Статистическое оценивание числовых характеристик случайной величины и параметров законов распределений

Параметрическое и непараметрическое оценивание

Параметрические критерии – это критерии, включающие в формулу расчета параметры распределения (среднее и дисперсию).

Непараметрические критерии – это критерии, не включающие в формулу расчета параметры распределения и основанные на оперировании частотами или рангами

* Параметрическое и непараметрическое оценивание

ПАРАМЕТРИЧЕСКИЕ	НЕПАРАМЕТРИЧЕСКИЕ
ДОСТОИНСТВА •Высокая мощность критериев (способны с большей достоверностью отвергать нулевую гипотезу) •Широкое применение	**ДОСТОИНСТВА** •не требуют никакого предположения о виде распределения •очень просты в использовании. •могут использовать качественную информацию (а не только количественную) и упорядоченные по рангам данные.
НЕДОСТАТКИ И ОГРАНИЧЕНИЯ Неприменимы в случае, если -распределение неизвестно или отличается от нормального -объемы выборок малы -данные являются неколичественными	**НЕДОСТАТКИ И ОГРАНИЧЕНИЯ** малая мощность и консервативность, т.е. с их помощью труднее отклонить нуль-гипотезу, для этого требуется большая выборка или большее число противоречащих альтернативных данных, чем при использовании параметрических критериев

* Параметрическое оценивание

Оценки числовых характеристик законов распределения вероятности случайных чисел или величин, изображаемые точкой на числовой оси, называются ***точечными***. В отличие от самих числовых характеристик оценки являются случайными, причем их значения зависят от объема экспериментальных данных, а законы распределения вероятности – от законов распределения вероятности самих случайных чисел или значений измеряемых величин.

Интервальной оценкой называют оценку, которая определяется двумя числами a_1 и a_2 – концами интервалов, накрывающего оцениваемый параметр a с заданной доверительной вероятностью.

Статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин.

Параметрическое оценивание

Несмещенной называют статистическую оценку a^*, мат.ожидание которой равно оцениваемому параметру a при любом объеме выборки, т.е. $M[a^*] = a$.

Эффективной называют статистическую оценку, которая (при заданном объеме выборки n) имеет наименьшую возможную дисперсию.

Состоятельной называют статистическую оценку, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру.

Оценка характеристик положения

Генеральной средней $\bar{x}_Г$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения x_1, x_2,…, x_N признака генеральной совокупности объема N различны, то

$$\bar{x}_Г = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Если значения признака x_1, x_2,…, x_k имеют соответственно частоты N_1, N_2,…, N_k, причем N_1, N_2,…, $N_k = N$, то

$$\bar{x}_Г = \frac{x_1 N_1 + x_2 N_2 + \ldots + x_k N_k}{N}$$

* Оценка характеристик положения

Выборочной средней $\bar{x}_B$ называют среднее арифметическое значений признака выборочной совокупности.

Если все значения x_1, x_2,…, x_n признака выборки объема n различны, то

$$\bar{x}_B = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Если значения признака x_1, x_2,…, x_k имеют соответственно частоты n_1, n_2,…, n_k, причем n_1, n_2,…, $n_k = n$, то

$$\bar{x}_B = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n}$$

т.е. выборочная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

* Оценка характеристик положения

Выборочная средняя – несмещенная оценка.

Свойство устойчивости выборочных средних: если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности найти выборочные средние, то они будут приближенно равны между собой.

Близость выборочных средних к генеральным не зависит от отношения выборки к объему генеральной совокупности, а зависит от объема: чем больше объем выборки, тем меньше выборочная средняя отличается от генеральной.

Групповой средней называют среднее арифметическое значений признака, принадлежащей группе.

Общей средней $\bar{x}$ называют среднее арифметическое значений признака, принадлежащих всей совокупности.

Общая средняя равна средней арифметической групповых средних, взвешенной по объемам групп.

* Оценка характеристик положения

Пример.

