

Практика 5.
Задания по теме:
Критерий Колмогорова-Смирнова для сравнения
распределений
и проверки соответствия распределения
нормальному.

Проверка типа распределения при помощи
статистических пакетов (MS Excel, Statistica)

Блок 1. Критерий Колмогорова–Смирнова: общие принципы

Задание 1 Эмпирическая функция распределения

По выборке объёма $n = 5$: 2, 5, 5, 7, 9.

- (а) Постройте эмпирическую функцию распределения $F_n(x)$ (ступенчатую) и нарисуйте её график.
- (б) Чему равно значение $F_n(5)$? $F_n(6)$?

Задание 2 Статистика D К-С

Статистика критерия Колмогорова–Смирнова для проверки согласия с теоретическим распределением определяется как $D = \sup_x |F_n(x) - F_0(x)|$, где $F_0(x)$ – теоретическая функция распределения.

- (а) Объясните словами, что означает «супремум модуля разности».
- (б) На каком графике можно визуально оценить эту разность?
- (с) Как связаны D и p -значение? (Чем больше D , тем _____ p -значение.)

Задание 3 Критические значения критерия

Для выборки объёма $n = 50$ и уровня значимости $\alpha = 0.05$ критическое значение D приблизительно равно $\frac{1.36}{\sqrt{n}}$.

- (а) Вычислите D .
- (б) Если вычисленное D оказалось равным 0.25, отвергается ли гипотеза о согласии с теоретическим распределением?

Блок 2. Проверка соответствия распределения нормальному (тест Колмогорова–Смирнова)

Задание 4 Применение теста к реальной выборке

Дана выборка результатов тестирования (баллы): 12, 15, 14, 18, 22, 17, 13, 19, 20, 21, 16, 1

- (a) Вычислите выборочное среднее $\bar{x}$ и выборочное стандартное отклонение s .
- (b) Нормируйте данные: $z_i = \frac{x_i - \bar{x}}{s}$. Сравните эмпирическую функцию распределения $F_n(z)$ с теоретической функцией стандартного нормального распределения $\Phi(z)$.
- (c) (Описательно) Оцените, достаточно ли мала максимальная разность D , чтобы не отвергать гипотезу о нормальности при $\alpha = 0.05$.

Задание 5 Интерпретация результатов из статистического пакета

Представьте, что в программе Statistica вы получили следующий вывод теста К-С на нормальность:

$$D = 0.124, \quad p = 0.082$$

Уровень значимости $\alpha = 0.05$.

- (a) Можно ли считать распределение нормальным? Аргументируйте.
- (b) Как изменится вывод, если выбрать $\alpha = 0.10$?
- (c) Что означает $p = 0.082$ для исследователя?

Задание 6 Сравнение теста К-С с другими критериями нормальности

Назовите два других критерия проверки нормальности (например, Шапиро–Уилка, Харке–Бера). В чём преимущество критерия К-С перед ними? В чём недостаток? (Указание: критерий К-С чувствителен к различиям в центре распределения, но может быть менее мощным на малых выборках.)

Блок 3. Критерий Колмогорова–Смирнова для сравнения двух распределений

Задание 7 Двухвыборочный тест К-С

Две выборки: Выборка А: 3, 5, 6, 8, 10, Выборка В: 7, 9, 11, 12, 13.

- (a) Постройте графики эмпирических функций распределения для А и В на одной координатной плоскости.
- (b) Найдите максимальное вертикальное расстояние между ними (визуально или расчётом).
- (c) Сделайте предположение: различаются ли распределения? (Без таблиц, качественно.)

Задание 8 Применение в медицине

Исследователь сравнил время восстановления после операции в двух группах пациентов (контроль и новая терапия). Распределение времени в обеих группах оказалось не нормальным. Почему для проверки гипотезы о

равенстве распределений можно использовать двухвыборочный критерий Колмогорова–Смирнова? Какие преимущества он даёт по сравнению с U-критерием Манна–Уитни?

