ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Цель : изучение методики предварительной обработки экспериментальных данных, проверки соответствия распределения результатов измерения закону нормального распределения; изучение возможностей пакета MS Excel при решении задач статистической обработки экспериментальных данных.

Общие положения

Объект исследования – это объект любого характера, который изучается экспериментальным путем.

Эксперимент – это специальным образом спланированная и организованная процедура изучения некоторого объекта исследования, при которой на этот объект оказывают запланированные воздействия и регистрируют его реакции на эти воздействия.

Экспериментальные данные – все исходные и выходные числовые данные эксперимента, сведенные в таблицу экспериментальных данных.

Обработка экспериментальных данных – различные методы построения математической модели объекта по таблице экспериментальных данных.

Основным «**рабочим инструментом**» эксперимента и обработки экспериментальных данных является численное значение факторов
воздействия и откликов объекта исследования, т. е. **число**.

Числа при экспериментировании получают тремя способами:

- **подсчетом;**
- **измерением;**
- **методом экспертных оценок**.

Предварительная обработка результатов измерений и наблюдений необходима для того, чтобы в дальнейшем, при построении эмпирических зависимостей, **эффективно использовать статистические методы и корректно анализировать полученные результаты**.

Содержание предварительной обработки в основном состоит в **отсеивании грубых погрешностей** измерения или погрешностей, неизбежно имеющих место при переписывании цифрового материала или при вводе на электронный носитель информации.

Другим важным моментом предварительной обработки данных является **проверка соответствия** распределения результатов измерения **закону нормального распределения**.

Если эта гипотеза неприемлема, то следует определить, какому закону распределения подчиняются опытные данные, и, если это возможно, **преобразовать данное распределение к нормальному**.

Только после выполнения перечисленных выше операций можно перейти **к построению эмпирических формул**, применяя, например, **метод наименьших квадратов**.

Генеральная совокупность и выборка.

Генеральной называют совокупность всех мыслимых наблюдений,
которые могли бы быть сделаны при данном комплексе условий.

Генеральная совокупность может быть **конечной и бесконечной**.

Данное выше определение генеральной совокупности можно считать строго обоснованным только для случаев конечных генеральных совокупностей

Понятие **бесконечной генеральной** совокупности – математическая **абстракция**, как и представление о том, что измерить случайную величину можно бесконечное число раз.

Приближенно бесконечную генеральную совокупность можно истолковать как **предельный случай** конечной генеральной совокупности.

В распоряжении исследователя, **никогда нет генеральной совокупности**, он может изучать только ее часть – выборку, причем всегда ограниченного объема.

Результаты ограниченного ряда наблюдений $x_1, x_2, ..., x_n$ случайной величины можно рассматривать как в**ыборку** из данной генеральной совокупности.

Выборка – любое конечное подмножество генеральной совокупности, предназначенное для непосредственных исследований,

Объем – количество единиц в выборке.

Относительной частотой случайного события, называется отношение числа появлений этого события к общему числу произведенных испытаний.

Мера объективной возможности случайного события называется вероятностью случайных событий.

Относительные частоты можно **истолковать как выборочные значения** вероятностей случайных событий.

Характеристики **теоретических распределений** можно рассматривать как характеристики, существующие в **генеральной совокупности**, а характеристики **эмпирических распределений** – как **выборочные характеристики.**

Можно встретить и другую терминологию. Характеристики распределения вероятностей в генеральной совокупности называют п**араметрами**, а выборочные (эмпирические) значения характеристик – **оценками или статистиками.**

Параметры обозначаются буквами греческого алфавита, а оценки – соответствующими буквами латинского алфавита.

Исходными данными при оценивании, как и при проверке любых предположений (статистических гипотез), касающихся неизвестного распределения случайной величины могут быть лишь только те результаты наблюдений, которые были получены **в ходе проведения опытов** (на выборке ограниченного объема).

Причем **предварительная обработка экспериментальных** данных обычно **начинается с подсчета тех или иных функций от результатов наблюдений (статистик).**

Оценивание – определение приближенного значения неизвестного параметра генеральной совокупности по результатам наблюдений.

К оценкам предъявляются требования состоятельности, **несмещенности, эффективности.**

Состоятельная оценка – оценка, сходящаяся по вероятности к значению оцениваемого параметра при безграничном возрастании объема выборки.

Несмещенная оценка – оценка, **математическое ожидание** которой равно значению оцениваемого параметра.

Оценка параметра называется э**ффективной**, если среди прочих оценок того же параметра она обладает **наименьшей дисперсией**.

