Лектор – *Сенько Олег Валентинович*

Курс «»
Прикладная Статистика и Анализ Данных часть I

Отдельные показатели используются для непрерывных, порядковых и номинальных переменных,
Для непрерывные переменных используется коэффициент корреляции Пирсона

$$\rho = \frac{cov_{X,Y}}{\sigma_X \sigma_Y}$$

где

$$cov_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)(y_i - y)$$

σ_X, σ_Y - выборочные стандартные отклонения для случайных переменных X,Y.

Коэффициент корреляции Пирсона может быть использован также и для оценки связанности двух порядковых переменныхю Предположим, что все объекты выборки имеют разные ранги по каждой из пременных X,Y. Введём обозначение $d_i = x_i - y_i$. Тогда коэффициент корреляции между порядковыми переменными X и Y приобретает вид

$$\rho^{Sp}_{X,Y} = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{1}$$

Коэффициент корреляции, вычсляемый согласно формуле (1) принято называть ранговым коэффициентом корреляции Спирмена

Наряду с коэффициентом корреляции Спирмена для оценки связанности двух ранговых переиенных X и Y используется также коэффициент корреляции Кендалла ($\tau-$статистика) и $\gamma-$статистика Пусть $\widetilde{M}$- множество всевозможных пар объектов из анализируемой выборки $\widetilde{S}$. Пару объектов $(s_{j'}, s_{j''})$ назовём согласованной при одновременном выполнении пар неравенств $X(s_{j'}) > X(s_{j''})$ и $Y(s_{j'}) > Y(s_{j''})$ или $X(s_{j'}) < X(s_{j''})$ и $Y(s_{j'}) < Y(s_{j''})$. Пару объектов $(s_{j'}, s_{j''})$ назовём несогласованной при одновременном выполнении пары неравенств $X(s_{j'}) > X(s_{j''})$ и $Y(s_{j'}) > Y(s_{j''})$ или пары $X(s_{j'}) < X(s_{j''})$ и $Y(s_{j'}) < Y(s_{j''})$. Пусть N_{con} - число согласованных пар, N_{unc} - число несогласованных пар.

$$\tau = \frac{N_{con} - N_{unc}}{|\widetilde{M}|}$$

$$\gamma = \frac{N_{con} - N_{unc}}{N_{con} + N_{unc}}$$

Связанность двух бинарных переменных $X \in \{x_1, x_2\}$, $Y \in \{y_1, y_2\}$ В ячейках таблицы 1 (таблицы сопряжённости) показаны количества объектов из анализируемой выборки при различных сочетаниях переменных.

Таблица: Таблица 1

	x_1	x_2
y_1	a	b
y_2	c	d

Известной мерой связанности является

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$

Коэффициент взаимосвязи λ (Goodman and Kruskal) позволяет оценить взаимосвязь двух категориальных переменных. Прогноз категориальной переменной $Y \in \{y_1, \ldots, y_m\}$ может производится исключительно по одномерному маргинальному распределению этой величины. При этом в качестве прогнозирования используется значение y_i, для которого $P(y_i)$ максимальна.
Прогноз может производится также с использованием переменной X. При этом в качестве прогноза при $X = x_i$ берётся мода условного распределения $P(Y|X = x_i)$, $i = 1, \ldots, m$.

- N_y - число ошибочных прогнозов при прогнозировании Y с испрльзованием $P(Y)$
- N_x - число ошибочных прогнозов при прогнозировании X с испрльзованием $P(X)$
- $N_{y,x}$ -число ошибочных прогнозов при прогнозировании Y с испрльзованием $P(Y|X)$
- $N_{x,y}$ -число ошибочных прогнозов при прогнозировании X с испрльзованием $P(X|Y)$

Мерой увеличения точности прогноза Y при условии использовании X может слухить

$$\lambda_{y,x} = \frac{N_y - N_{y,x}}{N_y}$$

Мерой увеличения точности прогноза X при условии использовании Y может слухить

$$\lambda_{x,y} = \frac{N_x - N_{x,y}}{N_x}$$

Коэффициент симметричной взаимосвязи

$$\lambda = \frac{N_y - N_{y,x} + N_x - N_{x,y}}{N_x + N_y}$$

Точечной оценкой параметра $\theta \in \Theta$ вероятностного распределения $\mathbf{P}$, называется значение функции $\hat{\theta}$, заданной на случайных выборках из $\mathbf{P}$, и принимающая значения из Θ. Например, $\frac{1}{m}\sum_{i=1}^{m} x_i$ является точечной оценкой $\mathbb{E}(X)$, вычисляемой по выборке $\{x_1, \ldots, x_m\}$

- Точечная оценка называется несмещённой, если

$$\mathbb{E}[\hat{\theta}(\widetilde{S})] = \theta$$

  в независимости от числа элементов в выборке $\widetilde{S}$
- Точечная оценка $\hat{\theta}(\widetilde{S})$ называется состоятельной, если $\hat{\theta}(\widetilde{S}) \to \theta$ при $m \to \infty$
  В определении состоятельности имеется ввиду сходимость по вероятности, то есть $\forall \epsilon \lim_{m\to\infty} \mathbf{P}\{|\hat{\theta}(\widetilde{S}_m) - \theta| > \epsilon\} = 0$

Несмещённая оценка $\hat{\theta}$ называется эффективной, если для любой несмещённой оценки $\hat{\theta}'$, отличной от $\hat{\theta}$ выполняется неравенство

$$\mathbb{E}(\hat{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta}' - \theta)^2$$

Эффективность, нижнияя граница для дисперсии оценок

Ограничения сверху на величину дисперсии несмещённых точечных оценок задаются неравенством Рао-Крамера

$$\mathbb{E}(\hat{\theta} - \theta)^2 \geq \frac{1}{mI(\theta)}$$

где

$$I(\theta) = \mathbb{E}[(\frac{\partial l(\widetilde{S}|\theta)}{\partial \theta})^2]$$

-информация Фишера

Метод максимального правдоподобия

В случае непрерывных распределений Функция $L(\widetilde{S}|\theta)$, равная многомерной плотности вероятности в точке, сответствующей выборке $\widetilde{S}$.. В случае дискретных распределений $L(\widetilde{S}|\theta)$ равна вероятности $\widetilde{S}$. Оценка максимального правдоподобия параметра θ:

$$\theta_{ML} = \arg\max_{\theta\in\Theta} L(\widetilde{S}|\theta)$$

Пусть выборка $\widetilde{S}$ состоит из независимых наблюдений,взятых из одного и того же вероятностного распределения $\widetilde{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Тогда

$$L(\widetilde{S}|\theta) = \prod_{j=1}^{m} f(\mathbf{x}_j|\theta)$$

На практике удобнее использовать логарифм функции правдоподобия

$$l(\widetilde{S}|\theta) = \log L(\widetilde{S}|\theta) = \sum_{i=1}^{m} \log[f(\mathbf{x}_j|\theta)]$$

Оценки максимального правдоподобия являются состоятельными. Вообще говоря оценки максимального правдоподобия не являются несмещёнными.

Имеет место слабая сходимость отклонения оценок максимального правдоподобия от истинного значения параметра распределения к нормальному распределению с нулевым математическим ожиданием:

$$\xi_m = \sqrt{m}(\hat{\theta}_m - \theta)) \to N[0, I(\theta)],$$

что соответствует сходимости $\forall x$

$$\mathbb{F}_{\xi_m}(x) \to \mathbb{F}_{N[0,I(\theta)]}(x)$$

при $m \to \infty$

Предполагается, что искомый параметр распределения θ сам является сучайной величинойю Предполагается, что на множестве значений параметра Θ заданы априорная вероятностная мера с плотностью $f_{pr}(\theta)$. Задана также функция потерь $L[\theta, \hat{\theta}(\widetilde{S})]$. Байесовский риск определяется как

$$\int_{\Theta} L[\theta, \hat{\theta}(\widetilde{S})] f(\widetilde{S}|\theta) f_{pr}(\theta) d\theta =$$

$$= f(\widetilde{S}) \int_{\Theta} L[\theta, \hat{\theta}(\widetilde{S})] f(\theta|\widetilde{S}) d\theta$$

Байесовской точечной оценке соответствует минимальное значение байесовского риска.

