Пример 1

Компания "МегаФон-Эконом" разработала три тарифных плана: "Базовый", "Оптимальный" и "Премиум". Маркетологи хотят проверить гипотезу о том, что выбор тарифа зависит от уровня месячного дохода клиента.

1. Формулировка гипотез:

Нулевая гипотеза (H_0): Признаки «Уровень дохода» и «Выбор тарифа» независимы.

$$H_0 : P(\text{Тариф} = i, \text{Доход} = j) = P(\text{Тариф} = i) \cdot P(\text{Доход} = j) \quad \forall i, j$$

Альтернативная гипотеза

H_1 : Признаки «Уровень дохода» и «Выбор тариф» зависимы.

$$H_1 : \exists i, j : P(\text{Тариф} = i, \text{Доход} = j) \neq P(\text{Тариф} = i) \cdot P(\text{Доход} = j)$$

Результаты опроса 300 клиентов:

Уровень дох. ╲ Тариф	Базовый	Оптимальн	Премиум	Σ
Низкий (до 40 тыс)	60	30	10	100
Средний (40-80 тыс)	30	50	20	100
Высокий (свыше 80 тыс)	10	20	70	100
Σ	100	100	100	300

3. Расчет ожидаемых частот (p_{ij})

Если гипотеза H_0 верна, то ожидаемая частота для ячейки (i, j) рассчитывается по формуле:

$$p_{ij} = \frac{(\text{сумма по строке } i) \times (\text{сумма по столбцу } j)}{\text{Общее кол-во } n}$$

Рассчитаем ожидаемые частоты для всех ячеек. Например, для ячейки "Низкий доход / Базовый тариф":

$$p_{11} = \frac{100 \times 100}{300} \approx 33.33$$

Полная таблица ожидаемых частот:

Уровень дох. ╲ Тариф	Базовый	Оптимальн	Премиум	Σ
Низкий (до 40 тыс)	33.33	33.33	33.33	100
Средний (40-80 тыс)	33.33	33.33	33.33	100
Высокий (свыше 80 тыс)	33.33	33.33	33.33	100
Σ	100	100	100	300

Тестовая статистика хи-квадрат Пирсона вычисляется по формуле:

$$\chi^2_в = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\nu_{ij} - n \cdot p_{ij})^2}{n \cdot p_{ij}}$$

где $m = 3$ — число строк, $k = 3$ — число столбцов, ν_{ij} — наблюдаемая частота, p_{ij} — ожидаемая частота.

Произведем расчет для всех ячеек:

$$\chi^2_в = \frac{(60-33.33)^2}{33.33} + \frac{(30-33.33)^2}{33.33} + \frac{(10-33.33)^2}{33.33} + \frac{(30-33.33)^2}{33.33} + \frac{(50-33.33)^2}{33.33} + \\ + \frac{(20-33.33)^2}{33.33} + \frac{(10-33.33)^2}{33.33} + \frac{(20-33.33)^2}{33.33} + \frac{(70-33.33)^2}{33.33} = 114$$

Число степеней свободы:

$$l = (m-1)(k-1) = (3-1)(3-1) = 4$$

Уровень значимости: $\alpha = 0.05$

Критическое значение: По таблице распределения χ^2 для $df = 4$ и $\alpha = 0.05$ находим:

$$\chi^2_{крит} = 9.49$$

$$\chi^2_в = 114.00 > \chi^2_{крит} = 9.49$$

Существует статистически значимая зависимость между уровнем дохода клиентов и выбором тарифа мобильной связи ($\chi^2 = 114.00$; $l = 4$; $p < 0.001$). Анализ таблицы сопряженности показывает, что клиенты с низким доходом предпочитают "Базовый" тариф, а клиенты с высоким доходом — "Премиум". Это позволяет компании направлять рекламные кампании на конкретные доходные группы для повышения эффективности маркетингового бюджета.

Критерий согласия Колмогорова.

Теорема. Пусть случайная величина Y равномерно распределена на $[0,1]$, пусть F^{-1} обратная функция к функции распределения $F(x)$. Тогда случайная величина $X = F^{-1}(Y)$ распределена по закону $F(x)$.

Доказательство. Для равномерно распределенной на $[0,1]$ случайной величины верно, что $P(Y < x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases}$. Следовательно, поскольку $0 \le F(x) \le 1$, то для всех значений $x \in (-\infty; +\infty)$

$$P(X < x) = P(F^{-1}(Y) < x) = P(Y < F(x)) = F(x).$$

Пусть $X_1, X_2, \ldots . X_n$ выборка из закона распределения, задаваемого ф.р. $F(x)$

Рассмотрим выборочную функцию распределения:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} I(x - X_k),\ x \in \mathbb{R},\ I(x) = \begin{cases} 1 & x > 0 \\ 0 & x \le 0 \end{cases},$$

Теорема 1. Если функция распределения непрерывна, то закон распределения статистики D_n не зависит от вида функции $F(x)$.

