Лекция 5. ОБРАБОТКА ЦЕНЗУРИРОВАННЫХ ВЫБОРОК

Учебные вопросы:

5.1. Цензурирование экспериментальных данных

5.2. Непараметрические методы оценивания

5.3. Параметрические методы оценивания

Литература

1.Ходасевич Г.Б., Пантюхин О.И., Ногин С.Б. Планирование эксперимента и обработка экспериментальных данных на ЭВМ. Ч. 1. Обработка экспериментальных данных на ЭВМ: учебное пособие; СПбГУТ.- СПб., 2014. – 88с.

2.Ходасевич Г.Б., Пантюхин О.И., Ногин С.Б. Планирование эксперимента и обработка экспериментальных данных на ЭВМ. Ч. 2. Планирование эксперимента: учебное пособие; СПбГУТ.- СПб., 2014. – 88с.

Основные понятия выборочного метода

- Вариационный ряд - последовательность вариант, записанных в возрастающем порядке.
- Статистическое распределение выборки – это перечень вариант x_i вариационного ряда и соответствующих им **частот** n_i ($\sum n_i = n$) или **относительных частот** $w_i = n_i / n$ ($\sum w_i = 1$),
- n – объём выборки.

Учебный вопрос:

5.1. Цензурирование экспериментальных данных

Некоторые статистические сведения могут быть представлены в виде цензурированных выборок [1,глава8]. Такие выборки характерны для ЭД по надежности объектов, достоверности обработки информации, защите информации от НСД.

Цензурированием называется событие, приводящее к прекращению наблюдений за изделием до наступления системного события (например, отказа) либо к свершению события в неизвестный момент времени в пределах некоторого интервала.

Цензурированной выборкой называется выборка, элементами которой являются полные наработки и наработки до цензурирования (неполные наработки).

Полной наработкой является наработка изделия от начала некоторого этапа его эксплуатации до системного события, например, наработка до отказа.

Неполная наработка характеризует наработку изделия: от начала эксплуатации до фиксированного момента времени, но до наступления системного события; от некоторого произвольного момента, не связанного с системным событием, до системного события или до конкретного момента времени.

Интервал, в котором произошло или произойдет системное событие, причем точное значение наработки до системного события неизвестно, называется **интервалом неопределенности.**
Этот интервал может быть ограниченным:

слева (цензурирование слева). Наблюдения за объектами прекращаются в какой-то момент времени. К моменту окончания наблюдений часть объектов отказала. Другая часть продолжает работать, причем неизвестно, как долго эти объекты проработают без отказа;

справа (цензурирование справа). К началу наблюдений объекты уже проработали некоторое неизвестное время без отказа. Отказавшие к моменту начала наблюдений объекты во внимание не принимаются;

слева и справа (цензурирование интервалом). Цензурирование интервалом является наиболее общим случаем цензурирования.

Применительно к задачам оценки надежности по результатам наблюдений в процессе эксплуатации цензурирование обычно связано с ограниченностью интервалов наблюдения.

Существует несколько типовых вариантов (планов) наблюдений. Краткое обозначение плана включает три элемента.
Первый элемент характеризует количество объектов N, предназначенных для наблюдений.
Второй – действия с отказавшими объектами:
U – отсутствие замены или восстановления отказавших объектов;
R – замена отказавших объектов;
M – восстановление отказавших объектов.
Третий элемент (одна или две буквы) определяет признак окончания наблюдений:
T – наблюдения заканчиваются по истечении фиксированного интервала времени;
r – наблюдения заканчиваются по достижении фиксированного количества реализаций (отказов, восстановлений);
z – наблюдения заканчиваются при наработке каждого объекта, равной ti.

План [NUT] указывает, что под наблюдением находится N объектов, отказавш объекты не заменяются и не восстанавливаются U, наблюдения заканчиваются по истечении заданного интервала времени T (однократно цензурированная выборка).
В отличие от [NUT] план [NUz] означает, что наблюдение за конкретным объектом заканчивается при возникновении его отказа или при достижении конкретного значения наработки (многократно цензурированная выборка).
План [NUT] соответствует цензурированию типа 1, при этом заранее фиксируется время проведения наблюдений, число событий представляет собой случайную величину. При цензурировании по плану [NUr] или при цензурировании типа 2 заранее задается число событий (доля событий), после наступления которых наблюдения прекращаются, время наблюдения заранее не фиксируется, т.е. оно случайно.

