Тема 5А. Проверка статистических гипотез. Критерий хи-квадрат и критерий согласия Колмогорова.

1. Основные определения.

В задачах математической статистики истинное вероятностное распределение P^* совокупности наблюдаемых величин $(\xi_1,\ldots,\xi_n)$, как правило, неизвестно, в некоторых случаях известно лишь множество вероятностных распределений $\wp$, которому принадлежит P^*. Поскольку распределение P^* неизвестно, то во многих задачах в отношении P^* формулируются различные предположения или иначе гипотезы. В основе всякой гипотезы располагается некоторое утверждение о том, что неизвестное распределение P^* обладает заданным свойством. В общем случае указанному свойству могут удовлетворять несколько различных распределений, образующих подмножество распределений $\wp_0 \subseteq \wp$. Отсюда возникает трактовка гипотезы как соответствующего подмножества $\wp_0$, и утверждения гипотезы как принадлежность истинного вероятностного распределения наблюдения P^* множеству $\wp_0$.

Определение 5А.1.

Неформально *статистической гипотезой* (*гипотезой*) называют некоторое утверждение о свойствах вероятностного распределения совокупности наблюдаемых величин P^*. Формально гипотезе соответствует подмножество $\wp_0 \subseteq \wp$ и утверждение гипотезы заключается в том, что $P^* \in \wp_0$.

Множество всех оставшихся распределений $\wp \setminus \wp_0$ разбивают на одно или несколько множеств $\wp_i$ $(i=\overline{1,k})$, которые ставят в соответствие гипотезам H_i.

Определение 5А.2.

Статистическая гипотеза H_i называется *простой*, если соответствующее ей множество $\wp_i$ состоит из одного распределения, либо *сложной*, если соответствующее ей множество $\wp_i$ состоит из более чем одного распределения.

Определение 5А.3.

Гипотезу H_0 выделяют особым образом и называют *основной* (*нулевой*) гипотезой. Гипотезы H_i $(i=\overline{1,k})$ называют *альтернативными гипотезами*, при этом распределения $P \in \wp_i$ называют *альтернативными распределениями*.

Задача проверки основной гипотезы H_0 против альтернативных гипотез H_i $(i=\overline{1,k})$ заключается в том, чтобы ответить на вопрос: согласуется ли основная гипотеза H_0 с полученной (в результате эсперимента) реализацией наблюдаемых величин, или иначе согласуется ли выдвинутое гипотетическое предположение с тем, что наблюдается в действительности.

Для решения задачи проверки гипотезы разрабатывается специальный метод обработки наблюдаемых величин, который позволяет принять обоснованное решение о том, принимается основная гипотеза или отклоняется.

Определение 5А.4.

Метод обработки наблюдений, согласно которому гипотеза принимается либо отклоняется, называется *статистическим критерием* (*критерием*).

Если статистический критерий выявил наличие согласованности между гипотезой и реализацией совокупности наблюдаемых величин, полученной в результате эксперимента, то говорят, что *гипотеза принимается*, в противном случае говорят, что *гипотеза отклоняется*. Следует иметь в виду, что статистический критерий ни в коем случае не доказывает гипотезу, а только лишь отвечает на вопрос: согласуется ли утверждение гипотезы с реализацией наблюдаемых величин. В этом смысле иногда статистические критерии называют *критериями согласия*.

Общий принцип организации статистических критериев является практически одинаковым для различных задач проверки статистических гипотез. Предположим, рассматривается задача проверки основной гипотезы H_0, с соответствующей ей множеством распределений $\wp_0$, против альтернативной гипотезы H_1, с соответствующей ей множеством распределений $\wp_1$.

Пусть X – множество всех возможных реализаций наблюдаемой совокупности величин $(\xi_1,..., \xi_n)$. Множество X можно разбить на два подмножества X_0 и X_1:

$$X_0 \cup X_1 = X \ ,$$
$$X_0 \cap X_1 = \varnothing \ ,$$

причем сделать это таким образом, чтобы для всех распределений $P_0 \in \wp_0$, соответствующих основной гипотезе H_0, вероятности события $\{(\xi_1,..., \xi_n) \in X_1\}$ составляли величины не более α, где α – некоторое малое число:

$$\forall P_0 \in \wp_0 : P\{(\xi_1,..., \xi_n) \in X_1 \mid P_0\} \le \alpha \ . \tag{5A.1}$$

Множество X_0 при этом определяется как дополнение множества X_1 до множества X :

$$X_0 = X \setminus X_1 .$$

Выбор множества X_1, вообще говоря, не является однозначным, тем не менее, неоднозначность выбора X_1 можно устранить, если стараться выбирать множество X_1 таким образом, чтобы для всех альтернативных распределений $P_1 \in \wp_1$ вероятности события $\{(\xi_1,..., \xi_n) \in X_1\}$ были как можно больше:

$$\forall P_1 \in \wp_1 : P\{(\xi_1,..., \xi_n) \in X_1 \mid P_1\} \to \max \ .$$

Заметим, что если основная гипотеза H_0 является верной и неизвестное истинное распределение P^* принадлежит множеству $\wp_0$, то в этом случае, благодаря особому выбору множества X_1, в силу (5A.1) вероятность события $\{(\xi_1,..., \xi_n) \in X_1\}$ оказывается малой величиной:

$$P\{(\xi_1,..., \xi_n) \in X_1 \mid P^*\} \le \alpha \ . \tag{5A.2}$$

а вероятность обратного события $\{(\xi_1,..., \xi_n) \in X_0\}$ соответственно оказывается близкой к единице:

$$P\{(\xi_1,..., \xi_n) \in X \mid P^*\} = 1 \ ,$$
$$P\{(\xi_1,..., \xi_n) \in X_0 \cup X_1 \mid P^*\} = 1 \ ,$$
$$P\{(\xi_1,..., \xi_n) \in X_0 \mid P^*\} + P\{(\xi_1,..., \xi_n) \in X_1 \mid P^*\} = 1 \ ,$$
$$P\{(\xi_1,..., \xi_n) \in X_0 \mid P^*\} = 1 - P\{(\xi_1,..., \xi_n) \in X_1 \mid P^*\} \ge 1 - \alpha \tag{5A.3}$$

Таким образом, если основная гипотеза H_0 верна, то событие $\{(\xi_1,..., \xi_n) \in X_1\}$ имеет малую вероятность близкую к нулю, а событие $\{(\xi_1,..., \xi_n) \in X_0\}$ наоборот имеет большую вероятность близкую к единице.

Теперь располагая разбиением (X_0, X_1) множества X, не составляет труда построить статистический критерий: действительно, представим, что в результате эксперимента получена реализация $(x_1,..., x_n)$ наблюдаемых величин $(\xi_1,..., \xi_n)$, которая попадает в множество X_1, другими словами имеет место появление события $\{(\xi_1,..., \xi_n) \in X_1\}$. С одной стороны появление события $\{(\xi_1,..., \xi_n) \in X_1\}$ при однократном проведении эксперимента означает, что вероятность события $\{(\xi_1,..., \xi_n) \in X_1\}$ не является малой (появление событий, имеющих малую вероятность, при однократном проведении эксперимента считается неправдоподобным). С другой стороны, если считать, что основная гипотеза H_0 верна, то в силу особого выбора разбиения (X_0, X_1) согласно (5А.2) вероятность события $\{(\xi_1,..., \xi_n) \in X_1\}$ не превышает малую величину α. Получается противоречие: если основная гипотеза H_0 верна, то вероятность события $\{(\xi_1,..., \xi_n) \in X_1\}$ мала, а результаты экперимента напротив указывают, что вероятность этого же события не может быть малой. Полученное противоречие является основанием для отклонения основной гипотезы H_0, поскольку показывает, что исходное предположение о том, что основная гипотеза H_0 верна, является ошибочным, и поэтому основную гипотезу H_0 следует отклонить, считая, что верной является альтернативная гипотеза H_1.

Если же в результате эксперимента получена реализация $(x_1,..., x_n)$ наблюдаемых величин $(\xi_1,..., \xi_n)$, которая попадает в множество X_0 (имеет место появление события $\{(\xi_1,..., \xi_n) \in X_0\}$), то никакого противоречия не происходит: появление события $\{(\xi_1,..., \xi_n) \in X_0\}$ указывает на то, что вероятность этого события не является малой, что согласуется с неравенством (5А.3), полученным в предположении, что основная гипотеза H_0 является верной. В этом случае нет никаких причин для отклонения основной гипотезы H_0, и она принимается.

Таким образом, принцип действия статистического критерия оказывается чрезвычайно простым: если происходит событие $\{(\xi_1,..., \xi_n) \in X_0\}$, то принимается основная гипотеза H_0, а если происходит событие $\{(\xi_1,..., \xi_n) \in X_1\}$, то принимается альтернативная гипотеза H_1:

$$\begin{aligned} &\{(\xi_1,..., \xi_n) \in X_0\} \to \text{ принимается } H_0, \\ &\{(\xi_1,..., \xi_n) \in X_1\} \to \text{ принимается } H_1. \end{aligned} \qquad (5А.4)$$

Один из возможных способов построения разбиения (X_0, X_1) множества X для статистического критерия заключается в конструировании особой статистики $T(\xi_1,..., \xi_n)$, которая называется *статистикой критерия*. Статистику критерия стараются выбрать таким образом, чтобы:

а) распределения статистики $P\{T \mid P_0\}$ при всех распределениях $P_0 \in \wp_0$ отличались от распределений статистики $P\{T \mid P_1\}$ при всех альтернативных распределениях $P_1 \in \wp_1$ (чем больше «отличие» между двумя множествами распределений статистики, тем более хороший критерий можно построить).

б) существовал способ вычисления распределений $P\{T \mid P_0\}$ при всех $P_0 \in \wp_0$ (в некоторых случаях это требование может быть ослаблено: достаточно располагать способом приближенного вычисления распределений статистики $P\{T \mid P_0\}$ при каждом $P_0 \in \wp_0$).

Статистика критерия $T(\xi_1,..., \xi_n)$ ставит в соответствие множеству X множество всех возможных значений статистики Γ:

$$\Gamma = \{T(x_1,..., x_n) : (x_1,..., x_n) \in X\}.$$

Подмножеству X_1 при этом соответствует подмножество значений $\Gamma_\alpha \subseteq \Gamma$:

$$\Gamma_\alpha = \{T(x_1,..., x_n) : (x_1,..., x_n) \in X_1\},$$

а подмножеству X_0 соответствуют все оставшиеся значения статистики, образующих подмножество $\Gamma \setminus \Gamma_\alpha$.

Легко видеть, что имеет место следующая эквивалентность условий:

$$\{(\xi_1,..., \xi_n) \in X_0\} \Leftrightarrow \{T(\xi_1,..., \xi_n) \notin \Gamma_\alpha\},$$

$$\{(\xi_1,\ldots,\xi_n)\in X_1\} \Leftrightarrow \{T(\xi_1,\ldots,\xi_n)\in \Gamma_\alpha\},$$

поэтому правило принятия решений (5А.4) для статистического критерия может быть переформулировано в терминах статистики $T(\xi_1,\ldots,\xi_n)$: если значение статистики $T(\xi_1,\ldots,\xi_n)$ не принадлежит Γ_α (принадлежит множеству $\Gamma \setminus \Gamma_\alpha$), то принимается основная гипотеза H_0, если же значение статистики $T(\xi_1,\ldots,\xi_n)$ принадлежит множеству Γ_α, то принимается альтернативная гипотеза H_1:

$$\begin{aligned} &\{T(\xi_1,\ldots,\xi_n)\notin \Gamma_\alpha\} \to \text{ принимается } H_0, \\ &\{T(\xi_1,\ldots,\xi_n)\in \Gamma_\alpha\} \to \text{ принимается } H_1. \end{aligned} \qquad (5А.5)$$

Две формулировки критерия (5А.4) и (5А.5) являются эквивалентными, поэтому, как правило, пользуются только одной из них (той, использование которой является наиболее простым). При использовании формулировки (5А.5) обычно оперируют только множеством Γ_α, в некоторых случаях даже не восстанавливая исходное разбиение (X_0, X_1) множества X.

Определение 5А.5.

Множество Γ_α значений статистики критерия, при которых основная гипотеза отклоняется, называется *критической областью гипотезы*.

Термин критическая область отражает факт отклонения (критики) основной гипотезы.

В соответствии с (5А.1) критическую область Γ_α выбирают таким образом, чтобы вероятности $P\{T\in\Gamma_\alpha \mid P_0\}$ составляли величину не более α (подобный выбор Γ_α позволяет сделать указанное ранее свойство б) статистики критерия):

$$\forall P_0 \in \wp_0 : P\{T(\xi_1,\ldots,\xi_n)\in\Gamma_\alpha \mid P_0\} \le \alpha .$$

В этом случае, если предположить, что основная гипотеза H_0 верна, то есть $P^* \in \wp_0$, то событие $T\in\Gamma_\alpha$ неизбежно имеет малую вероятность $P\{T\in\Gamma_\alpha \mid P^*\}$ (не более α). В этом случае если в результате эксперимента реализуется событие $T\in\Gamma_\alpha$, то считается, что вероятность $P\{T\in\Gamma_\alpha \mid P^*\}$ не является малой, что противоречит тому, что $P^*\in\wp_0$, и поэтому гипотеза H_0 отклоняется.

Статистический критерий за редким исключением не является безошибочной процедурой и может принимать ошибочные решения. Действительно, даже если $P^*\in\wp_0$ и гипотеза H_0 верна, то в общем случае с ненулевой вероятностью может произойти событие $T\in\Gamma_\alpha$, в этом случае критерий совершит ошибку и отклонит верную гипотезу H_0. Наоборот, если $P^*\in\wp_1$ и верна альтернативная гипотеза H_1, то в общем случае с ненулевой вероятностью может произойти событие $T\notin\Gamma_\alpha$, в этом случае критерий совершит ошибку и отклонит верную гипотезу H_1.

Определение 5А.6.

Ошибкой первого рода называется отклонение критерием верной основной H_0 гипотезы. *Ошибкой второго рода* называется отклонение критерием верной альтернативной гипотезы H_1.

Ошибки первого и второго рода имеют соответствующие вероятности.

Определение 5А.7.