При квадратичных потерях, то есть при

$$L[\theta, \hat{\theta}(\widetilde{S})] = [\hat{\theta}(\widetilde{S}) - \theta]^2$$

$$\int_{\Theta} [\hat{\theta}^2(\widetilde{S}) f(\theta|\widetilde{S}) - 2\hat{\theta}(\widetilde{S}) f(\theta|\widetilde{S})\theta +$$

$$+\theta^2 f(\theta|\widetilde{S})] d\theta$$

Принимая во внимание равенство $\int_{\Theta} f(\theta|\widetilde{S}) d\theta = 1$ и явную независимость $\hat{\theta}(\widetilde{S})$ от θ получаем, что байесовской точечной оценкой является апостериорное математическое ожидание

$$\hat{\theta}(\widetilde{S}) = \int_{\Theta} \theta \times f(\theta|\widetilde{S}) d\theta$$

Пусть у нас имеется семейство вероятностных распределений случайной величины U, задаваемых параметрами $\theta_1, \ldots, \theta_k$
Предположим, что моменты U связаны с параметрами $\theta_1, \ldots, \theta_k$

$$\mathbb{E}U = g_1(\theta_1, \ldots, \theta_k)$$

$$\mathbb{E}U^2 = g_2(\theta_1, \ldots, \theta_k)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\mathbb{E}U^r = g_r(\theta_1, \ldots, \theta_k).$$

Поиск неизвестных значений параметров $\theta_1, \ldots, \theta_k$ производится по следующей схеме. Рассчитаем выборочные оценки $M_1, \ldots, M_r$ моментов $\mathbb{E}U, \ldots, \mathbb{E}U^r$ Искомые оценки $\hat{\theta}_1, \ldots, \hat{\theta}_k$ ищутся как решене системы уравнений:

$$M_1 = g_1(\hat{\theta}_1, \ldots, \hat{\theta}_k)$$

$$M_2 = g_2(\hat{\theta}_1, \ldots, \hat{\theta}_k)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$M_r = g_r(\hat{\theta}_1, \ldots, \hat{\theta}_k).$$

Метод наименьших квадрато используется для оценок параметров регрессиооных моделей.

Предположим, что связь целевой переменной Y с переменной $X_1, \ldots, X_n$ описывается с помощью уравнения

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon$$

Вектор оценок регрессионных параметров $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_n)$ ищется по обучающей выборке $\widetilde{S} = \{y_1, \mathbf{x}_1, \ldots, (y_m, \mathbf{x}_m)\}$ через минимизацию суммы квадратов ошибок прогнозирования:

$$\hat{\boldsymbol{\beta}} = \arg\min \{ \sum_{j=1}^{m} [y_j - \beta_0 - \sum_{i=1}^{n} x_{ji}\beta_i]^2 \}$$

Можно показать, что

$$\hat{\boldsymbol{\beta}} = (\hat{X}^t\hat{X})^{-1}\hat{X}^t\boldsymbol{y}, \qquad (2)$$

где $\hat{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ \dots & \dots & \dots & \dots \\ 1 & x_{m1} & \dots & x_{mn} \end{pmatrix}$ - матрица плана,$\boldsymbol{y} = (y_1, \dots, y_m)$

Из уравнения (2) следует возможность представления $\hat{\beta}_i = \hat{c}_{1i}y_1 + \ldots + \hat{c}_{mi}y_m$, где $i = 0 \ldots, n$, а $\hat{c}_{ji}$ являются элементами матрицы $\hat{C} = (\hat{X}^t\hat{X})^{-1}\hat{X}^t$, зависящими от значений переменных $X_1, \ldots, X_n$, но независящих от значений целевой переменной Y.

- Всевозможные оценки параметро $\beta_0, \ldots, \beta_n$ вида $\tilde{\boldsymbol{\beta}} = \widetilde{C}\boldsymbol{y}$, где элементы матрицы $\widetilde{C}$ не зависят от значений переменной Y будем называть линейными.
- Оценка $\widetilde{\beta}_i$ называется несмещённой, если $\mathbb{E}\widetilde{\beta}_i = \beta_i$ при $i = 0, 1, \ldots, n$.
- Несмещённу оценку $\tilde{\boldsymbol{\beta}}$ назовём наилучшей, если при произвольном $\boldsymbol{\gamma} \in \mathbb{R}^{n+1}$ и произвольной несмещённой оценке $\widetilde{\boldsymbol{\beta}}'$ справедливо нестрогое неравенство

$$\mathbb{E}(\boldsymbol{\gamma}\widetilde{\boldsymbol{\beta}} - \boldsymbol{\gamma}\boldsymbol{\beta})^2 \le \mathbb{E}(\boldsymbol{\gamma}\widetilde{\boldsymbol{\beta}}' - \boldsymbol{\gamma}\boldsymbol{\beta})^2 \tag{3}$$

Очевидно, что справедливость неравенства (3) при произвольном $\gamma \in \mathbb{R}^{n+1}$ эквивалентна неорицательной определённости разности ковариационных матриц $\Sigma_{\widetilde{\boldsymbol{\beta}}}$ и $\Sigma_{\widetilde{\boldsymbol{\beta}}'}$:

$$\Sigma_{\widetilde{\boldsymbol{\beta}}} = \begin{pmatrix} D(\widetilde{\beta}_1,) & \dots & cov(\widetilde{\beta}_1, \widetilde{\beta}_n) \\ \dots & \dots & \dots \\ cov(\widetilde{\beta}_n, \widetilde{\beta}_1) & \dots & D(\widetilde{\beta}_n) \end{pmatrix}$$

$$\Sigma_{\widetilde{\boldsymbol{\beta}}'} = \begin{pmatrix} D(\widetilde{\beta}'_1,) & \dots & cov(\widetilde{\beta}'_1, \widetilde{\beta}'_n) \\ \dots & \dots & \dots \\ cov(\widetilde{\beta}'_n, \widetilde{\beta}'_1) & \dots & D(\widetilde{\beta}'_n) \end{pmatrix}$$

Наилучшие линейные несмещённые оценки в англоязычной статистической литературе принято называть best linear unbiased estimates (BLUE).

Ошибка ϵ является случайной величиной, которая в общем случае может стохастически зависеть от переменных $X_1, \ldots, X_n$. Однако при статистическом моделирование часто используется следующие ограничения на характер зависимости. Пусть $\epsilon_1, \ldots, \epsilon_m$ случайные ошибки на объектам выборки $\widetilde{S}$.

- $\mathbb{E}(\epsilon_j) = 0,\ j = 1, \ldots, m$
- условие гомоскедастичности: дисперсии случайной ошибок для всех объектов выборки одинаковы, то есть $\mathbb{E}\epsilon_j^2 = \sigma^2, j = 1, \ldots, m.$
- случайные ошибки на двух различных объектах не коррелируют между собой $\mathbb{E}\epsilon_i\epsilon_j = 0$, при $i \neq j$

.

Теорема.При соблюдении перечисленных выше условий оценки регрессионных параметров с использованием метода наименьших квадратов оказываются наилучшими линейными несмещёнными оценками (BLUE)

Точечные оценки некоторого параметра вероятностного распределения θ не дают представление о точности полученных оценок, о возможных отклонениях полученных оценок от истинных значений θ. Пусть $I(\widetilde{S}) = (\hat{\theta}_l(\widetilde{S}), \hat{\theta}_u(\widetilde{S}))$- интервал с границами, вычисленными с помощью некоторой процедуры. Пусть $\Omega_{\widetilde{S}}$- множество выборок, порождаемых тем же самым вероятностным процессом, что и $\widetilde{S}$, и совпадающих с $\widetilde{S}$ по размерую.
Будем говорить, что $I(\widetilde{S})$ является доверительным интервалом на уровне α, если

$$\mathbb{P}\{\theta \in I(\omega) | \omega \in \Omega_{\widetilde{S}}\} = 1 - \alpha,$$

где $I(\omega)$ вычисляется с помощью той же самой процедуры, что и $I(\widetilde{S})$

Распределение Стьюдента. Пусть

$$U_0, U_1, \ldots, U_n$$

-независимые случайные величин,имеющие распределение $N(0,1)$ Распределение случайной величины $V = \sum_{i=1}^{n} U_i^2$ называется распределением χ^2 с числом степеней свободы $df = n$. Распределение с

$$t = \frac{U_0}{\sqrt{\frac{1}{n}V}}$$

называется распределением Стьюдента с числом степенями свободы $df = n$,если случайная величина V имеет распределение χ^2 с числом степеней свободы $df = n$.