Вычисление характеристик эмпирических распределений (выборочных характеристик).

Здесь и в дальнейшем речь идет только о непрерывно распределенных случайных величинах.

Пусть имеется ограниченный ряд наблюдений x_1, x_2, ..., x_n случайной величины. Среднее значение наблюдаемого признака определяется по формуле

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где n – количество x_i значений выборки или объем выборки;
x_i - результат измерения i-й единицы.

Таким образом, x представляет собой эмпирическое или выборочное среднее. Если вычислено среднее, то легко найти отклонение каждого наблюдения от среднего

$$d_i = x_i - \bar{x}.$$

Величину $$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

называют дисперсией или вторым центральным моментом эмпирического распределения $S^2 = m_2$.

В случае одномерного эмпирического распределения произвольным моментом порядка k называется сумма k-х степеней отклонений результатов наблюдений от произвольного числа с, деленная на объем выборки п:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

где k может принимать любые значения натурального ряда чисел.

Первый центральный момент

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Второй центральный момент

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Несмещенную оценку для S^{2} (или σ^2 - дисперсия теоретического распределения) можно найти по формуле

$$\overline{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Выборочные среднеквадратические отклонения соответственно могут быть найдены по формулам

$$\overline{S} = \sqrt{\overline{S}^2}; S = \sqrt{S^2}.$$

Из других моментов чаще всего используют моменты третьего и четвертого порядка:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 \quad m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Выборочное значение коэффициента вариации V, являющееся мерой относительной изменчивости наблюдаемой случайной величины в %, определяют по формуле

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

Для нормальных и близких к нормальному распределений показатель V служит индикатором однородности выборочных наблюдений: принято считать, что при выполнимости неравенства V ≤ 33 % выборка является **количественно однородной по данному признаку**.

Выборочные значения характеристик распределения имеет смысл вычислять только в случае, если выборка является случайной.

Обычно на практике наблюдаемые значения x_1, x_2, ..., x_n величины случайные и отклонения их от среднего значения обусловлены погрешностями измерения и т. д. В свою очередь, погрешности – результат действия многих факторов.

Если имеет место такой редкий случай, когда в распоряжении исследователя имеется вся генеральная совокупность и необходимо сделать из нее выборку, то используют один из методов рандомизации (случайного выбора).

Отсев грубых погрешностей.

Можно встретить большое количество различных рекомендаций для проведения отсева грубых погрешностей наблюдения (аномальных значений).

Рассмотрим **наиболее простой метод отсева грубых погрешностей**. Если в распоряжении экспериментатора имеется выборка небольшого объема, то можно воспользоваться методом вычисления максимального относительного отклонения:

$$\tau = \frac{\left| x_{\min(\max)} - \bar{x} \right|}{\bar{S}} \leq \tau_{1-\alpha},$$

где $x_{min(max)}$ - крайний (наибольший или наименьший) элемент выборки, по которой подчитывается x, S и τ, вычисленной при доверительной вероятности $p = 1 - \alpha$.

Таким образом, для выделения аномального значения вычисляют τ, которое затем сравнивают с табличным значением $\tau_{1-\alpha}$.

Если это неравенство $\tau < \tau_{1-\alpha}$ соблюдается, то наблюдение не отсеивают, если не соблюдаются, то наблюдение исключают.

После исключения того или иного наблюдения или нескольких наблюдений характеристики эмпирического распределения должны быть пересчитаны по данным сокращенной выборки.

Квантили распределения статистики τ при уровнях значимости $\alpha = 0{,}10$, $\alpha = 0{,}05$, $\alpha = 0{,}025$, $\alpha = 0{,}01$ или доверительной вероятности $p = 1 - \alpha = 0{,}90; 0{,}95; 0{,}975; 0{,}99$ даны в справочниках.

На практике обычно используют уровень значимости $\alpha = 0{,}05$ (результат получается с 95%-й доверительной вероятностью).

Процедуру отсева нужно повторить и для следующего по абсолютной величине максимального относительного отклонения, но предварительно необходимо пересчитать x и S для выборки нового объема $(n-1)$.

Полигон и гистограмма частот распределения.

Если полученные экспериментальные данные разделить на классы, то можно построить полигон и гистограмму частот.

Разбиение на классы можно выполнить по правилу Штюргеса с округлением полученного значения до ближайшего целого числа.