Вероятностью ошибки первого рода называется функция $\alpha(P_0)$, определенная для распределений $P_0\in\wp_0$:

$$\alpha(P_0) = P\{T\in\Gamma_\alpha \mid P_0\}.$$

Вероятность ошибки первого рода также называют *уровнем значимости*.

Определение 5А.8.

Вероятностью ошибки второго рода называется функция $\beta(P_1)$, определенная для распределений $P_1 \in \wp_1$:

$$\beta(P_1) = P\{T \notin \Gamma_\alpha \mid P_1\}.$$

Функции вероятностей ошибки первого и второго родов являются основными характеристиками критериев проверки гипотез.

В общем случае, для проверки гипотезы могут быть предложены различные статистические критерии, основанные на различных статистиках T, поэтому необходимо располагать способом сравнения различных критериев, который позволил бы выяснить какой критерий из предложенных является наилучшим. Общепринятым является сравнение критериев на основе специальной характеристики критериев – функции мощности.

Определение 5А.9.

Функцией мощности критерия называется функционал, который для заданного распределения наблюдения $P \in \wp$ равен вероятности события $T \in \Gamma_\alpha$, которая вычисляется при условии, что наблюдение имеет функцию распределения P:

$$W(P) = P\{T \in \Gamma_\alpha \mid P\}.$$

Определение 5А.10.

Мощностью критерия при альтернативе $P_1 \in \wp_1$ называется значение функции мощности $W(P_1)$.

Функция мощности критерия является фундаментальной характеристикой критерия, поскольку отражает способность критерия принимать верные решения: принимать основную гипотезу в том случае, когда она оказывается верной, и отклонять в том случае, когда она оказывается неверной. Действительно, согласно определению функция мощности $W(P)$ равна вероятности отклонения основной гипотезы H_0, при условии наблюдаемые величины имеют распределение P. Если гипотеза H_0 верна, то есть $P^* \in \wp_0$, то значение функции мощности $W(P_0)$ для всех $P_0 \in \wp_0$ (в том числе и для $P^* \in \wp_0$) определяет вероятность отклонения верной гипотезы H_0 (вероятность принять неверное решение), желательно, чтобы эта вероятность была как можно меньше. Если основная гипотеза H_0 неверна и $P^* \in \wp_1$, то значение функции мощности $W(P_1)$ для всех $P_1 \in \wp_1$ (и для $P^* \in \wp_1$) показывает вероятность отклонения критерием неверной гипотезы H_0 (вероятность принять верное решение), желательно, чтобы эта вероятность была как можно больше.

Таким образом, функция мощности «хорошего» критерия:

а) имеет как можно меньшие значения для распределений $P_0 \in \wp_0$ (если истинное распределение наблюдения $P^* \in \wp_0$, то критерий с как можно меньшей вероятностью должен отклонять гипотезу H_0, поскольку она оказывается верной);

б) как можно быстрее возрастает до единицы при удалении от множества $\wp_0$ для распределений $P_1 \in \wp_1$ (если истинное распределение $P^* \notin \wp_0$, то критерий с как можно большей вероятностью должен отклонять гипотезу H_0, поскольку она оказывается неверной).

Определение 5А.11.

Критерий называется *несмещенным*, если мощность критерия при любом альтернативном распределении $P_1 \in \wp_1$ больше мощности критерия любом распределении $P_0 \in \wp_0$, соответствующем основной гипотезе H_0:

$$\forall P_1 \in \wp_1 \forall P_0 \in \wp_0 : W(P_1) > W(P_0).$$

Свойство несмещенности является желательным и говорит о том, что вероятность отклонения гипотезы, когда она неверна, больше вероятности отклонения гипотезы, когда она верна.

Определение 5А.12.

Критерий называется *состоятельным*, если мощность критерия при любом альтернативном распределении $P_1 \in \wp_1$ стремится к 1 с ростом количества наблюдаемых случайных величин n:

$$\forall P_1 \in \wp_1 : \lim_{n\to\infty} W(P_1) = 1 .$$

Наличие свойства состоятельности критерия показывает, что с ростом количества наблюдаемых случайных величин, то есть с увеличением количества поступающей информации, возрастает и способность критерия отклонять основную гипотезу в том случае, если она не верна.

На практике не всегда используют наилучшие в смысле функции мощности критерии, поскольку существенную роль может иметь сложность вычисления критерия. В условиях ограниченного времени, когда решение о том принимается гипотеза или отклоняется нужно сделать за короткий промежуток времени, зачастую применяются менее мощные критерии, но более простые в смысле вычисления.

2. Критерий хи-квадрат проверки простой гипотезы о вероятностях.

Пусть проводится серия независимых испытаний, в каждом из которых происходит в точности одно из событий A_1, A_2, …, A_k (события A_i образуют полную группу событий), имеющих неизвестные вероятности p_1^*, p_2^*, …, p_k^* ($0 < p_i^* < 1$). По результатам серии фиксируется количество ν_1 наступлений события A_1, количество ν_2 наступлений события A_2, и так далее до A_k, так что совокупность наблюдений представляет собой вектор случайных величин $(\nu_1, \nu_2, ..., \nu_k)$, имеющих полиномиальное распределение, которое будем обозначать $\Pi(p_1^*, ..., p_k^*; n)$:

$$P\{\nu_1 = y_1, \nu_2 = y_2, ..., \nu_k = y_k\} = \begin{cases} \dfrac{n!}{y_1! y_2! ... y_k!} p_1^{*y_1} p_2^{*y_2} ... p_k^{*y_k} & , \sum\limits_{i=1}^{k} y_i = n \\ 0 & , \sum\limits_{i=1}^{k} y_i \neq n \end{cases} .$$

Множество всех возможных распределений $\wp$ величин $(\nu_1, \nu_2, ..., \nu_k)$ образовано всеми полиномиальными распределениями:

$$\wp = \left\{ \Pi(p_1, p_2, ..., p_k; n) : \sum_{i=1}^{k} p_i = 1, p_i \geq 0 \right\} .$$

Основная гипотеза H_0 заключается в том, что неизвестные вероятности p_i равны заданным вероятностям p_i^0 ($0 < p_i^0 < 1$):

$$H_0 : p_1^* = p_1^0, \ p_2^* = p_2^0, \ …, \ p_k^* = p_k^0 .$$

Множество $\wp_0$, соответствующее основной гипотезе H_0, образовано единственным распределением $\Pi(p_1^0, ..., p_k^0; n)$:

$$\wp_0 = \{\Pi(p_1^0, ..., p_k^0; n)\} ,$$

поэтому основная гипотеза H_0 является простой.

Альтернативная гипотеза H_1 является сложной, поскольку соответствующее ей множество распределений $\wp_1$ представляет все оставшиеся полиномиальные распределения за вычетом распределения из множества $\wp_0$:

$$\wp_1 = \wp \setminus \wp_0 .$$

Таким образом, альтернативная гипотеза H_1 заключается в том, что нарушается хотя бы одно из равенств, утверждаемое основной гипотезой H_0:

$$H_1 : \exists j : p_j^* \neq p_j^0 .$$

В сформулированных условиях требуется построить статистический критерий проверки основной гипотезы H_0 против альтернативной гипотезы H_1.

Для решения задачи используется критерий хи-квадрат со статистикой критерия $X_n^2(\nu_1,..., \nu_k \mid p_1^0,..., p_k^0)$ следующего вида:

$$X_n^2(\nu_1,..., \nu_k \mid p_1^0,..., p_k^0) = n\sum_{i=1}^{k} \frac{\left(\frac{\nu_i}{n} - p_i^0\right)^2}{p_i^0} = \sum_{i=1}^{k} \frac{n^2\left(\frac{\nu_i}{n} - p_i^0\right)^2}{np_i^0} = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0)^2}{np_i^0} .$$

Статистика X_n^2 отражает «суммарное» отклонение наблюдаемых количеств ν_i наступлений событий A_i, от ожидаемых средних количеств наступлений событий – np_i^0, причем каждое отклонение $(\nu_i - np_i^0)^2$ входит в сумму с «весом» $\frac{1}{np_i^0}$, учитывающим величину гипотетической вероятности p_i^0.

Оказывается, что если величины $(\nu_1,\nu_2,..., \nu_k)$ имеют одно из альтернативных распределений, то статистика X_n^2 с большой вероятностью принимает «большие» значения (утверждение 5А.13). Если же величины $(\nu_1,\nu_2,..., \nu_k)$ имеют распределение $\Pi(p_1^0,..., p_k^0; n)$, то статистика X_n^2 с малой вероятностью принимает «большие» значения (это следует из теоремы 5А.14). Отсюда следует, что в качестве критической области Γ_α гипотезы H_0 следует брать области вида:

$$\Gamma_\alpha = \{x : x = X_n^2(\nu_1,..., \nu_k \mid p_1^0,..., p_k^0) \geq h_\alpha\},$$

где h_α – некоторый порог, поскольку такой вид критической области Γ_α максимизирует вероятности $P\{X_n^2 \in \Gamma_\alpha \mid \Pi_1\}$ для всевозможных альтернативных распределений $\Pi_1 \in \wp_1$.

Остается лишь найти способ вычисления порога h_α для заданного уровня значимости α. По определению уровень значимости есть вероятность:

$$P\{X_n^2 \in \Gamma_\alpha \mid \Pi(p_1^0,..., p_k^0; n)\} = \alpha$$

$$P\{X_n^2 \in \Gamma_\alpha \mid \Pi(p_1^0,..., p_k^0; n)\} = P\{X_n^2 \geq h_\alpha \mid \Pi(p_1^0,..., p_k^0; n)\} = 1 - F_{X_n^2}(h_\alpha \mid \Pi(p_1^0,..., p_k^0; n)) = \alpha ,$$

$$F_{X_n^2}(h_\alpha \mid \Pi(p_1^0,..., p_k^0; n)) = 1 - \alpha . \quad (5А.6)$$

откуда следует, что в качестве порога h_α следует брать квантиль уровня $1-\alpha$ того распределения $F_{X_n^2}(x \mid \Pi(p_1^0,..., p_k^0; n))$ статистики X_n^2, которое получает при условии что величины $(\nu_1,\nu_2,..., \nu_k)$ имеют распределение $\Pi(p_1^0,..., p_k^0; n)$, определяемое основной гипотезой H_0. Точное выражение для функции распределения $F_{X_n^2}(x \mid H_0)$ найти затруднительно, однако, можно показать (теорема 5А.14), что если гипотеза H_0 верна, то функция распределения $F_{X_n^2}(x \mid \Pi(p_1^0,..., p_k^0; n))$ при возрастании n стремится к функции распределения хи-квадрат с $k-1$ степенью свободы, то есть при больших n:

$$P\{X_n^2 < x \mid \Pi(p_1^0,..., p_k^0; n)\} = F_{X_n^2}(x \mid \Pi(p_1^0,..., p_k^0; n)) \approx F_{\chi^2_{k-1}}(x).$$

Таким образом, из (5A.6) получим приближенное равенство для вычисления порога h_α:

$$1-\alpha = F_{X_n^2}(h_\alpha \mid \Pi(p_1^0,..., p_k^0; n)) \approx F_{\chi^2_{k-1}}(h_\alpha),$$

откуда порог h_α приближенно равен квантилю уровня $1-\alpha$ распределения хи-квадрат с $k-1$ степенью свободы – $\chi^2(k-1)$.

Таким образом, проверка гипотезы H_0 сводится к следующей последовательности действий:

1) по заданному уровню значимости α определяется квантиль h_α уровня $1-\alpha$ распределения $\chi^2(k-1)$;

2) по реализации $(y_1,..., y_n)$ величин $(\nu_1, \nu_2,..., \nu_k)$ вычисляется значение статистики $X_n^2(y_1,..., y_n \mid p_1^0,..., p_k^0)$;

3) если $X_n^2(y_1,..., y_n \mid p_1^0,..., p_k^0) \geq h_\alpha$, тогда гипотеза H_0 отклоняется, если $X_n^2(y_1,..., y_n \mid H_0) < h_\alpha$, тогда гипотеза H_0 принимается.

Перейдем к доказательству основных фактов, использованных при формулировке критерия. Прежде всего, покажем, что если наблюдаемые величины $(\nu_1, \nu_2,..., \nu_k)$ имеют какое-либо альтернативное полиномиальное распределение, то значения статистики $X_n^2(\nu_1,..., \nu_n \mid p_1^0,..., p_k^0)$ неограниченно возрастают с ростом n.

Утверждение 5А.13.

Пусть величины $(\nu_1, \nu_2,..., \nu_k)$ имеют полиномиальное распределение $\Pi(p_1,..., p_k; n)$, $0 < p_i < 1$, которое не совпадает с распределением $\Pi(p_1^0,..., p_k^0; n)$, тогда последовательность (по n) случайных величин $X_n^2(\nu_1,..., \nu_n \mid p_1^0,..., p_k^0)$ не ограничена по вероятности, то есть:

$$X_n^2(\nu_1,..., \nu_n \mid p_1^0,..., p_k^0) \xrightarrow{P} \infty, \text{ при } n \to \infty.$$

Доказательство:

Пусть $\delta > 0$ и $\varepsilon > 0$ произвольно выбранные числа, покажем, что найдется N такое, что для всех $n \geq N$:

$$P\{X_n^2(\nu_1,..., \nu_n \mid p_1^0,..., p_k^0) > \varepsilon\} > 1-\delta,$$

это и будет означать, что $X_n^2(\nu_1,..., \nu_n \mid p_1^0,..., p_k^0) \xrightarrow{P} \infty$.

Заметим, что если совокупность величин $(\nu_1, \nu_2,..., \nu_k)$ имеет полиномиальное распределение $\Pi(p_1,..., p_k; n)$, то каждая случайная величина ν_i имеет биномиальное распределение $Bi(n, p_i)$, действительно, для ν_1 получим ($0 \leq y_1 \leq n$):

$$P\{\nu_1 = y_1\} = \sum_{y_2+...+y_k=n-y_1} P\{\nu_1 = y_1, \nu_2 = y_2,..., \nu_k = y_k\} = \sum_{y_2+...+y_k=n-y_1} \frac{n!}{y_1! y_2! ... y_k!} p_1^{y_1} p_2^{y_2} ... p_k^{y_k} =$$

$$= \sum_{y_2+...+y_k=n-y_1} \frac{n(n-1)\cdot ... \cdot(n-y_1+1)(n-y_1)!}{y_1! y_2! ... y_k!} p_1^{y_1} p_2^{y_2} ... p_k^{y_k} =$$

$$= \frac{n(n-1)\cdot ... \cdot(n-y_1+1)}{y_1!} p_1^{y_1} \sum_{y_2+...+y_k=n-y_1} \frac{(n-y_1)!}{y_2! ... y_k!} p_2^{y_2} ... p_k^{y_k} =$$

$$= \frac{n(n-1)\cdot ... \cdot(n-y_1+1)}{y_1!} p_1^{y_1}(p_2 + ... + p_k)^{n-y_1} = \frac{n!}{y_1!(n-y_1)!} p_1^{y_1}(1-p_1)^{n-y_1} = C_n^{y_1} p_1^{y_1}(1-p_1)^{n-y_1}.$$

Очевидно, то же самое может быть проделано и для любого i, поэтому $\nu_i \sim Bi(n, p_i)$ для любого i.