Выбор конкретного плана зависит от целей исследования. Далее рассматриваются планы типа […U…]. Обработка результатов по плану типа […R…] сводится к предыдущему типу путем переноса начала наблюдений каждого нового объекта к некоторому условному началу испытаний всех объектов. Планы типа […M…] можно рассматривать как планы типа […U…], если каждую наработку между отказами трактовать как наработку некоторого невосстанавливаемого объекта (полное восстановление ресурса объекта после отказа). Очевидно, что план типа [NUN] соответствует полной выборке.

Оценка надежности проводится с начала эксплуатации на некоторый (текущий) момент или за определенный интервал времени. **В первом случае** имеет место цензурирование слева по текущему моменту времени. Для невосстанавливаемых объектов часть из них к этому моменту времени может отказать, а другая часть продолжает работать, что соответствует плану наблюдения [NUT]. Значения наработок исправных объектов неизвестны, но очевидно, что они превышают интервал наблюдения. **Во втором случае** оценка надежности связана с цензурированием выборки справа (продолжительность работы средств точно неизвестна) и слева, часть средств может отказать к моменту начала наблюдения и не учитывается на текущем интервале, другая часть может отказать на текущем интервале, а третья продолжит работу и по завершении периода наблюдения. В рассмотренных вариантах цензурирование осуществляется по фиксированным моментам времени, и число наблюдений в выборке является случайным.
В некоторых случаях цензурирование осуществляется по конкретным событиям, например, при определенном числе отказов объектов, что характерно при проведении испытаний однотипных изделий в интересах определения показателей надежности, планы типа [NUr]. В планах наблюдения [NU(r,T)] прекращение наблюдений происходит после отказа r объектов или по достижении момента времени T в зависимости от того, какое из событий происходит ранее. В таких случаях объем выборки не является случайным, случайна продолжительность наблюдений.

Итак, формируемые в ходе эксплуатации выборки по надежности могут иметь: однократное цензурирование слева (например, период наблюдения от начала эксплуатации до текущего момента времени); цензурирование интервалом (период наблюдения определяется календарными сроками); многократное цензурирование слева; многократное цензурирование интервалом. Левая и правая границы цензурирования при этом определяются моментами времени или случайными событиями, например, моментом отказа какого-либо средства.

Для цензурированных выборок необходимо применять свои методы оценки показателей, проверки статистических гипотез. Теория обработки цензурированных выборок сложнее традиционных методов мате¬матической статистики и далека от своего завершения.

Следует отметить, что практически все выборки результатов наблюдения за функционированием объектов так или иначе цензурированы. Однако цензурирование следует учитывать только в тех случаях, когда интервал наблюдения соизмерим с наработкой на системное событие и количество неполных наблюдений составляет значительный процент в общем объеме.

5.2. Непараметрические методы оценивания

Непараметрические методы применяют тогда, когда закон распределения исследуемого показателя неизвестен и нет необходимости его аналитического описания. Эти методы проще в реализации, чем параметрические, но они не позволяют осуществлять прогноз значений показателей надежности. **К непараметрическим относят методы** последовательного перехода к новой системе координат, построения "множительной" оценки, ядерных оценок, Будстрепа и другие. С прикладной точки зрения методы различаются сложностью реализации и качеством получаемых оценок. Однако характеристики качества получаемых оценок исследованы не для всех методов, особенно слабо проработаны эти вопросы применительно к малым объемам выборок.

Построение эмпирической функции распределения наработки до отказа по формуле **$F_N(t)=i/N$ при $t>0$** (где N – объем выборки; i – количество наработок до отказа, попавших в интервал [0, t], i=1, 2, ..., N) применимо для планов [NUr], [NUT] и [NUz] в области полных наработок, но недопустимо в целом ко всей цензурированной выборке (так как этот подход предполагает использование информации по всей выборке). Если исключить все неполные наработки (наработки до цензурирования), то будут иметь место значительные ошибки в определении оценки $F_N(t)$. Наличие цензурирования приводит к неопределенности для $F_N(t)$ в области цензурирования, которая увеличивается с ростом числа неполных наработок.

