

1. Basic concepts and tasks of statistics. Types and sources of statistical data

Основные определения:

- Генеральная совокупность – сведения о всех анализируемых объектах;
- Выборка – множество результатов, отобранных из генеральной совокупности (репрезентативность);
- Объем совокупности – число единиц, образующих совокупность;
- Неопределенность и вариация (основные характеристики статистики) – при многократных измерениях происходят изменения;
- Признак – характеристика единицы совокупности;
- Показатель (индикатор) – количественная характеристика явления;
- Параметр – относительно постоянная величина, характеризующая генеральную совокупность;
- Выборочная характеристика (статистика) – эмпирический аналог параметра;
- Статистические выводы – заключения, формируемые анализом эмпирических данных.

Виды и источники статистических данных

Статистические данные разделяются на:

- Пространственные: сведения об объектах наблюдения с различным порядком;
- Временные: хронологический порядок (моментные – сумма только сумма значений и интервальные – суммирование дает общую характеристику и может быть проинтерпретирована);
- Пространственно-временные: набор объектов в хронологическом порядке со сведениями об объектах.

Статистические данные могут быть одномерными и многомерными, количественные и категориальные, первичные (регистрируемые для одного конкретного объекта) и агрегированные (объект – совокупность других объектов).

Шкалы измерения данных: качественные данные – номинальная (профессия, пол, город), порядковая (место в рейтинге), количественные (непрерывные и дискретные) – интервальная (температура воздуха), относительная (количество наличных денег, времени, объектов).

Источники статистических данных: непосредственные измерения, мнения экспертов, документированные значения.

- Статистическое наблюдение – планомерный и систематический сбор данных об исследуемых явлениях и процессах, бывает сплошным (на генеральной совокупности) и несплошным (на выборке).

Задачи статистики

В узком смысле: сжатие информации и наглядное представление результатов.

В широком смысле: обобщение результатов выборочного исследования на генеральную совокупность.

Первичная обработка: пример данных качественного характера

- таблица частот;
- столбиковая диаграмма;
- круговая диаграмма.

Первичная обработка: количественного характера.

- гистограмма

Этапы статистического моделирования:

Определение цели и задач моделирования

1. Формализация – преобразование объектов и отношений в математическую абстрактную модель;
2. Сбор и квантификация данных – предусматривает отражение данных в шкалах, их предварительная обработка – избавление от ошибок;
3. Спецификация модели – представление в виде формул;
4. Идентификация модели и ее анализ – оценка параметров модели, ее характеристик;
5. Верификация модели.

Выборочные статистики

Выборочная статистика – эмпирический аналог параметра. Выборочная характеристика является функцией от результатов наблюдений $\theta^* = \theta^*(x_1, \dots, x_n)$

Порядковые статистики

- Медиана – величина, разделяющая упорядоченный набор на 2 равные части – 50% всех наблюдений находятся ниже медианы, 50% выше.
- Первый квартиль – величина, разделяющая упорядоченную выборку, 25% всех наблюдений лежит ниже первого квартиля, 75%, соответственно, выше.
- Аналогично можно определить второй квартиль, причем, становится ясно, что понятия первого квартиля и медианы совпадают;
- Третий квартиль итеративно можно определить как разделяющую величину, ниже которой лежат 75% наблюдений, оставшиеся 25% выше.
- Первый дециль: 10% наблюдений лежат ниже, 90% выше.
- Интерквартильный размах IQR – разность между третьим и первым квартилями (служит мерой разброса)

Моментные характеристики положения и разброса

- Среднее значение – сумма значений признака, деленная на число его значений. (характеристика положения);
- Дисперсия – среднее значения квадрата отклонения результатов наблюдений от среднего значения (разброс);
- Среднее квадратическое отклонение – положительный квадратный корень из дисперсии.