По условию утверждения распределение $\Pi(p_1,..., p_k; n)$ не совпадает с распределением $\Pi(p_1^0,..., p_k^0; n)$, поэтому найдутся такие индексы i, при которых $p_i \neq p_i^0$, пусть j – один из таких индексов, то есть $p_j \neq p_j^0$. Поскольку $\nu_j \sim Bi(n, p_j)$, то в соответствии с теоремой Бернулли:

$$\frac{\nu_j}{n} \xrightarrow{P} p_j, \text{ при } n \to \infty,$$

отсюда следует, что для выбранного ранее δ и для $\varepsilon^* = \frac{|p_j - p_j^0|}{2}$ ($\varepsilon^* > 0$ поскольку $p_j \neq p_j^0$) найдется номер N^* такой, что для всех $n \geq N^*$:

$$P\left\{\omega : \left|\frac{\nu_j(\omega)}{n} - p_j\right| < \varepsilon^*\right\} > 1 - \delta.$$

Отсюда следует, что

$$P\left\{\omega : \left|\frac{\nu_j(\omega)}{n} - p_j^0 + p_j^0 - p_j\right| < \frac{|p_j - p_j^0|}{2}\right\} > 1 - \delta,$$

$$P\left\{\omega : \left|\left(\frac{\nu_j(\omega)}{n} - p_j^0\right) - (p_j - p_j^0)\right| < \frac{|p_j - p_j^0|}{2}\right\} > 1 - \delta,$$

$$P\left\{\omega : (p_j - p_j^0) - \frac{|p_j - p_j^0|}{2} < \left(\frac{\nu_j(\omega)}{n} - p_j^0\right) < (p_j - p_j^0) + \frac{|p_j - p_j^0|}{2}\right\} > 1 - \delta$$

Если $p_j - p_j^0 < 0$, тогда:

$$P\left\{\omega : -|p_j - p_j^0| - \frac{|p_j - p_j^0|}{2} < \left(\frac{\nu_j(\omega)}{n} - p_j^0\right) < -|p_j - p_j^0| + \frac{|p_j - p_j^0|}{2}\right\} > 1 - \delta,$$

$$P\left\{\omega : -\frac{3}{2}|p_j - p_j^0| < \left(\frac{\nu_j(\omega)}{n} - p_j^0\right) < -\frac{1}{2}|p_j - p_j^0|\right\} > 1 - \delta.$$

Если $p_j - p_j^0 > 0$, тогда:

$$P\left\{\omega : |p_j - p_j^0| - \frac{|p_j - p_j^0|}{2} < \left(\frac{\nu_j(\omega)}{n} - p_j^0\right) < |p_j - p_j^0| + \frac{|p_j - p_j^0|}{2}\right\} > 1 - \delta,$$

$$P\left\{\omega : \frac{1}{2}|p_j - p_j^0| < \left(\frac{\nu_j(\omega)}{n} - p_j^0\right) < \frac{3}{2}|p_j - p_j^0|\right\} > 1 - \delta.$$

В том и другом случаях:

$$P\left\{\omega : \frac{1}{2}|p_j - p_j^0| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right| < \frac{3}{2}|p_j - p_j^0|\right\} > 1 - \delta.$$

Из вложенности событий:

$$\left\{\omega : \frac{1}{2}|p_j - p_j^0| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right| < \frac{3}{2}|p_j - p_j^0|\right\} \subseteq \left\{\omega : \frac{1}{2}|p_j - p_j^0| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right|\right\},$$

следует неравенство для вероятностей событий:

$$1-\delta < P\left\{\omega : \frac{1}{2}\left|p_j - p_j^0\right| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right| < \frac{3}{2}\left|p_j - p_j^0\right|\right\} \le P\left\{\omega : \frac{1}{2}\left|p_j - p_j^0\right| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right|\right\}.$$

Пусть $C(n)$ есть событие:

$$C(n) = \left\{\omega : \frac{1}{2}\left|p_j - p_j^0\right| < \left|\frac{\nu_j(\omega)}{n} - p_j^0\right|\right\},$$

тогда $P\{C(n)\} > 1-\delta$, и для произвольного $\omega \in C(n)$:

$$X_n^2(\nu_1(\omega),\ldots,\nu_k(\omega) \mid p_1^0,\ldots,p_k^0) = n\sum_{i=1}^{k}\frac{\left(\frac{\nu_i(\omega)}{n} - p_i^0\right)^2}{p_i^0} \ge n\frac{\left(\frac{\nu_j(\omega)}{n} - p_j^0\right)^2}{p_j^0} > n\frac{\frac{1}{4}(p_j - p_j^0)^2}{p_j^0}.$$

Пусть $N > \max\left\{\frac{\varepsilon}{\frac{\frac{1}{4}(p_j - p_j^0)^2}{p_j^0}}, N^*\right\}$, тогда для $n \ge N$:

$$P\{C(n)\} > 1-\delta,$$

$$\omega \in C(n) : X_n^2(\nu_1(\omega),\ldots,\nu_k(\omega) \mid H_0) > n\frac{\frac{1}{4}(p_j - p_j^0)^2}{p_j^0} \ge N\frac{\frac{1}{4}(p_j - p_j^0)^2}{p_j^0} > \varepsilon.$$

Отсюда следует, что при $n \ge N$:

$$\{\omega : X_n^2(\nu_1(\omega),\ldots,\nu_k(\omega) \mid p_1^0,\ldots,p_k^0) > \varepsilon\} \supseteq C(n),$$

тогда

$$P\{\omega : X_n^2(\nu_1(\omega),\ldots,\nu_k(\omega) \mid p_1^0,\ldots,p_k^0) > \varepsilon) \ge P(C(n)) > 1-\delta.$$

Таким образом, для произвольных δ и ε найден способ определения числа N такого, что для всех $n \ge N$:

$$P\{\omega : X_n^2(\nu_1(\omega),\ldots,\nu_k(\omega) \mid p_1^0,\ldots,p_k^0) > \varepsilon) > 1-\delta.$$

Утверждение доказано.

Для того, чтобы статистику X_n^2 можно было использовать в качестве статистики критерия необходимо найти способ вычисления (хотя бы приближенного) значений функции распределения статистики X_n^2 при условии что величины $(\nu_1,\nu_2,\ldots,\nu_k)$ имеют распределение $\Pi(p_1^0,\ldots,p_k^0;n)$, определяемое основной гипотезой H_0.

Теорема 5А.14. (Пирсон)

Пусть величины $(\nu_1,\nu_2,\ldots,\nu_k)$ имеют полиномиальное распределение $\Pi(p_1^0,\ldots,p_k^0;n)$, тогда распределение статистики $X_n^2(\nu_1,\ldots,\nu_k \mid p_1^0,\ldots,p_k^0)$ стремится к распределению хи-квадрат с $k-1$ степенью свободы:

$$X_n^2(\nu_1,\ldots,\nu_k \mid p_1^0,\ldots,p_k^0) = \sum_{i=1}^{k}\frac{(\nu_i - np_i^0)^2}{np_i^0} \sim \chi^2(k-1),\ \text{при}\ n \to \infty.$$

Доказательство:

Преобразуем статистику $X_n^2(\nu_1,\ldots,\nu_k \mid p_1^0,\ldots,p_k^0)$ следующим образом:

$$X_n^2(\nu_1,\ldots,\nu_k \mid p_1^0,\ldots,p_k^0) = \sum_{i=1}^{k}\frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k}\left(\frac{\nu_i - np_i^0}{\sqrt{np_i^0}}\right)^2 = \nu^*\nu,$$

где вектор-столбец ν :

$$\nu = \begin{pmatrix} \dfrac{\nu_1 - np_1^0}{\sqrt{np_1^0}} \\ \dfrac{\nu_2 - np_2^0}{\sqrt{np_2^0}} \\ \ldots \\ \dfrac{\nu_k - np_k^0}{\sqrt{np_k^0}} \end{pmatrix},$$

и ν^* – транспонированный вектор ν .

Представим, что исходными наблюдаемыми величинами являются не величины $(\nu_1, \nu_2, ..., \nu_k)$, а выборка $(\xi_1, ..., \xi_n)$ объема n, в которой каждая случайная величина ξ_i отражает исход i-го испытания и принимает значения 1, 2, …, k в зависимости от того, событие с каким номером наступило в i-ом испытании:

$$\xi_i = \begin{cases} 1, & p_1^0 \\ 2, & p_2^0 \\ \ldots \\ k, & p_k^0 \end{cases}.$$

Пусть $I(\xi_i, j)$ – бинарная случайная величина:

$$I(\xi_i, j) = \begin{cases} 0 & , \xi_i \neq j \\ 1 & , \xi_i = j \end{cases}.$$

Заметим, что математическое ожидание $M[I(\xi_i, j)] = 1 \cdot P\{\xi_i = j\} = p_j^0$, и кроме того, легко видеть, что $\nu_j = \sum_{i=1}^{n} I(\xi_i, j)$, тогда:

$$\nu = \begin{pmatrix} \dfrac{\nu_1 - np_1^0}{\sqrt{np_1^0}} \\ \dfrac{\nu_2 - np_2^0}{\sqrt{np_2^0}} \\ \ldots \\ \dfrac{\nu_k - np_k^0}{\sqrt{np_k^0}} \end{pmatrix} = \begin{pmatrix} \dfrac{\sum_{i=1}^{n} I(\xi_i,1) - np_1^0}{\sqrt{np_1^0}} \\ \dfrac{\sum_{i=1}^{n} I(\xi_i,2) - np_2^0}{\sqrt{np_2^0}} \\ \ldots \\ \dfrac{\sum_{i=1}^{n} I(\xi_i,k) - np_k^0}{\sqrt{np_k^0}} \end{pmatrix} = \begin{pmatrix} \dfrac{\sum_{i=1}^{n} \dfrac{I(\xi_i,1)}{\sqrt{p_1^0}} - n\sqrt{p_1^0}}{\sqrt{n}} \\ \dfrac{\sum_{i=1}^{n} \dfrac{I(\xi_i,2)}{\sqrt{p_2^0}} - n\sqrt{p_2^0}}{\sqrt{n}} \\ \ldots \\ \dfrac{\sum_{i=1}^{n} \dfrac{I(\xi_i,k)}{\sqrt{p_k^0}} - n\sqrt{p_k^0}}{\sqrt{n}} \end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{i=1}^{n} \dfrac{I(\xi_i,1)}{\sqrt{p_1^0}} \\ \sum_{i=1}^{n} \dfrac{I(\xi_i,2)}{\sqrt{p_2^0}} \\ \ldots \\ \sum_{i=1}^{n} \dfrac{I(\xi_i,k)}{\sqrt{p_k^0}} \end{pmatrix} - \frac{n}{\sqrt{n}} \begin{pmatrix} \sqrt{p_1^0} \\ \sqrt{p_2^0} \\ \ldots \\ \sqrt{p_k^0} \end{pmatrix} =$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \dfrac{I(\xi_i,1)}{\sqrt{p_1^0}} \\ \dfrac{I(\xi_i,2)}{\sqrt{p_2^0}} \\ \ldots \\ \dfrac{I(\xi_i,k)}{\sqrt{p_k^0}} \end{pmatrix} - \frac{n}{\sqrt{n}} \begin{pmatrix} \sqrt{p_1^0} \\ \sqrt{p_2^0} \\ \ldots \\ \sqrt{p_k^0} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi^{(i)} - \frac{n}{\sqrt{n}} P = \frac{\sum_{i=1}^{n} \xi^{(i)} - nP}{\sqrt{n}},$$

где $\xi^{(i)}$ – вектор-столбец случайных величин и P – вектор столбец:

$$\xi^{(i)}=\begin{pmatrix}\dfrac{I(\xi_i,1)}{\sqrt{p_1^0}}\\ \dfrac{I(\xi_i,2)}{\sqrt{p_2^0}}\\ \dots\\ \dfrac{I(\xi_i,k)}{\sqrt{p_k^0}}\end{pmatrix},\quad P=\begin{pmatrix}\sqrt{p_1^0}\\ \sqrt{p_2^0}\\ \dots\\ \sqrt{p_k^0}\end{pmatrix}.$$

Таким образом, статистика $X_n^2(\nu_1,\dots,\nu_k \mid p_1^0,\dots,p_k^0)$:

$$X_n^2(\nu_1,\dots,\nu_k \mid p_1^0,\dots,p_k^0)=\nu^*\nu=\begin{pmatrix}\dfrac{\sum\limits_{i=1}^{n}\xi^{(i)}-nP}{\sqrt{n}}\end{pmatrix}^*\begin{pmatrix}\dfrac{\sum\limits_{i=1}^{n}\xi^{(i)}-nP}{\sqrt{n}}\end{pmatrix}.$$

Поскольку все случайные величины выборки ξ_i имеют одинаковое распределение, то все векторы $\xi^{(i)}$ $(i=\overline{1,n})$ имеют одинаковые моменты. Вычислим математическое ожидание $M[\xi^{(i)}]$:

$$M[\xi^{(i)}]=\begin{pmatrix}\dfrac{M[I(\xi_i,1)]}{\sqrt{p_1^0}}\\ \dfrac{M[I(\xi_i,2)]}{\sqrt{p_2^0}}\\ \dots\\ \dfrac{M[I(\xi_i,k)]}{\sqrt{p_k^0}}\end{pmatrix}=\begin{pmatrix}\dfrac{p_1^0}{\sqrt{p_1^0}}\\ \dfrac{p_2^0}{\sqrt{p_2^0}}\\ \dots\\ \dfrac{p_k^0}{\sqrt{p_k^0}}\end{pmatrix}=\begin{pmatrix}\sqrt{p_1^0}\\ \sqrt{p_2^0}\\ \dots\\ \sqrt{p_k^0}\end{pmatrix}=P.$$

