ЛЕКЦИЯ 5. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

1.1. ПЕРВИЧНАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА И ОЦЕНКА ОСНОВНЫХ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Математическая статистика позволяет с помощью математических методов обрабатывать, систематизировать и использовать численные результаты эксперимента для получения практических выводов.

Под *генеральной совокупностью* в математической статистике понимается множество (гипотетическое) всех возможных результатов измерения некоторой величины, которые могут быть получены в данных условиях. Тем же самым понятием в теории вероятностей является случайная величина X.

Реальная серия повторных измерений этой величины $x_1, x_2, ..., x_n$ трактуется как случайная выборка из генеральной совокупности, или просто *случайная выборка*. Число n называется *объемом* случайной выборки. Приведем примеры.

1. Проведена серия повторных измерений одной и той же физической величины в одних и тех же условиях. Разброс результатов обусловлен погрешностью измерительной аппаратуры.

2. Измеряется некоторая характеристика одинаковых изделий, изготовленных при поточном производстве. Разброс результатов обусловлен особенностями технологии производства.

3. Измеряется некоторая характеристика людей определенного пола и интервала возрастов, например, рост. Разброс результатов обусловлен природными факторами.

В статистике принята следующая математическая модель подобных экспериментов. *Каждый элемент случайной выборки рассматривается как отдельная случайная величина.* Относительно этих случайных величин, которые в дальнейшем будем обозначать заглавными буквами, известна некоторая априорная информация.

Случайная выборка называется *повторной*, если все входящие в нее случайные величины $X_1, X_2, ..., X_n$ взаимно независимы и имеют одинаковую функцию распределения $F(x)$, причем ту же, что и наблюдаемая случайная величина X. На практике это, в частности, означает, что измерения производятся *независимо* друг от друга (полученные результаты одних измерений не влияют на возможные результаты других). Величины имеют одинаковые математические ожидания $M(X_i) = a$, то есть, результаты измерений свободны от *систематических ошибок* (результаты в среднем не смещены относительно истинного значения $M(X) = a$), и одинаковые дисперсии $D(X_i) = D(X) = \sigma^2$, что называется *равноточностью* измерений (например, измерения физической величины проведены на одном и том же приборе при одинаковых условиях).

Если случайные величины $X_1, X_2, ..., X_n$, образующие повторную случайную выборку, имеют нормальное распределение с одинаковыми параметрами a, σ , то есть, $X_i \sim N(a;\ \sigma)$, то такая выборка называется *нормальной*, а соответствующая генеральная совокупность – *нормальной генеральной совокупностью.*

В математической статистике рассматривают и *неповторные* выборки, в которых нарушается хотя бы одно из указанных условий: взаимная независимость, одинаковость функции распределения. Слово «повторная» обычно опускается, и пишут просто «выборка». Для неповторной выборки обязательно пишут «неповторная выборка».

Распределение случайной величины X характеризуется рядом *параметров* (математическое ожидание, дисперсия и т.д.). Эти параметры называют *параметрами генеральной совокупности.* Важной задачей математической статистики является нахождение по случайной выборке приближенных значений каждого из параметров, называемых *точечными оценками параметров*, или просто *оценками*. Таким образом, *оценкой параметра* β называется функция $f(X_1, X_2, ..., X_n)$ от случайной выборки, значение которой принимается в качестве приближенного для данного параметра и обозначается $\tilde{\beta}$.

$$\beta \approx \tilde{\beta} = f(X_1, X_2, ..., X_n). \tag{1.1}$$

Так как оценка зависит от случайной выборки, то она, в свою очередь, является случайной величиной. Для одного и того же параметра β по одной и той же выборке можно построить много различных оценок. Для сравнения оценок между собой введены специальные характеристики.

Оценка называется *несмещенной*, если ее математическое ожидание равно истинному значению параметра, то есть $M(\tilde{\beta}) = \beta$. Несмещенная оценка обеспечивает близость в среднем значений оценки к значению оцениваемого параметра, то есть не дает систематической ошибки.

Оценка называется *состоятельной*, если при $n \to \infty$ она сходится по вероятности к истинному значению оцениваемого параметра, то есть

$$\forall \varepsilon > 0 \ \lim_{n \to \infty} P(|\tilde{\beta}_n - \beta| > \varepsilon) = 0, \tag{1.2}$$

где β_n – оценка параметра β, найденная по выборке объема n. Смысл понятия состоятельности заключается в том, что с увеличением объема выборки оценка стремится к истинному значению параметра.