Важные статистики

- Выборочной характеристикой называется функция от результатов наблюдений: Можем определить выборочное среднее значение двумя способами:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{среднее арифметическое}$$

$$\bar{x} = \sqrt[n]{\prod_{i=1}^n x_i} \quad \text{среднее геометрическое}$$

Так же можно определить выборочные дисперсию и среднее квадратичное отклонение:

$$\begin{aligned} \text{var}(x_i) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 && \text{выборочная дисперсия} \\ S &= \sqrt{\text{var}(x_i)} && \text{выборочное среднее квадратичное отклонение} \end{aligned}$$

Выборочные начальные и центральные моменты:

$$\begin{aligned} v_k^* &= \frac{1}{n} \sum_{i=1}^n x_i^k \\ \mu_k^* &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \end{aligned}$$

Коэффициенты асимметрии и эксцесса:

$$\begin{aligned} \beta_1 &= \frac{\mu_3^*}{S^3} \\ \beta_2 &= \frac{\mu_4^*}{S^4} - 3 \end{aligned}$$

Характеристики равномерности распределения количественного признака

- Кривая Лоренца;
- Коэффициент Джини:

$$G = 1 - 2 \sum_{i=1}^n x_i y_{i_{\text{нак}}} + \sum_{i=1}^n x_i y_i.$$

2. Some basic information from probability theory to construct the statistical models.

Понятие вероятности

Definition

Вероятность P – численная мера объективной возможности наступления события. Вероятностное пространство (Ω, \mathcal{F}, P) – состоит из пространства элементарных исходов Ω , сигма-алгебры на этом пространстве \mathcal{F} и самой вероятности P , формально, сигма-аддитивной меры.

Note

Требования к вероятности:

$$P(\Omega) = 1, \\ P(A \cup B) = P(A) + P(B)$$

Definition

События A и B несовместны, если $AB = \emptyset$

Definition

События образуют полную группу событий, если их сумма является достоверным событием.

Definition

Полная группа попарно несовместных событий определяется как произведение $A_i A_j = \emptyset$ при $i \neq j$ и $\sum_{i=1}^n A_i$ является достоверным событием, то есть $P\left(\sum_{i=1}^n A_i\right) = 1$.

Другое определение вероятности:

Definition

Вероятность наступления события A можно представить как предел относительной частоты при большом количестве испытаний:

$$P(A) \approx \frac{m}{n}.$$

Вероятность наступления сложных событий

Theorem

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Corollary:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Theorem: Inclusion-Exclusion

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P(A_1 \cap \dots \cap A_n)$$

Note

Частный случай для несовместных событий A и B :

$$P(A \cup B) = P(A) + P(B)$$

Аналогично для суммы попарно несовместных событий $\sum_{i=1}^n A_i$:

$$P(A_1 + \dots + A_n) = \sum_{i=1}^n P(A_i).$$

Theorem

Произведение зависимых событий

$$P(AB) = P(A)P(A|B) = P(B)P(B|A),$$

где $P(A|B)$ – условная вероятность события A .

Note

События A и B независимы, если

$$P(A|B) = P(A)$$

Вероятность произведения попарно независимых событий:

$$P(A_1 \dots A_n) = \prod_{i=1}^n P(A_i).$$

Definition

При известных априорных вероятностях $P(A_i)$, условная вероятность $P(B|A_i)$, определяется следующим образом:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)},$$

где A_i – событие, априорное событию B .

Theorem

Формула полной вероятности:

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i)$$

Theorem: Bayes

При известных априорных вероятностях $P(A)$ и $P(B)$ и условной вероятности $P(B|A)$, вероятность $P(A|B)$ может быть определена:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Схема Бернулли

Рассмотрим событие B_m , состоящее в том, что событие A в n повторных независимых испытаниях наступит ровно m раз. Вероятность такого события определяется следующим образом:

$$P_{n,m} = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m},$$

где p – вероятность наступления события A в каждом испытании.

