Todo list

Добавить ссылки . . . 3
В Задание и Заключение включить пункт “сделать/сделан сравнительный анализ прогр. пакетов статистической обработки данных” . . . 5
Вставить ссылку . . . 6
Ссылка на литературу . . . 7
Ссылка на литературу . . . 8
Ссылка на литературу . . . 8
Ссылки на литературу . . . 9
+cite: Shapiro S.S., Francia R.S. An appriximate analysis of variance test fo normality // J. Amer. Statist. Assoc., 337, 1972. – P.215-216. . . . 10
Shapiro S. S., Wilk M. B. An analysis of variance test for normality. — Biometrika, 1965, 52, №3 — p. 591-611. . . . 10
Shapiro S. S., Wilk M. B. An analysis of variance test for normality. — Biometrika, 1965, 52, №3 — p. 591-611. . . . 10
+cite: Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. — 238 с . . . 10
Ссылку на источник . . . 10
Ссылка . . . 11
Ссылки на литературу . . . 11
Ссылка . . . 14
Ссылка на листинг . . . 14
Проверить классические тестовые статистики . . . 16
Сделать вывод о полученных результатах . . . 16
Sturges . . . 16
не очень хорошее объяснение . . . 16
Ссылка . . . 20
+cite: Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58. . . . 22
Может стоит сделать нумерацию для критерия, чтобы делать на него ссылку? . . . 22
Думаю, здесь нужно сделать выводы о проделанной в данной части работе — подвести итог . . . 23

Введение

Данная работа посвящена статистическому анализу, обработке и исследованию реальных временных рядов. В настоящее время, выбор такой направленности соответствует необходимости в проведении анализа различных длительных наблюдений с математической и, в частности, статистической точек зрения. Поскольку наличие даже большого количества информации, полученной в процессе каких-либо наблюдений, не всегда способно раскрыть те или иные причины и следствия, представленные в конкретном случае. Наличие информации, без последующего всестороннего анализа, также не может раскрыть все скрытые проблемы и свойства. В свою очередь, математический аппарат и его конкретные прикладные части могут позволить не только проанализировать сложившуюся ситуацию, но и постараться дать некоторый прогноз по состоянию в будущем.

В качестве исследуемого материала в данной работе используются база данных с реальными наблюдениями, зафиксированными на озёрах, входящих в Нарочанский национальный парк, за период с 1955 по 2012 годы, полученная от учебно-научного центра «Нарочанская биологическая станция им. Г.Г.Винберга». В представленной базе данных присутсвовали наблюдения за следующими озёрами:

- Баторино
- Нарочь
- Мястро

Из представленных озёр для исследования было выбрано озеро Баторино. Данное озеро является уникальным природным объектом, изучение которого позволяет решать проблемы экологии не только в региональном, но и глобальном масштабе. Оно располагается у самой границы города Мядель и входит в состав Нарочанской озёрной группы. Кроме представленных ранее озер, в эту группу также входят озеро Белое и озеро Бледное.

В данной работе исследуемым показателем для озера Баторино был избран показатель температуры воды. Температура воды принадлежит к числу наиболее важных и фундаментальных характеристик любого водоёма. Её изменение во времени является одним из важнейших факторов, отражающих изменения в окружающей среде. Также нужно отметить, что исследование температуры воды является актуальным, вследствие зависимостей свойств воды от температуры. Так как данная характеристика оказывает сильное влияние на плотности воды, растворимость в ней газов, минеральных и органических веществ, в том числе одни из важнейших характеристик для обитания в ней живых организмов: растворимость и насыщенность воды кислородом. В частности от температуры воды в значительной мере зависит жизнедеятельность рыб: их распределение в водоёме, питание, размножение. К тому же температура тела рыб, как правило, не превышает температуры окружающей их воды. В то же время, любой водоём как экосистема является средой обитания различных, отличных от рыб, организмов. И поэтому отслеживание всех изменений и влияние этих изменений на их жизнь является крайне важным не только в экологическом смысле, но и в биологическом. Как следствие вышесказанного, изменения

температуры с течением времени следует считать одним из важнейших индикаторов изменений, происходящих в экосистеме озера. А исследование данного показателя, в свою очередь, является важнейшим в исследовании различных проблем, связанных с водоёмами. В подтверждение актуальности исследования представленной темы можно привести работы.

Добавить ссылки

В настоящее время, в условиях глобального потепления и крайне нестабильной климатической ситуации, наблюдения за состоянием озёрных экосистем представляют особую ценность как с научной, так и с практической стороны, поскольку только на основе таких наблюдений возможно выделить последствия антропогенного воздействия на фоне изменения природных факторов. А также получить некоторые заключения по экологической обстановке в определенной области.

Изначально, для решения этой задачи был выбран пакет **STATISTICA**. Данный программный пакет — это универсальная интегрированная система, предназначенная для статистического анализа и визуализации данных, управления базами данных и разработки пользовательских приложений, содержащая широкий набор процедур анализа для применения в научных исследованиях, технике, бизнесе, а также специальные методы добычи данных. В системе **STATISTICA** реализовано множество мощных языков программирования, которые снабжены специальными средствами поддержки. С их помощью легко создаются законченные пользовательские решения и встраиваются в различные другие приложения или вычислительные среды. В пакете представлены несколько сотен типов графиков 2D, 3D и 4D, матрицы и пиктограммы; предоставляется возможность разработки собственного дизайна графика. Средства управления графиками позволяют работать одновременно с несколькими графиками, изменять размеры сложных объектов, добавлять художественную перспективу и ряд специальных эффектов, разбивку страниц и быструю перерисовку. Например, 3D-графики можно вращать, накладывать друг на друга, сжимать или увеличивать.

Но пакет **STATISTICA**, являясь по сути узкоспециализированным пакетом статистического анализа, наравне с преимуществами имеет и свои недостатки. Главным из которых, на мой взгляд, являеется то, что **STATISTICA** — коммерческий продукт. И как следствие имеет закрытую платформу, закрытый исходный код и при этом является достаточно дорогим. Также, к недостаткам я бы отнёс довольно громоздкий интерфейс, и взаимодействие с программой только на уровне кнопок и таблиц. Недостаток в гибкости и открытости в некоторых ситуациях ограничивает возможности анализа. С другой стороны, такой подход позволяет быстро получать различные результаты без длительного изучения самого продукта. Но, объективно оценив все преимущества и недостатки, следует отметить, что пакет **STATISTICA** не является оптимальным, во всестороннем смысле, выбором.

Существует множество различных программных пакетов, с помощью которых можно осуществлять статистический анализ. Приведём основные из них:

- Excel
- STATISTICA
- SAS
- JMP
- SPSS
- SPlus
- R

- Mathematica
- Matlab

Каждый из представленных программных пакетов имеет свои преимущества и недостатки. Следует сразу отметить, что такие пакеты как **Mathematica** и **Matlab**, которые являются по сути общематематическими, хоть и имеют статистические модули, но все же не являются специализорованным решением. Прежде всего, я хотел бы отметить преимущества пакета **STATISTICA** перед программой **Excel**. Во-первых, **STATISTICA** является специализированным пакетом по обработке статистических данных, вследствие чего этот процесс намного быстрее аналогичного в **Excel**. Во-вторых, в выбранном пакете присутствуют все необходимые для анализа данных инструменты и формулы, многое из которого нет в **Excel**. Если вкратце подвести итог, то **Excel** следует рассматривать скорее как иструмент работы с таблицами, а **STATISTICA** является платным и довольно громоздким пакетом, сходным по представлению данных (в таблицах) с **Excel**. Каждый из оставшихся в списке пакетов заслуживает отдельного ознакомления. Но если рассмотреть весь этот список в совокупности, то у каждого пакета, за исключением одного — пакета **R** — можно выделить одну общую черту: все они являются проприетарными и являются платными. И каждая новая версия продукта, при необходимости в оной, нуждается в приобретении. На самом деле, данную черту можно трактовать и как недостаток, и как преимущество. Следует отметить, что крупные коммерческие компании, заинтересованные в статистическом анализе, предпочитают такие пакеты как **SAS**, **JMP**, **SPSS**, **SPlus**. Поскольку покупая тот или иной продукт, компания получает гарантию в точности полученных с помощью пакета результатов, а также наличие оплаченой технической поддержке. Тогда как *open-source* продукты предоставляются в комплекте "как есть". Но в совокупности с открытостью исходного кода и наличием огромных сообществ, отсутствие гарантии является слабым недостатком, так как у пользователя всегда есть возможность получить последнюю версию продукта, техническую помощь, или вовсе предложить свою помощь в разработке. Поэтому отдельные исследователи всё чаще и чаще выбирают пакет **R**.

