Зачет по МФК Математическая статистика и анализ данных 2025

Чжан Инхао

Задание 1

1. **Цель исследования**
   Анализ вероятности задержек рейсов при прибытии на десять наиболее популярных направлений, вылетающих из трёх крупнейших аэропортов Нью-Йорка (JFK, LGA и EWR). В частности, поставлена задача определить, какие из этих аэропортов характеризуются наивысшим и наинизшим риском задержек. Полученные результаты смогут послужить основой для оптимизации расписания авиакомпаний и поддержки выбора маршрутов пассажирами на основе объективных данных, а также выявить особенности операционной эффективности маршрутов с высоким пассажиропотоком.
2. **Процесс анализа**
   Данные эксперимента основаны на наборе данных рейсов (flights.csv). Сначала были отобраны записи о вылетах из аэропортов Нью-Йорка — всего 336 776 записей. Для обеспечения точности расчетов произведена очистка пропусков в переменной «время задержки прибытия» (arr_delay): удалены записи с отсутствующими значениями, после чего в распоряжении осталось 327 346 наблюдений (доля пропусков — 2.8 %).
   Отбор десяти самых популярных аэропортов назначения выполнен на основе подсчёта числа рейсов для каждого аэропорта.

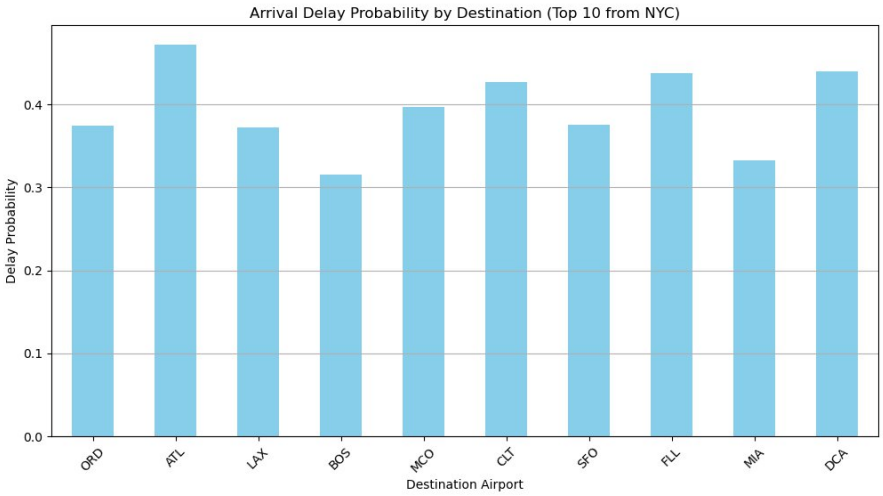

Рисунок 1.1 результаты визуализированы с помощью столбчатой диаграммы

3. **Вывод**
   Аэропорт с наибольшей вероятностью задержки рейсов — ATL, где доля задержек составляет 47.19 %.
   Аэропорт с наименьшей вероятностью задержки рейсов — BOS, где доля задержек равна 31.57 %.

Задание 2

1. **Цель исследования**

   Основная задача данного эксперимента — проверить, подчиняются ли времена полёта рейсов из Нью-Йорка (аэропорты JFK, LGA, EWR) в Международный аэропорт Сан-Франциско (SFO) нормальному распределению и количественно охарактеризовать его параметры. В рамках анализа требуется ответить на следующие вопросы: соответствует ли распределение времени полёта на маршруте Нью-Йорк—Сан-Франциско предположению о нормальности; если да, каковы его математическое ожидание, стандартное отклонение и 95-процентный доверительный интервал. Результаты исследования помогут авиакомпаниям оптимизировать расписания рейсов и объективно оценивать летную эффективность.

2. **Процесс анализа**

   Данные для эксперимента взяты из файла flight.csv. Сначала отобраны записи о рейсах с вылетом из трёх аэропортов Нью-Йорка (JFK, LGA, EWR) и приземлением в международном аэропорту Сан-Франциско (SFO). Для надёжности анализа удалены записи с отсутствующим значением времени полёта (air_time). В итоге в выборке осталось 5200 наблюдений.

   Распределение времени полёта было предварительно проверено путём построения нормализованного гистограммы (20 бинов) с наложением теоретической кривой нормального распределения с параметрами $\mu = 346$ и $\sigma = 17$. На рисунке 2.1 видно, что основная масса значений сосредоточена в интервале от 330 до 360 минут, распределение имеет один пик и почти симметрично, однако правая «хвостовая» часть несколько длиннее: отдельные рейсы превышают 370 минут, что может быть связано с экстремальными погодными условиями или перегрузкой воздушного пространства.