Аппроксимации в случае большого количества испытаний:

- Случай редких событий (событий, вероятность которых стремится к нулю). При этом интенсивность событий постоянна: $\lambda = np$. В данном случае можно воспользоваться теоремой Пуассона:

$$\lim_{n \rightarrow \infty} P_n(m) = \frac{\lambda^m}{m!} e^{-\lambda}$$

- Когда вероятность наступления события примерно равна 0.5, можно воспользоваться локальной теоремой Лапласа:

$$\lim_{n \rightarrow \infty} P_n(m) = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{1}{2} \frac{(m-np)^2}{npq}}$$

Random variables**Definition: (Random variables)**

Given an experiment with simple space S , a random variable is a function X of a kind $X : S \rightarrow \mathbb{R}$.

Пример: Число очков на игральной кости; оценка, полученная на экзамене; время ожидания автобуса на остановке.

Example: (Coin tosses) We toss a fair coin twice. The sample space is $S = \{HH, HT, TH, TT\}$. Here are some r.v.-s on this space:

- $X = \#$ of Heads:

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0$$

- $Y = \#$ of Tails: $Y = 2 - X$

- $I = \begin{cases} 1, & \text{if 1-st toss = Heads,} \\ 0, & \text{otherwise} \end{cases}$ – indicator random variable.

Distributions and probability mass functions

There are two main types of r.v.-s.: discrete and continuous.

Definition: (Discrete random variable)

A random variable X is said to be discrete if there is a finite list of values $\alpha_1, \alpha_2, \dots, \alpha_n$ or an infinite set $\alpha_1, \alpha_2, \dots$ s.t. $(\exists j) P(X = \alpha_j) = 1$. If X is a discrete r.v., then this finite or countably infinite set of values it takes and such that $P(X = x) > 0$ is called the support of X .

Note

Continuous r.v.-s can take any real value in an interval.

The distribution of an r.v. specifies the probabilities of all events associated with the r.v. For a discrete case the most natural way to do this is:

Definition: (Probability mass function)

The probability mass function (PMF) of a discrete r.v. X is the function p_X given by $p_X(x) = P(X = x)$. It is non-zero positive if $x \in (\text{support } X)$ and 0 otherwise.

Note

In writing $P(X = x)$, $X = x$ denotes an event. (Sometimes also written as $\{X = x\}$ – formally, $\{s \in S : X(s) = x\}$).

Example: (Two coin tosses). $X = \#$ of Heads, $Y = \#$ of Tails, $I = \begin{cases} 1, & \text{if 1-st toss = Heads,} \\ 0, & \text{otherwise} \end{cases}$ – indicator variable.

$$p_X(0) = P(X = 0) = \frac{1}{4}, \quad p_X(1) = \frac{1}{2}, \quad p_X(2) = \frac{1}{4},$$

$$Y = 2 - X, \text{ so same PMF. } p_I(0) = \frac{1}{2}, \quad p_I(1) = \frac{1}{2}.$$

Example: (sum of die rolls). Roll two fair 6-sided dice. Let $T = X + Y$, where X, Y are individual rolls. The sample space is $S = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$:

$$p_T(2) = p_T(12) = \frac{1}{36}, \quad p_T(3) = p_T(11) = \frac{2}{36}, \dots, \quad p_T(7) = \frac{6}{36}.$$

Theorem: (Valid PMFs)

Let X be a discrete random variable with support x_1, x_2, \dots . The PMF p_X of X must satisfy:

- Nonnegative: $p_X(x) > 0$ if $x = x_j$ for some j , $p_X(x) = 0$ otherwise;
- Sums to 1: $\sum_{j=1}^{\infty} p_X(x_j) = 1$.

Bernoulli and Binomial

Definition: (Bernoulli distribution)

A random variable X is said to have Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write $X \sim \text{Bern}(p)$ (X is Bernoulli-distributed). It is a family of distributions indexed by p .