Доверительные интервалы для средних значений

Предположим, что $X_1, \ldots, X_n$ - независимые случайные величины с распределением $N(\mu, \sigma^2)$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Случайная величина $Z = \frac{\overline{X}-\mu}{\sqrt{\frac{D}{n}}}$, где $D = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, имеет распределение Стьюдента с чисдом степеней свободы $df = n - 1$.
Пусть $t_{df,\beta}$-β-квантиль распределения Стьюдента с числом степеней свободы равным df
Очевидно, что неравенства $t_{n-1,\frac{\alpha}{2}} < Z < t_{n-1,\frac{1-\alpha}{2}}$ выполняются одновременно с вероятностью $1 - \alpha$.

Откуда следует, что с вероятностью α математическое ожидание μ принадлежит интервалу

$$(\overline{X} - t_{n-1,\frac{1-\alpha}{2}}\sqrt{\frac{D}{n}}, \overline{X} - t_{n-1,\frac{\alpha}{2}}\sqrt{\frac{D}{n}})$$

Из симметрии распределения Стьюдента следует $t_{n-1,\frac{\alpha}{2}} = -t_{n-1,\frac{1-\alpha}{2}}$. Откуда следует, что доверительный интерва для математического ожидания μ может быть записан в виде

$$(\overline{X} - t_{n-1,\frac{1-\alpha}{2}}\sqrt{\frac{D}{n}}, \overline{X} + t_{n-1,\frac{1-\alpha}{2}}\sqrt{\frac{D}{n}})$$

Доверительные интервалы для вероятности успехов в серии независимых испытаний

Предположим, что у нас имеется серия из n независимых испытаний, из которых k оказались успешными. Очевидно, что число успешных испытаний в серии подчиняется биномиальному распределению. Целью является поиск доверительного интервала (p_l, p_u) для вероятности успеха p на уровне α. То есть границы интервала должны подбираться таким образом, чтобы

$$Pr[p \in (p_l, p_u)] = 1 - \alpha$$

Метод Клоппера-Пирсона Метод вычисления границ доверительного интервала Клоппера-Пирсона основан на идеях теории проверки статистических гипотез.

Доверительные интервалы для вероятности успехов в серии независимых испытаний

Верхняя граница интервала p_u подбирается таким образом, чтобы нулевая гипотезао равенстве истинной вероятности успеха p_u отвергалась при величине ошибок первого рода $\frac{\alpha}{2}$, что соответствует выполнению равенства

$$\sum_{i=0}^{k} C_n^i p^i (1-p)^{n-i} = \sum_{i=0}^{k} C_n^i p_u^i (1-p_u)^{n-i} = \frac{\alpha}{2}$$

Доверительные интервалы для вероятности успехов в серии независимых испытаний

Нижняя граница интервала p_l подбирается таким образом, чтобы нулевая гипотезао равенстве истинной вероятности успеха p_l отвергалась при величине ошибок первого рода $\frac{\alpha}{2}$, что соответствует выполнению равенства

$$\sum_{i=0}^{k} C_n^i p^i (1-p)^{n-i} = \sum_{i=0}^{k} C_n^i p_l^i (1-p_l)^{n-i} = \frac{\alpha}{2}$$

Аппроксимация с помощью нормалного распределения

Аппроксимация $\hat{p} = \frac{k}{n}$ с помощью нормального распределения $N(p, \frac{\sigma}{\sqrt{n}})$ является оправданной при больших n благодаря центральной предельной теореме. Здесь $\sigma = \sqrt{p(1-p)}$ - стандартное отклонение для распределения Бернулли. Пусть z_β-β-квантиль нормального распределения $N(0,1)$.

Очевидно, что неравенства $z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{1-\alpha}{2}}$ выполняются одновременно с вероятностью $1-\alpha$.

Доверительные интервалы для вероятности успехов в серии независимых испытаний

Оценим стандартное отклонение как $\sqrt{\hat{p}(1-\hat{p})}$. Примем во внимание из-за симметрии нормального распределения, что

$$z_{\frac{\alpha}{2}} = -z_{\frac{1-\alpha}{2}}$$

Откуда следует одновременное выполнение с вероятностью $1-\alpha$ неравенств

$$\hat{p} - z_{\frac{1-\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{1-\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Доверительные интервалы для вероятности успехов в серии независимых испытаний

Таким образом доверительного интервала для p со значимостью на уровне α аппроксимируется интервалом

$$\hat{p} - z_{\frac{1-\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{1-\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Целью статистических критериев является оценивание, в какой степени наблюдения подтверждают статистических гипотезы - предположения о статистических характеристиках, процессов, генерирующего данные. В их число входят

- предположения о вероятностных распределениях отдельных переменных
- предположения о наиличии взаимосвязи между отдельными непрерывными или категориальными переменными
- предположения о различиях средних значений переменных в заранее заданных группах наблюдений
- предположейий о существовании регрессионных зависимостей
- предположейий о правомерности включения отдельных переменных в регрессионную модель

Можно выделить два основных подхода. В обоих этих подходах ипользуются статистики критерия- функции, зависящие от данных. Обычно статистика критерия $T(\widetilde{S})$ - это функция выборки $\widetilde{S}$, вычисляемой по значениям на объектах из $\widetilde{S}$ переменных, о которых делаются предположение, Передполагается, что различные значения $T(\widetilde{S})$ интуитивно соответствуют различным предположениям.

Первый подход основан на явном выделении двух сравниваемых гипотез о вероятностных распределениях, из которого генерируются данные:

- нулевая гипотеза $\mathbf{H}_0$ обычно соответствует базовому распределению, которое мы стремимся опровергнуть
- альтернативная гипотеза $\mathbf{H}_1$ обычно соответствует эффекту, существование которого мы стремимся доказать

Статистические критерии

Выбор решения о верности $\mathbf{H}_o$ или $\mathbf{H}_1$ может делаться, например, путём сравнения величины статистики $T(\widetilde{S})$ с некоторым пороговым значением δ

При выполнении неравенства $T(\widetilde{S}) \geq \delta$ верной считается гипотеза $\mathbf{H}_1$

При выполнении неравенства $T(\widetilde{S}) < \delta$ верной считается гипотеза $\mathbf{H}_0$

Таблица: Таблица 1

	$\mathbf{H}_o$ верна	$\mathbf{H}_1$ верна
Вывод о верности $\mathbf{H}_o$	Правильный вывод	ошибка первого рода
Вывод о верности $\mathbf{H}_1$	ошибка второго рода	Правильный вывод

Под вероятностью ошибки первого рода понимается вероятность ошибочного принятия гипотезы $\mathbf{H}_1$ при генерации выборок, совпадающих по размеру и структуре с выборкой $\widetilde{S}$, из вероятностного распределения, соответствующего гипотезе $\mathbf{H}_0$

Вероятность ошибки первого рода обозначим α

Под вероятностью ошибки второго рода понимается вероятность ошибочного принятия гипотезы $\mathbf{H}_0$ при генерации выборок,совпадающих по размеру и структуре с выборкой $\widetilde{S}$, из вероятностного распределения, соответствующего гипотезе $\mathbf{H}_1$

Вероятность ошибки второго рода обозначим β

В общем случае выделяется область значений статистики T, которую принято называть критической областью. При попадании $T(\widetilde{S})$ в критическую область $\mathbb{C}$ нулевая гипотеза $\mathbf{H}_0$ отвергается в пользу альтернативной гипотезы $\mathbf{H}_1$. В случае если $T(\widetilde{S})$ не принадлежит $\mathbb{C}$, принимается $\mathbf{H}_0$.
Величину $1-\beta$ принято называть мощностью критерия.

Несмещённость и состоятельность

- Критерий называется несмещённым, если $1-\beta > \alpha$
- Критерий называется состоятельным, если
  при фиксированном α
  $\beta \to 1$ при $|\widetilde{S}| \to \infty$

Статистический критерий является эффективным при одновременно низкими значения ошибок первого и второго рода. Возникает вопрос о оптимальности критерия в смысле минимальности β (или максимальности мощности критерия) при фиксированных значениях α. Ответ на этот вопрос даёт лемма Неймана-Пирсона.

