Основные понятия математической статистики

Малов Сергей Васильевич

Санкт-Петербургский государственный электротехнический университет

5/19 сентября 2020 г.

План

Введение

Статистическое исследование включает в себя

- Сбор данных
- Организацию собранных данных
- Анализ данных
- Интерпретацию полученных результатов

Введение

Планирование статистического исследования - важнейший этап статистического анализа. На этапе планирования происходит

- Осмысление целей предполагаемого исследования
- Разработка стратегии сбора статистических данных
- Выбор математической модели статистического эксперимента
- Формулировка задач статистического исследования

Введение

Статистический эксперимент может быть активным или пассивным

- При пассивном эксперименте исследователь не имеет возможности влиять на условия проведения статистического эксперимента и его роль ограничивается наблюдением за изучаемым явлением.
- Активный эксперимент позволяет исследователю самостоятельно формировать условия проведения эксперимента, чтобы оптимизировать возможности последующего анализа для достижения поставленных целей.

Введение

При интерпретации результатов статистического анализа следует учитывать

- Постановку статистического эксперимента
- Соответсвие выбранной математической модели реальным условиям статистического эксперимента
- Возможности использованных методов статистического анализа
- Все многообразие полученных результатов, если при анализе данных использовались различные методы.

План

Статистический эксперимент

В абстрактную модель статистического эксперимента входят три объекта $(\mathfrak{X},\mathfrak{F},\mathcal{P})$

- $\mathfrak{X}$ – множество допустимых результатов эксперимента
- $\mathfrak{F}$ – множество наблюдаемых событий или статистическая информация, получаемая в результате эксперимента
- $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ – семейство допустимых вероятностных распределений или параметризация.

Замечание. Следует отметить, что $(\mathfrak{X}, \mathfrak{F}, \mathrm{P}_\theta)$ при каждом фиксированном значении θ представляет собой вероятностный эксперимент. В основе методов математической статистики лежат вероятностные результаты.

Накопление статистической информации, выборка.

Одним из основных принципов статистического исследования является возможность накопления информации, позволяющая делать более достоверные выводы.

Выборочный принцип накопления статистической информации

- В математической статистике выборкой $X = (X_1, \ldots, X_n)$ принято называть набор независимых и одинаково распределенных случайных величин (НОРСВ) или векторов.
- Распределение выборки однозначно определяется распределением каждой ее компоненты, что позволяет параметризовать выборку распределением одной ее компоненты.

Выборочный принцип

- Выборка реализуется повторением эксперимента в одних и тех же условиях.
- Альтернативно, выборка может быть получена с использованием случайного выбора n элементов из генеральной совокупности N элементов, если N >> n.
- Следует отметить, что случайный выбор из генеральной совокупности без возвращения не дает независимых случайных величичн, однако если N >> n, то отклонение от НОРСВ несущественно.
- Если сама генеральная совокупность представляет собой набор НОРСВ, то случайный выбор n элементов из N представляет собой набор НОРСВ.

Накопление статистической информации

Выборочный принцип накопления статистической информации не всегда удовлетворяет потребностям статистического исследования.

- Например, если исследователь хочет контролировать условия проведения экспереимента, то речь по сути идет об условном распределении наблюдаемой величины при фиксированных условиях проведения эксперимента.
- Наблюдения могут оставаться независимыми, однако, если не делать никаких предположений о связи распределений наблюдений при различных условиях, то серьезные статистические выводы вообще говоря невозможны.

Накопление статистической информации

Простейшим обобщением выборочной модели является случай нескольких выборок

- Если существует лишь конечное число возможных условий проведения экспримента, то, в случае независимых наблюдений, получаем несколько выборок по числу условий.
- Помимо задач, связанных с распределением каждой из выборок, появляются задачи сравнения распределений различных выборок.
- В случае многообразия условий проведения экспереимента, для накопления статистической информации требуется установить связь между распределениями в различных условиях.

Регрессионные модели

Наиболее распространенными моделями такого типа являются регрессионные модели.

- Регрессией величины Y по величине X называется $E(Y|X) = g(X)$ – среднее значение величины Y в условиях X.
- В основе регрессионной модели лежит соотношение $E(Y|X) = g(X;\theta)$, $\theta \in \Theta_1$.
- Следует отметить, что параметр θ не определяет распределения набора наблюдений, поэтому в регрессионной модели также присутствует дополнительный (мешающий) параметр.

Асимптотический подход

Точные методы в статистике достаточно редки и неуниверсальны (требуют достаточно жестких предположений), в связи с чем огромную роль играют асимптотические методы.