Доказательство. Рассмотрим сначала случай строго монотонной функции $F(x)$. Наряду с выборкой $X_1, X_2, \ldots . X_n$, представляющей реализацию последовательности независимых случайных величин, распределенных по закону $F(x)$ рассмотрим $Y_1, Y_2, \ldots . Y_n$, где $Y_k = F(X_k)$.

Тогда, положив $y = F(x), x = F^{-1}(y)$, можем записать

$$\sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F(x) \right| = \sup_{y \in [0,1]} \left| \widehat{F_n}(F^{-1}(y)) - y \right| = \sup_{y \in [0,1]} \left| \widehat{\widehat{F}_n}(y) - y \right|$$

$$\widehat{F_n}(F^{-1}(y)) = \frac{1}{n}\sum_{k=1}^{n} I(X_k < F^{-1}(y)) = \frac{1}{n}\sum_{k=1}^{n} I(Y_k < y) = \widehat{\widehat{F}}_n(y).$$

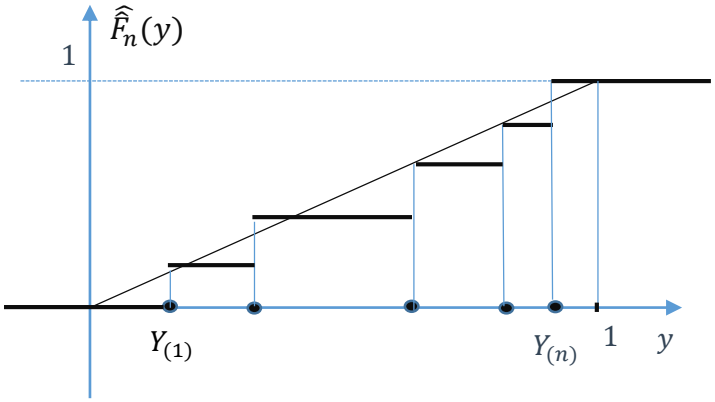
Здесь $\widehat{\widehat{F}}_n(y)$- эмпирическая функция распределения, выборки $Y_1, Y_2, \ldots . Y_n$, распределенной равномерно на $[0,1]$. Таким образом закон распределения статистики D_n не зависит от вида функции $F(x)$.

Если $F(x)$ не является строго монотонной, то в доказательстве теоремы $\sup_{x \in \mathbb{R}} \left| \widehat{F_n}(x) - F(x) \right|$ надо заменить на $\sup_{x \in M} \left| \widehat{F_n}(x) - F(x) \right|$, где M - множество строгой монотонности $F(x)$.

Замечание. Вычисление статистики Колмогорова.

Пусть $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$-вариационный ряд, построенный по пересчитанной выборке. Тогда

$$D_n = \sup_{y\in[0,1]} \left|\hat{\hat{F}}_n(y) - y\right| = \max_{1\le k\le n} \max\left\{\left|Y_{(k)} - \frac{k}{n}\right|, \left|Y_{(k)} - \frac{k-1}{n}\right|\right\} = \max_{1\le k\le n}\left\{\left|Y_{(k)} - \frac{2k-1}{2n}\right| + \frac{1}{2n}\right\}$$



В самом деле, для произвольного отрезка $[a,b]$ и любой точки $x \in [a,b]$ имеет место очевидное соотношение: $\max\{|x-a|, |x-b|\} = \left|x - \frac{a+b}{2}\right| + \frac{b-a}{2}$.

Определение. Статистикой Колмогорова называется $D_n = \sup_{x\in\mathbb{R}} \left|F_n(x) - F(x)\right|$

Критерий Колмогорова для $n \le 20$

Смирнов рассчитал и табулировал критические точки закона распределения D_n для различных уровней значимости , то есть такие $k_{1-\alpha}(n)$, что $P\left(D_n \ge k_{1-\alpha}(n)\right) = \alpha$

Если $D_n \ge k_{1-\alpha}(n)$, то основная гипотеза отклоняется

Если $D_n < k_{1-\alpha}(n)$, то основная гипотеза принимается

Теорема 2 (Колмогоров).

Определим функцию Колмогорова $K(x) = \sum_{j=-\infty}^{\infty} (-1)^n e^{-2k^2x^2}, x > 0$. Тогда для любой непрерывной $F(x)$. статистика $\sqrt{n}D_n$ при $n \to \infty$ по распределению сходится к функции Колмогорова:

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n < x\right) = K(x),\ x > 0.$$

Критерий Колмогорова для $n > 20$

Если $D_n \geq \frac{\lambda_{1-\alpha}}{\sqrt{n}}$, то основная гипотеза отклоняется

Если $D_n < \frac{\lambda_{1-\alpha}}{\sqrt{n}}$, то основная гипотеза принимается

где $\lambda_{1-\alpha}$-квантиль уровня $1-\alpha$ функции $K(x)$, то есть $K(\lambda_{1-\alpha}) = 1-\alpha$.