Вычислим дисперсионную матрицу $D[\xi^{(i)}]$:

$$\left\|D[\xi^{(i)}]\right\|_{l,m}=\operatorname{cov}\left(\frac{I(\xi_i,l)}{\sqrt{p_l^0}},\frac{I(\xi_i,m)}{\sqrt{p_m^0}}\right)=\frac{\operatorname{cov}(I(\xi_i,l),I(\xi_i,m))}{\sqrt{p_l^0p_m^0}}=$$

$$=\frac{M[(I(\xi_i,l)-M[I(\xi_i,l)])(I(\xi_i,m)-M[I(\xi_i,m)])]}{\sqrt{p_l^0p_m^0}}=$$

$$=\frac{M[(I(\xi_i,l)-p_l^0])(I(\xi_i,m)-p_m^0)]}{\sqrt{p_l^0p_m^0}}=\frac{M[I(\xi_i,l)I(\xi_i,m)]-p_l^0p_m^0}{\sqrt{p_l^0p_m^0}}.$$

Если $l\neq m$, то $I(\xi_i,l)I(\xi_i,m)=0$, поскольку случайная величина ξ_i не может принимать два различных значения l и m одновременно, и следовательно $M[I(\xi_i,l)I(\xi_i,m)]=0$. Если $l=m$, тогда $I(\xi_i,l)I(\xi_i,m)=(I(\xi_i,l))^2$, тогда:

$$M[I(\xi_i,l)I(\xi_i,m)]=M[(I(\xi_i,l))^2]=1^2\cdot P\{\xi_i=l\}=p_l^0.$$

Таким образом,

$$\left\|D[\xi^{(i)}]\right\|_{l,m}=\frac{M[I(\xi_i,l)I(\xi_i,m)]-p_l^0p_m^0}{\sqrt{p_l^0p_m^0}}=\begin{cases}\dfrac{p_l^0-p_l^0p_l^0}{\sqrt{p_l^0p_l^0}} & ,l=m\\ \dfrac{-p_l^0p_m^0}{\sqrt{p_l^0p_m^0}} & ,l\neq m\end{cases}=\begin{cases}1-\sqrt{p_l^0p_l^0} & ,l=m\\ -\sqrt{p_l^0p_m^0} & ,l\neq m\end{cases}.$$

Отсюда следует, что дисперсионную матрицу можно представить в виде:

$$D[\xi^{(i)}]=\begin{pmatrix} 1-\sqrt{p_1^0p_1^0} & -\sqrt{p_1^0p_2^0} & \dots & -\sqrt{p_1^0p_k^0} \\ -\sqrt{p_2^0p_1^0} & 1-\sqrt{p_2^0p_2^0} & \dots & -\sqrt{p_2^0p_k^0} \\ \dots & \dots & \dots & \dots \\ -\sqrt{p_k^0p_1^0} & -\sqrt{p_k^0p_2^0} & \dots & 1-\sqrt{p_k^0p_k^0} \end{pmatrix}=$$

$$=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}-\begin{pmatrix} -\sqrt{p_1^0p_1^0} & -\sqrt{p_1^0p_2^0} & \dots & -\sqrt{p_1^0p_k^0} \\ -\sqrt{p_2^0p_1^0} & -\sqrt{p_2^0p_2^0} & \dots & -\sqrt{p_2^0p_k^0} \\ \dots & \dots & \dots & \dots \\ -\sqrt{p_k^0p_1^0} & -\sqrt{p_k^0p_2^0} & \dots & -\sqrt{p_k^0p_k^0} \end{pmatrix}=$$

$$=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}-\begin{pmatrix} \sqrt{p_1^0} \\ \sqrt{p_2^0} \\ \dots \\ \sqrt{p_k^0} \end{pmatrix}\cdot\left(\sqrt{p_1^0} \quad \sqrt{p_2^0} \quad \dots \quad \sqrt{p_k^0}\right)=E_k-PP^*,$$

где E_k – единичная матрица порядка $k\times k$, P^* – транспонированный вектор P.

Как и ожидалось, дисперсионная матрица $D[\xi^{(i)}]$ является вырожденной. Действительно, что если дисперсионную матрицу умножить на вектор P, то получится нулевой вектор $\overline{0}$:

$$D[\xi^{(i)}]P=(E_k-PP^*)P=E_kP-PP^*P=P-PP^*P.$$

Легко видеть, что

$$P^*P=\left(\sqrt{p_1^0} \quad \sqrt{p_2^0} \quad \dots \quad \sqrt{p_k^0}\right)\begin{pmatrix} \sqrt{p_1^0} \\ \sqrt{p_2^0} \\ \dots \\ \sqrt{p_k^0} \end{pmatrix}=\sum_{i=1}^{k}\sqrt{p_i^0p_i^0}=\sum_{i=1}^{k}p_i^0=1,$$

тогда

$$D[\xi^{(i)}]P=P-PP^*P=P-P=\overline{0}.$$

Если бы матрица $D[\xi^{(i)}]$ была невырожденной, то равенство $D[\xi^{(i)}]\overline{a}=\overline{0}$ с некоторым вектором $\overline{a}$ выполнялось бы только в случае $\overline{a}=\overline{0}$, то есть не могло бы существовать ни одного ненулевого вектора $\overline{a}$, при котором выполнялось бы равенство $D[\xi^{(i)}]\overline{a}=\overline{0}$. Однако, найден ненулевой вектор P такой, что $D[\xi^{(i)}]P=\overline{0}$, тогда матрица $D[\xi^{(i)}]$ обязательно вырождена. Поскольку все векторы $\xi^{(i)}$ имеют вырожденную дисперсионную матрицу, то к сумме $\sum_{i=1}^{n}\xi^{(i)}$ не применима центральная предельная теорема для многомерного случая.

Преобразуем векторы $\xi^{(i)}$ в векторы $\eta^{(i)}$ с помощью ортогонального преобразования с матрицей Q (Q^* – транспонированная матрица Q):

$$\eta^{(i)}=Q\xi^{(i)},$$

$$Q^{-1}=Q^*,$$

тогда статистика $X_n^2(\nu_1,\dots,\nu_k\mid p_1^0,\dots,p_k^0)$ преобразуется к следующему виду:

$$X_n^2(\nu_1,\dots,\nu_k \mid p_1^0,\dots,p_k^0) = \left(\frac{\sum_{i=1}^{n}\xi^{(i)} - nP}{\sqrt{n}}\right)^* \left(\frac{\sum_{i=1}^{n}\xi^{(i)} - nP}{\sqrt{n}}\right) =$$

$$= \left(\frac{\sum_{i=1}^{n} Q^*\eta^{(i)} - nQ^{-1}QP}{\sqrt{n}}\right)^* \left(\frac{\sum_{i=1}^{n} Q^*\eta^{(i)} - nQ^{-1}QP}{\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^{n} Q^*\eta^{(i)} - nQ^*QP}{\sqrt{n}}\right)^* \left(\frac{\sum_{i=1}^{n} Q^*\eta^{(i)} - nQ^*QP}{\sqrt{n}}\right) =$$

$$= \left(Q^*\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right)^* \left(Q^*\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right)^* (Q^*)^* Q^* \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right) =$$

$$= \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right)^* QQ^{-1} \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right)^* \left(\frac{\sum_{i=1}^{n}\eta^{(i)} - nQP}{\sqrt{n}}\right).$$

Поскольку векторы $\xi^{(i)}$ имеют одинаковые математические ожидания и дисперсионные матрицы, то и векторы $\eta^{(i)}$ также имеют одинаковые математические ожидания и дисперсионные матрицы. Математическое ожидание $M[\eta^{(i)}]$:

$$M[\eta^{(i)}] = M[Q\xi^{(i)}] = QM[\xi^{(i)}] = QP. \tag{5А.7}$$

Дисперсионная матрица $D[\eta^{(i)}]$:

$$D[\eta^{(i)}] = M[(\eta^{(i)} - M[\eta^{(i)}])(\eta^{(i)} - M[\eta^{(i)}])^*] = M[(Q\xi^{(i)} - M[Q\xi^{(i)}])(Q\xi^{(i)} - M[Q\xi^{(i)}])^*] =$$
$$= M[(Q\xi^{(i)} - QM[\xi^{(i)}])(Q\xi^{(i)} - QM[\xi^{(i)}])^*] = M[Q(\xi^{(i)} - M[\xi^{(i)}])(Q(\xi^{(i)} - M[\xi^{(i)}]))^*] =$$
$$= M[Q(\xi^{(i)} - M[\xi^{(i)}])(\xi^{(i)} - M[\xi^{(i)}])^* Q^*] = QM[(\xi^{(i)} - M[\xi^{(i)}])(\xi^{(i)} - M[\xi^{(i)}])^*]Q^* = QD[\xi^{(i)}]Q^* =$$
$$= Q(E_k - PP^*)Q^* = QE_kQ^* - QPP^*Q^* = QQ^* - QP(QP)^* = E_k - QP(QP)^*.$$

Дисперсионная матрица $D[\eta^{(i)}]$ оказывается «почти единичной». Представим, что в матрице Q последняя строка совпадает с транспонированными вектором-столбцом P:

$$Q = \begin{pmatrix} Q_{k-1,k} \\ P^* \end{pmatrix},$$

где $Q_{k-1,k}$ – матрица порядка $(k-1)\times k$. Поскольку Q – ортогональная матрица, то её строки являются взаимно ортогональными векторами, отсюда следует, что строки матрицы $Q_{k-1,k}$ являются взаимно ортогональными векторами, которые к тому же ортогональны и вектору P^*, тогда:

$$QP = \begin{pmatrix} Q_{k-1,k} \\ P^* \end{pmatrix} P = \begin{pmatrix} Q_{k-1,k}P \\ P^*P \end{pmatrix} = \begin{pmatrix} \overline{0}_{k-1} \\ 1 \end{pmatrix}. \tag{5А.8}$$

где $\overline{0}_{k-1}$ – нулевой вектор-столбец порядка $k-1$, и следовательно:

$$QP\,(QP\,)^* = \begin{pmatrix} \overline{0}_{k-1} \\ 1 \end{pmatrix} \begin{pmatrix} \overline{0}^*_{k-1} & 1 \end{pmatrix} = \begin{pmatrix} \overline{0}_{k-1}\overline{0}^*_{k-1} & \overline{0}_{k-1}\cdot 1 \\ 1\cdot \overline{0}^*_{k-1} & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

то есть $QP\,(QP\,)^*$ – матрица все элементы, которой равны нулю, кроме элемента в k -ой строке и k -ом столбце, который равен 1. Таким образом, дисперсионная матрица D_η :

$$D[\eta^{(i)}] = E_k - QP\,(QP\,)^* = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} =$$
$$= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} E_{k-1} & 0 \\ 0 & 0 \end{pmatrix}. \qquad (5A.9)$$

где E_{k-1} единичная матрица порядка $(k-1)\times(k-1)$.

Заметим, что

$$\eta^{(i)} = Q\,\xi^{(i)} = \begin{pmatrix} Q_{k-1,k} \\ P^* \end{pmatrix} \xi^{(i)} = \begin{pmatrix} Q_{k-1,k}\,\xi^{(i)} \\ P^*\xi^{(i)} \end{pmatrix} = \begin{pmatrix} Q_{k-1,k}\,\xi^{(i)} \\ 1 \end{pmatrix},$$

поскольку,

$$P^*\xi^{(i)} = \begin{pmatrix} \sqrt{p_1^0} & \sqrt{p_2^0} & \dots & \sqrt{p_k^0} \end{pmatrix} \begin{pmatrix} \dfrac{I(\xi_i,1)}{\sqrt{p_1^0}} \\ \dfrac{I(\xi_i,2)}{\sqrt{p_2^0}} \\ \dots \\ \dfrac{I(\xi_i,k)}{\sqrt{p_k^0}} \end{pmatrix} = \sum_{j=1}^{k} \sqrt{p_j^0}\,\frac{I(\xi_i,\,j)}{\sqrt{p_j^0}} = \sum_{j=1}^{k} I(\xi_i,\,j) = 1\,,$$

в силу того, что случайная величина ξ_i принимает одно из целых значений от 1 до k , так что в сумме $\sum_{j=1}^{k} I(\xi_i,\,j)$ обязательно в точности одно слагаемое будет равно 1 и остальные будут равны 0.

