Теория Вероятностей и Статистика

Гипотезы о распределении и независимости

Потанин Богдан Станиславович

доцент, кандидат экономических наук

2024-2025

Тест Колмогорова

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.

Тест Колмогорова

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : D_X = D_0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sqrt{n}d_n = \sqrt{n}\left(\sup_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right) \approx \sqrt{n}\left(\max_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right), \quad T(X)|H_0 \xrightarrow{d} \mathcal{K}(n)$$

Где выборочная функция распределения $\hat{F}_n(t)$ считается по выборке X.

Тест Колмогорова

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : D_X = D_0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sqrt{n}d_n = \sqrt{n}\left(\sup_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right) \approx \sqrt{n}\left(\max_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right), \quad T(X)|H_0 \xrightarrow{d} \mathcal{K}(n)$$

  Где выборочная функция распределения $\hat{F}_n(t)$ считается по выборке X.
- Критическая область является правосторонней $\mathcal{T}_\alpha = \left(\mathcal{K}_n^{1-\alpha}, \infty\right)$, где $\mathcal{K}_n^{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения Колмогорова. В результате p-value $= 1 - F_{\mathcal{K}(n)}(T(x))$.

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : D_X = D_0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sqrt{n}d_n = \sqrt{n}\left(\sup_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right) \approx \sqrt{n}\left(\max_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right), \quad T(X)|H_0 \xrightarrow{d} \mathcal{K}(n)$$

Где выборочная функция распределения $\hat{F}_n(t)$ считается по выборке X.

- Критическая область является правосторонней $\mathcal{T}_\alpha = \left(\mathcal{K}_n^{1-\alpha}, \infty\right)$, где $\mathcal{K}_n^{1-\alpha}$ – квантиль уровня $1 - \alpha$ распределения Колмогорова. В результате p-value $= 1 - F_{\mathcal{K}(n)}(T(x))$.
- Поиск супремума фактически предполагает нахождение наибольшей разности между предполагаемой (в соответствии с нулевой гипотезой) и выборочной функцией распределения, что нетрудно сделать с помощью двух вспомогательных статистик:

$$d_n = \max(d_n^+, d_n^-)$$

Тест Колмогорова

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : D_X = D_0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sqrt{n}d_n = \sqrt{n}\left(\sup_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right) \approx \sqrt{n}\left(\max_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right), \quad T(X)|H_0 \xrightarrow{d} \mathcal{K}(n)$$

  Где выборочная функция распределения $\hat{F}_n(t)$ считается по выборке X.
- Критическая область является правосторонней $\mathcal{T}_\alpha = \left(\mathcal{K}_n^{1-\alpha}, \infty\right)$, где $\mathcal{K}_n^{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения Колмогорова. В результате p-value $= 1 - F_{\mathcal{K}(n)}(T(x))$.
- Поиск супремума фактически предполагает нахождение наибольшей разности между предполагаемой (в соответствии с нулевой гипотезой) и выборочной функцией распределения, что нетрудно сделать с помощью двух вспомогательных статистик:

$$d_n = \max(d_n^+, d_n^-)$$

$$d_n^+ = \max\left(|\frac{1}{n} - F_{D_0}(X_{(1)})|, |\frac{2}{n} - F_{D_0}(X_{(2)})|, \cdots, |\frac{n}{n} - F_{D_0}(X_{(n)})|\right)$$

Тест Колмогорова

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из распределения D_X с непрерывной функцией распределения $F_{D_X}(t)$. Также, рассмотрим вариационный ряд $X_{(1)}, ..., X_{(n)}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : D_X = D_0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sqrt{n}d_n = \sqrt{n}\left(\sup_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right) \approx \sqrt{n}\left(\max_{t\in R}|\hat{F}_n(t) - F_{D_0}(t)|\right), \quad T(X)|H_0 \xrightarrow{d} \mathcal{K}(n)$$

Где выборочная функция распределения $\hat{F}_n(t)$ считается по выборке X.