Именно согласно этим соображениям, в качестве инструмента исследования в данной работе был выбран пакет **R**. Рассмотрим его подробнее. **R** является функциональным интерпретируемым языком программирования с динамической типизацией данных для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом в рамках проекта *GNU*. Язык создавался как аналогичный языку **S**, разработанному в *Bell Labs* и является его альтернативной реализацией, хотя между языками есть существенные отличия, но в большинстве своём код на языке **S** работает в среде **R**.

R широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ. Доступен для широкого числа операционных систем: *Unix/Linux*, *Microsoft Windows*, *Mac OS X*.

R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения. В базовую поставку R включен основной набор пакетов, а всего по состоянию на момент написания данной работы доступно более 5000 пакетов, которые распространяются через *CRAN*(акроним *Comprehensive R Archive Network*). С помощью них можно значительно расширить возможности для статистического анализа. Причем многие из них написаны самими пользователями. Как следствие этого, при отсутствии какого-либо функционала, всегда можно реализовать его посредством создания своего пакета или просто функции. Следует также отметить возможность использования напрямую функций, написанных на языках программирования *C*, *C++*, *Java*.

Ещё одной особенностью **R** являются графические возможности, заключающиеся в возможности создания для публикаций качественной графики, которая может включать в себя математические символы. Динамические и интерактивные графики также доступны в качестве дополнительных пакетов.

R имеет свой собственный LaTeX-подобный формат документации, который может быть использован для всестороннего документирования, как в режиме онлайн, в различных вариантах, так и в бумажном носителе.

Следует также отметить, что уже упомянутые ранее программные пакеты статистического анализа, на данный момент почти все имеют поддержку кода, написанного на **R**. Это о многом говорит, и уже не стоит удивлятся, если известный коммерческий продукт в очередном выпуске добавит аналогичную поддержку.

Два-три раза в год выходит свободно-распространяемый информационный журнал *R Journal*. Он содержит информацию по статистической обработке данных и разработке, что может быть интересно как пользователям, так и разработчикам **R**.

Одна из самых популярных конференций, посвящённых языку — *useR! (The R User Conference)*, проходит ежегодно, начиная с 2004 года, собирает специалистов в различных областях.

Начиная с 2009 года каждой весной в Чикаго проводится конференция, посвящённая применению **R** в финансах (*R/Finance: Applied Finance with R*). В 2013 году прошла первая конференция, посвящённая применению **R** в страховании (*R in Insurance*).

В Задание и Заключение включить пункт “сделать/сделан сравнительный анализ прогр. пакетов статистической обработки данных”

1 Определения и вспомогательные результаты

Вставить ссылку

Приведём основные теоретические понятия из [], которыми будем пользоваться в дальнейшем.

Пусть имеется некоторый одномерный признак X. Из него извлечена выборка объёма n: $x_1, x_2, \ldots, x_n$.

Введём некоторые основные *описательные статистики*.

Характеристики положения

Среднее арифметическое значение является показателем центрального положения, вычисляется по формуле:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Квантилем Q_p порядка p, $0 < p < 1$, называется такое значение признака в упорядоченной совокупности, которое делит её в отношении $p : (1-p)$. К числу наиболее часто применяемых квантилей относятся:

1. *медиана* $(p = \frac{1}{2})$;
2. *квартиль* $(p = \frac{1}{4})$.

Пусть выборка упорядочена по возрастанию: $x_1^*, \ldots, x_n^*$. Тогда *медиана* вычисляется по формуле:

$$Me = \begin{cases} x_{l+1}^* & , \quad n = 2l+1, \\ \frac{x_l^* + x_{l+1}^*}{2} & , \quad n = 2l. \end{cases}$$

Характеристики рассеяния

Наиболее распространёнными мерами рассеяния являются *размах*, *дисперсия* и *среднеквадратическое отклонение*.

Размах определяется по формуле:

$$R = x_{max} - x_{min}.$$

Квартильный размах — интервал, содержащий медиану, в который попадает 50% выборки, вычисляется по формуле:

$$R_Q = q_3 - q_1,$$

где q_3, q_1 — соответственно, верхний и нижний квартили.

Выборочной дисперсии вычисляется по формуле:

$$S_x^2 = \frac{\sum (x_i - \overline{x})^2}{n}.$$

Также, для вычисления *выборочной дисперсии* используется несмещённая оценка, которая вычисляется по формуле:

$$s_x^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}.$$

Оценка *стандартного отклонения* вычисляется по формуле:

$$s_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}.$$

Значение стандартного отклонения можно представить, как среднее расстояние, на котором находятся элементы от среднего элемента выборки, и оно показывает, насколько хорошо среднее значение описывает всю выборку.

В качестве меры относительного разброса данных используют *коэффициент вариации*:

$$V = \frac{s_x}{\overline{x}} \cdot 100\%.$$

Данная мера показывает, какую долю среднего значения этой величины составляет её средний разброс. На основе значения коэффициента вариации, можно сделать вывод об однородности выборки:

- Если $V < 33\%$ — принято считать, что выборка однородна;
- Если $V > 33\%$ — не однородна.

Стандартная ошибка среднего значения вычисляется как:

$$s_{\overline{x}} = \frac{s_x}{\sqrt{n}}.$$

Данная величина оценивает выборочную изменчивость среднего значения, приближённо показывая, насколько выборочное среднее отличается отсреднего генеральной совокупности.

Характеристики формы распределения

Характеристики формы распределения применяются для выражения особенностей формы распределения.

Выборочный коэффициент асимметрии определяется следующим образом:

$$A_S = \frac{n \sum (x_i - \overline{x})^3}{(n-1)(n-2)s_x^3}. \tag{1.1}$$

Данный коэффициент характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричном распределении коэффициент асимметрии равен нулю.

- Если $|A_S| > 1$, то распределение является в значительной степени асимметричным.
- Если $\frac{1}{2} < |A_S| \leq 1$, то распределение незначительно асимметрично.
- Если $|A_S| \leq \frac{1}{2}$, то распределение является близким к симметричному.

Ссылка на литературу

Стандартная ошибка выборочного коэффициента асимметрии (1.1) вычисляется по формуле:

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}},$$

где n — объём выборки.

Для того, чтобы сделать в дальнейшем какие-либо выводы по значению коэффициента асимметрии введём *тестовую статистику*:

$$Z_{A_S} = \frac{A_S}{SES}.$$

- Если $|Z_{A_S}| \geq 2$, то асимметрия существенная и распределение признака в генеральной совокупности несимметрично;
- Если $|Z_{A_S}| < 2$, то асимметрия несущественна.

Ссылка на литературу

Данная тестовая статистика показывает: насколько существенным является коэффициент асимметрии данной выборки по отношению к генеральной совокупности.

Выборочный коэффициент эксцесса вычисляется по формуле:

$$K = \frac{n(n+1)\sum(x_i - \overline{x})^4 - 3(\sum(x_i - \overline{x})^2)^2(n-1)}{(n-1)(n-2)(n-3)s_x^4}.$$

В случае нормального распределения *коэффициент эксцесса* равен нулю. Положительный *коэффициент эксцесса* характеризует крутость(островершинность) кривой распределения относительно нормального распределения. Отрицательный, в свою очередь, пологость.