Точностью оценки $\tilde{\beta}$ называется средний квадрат отклонения оценки от β: $q^2(\tilde{\beta}) = M[(\tilde{\beta} - \beta)^2]$. Для несмещенных оценок точность определяется величиной дисперсии оценки: $q^2(\tilde{\beta}) = D(\tilde{\beta})$. Чем меньше $q = \sqrt{q^2(\tilde{\beta})}$, тем оценка лучше (точнее). *Наилучшей линейной оценкой* параметра β называется такая его линейная несмещенная оценка, которая имеет наименьшую дисперсию в классе всех линейных несмещенных оценок.

Пусть задана повторная случайная выборка $X_1, X_2, \ldots, X_n$. За *оценку математического ожидания* a принимается среднее арифметическое элементов выборки:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \, . \qquad (1.3)$$

Оценкой дисперсии σ^2 при *известном математическом ожидании* a является величина S_0^2

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2 \, . \qquad (1.4)$$

Оценкой дисперсии σ^2 при *неизвестном математическом ожидании* является величина S^2, которую называют *эмпирической дисперсией*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \, . \qquad (1.5)$$

Оценкой среднего квадратического отклонения σ при этом являются, соответственно, величины:

$$S_0 = \sqrt{S_0^2}\,; \qquad S = \sqrt{S^2}\, . \qquad (1.6)$$

Оценки математического ожидания и дисперсии, найденные по формулам (1.3), (1.4) и (1.5), являются несмещенными и состоятельными. Среднее арифметическое (1.3) является наилучшей линейной оценкой математического ожидания для повторной случайной выборки.

Оценка параметра $\sigma = \sqrt{\sigma^2}$ с помощью значений (1.6) является состоятельной, но смещенной (ее смещение убывает с увеличением n). Число $k = n - 1$ в формуле (1.5) называется *числом степеней свободы* оценки S^2.

Для практических расчетов формулу (1.5) удобно преобразовать к виду:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right). \tag{1.7}$$

Вычисление среднего значения $\overline{X}$ и оценки дисперсии S^2 упрощается, если отсчет значений X_i вести от подходящим образом выбранного начала отсчета C и в подходящем масштабе, то есть, если сделать линейную замену:

$$X_i = C + hU_i \,(i = 1, 2, ..., n). \tag{1.8}$$

При такой замене формулы (1.3) и (1.5) – (1.6) принимают вид:

$$\overline{X} = C + h\overline{U}, \qquad \overline{U} = \frac{1}{n}\sum_{i=1}^{n} U_i \ ; \tag{1.9}$$

$$S^2 = \frac{h^2}{n-1}\sum_{i=1}^{n}\left(U_i - \overline{U}\right)^2 = \frac{h^2}{n-1}\left(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2\right). \tag{1.10}$$

Для контроля вычислений весь расчет повторяют с другим началом отсчета *C,* результаты должны совпадать с точностью до возможных ошибок округления.

1.2. ОЦЕНКА МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ ПО НЕРАВНОТОЧНЫМ ИЗМЕРЕНИЯМ.

Часто встречающимся на практике случаем неповторной выборки является выборка, в которой случайные величины $X_1, X_2, \ldots, X_n$ взаимно независимы, имеют одинаковые математические ожидания, но различные дисперсии. Такие измерения называют *неравноточными.* Как правило, дисперсии каждой величины X_i не известны, но известны отношения дисперсий. Числа, обратно пропорциональные дисперсиям, называют *весами измерений* и обозначают w_i:

$$D(X_1) : D(X_2) : \ldots : D(X_n) = (1/w_1) : (1/w_2) : \ldots : (1/w_n)\,,$$

или

$$D(X_i) = \sigma^2 / w_i\,, \quad (i = 1, 2, \ldots, n)\,. \tag{1.13}$$

Коэффициент σ^2 в формуле (1.13) обычно не известен, он называется *дисперсией измерения с единичным весом*, веса w_i, как правило, известны.