Definition: (Indicator random variable)

Indicator r.v. of an event A = r.v. that equals 1 if A occurs and 0 otherwise. It denoted by I_A or $I(A)$.

Note

$$I_A \sim \text{Bern}(p) \text{ with } p = P(A).$$

An experiment that can result in a “success” or “failure” (but not both) is called a Bernoulli trial. A Bernoulli r.v. thus = indicator r.v. of success in Bernoulli trial.

Suppose n independent Bernoulli trials are run, each with $P(\text{success}) = p$. Let X = the number of successes. $X \sim \text{Bin}(n, p)$ – the Binomial distribution with parameters $n = 1, 2, \dots$ and $0 < p < 1$.

Note

$\text{Bern}(p)$ is the same as $\text{Bin}(1, p)$

Theorem: (Binomial PMF)

If $X \sim \text{Bin}(n, p)$, then the PMF of X :

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{for } k = 0, 1, \dots, n.$$

Theorem

Let $X \sim \text{Bin}(n, p)$ and $q = 1 - p$ (we often use q as a failure probability in Bernoulli trial). Then $n - X \sim \text{Bin}(n, q)$ (based on the binomials symmetry property).

Corollary: Let $X \sim \text{Bin}(n, p)$ with $p = \frac{1}{2}$ and even n . Then the distribution of X is symmetric about $\frac{n}{2}$ – that is:

$$P(X = \frac{n}{2} + j) = P(X = \frac{n}{2} - j)$$

Hypergeometric

Preface: Urn with w white and b black balls, drawing n balls with replacement yields $\text{Bin}(n, \frac{w}{w+b})$, for X – # of white balls in n trials. If we instead sample without replacement, then X follows a Hypergeometric distribution: $X \sim \text{HGeom}(w, b, n)$. In Bernoulli trials are independent, in Hypergeometric trials are dependent (cause of without replacement nature).

Theorem: (Hypergeometric PMF)

If $X \sim \text{HGeom}(w, b, n)$, then

$$P(X = k) = \frac{\binom{w}{k} \binom{b}{n-k}}{\binom{w+b}{n}}$$

Example: (Aces in a poker hand). In a 5-card hand from a well shuffled deck, the # of aces $\sim \text{HGeom}(4, 48, 5)$. Then

$$P(3 \text{ aces}) = \frac{\binom{4}{3} \binom{48}{2}}{\binom{52}{5}} \approx 0.0017.$$

Theorem

$\text{HGeom}(w, b, n)$ and $\text{HGeom}(n, w + b - n, w)$ are identical.

Discrete uniform

Preface: let C be a finite nonempty set of numbers. Choose one of these uniformly at random (i.e. all values are equally likely). Call the chosen number X . Then X is said to have the Discrete Uniform distribution with parameter C , which one can be denoted as $X \sim \text{DUnif}(C)$.

Note

The PMF is $P(X = x) = \frac{1}{|C|}$ for $x \in C$ (and 0 otherwise). For any subset $A \subset C$, $P(X \in A) = \frac{|A|}{|C|}$.

Числовые статистики для дискретных и непрерывных случайных величин

Начальные и центральные моменты для дискретных случайных величин:

$$\begin{aligned} \nu_k^* &= \sum_{i=1}^n x_i^k p_i \\ \mu_k^* &= \sum_{i=1}^n (x_i - \bar{x})^k p_i. \end{aligned}$$

Для непрерывной случайной величины:

$$\begin{aligned} \nu_k^* &= \int_{-\infty}^{\infty} x^k f(x) dx \\ \mu_k^* &= \int_{-\infty}^{\infty} (x - \bar{x})^k f(x) dx. \end{aligned}$$

