2.3. ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ И ИХ СВОЙСТВА

Пусть требуется определить значение параметра **Θ** по экспериментальным данным.

Значение параметра, вычисленное по ограниченному объему ЭД, является случайной величиной, т. е. значение такой величины от выборки к выборке может меняться заранее не предвиденным образом. Следовательно, в результате обработки ЭД определяется не значение параметра **Θ**, а только лишь его приближенное значение – статистическая оценка параметра **Θ***. Оценка есть неточное, неполное, приближенное отображение параметра. Получить статистическую оценку параметра теоретического распределения означает найти функцию от имеющихся результатов наблюдения, которая и даст приближенное значение искомого параметра:

$$\boldsymbol{\Theta}_n^* = \boldsymbol{\Theta}^*(x_1, x_2, \ldots, x_n)$$

Различают два вида оценок – точечные и интервальные. Точечными называют такие оценки, которые характеризуются одним числом. При малых объемах выборки точечные оценки могут значительно отличаться от истинных значений параметров, поэтому их применяют при большом объеме выборки. Интервальные оценки задаются двумя числами, определяющими вероятный диапазон возможного значения параметра. Эти оценки применяются как для малых, так и для больших выборок. Рассмотрим вначале точечные оценки.

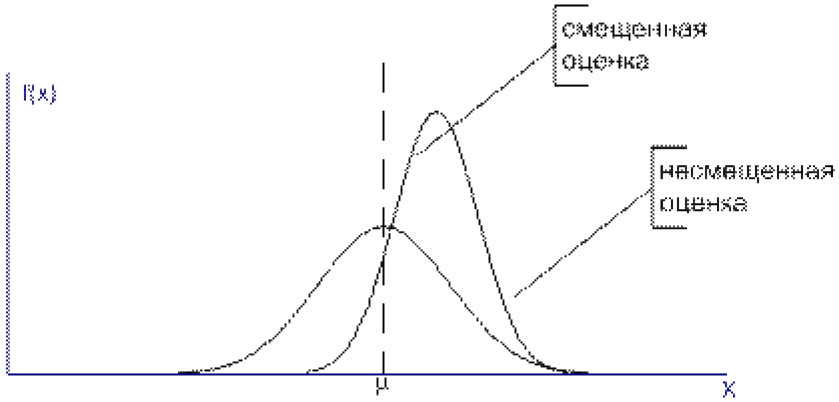
Применительно к каждому оцениваемому параметру закона распределения генеральной совокупности существует множество функций, позволяющих вычислить искомые значения. Например, оценку математического ожидания можно вычислить, взяв среднее арифметическое выборочных значений, половину суммы крайних членов вариационного ряда, средний член выборки и т.д. Указанные функции отличаются качеством оценок и трудоемкостью реализации.

Качество оценок характеризуется такими свойствами, как *несмещенность, состоятельность, эффективность и достаточность*.

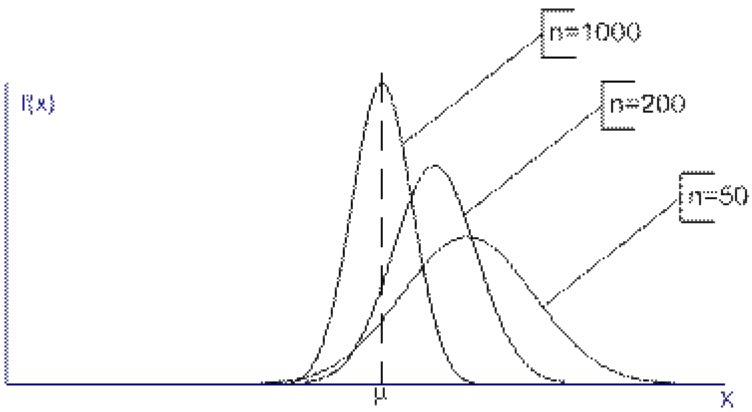
Несмещенность характеризует отсутствие систематических (в среднем) отклонений оценки от параметра при любом конечном, в том числе и малом, объеме выборки, т. е.