Постановка задачи определения показателей надежности по цензурированным выборкам формулируется следующим образом.

Имеются выборочные значения наработки до отказа ***t1 , t2 , …, tr***
и до цензурирования **t1, t2, ..., tk.**
Количество наработок до отказа r и до цензурирования k, объем выборки **N=r+k.**
Необходимо определить: эмпирическую функцию распределения наработок до отказа, оценку вероятности безотказной работы, среднее значение (оценку математического ожидания) наработки до отказа.
Допущения: результаты получены с использованием одного из планов типа [NUr], [NUT] или [NUz].
Решение задачи включает выполнение следующих этапов:
предварительная обработка ЭД;
построение эмпирической функции распределения $F_N(t)$;
определение оценки вероятности безотказной работы p*(t) и средней наработки до отказа To.
Предварительная обработка ЭД предусматривает построение общего вариационного ряда, для этого наработки на отказ и на цензурирование упорядочивают в порядке неубывания. Если отдельные наработки до отказа равны наработкам до цензурирования, то в вариационном ряду первыми ставятся наработки до отказа.

К числу основных методов построения эмпирической функции распределения относятся методы: последовательного перехода к новой системе координат и множительной оценки.

Рассмотрим метод множительной оценки.

Пример 5.1. Проведено испытание десяти объектов по плану [NUz]. Наработки шести объектов до отказа составили **1922, 2576, 2314, 1873, 2135, 2018** часов. К моменту оценки четыре объекта безотказно проработали **2107, 3936, 2010, 2397** часов. Необходимо построить эмпирическую функцию распределения наработки до отказа.

Решение. Построим общий вариационный ряд (таблица) (звездочками помечены наработки на цензурирование).

n	1	2	3	4	5	6	7	8	9	10
tn	1873	1922	2010 *	2018	2107 *	2135	2314	2397 *	2576	3936 *
tr	t1	t2		t4		t6	t7		t9	
tk			t3		t5			t8		t10
i	1			2		3			4	
ri	2			1		2			1	
ki	1			1		1			1	

Пример 5.2. Используя метод множительных оценок для условий примера 5.1, **построить эмпирическую функцию распределения наработки до отказа, оценить среднюю наработку до отказа и вероятность безотказной работы за наработку 2000 часов.**

Решение. Воспользуемся методом множительной оценки вероятностей безотказной работы и эмпирической функции распределения наработки до отказа:

p*(t <1873)=1;	F10(t <1873)= 1- p*(t <1873)=0;
В момент t1 отказал 1й объект из 10ти	
p*(t1) = p*(1873) = **1– 1/10=0,9**;	F10(t1) = 1– p*(1873) = 1-0,9= 0,1;
В момент t2 отказал 2й объект из 9ти оставшихся	
p*(t2) = p*(1922) = **0,9(1–1/9)=0,8;**	F10(t2) = 1-0,8= 0,2;
В момент t3 НЕ отказал 3й объект, переходим к t4, осталось 7 объектов. Так же решаем для следующих моментов времени	
p*(t4) = p*(2018) = **0,8(1–1/7)=0,686;**	F10(t4) = 1-0,686= 0,314;
p*(t6) = p*(2135) = **0,686(1 – 1/5)=0,549;**	F10(t6) = 1-0,549= 0,451;
p*(t7) = p*(2314) = **0,549(1 – 1/4)=0,411;**	F10(t7) = 1-0,411= 0,589;
p*(t9) = p*(2576) = **0,397(1 – 1/2)=0,206;**	F10(t9) = 1-0,206= 0,794.

Так как последний из объектов не отказал, то p*(t) не равна 0, а $F10(t)$ не равна 1.

Оценка средней наработки до отказа находится по формуле:

$$T_o = \mu_1(t) = \sum_{i=1}^{r} t_i [F_N(t_i) - F_N(t_{i-1})] + [1 - F_N(t_r)]z,$$ где z = max (tr, tk); $t0$=0.