- Критическая область является правосторонней $\mathcal{T}_\alpha = \left(\mathcal{K}_n^{1-\alpha}, \infty\right)$, где $\mathcal{K}_n^{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения Колмогорова. В результате p-value $= 1 - F_{\mathcal{K}(n)}(T(x))$.
- Поиск супремума фактически предполагает нахождение наибольшей разности между предполагаемой (в соответствии с нулевой гипотезой) и выборочной функцией распределения, что нетрудно сделать с помощью двух вспомогательных статистик:

$$d_n = \max(d_n^+, d_n^-)$$

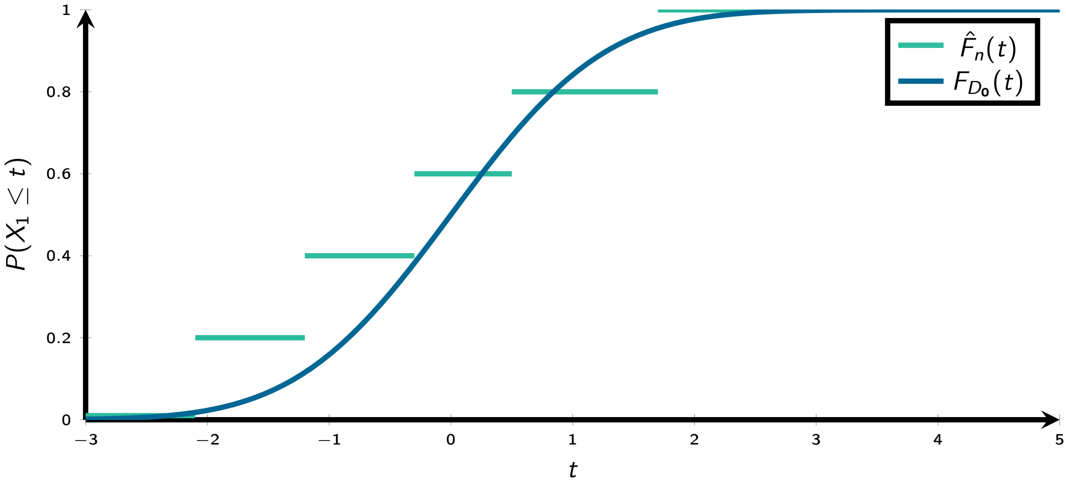
$$d_n^+ = \max\left(|\frac{1}{n} - F_{D_0}(X_{(1)})|, |\frac{2}{n} - F_{D_0}(X_{(2)})|, \cdots, |\frac{n}{n} - F_{D_0}(X_{(n)})|\right)$$

$$d_n^- = \max\left(|F_{D_0}(X_{(1)}) - \frac{1-1}{n}|, |F_{D_0}(X_{(2)}) - \frac{2-1}{n}|, \cdots, |F_{D_0}(X_{(n)}) - \frac{n-1}{n}|\right)$$

Тест Колмогорова

Графическая интерпретация

- Тестовая статистика пропорциональна наибольшему расстоянию между графиками. Смысл расчета d_n^+ и d_n^- заключается в том, что наибольшее расстояние имеет смысл искать лишь около точек разрыва выборочной функции распределения.



Тест Колмогорова

Пример

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

Тест Колмогорова

Пример

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

- Для удобства запишем реализацию вариационного ряда $(-1, 0, 2)$ и рассчитаем тестовую статистику:

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

- Для удобства запишем реализацию вариационного ряда $(-1, 0, 2)$ и рассчитаем тестовую статистику:

$$d_n^+(x) = \max(|1/3 - \Phi(-1)|, |2/3 - \Phi(0)|, |3/3 - \Phi(2)|) \approx \max(0.175, 0.167, 0.023) = 0.175$$