Стандартная ошибка коэффициента эксцесса может быть вычислена по формуле:

$$SEK = 2(SES)\sqrt{\frac{(n^2-1)}{(n-3)(n+5)}},$$

где n — объём выборки.

По аналогии с коэффициентом асимметрии, введём также *тестовую статистику* Z_K для коэффициента эксцесса.

$$Z_K = \frac{K}{SEK}.$$

- Если $|Z_K| > 2$, то коэффициент эксцесса является значимым;
- Если $|Z_K| \leq 2$, то коэффициент эксцесса не является значимым и нельзя сделать никаких заключений о коэффициенте эксцесса генеральной совокупности.

Ссылка на литературу

Выборочный коэфициент корреляции

Для оценки тесноты линейной связи между признаками используется парный линейный *коэффициент корреляции Пирсона*, который определяется формулой:

$$r_{xt} = \frac{S_{xt}}{S_x S_t}, \tag{1.2}$$

где $S_{xt} = \frac{1}{n}\sum(x_i - \overline{x})(t_i - \overline{t})$, $S_x = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$, $S_t = \sqrt{\frac{1}{n}\sum(t_i - \overline{t})^2}$.

Коэффициент корреляции характеризует силу связи и направление. Если $-1 \leq r_{xt} < 0$, то наблюдается отрицательная корреляция, если $0 < r_{xt} \leq 1$, то наблюдается положительная корреляция.

Для описания силы связи *коэффициента корреляции* используются градации, обозначенные в таблице 1, где в качестве значения коэффициента корреляции указано абсолютное значение $|r_{xt}|$.

Таблица 1 — Абсолютные значения коэффициента корреляции

Абсолютное значение	**Интерпретация**
0—0.2	Очень слабая зависимость
0.2—0.5	Слабая зависимость
0.5—0.7	Средняя зависимость
0.7—0.9	Высокая зависимость
0.9—1	Очень высокая зависимость

Критерий значимости выборочного коэффициента корреляции

Пусть двумерная генеральная совокупность (X, t) распределена нормально. Из этой совокупность извлечена выборка объёма n и по ней найден выборочный коэффициент корреляции $r_{xt} \neq 0$.

Требуется проверить нулевую гипотезу $H_0 : r = 0$, при конкурирующей гипотезе $H_1 : r \neq 0$.

Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0, надо вычислить наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{r_{xt}\sqrt{n-2}}{\sqrt{1-r_{xt}^2}},$$

где r_{xt} — выборочный коэффициент корреляции, и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n - 2$, где n — число пар значений выборки, найти критическую точку $t_{\text{кр}}(\alpha, k)$ для двусторонней критической области.

- Если $|T_{\text{набл}}| < t_{\text{кр}(\alpha,k)}$ — нет оснований отвергнуть нулевую гипотезу.
- Если $|T_{\text{набл}}| > t_{\text{кр}(\alpha,k)}$ — нулевую гипотезу отвергают.

Величина $T_{\text{набл}}$ при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы, если нулевая гипотеза отвергается, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля, а исследуемые переменные коррелированы.

Приведём некоторые критерии, которые понадобятся в нашем исследовании. Для этого воспользуемся литературой.

Ссылки на литературу

Критерий Шапиро–Уилка

Критерий Шапиро–Уилка используется для проверки гипотезы H_0: «генеральная совокупность распределена нормально» и является одним наиболее эффективных критериев проверки нормальности.

+cite: Shapiro S.S., Francia R.S. An appriximate analysis of variance test fo normality // J. Amer. Statist. Assoc., 337, 1972. – P.215-216.

Пусть имеется вариационный ряд $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, построенный по извлечённой из генеральной совокупности выборки $x_1, x_2, \ldots, x_n$.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0 необходимо вычислить статистику:

$$W = \frac{\left(\sum_{i=1}^{k} a_{(n-i+1)}(x_{(n-i+1)} - x_{(i)})\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

где коэффициенты $a_{(n-i+1)}$ — известные константы, представленные в таблицах из [], индекс k вычисляется следующим образом:

Shapiro S. S., Wilk M. B. An analysis of variance test for normality. — Biometrika, 1965, 52, №3 — p. 591-611.

$$k = \begin{cases} (n-1)/2 & , \quad n = 2l+1, \\ n/2 & , \quad n = 2l. \end{cases}$$

И по заданному уровню значимости α из таблицы критических значений статистики $W(\alpha)$ [] найти критическую точку $W_{\text{кр}}(\alpha)$.

Shapiro S. S., Wilk M. B. An analysis of variance test for normality. — Biometrika, 1965, 52, №3 — p. 591-611.

- Если $W < W_{\text{кр}}(\alpha)$ — нулевую гипотезу отклоняют.
- Если $W > W_{\text{кр}}(\alpha)$ — нет оснований отклонять нулевую гипотезу.

+cite: Кобзарь А. И. Прикладная математическая статистика. — М.: Физматлит, 2006. — 238 с

Критерий χ^2

Пусть нулевая гипотеза H_0 состоит в том, что генеральная совокупность распределена нормально.

Для того, чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, необходимо вычислить наблюдаемое значение критерия

$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n'_i)^2}{n'_i},$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = s - 3$ найти критическую точку $\chi^2_{\text{кр}}(\alpha, k)$, где n_i — эмпирические частоты, а n'_i — теоретические, s — число групп (частичных интервалов) выборки[].

Ссылку на источник

- Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha, k)$ — нет оснований отвергать нулевую гипотезу.
- Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(\alpha, k)$ — нулевую гипотезу отвергают.

Критерий Колмогорова—Смирнова

Пусть нулевая гипотеза H_0 состоит в том, что генеральная совокупность распределена нормально.

Критерий заключается в том, что можно сравнивать эмпирическую функцию распределения $F^*(x)$ с гипотетической $F(x)$ и, если мера расхождения между ними мала, то считать справедливой гипотезу H_0.

Для того, чтобы при заданном уровне значимости проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально, необходимо вычислить статистику Колмогорова-Смирнова

$$D = \sqrt{n} \max_x |F^*(x) - F(x)|$$

и по таблице критических значений статистики Колмогорова–Смирнова, по заданному уровню значимости α найти критическую точку $D_{\text{кр}}(\alpha)$.

- Если $D < D_{\text{кр}}(\alpha)$ — нет оснований отвергать нулевую гипотезу.
- Если $D > D_{\text{кр}}(\alpha)$ — нулевую гипотезу отвергают.

Критерий Граббса

Ссылка

Критерий Граббса основан на предположении о нормальности исходных данных. То есть, перед применением данного критерия необходимо убедиться, что данные могут быть в разумных пределах аппроксимированы нормальным распределением.

Критерий Граббса обнаруживает один выброс за одну процедуру проверки. Найденный выброс исключается из выборки и процедура проверки критерия проверяется пока не будут исключены все выбросы. Не рекомендуется использовать данный критерий для выборок объёмом ниже 7.

Данный критерий заключается в проверке нулевой гипотезы H_0 : в выборке нет выбросов, при конкурирующей гипотезе H_1 : в выборке есть по крайней мере 1 выброс.

Статистика критерия Граббса определяется следующим образом:

$$G = \frac{\max |y_i - \overline{y}|}{s_y},$$

где $\overline{y}$ — выборочное среднее, s_y - выборочное стандартное отклонение.

Гипотеза H_0 отклоняется (значение y_i является выбросом) при заданном уровне значимости α, если

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\left(\frac{\alpha}{2n}, n-2\right)}}{n - 2 + t^2_{\left(\frac{\alpha}{2n}, n-2\right)}}},$$

где n — объём выборки, $t_{\left(\frac{\alpha}{2n}, n-2\right)}$ является критическим значением t-распределения с $n-2$ степенями свободы и уровнем значимости $\frac{\alpha}{2n}$.