критерий отношения правдоподобия Предположим, что гипотезы $\mathbf{H}_0$ и $\mathbf{H}_1$ различаются значением параметра θ: $\theta = \theta_0$ при $\mathbf{H}_0$, $\theta = \theta_1$ при $\mathbf{H}_1$. Статистикой критерия отношения правдоподобия является отношение функцмй правдоподобия

$$T(\widetilde{S}) = \frac{L(\widetilde{S}|\theta_0)}{L(\widetilde{S}|\theta_1)}$$

Решение о верности $\mathbf{H}_0$ при $T(\widetilde{S}) > \delta$

Лемма Нейианна-Пирсона Пусть δ_α - значения порога δ, при которо вероятность ошибок первого рода составила α. Критерий отношения правдоподобия с порогом δ_α имеет максимальную мощность среди всех критериев с фиксированным α.

Подход, основанный, на сравнении доставерности нулевой и альтернативной статистической гипотезы по имеющимся данным, связан прежде всего с именами Пирсона (E.S.Pearson) и Нейманна (J. Neyman) Подход является логичным и строгим. Вместе с тем, для многих прикладных задач применение данного подхода усложняется из-за обилия всевозможных вариантов выбора альтернативной гипотезы. Произвол в выборе затрудняет статистические расчёты и интерпретацию результатов.

Вполне возможно оценитвать достоверность нулевой гипотезы отдельно без привлечения альтернативных гипотез.

В современных исследованих значительную популярность завоевал подход, использующий для оценок достоверности $p-$значения.

Под $p-$значением понимается вероятность $\mathbf{P}[T(\widetilde{\omega}) \geq T(\widetilde{S})]$ при генерации выборок $\widetilde{\omega}$, совпадающих по размеру и структуре с выборкой $\widetilde{S}$ из вероятностного распределения, соответствующего нулевой гипотезе $\mathbf{H}_0$. Иными словами под $p-$значением понимается максимально возможная ошибка первого рода, при которой гипотеза $\mathbf{H}_0$ для выборки $\widetilde{S}$ может быть отвергнута.
Результат считается достоверным, если рассчитанное $p-$значение оказывается ниже заранее фиксированного уровня

$$p < 0.05, p < 0.01, p < 0.001...$$

. При выбороке конкретного значения порога следует учитывать эффект множественного тестирования. В целом выбор порога определяется многими факторами: имеющимися ресурсами для дальнейших исследований; политикой журнала, куда планируется представить публикацию; соответствие результата существующим теоретическим представлениям и т.д.

Расчёт $p-$значений может производится с использованием функций распределений, которые в рамках нулевых гипотез могут вычисляться точно для специально подобранных статистик. Также могут использоваться теоретические аппроксимации функций распределения, приближающиеся к истинным функциям при увеличении размеров выборок. В последнии годы для аппоксимации используется компьютерная генерация выборок с использованием датчиков случайных чисел.

Вычисление точных величин $p-$значений стало возможным только при развитии компьютерной техники. Ранее использовался подход посннованный на использований уровней значимости. Поскольку уровень значимость фактически обозначает допустимую величину ошибок первого рода, то будем использовать для его обозначения α. Обычно использовался фиксированный набор значений α

$$\alpha = 0.05, 0.01, 0.001\ldots$$

.

Заранее формировались таблицы, в которых каждому фикированному уровню значимости в зависимости от размера выборки $\widetilde{S}$ ставилась в соответствие квантиль распределения. Нулевая гипотеза отвергалась в случае, если величина статистики превышала указанную в таблице квантиль.

Критериями, когда $\mathbf{H}_0$ отвергается при выполнении неравенства $T(\widetilde{S} > \delta)$. Однако протеворечить нулевой гипотезе могут не тоько высокие значения T, но и низкие значения T. В таких случаях целесообразно использовать двухсторонние критерии.
$\mathbf{H}_0$ отвергается при выполнении по крайней мере одного из двух неравенств:

- $T(\widetilde{S}) \geq \delta_u$
- $T(\widetilde{S}) \leq \delta_l$

В случае симметричных распределений левый и правый пороги выбираются равными по модулю. В этом случае $\mathbf{H}_0$ отвергается при выполнении неравенств

- $T(\widetilde{S}) \geq \delta$
- $T(\widetilde{S}) \leq -\delta$

Односторонний и двухсторонний критерии

Соответственно под двух сторонним $p-$значением понимается вероятность
$$\mathbf{P}[|T(\widetilde{\omega})| > |T(\widetilde{S})|]$$

Под параметрическими критериями называются критерии, в которых предполагаемые вероятностные распределения могут целеком задаваться несколькими параметрами.

Тест Стьюдента для оценки значимости коэффициента корреляции Пирсона

Нулевая гипотеза $\mathbf{H_0}$:

- переменные X_1 и X_2 подчиняются двумерному нормальному распределению
- коэффициент корреляции ρ между ними равен 0

Пусть $\widetilde{S} = \{(x_{11}, x_{12}), \ldots, (x_{m1}, x_{m2})\}$ множесто пар соответствующих наблюдений переменных X_1 и X_2. Предположим, что $\hat{\rho}(\omega)$ - оценка коэффициента корреляции, полученная по случайной выборке ω, сгенерированной из предполагаемого распределения. Предполагается, что ω имеет тот же вид, что и выборка $\widetilde{S}$ и все пары генерируются независимо.

Стюдентом было показано, что статистика

$$T(\omega) = \hat{\rho}(\omega)\sqrt{\frac{m-2}{1-\hat{\rho}^2(\omega)}}$$

подчиняется распределению Стьюдента с числом степеней свободы, равным $m-2$. Пусть $\hat{\rho}(\widetilde{S}) > 0$. Тогда в качестве альтернативы $\mathbf{H}_0$ естественно предположить, что $\rho > 0$. Тогда одностороннее $p-$значение вычисляется по формуле $p = 1 - \mathbb{F}^{st}_{m-2}[T(\widetilde{S})]$, где $\mathbb{F}^{st}_{m-2}-$фунция распределения Стьдента с числом степеней свободы $m-2$. Из симметрии распределения Стьюдента следует, что одностороннее $p-$значение вычисляется по этой же формуле, а двухстороннее $p-$значение равно удвоенному одностороннему.

Оценка достоверности коэффициента корреляции с использованием $Z-$преобразования Фишера

Предположим, что переменные X_1 и X_2 подчиняются двумерному нормальному распределению с коэффициентом корреляции равным ρ. Тогда распределение статистики

$$z(\omega) = \frac{1}{2}\log\frac{1+\hat{\rho}(\omega)}{1-\hat{\rho}(\omega)}$$

хорошо аппроксимируется нормальным распределением с математическим ожиданием

$$\frac{1}{2}\log\frac{1+\rho}{1-\rho}$$

и дисперсией $\sqrt{\frac{1}{m-3}}$.

Оценка достоверности коэффициента корреляции с использованием $Z-$преобразования Фишера

Для проверки гипотезы о равенстве коэффициента корреляции ρ_0 достаточно

- вычислить $z_0 = \frac{1}{2}\sqrt{\frac{1+\rho}{1-\rho}}$
- вычислить $z(\widetilde{S}) = \frac{1}{2}\sqrt{\frac{1+\hat{\rho}(\widetilde{S})}{1-\hat{\rho}(\widetilde{S})}}$
- вычислить $T(\widetilde{S}) = [z(\widetilde{S}) - z_0]\sqrt{m-3}$

Одностороннее $p-$значение при альтернативе $\rho > 0$ вычисляется по формуле $p = 1 - \mathbb{F}_N[T(\widetilde{S})]$, где $\mathbb{F}_N-$ кумулятивная фунция нормального распределения.

Из симметрии нормального распределения следует, что одностороннее $p-$значение при альтернативе $\rho > 0$ также вычисляется по формуле $p = 1 - \mathbb{F}_N[-T(\widetilde{S})]$, а двухстороннее $p-$значение равно удвоенному одностороннему.