Тест Колмогорова

Пример

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

- Для удобства запишем реализацию вариационного ряда $(-1, 0, 2)$ и рассчитаем тестовую статистику:

$$d_n^+(x) = \max(|1/3 - \Phi(-1)|, |2/3 - \Phi(0)|, |3/3 - \Phi(2)|) \approx \max(0.175, 0.167, 0.023) = 0.175$$
$$d_n^-(x) = \max(|\Phi(-1) - 0/3|, |\Phi(0) - 1/3|, |\Phi(2) - 2/3|) \approx \max(0.159, 0.167, 0.311) = 0.311$$

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

- Для удобства запишем реализацию вариационного ряда $(-1, 0, 2)$ и рассчитаем тестовую статистику:

$$d_n^+(x) = \max(|1/3 - \Phi(-1)|, |2/3 - \Phi(0)|, |3/3 - \Phi(2)|) \approx \max(0.175, 0.167, 0.023) = 0.175$$
$$d_n^-(x) = \max(|\Phi(-1) - 0/3|, |\Phi(0) - 1/3|, |\Phi(2) - 2/3|) \approx \max(0.159, 0.167, 0.311) = 0.311$$
$$T(x) = \sqrt{3}d_n(x) \approx \sqrt{3}\max(0.175, 0.311) = 0.311\sqrt{3} \approx 1.457$$

Имеется реализация выборки $x = (2, -1, 0)$. На 10%-м уровне значимости протестируем гипотезу $H_0 : X_1 \sim \mathcal{N}(0, 1)$. Отметим. что объем данной выборки слишком мал для того, чтобы асимптотическое распределение статистики было достаточно близко к истинному, однако, малая выборка рассматривается из соображений снижения вычислительной нагрузки.

- Для удобства запишем реализацию вариационного ряда $(-1, 0, 2)$ и рассчитаем тестовую статистику:

$$d_n^+(x) = \max(|1/3 - \Phi(-1)|, |2/3 - \Phi(0)|, |3/3 - \Phi(2)|) \approx \max(0.175, 0.167, 0.023) = 0.175$$

$$d_n^-(x) = \max(|\Phi(-1) - 0/3|, |\Phi(0) - 1/3|, |\Phi(2) - 2/3|) \approx \max(0.159, 0.167, 0.311) = 0.311$$

$$T(x) = \sqrt{3}d_n(x) \approx \sqrt{3}\max(0.175, 0.311) = 0.311\sqrt{3} \approx 1.457$$

- Поскольку p-value $= 1 - F_{\mathcal{K}(3)}(1.457) \approx 0.857 > 0.1$, то нулевая гипотеза не отвергается на 10%-м уровне значимости.

- Иногда тест Колмогорова также именуют тестом Колмогорова-Смирнова.

- Иногда тест Колмогорова также именуют тестом Колмогорова-Смирнова.
- Найти таблицу распределения для распределения Колмогорова крайне сложно. Как правило, вместо таблицы с квантилями распределения Колмогорова, в учебниках и в интернете приводят таблицу критических значений для упрощенного вида тестовой статистики $T(X) = d_n$, то есть без $\sqrt{n}$. При использовании подобных таблиц следует либо использовать статистику $T(X) = d_n$, либо умножать критические значения на $\sqrt{n}$.

Несколько технических примечаний

- Иногда тест Колмогорова также именуют тестом Колмогорова-Смирнова.
- Найти таблицу распределения для распределения Колмогорова крайне сложно. Как правило, вместо таблицы с квантилями распределения Колмогорова, в учебниках и в интернете приводят таблицу критических значений для упрощенного вида тестовой статистики $T(X) = d_n$, то есть без $\sqrt{n}$. При использовании подобных таблиц следует либо использовать статистику $T(X) = d_n$, либо умножать критические значения на $\sqrt{n}$.
- Разбираться с данными таблицами не нужно. На контрольной работе будут непосредственно указаны квантили распределения Колмогорова (необходимо подобрать нужную в зависимости от уровня значимости, p-value считать для этого теста вручную не нужно, так как слишком сложно).