$$M[\boldsymbol{\Theta}^*] = \boldsymbol{\Theta}.$$

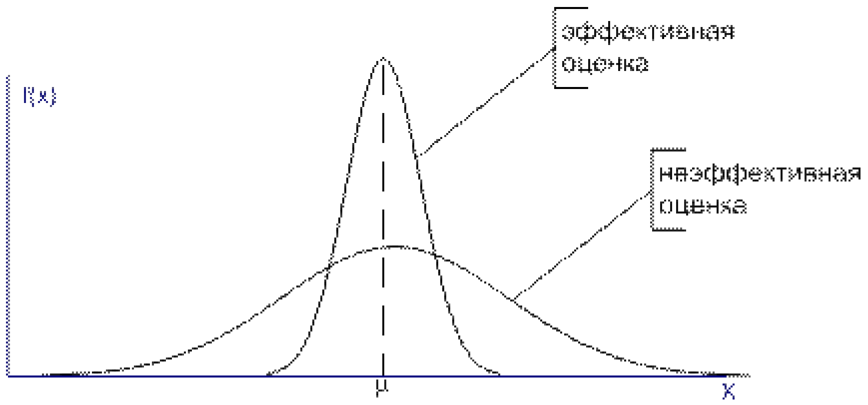
Использование статистической оценки, математическое ожидание которой не равно оцениваемому параметру, приводит к систематическим ошибкам. Не всегда наличие смещения плохо. Оно может быть существенно меньше погрешности регистрации значений параметра или давать дополнительную гарантию выполнения требований к значению параметра (если даже при положительном смещении оценка **Θ*** меньше предельно допустимого значения, то несмещенное значение тем более будет отвечать этому условию). В таких ситуациях допустимо применение смещенных оценок, если они вычисляются проще, чем несмещенные. Но даже несмещенная оценка может быть удалена от истинного значения.



<u>Состоятельность</u> характеризует сходимость по вероятности оценки **Θ*** к истинному значению параметра **Θ** при неограниченном увеличении объема выборки **n**. Для состоятельности оценки достаточно, но не обязательно, чтобы математическое ожидание квадрата отклонения оценки от параметра $M[\Theta - \Theta^*]^2$ стремилось к нулю с увеличением объема выборки (здесь и далее символ М означает математическое ожидание). Свойство состоятельности проявляется при неограниченном увеличении **n**, а при небольших объемах ЭД наличие этого свойства еще недостаточно для применения оценки.



<u>Эффективность</u> характеризует разброс случайных значений оценки около истинного значения параметра. Среди всех оценок следует выбрать ту, значения которой теснее сконцентрированы около оцениваемого параметра.



Для многих применяемых способов оценивания выборочные распределения параметров асимптотически нормальны, поэтому часто мерой эффективности служит дисперсия оценки. В таком понимании эффективная оценка – это оценка с минимальной дисперсией. При неограниченном увеличении n эффективная оценка является и состоятельной. В случае оценивания одного параметра дисперсия несмещенной оценки отвечает условию Рао – Крамера

$$D(\Theta^*) = \geq \left[-nM \left(\frac{\partial^2 \ln f\,(x, \Theta)}{\partial \Theta^2} \right) \right]^{-1}$$

где $f(x, \mathbf{\Theta})$ – плотность распределения варианты; n – количество наблюдений.

Сравнительная эффективность оценки с дисперсией $D_n[\mathbf{\Theta}^*]$ измеряется коэффициентом эффективности

$$\varepsilon = D[\mathbf{\Theta}^*]/D_k[\mathbf{\Theta}^*],$$

который не превышает единицы. Чем ближе коэффициент ε к единице, тем эффективнее оценка. Отмеченное ограничение применимо и к дискретным распределениям, если вместо плотности распределения подставить в него функцию вероятности.