Группа	1-я группа		2-я группа	
Значение признака	1	6	1	5
Частота	10	15	20	30
Объем	10+15 = 25		20+30 = 50	

Групповые средние:

$$\bar{x}_1 = \frac{10 \cdot 1 + 15 \cdot 6}{25} = 4; \quad \bar{x}_2 = \frac{20 \cdot 1 + 30 \cdot 5}{50} = 3,4.$$

Общая средняя

$$\bar{x} = \frac{25 \cdot 4 + 50 \cdot 3,4}{25 + 50} = 3,6$$

Оценка характеристик положения

Отклонением называют разность между значением признака и общей средней $\bar{x}_i - \bar{x}$

Теорема. Сумма произведений отклонений на соответствующие частоты равна нулю:

$$\sum n_i \left(\bar{x}_i - \bar{x} \right) = 0$$

Следствие. Среднее значение отклонения равно нулю.

* Оценка характеристик положения

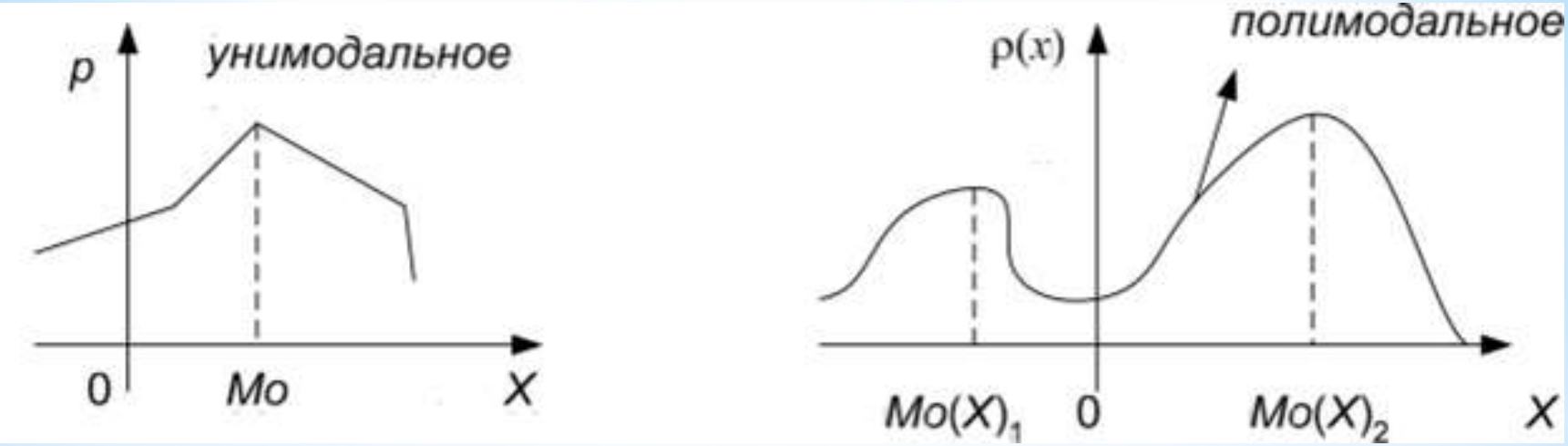
Модой Mo(X) с.в. X называется ее наиболее вероятное *значение* (для которого вероятность p_t или плотность вероятности *p(x)* достигает максимума).
Например, для ряда

Варианта	**1**	**4**	**7**	**9**	**12**
частота	5	1	20	6	10

мода равна 7.