Математическое ожидание – начальный момент первого порядка:

$$\mathbb{E}X = \sum_{i=1}^n x_i p_i;$$

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x) dx$$

Theorem: (Properties of expectation)

$$\begin{aligned}\mathbb{E}[\text{const}] &= \text{const}; \\ \mathbb{E}[\text{const } X] &= \text{const } \mathbb{E}X; \\ \mathbb{E}[X_1 + \dots + X_n] &= \mathbb{E}X_1 + \dots + \mathbb{E}X_n; \\ (\text{independent:}) \mathbb{E}[X_1 \cdot \dots \cdot X_n] &= \mathbb{E}X_1 \cdot \dots \cdot \mathbb{E}X_n; \\ \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] &= \mathbb{E}X_i\end{aligned}$$

Дисперсия – центральный момент 2-го порядка:

$$\text{var } X = \sum_{i=1}^n (x_i - \bar{x})^2 p_i$$

$$\text{var } X = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x) dx$$

Theorem: (Properties of the variance)

$$\begin{aligned}\text{var } X &= \mathbb{E}[X^2] - \mathbb{E}X^2 \\ \text{var}[\text{const}] &= 0 \\ \text{var}[\text{const } X] &= \text{const}^2 \text{var } X \\ (\text{independent:}) \text{var}[X_1 + \dots + X_n] &= \text{var } X_1 + \dots + \text{var } X_n; \\ \text{var}[X \pm C] &= \text{var } X \\ \text{var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] &= \frac{\text{var } X_i}{n}\end{aligned}$$

Взаимосвязи начальных и центральных моментов:

$$\begin{aligned}\mu_2 &= v_2 - v_1^2 \\ \mu_3 &= v_3 - 3v_1 v_2 + 2v_1^3\end{aligned}$$

Предельные теоремы

Theorem: (Bernoulli theorem)

Если вероятность появления события A в одном испытании равна p , число наступлений события при n незави-

симых испытаниях n , то $\forall \varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{m}{n} - P(A) \right| < \varepsilon \right) = 1.$$

Note

Для оценки вероятности по теореме Бернулли используется формула:

$$P \left(\left| \frac{m}{n} - P(A) \right| < \varepsilon \right) \geq \frac{\text{var } X}{n\varepsilon^2}.$$

Theorem

Пусть событие A может произойти в любом из n независимых испытаний с одной и той же вероятностью p и $v_n(A)$ – число осуществлений события A в n испытаниях. Тогда

$$\frac{v_n(A) - np}{\sqrt{np(1-p)}} \rightarrow N(0; 1)$$

Theorem: (Центральная предельная теорема)

Пусть X_1, \dots, X_n – последовательность независимых одинаково распределенных случайных величин, имеющих конечные математические ожидания μ и дисперсии σ^2 . Тогда

$$\frac{\sum_{i=1}^n X_i - \mu n}{\sigma \sqrt{n}} \rightarrow N(0, 1)$$

Другие предельные теоремы:

- Лемма Маркова

$$P(X \geq \tau) \leq \frac{M(X)}{\tau}, \quad \tau > 0.$$

- Неравенство Чебышева

$$P\{|X - E(X)| \leq \varepsilon\} \geq 1 - \frac{\text{var } X}{\varepsilon^2}$$

- Теорема Чебышева

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n EX_i \right| < \varepsilon \right\} = 1$$

3. Оценивание параметров в практике статистического анализа

В статистике наблюдения $\mathbf{x} = (x_1, \dots, x_n)$ рассматриваются как реализация случайного вектора $\mathbf{X}_n = (X_1, \dots, X_n)$, который имеет определенный закон распределения. Задача статистического оценивания заключается в оценке характеристик неизвестного распределения случайного вектора \mathbf{X}_n , используя его реализацию \mathbf{x} .

Definition: (Оценка)

Оценка $\hat{\theta}$ параметра θ – некоторая функция от наблюдений, принимающая значение параметра, которое на практике используют вместо неизвестного значения параметра θ .