Достаточность характеризует полноту использования информации, содержащейся в выборке. Другими словами, оценка **Θ*** будет достаточной, если все другие независимые оценки на основе данной выборки не дают дополнительной информации об оцениваемом параметре. Эффективная оценка обязательно является и достаточной.

Рассмотренные свойства применимы также и к ЭД, которые характеризуются многомерными распределениями вероятностей.

Подходы к формированию оценок разработаны в теории несмещенных оценок, предложенной А. Н. Колмогоровым и С. Рао. В данной теории предполагается известным с точностью до параметра **Θ** вид функции плотности распределения наблюдаемой случайной величины $f(x, \mathbf{\Theta})$. Вид распределения устанавливается исходя из априорных соображений, например, на основе общепринятых суждений о характере безотказной работы технических средств. Тогда задача сводится к нахождению такой функции от результатов наблюдений, которая дает несмещенную и эффективную оценку.

2.4. ОЦЕНКИ МОМЕНТОВ И КВАНТИЛЕЙ РАСПРЕДЕЛЕНИЯ

Для характеристики эмпирического распределения можно использовать оценки центральных и начальных моментов. Применение находят моменты до четвертого порядка включительно, так как точность выборочных моментов резко падает с увеличением их порядка, в частности, дисперсия начальных моментов порядка *r* зависит от моментов порядка *2r*. Она становится значительной для моментов высокого порядка даже при больших объемах выборки. Выборочные значения моментов определяют непосредственно по выборке или по сгруппированным данным.

Выборочные значения центральных моментов случайной величины *X* вычисляются по выборке с применением с формул

$$\alpha_1 = \overline{x} = \frac{1}{n}\sum_{i-1}^{n} x_i$$

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha_1)^k\,, k = 2,3,4,\ldots \text{(2.1)}$$

Указанные величины являются оценками соответствующих теоретических моментов распределения и должны рассматриваться как случайные. Вычисления по формулам (2.1) дают состоятельные, но смещенные оценки моментов старше первого. Смещение удается устранить введением поправочных коэффициентов, зависящих от объема выборки. Несмещенными и состоятельными будут оценки

$$\mu_2{}^* = S^2 = \frac{1}{n-1}\sum_{i=}(x_i - \overline{x})^2$$

$$\mu_3{}^* = \frac{n^2}{(n-1)(n-2)}\tilde{\mu}_3;$$

$$\mu_4{}^* = \frac{n(n^2-2n+3)\tilde{\mu}_4 - 3n(2n-3)\mu_2^2}{(n-1)(n-2\_(n-3)} \text{ (2.2)}$$

Оценки моментов по сгруппированным ЭД

$$\alpha_{1,g} = \overline{x_g} = \frac{1}{n}\sum_{j=1}^{g} n_j \bar{x}_j$$

$$\tilde{\mu}_{k,g} = \frac{1}{n}\sum_{j=1}^{g} n_j(\bar{x}_j - \overline{x_g})^k\,, k = 2,3,4,\ldots \text{ (2.3)}$$

где $\bar{x}_j$ – центр j-го интервала; g – количество интервалов. Первый и последний интервалы могут быть открытыми. Тогда для них центры устанавливаются экспертным путем, исходя из сущности параметра.

Группирование и приписывание соответствующей частости значения варианты в середине интервала группирования вносят некоторые искажения (при этом подразумевается, что распределение результатов наблюдения внутри интервала равномерное). Если распределение непрерывно и имеет достаточно высокий порядок соприкосновения с осью абсцисс (значения функции плотности распределения быстро убывают при удалении от центра распределения), то для снижения ошибок группирования используют поправки Шеппарда. Уточненные значения выборочных моментов для случая равной длины всех интервалов определяются через оценки моментов, вычисленные по сгруппированным данным:

$$a_1 = a_{1,g};$$

$$\mu_2 = \tilde{\mu}_g - h^2/12;$$

$$\mu_3 = \tilde{\mu}_{3,g};$$

$$\mu_4 = \tilde{\mu}_{4,g} - \mu_{2,g}h^2/2 + 7h^4/240. \qquad (2.4)$$

где h – длина интервала группирования. Указанные поправки ведут к уточнению только при соблюдении указанного условия, в противном случае они могут привести к еще большей ошибке.