Если вероятность или плотность вероятности достигает максимума в одной точке, распределение называется *унимодальным*, если же максимум достигается в нескольких точках, распределение называется *полимодальным*

* Оценка характеристик положения



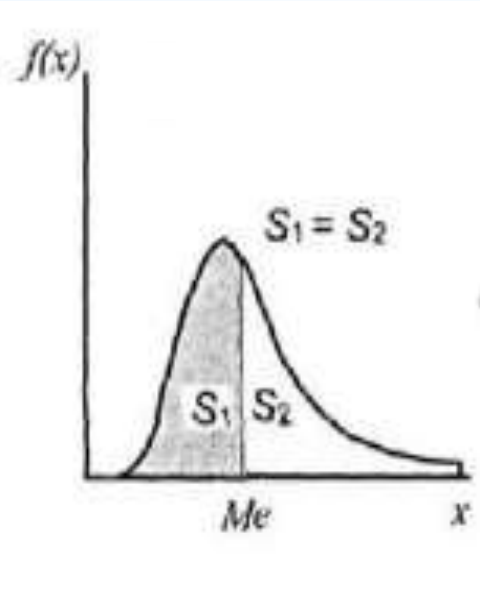
Медианой Me(X) с.в. X называется такое ее значение, при котором вероятность того, что СВ X < Me, одинаково вероятна тому, что СВ X > Me, и будет равна 0,5.

$$P\left(-\infty \le x \le Me\right) = P\left(Me \le x \le \infty\right) = 0,5.$$

$$\int_{-\infty}^{Me} \rho(x)dx = \int_{Me}^{\infty} \rho(x)dx = 0,5.$$

* Оценка характеристик положения

Для непрерывной случайной величины медиана это абсцисса точки в которой площадь под кривой распределяется пополам.



Медиана – это уровень показателя, который делит набор данных на две равные половины. Значения в одной половине меньше, а в другой больше медианы.

Оценка характеристик положения

Для дискретной с.в. значение медианы зависит от того четное или нечетное значение случайной величины.

Если количество значений нечетно, то медиана будет соответствовать центральному значению ряда, номер которого можно определить по формуле:

$$№_{Me} = \frac{N+1}{2}$$

Или если N=2k+1, то $Me=x_{к+1}$ (среднее по порядку значение).

Если значение случайных величин четное, т.е. N=2k, то

$$Me = \frac{x_k + x_{k+1}}{2}$$

* Оценка характеристик положения

Например, для ряда 4 8 10 12 18 медиана равна 10.
Для ряда 4 8 10 12 18 23
Me = (10+12)/2 = 11.
Внимание! Если число случаев четное и в центре ряда находятся два разных числа, то медианой будет среднее между ними (даже если такого значения нет в самом ряду исследуемых случаев).
Медиана обладает высокой *робастностью*, то есть нечувствительностью к неоднородностям и ошибкам выборки.
Сумма разностей между членами ряда выборки и медианой меньше, чем сумма этих разностей с любой другой величиной. В том числе с арифметическим средним.

* Оценка характеристик положения

Медиана интервальных рядов определяется по формуле

$$M_e = X_{Me} + i_{Me} \cdot \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}}$$

где X_{Me} – нижняя граница медианного интервала (того интервала, накопленная частота которого превышает полусумму всех частот);

i_{Me} – величина медианного интервала;

f – частота (сколько раз в ряду встречается то или иное значение);

S_{Me-1} – сумма частот интервалов предшествующих медианному интервалу;

f_{Me} – число значений в медианном интервале (его частота).

* Оценка характеристик положения

Пример.

Цена, руб.	Количество, шт.
100-200	20
200-300	50
300-400	**60**
400-500	40
500-600	30

Цена, руб.	Количество, шт.	Накопленная частота, шт.	Накопленная доля
100-200	20	20	10,0%
200-300	50	70	35,0%
300-400	60	130	***65,0%***
400-500	40	170	85,0%
500-600	30	200	100,0%
Итого	**200**		

$$Me = x_{Me} + i_{Me}\frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}} = 300 + 100 * \frac{\frac{200}{2} - 70}{60} = 350 \text{ руб.}$$

* Оценка характеристик положения

Квартили - это значения признака в ранжированном ряду, выбранные таким образом, что 25% единиц совокупности будут меньше величины Q_1, 25% единиц будут заключены между Q_1 и Q_2, 25% - между Q_2 и Q_3, остальные 25% превосходят Q_3. Квартили определяются по формулам, аналогичным формуле для расчета медианы. Для интервального ряда:

$$Q_1 = x_{Q_1} + i_{Q_1} \cdot \frac{0{,}25 \cdot \sum f_i - S_{Q_1 - 1}}{f_{Q_1}}$$

$$Q_2 = Me$$

$$Q_3 = x_{Q_3} + i_{Q_3} \cdot \frac{0{,}75 \cdot \sum f_i - S_{Q_3 - 1}}{f_{Q_3}}$$

Децилем называется структурная переменная, делящая распределение на 10 равных частей по числу единиц в совокупности. Децилей 9, а децильных групп 10.