Пусть $\eta^{(i)} = Q_{k-1,k}\,\xi^{(i)}$, тогда $\eta^{(i)} = \begin{pmatrix} Q_{k-1,k}\,\xi^{(i)} \\ 1 \end{pmatrix} = \begin{pmatrix} \eta^{(i)} \\ 1 \end{pmatrix}$, причем из (5A.7) и (5A.8) следует:

$$M\,[\eta^{(i)}] = \begin{pmatrix} M\,[\eta^{(i)}] \\ 1 \end{pmatrix} = QP\; = \begin{pmatrix} \overline{0}_{k-1} \\ 1 \end{pmatrix}. \qquad (5A.10)$$

и в силу (5A.9),

$$D[\eta^{(i)}] = \begin{pmatrix} D[\eta^{(i)}] & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} E_{k-1} & 0 \\ 0 & 0 \end{pmatrix}$$

$$D[\eta^{(i)}] = E_{k-1}\,. \qquad (5A.11)$$

Заметим, что

$$\frac{\sum_{i=1}^{n}\eta^{(i)}-nQP}{\sqrt{n}}=\frac{\sum_{i=1}^{n}\begin{pmatrix}\eta^{(i)}\\1\end{pmatrix}-n\begin{pmatrix}\overline{0}_{k-1}\\1\end{pmatrix}}{\sqrt{n}}=\frac{\begin{pmatrix}\sum_{i=1}^{n}\eta^{(i)}\\n\end{pmatrix}-\begin{pmatrix}\overline{0}_{k-1}\\n\end{pmatrix}}{\sqrt{n}}=\frac{\begin{pmatrix}\sum_{i=1}^{n}\eta^{(i)}\\0\end{pmatrix}}{\sqrt{n}},$$

тогда статистика $X_n^2(\nu_1,...,\nu_k \mid p_1^0,...,p_k^0)$:

$$X_n^2(\nu_1,...,\nu_k \mid p_1^0,...,p_k^0)=\left(\frac{\sum_{i=1}^{n}\eta^{(i)}-nQP}{\sqrt{n}}\right)^*\left(\frac{\sum_{i=1}^{n}\eta^{(i)}-nQP}{\sqrt{n}}\right)=$$

$$=\frac{\begin{pmatrix}\sum_{i=1}^{n}\eta^{(i)}\\0\end{pmatrix}^*\begin{pmatrix}\sum_{i=1}^{n}\eta^{(i)}\\0\end{pmatrix}}{\sqrt{n}\sqrt{n}}=\frac{\left(\sum_{i=1}^{n}\eta^{(i)}\right)^*\left(\sum_{i=1}^{n}\eta^{(i)}\right)}{\sqrt{n}\sqrt{n}}=\left(\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}\right)^*\left(\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}\right). \tag{5A.12}$$

Векторы $\eta^{(i)}$ имеют одинаковые распределения (поскольку векторы $\xi^{(i)}$ и следовательно $\eta^{(i)}$ имеют одинаковые распределения), математические ожидание $M[\eta^{(i)}]=\overline{0}_{k-1}$ (5A.10) и невырожденные дисперсионные матрицы $D[\eta^{(i)}]=E_{k-1}$ (5A.11), поэтому к последовательности случайных величин $\eta^{(i)}$ применима центральная предельная теорема для многомерного случая, согласно которой нормированная сумма $\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}$ имеет асимптотически многомерное нормальное распределение $N(\overline{0}_{k-1},E_{k-1})$:

$$\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}\sim N(\overline{0}_{k-1},E_{k-1}),\text{ при } n\to\infty .$$

Пусть вектор-столбец случайных величин $\zeta=(\zeta_1,...,\zeta_{k-1})\sim N(\overline{0}_{k-1},E_{k-1})$, поскольку распределение вектора $\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}$ стремится к распределению случайной величины ζ, то распределение случайной величины $\left(\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}\right)^*\left(\frac{\sum_{i=1}^{n}\eta^{(i)}}{\sqrt{n}}\right)$ стремится к распределению случайной величины $\zeta^*\zeta$. Таким образом, из (5A.12) распределение статистики $X_n^2(\nu_1,...,\nu_k \mid p_1^0,...,p_k^0)$ стремится к распределению суммы квадратов $\zeta^*\zeta$:

$$\zeta^*\zeta=\begin{pmatrix}\zeta_1 & ... & \zeta_{k-1}\end{pmatrix}\begin{pmatrix}\zeta_1\\...\\\zeta_{k-1}\end{pmatrix}=\sum_{i=1}^{k-1}\zeta_i^2 .$$

Взятые по отдельности случайные величины $\zeta_i\sim N(0,1)$ имеют нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, и кроме того независимы поскольку являются некоррелированными (дисперсионная матрица $D[\zeta]=E_{k-1}$ является единичной, так что все ковариации $\mathrm{cov}(\zeta_i, \zeta_j) = 0$ при $i \neq j$) нормальными случайными величинами. Отсюда следует, что случайная величина $\sum_{i=1}^{k-1} \zeta_i^2$ имеет распределение $\chi^2(k-1)$, тогда и статистика $X_n^2(\nu_1, ..., \nu_k \mid p_1^0, ..., p_k^0)$ при $n \to \infty$ имеет распределение $\chi^2(k-1)$:

$$X_n^2(\nu_1, ..., \nu_k \mid p_1^0, ..., p_k^0) \sim \chi^2(k-1), \text{ при } n \to \infty.$$

Теорема доказана.

Следующее утверждение показывает, что критерий хи-квадрат является состоятельным.

Утверждение 5А.15.

Пусть величины $(\nu_1, ..., \nu_k)$ имеют полиномиальное распределение $\Pi(p_1^0, ..., p_k^0; n)$, тогда при всяком альтернативном распределении $\Pi(p_1, ..., p_k; n) \in \wp_1$ значение функции мощности $W(\Pi(p_1, ..., p_k; n))$:

$$W(\Pi(p_1, ..., p_k; n)) = P\{X_n^2(\nu_1, ..., \nu_k \mid p_1^0, ..., p_k^0) \geq h_\alpha \mid \Pi(p_1, ..., p_k; n)\}$$

стремится к 1 при $n \to \infty$:

$$\lim_{n \to \infty} W(\Pi(p_1, ..., p_k; n)) = 1.$$

Без доказательства.

Ранее было показано, что если гипотеза H_0 верна, то распределение статистики X_n^2 при увеличении n стремится к распределению $\chi^2(k-1)$, можно также установить, что если гипотеза H_0 не верна, то распределение статистики X_n^2 при увеличении n стремится к нецентральному распределению хи-квадрат с $k-1$ степенью свободы и параметром нецентральности a – $\chi^2(k-1, a)$.

Случайная величина $\chi_k^2(a)$ имеет нецентральное распределение $\chi^2(k, a)$, если:

$$\chi_k^2(a) = \sum_{i=1}^{k} \alpha_i^2,$$

где α_i – совместно независимые случайные величины, $\alpha_i \sim N(a_i, 1)$ и $a = \sum_{i=1}^{k} a_i^2$, при этом плотность вероятности случайной величины $\chi_k^2(a)$ зависит только от величины $a = \sum_{i=1}^{k} a_i^2$, но не по отдельности от $a_1, \ldots, a_k$.

Утверждение 5А.16.

Пусть величины $(\nu_1, \nu_2, ..., \nu_k)$ имеют полиномиальное распределение $\Pi(p_1, ..., p_k; n)$ ($0 < p_i < 1$), тогда распределение статистики $X_n^2(\nu_1, ..., \nu_k \mid p_1^0, ..., p_k^0)$:

$$X_n^2(\nu_1, ..., \nu_k \mid p_1^0, ..., p_k^0) = n \sum_{i=1}^{k} \frac{\left(\frac{\nu_i}{n} - p_i^0\right)^2}{p_i^0}$$

стремится при $n \to \infty$ к нецентральному распределению $\chi^2\left(k-1, n\sum_{i=1}^{k} \frac{(p_i - p_i^0)^2}{p_i^0}\right)$.

Без доказательства.

Условия применимости на практике.

Поскольку известно только предельное (при $n \to \infty$) распределение статистики X_n^2 (теорема 5А.14), то для конечного n использование распределения $\chi^2(k-1)$ в качестве распределения χ^2_n является приближенным. Замечено, что «хорошее» приближение достигается в тех случаях, когда все произведения ($i = \overline{1,k}$),

$$np_i^0 \geq 5 .$$

Проверка гипотезы о распределении полностью известном.

Рассмотрим следующую задачу проверки гипотезы: пусть $(\xi_1,..., \xi_n)$ – выборка из неизвестного распределения $F_\xi(x)$, основная гипотеза H_0 заключается в том, что $F_\xi(x) = F_0(x)$, где $F_0(x)$ – известная функция распределения, а альтернативная гипотеза H_1 – в том, что $F_\xi(x) \neq F_0(x)$. Требуется предложить критерий проверки освноной гипотезы H_0 против альтернативной H_1.

Воспользоваться критерием хи-квадрат для решения непосредственно поставленной задачи невозможно, тем не менее, имеется возможность сформулировать «близкую» к поставленной задачу, для решения которой использовать критерий хи-квадрат.

Пусть $x_1 < x_2 < ... < x_{k-2} < x_{k-1}$ некоторые числа, рассмотрим разбиение числовой оси на интервалы и полуинтервалы:

$$L_1 = (-\infty; x_1),$$
$$L_2 = [x_1; x_2),$$
$$\ldots,$$
$$L_{k-1} = [x_{k-2}; x_{k-1}),$$
$$L_k = [x_{k-1}; \infty).$$

Зафиксируем некоторый номер i и определим события,

$$A_1 = \{\omega : \xi_i(\omega) \in L_1\},$$
$$A_2 = \{\omega : \xi_i(\omega) \in L_2\},$$
$$\ldots,$$
$$A_k = \{\omega : \xi_i(\omega) \in L_k\}.$$

Легко видеть, что события A_1, A_2, …, A_k вообще говоря при всех i одинаковы, поскольку все случайные величины ξ_i выборки одинаковы (имеют одну и ту же функцию распределения $F_\xi(x)$), и кроме того образуют полную группу событий, поскольку несовместны и их объединение есть множество всех элементарных событий. Определим вероятности p_1, p_2, …, p_k событий A_1, A_2, …, A_k:

$$p_1 = P(A_1) = P\{\xi_i < x_1\} = F_\xi(x_1),$$
$$p_2 = P(A_2) = P\{x_1 \leq \xi_i < x_2\} = F_\xi(x_2) - F_\xi(x_1),$$
$$\ldots,$$
$$p_k = P(A_k) = P\{\xi_i > x_{k-1}\} = 1 - F_\xi(x_k).$$

Рисунок 5А.1. Разбиение и вероятности.

Из исходной выборки $(\xi_1,..., \xi_n)$ – сформируем вектор $(\nu_1,..., \nu_k)$ по правилу:

$$\nu_j = \sum_{i=1}^{n} I(\xi_i, L_j),$$

$$I(\xi_i, L_j) = \begin{cases} 1 & ,\xi_i \in L_j \\ 0 & ,\xi_i \notin L_j \end{cases},$$

то есть ν_j – случайное количество величин выборки $(\xi_1,..., \xi_n)$ попавших в интервал (полуинтервал) L_j.

В качестве основной гипотезы рассмотрим «расширенную» гипотезу $\tilde{H}_0$:

$$p_i = p_i^0,\ i = \overline{1,k},$$
$$p_1^0 = F_0(x_1),$$
$$p_2^0 = F_0(x_2) - F_0(x_1), \qquad (5A.13)$$
$$\ldots,$$
$$p_k^0 = 1 - F_0(x_k).$$
А в качестве альтернативной – расширенную гипотезу $\tilde{H}_1$:
$$\tilde{H}_1 : \exists j : p_j \neq p_j^0.$$
Теперь для проверки «расширенной» гипотезы $\tilde{H}_0$ против альтернативной $\tilde{H}_1$ может быть использован критерий хи-квадрат, рассмотренный выше.

Из (5А.13) следует, что гипотеза $\tilde{H}_0$ заключается в том, что:
$$\begin{matrix} F_\xi(x_1) = F_0(x_1) & & F_\xi(x_1) = F_0(x_1) \\ F_\xi(x_2) - F_\xi(x_1) = F_0(x_2) - F_0(x_1) & \Leftrightarrow & F_\xi(x_2) = F_0(x_2) \\ \ldots & & \ldots \\ 1 - F_\xi(x_k) = 1 - F_0(x_k) & & F_\xi(x_k) = F_0(x_k) \end{matrix}$$

Таким образом, «расширенная» гипотеза $\tilde{H}_0$ утверждает, что $F_\xi(x_i) = F_0(x_i)$ только для точек x_i, а гипотеза H_0 утверждает, что $F_\xi(x) = F_0(x)$ для всех x, поэтому H_0 и $\tilde{H}_0$, вообще говоря, различные гипотезы. Фактически, $\tilde{H}_0$ утверждает, что истинное распределение $F_\xi(x)$ принадлежит некоторому множеству G_0:
$$\tilde{H}_0 : F_\xi(x) \in G_0,$$
где G_0 – множество таких функций распределения $F(x)$, что $F_\xi(x_i) = F(x_i)$:
$$G_0 = \{F(x) : F_\xi(x_i) = F(x_i), i = \overline{1, k-1}\}.$$
Конечно, $F_0(x) \in G_0$, однако, в G_0 могут оказаться и другие функции $F(x)$, отличные от $F_0(x)$, поэтому гипотеза $\tilde{H}_0$ «расширенная».

Остается вопрос о выборе точек x_0, …, x_{k-1}, которые определяют интервалы и события A_1, …, A_k: на практике количество точек выбирают так чтобы,
$$np_i^0 \geq 5,$$
при этом местоположение точек выбирают так, чтобы все гипотетические вероятности p_i^0 оказались приближенно равны между собой:
$$p_1^0 \approx p_2^0 \approx \ldots \approx p_k^0.$$

3. Критерий хи-квадрат проверки сложной гипотезы о вероятностях.

Пусть проводится серия из n независимых испытаний, в каждом из которых может произойти в точности одно из событий A_1, …, A_k, имеющих неизвестные вероятности p_1, …, p_k. По результатам серии фиксируются количества наступлений событий A_1, …, A_k, так что наблюдаемые величины образуют вектор $(\nu_1, \ldots, \nu_k)$, имеющий полиномиальное распределение $\Pi(p_1^*, \ldots, p_k^*; n)$.

Множество всех возможных распределений $\wp$ величин $(\nu_1, \nu_2, \ldots, \nu_k)$ образовано всеми полиномиальными распределениями:

$$\wp = \left\{ \Pi(p_1, p_2, \ldots, p_k; n) : \sum_{i=1}^{k} p_i = 1, p_i \ge 0 \right\}.$$

Основная гипотеза H_0 заключается в том, что неизвестные вероятности p_i равны заданным выражениям $p_i^0(\theta)$ при некотором значении параметра $\theta \in \Theta$, где Θ – множество допустимых значений параметра (в общем случае параметр θ является d -мерным):

$$H_0 : p_1 = p_1^0(\theta), \ldots, p_k = p_k^0(\theta).$$

Множество $\wp_0$, соответствующее основной гипотезе H_0, образовано полиномиальными распределениями $\Pi(p_1^0(\theta),\ldots, p_k^0(\theta); n)$, получаемыми при всевозможных допустимых значениях параметра θ:

$$\wp_0 = \{\Pi(p_1^0(\theta),\ldots, p_k^0(\theta); n) : \theta \in \Theta\}.$$

Таким образом, гипотеза H_0 утверждает, что неизвестное распределение $\Pi(p_1^*,\ldots, p_k^*; n)$ совпадает с одним из распределений вида $\Pi(p_1^0(\theta),\ldots, p_k^0(\theta); n)$ при некотором допустимом значении параметра $\theta^* \in \Theta$.