Тест Хи-квадрат Пирсона о распределении

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из дискретного распределения (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения принимают значения $v_1, ..., v_m$ с вероятностями $p_1, ..., p_m$ соответственно, где $p_1 + ... + p_m = 1$.

Тест Хи-квадрат Пирсона о распределении

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из дискретного распределения (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения принимают значения $v_1, ..., v_m$ с вероятностями $p_1, ..., p_m$ соответственно, где $p_1 + ... + p_m = 1$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : p_1 = p_1^0, ..., p_m = p_m^0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m} \frac{\left(V_i - np_i^0\right)^2}{np_i^0}, \qquad V_i = \sum_{j=1}^{n} I(X_j = v_i), \qquad T(X)|H_0 \xrightarrow{d} \chi^2(m-1)$$

Тест Хи-квадрат Пирсона о распределении

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из дискретного распределения (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения принимают значения $v_1, ..., v_m$ с вероятностями $p_1, ..., p_m$ соответственно, где $p_1 + ... + p_m = 1$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : p_1 = p_1^0, ..., p_m = p_m^0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m} \frac{\left(V_i - np_i^0\right)^2}{np_i^0}, \qquad V_i = \sum_{j=1}^{n} I(X_j = v_i), \qquad T(X)|H_0 \xrightarrow{d} \chi^2(m-1)$$

Где V_i – количество наблюдений в выборке, принявших значение v_i (частота появления v_i).

Тест Хи-квадрат Пирсона о распределении

Формулировка

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из дискретного распределения (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения принимают значения $v_1, ..., v_m$ с вероятностями $p_1, ..., p_m$ соответственно, где $p_1 + ... + p_m = 1$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : p_1 = p_1^0, ..., p_m = p_m^0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m} \frac{\left(V_i - np_i^0\right)^2}{np_i^0}, \qquad V_i = \sum_{j=1}^{n} I(X_j = v_i), \qquad T(X)|H_0 \xrightarrow{d} \chi^2(m-1)$$

  Где V_i – количество наблюдений в выборке, принявших значение v_i (частота появления v_i).
- Тестовая статистика измеряет, насколько математическое ожидание (при верной нулевой гипотезе) частоты появления значения v_i, то есть $E(V_i|H_0) = np_i^0$, отклоняется от фактически наблюдаемой частоты, то есть V_i. Чем больше соответствующие отклонения, тем менее правдоподобной кажется нулевая гипотеза, что мотивирует рассмотрение правосторонней критической области $\mathcal{T}_\alpha = \left(\chi^2_{m-1,1-\alpha}, \infty\right)$. Где $\chi^2_{m-1,1-\alpha}$ – квантиль уровня $1-\alpha$ распределения $\chi^2(m-1)$.

- Рассмотрим выборку $X = (X_1, ..., X_n)$ из дискретного распределения (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения принимают значения $v_1, ..., v_m$ с вероятностями $p_1, ..., p_m$ соответственно, где $p_1 + ... + p_m = 1$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : p_1 = p_1^0, ..., p_m = p_m^0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m} \frac{\left(V_i - np_i^0\right)^2}{np_i^0}, \qquad V_i = \sum_{j=1}^{n} I(X_j = v_i), \qquad T(X)|H_0 \xrightarrow{d} \chi^2(m-1)$$

  Где V_i – количество наблюдений в выборке, принявших значение v_i (частота появления v_i).
- Тестовая статистика измеряет, насколько математическое ожидание (при верной нулевой гипотезе) частоты появления значения v_i, то есть $E(V_i|H_0) = np_i^0$, отклоняется от фактически наблюдаемой частоты, то есть V_i. Чем больше соответствующие отклонения, тем менее правдоподобной кажется нулевая гипотеза, что мотивирует рассмотрение правосторонней критической области $\mathcal{T}_\alpha = \left(\chi^2_{m-1,1-\alpha}, \infty\right)$. Где $\chi^2_{m-1,1-\alpha}$ – квантиль уровня $1-\alpha$ распределения $\chi^2(m-1)$.
- Очевидно, что p-value $= 1 - F_{\chi^2(m-1)}(T(x))$.