Начальный эмпирический момент порядка r по несгруппированным данным определяется соотношением

$$\alpha_r = \frac{1}{n}\sum_{j=1}^{n} x_i^r\ ,\ r = 1,2,3,\ldots \qquad (2.5)$$

Центральные и начальные оценки моментов связаны между собой следующими зависимостями:

$$\tilde{\mu}_2 = \alpha_2 - \alpha_1^2;$$
$$\tilde{\mu}_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^2;$$
$$\tilde{\mu}_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4. \qquad (2.6)$$

В процессе обработки ЭД проще сначала определить оценки начальных моментов, потом перейти к смещенным оценкам центральных моментов и затем вычислить несмещенные оценки.

Оценки моментов (в том числе среднее арифметическое) могут быть дробными величинами, даже если значения параметров являются целыми.

Квантилью, отвечающей уровню вероятности G, называют такое значение варианты x_G, при котором функция распределения случайной величины принимает значение G, т. е. квантиль – это значение аргумента x_G функции распределения, при котором $F(x_G)=G$.

Эмпирическую квантиль находят по заданному значению вероятности G, используя вариационный ряд или ступенчатую ломаную линию.

Наряду с указанными параметрами для описания распределений применяются и другие характеристики:

- среднеквадратическое отклонение;
- коэффициенты асимметрии и эксцесса;
- стандартизованные переменные $u=(x-\alpha_1)/S$ (S—средне квадратичное отклонение).

Стандартизация переменной позволяет упростить расчеты, кроме того, в литературе многие справочные статистические таблицы приводятся именно для стандартизованных переменных. Нетрудно показать, что математическое ожидание стандартизованной переменной равно нулю, а дисперсия равна единице, т.е. после такого преобразования ЭД справедливы следующие соотношения:

$$M(u) = \frac{1}{n}\sum_{i=1}^{n} u_i = 0;$$

$$D(u) = \frac{1}{n}\sum_{i=1}^{n} (u_i - 0)^2 = 1.$$

Величина u называется центрированной и нормированной. Переход от центрированной и нормированной величины к исходной осуществляется простым преобразованием

$$x=uS+m_1.$$

Потери информации при стандартизации и обратном преобразовании не происходит.

Анализируя назначение рассмотренных параметров, необходимо отметить следующее. Одни параметры характеризует средние величины, а другие – вариацию. Главное назначение средних величин (оценок начальных моментов и в первую очередь первого момента распределения) состоит в их обобщающей функции. Это обобщение позволяет заменить множество различных индивидуальных значений показателя средней величиной, характеризующей всю однородную совокупность. Иначе говоря, средняя величина является типической характеристикой варианты в конкретной выборке. Иногда средняя величина обобщает и неоднородные совокупности данных. Например, может применяться такой показатель как среднее количество обработанных запросов на сервере в течение суток, хотя очевидно, что дневная загрузка сервера сильно отличается от загрузки в ночное время. Указанный показатель имеет смысл для оценки ресурса накопителей на жестких дисках. Наряду оценками математического ожидания (средними величинами в формулах 2.1 и 2.5) находят применение и другие оценки – среднее геометрическое, среднее гармоническое значение.