Характеристики линейной регрессии

Ссылки на литературу

Для введения следующих понятий воспользуемся.

Предположим, что уравнение регрессии имеет вид: $x^*(t) = at + b$. Тогда *объяснённая уравнением регрессии дисперсия*, характеризующая изменчивость линии регрессии относительно среднего значения, вычисляется по формуле:

$$\overline{\sigma^2} = \frac{1}{n} \sum_{j=1}^{n} (x^*(t_j) - \overline{x})^2, \tag{1.3}$$

где x — выборка, $\overline{x}$ — выборочное среднее.

Остаточная дисперсия $\overline{D}$, характеризующая отклонение уравнения регрессии от результатов наблюдений x вычисляется по формуле:

$$\overline{D} = \frac{1}{n}\sum_{j=1}^{n}(x_j - x^*(t_j))^2. \tag{1.4}$$

Тогда, *общая дисперсия* введённой ранее выборки x будет равна сумме 1.3 и 1.4:

$$S_x^2 = \overline{D} + \overline{\sigma^2}.$$

Дисперсия отклонения вычисляется по формуле:

$$\sigma_\varepsilon^2 = s_x^2(1 - r_{xt}^2),$$

где r_{xt} — выборочный коэффициент корреляции.

Стандартные случайные погрешности параметров a и b:

$$\sigma_a = \frac{\sigma_\varepsilon}{S_x\sqrt{n-2}}, \quad \sigma_b = \frac{\sigma_\varepsilon}{\sqrt{n-2}}\sqrt{1 + \frac{\overline{x}^2}{S_x^2}},$$

где S_x, S_x^2 — стандартное отклонение и дисперсия соответственно.

Коэффициент детерминации показывает долю дисперсии исходного ряда, которая описывается моделью регрессии, вычисляется по формуле:

$$\eta_{x(t)}^2 = \frac{\overline{\sigma^2}}{s_x^2}.$$

Применяя неравенство $\eta_{y(x)}^2 - r_{xy}^2 \leq 0.1$, можно сделать вывод об отклонении от линейности.

Критерий значимости коэффициентов регрессии

Пусть нулевая гипотеза заключается в равенстве нулю коэффициентов линейной регрессии $H_0 : a = 0, b = 0$.

Для проверки нулевой гипотезы $H_0 : a = 0, b = 0$ необходимо рассчитать:

$$T_a = \frac{a}{\sigma_a}, \quad T_b = \frac{b}{\sigma_b},$$

по статистической таблице определить $t_{\text{кр}}(k, \alpha)$ — критическую точку t-распределения Стьюдента при заданном уровне значимости α и числе степеней свободы $k = n - 2$.

- Если $|T_a| > t_{\text{кр}}(k, \alpha)$, то нулевая гипотеза отвергается и отклонение a от нуля носит неслучайный характер, и, следовательно, величина a значима;
- Если $|T_b| > t_{\text{кр}}(k, \alpha)$, то нулевая гипотеза отвергается, отклонение b от нуля носит неслучайный характер, и, следовательно, величина b значима.

Критерий Фишера

Данный критерий используется для оценки адекватности регрессионной модели. Пусть выдвинута нулевая гипотеза о равенстве дисперсий

$$H_0 : \overline{\sigma^2} = \frac{\overline{D}}{n-2}.$$

Для проверки данной гипотезы используется F-критерий Фишера. Необходимо вычислить дисперсионное отношение

$$F_{\text{крит}} = \frac{(n-2)\overline{\sigma^2}}{\overline{D}},$$

которое сравнивается с $F_{\text{табл}}(v_1, v_2, \alpha)$ при заданном уровне значимости α, и степенях свободы $v_1 = 1, v_2 = n - 2$.

Если $F_{\text{крит}} > F_{\text{табл}}$, то нулевая гипотеза о равенстве дисперсий отвергается, что означает в рассматриваемом случае адекватность регрессионной модели.

2 Обработка реального временного ряда с помощью R

2.1 Вычисление основных описательных статистик

В качестве исходных данных примем выборку из полученной от учебно-научного центра базы данных, путём отбора наблюдений в июле месяце за период с 1975 по 2012 год. Выборка представлена в приложении А в таблице А.1. Графически исходные данные представлены на рисунке 1.

Ссылка

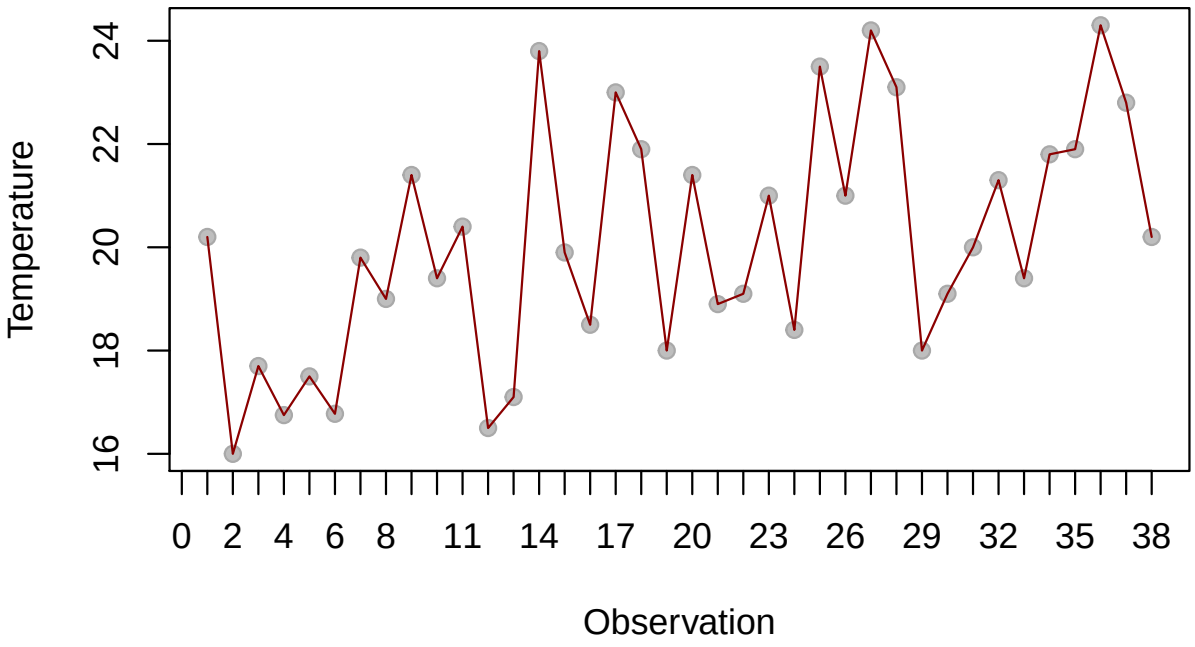

Рисунок 1 — График исходных данных.

Начнём исследование временного ряда с вычисления описательных статистик. **R** предоставляет в пакете *base* такие функции как: *var* — дисперсия, *mean* — среднее, *sd* — стандартное отклонение, *median* — медиана, *quantile* — квантили, *range* — размах, *min, max*. Также, в различных пакетах можно найти другие интерисующие функции, как статистические, так и математические. Но в целях удобства, компактности и контроля за функциональностью мной был написан модуль *dstats*. Данный модуль позволяет вычислять все необходимые в данной работе описательные статистики.

Ссылка на листинг

Для получения всех описательных статистик воспользуемся представленным в [] модулем. Полученные результаты представлены в таблице 1.