Альтернативная гипотеза H_1, напротив, заключается в том, что каждом допустимом значении параметра $\theta \in \Theta$ нарушается хотя бы одно из равенств, утверждаемых основной гипотезой H_0:

$$H_1 : \forall \theta \in \Theta \exists j : p_j^* \ne p_j^0(\theta),$$

другии словами, неизвестное распределение $\Pi(p_1^*,\ldots, p_k^*; n)$ не совпадает ни с одним из распределений вида $\Pi(p_1^0(\theta),\ldots, p_k^0(\theta); n)$ при $\theta \in \Theta$. Множество распределений $\wp_1$, соответствующее H_1 представляет все оставшиеся полиномиальные распределения за вычетом распределения из множества $\wp_0$:

$$\wp_1 = \wp \setminus \wp_0.$$

В приведенных условиях требуется выполнить проверку освноной гипотезы H_0 против альтернативной гипотезы H_1.

Заметим, что сформулированная задача, схожа с задачей, рассмотренной в пункте 2, отличие заключается в том, что гипотетические вероятности $p_1^0(\theta)$ являются не числовыми значениями, а некоторыми функциями параметра θ. Указанное отличие не позволяет в качестве статистики критерия использовать функцию $X_n^2(\nu_1,\ldots, \nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta))$:

$$X_n^2(\nu_1,\ldots,\nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta)) = \sum_{i=1}^{k} \frac{(\nu_i - np_i^0(\theta))^2}{np_i^0(\theta)},$$

поскольку функция $X_n^2(\nu_1,\ldots, \nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta))$ оказывается зависимой от параметра θ, теорема Пирсона (5А.14) не может быть применима и как следствие предельное (при $n \to \infty$) распределение величины $X_n^2(\nu_1,\ldots, \nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta))$ неизвестно. Более того, следует ожидать, что это распределение окажется различным при различных значениях параметра θ. Тем не менее, при специальном выборе параметра θ удается найти предельное распределение.

Предположим, что при каждой реализации наблюдаемых величин $(\nu_1,\ldots, \nu_n)$ значение параметра θ выбирается таким образом, чтобы минимизировать значение функции $X_n^2(\nu_1,\ldots, \nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta))$. Минимальные значения $X_n^2(\nu_1,\ldots, \nu_k \mid p_1^0(\theta),\ldots, p_k^0(\theta))$ образуют статистику $\tilde{X}_n^2(\nu_1,\ldots, \nu_k)$, не зависящую от параметра:

$$\tilde{X}_n^2(\nu_1,\ldots,\nu_k)=\min_\theta X_n^2(\nu_1,\ldots,\nu_k\mid p_1^0(\theta),\ldots,p_k^0(\theta))=\min_\theta\sum_{i=1}^k\frac{(\nu_i-np_i^0(\theta))^2}{np_i^0(\theta)}.$$

Пусть $\tilde{\theta}(\nu_1,\ldots,\nu_n)$ – значение параметра θ, при котором достигается минимальное значение функции $X_n^2(\nu_1,\ldots,\nu_k\mid p_1^0(\theta),\ldots,p_k^0(\theta))$, тогда:

$$\tilde{X}_n^2(\nu_1,\ldots,\nu_k)=\min_\theta\sum_{i=1}^k\frac{(\nu_i-np_i^0(\theta))^2}{np_i^0(\theta)}=\sum_{i=1}^k\frac{(\nu_i-np_i^0(\tilde{\theta}))^2}{np_i^0(\tilde{\theta})}=X_n^2(\nu_1,\ldots,\nu_k\mid p_1^0(\tilde{\theta}),\ldots,p_k^0(\tilde{\theta}))$$

Теорема 5А.17. (Фишер)

Пусть совокупность величин $(\nu_1,\ldots,\nu_k)$ имеет полиномиальное распределение $\Pi(p_i^0(\theta^*),\ldots,p_k^0(\theta^*);n)$ при некотором допустимом значении параметра $\theta^*\in\Theta$, тогда распределение статистики:

$$\tilde{X}_n^2(\nu_1,\ldots,\nu_k)=\min_\theta\sum_{i=1}^k\frac{(\nu_i-np_i^0(\theta))^2}{np_i^0(\theta)},$$

стремится при $n\to\infty$ к распределению $\chi^2(k-1-d)$, где d – размерность множества значений параметра Θ.

Без доказательства.

Вычисление статистики $\tilde{X}_n^2(\nu_1,\ldots,\nu_k)$ требует трудоемкой операции нахождения минимума, а для решения в общем виде требует нахождения функции $\tilde{\theta}(\nu_1,\ldots,\nu_n)$ доставляющей минимум $X_n^2(\nu_1,\ldots,\nu_k\mid p_1^0(\theta),\ldots,p_k^0(\theta))$, что существенно затрудняет использование статистического критерия.

Оказывается, сформулированная выше теорема Фишера справедлива и в том случае, когда вместо функции $\tilde{\theta}(\nu_1,\ldots,\nu_n)$ используется МП-оценка $\hat{\theta}(\nu_1,\ldots,\nu_n)$ параметра θ, вычисляемая по функции правдоподобия, составленной в соответствии с распределением $\Pi(p_i^0(\theta),\ldots,p_k^0(\theta);n)$.

Теорема 5А.18. (Фишер)

Пусть наблюдаемые величины $(\nu_1,\ldots,\nu_k)$ имеют полиномиальное распределение $\Pi(p_i^0(\theta^*),\ldots,p_k^0(\theta^*);n)$ при некотором допустимом значении параметра $\theta^*\in\Theta$, и функции $p_i^0(\theta)$ при $\theta\in\Theta$ таковы, что:

1) $p_i^0(\theta)\ge c>0$ ($i=\overline{1,k}$),

2) существуют и непрерывны производные $\dfrac{\partial p_i(\theta)}{\partial\theta_j}$ ($i=\overline{1,k}$, $j=\overline{1,d}$),

3) существуют и непрерывны производные $\dfrac{\partial^2 p_i(\theta)}{\partial\theta_j\partial\theta_l}$ ($i=\overline{1,k}$, $j=\overline{1,d}$, $l=\overline{1,d}$),

4) для всех $\theta\in\Theta$ ранг матрицы, образованной частными производными, $\left\|\dfrac{\partial p_i(\theta)}{\partial\theta_j}\right\|$ равен d (где d – размерность множества значений параметра Θ).

Если $\hat{\theta}(\nu_1,\ldots,\nu_n)$ – МП-оценка параметра θ, тогда распределение статистики:

$$\hat{X}_n^2(\nu_1,\ldots,\nu_k)=X_n^2(\nu_1,\ldots,\nu_k\mid p_1^0(\hat{\theta}),\ldots,p_k^0(\hat{\theta}))=\sum_{i=1}^k\frac{(\nu_i-np_i^0(\hat{\theta}))^2}{np_i^0(\hat{\theta})}$$

стремится при $n\to\infty$ к распределению $\chi^2(k-1-d)$.

Без доказательства.

В остальном статистический критерий аналогичен статистическому критерию хи-квадрат, рассмотренному в пункте 2: в качестве критической области Γ_α выбирается область вида:

$$\Gamma_\alpha = \{x : x = X_n^2(\nu_1,..., \nu_k \mid p_1^0(\hat{\theta}),..., p_k^0(\hat{\theta})) \geq h_\alpha\},$$

где пороговое значение h_α выбирается исходя из заданного уровня значимости α как квантиль уровня $1-\alpha$ распределения $\chi^2(k-1-d)$.

Проверка гипотезы о распределении с неизвестным параметром.

Пусть $(\xi_1, \xi_2,..., \xi_n)$ – выборка из неизвестного распределения $F_\xi(x)$, основная гипотеза H_0 заключается в том, что $F_\xi(x) = F_0(x \mid \theta)$, где $F_0(x \mid \theta)$ – функция распределения известная с точностью до значения параметра θ, а альтернативная H_1 – в том, что неизвестная фунция $F_\xi(x)$ не совпадает ни одной из функций вида $F_0(x \mid \theta)$. Требуется предложить критерий проверки основной гипотезы H_0 против альтернативной гипотезы H_1.

На практике сформулированную задачу заменяют другой «близкой» задачей: выбираются точки $x_1 < ... < x_{k-1}$ и рассматривается разбиение числовой оси на полуинтервалы и интервалы:

$$L_1 = (-\infty; x_1),\ L_2 = [x_1; x_2),\ \ldots,\ L_k = [x_{k-1}; \infty).$$

Рассматриваются события $A_1, \ldots, A_k$:

$$A_j = \{\omega : \xi_i(\omega) \in L_j\}.$$

Легко видеть, что,

$$p_1 = P(A_1) = F_\xi(x_1),$$
$$p_2 = P(A_2) = F_\xi(x_2 \mid \theta) - F_\xi(x_1 \mid \theta),$$
$$\ldots,$$
$$p_k = P\{A_k\} = 1 - F_\xi(x_k \mid \theta).$$

Для исходной выборки $(\xi_1,..., \xi_n)$ определяется вектор $(\nu_1,..., \nu_k)$ так, что:

$$\nu_j = \sum_{i=1}^{n} I(\xi_i, L_j),$$

$$I(\xi_i, L_j) = \begin{cases} 1 & ,\xi_i \in L_j \\ 0 & ,\xi_i \notin L_j \end{cases}.$$

В качестве основной гипотезы рассматривается «расширенная» гипотеза $\tilde{H}_0$:

$$p_i = p_i^0(\theta),\ i = \overline{1,k},$$
$$p_1^0 = F_0(x_1 \mid \theta),$$
$$p_2^0 = F_0(x_2 \mid \theta) - F_0(x_1 \mid \theta),$$
$$\ldots,$$
$$p_k^0 = 1 - F_0(x_k \mid \theta).$$

Для проверки гипотезы $\tilde{H}_0$ используется статистический критерий со статистикой $X_n^2(\nu_1,..., \nu_k \mid p_1^0(\hat{\theta}),..., p_k^0(\hat{\theta}))$, где $\hat{\theta}(\nu_1,..., \nu_k)$ – МП-оценка параметра θ.

В качестве критической области Γ_α выбирается область вида:

$$\Gamma_\alpha = \{x : x = X_n^2(\nu_1,..., \nu_k \mid p_1^0(\hat{\theta}),..., p_k^0(\hat{\theta})) \geq h_\alpha\},$$

где h_α – квантиль уровня $1-\alpha$ распределения $\chi^2(k-1-d)$ и α – заданный уровень значимости.

Проверка гипотезы о независимости признаков.

Пусть проводится серия из n независимых испытаний, в каждом из которых происходит в точности одно из событий $A_1, \ldots, A_k$ и в точности одно из событий $B_1, \ldots, B_m$, причем вероятности совместного наступления событий $P(A_i B_j) = P_{ij}$ неизвестны. По результатам серии фиксируется количество ν_{ij} наступлений каждой пары $A_i B_j$, таким образом, наблюдаемые величины $(\nu_{11},..., \nu_{1m}, \nu_{21},..., \nu_{2m},..., \nu_{k1},..., \nu_{km})$ имеют полиномиальное распределение $\Pi(P_{11},..., P_{1m}, P_{21},..., P_{2m}, P_{k1},..., P_{km}; n)$.

Основная гипотеза H_0 заключается в том, что события A_i и B_j попарно независимы, то есть вероятности $P_{ij} = P(A_i B_j) = P(A_i)P(B_j)$, или иначе неизвестные вероятности $P_{ij} = \theta_{A,i}\theta_{B,j}$ при некоторых числах $\theta_{A,i}$ и $\theta_{B,j}$, где вектор вероятностей $\theta = (\theta_{A,1},..., \theta_{A,k-1}, \theta_{B,1},..., \theta_{B,m-1})$ играет роль параметра:

$$H_0: P_{ij} = p_{ij}^0(\theta) = \theta_{A,i}\theta_{B,j},$$

$$i = \overline{1,k}, \ j = \overline{1,q}.$$

Заметим, что $\theta_{A,k} = 1 - \sum_{i=1}^{k-1} \theta_{A,i}$ и $\theta_{B,m} = 1 - \sum_{j=1}^{m-1} \theta_{B,j}$, поэтому эти вероятности не входят в вектор параметров $\theta = (\theta_{A,1},..., \theta_{A,k-1}, \theta_{B,1},..., \theta_{B,m-1})$.

Альтернативная гипотеза H_1 утверждает, что ни при каких числах $\theta_{A,i}$ и $\theta_{B,j}$ одновременное не выполняются все равенства $P_{ij} = \theta_{A,i}\theta_{B,j}$.

Требуется предложить статистический критерий проверки основной гипотезы H_0 против альтернативной гипотезы H_1.