- Вместо одного статистического эксперимента рассматривается последовательность статистических экспериментов $(\mathfrak{X}_n, \mathfrak{F}_n, \mathcal{P}_n)$, $\mathcal{P}_n = \{P_{\theta,n}, \theta \in \Theta\}$.
- Отметим, что параметрическое множество не зависит от n, часто это позволяет делать все более точные выводы о параметре в процессе накопления статистической информации при увеличении n.
- В качестве примера можно привести выборку $X_1, \ldots, X_n$ из распределения P_θ, $\theta \in \Theta$. С ростом n происходит накопление статситической информации.

Параметризация

Важной составляющей модели статистического эксперимента является параметризация – семейство допустимых распределений наблюдений и его формализация с помощью параметра.

- Параметр должен быть идентифицируемым, т.е. каждому распределению множества допустимых распределений должно соответствовать единственное значение параметра – свойство инъективности.
- Близким значениям параметра (если параметрическое множество плотное) должны соответствовать близкие распределения – свойство непрерывности.

Типы параметрических семейств

Методы анализа статистических данных существенно зависят от типа параметризации. Статистические эксперименты по типу параметризации классифицируют в три группы.

- Параметрические – параметр представлят собой вещественное число $\theta \in \Theta \subseteq \mathbb{R}$ или вектор $\theta = (\theta_1, \ldots, \theta_{\mathrm{d}}) \in \Theta \subseteq \mathbb{R}^{\mathrm{d}}$.
- Семипараметрические – параметр состоит из вещественного числа или вектора и дополнительной (обычно функциональной) компоненты $\theta = (\theta_1, \theta_2)$ и $\theta_1 \in \mathbb{R}^{\mathrm{d}}$.
- Непараметрические – все остальные статистические эксперименты.

План

Задачи статистического анализа

Цель статистического анализа данных в условиях статистического эксперимента $(\mathfrak{X}, \mathfrak{F}, \mathcal{P})$, $\{\mathrm{P}_\theta : \theta \in \Theta\}$ — сделать те или иные выводы об исинном значении параметра θ. Различают три типа задач статистического анализа

- Точечное оценивание параметра
- Доверительное оценивание параметра
- Проверка статистических гипотез

Для решения задач статистического анализа используются статистики – функции от наблюдений $\mathrm{T} : \mathfrak{X} \to \mathrm{E}$. По сути результаты статистического анализа строятся на базе значений набора статистик.

Достаточные и подчиненные статистики

Среди статистик в условиях статистического эксперимента $(\mathfrak{X},\mathfrak{F},\mathcal{P})$ можно выделить два типа:

- Подчиненной называется статистика, распределение которой не зависит от θ. Подчиненная статистика не несет никакой информации о параметре и не может быть использована для статистического анализа.
- Если условное распределение исходного набора наблюдений при условии T не зависит от параметра модели θ, то такая статистика называется достаточной. Достаточная статистика содержит ту же информацию о параметре распределения, что и исходный набор данных. Имеет смысл строить результаты анализа с использованием достаточных статисик.
- Весь исходный набор наблюдений всегда является достаточной статистикой. Достаточная статистика, обеспечивающая максимальную редукцию данных называется минимальной.

Точечное оценивание

Задача точечного оценивания состоит в том, чтобы по результатам наблюдений найти приближенное значение параметра.

- Точечная оценка – статистика $\delta : \mathfrak{X} \to \Theta$.
- Гауссовский риск точечной оценки $\mathrm{R}_\delta(\theta) = \mathrm{E}_\theta(\delta(\mathrm{X}) - \theta)^2$ обычно используют для измерения точности оценки.
- Стараются выбрать оценку с мимнимальным риском, но в классе всех оценок такой не существует. Оптимальную оценку можно искать
  - В ограниченном классе разумных оценок (н-р, среди несмещенных оценок)
  - С использованием функционалов риска (байесовский, минимаксный подходы)

Доверительное оценивание

Задача доверительного оценивания состоит в том, чтобы по результатам наблюдений выбрать область значений параметра, накрывающую теоретическое значение параметра с достаточно большой вероятностью.

- Доверительная оценка – статистика $\hat{\Theta} : \mathfrak{X} \to \mathcal{Y}$, где $\mathcal{Y}$ – совокупность подмножеств множества Θ определенной формы.
- Доверительная оценка $\hat{\Theta}$ имеет уровень доверия $1 - \alpha$, если

$$\mathrm{P}(\theta \in \hat{\Theta}) \geq 1 - \alpha \quad \text{при любом} \quad \theta \in \Theta.$$

Доверительное оценивание

Наиболее распространенные доверительные множества — доверительные интервалы.