* Оценка характеристик рассеяния

Размах вариации определяется как разность между максимальным и минимальным значением признака в изучаемой совокупности:

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

Характеристика проста в вычислении, но малоинформативна.

Дисперсия - средний квадрат отклонения значения с.в. от ее среднего значения. Дисперсия есть характеристика рассеяния, разбросанности значений величины около ее среднего значения.

Генеральной дисперсией $D_г$ называют среднее арифметическое квадратов отклонений признака генеральной совокупности от их среднего значения $\overline{x}_Г$.

* Оценка характеристик рассеяния

Если все значения x_1, x_2,…, x_N признака генеральной совокупности объема N различны, то

$$D_г = \left(\sum_{i=1}^{N} (x_i - \overline{x}_г)^2 \right) / N.$$

Если значения признака x_1, x_2,…, x_k имеют соответственно частоты N_1, N_2,…, N_k, причем N_1, N_2,…, $N_k = N$, то

$$D_г = \left(\sum_{i=1}^{k} N_i (x_i - \overline{x}_г)^2 \right) / N,$$

Генеральным средним квадратическим отклонением (стандартом) называют квадратный корень из генеральной дисперсии:

$$\sigma_г = \sqrt{D_г}.$$

* Оценка характеристик рассеяния

Пример.

Варианта	2	4	5	6
частота	8	9	10	3

$$\overline{x}_г = \frac{8\cdot 2 + 9\cdot 4 + 10\cdot 5 + 3\cdot 6}{8+9+10+3} = \frac{120}{30} = 4.$$

$$D_г = \frac{8\cdot(2-4)^2 + 9\cdot(4-4)^2 + 10\cdot(5-4)^2 + 3\cdot(6-4)^2}{30} = 54/30 = 1{,}8.$$

$\sigma_г = 1{,}34.$

Теорема. Дисперсия равна среднему квадратов значений признака минус квадрат общей средней:

$$D = \overline{x^2} - [\overline{x}]^2.$$

* Оценка характеристик рассеяния

Пример.

Варианта	1	2	3	4
частота	20	15	10	5

$$\bar{x} = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{20 + 15 + 10 + 5} = \frac{100}{50} = 2.$$

$$\bar{x^2} = \frac{20 \cdot 1^2 + 15 \cdot 2^2 + 10 \cdot 3^2 + 5 \cdot 4^2}{50} = 5.$$

$$D = \bar{x^2} - [\bar{x}]^2 = 5 - 2^2 = 1.$$

* Оценка характеристик рассеяния

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\overline{x}_B$.

Если все значения x_1, x_2,…, x_N признака генеральной совокупности объема N различны, то

$$D_в = \left(\sum_{i=1}^{n} (x_i - \overline{x}_в)^2 \right) / n.$$

Если значения признака x_1, x_2,…, x_k имеют соответственно частоты N_1, N_2,…, N_k, причем N_1, N_2,…, $N_k = N$, то

$$D_в = \left(\sum_{i=1}^{k} n_i (x_i - \overline{x}_в)^2 \right) / n,$$

Выборочным средним квадратическим отклонением (стандартом) (С.К.О.) называют квадратный корень из выборочной дисперсии:

$$\sigma_в = \sqrt{D_в}.$$

* Оценка характеристик рассеяния

Замечания.