Число классов определяется по формуле

Jock Sturges

$$k \approx 1 + 3{,}32 \cdot \lg(n).$$

Далее определяют размах варьирования:

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

Определяют ширину интервала $h = \frac{R}{k}.$

Затем устанавливают границы интервалов и подсчитывают число попаданий случайной величины в каждый из выбранных интервалов (абсолютные частоты Bj), для этого значения экспериментальных данных просматривают по порядку от первой до последней строчки, и при чтении каждого результата соответствующую метку (точку или черточку) заносят в тот класс, к которому относится данное наблюдение. Каждая метка соответствует одному значению из выборки.

Затем определяют относительные частоты попаданий в j-й интервал (класс) как (B_j / n) и относительные накопленные частоты как $\Sigma(B_j / n)$.

Для проверки, сумма B_j равна количеству экспериментальных данных (опытов) n.

Гистограмма и полигон распределений являются графическим отображением частот, которые, в свою очередь, представляют собой оценки плотностей вероятностей.

Кумулятивная линия – график накопленных частот, в свою очередь оценивающих функцию распределения F(x) в точке x. Многие наблюдения в природе при такой обработке дают колоколообразные полигоны распределения.

Если распределение случайной величины подчиняется определенному закону и может быть хотя бы **приближенно описано кривой** $y = ae^{-bx^2}$, то такое распределение называют нормальным.

Так как к коэффициентам a и b предъявляется только одно требование, а именно: a, b > 0, то можно говорить о семействе кривых нормального распределения. С увеличением коэффициента a кривая «вытягивается» в высоту; при увеличении коэффициента b кривая «сплющивается».

Нормальное распределение обладает и другими важными свойствами, которые позволяют считать это распределение основой математической статистики. Рассмотрим эти свойства.

1. Ордината y, которая определяет высоту кривой для каждой точки оси Ox (абсциссы), представляет собой плотность вероятности некоторого значения переменной x и определяется следующей формулой

$$y = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)}{2\sigma}}$$

$$-\infty < x < +\infty, \sigma > 0,$$

где σ – среднеквадратическое отклонение теоретического распределения;
μ – среднее значение (математическое ожидание) теоретического распределения.

Из формулы (16) следует, что нормальное распределение полностью определяется величинами μ и σ
(π = 3,141593... и e = = 2,718282... – математические постоянные).

Математическое ожидание μ определяет положение кривой распределения относительно оси Ox.

Среднеквадратическое отклонение σ определяет форму кривой.

Чем больше σ (разброс данных), тем кривая становится более пологой (ее основание более широкое).

2. Кривая нормального распределения симметрична относительно среднего значения.

3. Максимум ординаты кривой

$$y_{\max} = \frac{1}{\sqrt{2\pi\sigma^2}}$$

что при σ = 1 составляет примерно 0,4. Если x → ± ∞, то y →0 (асимптотически).

Другими словами, очень большие и очень малые значения переменной x маловероятны.

Примерно 2/3 всех наблюдений лежит в площади, отсекаемой перпендикулярами к оси Ox ($\mu \pm \sigma$).

При большом объеме выборки примерно 90 % всех наблюдений лежит между $-1{,}64\sigma$ и $+1{,}64\sigma$. Границы $-0{,}675\sigma$ и $+0{,}675\sigma$ называют вероятными отклонениями: в этом интервале находится около 50 % всех наблюдений.

Для нормального распределения среднее, мода и медиана совпадают.

Медианой выборки является среднее значение из всего упорядоченного набора значений.

Модой выборки называется значение, которое встречается большее число раз в выборке.

Для статистических методов построения эмпирических зависимостей очень важно, чтобы результаты наблюдений подчинялись нормальному закону распределения, поэтому проверка нормальности распределения – основное содержание предварительной обработки результатов наблюдений.

Проверка гипотезы нормальности распределения.

1. Среднее абсолютное отклонение.

Для небольших выборок (n < 120) можно найти простые рекомендации по проверке нормальности распределения.

Для этого необходимо вычислить **среднее абсолютное отклонение (САО)** по формуле

$$CAO = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

Для выборки, имеющей приближенно нормальный закон распределения, должно быть справедливо выражение

$$\left|\frac{CAO}{\overline{S}} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{n}}.$$

Пользуясь САО, можно также с 95%-й доверительной вероятностью оценить μ (среднее значение теоретического распределения) по : $\overline{x}$

$$\mu = \overline{x} \pm (0{,}71 \div 0{,}6) \cdot CAO.$$

Коэффициент (0,71 ÷ 0,6) зависит от величины выборки n (в данном случае n = 15÷ 20) и 1 - α = 0,95.