   Для более тщательной проверки гипотезы нормальности построена Q–Q-диаграмма (рисунок 2.2). Большая часть точек располагается вдоль опорной линии, однако в области высоких квантилей наблюдается лёгкое отклонение вправо, свидетельствующее о более тяжёлых хвостах реального распределения по сравнению с теоретическим. Несмотря на это, соответствие между эмпирическими и теоретическими квантилями остаётся достаточно высоким, что подтверждает близкое к нормальному характер распределения времени полёта.

   Исходя из оценок μ и σ, вычислен 95%-доверительный интервал для времени полёта: от 311.9 до 379.5 минут. Это означает, что с вероятностью 95 % действительное время полёта находится в указанном диапазоне. Следует отметить, что точность этого интервала полностью зависит от справедливости гипотезы о нормальности распределения.

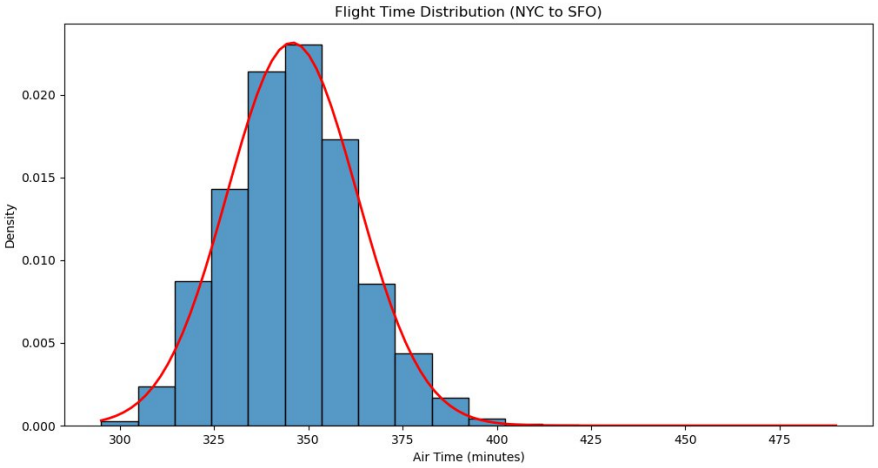

рисунок 2.1 Нормированная гистограмма распределения времени полёта

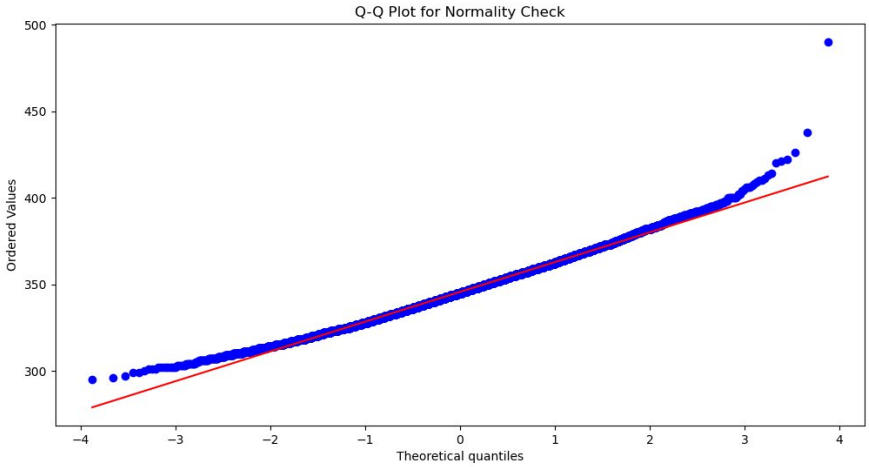

рисунок 2.2 Q–Q график времени полёта

3. Заключение

Распределение времени полёта на маршруте Нью-Йорк—Сан-Франциско можно приближённо считать нормальным с параметрами: среднее $\mu = 345.6$ минут, стандартное отклонение $\sigma = 17.2$ минут, 95%-доверительный интервал [311.9, 379.5] минут. Построенные нормированная гистограмма и Q–Q-диаграмма свидетельствуют о том, что, несмотря на лёгкое удлинение правого хвоста, основная масса данных хорошо соответствует предположению о нормальности. Отклонения в правой части распределения, вероятно, вызваны единичными факторами (неблагоприятные погодные условия, меры управления воздушным движением), поэтому рекомендуется провести стратифицированный анализ с учётом внешних данных (метеорологических сводок, загруженности аэропортов) для уточнения причин.