Эти оценки не являются несмещенными. Доказано, что сумма квадратов отклонений значений признака для выборочного среднего арифметического меньше, чем сумма квадратов отклонений от любой другой величины, в том числе от истинного среднего (математического ожидания). Поэтому результат будет содержать систематическую ошибку, и оценочное значение дисперсии окажется заниженным. Необходим поправочный коэффициент.

$$D_B = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}_в\right)}{n-1}; \quad D_B = \frac{\sum_{i=1}^{k} n_i\left(x_i - \overline{x}_в\right)}{n-1}$$

С.К.О. имеет те же единицы измерения, что и результаты измерения, оно характеризует степень отклонения признака от ср. арифм., т.е. показывает, как расположена основная часть вариант относительно ср. арифм.

* Оценка характеристик рассеяния

Коэффициент вариации равен отношению выборочного С.К.О. к выборочной средней, выраженному в процентах:

$$V = \frac{\sigma_B}{\overline{x}_в} \cdot 100\%$$

Служит для сравнения величин рассеяния двух вариационных рядов. Коэффициент вариации – безразмерная величина.

Аналогично определяется коэффициент вариации для генеральной совокупности.

* Оценка характеристик формы

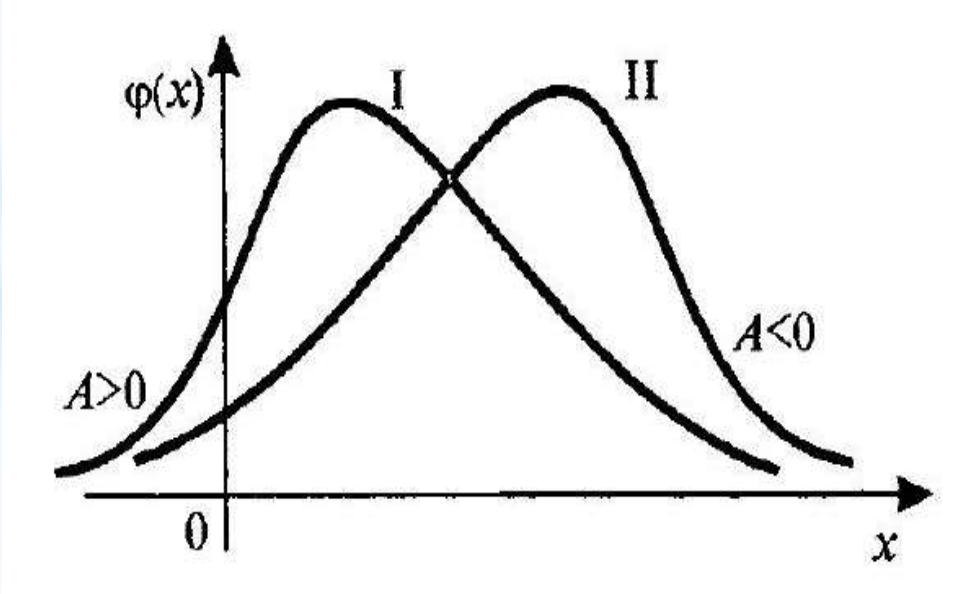
Коэффициент асимметрии, асимметрия - отношение центрального момента третьего порядка к кубу С.К.О.:

$$a_s = \frac{\mu_3}{\sigma^3}$$

Коэффициент асимметрии характеризует скошенность распределения по отношению к мат.ожиданию. Асимметрия положительна, если «длинная часть» кривой распределения расположена справа от математического ожидания; асимметрия отрицательна, если «длинная часть» кривой расположена слева от математического ожидания.