	Значение
Среднее	20.08
Медиана	19.95
Нижний квартиль	18.42
Верхний квартиль	21.70
Минимум	16.00
Максимум	24.30
Размах	8.30
Квартильный размах	3.28
Дисперсия	5.24
Стандартное отклонение	2.29
Коэффициент вариации	26.10
Стандартная ошибка	0.37
Асимметрия	0.14
Ошибка асимметрии	0.38
Эксцесс	-0.85
Ошибка эксцесса	0.75

Таблица 1 — Описательные статистики для наблюдаемых температур.

Рассмотрим подробнее некоторые полученные статистики.

Как видно из таблицы, *средняя* температура в июле месяце за период с 1975 по 2012 составляет приблизительно 20°C. При этом *размах* температур равен 8.3°C. *Дисперсия* в данном случае равна 5.24.

Стандартное отклонение оказалось равным приблизительно 2.3. Полученное значение не велико, а значит можно сказать, что среднее значение хорошо описывает выборку. И что в среднем, температура воды озера Баторино отличается от полученной ранее *средней* температуры на 2.3°C.

Коэффициент вариации в нашем случае равен 26.1%. Из этого следует, что выборку можно счиать однородной, так как полученное значение является меньшим 33%.

Стандартная ошибка среднего значения равна 0.37.

Коэффициент асимметрии — мера симметричности распределения. Полученное значение: 0.14. Данное значение говорит о незначительном коэффициенте асимметрии выборки. То есть о том, что выборочное распределение можно считать близким к симметричному.

Стандартная ошибка асимметрии равна 0.38.

Коэффициент эксцесса в рассматриваемом случае равен -0.85. Так как коэффициент эксцесса нормального распределения равен 0, то в данном случае можно говорить о пологости пика распределения выборки по отношению к нормальному распределению.

Стандартная ошибка коэффициента эксцесса равна 0.75.

По данному ранее определению тестовых статистик для коэффициента асимметрии и эксцесса, проверим значимость полученных значений для генеральной совокупности. Для этого в модуле *dstats* мной реализованы функции *dstats.test.skew* и *dstats.test.kurtosis*:

Полученная тестовая статистика для коэффициента асимметрии:

$$Z_{A_S} = \frac{A_S}{SES} = 0.3630143.$$

Данное значение попадает под случай $|Z_{A_S}| \leq 2$, а значит, выборочный коэффициент асимметрии не является значимым. Из чего, в свою очередь, следует, что по нему нельзя судить о коэффициенте асимметрии генеральной совокупности.

Полученная тестовая статистика для коэффициента эксцесса:

$$Z_K = \frac{K}{SEK} = -1.135476.$$

Данное значение попадает под случай $|Z_K| \leq 2$, а значит, в данном случае выборочный коэффициент эксцесса не является значимым и нельзя ничего сказать о коэффициенте эксцесса генеральной совокупности.

Проверить классические тестовые статистики

Сделать вывод о полученных результатах

2.2 Исследование статистических данных

В **R** можно найти различные пакеты, позволяющие строить разнообразные гистограммы, диаграммы рассеяния, вероятностные графики, линейные графики, диаграммы диапазонов, размахов, круговые диаграммы, столбчатые диаграммы, последовательные графики и т.д., позволяющие увидеть специфику данных.

В пакет *base* для визуализации входят такие функции как:

- *plot*: общая функция для построения графиков $y(x)$;
- *barplot*: столбцовые диаграммы;
- *boxplot*: график “ящик-с-усами”;
- *hist*: гистограммы;
- *pie*: круговые диаграммы;
- *dotchart*: точечные графики;
- *image, heatmap, contour, persp*: функции для генерации трёхмерных графиков;
- *qqnorm, qqline, qqplot*: графики квантилей;
- *pairs, coplot*: отображают на графиках несколько выборок.

С помощью функции *hist* построим гистограмму для отображения вариационного ряда исходных данных. Гистограммы позволяют увидеть, как распределены значения переменных по интервалам группировки, то есть как часто переменные принимают значения из различных интервалов. А также, что бывает более важным, повзоляет сделать предположение о разновидности распределения. Полученная гистограмма отражена на рисунке 2.

Представленная гистограмма построена с автоматически рассчитанным количеством интервалов разбиения. Воспользуемся *формулой Стерджеса* для вычисления этого количества. Из [] количество интервалов разбиения рассчитывается по формуле:

Sturges

$$k = \lceil log_2 n \rceil + 1 = \lceil log_2 38 \rceil + 1 = 7. \tag{2.1}$$

Следует отметить, что в построенной гистограмме, на рисунке 2, получилось 9 интервалов. Данный результат обосновывается особенностями реализации функции *hist*. Указанная особенность заключается в том, что эта функция вычисляет количество интервалов по формуле Стерджеса 2.1 и при построении интервалов пользуется принципом “красивого” разбиения.

не очень хорошее объяснение

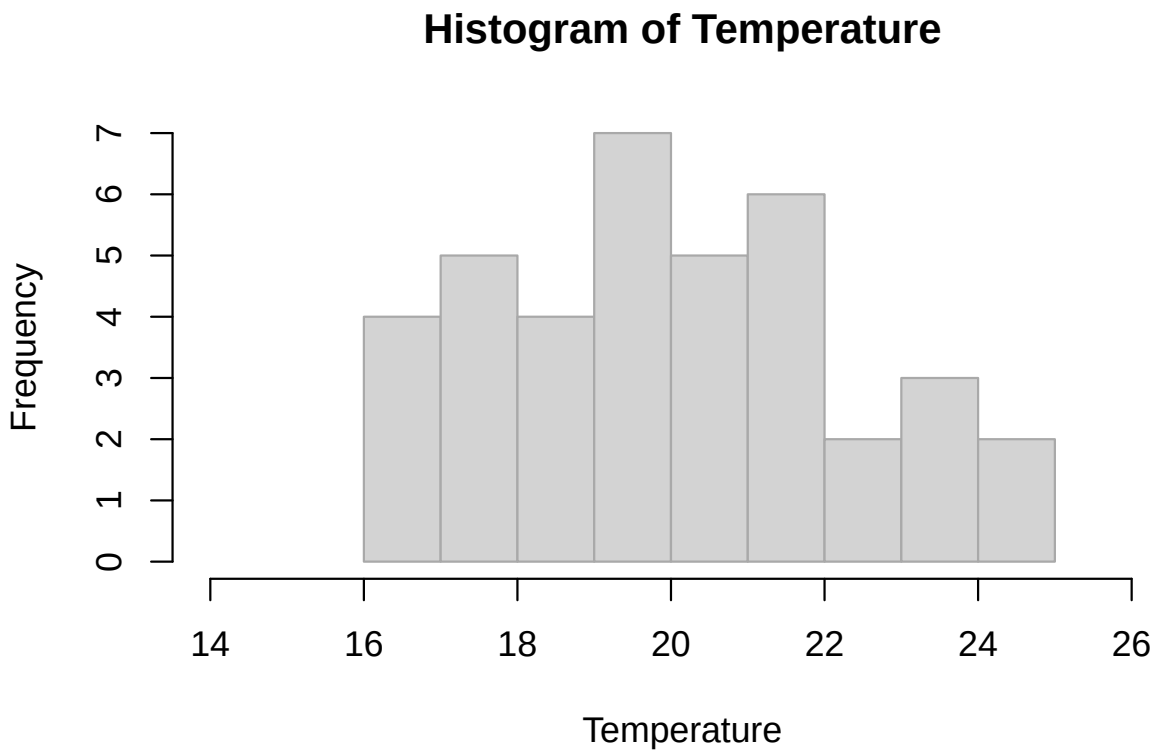

Рисунок 2 — Гистограмма наблюдаемых температур.

По полученной гистограмме можно визуально предположить близость выборочного распределения к нормальному распределению. Исследуем подробнее данное наблюдение.

Для этого построим гистограмму с кривой плотности нормального распределения. Построенная гистограмма отображена на рисунке 3. На основании этой диаграммы уже можно сказать больше. Во-первых, здесь нагляднее представлена близость выборочного распределения к нормальному. Во-вторых, по этой гистограмме можно подтвердить или опровергнуть результаты, полученные на этапе вычисления описательных статистик в параграфе 2.1.