- Если параметр распределения — вещественное число ($\Theta \subseteq \mathbb{R}$), то для доверительного оценивания используются интервалы $\hat{\Theta} = [T_1(X), T_2(x)]$.
  - Различают односторонние ($T_1 \equiv \inf\{\Theta\}$ – правосторонний; $T_2 \equiv \sup\{\Theta\}$ – левосторонний) и двухсторонние доверительные интервалы.
  - Чем короче длина двухстороннего доверительного интервала при фиксированном уровне доверия, тем точнее оценка.
  - Точность доверительной оценки зависит от эффективности использования имеющейся статистической информации.

Доверительное оценивание

- В случае многомерного параметра можно говорить о доверительных интервалах для отдельных параметров или функций параметров. Следует учитывать, что результат можно интерпретировать для любого из параметров, но не для всех парамеров в совокупности.
- Иногда удается построить совместные доверительные интервалы для нескольких параметров или функций параметров.
- Метод множественного оценивания Шеффе устанавливает связь между доверительными эллипсоидами и совместными доверительными интервалами в случае совместного нормального распределения оценок параметров или функций

Построение доверительных интервалов

Рассмотрим метод построения доверительных интервалов, широко использующийся в математической статистике

- Пусть $G(X, \theta)$ функция параметра и наблюдений, удовлетворяющая условиям
  - Распределение $G(X, \theta)$ не зависит от θ.
  - Для любого α существует I_α такое, что $P_\theta(G(X, \theta) \in I_\alpha) \geq 1 - \alpha$ и $\hat{\Theta} = \{\theta : G(X, \theta) \in I_\alpha\} \in \mathcal{Y}$.
- Тогда $\hat{\Theta}$ — доверительное множество уровня доверия $1 - \alpha$.
- Функцию $G(X, \theta)$ будем называть генератором доверительного множества $\hat{\Theta}$.

Проверка статистических гипотез

Статистической гипотезой называется утверждение о теоретическом распределении, выражаемое в терминах параметра $\theta \in \Theta_{\mathrm{H}}$, $\Theta_{\mathrm{H}} \subseteq \Theta$. В классической постановке выдвигают

- основную (или нулевую) гипотезу $\mathrm{H}_0 : \theta \in \Theta_0$
- и альтернативную гипотезу (альтернативу) $\mathrm{H}_{\mathrm{A}} : \theta \in \Theta_{\mathrm{A}}$: $\Theta \cap \Theta_{\mathrm{A}} = \varnothing$.

Задача исследователя – по результатам наблюдений сделать выбор между основной гипотезой и альтернативой.

Проверка статистических гипотез

Различные варианты истинного положения дел и решения исследователя приведены в следующей таблице.

	Принята H_0	Отвергнута H_0
H_0 верна	+	Ошибка I рода
H_0 не верна	Ошибка II рода	+

Проверка статистических гипотез

Исследователь старается минимизировать ошибки

- Подход Неймана–Пирсона заключается в том, чтобы ограничить ошибку I рода малым наперед заданным числом α.
- Критерий – правило, согласно которому принимается или отвергается основная гипотеза.
- Формально критерий – статистика $\phi: \mathfrak{X} \to [0, 1]$, определяющая вероятность отвергнуть основную гипотезу по результатам наблюдений (0 – принимаем основную гипотезу, 1 – отвергаем).
- Критерий называется нерандомизованным, если результаты наблюдений однозначно определяют решение $\phi(\mathfrak{X}) = \{0, 1\}$.
- Значение α, ограничивающее вероятность ошибки I рода называется уровнем значимости критерия.

Проверка статистических гипотез

- Мощность критерия равна $1 - \mathrm{P}_\theta$(ош. II рода).
- Важную роль играет среднее значение $\mathrm{E}_\theta\phi(\mathrm{X})$

$$\mathrm{b}(\theta) = \mathrm{E}_\theta\phi(\mathrm{X}) = \begin{cases} \text{вероятность ош. I рода, при } \theta \in \Theta_0 \\ \text{мощность, при } \theta \in \Theta_\mathrm{A} \end{cases}$$

Проверка статистических гипотез

Обычно в основе статистичекого критерия лежит статистика критерия. Статистика критерия удовлетворяет следующим условиям

- Распределение статистики критерия не зависит от параметра при справедливости нулевой гипотезы $\theta \in \Theta_0$.
- Это распределение изучено (существуют таблицы)
- Распределение статистики критерия при альтернативе $\theta \in \Theta_A$ отличается от ее распределения при справедливости нулевой гипотезы.

Замечание. Рассматриваемый способ построения статистического критерия наиболее распостраненный, но далеко не единственный.