Тест Хи-квадрат Пирсона о распределении

Пример

Лесничий утверждает, что бобры в лесу встречают вдвое реже, чем зайцы и вдвое чаще, чем белки. На летних каникулах Лаврентий часто посещал лес и встретил 30 бобров, 50 зайцев и 20 белок. Поможем Лаврентию, на уровне значимости 10%, протестировать гипотезу о том, что Лесничий говорит правду.

Тест Хи-квадрат Пирсона о распределении

Пример

Лесничий утверждает, что бобры в лесу встречают вдвое реже, чем зайцы и вдвое чаще, чем белки. На летних каникулах Лаврентий часто посещал лес и встретил 30 бобров, 50 зайцев и 20 белок. Поможем Лаврентию, на уровне значимости 10%, протестировать гипотезу о том, что Лесничий говорит правду.

- У Лаврентия есть выборка из $n = 100$ наблюдений. Без потери общности будем обозначать бобров как 1, зайцев как 2, а белок как 3, откуда $x = (\underbrace{1, ..., 1}_{\text{30 раз}}, \underbrace{2, ..., 2}_{\text{50 раз}}, \underbrace{3, ..., 3}_{\text{20 раз}})$, следовательно $V_1(x) = 30$, $V_2(x) = 50$ и $V_3(x) = 20$.

Лесничий утверждает, что бобры в лесу встречают вдвое реже, чем зайцы и вдвое чаще, чем белки. На летних каникулах Лаврентий часто посещал лес и встретил 30 бобров, 50 зайцев и 20 белок. Поможем Лаврентию, на уровне значимости 10%, протестировать гипотезу о том, что Лесничий говорит правду.

- У Лаврентия есть выборка из $n = 100$ наблюдений. Без потери общности будем обозначать бобров как 1, зайцев как 2, а белок как 3, откуда $x = (\underbrace{1, ..., 1}_{30 \text{ раз}}, \underbrace{2, ..., 2}_{50 \text{ раз}}, \underbrace{3, ..., 3}_{20 \text{ раз}})$, следовательно $V_1(x) = 30$, $V_2(x) = 50$ и $V_3(x) = 20$.
- Поскольку $p_1^0 = 0.5p_2^0$ и $p_1^0 = 2p_3^0 = 2(1 - p_1^0 - p_2^0)$, то $p_1^0 = 2/7$, $p_2^0 = 4/7$ и $p_3^0 = 1/7$. Следовательно тестируется гипотеза $H_0 : p_1 = 2/7, p_2 = 4/7, p_3 = 1/7$.

Лесничий утверждает, что бобры в лесу встречают вдвое реже, чем зайцы и вдвое чаще, чем белки. На летних каникулах Лаврентий часто посещал лес и встретил 30 бобров, 50 зайцев и 20 белок. Поможем Лаврентию, на уровне значимости 10%, протестировать гипотезу о том, что Лесничий говорит правду.

- У Лаврентия есть выборка из $n = 100$ наблюдений. Без потери общности будем обозначать бобров как 1, зайцев как 2, а белок как 3, откуда $x = (\underbrace{1, ..., 1}_{30\text{ раз}}, \underbrace{2, ..., 2}_{50\text{ раз}}, \underbrace{3, ..., 3}_{20\text{ раз}})$, следовательно $V_1(x) = 30$, $V_2(x) = 50$ и $V_3(x) = 20$.
- Поскольку $p_1^0 = 0.5p_2^0$ и $p_1^0 = 2p_3^0 = 2(1 - p_1^0 - p_2^0)$, то $p_1^0 = 2/7$, $p_2^0 = 4/7$ и $p_3^0 = 1/7$. Следовательно тестируется гипотеза $H_0 : p_1 = 2/7, p_2 = 4/7, p_3 = 1/7$.
- Рассчитаем реализацию тестовой статистики:

$$T(x) = \frac{(30 - 100 \times 2/7)^2}{100 \times 2/7} + \frac{(50 - 100 \times 4/7)^2}{100 \times 4/7} + \frac{(20 - 100 \times 1/7)^2}{100 \times 1/7} = 3.25$$