Следует отметить согласованность полученных описательных статистик с полученной гистограммой. Во-первых, по коэффициенту асимметрии мы предположили о близости распределения к симметричному. Это подтверждается гистограммой: на ней можно заметить небольшую скошенность вправо, что также согласовывается со знаком коэффициента. Во-вторых, коэффициент эксцесса указывал на пологость пика распределения. Данное заключение подтверждается кривой плотности — она имеет чуть более растянутую колоколобразную форму.

Другим очень часто используемым графическим способом проверки характера распределения данных является построение т.н. *графиков квантилей* (*Q-Q plots*, *Quantile-Quantile plots*). На таких графиках изображаются квантили двух распределений — эмпирического (т.е. построенного по анализируемым данным) и теоретически ожидаемого стандартного нормального распределения. При нормальном распределении проверяемой переменной точки на графике квантилей должны выстраиваться в прямую линию, исходящую под улом 45 градусов из левого нижнего угла графика. Графики квантилей особенно полезны при работе с небольшими по размеру совокупностями, для которых невозможно построить гистограммы, принимающие какую-либо выраженную форму.

В **R** для построения графиков квантилей можно использовать базовую функцию

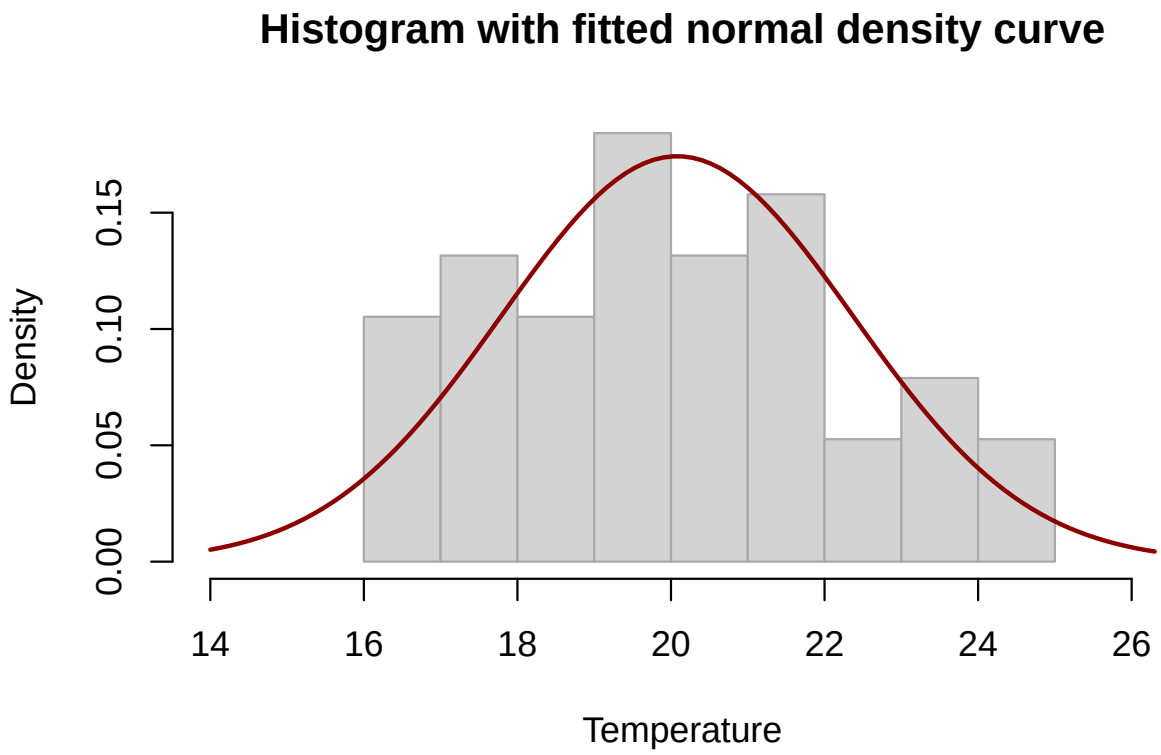

Рисунок 3 — Гистограмма наблюдаемых температур с кривой плотности нормального распределения.

qqnorm (Рисунок 4) На этом графике можно визуально обнаружить аномальное положение наблюдаемых значений по отношению к нормальному распределению. В данном случае отклонения можно наблюдать на концах рассматриваемого промежутка. Остальные значения образуют отчетливую прямую. А значит, подтверждается предположение о нормальности выборочного распределения.

Далее следует проверить полученные результаты с помощью некоторых формальных тестов. Существует целый ряд статистических тестов, специально разработанных для проверки нормальности выборочного распределения. В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулиировать следующим образом: “Анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение”. Если получаемая при помощи того или иного теста вероятность ошибки P оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.

В **R** реализованы практически все имеющиеся тесты на нормальность — либо в виде стандарных функций, либо в виде функций, входящих в состав отдельных пакетов. Примером базовой функции является *shapiro.test()*, при помощи которой можно выполнить широко используемый *тест Шапиро-Уилка*:

```
Shapiro-Wilk normality test

data:  Temperature
W = 0.9727, p-value = 0.4706
```

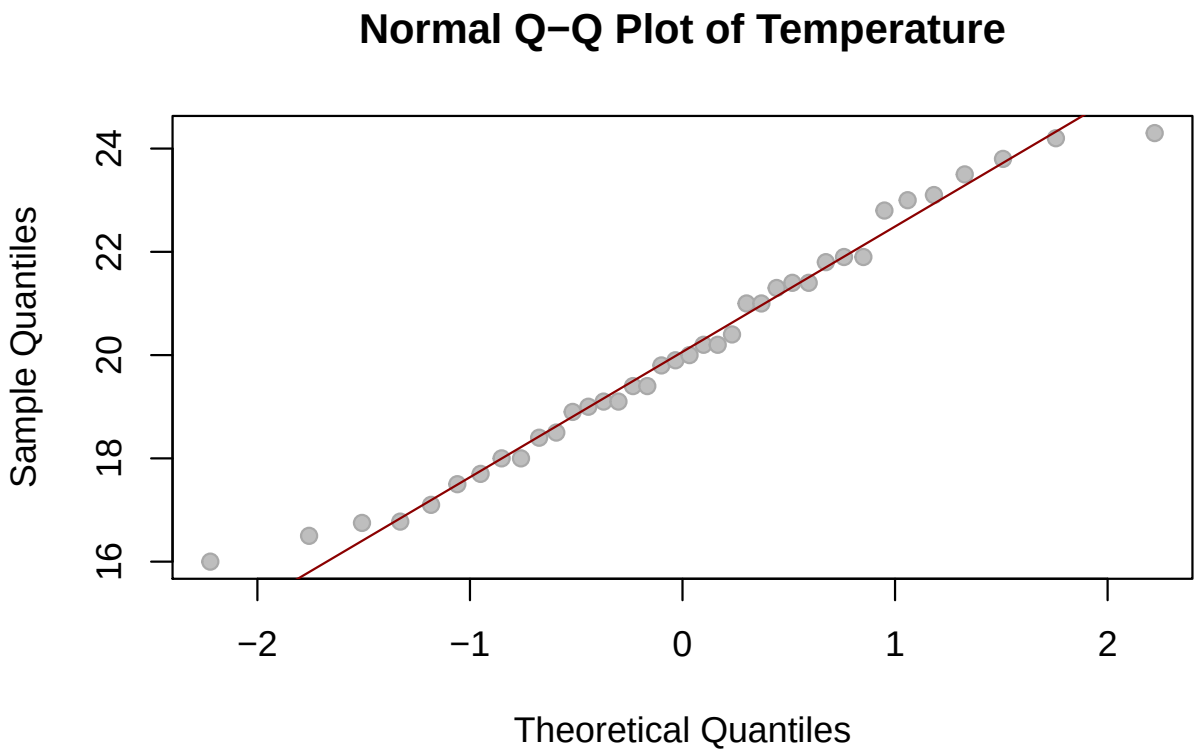

Рисунок 4 — График квантилей для наблюдаемых температур.