Существует два подхода к статистическому оцениванию параметров:

- 1) θ – неслучайная величина, неизвестная;
- 2) θ – случайная величина с известной плотностью распределения $f(\theta)$ (Байесовский подход)

Theorem

Байесовский критерий минимального среднего риска оценивания:

$$\bar{r} = \int \int_{\theta \hat{\theta}} r(\theta; \hat{\theta}) f(\theta; \mathbf{x}) d\mathbf{x} d\theta$$

$$f(\theta; \mathbf{x}) = f(\theta) f(\mathbf{x}|\theta)$$

$$\bar{r} = \int \int_{\theta -\infty}^{\infty} r(\theta; \hat{\theta}) f(\theta) f(\mathbf{x}|\theta) d\mathbf{x} d\theta,$$

где

$$\int_{-\infty}^{\infty} r(\theta; \hat{\theta}) f(\theta) f(\mathbf{x}|\theta) d\mathbf{x} \rightarrow \min$$

условный средний риск.

Два вида представления оценок:

- Точечный $\hat{\theta}$
- Интервальный $P(\theta_{\min} \leq \theta \leq \theta_{\max}) = r$

Точечное оценивание параметров

Note

Свойства оценок:

- Несмещенность оценки – математическое ожидание оценки совпадает с оцениваемым параметром:

$$\mathbb{E}\hat{\theta} = \theta.$$

- Состоятельность – при увеличении числа наблюдений оценка сходится по вероятности к оцениваемому параметру;
- Эффективность – несмещенная оценка, дисперсия которой является минимальной по сравнению с другими оценками

Метод моментов: выборочные моменты используются в качестве оценок моментов генеральной совокупности.

Рассмотрим непрерывную случайную величину X , которая в результате n испытаний принимает значения x_1, \dots, x_n . Нам известен вид функции плотности $f(x; \theta)$ с неизвестным параметром θ . Для нахождения оценок ставят оптимизационную задачу максимизации функции правдоподобия вида:

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \dots f(x_n; \theta)$$

Для работы с производными рассматривают логарифм функции правдоподобия.

Пример: Exponential distribution:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Likelihood:

$$L(x; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i}$$

Log-likelihood:

$$\ln L(x_1, \dots, x_n, \lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

The result of partial derivation:

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

Now we can obtain estimated value:

$$\lambda^* = \frac{n}{\sum_{i=1}^n x_i}.$$

Примеры точечных оценок наиболее важных характеристик

- Оценка математического ожидания – выборочное среднее:

$$\mathbb{E}X \approx \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Оценка генеральной дисперсии – выборочная дисперсия:

$$\text{var } X \approx S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- Несмещенная оценка дисперсии:

$$\text{var } X \approx S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- Выборочная оценка генеральной доли:

$$p = \frac{m}{n}.$$

Для обеспечения робастности используется метод максимального квази-правдоподобия

Интервальное оценивание параметров

Пусть имеется выборка $x = (x_1, \dots, x_n)$ из генеральной совокупности $X = (X_1, \dots, X_n)$. Целью является построения доверительного интервала $(\theta_{\min}; \theta_{\max})$, такого что:

$$P(\theta_{\min} < \theta_{\max}) = \gamma,$$

где γ – надежность.

Этапы построения интервальных оценок:

- Выбор γ ;
- $\zeta = \mu(\theta)$, причем подразумевается, что $P(\zeta)$ известно;
- Построение доверительного интервала для статистики:

$$P(\zeta_{\min} < \zeta < \zeta_{\max}) = \gamma.$$

- Построение доверительного интервала для оценки:

$$P(\theta_{\min} < \theta < \theta_{\max}) = \gamma.$$

Пример: (для построения доверительного интервала для среднего) Пусть имеется выборка $x = (x_1, x_2, \dots, x_n)$, кроме того $x_i \sim N(a, \sigma^2)$ i.i.d и σ^2 известно.