При этом ошибка первого рода критерия равна

$$P\left(D_n \geq \frac{\lambda_{1-\alpha}}{\sqrt{n}} | H_0\right) = P\left(\sqrt{n}D_n \geq \lambda_{1-\alpha}\right) = 1 - K(\lambda_{1-\alpha}) = \alpha\ .$$

Квантили функции Колмогорова

$1-\alpha$	0.9	0.95	0.98	0.99
$\lambda_{1-\alpha}$	1.224	1.358	1.515	1.628

$$D_n = \sup_{y\in[0,1]} \left|\hat{F}_n(y) - y\right| = \max_{1\leq k\leq n} \max\left\{\left|Y_{(k)} - \frac{k}{n}\right|, \left|Y_{(k)} - \frac{k-1}{n}\right|\right\} = \max_{1\leq k\leq n}\left\{\left|Y_{(k)} - \frac{2k-1}{2n}\right| + \frac{1}{2n}\right\}$$

Пример 2

Пассажир, приходящий в случайные моменты времени на автобусную остановку, в течение пяти поездок фиксировал своё время ожидания автобуса: 5,1; 3,7; 1,2; 9,2; 4,8 мин. Проверить гипотезу о том, что время ожидания автобуса равномерно распределено на отрезке [0; 10] на уровне значимости 0,05.

Решение. Здесь $F(x) = \frac{x}{10}, 0 \leq x \leq 10,$ - функция распределения проверяемого закона. Упорядочим и пересчитаем выборку: $Y_k = F(X_k) = \frac{X_k}{10}$.

$X_{(k)}$	$Y_{(k)}$	$\frac{2k-1}{2n}$	$\left\|Y_{(k)}-\frac{2k-1}{2n}\right\| + \frac{1}{2n}$
1.2	0.12	0.1	0.12
3.7	0.37	0.3	0.17
4.8	0.48	0.5	0.12
5.1	0.51	0.7	0.29
9.2	0.92	0.9	0.02

Таким образом, $D_5 = 0.29$, в то время, как критическая точка $k_{0.95}(5) = 0.56$

Таким образом, гипотеза о равномерном законе распределения времени ожидания принимается.

Критерий Смирнова

Пусть имеется две выборки

$X_1, X_2, \ldots . X_n$ выборка из закона распределения, задаваемого ф.р. $F(x)$

$Y_1, Y_2, \ldots . Y_m$ выборка из закона распределения, задаваемого ф.р. $G(x)$

Проверяется гипотеза об однородности выборок:

H_0 : законы распределения выборок совпадают $F(x) = G(x)$

Теорема 3 (Смирнов)

Пусть $F_n(x)$ и $G_m(x)$ – эмпирические функции распределения двух рассматриваемых выборок, пусть $D_{n,m} = \sup\limits_{-\infty<x<\infty} \left|F_n(x) - G_m(x)\right|$ - статистика Смирнова. Тогда в предположении непрерывности общего закона распределения двух выборок

$F(x)=G(x)$ статистика $\sqrt{\frac{nm}{n+m}}D_{n,m}$ при $n\to\infty, m\to\infty$ по распределению сходится к функции Колмогорова:

$$\lim_{\substack{n\to\infty\\ m\to\infty}} P\left(\sqrt{\frac{nm}{n+m}}D_{n,m} < x\right) = K(x),\ x>0.$$

Пример 3

По приведенным в таблице данным проверить по критерию Смирнова что законы распределения выборок совпадают. Принять $\alpha = 0.05$

Первая выборка	Вторая выборка
-1.723	0.985
-1.517	0.862
-0.442	0.916
-1.245	0.673
-0.91	-1.044
0.262	0.069
-1.706	-0.756
-0.998	0.697
0.108	-0.182
-0.107	-0.644