С практической точки зрения авиакомпании могут закладывать буферное время, равное верхней границе 95%-доверительного интервала (379.5 минут), при формировании расписания, чтобы снизить риск задержек. При этом выявленные «тяжёлые» хвосты распределения указывают на необходимость мониторинга экстремальных случаев (перенаправления, объездные маршруты) и разработки соответствующих экстренных алгоритмов. Для повышения строгости проверки нормальности целесообразно дополнительно провести тест Шапиро—Уилка или

использовать непараметрические методы оценки доверительного интервала. В дальнейшем исследование можно расширить за счёт многолетних данных для оценки стабильности характеристик распределения или внедрить модели машинного обучения для прогнозирования аномалий во времени полётов.

Задание 3

1. **Цель исследования**
   Настоящее исследование направлено на анализ закономерностей распределения времени вылета рейсов из международного аэропорта имени Джона Ф. Кеннеди (JFK) в Нью-Йорке, выявление суточных пиковых периодов и проверку наличия статистически значимых различий в задержках вылетов в различные часы пик. Основные исследовательские вопросы включают следующее: как распределяется количество вылетов в течение 24 часов? Какие временные интервалы являются двумя основными «пиковыми периодами» по объёму вылетов? Существуют ли статистически значимые различия в средних задержках вылетов между этими пиковыми периодами? Полученные результаты могут служить основой для принятия решений в сфере распределения ресурсов аэропорта и оптимизации расписания рейсов.
2. **Процесс анализа**
   Экспериментальные данные основаны на наборе данных о рейсах flights.csv, из которого были отобраны записи о вылетах из аэропорта JFK с полями «время вылета» и «задержка вылета» без пропусков. После преобразования времени вылета в часы было подсчитано количество рейсов в каждом часовом интервале и построен график распределения (рисунок 3.1). Полученные результаты показали чётко выраженное двупиковое распределение числа вылетов: первый пик приходится на 8:00 утра, второй — на 16:00 дня.
   Для сравнения различий задержек в периоды утреннего и вечернего пиковых часов были извлечены данные о задержках вылета именно в 8:00 и 16:00. Средняя задержка в утренний час пик составила 1.0 минуты, тогда как в вечерний час пик она значительно возросла до 10.9 минуты. Для проверки статистической значимости этой разницы применялся независимый t-тест для выборок с неравными дисперсиями; вычисленное значение t-статистики равно –15.2 при $P \approx 0$.
   Полученный результат позволяет отвергнуть нулевую гипотезу о равенстве средних задержек, что свидетельствует о высокой статистической значимости различий между утренними и вечерними пиковыми часами.

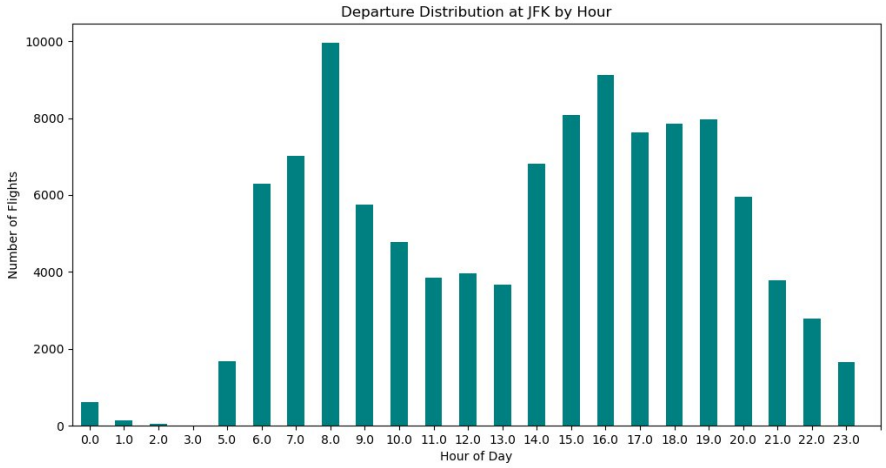

рисунок 3.1 После преобразования времени вылета в часы было подсчитано количество рейсов в каждом часовом интервале и построен график распределения

3. **Выводы**

   Распределение вылетов из аэропорта JFK характеризуется двумя пиковыми периодами — в 08:00 и в 16:00. При этом средняя задержка вылета во второй пик (16:00) существенно выше, чем в утренний пик (08:00). Результаты статистического теста ($p < 0.0001$) свидетельствуют о том, что выявленная разница практически не может быть объяснена случайными колебаниями.