- Выберем \bar{X} в качестве статистики.

$$\mathbb{E}(\bar{X}) = a \quad \text{var } \bar{X} = \frac{1}{n} \sigma^2$$

$$\bar{X} \sim N\left(a, \frac{\sigma^2}{n}\right):$$

$$\bar{X} - \frac{\sigma}{\sqrt{n}} < a < \bar{X} + \frac{\sigma}{\sqrt{n}}$$

$$P\left(-t_\gamma < \frac{\bar{X} - a}{\sigma/\sqrt{n}} < t_\gamma\right) = \gamma$$

$$P\left(\bar{X} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{X} + t_\gamma \frac{\sigma}{\sqrt{n}}\right) = \gamma.$$

4. Статистическая проверка гипотез и ее приложения

Definition: (Формальное определение статистической гипотезы)

Статистическая гипотеза – утверждение, относительно распределения и свойств случайной величины, которое можно опровергнуть или не опровергать на основе выборки.

Статистическая гипотеза позволяет проверять утверждения относительно всей генеральной совокупности, располагая лишь случайной выборкой.

Note

Задача проверки статистической гипотезы формируется в виде двух гипотез:

- H_0 : – нулевая гипотеза;
- H_1 : – альтернативная гипотеза.

Этапы проверки статистической гипотезы и ее возможные результаты

1. Необходимо сформулировать утверждение, которое необходимо проверить – гипотезу H_0 и альтернативную гипотезу H_1 ;
2. Задать уровень значимости α ;
3. Рассчитать статистику критерия φ , которая зависит от результатов наблюдений и является случайной величиной;
4. Построить критическую область, выделяя из области значений φ множество C таким образом, что выполняется условие:

$$P(\varphi \in C) = \alpha.$$

5. сделать вывод о ложности проверяемого утверждения на основе попадания или непадения значения φ , рассчитанное на основе выборки, в критическую область критерия.

Гипотезы бывают односторонними и двусторонними:

1. Односторонняя гипотеза:

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta < \theta_0 \end{aligned}$$

2. Двусторонняя гипотеза:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

Проверка гипотезы может иметь два результата:

1. Отвергнуть гипотезу H_0 и принять H_1 ;
2. Не отвергать гипотезу H_0 .

Note

Невозможность отвергнуть гипотезу H_0 не означает, что она верна.

Ошибки первого и второго рода

Возможные результаты

Пример: Проверка гипотезы о генеральном среднем.

$$H_0 : \mu = \mu_0$$

1. В случае, когда $n > 30$, σ известна:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0; 1).$$

2. В случае, когда $n > 30$, σ неизвестна:

$$Z = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0; 1),$$

где S – точечная оценка дисперсии.

3. В случае $n < 30$, σ известна:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0; 1)$$

4. $n < 30$, σ неизвестна:

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim \text{St}(n - 1)$$

Проверка гипотезы о генеральной совокупности:

$$H_0 : \sigma^2 = \sigma_0^2.$$

Используется критерий:

$$\Xi^2 = \frac{(n - 1)s^2}{\sigma^2} \sim \Xi^2(n - 1).$$

Проверка гипотезы о генеральной доле:

$$H_0 : p = p_0.$$

Используется критерий:

$$z = \frac{p - p_0}{\sqrt{\left(\frac{p_0(1 - p_0)}{n}\right)}} \sim \mathcal{N}(0; 1).$$

Definition: (Мощность теста)

Мощностью теста называется вероятность отвержения H_0 при условии, что альтернативная гипотеза верна:

$$\text{Power} = P(\text{отвергнуть } H_0 | H_1 \text{ верна}) = 1 - \beta.$$

Note

С ростом мощности теста вероятность совершить ошибку второго рода сокращается.

.....

Гипотезы бывают простыми и сложными:

- Простая гипотеза:

$$H_0 : \theta = \theta_0;$$

- Сложная гипотеза:

$$H_0 : \theta \neq \theta_0;$$

$$H_0 : \theta > \theta_0;$$

$$H_0 : \theta < \theta_0$$