Для решения задачи используется критерий хи-квадрат проверки сложной гипотезы со статистикой,

$$\hat{X}_n^2(\nu_{11},..., \nu_{km}) = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(\nu_{ij} - np_{ij}^0(\hat{\theta}))^2}{np_{ij}^0(\hat{\theta})} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(\nu_{ij} - n\hat{\theta}_{A,i}\hat{\theta}_{B,j})^2}{n\hat{\theta}_{A,i}\hat{\theta}_{B,j}},$$

где вектор вероятностей $\hat{\theta} = (\hat{\theta}_{A,1},..., \hat{\theta}_{A,k-1}, \hat{\theta}_{B,1},..., \hat{\theta}_{B,m-1})$ является МП-оценкой параметра $\theta = (\theta_{A,1},..., \theta_{A,k-1}, \theta_{B,1},..., \theta_{B,m-1})$ и $\hat{\theta}_{A,k} = 1 - \sum_{i=1}^{k-1} \hat{\theta}_{A,i}$, $\hat{\theta}_{B,m} = 1 - \sum_{j=1}^{m-1} \hat{\theta}_{B,j}$. Гипотеза H_0 определяет функцию распределения величин $(\nu_{11},..., \nu_{km})$ как полиномиальное распределение $\Pi(p_{11}^0(\theta),..., p_{km}^0(\theta); n)$ с вероятностями:

$$P(y_{11},..., y_{km}; p_{11}^0(\theta),..., p_{km}^0(\theta); n) = \begin{cases} \dfrac{n!}{\prod_{i=1}^{k}\prod_{j=1}^{m} y_{ij}!} \prod_{i=1}^{k}\prod_{j=1}^{m} p_{ij}(\theta)^{y_{ij}} & , \sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} = n \\ 0 & , \sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} \neq n \end{cases} =$$

$$= \begin{cases} \dfrac{n!}{\prod_{i=1}^{k}\prod_{j=1}^{m} y_{ij}!} \prod_{i=1}^{k}\prod_{j=1}^{m} (\theta_{A,i}\theta_{B,j})^{y_{ij}} & , \sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} = n \\ 0 & , \sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} \neq n \end{cases}.$$

Отсюда функция правдоподобия $L(y_{11},..., y_{km} | \theta) = P(y_{11},..., y_{km}; p_{11}^0(\theta),..., p_{km}^0(\theta); n)$ и МП-оценка $\hat{\theta}$ доставляет максимальное значение функции $\ln L(y_{11},..., y_{km}; \theta; n)$ (или минимальное значение $-\ln L$) при условиях $\sum_{i=1}^{k}\theta_{A,i}=1$ и $\sum_{j=1}^{m}\theta_{B,j}=1$. Для нахождения МП-оценки $\hat\theta$ воспользуемся методом множителей Лагранжа с функцией,

$$\Phi(\theta,\lambda_0,\lambda_1)=-\ln L(\nu_{11},...,\nu_{km})+\lambda_0\left(\sum_{i=1}^{k}\theta_{A,i}-1\right)+\lambda_1\left(\sum_{j=1}^{m}\theta_{B,j}-1\right)=$$

$$=-\ln\left(\frac{n!}{\prod_{i=1}^{k}\prod_{j=1}^{m}\nu_{ij}!}\prod_{i=1}^{k}\prod_{j=1}^{m}(\theta_{A,i}\theta_{B,j})^{\nu_{ij}}\right)+\lambda_0\left(\sum_{i=1}^{k}\theta_{A,i}-1\right)+\lambda_1\left(\sum_{j=1}^{m}\theta_{B,j}-1\right)=$$

$$=-\ln\left(\frac{n!}{\prod_{i=1}^{k}\prod_{j=1}^{m}\nu_{ij}!}\right)-\sum_{i=1}^{k}\sum_{j=1}^{m}\nu_{ij}\ln(\theta_{A,i}\theta_{B,j})+\lambda_0\left(\sum_{i=1}^{k}\theta_{A,i}-1\right)+\lambda_1\left(\sum_{j=1}^{m}\theta_{B,j}-1\right).$$

Для определения $\hat\theta=(\hat\theta_{A,1},...,\hat\theta_{A,k},\hat\theta_{B,1},...,\hat\theta_{B,m})$ требуется решить систему:

$$\begin{cases}\left.\dfrac{\partial\Phi}{\partial\theta_{A,i}}\right|_{\theta=\hat\theta}=0\\ \left.\dfrac{\partial\Phi}{\partial\theta_{B,j}}\right|_{\theta=\hat\theta}=0\\ \left.\dfrac{\partial\Phi}{\partial\lambda_0}\right|_{\theta=\hat\theta}=0\\ \left.\dfrac{\partial\Phi}{\partial\lambda_1}\right|_{\theta=\hat\theta}=0\end{cases}\Leftrightarrow\begin{cases}-\sum_{j=1}^{m}\nu_{ij}\dfrac{\hat\theta_{B,j}}{\hat\theta_{A,i}\hat\theta_{B,j}}+\lambda_0=0\\ -\sum_{i=1}^{k}\nu_{ij}\dfrac{\hat\theta_{A,i}}{\hat\theta_{A,i}\hat\theta_{B,j}}+\lambda_1=0\\ \sum_{i=1}^{k}\hat\theta_{A,i}-1=0\\ \sum_{j=1}^{m}\hat\theta_{B,j}-1=0\end{cases}\Leftrightarrow\begin{cases}-\sum_{j=1}^{m}\nu_{ij}\dfrac{1}{\hat\theta_{A,i}}+\lambda_0=0\\ -\sum_{i=1}^{k}\nu_{ij}\dfrac{1}{\hat\theta_{B,j}}+\lambda_1=0\\ \sum_{i=1}^{k}\hat\theta_{A,i}-1=0\\ \sum_{j=1}^{m}\hat\theta_{B,j}-1=0\end{cases}\Leftrightarrow\begin{cases}\hat\theta_{A,i}=\dfrac{\sum_{j=1}^{m}\nu_{ij}}{\lambda_0}\\ \hat\theta_{B,j}=\dfrac{\sum_{i=1}^{k}\nu_{ij}}{\lambda_1}\\ \sum_{i=1}^{k}\dfrac{\sum_{j=1}^{m}\nu_{ij}}{\lambda_0}-1=0\\ \sum_{j=1}^{m}\dfrac{\sum_{i=1}^{k}\nu_{ij}}{\lambda_1}-1=0\end{cases}\Leftrightarrow$$

$$\begin{cases}\hat\theta_{A,i}=\dfrac{\sum_{j=1}^{m}\nu_{ij}}{\lambda_0}\\ \hat\theta_{B,j}=\dfrac{\sum_{i=1}^{k}\nu_{ij}}{\lambda_1}\\ \lambda_0=\sum_{i=1}^{k}\sum_{j=1}^{m}\nu_{ij}\\ \lambda_1=\sum_{j=1}^{m}\sum_{i=1}^{k}\nu_{ij}\end{cases}\Leftrightarrow\begin{cases}\hat\theta_{A,i}=\dfrac{\sum_{j=1}^{m}\nu_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{m}\nu_{ij}}\\ \hat\theta_{B,j}=\dfrac{\sum_{i=1}^{k}\nu_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{k}\nu_{ij}}\\ \lambda_0=\sum_{i=1}^{k}\sum_{j=1}^{m}\nu_{ij}\\ \lambda_1=\sum_{j=1}^{m}\sum_{i=1}^{k}\nu_{ij}\end{cases}\Leftrightarrow\begin{cases}\hat\theta_{A,i}=\dfrac{\sum_{j=1}^{m}\nu_{ij}}{n}\\ \hat\theta_{B,j}=\dfrac{\sum_{i=1}^{k}\nu_{ij}}{n}\\ \lambda_0=n\\ \lambda_1=n\end{cases}.$$

Таким образом, статистика $\hat X_n^2(\nu_{11},...,\nu_{km})$ имеет вид:

$$\hat{X}_n^2(\nu_{11},\ldots,\nu_{km}) = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(\nu_{ij} - n\theta_{A,i}\theta_{B,j})^2}{n\theta_{A,i}\theta_{B,j}},$$

$$\theta_{A,i} = \frac{\sum_{j=1}^{m}\nu_{ij}}{n},\ \theta_{B,j} = \frac{\sum_{i=1}^{k}\nu_{ij}}{n}.$$

Согласно теореме Фишера 5А.18 распределение статистики $\hat{X}_n^2(\nu_{11},\ldots,\nu_{km})$ при $n \to \infty$ стремится к распределению $\chi^2(km - 1 - ((k-1)+(m-1)))$, где km – количество вероятностей P_{ij} и $(k-1)+(m-1)$ – количество параметров ($k-1$ параметров $\theta_{A,i}$ и $m-1$ параметров $\theta_{B,j}$). Легко видеть, что:

$$km - 1 - (k-1) - (m-1) = km - k - m + 1 = k(m-1) - (m-1) = (m-1)(k-1),$$

поэтому распределение статистики X_n^2 стремится при $n \to \infty$ к распределению $\chi^2((k-1)(m-1))$.

В качестве критической области Γ_α выбирается область вида:

$$\Gamma_\alpha = \{x : x = \hat{X}_n^2(\nu_{11},\ldots,\nu_{km}) \geq h_\alpha\}$$

где h_α – квантиль уровня $1-\alpha$ распределения $\chi^2((k-1)(m-1))$ и α – заданный уровень значимости.

Проверка гипотезы об однородности.

Пусть проводится m независимых серий испытаний: в первой серии проводится n_1 независимых испытаний, в каждом из которых происходит в точности одно из событий A_{11}, …, A_{1k}, во второй серии проводится n_2 независимых испытаний, в каждом из которых происходит в точности одно из событий A_{21}, …, A_{2k}, и так далее, в m-ой серии проводится n_m независимых испытаний, в каждом из которых происходит в точности одно из событий A_{m1}, …, A_{mk}. По результатам серии фиксируются количества ν_{ij} наступлений каждого события A_{ij}, при этом вероятности событий $P_{ij} = P(A_{ij})$ неизвестны.

В соответствии с условиями функция распределения вектора $(\nu_{11},\ldots,\nu_{mk})$ является произведением полиномиальных распределений $\prod_{i=1}^{m}\Pi(P_{i1},\ldots,P_{ik};n_i)$.

Основная гипотеза H_0 заключается в том, что при фиксированном j и переменном i события A_{ij} имеют одинаковые вероятности, то есть выполняются равенства,

$$P_{11} = P_{21} = \ldots = P_{m1},$$
$$\ldots,$$
$$P_{1k} = P_{2k} = \ldots = P_{mk},$$

или, что тоже самое, при фиксированном j и переменном i вероятности $P_{ij} = \theta_j$ при некоторых θ_j, где вектор вероятностей $\theta = (\theta_1,\ldots,\theta_{k-1})$ играет роль параметра:

$$H_0:\ P_{ij} = p_j^0(\theta) = \theta_j,$$
$$i = \overline{1,m},\ j = \overline{1,k}.$$

Вероятность $\theta_k = 1 - \sum_{j=1}^{k-1}\theta_i$, поэтому θ_k не входит в вектор параметров $\theta = (\theta_1,\ldots,\theta_{k-1})$.

Альтернативная гипотеза H_1 заключается в том, что нарушается хотя бы одно равенство, утверждаемое основной гипотезой H_0.

В приведенных условиях требуется предложить критерий проверки освновной гипотезы H_0 против альтернативной H_1.

Для решения задачи используется статистика,

$$\hat{X}_n^2(\nu_{11},\ldots,\nu_{mk}) = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij} - n_i p_j^0(\hat{\theta}))^2}{n_i p_j^0(\hat{\theta})} = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij} - n_i\hat{\theta}_j)^2}{n_i\hat{\theta}_j},$$

где вектор вероятностей $\hat{\theta} = (\hat{\theta}_1,\ldots,\hat{\theta}_{k-1})$ является МП-оценкой параметра $\theta = (\theta_1,\ldots,\theta_{k-1})$ и $\hat{\theta}_k = 1 - \sum_{j=1}^{k-1}\hat{\theta}_i$. Гипотеза H_0 определяет функцию распределения величин $(\nu_{11},\ldots,\nu_{mk})$ как произведение полиномиальных распределений:

$$F_0(y_{11},\ldots,y_{km}; p_1^0(\theta),\ldots,p_k^0(\theta); n_1,\ldots,n_m) = \begin{cases}\prod_{i=1}^{m}\left(\dfrac{n_i!}{\prod_{j=1}^{k} y_{ij}!}\prod_{j=1}^{k} P_{ij}^{\,y_{ij}}\right) & ,\sum_{j=1}^{k} y_{ij} = n_i,\ i=\overline{1,m}\\ 0 & ,\text{иначе}\end{cases} =$$

$$= \begin{cases}\prod_{i=1}^{m}\left(\dfrac{n_i!}{\prod_{j=1}^{k} y_{ij}!}\prod_{j=1}^{k}\theta_j^{\,y_{ij}}\right) & ,\sum_{j=1}^{k} y_{ij} = n_i,\ i=\overline{1,m}\\ 0 & ,\text{иначе}\end{cases}.$$

Таким образом, функция правдоподобия $L(y_{11},\ldots,y_{km}\,|\,\theta) = F_0(y_{11},\ldots,y_{km}; p_1^0(\theta),\ldots,p_k^0(\theta); n_1,\ldots,n_m)$ и МП-оценка $\hat{\theta}$ доставляет максимальное значение функции $\ln L$ (или минимальное значение функции $-\ln L$) при условии $\sum_{j=1}^{k}\hat{\theta}_j = 1$. Для нахождения МП-оценки θ^* используется метод множителей Лагранжа с функцией,

$$\Phi(\theta,\lambda_0) = -\ln L(\nu_{11},\ldots,\nu_{km}) + \lambda_0\left(\sum_{j=1}^{k}\theta_j - 1\right) = -\ln\left(\prod_{i=1}^{m}\frac{n_i!}{\prod_{j=1}^{k} y_{ij}!}\prod_{j=1}^{k}\theta_j^{\,\nu_{ij}}\right) + \lambda_0\left(\sum_{j=1}^{k}\theta_j - 1\right) =$$

$$= -\sum_{i=1}^{m}\ln\left(\frac{n_i!}{\prod_{j=1}^{k} y_{ij}!}\right) - \sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}\ln\theta_j + \lambda_0\left(\sum_{j=1}^{k}\theta_j - 1\right).$$

Для определения $\theta^* = (\theta_1^*,\ldots,\theta_k^*)$ требуется решить систему:

$$\begin{cases}\left.\dfrac{\partial\Phi}{\partial\theta_j}\right|_{\theta=\hat{\theta}} = 0\\ \left.\dfrac{\partial\Phi}{\partial\lambda_0}\right|_{\theta=\hat{\theta}} = 0\end{cases} \Leftrightarrow \begin{cases}-\sum_{i=1}^{m}\nu_{ij}\dfrac{1}{\hat{\theta}_j} + \lambda_0 = 0\\ \sum_{j=1}^{k}\hat{\theta}_j - 1 = 0\end{cases} \Leftrightarrow \begin{cases}\hat{\theta}_j = \dfrac{\sum_{i=1}^{m}\nu_{ij}}{\lambda_0}\\ \sum_{j=1}^{k}\dfrac{\sum_{i=1}^{m}\nu_{ij}}{\lambda_0} - 1 = 0\end{cases} \Leftrightarrow \begin{cases}\hat{\theta}_j = \dfrac{\sum_{i=1}^{m}\nu_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}}\\ \lambda_0 = \sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}\end{cases} \Leftrightarrow \begin{cases}\hat{\theta}_j = \dfrac{\sum_{i=1}^{m}\nu_{ij}}{\sum_{i=1}^{m} n_i}\\ \lambda_0 = \sum_{i=1}^{m} n_i\end{cases}.$$

Таким образом, статистика $\hat{X}_n^2(\nu_{11},\ldots,\nu_{mk})$ имеет вид:

$$\hat{X}_n^2(\nu_{11},\ldots,\nu_{mk}) = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij}-n_i\theta_j)^2}{n_i\theta_j},$$

$$\hat{\theta}_j = \frac{\sum_{i=1}^{m}\nu_{ij}}{\sum_{i=1}^{m}n_i}.$$

Можно показать, что распределение статистики $\hat{X}_n^2(\nu_{11},\ldots,\nu_{mk})$ при $n\to\infty$ стремится к распределению $\chi^2(m(k-1)-(k-1))$, где $m(k-1)$ – количество «независимых» ν_{ij} ($i=\overline{1,m}$, $j=\overline{1,k-1}$ – при фиксированном i: $\nu_{ik}=n_i-\sum_{j=1}^{k-1}\nu_{ij}$) и $k-1$ – количество параметров θ_i (заметим, что $m(k-1)-(k-1)=(k-1)(m-1)$).