<u>*Средняя геометрическая*</u>. Может применяться при определении темпов роста когда индивидуальные значения представлены в относительных величинах. Существуют формулы для простой средней геометрической

$$\bar{x} = \sqrt[n]{nx_i} \ .$$

Для взвешенной средней геометрической:

$$\bar{X} = \sqrt[\sum fn_i]{nX_i^{n_i}}$$

<u>*Средняя гармоническая.*</u> Простая средняя гармоническая используется в том случае, когда веса значений признака одинаковы.

$$\bar{X} = \frac{n}{\sum \frac{1}{X}}$$

Гармоническая взвешанная используется в том случае, когда веса (или объемы выборок) по каждому признаку не равны.

$$\bar{X} = \frac{\sum n_i}{\sum \frac{n_i}{X_i}}$$

Каждый элемент ЭД формируется под влиянием как общих закономерностей, так и особых условий и случайных событий. Следовательно, в обработке ЭД большой интерес представляют вопросы оценки величин, характеризующих вариацию значений параметра у разных объектов или у одного и того же объекта в разные моменты времени. Вариацией какого-либо параметра (показателя) в совокупности наблюдений называется различие его значений у разных элементов этой совокупности.

2.5. ТОЧЕЧНАЯ ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.5.1. Сущность задачи точечного оценивания параметров

<u>*Точечная оценка*</u> предполагает нахождение единственной числовой величины, которая и принимается за значение параметра. Такую оценку целесообразно определять в тех случаях, когда объем ЭД достаточно велик. Причем не существует единого понятия о достаточном объеме ЭД, его значение зависит от вида оцениваемого параметра (к этому вопросу предстоит вернуться при изучении методов интервальной оценки параметров, а предварительно будем считать достаточной выборку, содержащую не менее чем 10 значений). При малом объеме ЭД точечные оценки могут значительно отличаться от истинных значений параметров, что делает их непригодными для использования.

Задача точечной оценки параметров в типовом варианте постановки состоит в следующем.

Имеется: выборка наблюдений ($x_1, x_2, \ldots, x_n$) за случайной величиной X. Объем выборки n фиксирован.

Известен вид закона распределения величины X, например, в форме плотности распределения $f(\mathbf{\Theta}, x)$, где $\mathbf{\Theta}$ – неизвестный (в общем случае векторный) параметр распределения. Параметр является неслучайной величиной.

Требуется найти оценку $\mathbf{\Theta^*}$ параметра $\mathbf{\Theta}$ закона распределения.

Ограничения: выборка представительная.

Существует несколько методов решения задачи точечной оценки параметров, наиболее употребительными из них являются методы максимального (наибольшего) правдоподобия, моментов и квантилей.

2.5.2. Метод максимального правдоподобия

Метод предложен Р. Фишером в 1912 г. Метод основан на исследовании вероятности получения выборки наблюдений $(x_1, x_2, \ldots, x_n)$. Эта вероятность равна

$$f(x_1, \Theta)\, f(x_2, \Theta)\, \ldots f(x_n, \Theta)\, dx_1\, dx_2 \ldots dx_n.$$

Совместная плотность вероятности

$$L(x_1, x_2 \ldots, x_n ; \Theta) = f(x_1, \Theta)\, f(x_2, \Theta)\, \ldots f(x_n, \Theta), \qquad (2.7)$$

рассматриваемая как функция параметра $\mathbf{\Theta}$, называется *функцией правдоподобия*.

В качестве оценки $\mathbf{\Theta^*}$ параметра $\mathbf{\Theta}$ следует взять то значение, которое обращает функцию правдоподобия в максимум. Для нахождения оценки необходимо заменить в функции правдоподобия Θ на q и решить уравнение

$$dL/d\,\mathbf{\Theta^*} = 0.$$

Для упрощения вычислений переходят от функции правдоподобия к ее логарифму $\ln L$. Такое преобразование допустимо, так как функция правдоподобия – положительная функция, и она достигает максимума в той же точке, что и ее логарифм. Если параметр распределения векторная величина

$$\mathbf{\Theta^*} = (q_1, q_2, \ldots, q_n),$$

то оценки максимального правдоподобия находят из системы уравнений

$$\begin{cases} d \ln L(q_1, q_2, \ldots, q_n) / dq_1 = 0; \\ d \ln L(q_1, q_2, \ldots, q_n) / dq_2 = 0; \\ \quad . \; . \; . \; . \; . \; . \; . \; . \; . \\ d \ln L(q_1, q_2, \ldots, q_n) / dq_n = 0. \end{cases}$$

Для проверки того, что точка оптимума соответствует максимуму функции правдоподобия, необходимо найти вторую производную от этой функции. И если вторая производная в точке оптимума отрицательна, то найденные значения параметров максимизируют функцию.