Блок 4. Проверка типа распределения в статистических пакетах (MS Excel, Statistica)

Задание 9 Работа в MS Excel: гистограмма и тест на нормальность

Опишите пошагово (в виде инструкции), как в MS Excel:

- Построить гистограмму для выборки (используя «Пакет анализа» -> «Гистограмма»).
- Визуально оценить, похоже ли распределение на нормальное.
- Выполнить тест на нормальность (например, с помощью функции = НОРМ.ТЕСТ? В Excel нет встроенного теста К-С, но можно использовать надстройку «Пакет анализа» или описать вычисление через сортировку и расчёт эмпирической функции). Уточните: в Excel прямого теста К-С нет, но можно вычислить статистику вручную или использовать сторонние надстройки.
- Какие ограничения есть у Excel для проверки типа распределения по сравнению со специализированными пакетами?

Задание 10 Работа в Statistica

В пакете Statistica есть модуль «Базовые статистики и таблицы» -> «Проверка нормальности».

- Перечислите не менее трёх критериев нормальности, которые можно получить в этом модуле (включая К-С).
- Для выборки из задания 4 (Блок 2) предположите, какие значения р-значений выдали бы разные критерии. Могут ли они противоречить друг другу? Почему?
- Как в Statistica интерпретировать Q-Q график (квантиль-квантиль) при проверке нормальности? Что означает отклонение точек от прямой линии?

Задание 11 Сравнение пакетов

Заполните таблицу:

Функция / возможность	MS Excel	Statistica
Построение гистограммы	_____	_____
Q-Q plot (график нормальности)	_____	_____
Тест Колмогорова–Смирнова (встроенный)	_____	_____
Тест Шапиро–Уилка	_____	_____
Возможность сравнения двух распределений (К-С)	_____	_____

(Ответ: +/– или «требуется надстройка», «есть», «нет».)

Комплексное задание (по всем темам)

Задание 12 Анализ реальных данных с использованием пакетов

Вам даны результаты измерения уровня сахара в крови (ммоль/л) у 40 пациентов:

5.2, 5.5, 5.3, 5.8, 6.1, 5.9, 5.7, 5.4, 6.0, 5.6, 5.2, 5.9, 6.2, 5.5, 5.3, 5.7, 6.1, 5.8, 5.4, 6.0, 5.6, 5.5, 5.9

- (a) В MS Excel постройте гистограмму (интервалы выберите самостоятельно) и Q-Q график (с использованием инструментов «Надстройки» либо описательно). Сделайте вывод о близости к нормальному распределению.
- (b) В Statistica (или мысленно) выполните тест Колмогорова–Смирнова на нормальность. Предположим, получено $p = 0.32$. Что это значит?
- (c) Если бы p оказалось меньше 0.05, как бы вы поступили? Какие дальнейшие действия можно предпринять (например, преобразование данных, использование непараметрических методов)?

Задание 13 Сравнение двух выборок с помощью К-С в статистическом пакете

Исследователь хочет проверить, различаются ли распределения роста у мужчин и женщин. Он получил две выборки объёмом по 100 человек. В пакете Statistica применил двухвыборочный тест Колмогорова–Смирнова и получил $D = 0.22$ и $p = 0.003$.

- (a) Сформулируйте H_0 и H_1 для этого теста.
- (b) Отвергается ли H_0 на уровне $\alpha = 0.05$? А на уровне $\alpha = 0.01$? Интерпретируйте результат содержательно.
- (c) Какой график лучше всего проиллюстрирует различие между двумя распределениями? (Назовите тип графика, который строят в статистических пакетах для визуализации двух эмпирических распределений.)

Задание 14 Практическое задание: пошаговый расчёт К-С в Excel

Дана малая выборка: 1, 2, 3, 4, 5. Проверяем гипотезу о том, что данные извлечены из равномерного распределения на $[0, 6]$. Теоретическая функция распределения $F_0(x) = x/6$ для $0 \leq x \leq 6$.

- (a) Вычислите $F_n(x)$ для каждого значения выборки.
- (b) Найдите максимальную разность $|F_n(x) - F_0(x)|$ (рассчитайте таблицу со значениями x_i , $F_n(x_i - 0)$, $F_0(x_i)$, ...).
- (c) Сравните полученный D с критическим значением для $n = 5$ при $\alpha = 0.05$ (табличное значение приблизительно 0.563). Отвергается ли гипотеза?
- (d) Опишите, как выполнить аналогичные вычисления в MS Excel (с использованием функций сортировки, ЕСЛИ и т.д.).