Среднее арифметическое (1.3) для неравноточных измерений является несмещенной и состоятельной оценкой математического ожидания, но не является наилучшей линейной оценкой. Наилучшей линейной оценкой в этом случае будет *среднее взвешенное*:

$$\overline{X}_{взв} = \frac{\left(\sum_{i=1}^{n} X_i w_i \right)}{\sum_{i=1}^{n} w_i}, \tag{1.14}$$

эта оценка будет несмещенной и состоятельной. Она и используется на практике для неравноточных измерений.

1.3. ОЦЕНКА ДИСПЕРСИИ ПО НЕСКОЛЬКИМ СЕРИЯМ ЭКСПЕРИМЕНТОВ

Пусть заданы L независимых повторных выборок – L серий измерений. Случайные величины различных выборок имеют, в общем случае, различные математические ожидания, но дисперсии всех величин во всех выборках одинаковы. Такая ситуация возникает, когда одним и тем же прибором производят измерения различных величин (например, измерения значений функции для различных значений аргумента).

В этом случае для оценки единой дисперсии можно использовать значения измерений всех серий. По каждой выборке находят эмпирическую дисперсию S_j^2 с числом степеней свободы k_j (j = 1, 2, ..., L). В качестве оценки единой дисперсии принимают сводную эмпирическую дисперсию:

$$S_{св}^2 = \left(\sum_{j=1}^{L} S_j^2 k_j \right) \Bigg/ \sum_{j=1}^{L} k_j \tag{1.15}$$

с числом степеней свободы $k_{св} = \sum_{j=1}^{L} k_j$. Сводная оценка дисперсии (1.15) является несмещенной, она более точная, чем каждая из эмпирических дисперсий S_j^2.

1.4. НЕКОТОРЫЕ ИСПОЛЬЗУЕМЫЕ В СТАТИСТИКЕ РАСПРЕДЕЛЕНИЯ.

В статистике широко используются следующие законы распределения, связанные с обработкой экспериментов из нормальной генеральной совокупности:

1) стандартное нормальное распределение;

2) распределение Пирсона;

3) распределение Стьюдента;

4) распределение Фишера.

Стандартное нормальное распределение было рассмотрено на предыдущей лекции. Если случайная величина U имеет стандартное нормальное распределение, вероятность ее попадания в интервал (t_1, t_2) вычисляется по формуле:

$$P(t_1 < U < t_2) = \Phi(t_2) - \Phi(t_1), \tag{1.16}$$

где $\Phi(x)$ – значение интеграла вероятностей. Если случайная величина X имеет нормальное распределение с параметрами a, σ ($X \sim N(a; \sigma)$), вероятность ее попадания в интервал (x_1, x_2) вычисляется по формуле:

$$P(x_1 < X < x_2) = \Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right). \tag{1.17}$$

Распределением Пирсона с к степенями свободы называется распределение суммы квадратов $к$ взаимно независимых случайных величин, каждая из которых имеет стандартное нормальное распределение. Можно также говорить, что это распределение квадрата длины случайного вектора, имеющего k координат, каждая из которых имеет стандартное нормальное распределение U_i. Случайная величина, имеющая распределение Пирсона, часто обозначается $\chi^2(k)$:

$$\chi^2(k) = U_1^2 + U_2^2 + \ldots + U_k^2$$

и само распределение называется χ^2 - распределением.

Математическое ожидание и дисперсия величины $\chi^2(k)$ равны соответственно:

$$M(\chi^2) = k; \qquad D(\chi^2) = 2k.$$

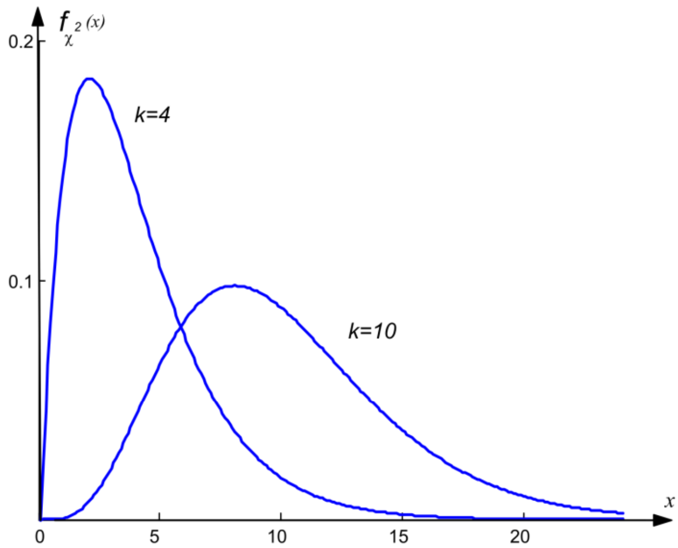

Рис. 1.2. График функции плотности вероятностей случайной величины, имеющей распределение Пирсона (χ^2) с k степенями свободы