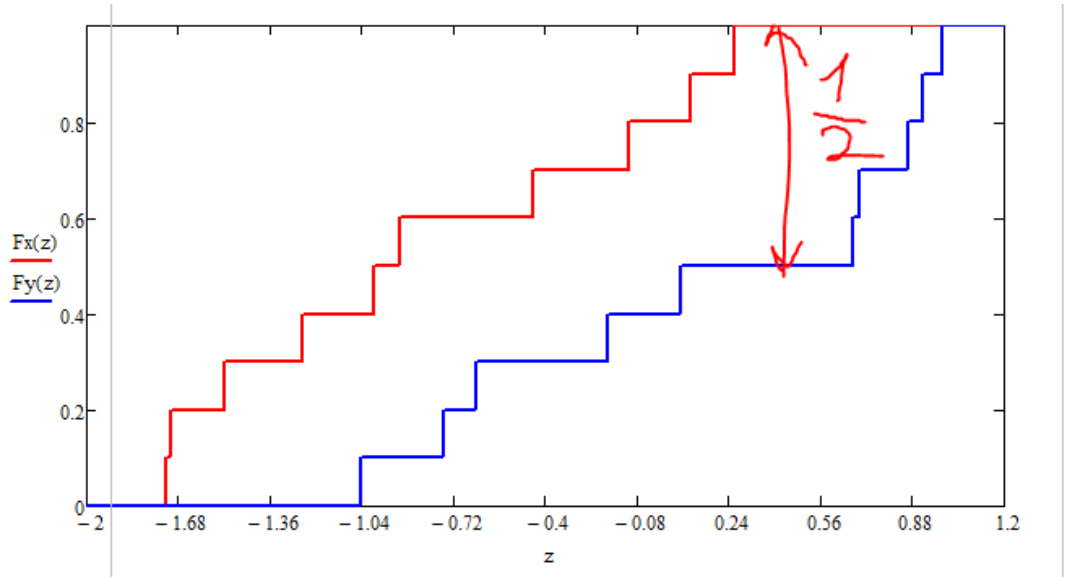
Составим функцию распределения для заданных выборок

$$x := \begin{pmatrix} -1.723 \\ -1.517 \\ -0.442 \\ -1.245 \\ -0.91 \\ 0.262 \\ -1.706 \\ -0.998 \\ 0.108 \\ -0.107 \end{pmatrix} \qquad y := \begin{pmatrix} 0.985 \\ 0.862 \\ 0.916 \\ 0.673 \\ -1.044 \\ 0.069 \\ -0.756 \\ 0.697 \\ -0.182 \\ -0.644 \end{pmatrix}$$

$$ind(k) := \begin{cases} 0 & \text{if } k < 0 \\ 1 & \text{if } k \geq 0 \end{cases}$$

$$Fx(z) := \frac{1}{10}\left(\sum_{i=0}^{9} ind\left(z - x_{i,0}\right)\right)$$

$$Fy(z) := \frac{1}{10}\cdot\sum_{i=0}^{9} ind\left(z - y_{i,0}\right)$$



В данном случае по графику видно в какой точке будет находится максимум разности функций распределения. Найдём статистику Колмогорова следующим образом, каждый скачок функции распределения равен $\frac{1}{n}$, в данном примере $n = 10$, эти скачки происходят в точках равных элементам выборки, отсортируем обе выборки и посмотрим насколько одна «убегает» от другой

Первая выборка	Вторая выборка
-1.723	
-1.706	
-1.517	

-1.245

-1.044

-0.998 -0.756

-0.91

-0.756

-0.644

-0.442

-0.182

-0.107

0.069

0.108

0.262

0.673

0.697

0.862

0.916

0.985

Например в точке -1.246 функция распределения первой выборки равна $\frac{4}{10}$ (4 элемента выборки из 10 меньше, чем -1.246), а функция распределения второй выборки в этой точке равна 0, найдём участок на котором одна из функций дальше всего убежала, это происходит после значения 0.262, в этой точке функция распределения первой выборки равна 1, а функция распределения второй выборки равна $\frac{5}{10}$, таким образом статистика Колмогорова равна $\frac{1}{2}$

поскольку $D_{10,10} = 0.5 > 0.41$ гипотеза отклоняется

Критические точки для статистики Колмогорова D_n

Объем выборки n	Уровень значимости α			
	0,10	0,05	0,02	0,01
1	0,95	0,98	0,99	0,995
2	0,78	0,84	0,90	0,93
3	0,64	0,71	0,78	0,83
4	0,57	0,62	0,69	0,73
5	0,51	0,56	0,62	0,67
6	0,47	0,52	0,58	0,62
7	0,44	0,48	0,54	0,58
8	0,41	0,45	0,51	0,54
9	0,39	0,43	0,48	0,51
10	0,37	0,41	0,46	0,49
11	0,35	0,39	0,44	0,47
12	0,34	0,38	0,42	0,45
13	0,33	0,36	0,40	0,43
14	0,31	0,35	0,39	0,42
15	0,30	0,34	0,38	0,40
16	0,29	0,33	0,37	0,39
17	0,29	0,32	0,36	0,38
18	0,28	0,31	0,34	0,37
19	0,27	0,30	0,34	0,36
20	0,26	0,29	0,33	0,35

Критические точки распределения Колмогорова

$$Q(\lambda) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}$$

α	0,10	0,05	0,02	0,01
$\lambda_{кр}$	1,23	1,36	1,52	1,63