Имеются две выборки, содержащие независимые наблюдения переменной X:

$$\widetilde{S}_1 = (x_1, \ldots, x_{m_1})$$

$$\widetilde{S}_2 = (x_{m_1+1}, \ldots, x_{m_2})$$

Нулевая гипотеза $\mathbf{H}_0$:
Выборки $\widetilde{S}_1$ и $\widetilde{S}_2$ сгенерированы из распределений $N(\mu_1, \sigma_1^2)$ и $N(\mu_2, \sigma_2^2$ соответственно. При этом $\sigma_1 = \sigma_2 = \sigma$, $\mu_1 - \mu_2 = \delta$. Пусть
$\hat{\mu}_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} x_i$, $\hat{\mu}_2 = \frac{1}{m_2 - m_1} \sum_{i=m_1+1}^{m_2} x_i$,
$\hat{\sigma}_1 = \frac{1}{m_1 - 1} \sum_{i=1}^{m_1} (x_i - \mu_1)^2$, $\hat{\sigma}_2 = \frac{1}{m_2 - m_1 - 1} \sum_{i=m_1+1}^{m_2} (x_i - \mu_2)^2$

Проверка равества разности математических ожиданий в двух независимых группах фиксированной константе

Статистика двухвыборочного критерия

$$T = \frac{\hat{\mu_1} - \hat{\mu_2} - \delta}{\hat{\sigma}_{12}\sqrt{\frac{1}{m_1} + \frac{1}{m_2 - m_1}}},$$

где

$$\sigma_{12} = \frac{(m_1 - 1)\hat{\sigma_1} + (m_2 - 1)\hat{\sigma_2}}{m_2 - 2}$$

При генерации пар независимых выборок размером m_1 и $m_2 - m_1$ согласно $\mathbf{H}_0$ статистика T подчиняется распределению Стьюдента с числом степеней свободы $m_2 - 2$.

Проверка равенства разности математических ожиданий в двух независимых группах фиксированной константе

Одностороннее $p-$значение при альтернативах $\mu_1 - \mu_2 > \delta$ или $\mu_1 - \mu_2 < \delta$ рассчитывается как $1 - \mathbb{F}^{st}_{m-2}[T(\widetilde{S}_1, \widetilde{S}_2)]$. Из-за симметрии распределения Стьюдента вухстороннее $p-$значение равно удвоенному одностороннему. В случае, если гипотеза о равенстве дисперсий в двух сравниваемых группах является несостоятельной может быть использован критерий Уэлча. То есть используется статистика

$$T = \frac{\hat{\mu_1} - \hat{\mu_2} - \delta}{\sqrt{\frac{\hat{\sigma}_1^2}{m_1} + \frac{\hat{\sigma}_2^2}{m_2 - m_1}}}$$

Проверка равества разности математических ожиданий в двух независимых группах фиксированной константе

В случае, если данные генерируются согласно $\mathbf{H}_0$,при увеличение размеров выборок ω_1 и ω_2 распределение статистики $T(\omega_1, \omega_2)$ стремится к распределению Стьюдента с числом степеней свободы

$$\nu = \frac{\left(\frac{\hat{\sigma}_1^2}{m_1} + \frac{\hat{\sigma}_2^2}{m_2 - m_1}\right)^2}{\frac{\hat{\sigma}_1^4}{m_1^2(m_1 - 1)} + \frac{\hat{\sigma}_2^4}{(m_2 - m_1)^2(m_2 - m_1 - 1)}}$$

Значимость регрессионных коэффициентах в многомерной линейной регрессии

Пусть

$$y_1 = \beta_0 + \sum_{i=1}^{n} x_{1i}\beta_i + \epsilon_1$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$y_m = \beta_0 + \sum_{i=1}^{n} x_{mi}\beta_i + \epsilon_m$$

Случайные ошибки $\epsilon_1, \ldots, \epsilon_m$ независимы и распределены по $N(0, \sigma^2)$
Требуется проверить нулевую гипотезу о равенстве $\beta_i = \beta_i'$.
Предположим, что с помощью МНК построена модель, вычисляющая оценку Y в точке $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})$ в виде

$$\hat{y}_j = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_i x_{ji}$$

Значимость регрессионных коэффициентах в многомерной линейной регрессии

Рассчитаем стандартное отклонение для оценки регрессионного коэффициента β_i

$$\hat{\sigma}(\hat{\beta}_i) = \frac{\hat{\sigma}_{err}}{\sum_{j=1}^{m} (x_{ji} - \bar{x}_i)^2},$$

где $\bar{x}_i = \sum_{j=1}^{m} x_{ji}$

$$\sigma_{err} = \frac{\sum_{j=1}^{m} (y_j - \hat{y}_j)^2}{m - n - 1}$$

Значимость регрессионных коэффициентах в многомерной линейной регрессии

Статистика критерия

$$T = \frac{\hat{\beta}_i - \beta_i'}{\hat{\sigma}(\hat{\beta}_i)}$$

при справедливости нулевой гипотезы подчиняется распределению Стьюдента С числом степеней свободы равным $m - n - 1$. Одностороннее $p-$значение при альтернативах $\beta_i > \beta_i'$ или $\mu_1 - \mu_2 < \delta$ рассчитывается как $1 - \mathbb{F}^{st}_{m-n-1}[T(\widetilde{S})]$. Двухстороннее $p-$значение рас вно удвоенному одностороннему.

Предположим, что переменные U_1 и U_2 являются независимыми и подчиняются распределению χ^2 с числом степеней свободы d_1 и d_2 соответственно. Тогда переменная $V = \frac{U_1 d_1}{U_2 d_2}$ подчиняется $F-$распределению с параметрами d_1, d_2 -$F(d_1, d_2)$.
$F-$распределение может быть использовано для проверки нулевой гипотезы

$$\beta_1 = \beta_2 = \ldots = \beta_n = 0$$

против альтернативы, что $\beta_i =\neq 0$ хотя бы для одного $i \in \{1, \ldots, n\}$.
Каи и при рассмотрении предыдущего критерия предполагается независимость и распределенность по $N(0, \sigma^2)$ случайных ошибок.

Введем определения

- $SSM = \sum_{j=1}^{m}(\hat{y}_j - \bar{Y})$
- $SSR = \sum_{j=1}^{m}(\hat{y}_j - y_j)^2$
- $MSM = \frac{SSM}{n}$
- $MSR = \frac{SSR}{m-n-1}$

Статистика $T = \frac{MSM}{MSR}$ имеет распределение $F(n, m-n-1)$.
$p-$значение рассчитывается как $1 - \mathbb{F}^F_{(n,m-n-1)}[T(\widetilde{S})]$, где $\mathbb{F}^F_{(n,m-n-1)}$ кумулятивная функция $F-$распределения.

Предположим, что некоторый фактор X задаёт I популяций с математическими ожиданиями переменной Y равными $\mu_1, \ldots, \mu_I$.
Предполагается, что внутри популяции i Y подчиняется нормальному распределению $N(\mu_i, \sigma^2)$.
Предполжим, что чило объектов в популяциях $J_1, \ldots, J_I$. Пусть

$$\mu = \frac{1}{m}\sum_{i=1}^{I} J_i\mu_i,$$

где $m = \sum_{i=1}^{I} J_i$ Пусть $\alpha_i = \mu_i - \mu$. Однофакторная модель дипердсионного анализа может быть записана в виде

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$j = 1, \ldots, J_i$, $i = 1, \ldots, I$

Однофакторный дисперсионный анализ. Фиксированные эффекты

В модели дисперсионного анализа с фиксированными эффектами предполагается, что величины (эффекты) $\alpha_1, \dots, \alpha_i$ явволяются детерминированными параметрами.
Проверяется нулевая гипотеза $\mathbf{H}_0 : \alpha_1 = \dots = \alpha_I = 0$.
На первом шаге вычисляются оценки $\hat{\mu}, \hat{\alpha}_1, \dots, \hat{\alpha}_I$. Для этих целей может быть использован метод МНК с минимизацией функционала

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ji} - \hat{\mu} - \hat{\alpha}_i).$$

В связи с неоднозначностью экстремума вводится дополнительное ограничение $\sum_{i=1}^{I} J_i \alpha_i = 0$.

Однофакторный дисперсионный анализ. Фиксированные эффекты.