Тогда:

To = m1(t) =1873·(0,1 – 0)+1922·(0,2 – 0,1)+2018·(0,314 – 0,2)+2135·(0,451 – 0,314)+ +2314·(0,589 – 0,451)+ 2576·(0,794 – 0,589)+(1 – 0,794)·3936 = 2559,9 ч.

Простое вычисление среднего значения по всем наработкам дает величину, равную 2328,8 ч, что меньше To.

Оценка вероятности безотказной работы за наработку 2000 ч:

Находим множитель d на отрезке между соседними наработками на отказ:

d = (2000 – 1922)/(2018 – 1922)=0,813;

Находим вероятность БОР в этот момент времени как сумму произведений значения вероятности БОР на d (в момент после указанного) и на 1- d (в момент до указанного) :

p*(2000)=(0,813)·(0,686)+(1– 0,813)·0,8=0,707.

Рассмотренные подходы к построению эмпирической функции распределения просты в реализации, не требуют большого объема данных и сложных вычислений. Они позволяют получить (за исключением ряда ситуаций, в которых Nu,i=0) несмещенные, состоятельные, асимптотически нормальные оценки значений функции распределения наработки изделия до отказа. Существенным недостатком оценок является невозможность их применения в интересах прогнозирования надежности исследуемых изделий. Преодоление данного недостатка воз¬можно на основе параметрического оценивания показателей, которое позволяет получать оценки с более высокой точностью, чем непараметрические методы.

5.3. Параметрические методы оценивания

Применение параметрических методов предполагает априор¬ное знание теоретического закона распределения исследуемой ве¬личины или его определение по эмпирическим данным, что обусловливает необходимость проверки согласованности ЭД и выбранного теоретического закона. Параметрическая оценка по цензурированным выборкам основывается на традиционных методах математической статистики (максимального правдоподобия, моментов, квантилей), методах линейных оценок и ряде других.

Обработка многократно цензурированных выборок ме¬тодом максимального правдоподобия допускается при следующих условиях:

$6 < N < 10, \quad r/N >= 0{,}5;$

$10 <= N < 20, \quad r/N >= 0{,}3;$

$20 <= N < 50, \quad r/N >= 0{,}2;$

$50 <= N < 100, \quad r/N >= 0{,}1.$

Когда эти ограничения не выполняются, можно вычислять только нижнюю доверительную границу параметров распределения

Оценки, получаемые по методу максимального правдоподобия, при относительно нежестких ограничениях асимптотически эффективны, не смещены и распределены асимптотически нормально. Если непрерывная переменная с функцией плотности f(x, t) цензурирована в точках a и b (a<b), то функция плотности распределения при цензурировании определяется как

$$f(x,T)/\left[\int_a^b f(x,T)\,dx\right]$$

Функция правдоподобия при N наблюдениях:

$$L_1(x,T)=\prod_{i=1}^{N} f(x_i,T)/\left[\int_a^b f(x,T)\,dx\right]^N$$

Решение уравнения правдоподобия при различных схемах цензурирования является достаточно сложной задачей. В явном виде такие решения можно получить только для однопараметрических законов распределения. Известны уравнения для нахождения параметров типовых законов распределения показателей надежности по цензурированным слева выборкам.

Экспоненциальное распределение.

Точечные оценки параметра распределения λ при различных планах наблюдения:

$$\frac{r\,N}{(N-1)\left(\sum_{i=1}^{r} t_i + \sum_{j=1}^{k} \tau_j\right)}, \quad r > 1, \quad \text{для } (NUz);$$

$$\frac{r}{\sum_{i=1}^{r} t_i + (N-r)\,T}, \quad r > 0, \quad \text{для } (NUT),$$

$$\frac{r-1}{\sum_{i=1}^{r} t_i + (N-r)t_r}, \quad r > 1, \quad \text{для } (NUr).$$

Нормальное распределение . Оценки параметров распределения m и s для планов наблюдения [NUr], [NUT] и [NUz] находятся из системы уравнений:

$$\sum_{j=1}^{k}(t_j-\mu)/\sigma+\sum_{j=1}^{k}f\left(\frac{\mu-\tau_j}{\sigma}\right)/\Phi\left(\frac{\mu-\tau_j}{\sigma}\right)=0,$$

$$r-\sum_{j=1}^{k}\frac{(t_j-\mu)^2}{\sigma^2}+\sum_{j=1}^{k}\left(\frac{\mu-\tau_j}{\sigma}\right)f\left(\frac{\mu-\tau_j}{\sigma}\right)/\Phi\left(\frac{\mu-\tau_j}{\sigma}\right)=0,$$

где Φ(x) – функция нормального распределения, f(x) – функция плотности нормального распределения.