График функции плотности вероятностей случайной величины, имеющей распределение Пирсона, представлен на рис. 1.2. Таблица квантилей $\chi_P^2(k)$ распределения Пирсона приведена в таблице П3 приложения.

Распределением Стьюдента (или ***t-распределением***) ***с к степенями свободы*** называется распределение отношения

$$T(k) = \frac{U}{\sqrt{(U_1^2 + U_2^2 + \ldots + U_k^2)/k}} = \frac{U}{\sqrt{\chi^2(k)/k}}, \tag{1.18}$$

где все случайные величины $U, U_1, U_2, \ldots, U_к$ – взаимно независимы и имеют стандартное нормальное распределение. Распределение Стьюдента симметрично относительно центра $M\left(T(k)\right) = 0$. Кривая распределения Стьюдента (рис. 1.3) внешне похожа на кривую стандартного нормального распределения. При $k \to \infty$ распределение Стьюдента стремится к стандартному нормальному распределению. Но при малых значениях $к$ числовые значения функций плотности распределения вероятностей существенно различаются, что особенно сказывается на квантилях $t_P(k)$ при вероятностях P, близких к 0 и к 1. Квантили t-распределения Стьюдента $t_P(k)$ приведены в приложении (табл. П2).

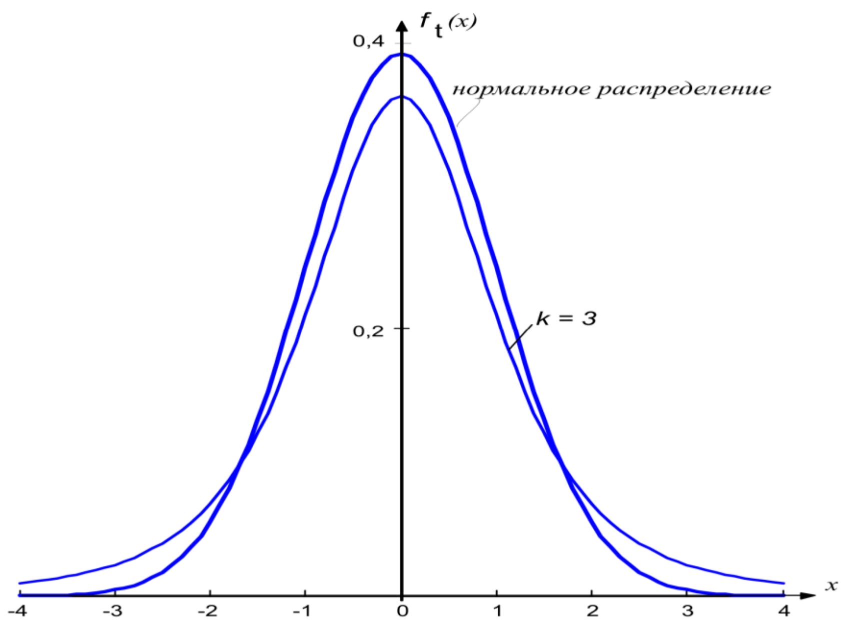

Рис. 1.3. Графики функций плотности вероятностей распределения случайных величин, имеющих стандартное нормальное распределение и распределение Стьюдента (t) с k степенями свободы

Распределением Фишера (или ***F-распределением***) ***с $к_1$ и $к_2$ степенями свободы*** называется распределение отношения

$$F = \frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}.$$

Кривая распределения Фишера внешне похожа на кривую распределения Пирсона (рис. 1.2). Квантили F-распределения Фишера $F_P(k_1, k_2)$ приведены в приложении (табл. П4).