В этом случае

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ji}$$

$$\hat{\alpha}_i = \frac{1}{J_i}\sum_{j=1}^{J_i} y_{ji}$$

Вычислим

- $SSB = \sum_{i=1}^{I} J_i \hat{\alpha}_i^2$
- $SSR = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ji} - \hat{\mu} - \hat{\alpha}_i)^2$
  Отношение $\frac{SS_B(m-I)}{SS_R I}$ подчиняется $F-$распределению $F(I-1, m-I)$

В модели дисперсионного анализа со случайными эффектами предполагается, что величины (эффекты) $\alpha_1, \dots, \alpha_i$ яввляются случайными величинами, распределёнными по закону $N(0, \sigma_a^2)$. Целью анализа является оценивание μ дисперсий σ и σ_a, проверка нулевой гипотезы $\mathbf{H}_0$: $\sigma_a = 0$.
При этом

- $\hat{\sigma}^2 = \frac{1}{m-I} SS_R$
- $\hat{\sigma}_a^2 = \frac{1}{k}[\frac{1}{I-1} SS_B - \frac{1}{m-I} SS_R]$

При $J_1 = J_2 = \ldots = J_I = J$ $k = J$ Проверка нулевой гипотезы производится совершенно аналогично тому, как это делается в случае проверке нулевой гипотезы при фиксированных эффектах.
.

Проверка соответствия эмпирического распределения предполагаемой модели

Критерий χ^2 Требуется по выборке значений переменной X $\widetilde{S} = (x_1, \ldots, x_m)$ оценить соответствует ли эмприческое распределеие предполагаемой теоретической модели Предполагается, что множество значений переменной X разбито на k подмножеств - ячеек. Предположим, что согласно теоретической модели вероятности ячеек составляют $P_1, \ldots, P_k$

Статистика критерия

$$T_{\chi^2}(\widetilde{S}) = \sum_{i=1}^{k} \frac{(mP_i - m_i)^2}{mP_i}$$

При больших m распределение статистики T_{χ^2} при генерации данных согласно теоретической модели аппроксимируется распределением χ^2 с числом степеней свободы равным $k-1$.

Возможность аппроксимации распределения T_{χ^2} при больших m распределением χ^2. Справедливости аппроксимации при равенстве числа ячеек двумвытекает из возможности аппроксимации при больших m биномиального распределения $B(m,p)$ нормальным распределением $N(mp, mp(1-p))$. То есть вероятность попадания m_1 наблюдений в ячейку 1 или соответственно $m - m_1$ наблюдений в ячейку 2 может быть аппроксимирована плотностью нормального распределения $N(mp, mp(1-p))$. Очевидно также, что

$$T_{\chi^2} = \frac{(m_1 - mp)^2}{mp} + \frac{[m - m1 - m(1-p)]^2}{m(1-p)} =$$

$$= [\frac{m_1 - mp}{\sqrt{mp(1-p)}}]^2$$

Из возможности аппроксимации m_1 нормальным распределением следует возможность аппроксимации величины $\frac{m_1 - mp}{\sqrt{mp(1-p)}}$ нормальным распределением $N(0,1)$. Откуда следует возможность аппроксимации распределения статистики T_{χ^2} распределением χ^2 с одной степенью свободы. Применение теста χ^2 возможно при достаточно больших объёмах выборки $\widetilde{S}$. При этом достаточно большим должно быть число наблюдений в каждой из ячеек. Рекомендуемым требованием является выполнение неравенства $mP_i \geq 10$.

Проверка соответствия эмпирического распределения предполагаемой модели

G-критерий Альтернативой критерию χ^2 является $G-$тест со статистикой критерия

$$T_G(\widetilde{S}) = 2\sum_{i=1}^{k} m_i \log \frac{m_i}{mp_i}$$

При больших m распределение статистики T_G при генерации данных согласно теоретической модели также аппроксимируется распределением χ^2 с числом степеней свободы равным $k-1$.

$G-$критерий являктся критерием максимального правдоподобия. Нулевая гипотеза о том, что верояности ячеек задаются равенствами

$$p_1 = p_1^0, \ldots, p_k = p_k^0,$$

проверяются против альтернативной гипотезы, что вероятности ячеек соответствуют оценкам максимального правдоподобия, то есть задаются равенствами

$$p_1 = \frac{m_1}{m}, \ldots, p_k = \frac{m_k}{m}$$

.

Критерий отношения правдоподобий основан на сравнении отношения правдоподобия

$$l_{rel}(\widetilde{S}) = \log[\frac{L(p_1^0,\dots,p_k^0|\widetilde{S})}{L(\frac{m_1}{m},\dots,\frac{m_k}{m}|\widetilde{S})}]$$

с некоторым порогом δ

$$l_{rel}(\widetilde{S}) = \log[\frac{\prod_{i=1}^{k} p_i^{m_i}}{\prod_{i=1}^{k}(\frac{m_i}{m})^{m_i}}] =$$

$$= \sum_{i=1}^{k} m_i \log\frac{mp_i}{m_i}$$

Вместо правила $l_{rel}(\widetilde{S}) \geq \delta$ очевидно можно использовать правило $l_{rel}(\widetilde{S})\kappa \geq \frac{\delta}{\kappa}$ при $\kappa > 0$ и $l_{rel}(\widetilde{S})\kappa \leq \frac{\delta}{\kappa}$ при $\kappa < 0$.
Разложим величину $l_{rel}(\widetilde{S})\kappa$ в окрестности точки $\frac{mp_i}{m_i} = 1$ в ряд Тейлора. Обозначим $\delta_i = \frac{mp_i - m_i}{m_i}$. Тогда $\frac{mp_i}{m_i} = 1 + \delta_i$ $m_i = \frac{mp_i}{1+\delta_i}$.
Разложим каждый из $\log(1 + \delta_i)$ в ряд Тейлора

$$\sum_{i=1}^{k} \frac{mp_i}{1+\delta_i} \log(1+\delta_i)$$

Сумма первых членов разложения: $\sum_{i=1}^{k} \frac{mp_i\delta_i}{1+\delta_i} = \sum_{i=1}^{k} m_i \frac{mp_i - m_i}{m_i}\kappa = 0.$

Сумма вторых членов разложения:

$$\sum_{i=1}^{k}\frac{-\delta_i^2}{2}\frac{mp_i}{1+\delta_i}\kappa=-\sum_{i=1}^{k}\frac{(mp_i-m_i)^2(1+\delta_i)}{2mp_i}\kappa$$

в окрестности точки $\frac{mp_i}{m_i}=1$ близка к сумме

$$\sum_{i=1}^{k}\frac{-\kappa(mp_i-m_i)^2}{2mp_i}$$

и равна статистике критерия χ^2 при $\kappa=-2$.

Проверка соответствия эмпирического распределения предполагаемой модели

Критерий Колмогорова-Смирнова Требуетс я по выборке значений переменной X $\widetilde{S} = (x_1^t, \ldots, x_m^t)$ проверить нулевую гипотезу о независимой генерации наблюдений из $\widetilde{S}$ из распределения с предполагаемой кумулятивной функцией распределения $\mathbb{F}(x)$. Предположим, что $\mathbb{F}_m(x)$ является эмприческим распределеием, рассчитанными по выборке $\omega = (x_1, \ldots, x_m)$, сгенерированной согласно нулевой гипотезе.

$$\mathbb{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} I(i,x),$$

где $I(i,x) = 1$ при $x_i \leq x$ и $I(i,x) = 0$ при $x_i > 0$
Статистика Колмогорова

$$D_m = \sup_x |\mathbb{F}_m(x) - \mathbb{F}(x)|$$

$\sup_x |\mathbb{F}_m(x) - \mathbb{F}(x)|$ -точная верхняя грань множества $\{z = |\mathbb{F}_m(x) - \mathbb{F}(x)|, x \in \mathbb{R}\}$
Справедлива Теорема Гливенко-Кантелли При генерации данных из распределения с функцией $\mathbb{F}(x)$ $\forall \epsilon \lim_{m \to \infty} \mathbf{P}\{D_m > \epsilon\} = 0$
Броуновский мост $B(t)$ Броуновский мост является непрерывным по времени стохастическим процессом Для броуновского моста, заданного на отрезке $[0, T]$, справедливо равенство $B(0) = B(1)$. Пусть $W(t)$ - винеровский процесс. Вероятностное распределение для $B(t)$ выражается через условное распределением для винеровского процесса: $\forall t \in [0, T]$ и для произвольно интервала a

$$\mathbf{P}\{B(t) \in a\} = \mathbf{P}\{W(t) \in a | W(T) = 0\}$$

Распределение Колмогорова - распеделение случайной переменной

$$K = \sup_{t \in [0,1]} |B(t)|$$

Случайная величина K имеет распределение Колмогорова. Её функция распределения может быть записана как $\sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}$ при $x > 0$ и равна 0 при $x \leq 0$