В полученных результатах W — статистика Шапиро-Уилка. Вероятность ошибки $p = 0.4706 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение на основе данного теста нельзя.

Попробуем опровегнуть наше предположение на основе проверки критерия χ^2 Пирсона. Для этого воспользуемся пакетом *nortest* и функцией *pearson.test*:

```
Pearson chi-square normality test

data:  Temperature
P = 1.7895, p-value = 0.938
```

В полученных результатах P — статистика χ^2 Пирсона. Вероятность ошибки $p = 0.938 > 0.05$, а значит нулевая гипотеза не отвергается. Следовательно опровергнуть предположение о нормальности на основе данного теста также нельзя. Проверим критерий: примем уровень значимости $\alpha = 0.05$, тогда из таблицы распределения χ^2 найдём критическое значение критерия $\chi^2_{\text{кр}}(\alpha, k) = 7.8$. Отсюда следует, что

$$\chi^2_{\text{набл}} < \chi^2_{\text{кр}},$$

где $\chi^2_{\text{набл}} = P = 1.7895$.

А значит, нулевую гипотезу при уровне значимости $\alpha = 0.05$ не отвергаем и подтверждаем сделанный вывод на основании вычисленной вероятности ошибки.

Воспользуемся для тех же целей критерием Колмогорова–Смирнова. Как в предыдущем случае воспользуемся представленной в пакете *nortest* функцией *ks.test*:

```
Two-sample Kolmogorov-Smirnov test

data:  Temperature and test.nsample
D = 0.0645, p-value = 0.9975
alternative hypothesis: two-sided
```

Вероятность ошибки $p > 0.05$, а значит нулевую гипотезу отвергнуть нельзя. Следовательно опровергнуть предположение о нормальности, как и предыдущих случаях, также нельзя. Проверим критерий: примем так же уровень значимости $\alpha = 0.05$, тогда критическое значение $D_{\text{кр}}(\alpha) = 1.358$. Следовательно,

$$D < D_{\text{кр}}(\alpha),$$

и подтверждаем сделанные ранее заключения: нельзя отвергнуть нулевую гипотезу о нормальности выборочного распределения.

В соответствии с результатами проверки критериев и на основе построенных гистограммы и графика квантилей, можно сделать заключение о том, что распределение температуры воды озера Баторино в июле 1975–2012 годов является близким к нормальному закону распределения.

Воспользуемся ещё одним из графических инструментов анализа данных — Bag Plot (диаграмма концентрации). Диаграмма концентрации является двумерным обобщением широко известного графика «ящик-с-усами». Данный инструмент применяется для поиска нетипичных наблюдений по сочетанию пары количественных признаков (в нашем случае это дата и температура). Основными компонентами данной диаграммы являются «мешок», который содержит 50% значений выборки, граница, которая отделяет внутренние точки от выбросов, и контур, показывающий точки снаружи «мешка», но внутри границы. Аналогично диаграмме размаха, диаграмма концентрации визуализирует некоторые характеристики выборочных данных: положение выборки (расположением медианы), распространение (размер «мешка»), корреляция (ориентация «мешка»), асимметрия (форма «мешка» и контура), хвосты (точки на границе контура и выбросы)[].

Ссылка

Результат построения диаграммы проиллюстрирован на рисунке 5. В центре полученной диаграммы находится медиана, «мешок» обозначен темным оттенком, контур светлым, выбросы обозначены перекрестием. Следовательно, выбросы не обнаружены, но следует отметить, что несколько точек находятся на самой границе контура — будем считать их подозрительными на выброс. По диаграмме также можно сказать о корреляции рассматриваемых переменных: «мешок» ориентирован вверх, что говорит о положительной корреляции. Также можно сделать заключения по асимметрии и проверить результаты, полученные на этапе вычисления описательных статистик. На диаграмме можно видеть, полученная фигура очень похожа на эллипс с центром, обозначенным медианой. Как следствие этого, можно судить о близости выборочного распределения с симметричным распределением, что подтверждает полученные ранее в анализе описательных статистик, представленных в таблице 1, результаты.

На данном этапе по результатам полученных на основе графиков 4 и 5 возникли подозрения о выбросах в исходной выборке. Выявление таких аномальных значений важно, так как их наличие, как правило, сильно влияет на всю выборку, в частности, на коэффициент корреляции. Проверим наличие выбросов с помощью статистических критериев.

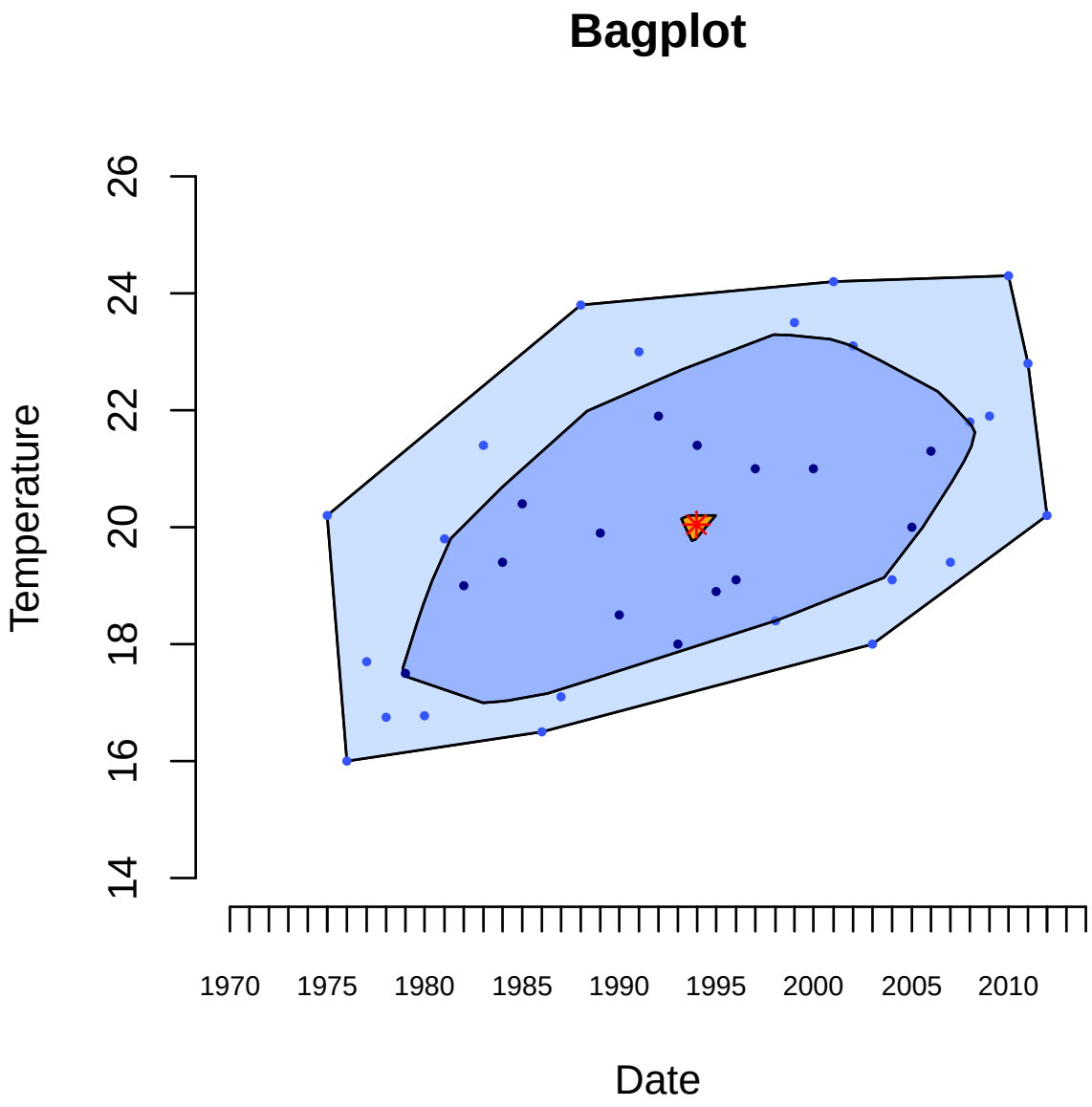

Рисунок 5 — Диаграмма концентрации.