Проверка статистических гипотез

Пусть T – статистика критерия; P_0 – ее распределение при нулевой гипотезе.

- Для построения статистического критерия потребуется дополнительно набор множеств $\mathcal{I}$ (доверительных), удовлетворяющих условию: для любого α существует $I_\alpha \in \mathcal{I}$: $P_0(T \in I_\alpha) \geq (=) 1 - \alpha$.
- Исходя из распределения статистики критерия строим нерандомизованный критерий

$$\phi(X) = \begin{cases} 0, & T \in I_\alpha \\ 1, & T \notin I_\alpha \end{cases}$$

P-значение

Рассмотренный метод построения статистического критерия позволяет определить P-значение — наименьшее α, такое что $T \notin I_\alpha$.

- При положительном распределении T наиболее часто используется $I_\alpha = [0, x_\alpha]$, где x_α удовлетворяет условию $P_{\theta_0}(T > x_\alpha) = 1 - F_T(x_\alpha) = \alpha$, F_0 функция распределения T при нулевой гипотезе.
- В этом случае P-значение равно $PV = 1 - F_T(T)$ и имеет равномерное $U(0,1)$ распределения (преобразование Смирнова) при нулевой гипотезе, если F_0 непрерывная функция.
- Аналогичное свойство выполнено и для других тестов,

Распределение P-значения

Утверждение (обобщение преобразования Смирнова).

Пусть G статистика критерия для проверки статистической гипотезы H_0, $\{I_\alpha\}_{\alpha\in[0,1]}$ — семейство вложеных замкнутых множеств $I_\alpha \subseteq I_{\alpha_1}$ при любых $\alpha > \alpha_1$ и

$$P(T \in I_\alpha) = 1 - \alpha \quad \text{при всех} \quad \alpha \in [0,1],$$

P-значение — $PV = \inf\{\alpha : T \notin I_\alpha\}$. Тогда, $PV \sim U(0,1)$.

Доказательство.

Отметим, что для любого $\epsilon > 0$: $\alpha + \epsilon \leq 1$,

$$P(PV \in [\alpha, \alpha+\epsilon)) = P(T \in I_\alpha \backslash I_{\alpha+\epsilon}) = P(T \in I_\alpha) - P(T \in I_{\alpha+\epsilon}) = \epsilon.$$

Следовательно, распределение PV – аюсолютно непрпрерывно и имеет $U(0,1)$ распределение.

Распределение P-значения

Утверждение (обобщение преобразования Смирнова).

Пусть G статистика критерия для проверки статистической гипотезы H_0, $\{I_\alpha\}_{\alpha\in[0,1]}$ — семейство вложеных замкнутых множеств $I_\alpha \subseteq I_{\alpha_1}$ при любых $\alpha > \alpha_1$ и

$$P(T \in I_\alpha) = 1 - \alpha \quad \text{при всех} \quad \alpha \in [0,1],$$

P-значение — $PV = \inf\{\alpha : T \notin I_\alpha\}$. Тогда, $PV \sim U(0,1)$.

Доказательство.

Отметим, что для любого $\epsilon > 0$: $\alpha + \epsilon \leq 1$,

$$P(PV \in [\alpha, \alpha+\epsilon)) = P(T \in I_\alpha \backslash I_{\alpha+\epsilon}) = P(T \in I_\alpha) - P(T \in I_{\alpha+\epsilon}) = \epsilon.$$

Следовательно, распределение PV – аюсолютно непрпрерывно и имеет $U(0,1)$ распределение.

Доверительные множества и критерии

Рассмотренный ранее метод построения доверительных интервалов может быть использован для построения критериев.

- Пусть $H_0 : \theta = \theta_0$ – простая гипотеза; $G(X, \theta)$ — функция-генератор доверительного множества $\hat{\Theta} = \{\theta : G(X, \theta) \in I_\alpha\}$ уровня доверия $1 - \alpha$.
- Дополнительно предположим, что распределение $G(X, \theta)$ отличается от распределения $G(X, \theta_0)$ при любом $\theta \in \Theta_A$.
- Тогда,
$$\phi(X) = \begin{cases} 0, & G(X, \theta_0) \in I_\alpha \\ 1, & G(X, \theta_0) \notin I_\alpha \end{cases} \quad \text{или} \quad \phi(X) = \begin{cases} 0, & \theta_0 \in \hat{\Theta} \\ 1, & \theta_0 \notin \hat{\Theta} \end{cases}$$
— статистический критерий для проверки H_0 уровня значимости α.
- Данный метод позволяет строить критерии и для сложных гипотез, но при их построении следует находить I_α исходя из наибольшего значения вероятности при $\theta \in \Theta_0$.