Теорема Колмогорова При генерации данных из $\mathbb{F}(x)$

$$\sqrt{m} D_m \to K$$

по распределению при $m \to \infty$, то есть $\forall x$

$$\lim_{m\to\infty} \mathbf{P}(\sqrt{m} D_m \le x) = \mathbf{P}(K \le x)$$

Как видно,распределение статистики Колмогорова при $m \to \infty$ не зависит от от конкретного вида распределения Для того, чтобы оценить соответствие данных распределению $\mathbb{F}(x)$ на уровне α достаточно сравнить рассчитанное по данным значени $\sqrt{m} D_m(\widetilde{S})$ с $K_{1-\alpha}$, то есть с $(1-\alpha)-$квантилью распределения Колмогорова

При $\sqrt{m}D_m(\widetilde{S}) \geq K_{1-\alpha}$ нулевая гипотеза о генерации данных из распределения $F(x)$ отвергается на уровне α. В противном случае нулевая гипотеза на этом уровне значимости отвергнута быть не может. В качестве $p-$значений используются вероятности $\mathbf{P}(K \geq \sqrt{m}D_m(\widetilde{S}))$.

Представленная выше схема является корректной, если производится проверка одного единственного распределения (например,при фиксированных параметрах). В тех случаях, когда оценивание параметров производится по обучающей выборке $\widetilde{S}$, прямое применение критерия Колмогорова-Смирнова приводит к существенному занижению $p-$значений и, как следствие, к ошибочному принятию нулевой гипотезы о соответствии проверяемого распределения данным.

Критерий Лиллеофорса

Для проверки соответствия $\widetilde{S}$ нормального распределения, параметры которого также оцениваются по выборке $\widetilde{S}$,вместо теоретически полученного распределения Колмогорова, использется распределение Лиллеофорса, рассчитанное с помощью методов Монте-Карло.

Критерий Колмогорова-Смирнова Имеется две выборки

$$\widetilde{S}_1 = \{x^t_1, \dots, x^t_{m_1}\}$$

$$\widetilde{S}_2 = \{x^t_{m_1+1}, \dots, x^t_{m_2}\}$$

Требуется проверить нулевую гипотезу о независимой генерации объектов выборок $\widetilde{S}_1$ и $\widetilde{S}_2$ из одного и того же вероятностного распределения по эмпирическим функциям распределениям $\mathbb{F}_{m_1}(x)$ и $\mathbb{F}_{m_2-m_1}(x)$ Пусть

$$D_{m_1,m_2-m1} = \sup_x |\mathbb{F}_{m_1}(x) - \mathbb{F}_{m_2-m_1}(x)|$$

- статистика, вычисляемая по двум выборкам $\widetilde{\omega}_1 = (x_1, \dots, x_{m_1})$ и $\widetilde{\omega}_2 = (x_{m_1+1}, \dots, x_{m_2})$.

Сравнение независимых групп с помощью критерия Колмогорова-Смирнова

Теорема Колмогорова При независимой генерации объектов выборок $\widetilde{\omega}_1$ и $\widetilde{\omega}_2$ из одного и того же распределения

$$\sqrt{\frac{m_1(m_2-m_1)}{m_2}}D_{m_1,m_2-m_1} \to K$$

по распределению при $m_1, m_2 \to \infty$, то есть $\forall x$

$$\lim_{m_1,m_2\to\infty} \mathbf{P}\Big(\sqrt{\frac{m_1(m_2-m_1)}{m_2}}D_{m_1,m_2-m_1} \leq x\Big) = \mathbf{P}(K \leq x)$$

Сравнение независимых групп с помощью критерия Колмогорова-Смирнова

Как видно,эмпирическое распределение $\sqrt{\frac{m_1(m_2-m_1)}{m_2}}D_{m_1,m_2-m_1}$ а при $m_1, m_2 \to \infty$ не зависит от конкретного вида истинного распределения, из которого генерируются данные. Для того, чтобы оценить соответствие выборок $\widetilde{S}_1$ и $\widetilde{S}_2$ нулевой гипотезе на уровне α достаточно сравнить рассчитанное по данным значени $\sqrt{\frac{m_1(m_2-m_1)}{m_2}}D_{m_1,m_2-m_1}$ с $K_{1-\alpha}$, то есть с $(1-\alpha)-$квантилью распределения Колмогорова В качестве $p-$значений используются вероятности $\mathbf{P}(K \geq \sqrt{\frac{m_1(m_2-m_1)}{m_2}}D_{m_1,m_2-m_1}(\widetilde{S}_1, \widetilde{S}_2))$.

Тест Манна-Уиттни Имеется две выборки

$$\widetilde{S}_1 = \{x_1^t, \dots, x_{m_1}^t\}$$

$$\widetilde{S}_2 = \{x_{m_1+1}^t, \dots, x_{m_2}^t\}$$

Нулевая гипотеза $\mathbf{H}_0$ - обе выборки независимо извлечены из одной и той же генеральной совокупности. При генерации выборок $\widetilde{\omega}_1 = \{x_1, \dots, x_{m_1}\}$ и $\widetilde{\omega}_2 = \{x_{m_1+1}, \dots, y_{m_2}\}$ в соответствии с $\mathbf{H}_0$ очевидно справедливо равенство $P(x_i > y_j) = P(x_i < y_j)$

Статистика критерия Используем следующую процедуру для вычисления статистики критерия.

- Сформируем объединённую выборку $\widetilde{\omega}_{1,2} = \widetilde{\omega}_1 \cup \widetilde{\omega}_2$
- Для каждого наблюдения из $\widetilde{\omega}_{1,2}$ вычислим его ранг. Если несколько наблюдений совпадают, то ранжирование на этом участке выбирается произвольно. Далее в группе совпадающих наблюдений $\widetilde{g}$ вычисляется средний ранг, который и присваивается каждому наблюдению из $\widetilde{g}$.
- Вычислим $R_1 = \sum_{j=1}^{m_1} r_{1,2}(x_j)$.
- Вычислим $U_1 = R_1 - \frac{m_1(m_1+1)}{2}$
- Вычислим $R_2 = \sum_{j=1}^{m_2-m_1} r_{1,2}(x_j)$.
- Вычислим $U_2 = R_2 - \frac{(m_2-m_1)(m_2-m_1+1)}{2}$

Через $r_{1,2}(x_j)$ обозначен ранг наблюдения x_j в объединённой выборке $\widetilde{\omega}_1 \cup \widetilde{\omega}_2$.
Справедливо равенство $U_1 + U_2 = m_1(m_2 - m_1)$
В качестве статистики критерия U-теста используется

$$T_u(\widetilde{\omega}_1, \widetilde{\omega}_2) = \min(U_1, U_2)$$

При условии $m_1(m_2 - m_1) > 20$ и при отсутствии совпадающих наблюдений в выборках распределение статистики T_u хорошо аппроксимируется нормальным распределением $N(\mu, \sigma)$, где $\mu = \frac{m_1(m_2 - m_1)}{2}$ $\sigma = \sqrt{\frac{m_1(m_2 - m_1)(m_2 + 1)}{12}}$

В случае наличия совпадающих наблюдений распределение статистики T_u также о аппроксимируется нормальным распределением $N(\mu, \sigma)$. Однако стандартное отклонение вычисляется по формуле

$$\sigma = \sqrt{\frac{m_1(m_2 - m_1)}{m_2(m_2 - 1)} \times [\frac{m_2^3 - m_2}{12} - \sum_{j=1}^{k} \frac{t_j^3 - t_j}{12}]},$$

где

- k-число групп совпадающих наблюдений
- t_j- число наблюдений в группе j

В случае малых выборок используются заранее рассчитанные точные значения квантилей распределения статистики T_u, собранные в специальные таблицы. В этом случае для оценки значимости отклонения $\mathbf{H}_0$ на уровне α достаточно сравнить $T_u(\widetilde{S}_1, \widetilde{S}_2)$ с $(1-\alpha)-$квантилью,представленной в таблице. При достижении и ли превышении указанной квантили нулевая гипотеза отвергается. В случае аппроксимации нормальным распределением $p = 1 - \mathbb{F}_N[T_u(\widetilde{S}_1, \widetilde{S}_2)]$

Оценка достоверности связи бинарных показателей по таблице сопряжённости

Точный тест Фишера Требуется по таблице сопряжённости двух бинарных признаков $X \in \{x_1, x_2\}$, $Y \in \{y_1, y_2\}$ оценить достоверность наличия связи между ними.