Итак, нахождение оценок максимального правдоподобия включает следующие этапы: построение функции правдоподобия (ее натурального логарифма); дифференцирование функции по искомым параметрам и составление системы уравнений; решение системы уравнений для нахождения оценок; определение второй производной функции, проверку ее знака в точке оптимума первой производной и формирование выводов.

Пример 2.3. Будем считать, что случайная величина *X* имеет нормальное распределение. Необходимо найти оценки максимального правдоподобия параметров m и S этого распределения.

Решение. Функция правдоподобия для выборки ЭД объемом n

$$L(\alpha,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} exp\left[-\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2}\right]$$

.

Логарифм функции правдоподобия

$$ln\, L\,(\alpha,\sigma) = n\, ln(\sqrt{2\pi}) - n\, ln\,\sigma - \left\{\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{2\sigma^2}\right\}.$$

Система уравнений для нахождения оценок параметров

$$\frac{\partial\, ln\, L\,(\alpha,\sigma)}{\partial\alpha} = \left\{\sum_{i=1}^{n}\frac{(x_i-\alpha)}{\sigma^2}\right\} = 0;$$

$$\frac{\partial\, ln\, L\,(\alpha,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{\sigma^3} = 0.$$

Из первого уравнения следует: $\sum_{i=1}^{n}(x_i-\alpha) = 0,$

или окончательно $\alpha = \frac{\sum_{i=1}^{n}x_i}{n}.$

Таким образом, среднее арифметическое является оценкой максимального правдоподобия для математического ожидания.

Из второго уравнения можно найти

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\alpha)^2}{n}.$$

Эмпирическая дисперсия является смещенной. После устранения смещения

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\alpha)^2}{n-1}.$$

Для проверки того, что полученные оценки максимизируют значение функции правдоподобия, возьмем вторые производные

$$\frac{\partial^2 \ln L\,(\alpha,\sigma)}{\partial\alpha^2} = -\frac{n}{\sigma^2};$$

$$\frac{\partial^2 \ln L\,(\alpha,\sigma)}{\partial\sigma^2} = \frac{n}{\sigma^2} - 3\sum_{i=1}^{n}\frac{(x_i-\alpha)^2}{\sigma^4} = \frac{n}{\sigma^2} - \frac{3n}{\sigma^2} = -2n\sigma^2.$$

Вторые производные от функции $\ln(L(m,S))$ независимо от значений параметров меньше нуля, следовательно, найденные значения параметров являются оценками максимального правдоподобия.

Метод максимального правдоподобия позволяет получить состоятельные, эффективные (если таковые существуют, то полученное решение даст эффективные оценки), достаточные, асимптотически нормально распределенные оценки. Этот метод может давать как смещенные, так и несмещенные оценки. Смещение удается устранить введением поправок. Метод особенно полезен при малых выборках.

2.5.3. Метод моментов

Метод предложен К. Пирсоном в 1894 г. Сущность метода:

- выбирается столько эмпирических моментов, сколько требуется оценить неизвестных параметров распределения. Желательно применять моменты младших порядков, так как погрешности вычисления оценок резко возрастают с увеличением порядка момента;
- вычисленные по ЭД оценки моментов приравниваются к теоретическим моментам;
- параметры распределения определяются через моменты, и составляются уравнения, выражающие зависимость параметров от моментов, в результате получается система уравнений. Решение этой системы дает оценки параметров распределения генеральной совокупности.

Пример 2.4. Предположим, что случайная величина X имеет гамма-распределение. Необходимо найти оценки параметров этого распределения (можно отметить, что нормальное распределение является частным случаем гамма-распределения).