Оценка характеристик формы

$$As = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3}; \quad As = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{n\sigma^3}$$



* Оценка характеристик формы

Для быстрой предварительной оценки используют меру скошенности. **Мера скошенности** (оэффициент асимметрии Пирсона) определяется как отклонение среднего арифметического от моды

$$A = \frac{\overline{x} - Mo}{\sigma}$$

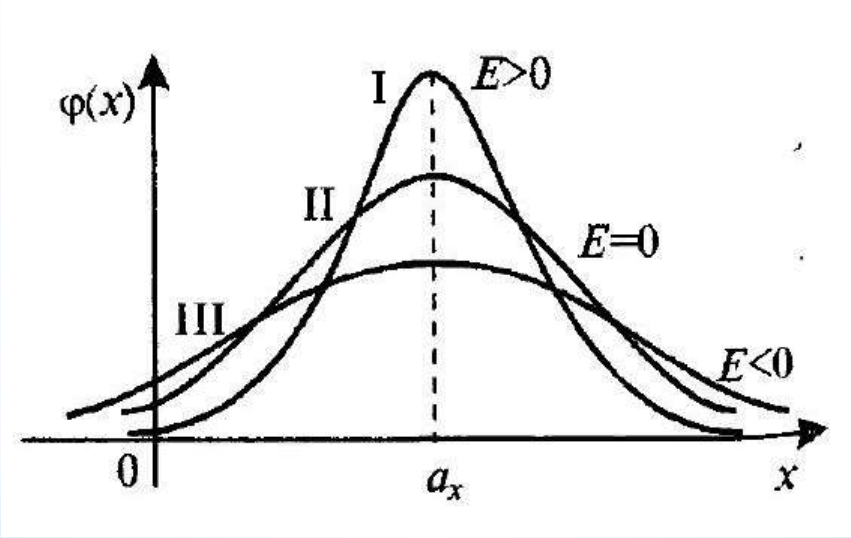
Характеризует асимметрию только в центральной части распределения.

Эксцессом (или коэффициентом эксцесса) случайной величины называется число:

$$e_s = \frac{\mu_4}{\sigma^4} - 3$$

Кривые, более островершинные, чем нормальная, обладают положительным эксцессом, более плосковершинные - отрицательным эксцессом.

* Оценка характеристик формы



Пример. Для заданного вариационного ряда вычислить эксцесс и асимметрию.

x_i	0-0.4	0.4-0.8	0.8-1.2	1.2-1.6	1.6-2.0	2.0-2.4	Итого
n_i	3	17	21	25	24	10	100

* Оценка характеристик формы

Интервалы	Середина интервала, x_i	n_i	$x_i n_i$	$(x_i - \overline{x})^2 n_i$	$(x_i - \overline{x})^3 n_i$	$(x_i - \overline{x})^4 n_i$
0 – 0.4	0.2	3	0.6	3.763	-4.215	4.721
0.4 – 0.8	0.6	17	10.2	8.813	-6.345	4.569
0.8 – 1.2	1.0	21	21.0	2.150	-0.688	0.220
1.2 – 1.6	1.4	25	35.0	0.160	0.013	0.001
1.6 – 2	1.8	24	43.2	5.530	2.654	1.274
2 – 2.4	2.2	10	22.0	7.744	6.815	5.997
Итого	--	100	132.0	28.160	-1.766	16.782

$$\overline{x} = \frac{\sum x_i n_i}{\sum n_i} = \frac{132}{100} = 1.32$$

$$Mo = x_{Mo} + h_{Mo} \cdot \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} = 1.2 + 0.4 \cdot \frac{25 - 21}{(25 - 21) + (25 - 24)} = 1.52$$

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 n_i}{\sum n_i} = \frac{28.16}{100} = 0.282$$

* Оценка характеристик формы

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.282} = 0.531$$

Коэффициент асимметрии Пирсона:

$$A = \frac{\overline{x} - Mo}{\sigma} = \frac{1.32 - 1.52}{0.531} = -0{,}377$$

$$\mu_3 = \frac{\sum (x_i - \overline{x})^3 n_i}{\sum n_i} = \frac{-1.766}{100} = -0.018$$

$$\alpha_s = \frac{-0.018}{0.531^3} = -0.12$$

$$\mu_4 = \frac{\sum (x_i - \overline{x})^4 n_i}{\sum n_i} = \frac{16.782}{100} = 0.168$$

$$\varepsilon_k = \frac{0.168}{0.531^4} - 3 = -0.887$$