Данная система уравнений допускает только численное решение. При таком решении уравнений в качестве начальных приближений неизвестных параметров обычно берут оценки математического ожидания и среднеквадратического отклонения, вычисленные по объединенной выборке.
Численное решение уравнения относительно неизвестных можно произвести с помощью функций root и Find пакета MathCAD.

Распределение Вейбулла.

Оценки параметров d и b для плана [NUz] вычисляются на основе системы уравнений:

$$(r/\beta + \sum_{i=1}^{r} \ln t_i)(\sum_{i=1}^{r} t_i^{\beta} + \sum_{j=1}^{k} \tau_j^{\beta}) - (\sum_{i=1}^{r} t_i^{\beta} \ln t_i + \sum_{j=1}^{k} \tau_j^{\beta} \ln \tau_j) r = 0,$$

$$\delta = \left[(\sum_{i=1}^{r} t_i^{\beta} + \sum_{j=1}^{k} \tau_j^{\beta}) / r \right]^{1/\beta}.$$

Системы уравнений не имеют аналитического решения и требуют применения численных методов: вначале находится корень первого уравнения (оценка параметра b), затем прямой подстановкой значение оценки параметра d. Для двухпараметрического распределения Вейбулла большие (b>4) или малые (b<0,5) значения параметра свидетельствуют о том, что ЭД не подчиняются этому закону или отношение r/N мало. В таких случаях следует применить непараметрические методы оценивания или перейти к трехпараметрическому закону распределения Вейбулла.

Трудности применения метода максимального правдопо¬добия обусловливают разработку других методов. Метод моментов обычно приводит к простым вычислительным процедурам, позволяет получить асимптотически эффективные, несмещенные и нормально распределенные оценки, но требует учета типа цензурирования и применим при относительно большом объеме выборки (не менее 30). Использование метода квантилей для оценок параметров законов распределений менее критично к типу цензурирования. Высокая точность оценок достигается оптимальным подбором квантилей, хотя такой подбор не всегда удается осуществить.

Метод линейных оценок применяют при небольшом объеме выборки, он обеспечивает высокую эффективность, состоятельность и несмещенность оценок параметров распределения. Этот метод основан на нахождении линейной функции от порядковых статистик (упорядоченных элементов выборки), которая была бы несмещенной оценкой искомого параметра. Применение связано с необходимос¬тью использования специальных видов распределений, что вызывает определенные неудобства и затрудняет автоматизацию расчетов.

Заключение

В целом следует отметить, что нет единого метода, лучшего для всех ситуаций оценивания. В каждом конкретном случае необходимо выбирать метод, наиболее подходящий по своим возможностям для заданного типа выборки и требований к оценкам показателей надежности, наличного ресурса по обработке данных, целесообразной степени автоматизации. Рациональным является комбинированное использование методов, например нахождение приближенных оценок на основе квантилей с последующим их уточнением по формулам, полученным методом максимального правдоподобия.

Задание на самостоятельную работу

1. Повторить материал лекции.
2. Изучить задание на лабораторную работу №3 (файл ЛР3_бакалаврам теория и задание ЛабРаб по ОЭД.doc).

3. Выполнить задание на занятие по вариантам:

В результате испытаний однотипных невосстанавливаемых изделий на безотказность функционирования получены значения наработок до отказа. К моменту завершения испытаний часть изделий отказала, а другая – сохранила работоспособность. Необходимо определить показатели безотказности изделий на основе непараметрических и параметрических методов, а именно оценить:

среднюю наработку до отказа To;

вероятность безотказной работы для значений наработок t, равных 0,5To, To, 1,5 To и 2To.

4. Подготовить ответы на контрольные вопросы