Для этих целей воспользуемся критерием Граббса. Воспользуемся им для определения наличия выбросов в исходной выборке.

Полученные результаты проверки критерия Граббса:

```
Grubbs test for one outlier

data:  Temperature
G = 1.8435, U = 0.9057, p-value = 1
alternative hypothesis: highest value 24.3 is an outlier
```

Данный результат (p-value$=1$) однозначно говорит нам о том, что следует отклонить альтернативную гипотезу H_1 и принять гипотезу H_0. Другими словами, это говорит о том, что в исходной выборке нету выбросов. А значит выборка однородна. Значит, наши

предположения о выбросах на основе графического представления выборки оказались не подтвердились проверкой критерия.

+cite: Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21, 1, 27-58.

Исследуем теперь зависимость температуры воды от времени, построив диаграмму рассеяния и вычислив коэффициент соответствующих переменных.

Для начала построим корреляционную матрицу. Как видно из таблицы 2, коэффициент

	Temperature	Date
Temperature	1.00	0.52
Date	0.52	1.00

Таблица 2 — Корреляционная матрица.

корреляции $r_{xt} = 0.52$. Этим подтверждается наши выводы из диаграммы концентрации о положительной корреляции, поскольку полученный коэффициент корреляции является положительным по характеру и по таблице 1 средним (умеренным) по силе: $0.5 < r_{xt} < 0.7$.

Попробуем оценить значимость полученного выборочного коэффициента корреляции с помощью возможностей пакета **R** и описанного ранее критерия значимости.

Может стоит сделать нумерацию для критерия, чтобы делать на него ссылку?

Пакет **R** предоставляет с помощью функции *cor.test* различные методы для проверки значимости выборочного коэффициента корреляции. Воспользуемся проверкой теста методом Пирсона:

```
Pearson's product-moment correlation

data:  Temperature and Date
t = 3.6801, df = 36, p-value = 0.0007579
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2439316 0.7218701
sample estimates:
      cor
0.5228432
```

Как видно из полученных результатов $p - value < 0.05$, следовательно это говорит о том, что необходимо отвергнуть нулевую гипотезу.

Проверим с помощью критерия:

$$T_{\text{набл}} = \frac{r_{xt}\sqrt{n-2}}{\sqrt{1-r_{xt}^2}} \approx 3.98095.$$

Примем уровень значимости $\alpha = 0.05$. Число степеней свободы $k = n - 2 = 36$ (что подтверждает вычисленное функцией значение df). Тогда из таблицы критических точек распределения Стьюдента $t_k \approx 2,03$. Следовательно,

$$T_{\text{набл}} > t_k$$

Значит, нулевую гипотезу отвергаем и подтверждаем правильность полученных с помощью **R** результатов. Другими словами, выборочный коэффициент значимо отличается от нуля, т.е. температура воды и время при уровне значимости $\alpha = 0,05$ имеют зависимость.

Проиллюстрируем предположенную зависимость с помощью двумерной диаграммы рассеяния. Данные диаграммы используются для визуального исследования зависимости

между двумя переменными. Если переменные сильно связаны, то множество точек данных принимает определённую форму. С помощью таких диаграмм можно наглядно изучить знак коэффициента корреляции. Если точки на диаграмме расположены хаотически, то это говорит о независимости рассматриваемых переменных. Если с ростом переменной t возрастает переменная x то имеет место положительная корреляция. Если же с ростом переменной t переменная x убывает, то это указывает на отрицательную корреляцию. Построим такую диаграмму и проверим полученные ранее результаты в таблице 2.

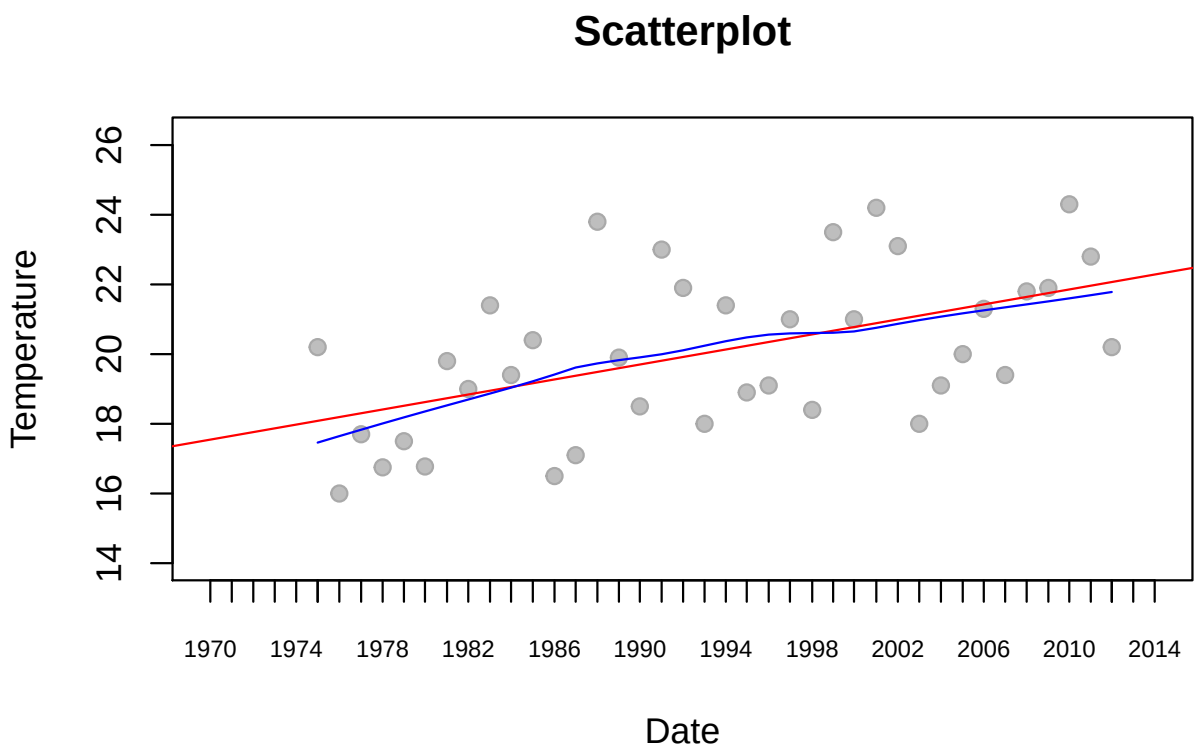

Рисунок 6 — Диаграмма рассеяния.

Из рисунка 6 видно, что точки образуют своеобразное «облако», ориентированное по диагонали вверх. Что, в свою очередь, подтверждает полученные ранее результаты о коррелированности температуры и времени, в том числе и положительность коэффициента корреляции (ориентированность вверх). Но при этом, данная диаграмма наглядно показывает степень коррелированности: так как точки не образуют чёткой формы, а разбросаны относительно диагонали, то можно говорить о наличии умеренной корреляции. То есть, нельзя сказать, что корреляция сильная, но и нельзя сказать, что связь между переменными отсутствует.

Думаю, здесь нужно сделать выводы о проделанной в данной части работе — подвести итог