В качестве критической области Γ_α выбирается область вида:

$$\Gamma_\alpha = \{x : x = \hat{X}_n^2(\nu_{11},\ldots,\nu_{mk}) \geq h_\alpha\},$$

где h_α – квантиль уровня $1-\alpha$ распределения $\chi^2((k-1)(m-1))$ и α – заданный уровень значимости.

4. Критерий согласия Колмогорова.

Пусть $(\xi_1,\ldots,\xi_n)$ – выборка из неизвестного распределения $F_\xi(x)$, основная гипотеза H_0 заключается в том, что:

$$H_0:\ F_\xi(x)=F_0(x),$$

а альтернативная гипотез H_1 напротив, утверждает , что:

$$H_1:\ F_\xi(x)\neq F_0(x).$$

Требуется составить критерий проверки гипотезы H_0 против гипотезы H_1.

Если функция $F_\xi(x)$ является непрерывной и возрастающей функцией, тогда в качестве критерия может использоваться критерий согласия Колмогорова, статистика которого $D_n(\xi_1,\ldots,\xi_n\mid F_0(x))$ имеет вид:

$$D_n(\xi_1,\ldots,\xi_n\mid F_0(x)) = \sqrt{n}\sup_{-\infty<x<\infty}\left|F_n^*(x;\xi_1,\ldots,\xi_n)-F_0(x)\right|,$$

где $F_n^*(x;\xi_1,\ldots,\xi_n)$ – эмпирическая функция распределения, построенная по выборке $(\xi_1,\ldots,\xi_n)$.

Если распределение $F_\xi(x)\neq F_0(x)$, тогда точная верхняя грань $\sup_{-\infty<x<\infty}\left|F_n^*(x;\xi_1,\ldots,\xi_n)-F_0(x)\right|$ не стремится к нулю с ростом n, а стремится к некоторому конечному числу, которое затем умножается на $\sqrt{n}$. Отсюда, статистика $D_n(\xi_1,\ldots,\xi_n\mid F_0(x))$ неограниченно возрастает с ростом n, то есть с большой вероятностью принимает «большие» значения. Если же $F_\xi(x)=F_0(x)$, то из теоремы Колмогорова 5А.20, рассматриваемой далее, следует, что статистика $D_n(\xi_1,\ldots,\xi_n\mid F_0(x))$ с малой вероятностью принимает «большие» значения, поэтому в критическую область Γ_α гипотезы H_0 следует отнести «большие» значения статистики D_n:

$$\Gamma_\alpha = \{d : d = D_n(\xi_1,\ldots,\xi_n\mid F_0(x)) \geq h_\alpha\},$$

где $h_\alpha = h(n,\alpha)$ – пороговое значение, определяемое распределением статистики D_n, объемом выборки n и уровнем значимости α. Если основная гипотеза H_0 верна, то есть $F_\xi(x) = F_0(x)$, то при сравнительно небольших объемах выборки ($n < 20$) распределение статистики D_n известно точно, при больших объемах выборки ($n \geq 20$) для функции распределения известно приближенное выражение, основанное на сходимости функции распределения D_n к известной функции (теорема Колмогорова 5А.20).

Прежде всего, покажем, что статистика $D_n(\xi_1,..., \xi_n \mid F_0(x))$ неограниченно возрастает в том случае, когда верна альтернативная гипотеза H_1.

Утверждение 5А.19.

Пусть $(\xi_1,..., \xi_n)$ – выборка из распределения $F_\xi(x)$ и $F_\xi(x) \neq F_0(x)$, тогда последовательность (по n) случайных величин $D_n(\xi_1,..., \xi_n \mid F_0(x))$:

$$D_n(\xi_1,..., \xi_n \mid F_0(x)) = \sqrt{n} \sup_{-\infty < x < \infty} \left| F_n^*(x;\xi_1,..., \xi_n) - F_0(x) \right|,$$

является не ограниченной по вероятности.

Доказательство:

Пусть $\delta > 0$ и $\varepsilon > 0$ произвольно выбранные числа, покажем, что найдется N такое, что для всех $n \geq N$:

$$P\{\omega : D_n(\xi_1,..., \xi_n \mid F_0(x)) > \varepsilon\} \geq 1 - \delta .$$

Если функция распределения $F_\xi(x)$ и $F_0(x)$ не совпадают, то найдется хотя бы одно число x_0 такое, что $F_\xi(x_0) \neq F_0(x_0)$, и, следовательно, число $c = \dfrac{\left|F_\xi(x_0) - F_0(x_0)\right|}{2} > 0$. Заметим, что при всех элементарных событиях $\omega \in \Omega$:

$$D_n(\xi_1,..., \xi_n \mid F_0(x)) = \sqrt{n} \sup_{-\infty < x < \infty} \left|F_n^*(x) - F_0(x)\right| \geq \sqrt{n}\left|F_n^*(x_0) - F_0(x_0)\right| =$$

$$= \sqrt{n}\left|F_n^*(x_0) - F_\xi(x_0) + F_\xi(x_0) - F_0(x_0)\right| > \sqrt{n}\left(\left|F_\xi(x_0) - F_0(x_0)\right| - \left|F_n^*(x_0) - F_\xi(x_0)\right|\right).$$

Значение эмпирической функции распределения $F_n^*(x;\xi_1,..., \xi_n)$ при каждом x сходится по вероятности к значению $F_\xi(x)$, отсюда, для $\delta > 0$ и $c > 0$ найдется N^* такое, что для всех $n \geq N^*$:

$$P\{\omega : \left|F_n^*(x_0;\xi_1,..., \xi_n) - F_\xi(x)\right| < c\} \geq 1 - \delta .$$

Пусть для всех $n \geq N^*$ события $A(n)$ образованы теми элементарными событиями ω, при которых $\left|F_n^*(x_0;\xi_1,..., \xi_n) - F_\xi(x)\right| < c$:

$$A(n) = \{\omega : \left|F_n^*(x_0;\xi_1,..., \xi_n) - F_\xi(x)\right| < c\},$$

$$P(A(n)) > 1 - \delta .$$

Если $n \geq N^*$ и $\omega \in A(n)$, тогда,

$$D_n(\xi_1,..., \xi_n \mid F_0(x)) > \sqrt{n}\left(\left|F_\xi(x_0) - F_0(x_0)\right| - \left|F_n^*(x_0) - F_\xi(x_0)\right|\right) > \sqrt{n}\left(\left|F_\xi(x_0) - F_0(x_0)\right| - c\right) =$$

$$= \sqrt{n}\left(\left|F_\xi(x_0) - F_0(x_0)\right| - \frac{\left|F_\xi(x_0) - F_0(x_0)\right|}{2}\right) = \sqrt{n}\,\frac{\left|F_\xi(x_0) - F_0(x_0)\right|}{2}.$$

Пусть номер $N = \max\left\{\dfrac{4\varepsilon^2}{\left|F_\xi(x_0) - F_0(x_0)\right|^2}, N^*\right\}$, тогда для каждого $n \geq N$ и всех $\omega \in A(n)$:

$$D_n(\xi_1,\dots,\xi_n \mid F_0(x)) > \sqrt{n}\,\frac{\left|F_\xi(x_0) - F_0(x_0)\right|}{2} > \frac{2\varepsilon}{\left|F_\xi(x_0) - F_0(x_0)\right|}\,\frac{\left|F_\xi(x_0) - F_0(x_0)\right|}{2} > \varepsilon$$

Отсюда следует, что при $n \geq N$:

$$\{\omega : D_n(\xi_1,\dots,\xi_n \mid F_0(x)) > \varepsilon\} \supseteq A(n),$$

тогда при $n \geq N$,

$$P\{\omega : D_n(\xi_1,\dots,\xi_n \mid F_0(x)) > \varepsilon\} \geq P(A(n)) > 1 - \delta .$$

Таким образом, для произвольных $\delta > 0$ и $\varepsilon > 0$ найден номер N такой, что для всех $n \geq N$:

$$P\{\omega : D_n(\xi_1,\dots,\xi_n \mid F_0(x)) > \varepsilon\} \geq 1 - \delta .$$

Утверждение доказано.

Для вычисления порогового значения $h(n,\alpha)$ требуется определить функцию распределения статистики D_n в том случае, когда гипотеза H_0 верна. Покажем, что если гипотеза H_0 верна и функция $F_\xi(x)$ непрерывна и возрастает, тогда распределение D_n вовсе не зависит от $F_\xi(x)$. Действительно, если H_0 верна и $F_0(x) = F_\xi(x)$, то статистика D_n имеет вид:

$$D_n(\xi_1,\dots,\xi_n \mid F_0(x)) = \sqrt{n} \sup_{-\infty < x < \infty} \left|F_n^*(x;\xi_1,\dots,\xi_n) - F_\xi(x)\right|$$

Поскольку $F_\xi(x)$ непрерывна и возрастает, то существует обратная функция $F_\xi^{-1}(u)$ для u ($0 \leq u \leq 1$), тогда,

$$D_n(\xi_1,\dots,\xi_n \mid F_0(x)) = \sqrt{n} \sup_{0 \leq u \leq 1} \left|F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) - F_\xi(F_\xi^{-1}(u))\right| =$$

$$= \sqrt{n} \sup_{0 \leq u \leq 1} \left|F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) - u\right| .$$

Эмпирическую функцию распределения $F_n^*(x;\xi_1,\dots,\xi_n)$ можно представить как сумму:

$$F_n^*(x;\xi_1,\dots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n} I(x - \xi_i),$$

$$I(t) = \begin{cases} 1 & ,t \geq 0 \\ 0 & ,t < 0 \end{cases}.$$

Отсюда,

$$F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n} I(F_\xi^{-1}(u) - \xi_i)$$

Поскольку $F_\xi(x)$ возрастает, то из $x \leq y$ следует, что $F_\xi(x) \leq F_\xi(y)$, поэтому если $F_\xi^{-1}(u) \leq \xi_i$, то $F_\xi(F_\xi^{-1}(u)) \leq F_\xi(\xi_i)$ или $u \leq F_\xi(\xi_i)$, если же $F_\xi^{-1}(u) \geq \xi_i$, тогда $F_\xi(F_\xi^{-1}(u)) \geq F_\xi(\xi_i)$ или $u \geq F_\xi(\xi_i)$. Отсюда следует, что:

$$F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n} I(F_\xi^{-1}(u) - \xi_i) = \frac{1}{n}\sum_{i=1}^{n} I(u - F_\xi(\xi_i)) .$$

Пусть случайные величины $\eta_i = F_\xi(\xi_i)$, тогда:

$$F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) = \frac{1}{n}\sum_{i=1}^{n} I(u - \eta_i) = G_n^*(u;\eta_1,\dots,\eta_n),$$

где $G_n^*(u;\eta_1,\dots,\eta_n)$ – эмпирическая функция распределения выборки $(\eta_1,\dots,\eta_n)$. Таким образом,

$$D_n(\xi_1,\dots,\xi_n \mid F_0(x)) = \sqrt{n} \sup_{0 \leq u \leq 1} \left|F_n^*(F_\xi^{-1}(u);\xi_1,\dots,\xi_n) - u\right| = \sqrt{n} \sup_{0 \leq u \leq 1} \left|G_n^*(u;\eta_1,\dots,\eta_n) - u\right| .$$

Заметим, что случайные величины $\eta_i = F_\xi(\xi_i)$ имеют равномерное распределение $R[0;1]$, поскольку $F_\xi(x)$ непрерывная и возрастающая функция (тема 4 пункт 6), поэтому $G_n^*(u;\eta_1,...,\eta_n)$ – эмпирическая функция распределения для выборки из равномерного распределения $R[0;1]$ независимо от того какова функция $F_\xi(x)$, отсюда распределение статистики:

$$D_n(\xi_1,..., \xi_n \mid F_0(x)) = \sqrt{n} \sup_{0 \le u \le 1} \left| G_n^*(u; F_\xi(\xi_1),..., F_\xi(\xi_1)) - u \right|$$

зависит только от n и не зависит от $F_\xi(x)$. При сравнительно небольших n распределение D_n может быть вычислено, и значения сведены в таблицы для различных n. При больших n используется приближенное вычисление функции распределения D_n, основанное на теореме Колмогорова.

Теорема 5А.20 (Колмогоров)

Пусть $(\xi_1,..., \xi_n)$ – выборка из распределения $F_\xi(x)$, $F_n^*(x;\xi_1,...,\xi_n)$ – эмпирическая функция распределения выборки $(\xi_1,..., \xi_n)$ и статистика $D_n(\xi_1,..., \xi_n; F_\xi(x))$:

$$D_n(\xi_1,..., \xi_n; F_\xi(x)) = \sqrt{n} \sup_{-\infty < x < \infty} \left| F_n^*(x;\xi_1,..., \xi_n) - F_\xi(x) \right|.$$

Если $F_\xi(x)$ – непрерывная функция, тогда для любого фиксированного $t > 0$:

$$\lim_{n \to \infty} P\left\{ D_n(\xi_1,...,\xi_n; F_\xi(x)) \le t \right\} = K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}.$$

Без доказательства.

Из теоремы Колмогорова следует, что если гипотеза H_0 верна, тогда уровень значимости α:

$$\alpha = P\left\{ D_n(\xi_1,..., \xi_n \mid F_0(x)) \ge h_\alpha \right\} = 1 - P\left\{ D_n(\xi_1,..., \xi_n \mid F_0(x)) < h_\alpha \right\} \approx 1 - K(h_\alpha),$$

$$K(h_\alpha) \approx 1 - \alpha,$$

где $K(t)$ известная функция, для которой составлены таблицы значений, позволяющие вычислить значение h_α.