1.5. ПОСТРОЕНИЕ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

В разделе 1.1. были рассмотрены точечные оценки параметров Напомним, что точечная оценка $\tilde{\beta}$ неизвестного параметра β для каждой случайной выборки дает лишь одно числовое значение, которое мы принимаем за приближенное значение этого параметра. Важная задача – определить точность полученного приближения, то есть на сколько точечная оценка может отклоняться от истинного значения параметра. Ответить на этот вопрос позволяют доверительные интервалы.

Доверительным интервалом параметра β называется интервал со случайными границами $(\tilde{\beta}-\varepsilon_1;\ \tilde{\beta}+\varepsilon_2)$, который накрывает истинное значение параметра β с заданной вероятностью Р, которая называется *доверительной вероятностью.* Величина $\alpha = 1 - \mathrm{P}$ называется *уровнем значимости*. При этом обычно требуют, чтобы вероятности выхода за границы доверительного интервала в обе стороны были равны между собой, а именно:

$$P(\beta < \tilde{\beta}-\varepsilon_1) = P(\beta > \tilde{\beta}+\varepsilon_2) = (1-\mathrm{P})/2 = \alpha/2\,.$$

Это дополнительное требование обеспечивает единственность решения задачи.

Пусть задана повторная случайная выборка $X_1, X_2, ..., X_n$ из нормальной генеральной совокупности. Это означает, что результаты эксперимента независимы и подчиняются нормальному закону распределения с одинаковыми параметрами $X_i \sim N(a;\sigma)$. Доверительные интервалы для параметров нормального распределения находят следующим образом.

1.5.1. Доверительный интервал для математического ожидания a.

С вероятностью Р математическое ожидание a принадлежит интервалу

$$a \in (\overline{X}-\varepsilon;\ \overline{X}+\varepsilon)\,, \tag{1.19}$$

– если σ известно, то

$$\varepsilon = u_{1-\alpha/2}\,\sigma\big/\sqrt{n}\,, \tag{1.20}$$

– если σ не известно, то

$$\varepsilon = t_{1-\alpha/2}(k)\,S\big/\sqrt{n}\,, \tag{1.21}$$

где $\overline{X}$ – оценка математического ожидания (1.3); $S=\sqrt{S^2}$ – оценка среднего квадратического отклонения σ (1.6) при неизвестном математическом ожидании; $u_{1-\alpha/2}$ – квантиль стандартного нормального распределения; $t_{1-\alpha/2}(k)$ – квантиль распределения

Стьюдента с $к$ степенями свободы; n – объем выборки; $к$ – число степеней свободы при вычислении оценки S.

Часто доверительный интервал для математического ожидания записывают символически:

$$a = \overline{X} \pm \varepsilon . \tag{1.22}$$

1.5.2. Доверительный интервал для среднего квадратического отклонения

Доверительный интервал для среднего квадратического отклонения σ при доверительной вероятности $\mathrm{P} = 1 - \alpha$ имеет вид:

– если математическое ожидание a известно, то

$$S_0 \sqrt{\frac{n}{\chi^2_{1-\alpha/2}(n)}} < \sigma < S_0 \sqrt{\frac{n}{\chi^2_{\alpha/2}(n)}} , \tag{1.23}$$

где S_0 – оценка среднего квадратического отклонения σ (1.6) при известном математическом ожидании;

– если математическое ожидание a не известно, то

$$S \sqrt{\frac{k}{\chi^2_{1-\alpha/2}(k)}} < \sigma < S \sqrt{\frac{k}{\chi^2_{\alpha/2}(k)}} , \tag{1.24}$$

где S – оценка среднего квадратического отклонения σ (1.6) при неизвестном математическом ожидании; $\chi^2_P(k)$ – квантиль распределения Пирсона с $к$ степенями свободы; $к$ – число степеней свободы оценки S.

1.5.3. Доверительный интервал для дисперсии

Доверительный интервал для дисперсии σ^2 при доверительной вероятности $\mathrm{P} = 1 - \alpha$, соответственно, находится по формулам:

– если математическое ожидание a известно, то

$$S_0^2 \frac{n}{\chi^2_{1-\alpha/2}(n)} < \sigma^2 < S_0^2 \frac{n}{\chi^2_{\alpha/2}(n)} , \tag{1.25}$$

– если математическое ожидание a не известно, то

$$S^2 \frac{k}{\chi^2_{1-\alpha/2}(k)} < \sigma^2 < S^2 \frac{k}{\chi^2_{\alpha/2}(k)} \tag{1.26}$$