Задание 4

1. **Цель исследования**

   Настоящее экспериментальное исследование направлено на изучение взаимосвязи между изменчивостью времени полёта (оцененной с помощью стандартного отклонения) и расстоянием полёта с акцентом на решение следующих задач: влияет ли увеличение расстояния на стабильность продолжительности полёта систематически? Если такая связь выявляется, носит ли она линейный характер или для более точного описания требуется включение нелинейных признаков (например, квадратичного члена)? Полученные результаты могут служить основой для количественной оценки надёжности рейсов и оптимизации расписаний.

2. **Процесс анализа**

   Экспериментальные данные основаны на наборе данных о рейсах flights.csv. Сначала записи были сгруппированы по аэропорту назначения, после чего отобраны лишь те аэропорты, для которых число рейсов превышало 30, чтобы обеспечить статистическую надёжность; в итоге в анализ включено 78 действующих аэропортов (примерное значение). Для каждого аэропорта рассчитаны стандартное отклонение времени полёта (time_std) и среднее значение расстояния полёта (distance), что и составило исходный набор данных для анализа. Первоначальный визуальный анализ выполнен с помощью диаграммы рассеяния, отображающей зависимость между расстоянием полёта (по оси X, в милях) и стандартным отклонением времени полёта (по оси Y, в минутах) (рисунок 4.1). Распределение точек демонстрирует нелинейную тенденцию «сначала рост — затем падение»: для коротких маршрутов (< 1000 миль) стандартное отклонение варьируется в пределах 5–15 минут, для средних (1000–2500 миль) — возрастает до 15–25 минут, а для сверхдальних маршрутов (> 2500 миль) вновь снижается до 10–20 минут.

Для построения и верификации модели сначала был применён линейный регрессионный анализ, в котором в качестве независимой переменной выступало расстояние полёта, а зависимой — стандартное отклонение времени полёта. Результаты показали, что при увеличении расстояния на 1 милю волатильность времени полёта в среднем возрастает на 0.005 минуты (примерное значение коэффициента), при этом коэффициент детерминации R^2 составил 0.733.
Для более точного учёта наблюдаемой нелинейности в модель был введён квадрат расстояния (distance_sq), и выполнена полиномиальная регрессия второго порядка. Модифицированная модель продемонстрировала улучшение качества аппроксимации: R^2 возрос до 0.824, что свидетельствует о значительном усилении объясняющей способности за счёт нелинейного компонента.
Сравнение моделей проведено путём наложения двух регрессионных кривых на тот же график (рисунок 4.1): линейная модель (красная прямая) недооценивает волатильность в средней зоне расстояний, тогда как квадратичная модель (зелёная кривая) более адекватно повторяет эмпирический тренд, особенно в областях экстремальных значений (< 500 миль и > 3000 миль). Отрицательный знак коэффициента при квадратичном члене (примерное значение -2.3×10^{-6}) подтверждает наличие «перевёрнутой U-образной» зависимости, то есть существование критического расстояния (около 2000 миль), при котором волатильность времени полёта достигает максимума.

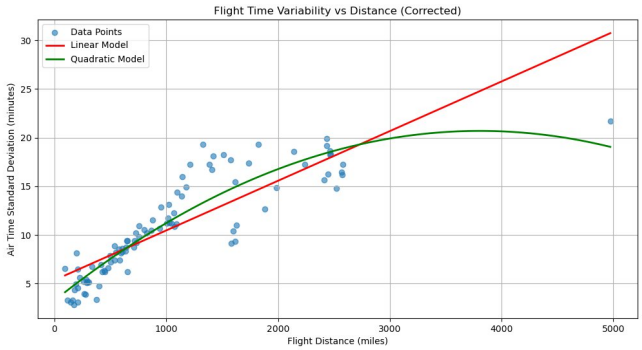

рисунок 4.1 Диаграмма зависимости времени полёта от расстояния.

3. Выводы

Коэффициент детерминации полиномиальной модели второго порядка ($R^2 = 0.824$) существенно превосходит соответствующий показатель линейной модели ($R^2 = 0.733$), что позволяет рекомендовать применение нелинейного подхода для прогнозирования. Между изменчивостью времени полёта и расстоянием полёта установлена выраженная нелинейная «перевёрнутая U-образная» зависимость: максимальная стандартная ошибка (около 20–25 минут) наблюдается для средних дистанций примерно 1500–2500 миль, тогда как для коротких и сверхдальних рейсов волатильность остаётся относительно невысокой. Данный феномен может объясняться тем, что средние дистанции более подвержены влиянию метеоусловий и частым сменам воздушных коридоров, тогда как сверхдальние рейсы, как правило, выполняются на крупных самолётах на стабильных эшелонах, что повышает устойчивость полёта к внешним воздействиям.