Таблица: Таблица 1

	x_1	x_2
y_1	a	b
y_2	c	d

Проверяется нулевая гипотеза $\mathbf{H}_0$ о независимости Y и X. В этом случае вероятность комбинации (a, b, c, d) подчиняется гипергеометрическому распределению, описывающего вероятность возникновения различных комбинаций двух групп объектов при выборе без возвращения из исходной выборки. Считаем общие количества наблюдений с $Y = y_1$, $Y = y_2$, $X = x_1$, $X = x_2$ фиксированными и равными их значениям на исходной выборке $\widetilde{S}$.

Из независимости Y от X следует равновероятность всевозможных перестановое позиций X относительно фиксированных позиций Y. Откуда следует, что вероятность a раз встретить комбинацию $Y = y_1, X = x_1$ может быть определена по представленной далее формуле. Очевидно, что задание aоднозначно задаёт b,c и d.

$$P(a,b,c,d) = \frac{C_{a+c}^{a} C_{b+d}^{b}}{C_m^{a+b}}$$

Пусть $m = a + b + c + d$. Для вычисление $p-$значений может быть использована процедура, связанная с выбором экстремальных значений в таблице сопряжённости.Например, мы можем выбрать минимальное из четырёх значений. Допустим,что минимальным значением является a.

Тогда $p-$значение вычисляется как сумма

$$p = \sum_{(z_1,z_2,z_3,z_4)\in\widetilde{Z}} P(z_1,z_2,z_3,z_4) =$$

$$\sum_{(z_1,z_2,z_3,z_4)\in\widetilde{Z}} \frac{C_{a+c}^{z_1} C_{b+d}^{z_2}}{C_m^{a+b}}$$

где $\widetilde{Z}-$ множество векторов (z_1, z_2, z_3, z_4), для которых

- $z_1 \in \{0, \ldots, a\}$
- $z_1 + z_2 = a + b$
- $z_1 + z_3 = a + c$

Оценка достоверности связи бинарных показателей по таблице сопряжённости

критерий χ^2 Соответствие эмпирического распределения в ячейках таблицы сопряжённости бинарных переменных X и Y истинной вероятности попадания в ячейки может быть оценено с помощью статистики критерия χ^2:

$$T_{\chi^2} = \sum_{k=1}^{2} \sum_{k'=1}^{2} \frac{(m_{kk'} - mp_{kk'})^2}{mp_{kk'}}$$

В случае независимости переменных X и Y очевидно $p_{kk'} = p_k^x p_{k'}^y$, где

- $p_k^x = P(X = x_k)$
- $p_{k'}^y = P(Y = y_{k'})$

Оценка достоверности связи бинарных показателей с помощью критерия χ^2

Используем очевидные оценки вероятностей $p_k^x, p_{k'}^y$

- $p_1^x = \frac{a+c}{a+b+c+d}$
- $p_2^x = \frac{b+d}{a+b+c+d}$
- $p_1^y = \frac{a+b}{a+b+c+d}$
- $p_2^y = \frac{c+d}{a+b+c+d}$

Получаем, что

$$T_{\chi^2} = \frac{(a - mp_1^x p_1^y)^2}{mp_1^x p_1^y} + \frac{(b - mp_2^x p_1^y)^2}{mp_2^x p_1^y} + \frac{(c - mp_1^x p_2^y)^2}{mp_1^x p_2^y} + \frac{(d - mp_2^x p_2^y)^2}{mp_2^x p_2^y}$$

Оценка достоверности связи бинарных показателей с помощью критерия χ^2

После преобразований получаем

$$T_{\chi^2} = \frac{(ad - bc)^2}{(a+b)(c+d)(b+d)(a+c)} = \phi^2 * m$$

Предположим, что имеется последовательность элементов, принимающих два различных значения $\{+, -\}$. Целью является проверка нулевой гипотезы о независимой генерации элементов последовательности из некоторого фиксированного распределения Бернулли. Под сериями будем понимать нерасширяемые подпоследовательности одинаковых элементов.
Например, в последовательности

$$+++--+---++++----$$

сериями являются

- $+$ в первых трёх позициях
- $-$ в позициях 4 и 5
- $+$ в позиции 6
- $-$ в позициях 7-9

- $+$ в позициях 10-13
- $-$ в позициях 14-17

Таким образом, общее число серий составляет 6, из которых 3 состоят из $+$ и 3 состоят из $-$.
Пусть

- n_+ - число $+$ в последовательности
- n_- - число - в последовательности
- R - число серий в последовательности

Статистика критерия

$$T_{ww} = \frac{R - \bar{R}}{\hat{\sigma}_R},$$

где

$$\bar{R} = \frac{2n_+n_-}{n_+ + n_-} + 1,$$

$$\hat{\sigma}_R^2 = \frac{(2n_+n_-)(2n_+n_- - n_+ - n_-)}{(n_+ + n_-)^2(n_+ + n_- - 1)}$$

При $n_+ > 10$, $n_- > 10$ распределение статистики T_{ww} хорошо аппроксимируется нормальным распределением $N(0,1)$. Для малых выбоок значимость может быть оцененa с использованием специальных таблиц, содержащих рассчитанные значения соответствующих квантилей.

Оценка статистической значимости эффектов по повторным измерениям

Целью многих исследований является статистическая оценка влияния некоторого воздействия на показатель X. При этом оценка производится по выборке вида $\widetilde{S} = \{(x^t_{1,1}, x^t_{2,1}), \dots, (x^t_{1,m}, x^t_{2,m})\}$, где $x_{1,j}$ значение показателя X до произведённого воздействия, $x^t_{2,j}$ значение показателя X после произведённого воздействия.

Знаковый ранговый критерий Уилкоксона

Нулевая гипотеза: каждая пара наблюдений, сделанных до и после произведённого воздействия, генерируется независимо из одного и того же вероятностного распределения. При этом разности значений X до и после произведённого воздействия распределены симметрично относительно 0.

Пусть $\widetilde{\omega} = \{(x_{1,1}, x_{2,1}), \ldots, (x_{1,m}, x_{2,m})\}$ - произвольная выборка, генерируемая в соответствии с нулевой гипотезой. Исключим из выборки все пары, для которых выполняется равенство $x_{1,i} = x_{2,i}$, и упорядочим оставшиеся m_r пар по величине $|x_{2,i} - x_{1,i}|$. Обозначим ранг пары i в образовавшейся последовательности через R_i.
Статистикой критерия является сумма

$$T_{wsr}(\widetilde{\omega}) = \sum_{i=1}^{m_r} R_i sgn(x_{2,i} - x_{1,i}).$$

При $m_r \to \infty$ распределение статистики $z(\widetilde{\omega}) = \frac{T_{wsr}}{\hat{\sigma}_{wsr}}$, где $\hat{\sigma}_{wsr} = \sqrt{\frac{m_r(m_r+1)(2m_r+1)}{6}}$ стремится к нормальному распределению $N(0,1)$

При $m_r > 20$ аппроксимация нормальным распределением считается удовлетворительной и в качестве одностороннего $p-$значения могут быть использованоа величина $1 - \mathbb{F}_N[T(\widetilde{S})]$. Двустороннее $p-$значение равно удвоенному одностороннему. Для малых выбоок значимость может быть оценена с использованием специальных таблиц, содержащих рассчитанные значения соответствующих квантилей.