица 16: Результаты критерия Колмогорова (простая гипотеза, геометрическое)

Характеристика	**Значение**
Статистика D_n	0.4280
Критическое значение $k_{0.95}/\sqrt{n}$	0.0608
Вывод	H_0 **отвергается**

Критерий хи-квадрат.
Статистика критерия:

$$\chi^2 = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j} \tag{103}$$

где ν_j — наблюдаемые частоты, np_j — ожидаемые частоты при H_0.

Таблица 17: Результаты критерия χ^2 (простая гипотеза, геометрическое)

Характеристика	**Значение**
Статистика χ^2	458.6966
Число степеней свободы ν	9
Критическое значение $\chi^2_{0.95}(9)$	16.9190
Вывод	H_0 **отвергается**

Проверка сложной гипотезы. Гипотеза:

$$H_0 : \xi \sim \text{Geom}(\theta),\ \theta \text{ неизвестен} \quad \text{против} \quad H_1 : \xi \not\sim \text{Geom}(\theta) \tag{104}$$

Оценка параметра по данным: $\hat{\theta} = 1/\bar{X} = 0.4677$.

Примечание: При проверке сложной гипотезы число степеней свободы для критерия χ^2 уменьшается на 1 (число оцениваемых параметров).

Таблица 18: Результаты проверки сложной гипотезы (геометрическое)

Критерий	**Статистика**	**Крит. знач.**	**Вывод**
Колмогоров-Смирнов	$D_n = 0.4677$	0.0608	H_0 **отвергается**
Хи-квадрат	$\chi^2 = 387.90$, $\nu = 6$	12.5916	H_0 **отвергается**

Вывод по геометрическому распределению. Оба критерия (Колмогорова-Смирнова и хи-квадрат) **отвергают** гипотезу о геометрическом распределении данных как при известном $\theta = 0.4$, так и при оценённом $\hat{\theta} = 0.47$.

Это объясняется особенностями реальных данных:

- Наличие значительных выбросов (максимальное значение 49 при теоретическом ожидании около 2–3).
- Высокая дисперсия данных ($S^2 = 18.2$), существенно превышающая теоретическую ($\mathbb{D}[\xi] = 3.75$).
- Реальное поведение клиентов не полностью соответствует модели независимых испытаний Бернулли.

Несмотря на отвержение гипотезы, геометрическое распределение может использоваться как *приближённая модель* для описания данных о конверсии, однако для точного моделирования требуются более сложные модели.

4.4.2 Нормальное распределение II: температурные данные

Источник данных. Используются данные FiveThirtyEight Weather (раздел 3.3): среднесуточная температура в Хьюстоне за весенний период (март–май). Объём выборки $n = 92$, выборочное среднее $\bar{X} = 21.59°\text{C}$.

Проверка простой гипотезы. Гипотеза:

$$H_0 : \xi \sim N(\mu = 22.5, \theta^2 = 16) \quad \text{против} \quad H_1 : \xi \not\sim N(\mu = 22.5, \theta^2 = 16) \tag{105}$$

Критерий Колмогорова-Смирнова.

Таблица 19: Результаты критерия Колмогорова (простая гипотеза, нормальное)

Характеристика	**Значение**
Статистика D_n	0.0972
p-value	0.3289
Критическое значение $k_{0.95}/\sqrt{n}$	0.1418
Вывод	H_0 **принимается**

Критерий хи-квадрат.

Таблица 20: Результаты критерия χ^2 (простая гипотеза, нормальное)

Характеристика	**Значение**
Статистика χ^2	10.3500
Число степеней свободы ν	4
Критическое значение $\chi^2_{0.95}(4)$	9.4877
Вывод	H_0 **отвергается**

Проверка сложной гипотезы. Гипотеза:

$$H_0 : \xi \sim N(\mu, \theta^2),\ \mu, \theta \text{ неизвестны} \quad \text{против} \quad H_1 : \xi \not\sim N(\mu, \theta^2) \tag{106}$$

Оценки параметров по данным: $\hat{\mu} = 21.5882$, $\hat{\theta} = 4.7372$.

Таблица 21: Результаты проверки сложной гипотезы (нормальное)

Критерий	**Статистика**	**Крит. знач.**	**Вывод**
Колмогоров-Смирнов	$D_n = 0.1220$, $p = 0.119$	0.1418	H_0 **принимается**
Хи-квадрат	$\chi^2 = 18.60$, $\nu = 3$	7.8147	H_0 **отвергается**

Примечание: При проверке сложной гипотезы число степеней свободы для критерия χ^2 уменьшается на 2 (число оцениваемых параметров).

Вывод по нормальному распределению. Критерии дают **противоречивые результаты**:

- **Критерий Колмогорова-Смирнова принимает** гипотезу о нормальности как для простой ($p = 0.33$), так и для сложной ($p = 0.12$) гипотез.
- **Критерий хи-квадрат отвергает** гипотезу в обоих случаях.

Такое расхождение объясняется следующими факторами:

1. Критерий Колмогорова-Смирнова более чувствителен к отклонениям в центральной части распределения, тогда как критерий χ^2 учитывает отклонения на хвостах.

2. Небольшой объём выборки ($n = 92$) и выбор числа интервалов влияют на мощность критерия χ^2.

3. Температурные данные имеют слабую сезонную компоненту, которая нарушает строгую нормальность.

В целом, **нормальное распределение является приемлемой моделью** для температурных данных, что подтверждается критерием Колмогорова-Смирнова. Отвержение критерием χ^2 может быть связано с особенностями разбиения на интервалы и малым числом степеней свободы.

4.4.3 Общий вывод

Проведённый анализ реальных данных с использованием критериев проверки гипотез показал:

1. **Геометрическое распределение** (данные о конверсии клиентов): гипотеза о геометрическом распределении **отвергается** обоими критериями. Реальные данные имеют большую дисперсию и более тяжёлые хвосты, чем предсказывает модель. Для практических целей геометрическое распределение может использоваться как приближение, но точное моделирование требует более сложных моделей.

2. **Нормальное распределение II** (температурные данные): критерий Колмогорова-Смирнова **принимает** гипотезу о нормальности ($p > 0.05$), тогда как критерий χ^2 её отвергает. Учитывая большую мощность критерия Колмогорова для непрерывных распределений, можно заключить, что нормальное распределение является **адекватной моделью** для температурных данных.

3. Использование реальных данных продемонстрировало, что теоретические модели распределений являются идеализацией, и реальные явления могут лишь приближённо описываться классическими распределениями.