Коэффициенты для определения 95%-х доверительных границ для среднего значения по САО приведены в справочниках.

Размах варьирования R.

Быструю проверку гипотезы нормальности распределения для сравнительно широкого класса выборок 3 < n < 1000 можно выполнить, используя размах варьирования R.

Подсчитываем отношение $\frac{R}{\overline{S}}$

и сопоставляем с критическими верхними и нижними границами этого отношения, приведенными в справочниках

Если $k_{i'} \le \frac{R}{\overline{S}} \le k_{\hat{a}}.$

меньше нижней или больше верхней границы, то нормального распределения нет. Особенно важно, чтобы это условие соблюдалось при α = 0,10 (10%-й уровень значимости).

Показатели асимметрии и эксцесса.

Некоторое представление о близости эмпирического распределения к нормальному может дать анализ показателей асимметрии и эксцесса. Показатель асимметрии можно определяют по формуле

$$g_1 = \frac{m_3}{m_2^{3/2}}.$$

Для симметричных распределений $m_3 = 0$ и $g_1 = 0$.

Для нормального распределения $m_4 / m_2^2 = 3$.

Для удобства сравнения эмпирического распределения и нормального в качестве показателя эксцесса принимают величину

$$g_2 = \frac{m_4}{m_2^2} - 3.$$

Несмещенные оценки для показателей асимметрии G_1 и эксцесса G_2 определяют соответственно по формулам:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1;$$

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right].$$

Для проверки гипотезы нормальности распределения следует также вычислить среднеквадратические отклонения для показателей асимметрии и эксцесса соответственно:

$$S_{G_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}};$$

$$S_{G_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}.$$

Если выполняются условия $G_1 \leq 3S_{G_1}$, $G_2 \leq 5S_{G_2}$, то гипотеза нормальности исследуемого распределения может быть принята.

По критерию χ^2 (хи-квадрат)

Рассмотрим методику проверки гипотезы нормальности распределения по χ^2 критерию. Применение критерия χ^2 предполагает также использование свойств так называемого стандартного нормального распределения, которое имеет вид:

$$y = f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \cong 0{,}4e^{\frac{-z^2}{2}}.$$

Расчеты выполняют в табличной форме. Значения χ^2 определяют по формуле

$$\chi^2 = \sum_{j=1}^{n} \frac{(B_j - E_j)^2}{E_j},$$

где B_j – наблюдаемая частота;

E_j – ожидаемая по стандартному нормальному распределению частота.

$$E_j = f(z_j) \cdot k',$$

$$k' = \frac{nn_{\bar{s}}}{\bar{S}},$$

где f(zj) – уравнение кривой стандартного нормального распределения:

$$f(z_j) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_j^2}{2}};$$

z_j – степень функции кривой нормального распределения:

$$z_j = \frac{\left| x_i - \bar{x}_{ожид} \right|}{\bar{S}_{ожид}}$$

$\bar{x}_{ожид}$ ожидаемое среднее значение наблюдаемого признака

$$\bar{x}_{ожид} = \frac{1}{n}\sum_{j=1}^{n_{кл}} B_j x_j;$$

$\bar{S}_{ожид}$ ожидаемая дисперсия:

$$\bar{S}_{ожид}\sqrt{\frac{\sum_{j=1}^{n_{кл}} B_j x_j^2 - \frac{(\sum_{j=1}^{n_{кл}} B_j x_j)^2}{n}}{n-1}};$$

$n_{кл}$ - число классов (интервалов).

Полученное значение χ^2 сравнивают с табличным или критическим значением $\chi^2_{\text{пк}\alpha}$.

Число степеней свободы v определяют по формуле

$$\nu = n_{кл} - 1 - \kappa,$$

где r – число параметров распределения (для нормального распределения r = 2, так как оцениваются два параметра $\overline{x}, \overline{S}$

Гипотеза нормальности распределения принимается в случае выполнения условия $\chi^2 \leq \chi^2_{\text{пк}\alpha}$.

Методика проверки нормальности распределения по показателям асимметрии и эксцесса очень хорошо иллюстрирует использование моментов, а также удобна при использовании компьютерных технологий.

Для практического применения (особенно при расчетах с использованием компьютерных технологий) рекомендуются в основном две методики: по размаху варьирования и по χ^2-критерию, причем первая служит для быстрой «прикидочной» проверки, а вторая – для основательной проверки нормальности распределения.

В настоящее время обработку экспериментальных данных существенно облегчают современные компьютерные технологии, современное программное обеспечение. Например, электронные таблицы MS Excel.