Проверка значимости

Наиболее распространенным методом в медицинских исследованиях является проверка значимости отклонений от основной гипотезы.

- Существенным результатом является отвержение основной гипотезы – выявление значимых отклонений.
- Реальные значения отклонений от основной гипотезы не изучаются.
- Наличие богатой статистической информации позволяет выявлять даже несущественные с практической точки зрения различия.
- Если проверяется несколько статистических гипотез, то необходима поправка.

Типы статистических гипотез

В зависимости от имеющихся статистических данных и выбранной модели эксперимента существует четыре основных типа статситических гипотез.

- Гипотеза согласия ставится в терминах распределения отдельных наблюдений (например, элементов выборки) о согласии с некоторым фиксированным распределением (простая) или принадлежности некоторому множеству распределений (сложная).

Типы статистических гипотез

- При наличии двух или нескольких выборок или неодинаково распределенных результатов наблюдений, распределение которых контролируется значением сопутствующего фактора, гипотеза однородности состоит в том, что распределения в различных выборках или при различных значениях сопутствующего фактора совпадают.
- При наличии пары или нескольких измеряемых величин часто выдвигают гипотезу их независимости.
- Гипотеза случайности набора наблюдений $X_1, \ldots, X_n$ состоит в том, что он является выборкой.

Различение гипотез, «тонкие» гипотезы

Если нулевая гипотеза и альтернатива имеют общую границу, то их различение не представляется возможным, если теоретическое значение параметра находится на границе.

- Нулевую гипотезу будем называть «тонкой», если множество параметров Θ_0 совпадает со своей границей.
- Большинство классических задач проверки статистических гипотез имеют дело с тонкими гипотезами. Например,
  - простая гипотеза согласия $\mathrm{H}_0 : \theta = \theta_0$,
  - гипотеза однородности,
  - гипотеза независимости,
  - гипотеза случайности.

«Тонкие» гипотезы

Обычно проще проверять именно «тонкие» гипотезы, однако для получения серьезных статистических выводов следует обязательно принимать во внимание альтернативу.

- Приведенный ранее метод построения статистических гипотез обычно удается реализовать только для «тонких» гипотез.
- При проверке «тонкой» гипотезы следует определиться, какое именно отклонение мы будем считать значимым, поскольку, имея достаточно большое число статистических данных, можно выявить даже незначительные отклонения от основной гипотезы.
- Выбор минимального значимого отклонения позволяет спланировать эксперимент таким образом, чтобы различить нулевую гипотезу и значимое отклонение от нее с достаточно большой вероятностью, при этом не потратив лишних средств на получение чрезмерной статистической информации.

План

Наблюдаемые и изучаемые переменные

Цели статистического исследования и статистические данные должны соответствеовать друг другу.

- Цель статистического исследования (параметр) обычно ставится в соответствие распределению некоторой случайной величины (или вектора), которую будем называть изучаемой переменной.
- Наблюдаемыми переменными называются случайные величины, реализации которых представляют собой статистические данные.
- Наиболее часто изучаемые и наблюдаемые переменные совпадают.

Типы наблюдаемых и изучаемых переменных

Наблюдаемые/изучаемые переменные можно классифицировать по их типу

- Количественные – характеризующие числовое значение изучаемой характеристики.
- Порядковые – характеризующие порядок, но не величину наблюдаемой характеристики.
- Категориальные – переменные группировки, числовые значения категориальных переменных не важны.

Типы наблюдаемых и изучаемых переменных

Количественные переменные:

- Могут быть как непрерывными, так и дискретными.
- Распределение самой случайной величины и ее числовые характеристики могут представлять интерес для статистического анализа.

Порядковые переменные:

- Дискретные целочисленные.
- Интерес представляют лишь атомы соответствующего дискретного распределения с учетом порядка.

Категориальные переменные:

- Дискретные целочисленные или вовсе нечисловые.
- Интерес представляют лишь атомы соответствующего дискретного распределения.

Типы статистических данных

Можно выделить несколько типов статистических данных

- Простые налбюдения – случайные величины, объединяемые в векторы. В зависимости от характера наблюдаемой переменной, они могут быть количественными, порядковыми или категориальными.
- Временные ряды – наборы переменных, характеризующих течение изучаемого процесса во времени, которые имеет смысл интерпретировать как случайный, т.е. значения случайного процесса в выбранных точках.
- Данные типа времени жизни – измерения времен до определенного события (например, заболевания). Часто момент интересующего события не наблюдается, что мотивирует вводить в модель цензурирование справа.