Решение. Функция плотности гамма-распределения имеет вид

$$f(\chi,\lambda) = \frac{\lambda^\nu}{\Gamma(\nu)}\chi^{\nu-1}\,exp(-\lambda\chi), \qquad \chi \geq 0, \lambda > 0, \nu \geq 0.$$

Распределение характеризуется двумя параметрами ν и λ, поэтому следует выразить один параметр через оценку математического ожидания, а другой – через оценку дисперсии. Математическое ожидание и дисперсия этого распределения равны ν/λ и ν/λ^2 соответственно. Пусть их оценки определены равны:

$$\alpha_1 = 27{,}51,\ \mu_2 = 0{,}91.$$

Составим систему уравнений для оцениваемых параметров

$$\frac{\nu}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \alpha_1,$$

$$\frac{\nu}{\lambda^2} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \alpha_1)^2 = \mu_2.$$

Разделив оценку математического ожидания на оценку дисперсии, получим

$$\lambda = \alpha_1/\mu_2 = 30{,}12,$$

Метод моментов позволяет получить состоятельные, достаточные оценки, они при довольно общих условиях распределены асимптотически нормально. Смещение удается устранить введением поправок. Эффективность оценок невысокая, т.е. даже при больших объемах выборок дисперсия оценок относительно велика (за исключением нормального распределения, для которого метод моментов дает эффективные оценки). В реализации метод моментов проще метода максимального правдоподобия. Напомним, что метод целесообразно применять для оценки не более чем четырех параметров, так как точность выборочных моментов резко падает с увеличением их порядка.

2.5.4. Метод квантилей

Сущность метода квантилей схожа с методом моментов: выбирается столько квантилей, сколько требуется оценить параметров; неизвестные теоретические квантили, выраженные через параметры распределения, приравниваются к эмпирическим квантилям. Решение полученной системы уравнений дает искомые оценки параметров.

Дисперсия $D(x_G)$ выборочной квантили обратно пропорциональна квадрату плотности распределения

$$D(x_G) = [G(1-G)]/[nf^2(x_G)]$$

в окрестностях точки x_G. Поэтому следует выбирать квантили вблизи тех значений x, в которых плотность вероятности максимальна.

Пример 2.5. Оценить методом квантилей параметры нормального распределения случайной величины.

Решение. Так как требуется определить два параметра распределения m и S, то выберем из вариационного ряда две эмпирические квантили. Например, можно взять

$$G_1 = 5/44 = 0{,}114;\ \ x_{G1} = 26{,}13;$$

$$G_2 = 31/44 = 0{,}705;\ \ x_{G2} = 28{,}01.$$

Используя стандартные функции математических пакетов, для выбранных значений G_1 и G_2 определим значения аргументов теоретической функции распределения для стандартизованной переменной

$$U_{G1} = -1,207; \ \ U_{G2} = 0,538.$$

Составим систему из двух уравнений

$$U_{G1} = (x_{G1} - m)/S;$$

$$U_{G1} = (x_{G2} - m)/S.$$

Решение системы позволит найти искомые оценки параметров

$$m = (U_{G2}\ x_{G1} - U_{g1}\ x_{G2})/(U_{g2} - U_{g1}) = 27{,}42; \ \ S = (x_{G1} - m)/U_{g1} = 1{,}07.$$

Метод квантилей позволяет получить асимптотически нормальные оценки, однако они несут в себе некоторый субъективизм, связанный с относительно произвольным выбором квантилей. Эффективность оценок не выше метода моментов. Определение оценок может приводить к необходимости численного решения достаточно сложных систем уравнений.

Оценки, вычисленные на основе различных методов, различаются. Универсального ответа на вопрос, какой из рассмотренных методов лучше или следует ли положиться на данный метод при решении любой задачи, нет. Значение оценки в каждом конкретном случае (для разных выборок) отличается от истинного значения параметра на неизвестную величину, иначе говоря, существует некоторая доля неопределенности в знании действительного значения параметра. Качество оценок можно определить косвенно путем проверки согласованности эмпирических данных и теоретического закона распределения.