Лесничий утверждает, что бобры в лесу встречают вдвое реже, чем зайцы и вдвое чаще, чем белки. На летних каникулах Лаврентий часто посещал лес и встретил 30 бобров, 50 зайцев и 20 белок. Поможем Лаврентию, на уровне значимости 10%, протестировать гипотезу о том, что Лесничий говорит правду.

- У Лаврентия есть выборка из $n = 100$ наблюдений. Без потери общности будем обозначать бобров как 1, зайцев как 2, а белок как 3, откуда $x = (\underbrace{1, ..., 1}_{\text{30 раз}}, \underbrace{2, ..., 2}_{\text{50 раз}}, \underbrace{3, ..., 3}_{\text{20 раз}})$, следовательно $V_1(x) = 30$, $V_2(x) = 50$ и $V_3(x) = 20$.
- Поскольку $p_1^0 = 0.5p_2^0$ и $p_1^0 = 2p_3^0 = 2(1 - p_1^0 - p_2^0)$, то $p_1^0 = 2/7$, $p_2^0 = 4/7$ и $p_3^0 = 1/7$. Следовательно тестируется гипотеза $H_0 : p_1 = 2/7, p_2 = 4/7, p_3 = 1/7$.
- Рассчитаем реализацию тестовой статистики:

$$T(x) = \frac{(30 - 100 \times 2/7)^2}{100 \times 2/7} + \frac{(50 - 100 \times 4/7)^2}{100 \times 4/7} + \frac{(20 - 100 \times 1/7)^2}{100 \times 1/7} = 3.25$$

- Поскольку p-value $= 1 - F_{\chi^2(3-1)}(3.25) \approx 0.197 > 0.1$, то нулевая гипотеза не отвергается на 10%-м уровне значимости.

- Имеются выборки $X=(X_1,...,X_n)$ и $Y=(Y_1,...,Y_n)$, такие, что $Corr(X_i,Y_j)=\rho$ при $i=j$ и X_i,Y_j независимы при любых $i\neq j$.

Выборочная корреляция

Формулировка

- Имеются выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$, такие, что $Corr(X_i, Y_j) = \rho$ при $i = j$ и X_i, Y_j независимы при любых $i \neq j$.
- Состоятельная оценка корреляции $\hat{\rho}_n \xrightarrow{p} \rho$, именуемая выборочной корреляцией, имеет вид:

$$\hat{\rho}_n = \frac{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2}} = \frac{\left(\frac{1}{n}\sum\limits_{i=1}^{n} X_i Y_i\right) - \overline{X}_n \overline{Y}_n}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \frac{1}{n}\sum\limits_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2}}$$

Выборочная корреляция

Формулировка

- Имеются выборки $X=(X_1,...,X_n)$ и $Y=(Y_1,...,Y_n)$, такие, что $Corr(X_i,Y_j)=\rho$ при $i=j$ и X_i,Y_j независимы при любых $i\neq j$.
- Состоятельная оценка корреляции $\hat{\rho}_n\xrightarrow{p}\rho$, именуемая выборочной корреляцией, имеет вид:

$$\hat{\rho}_n=\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}_n\right)\left(Y_i-\overline{Y}_n\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\sum\limits_{i=1}^{n}\left(Y_i-\overline{Y}_n\right)^2}}=\frac{\left(\frac{1}{n}\sum\limits_{i=1}^{n}X_iY_i\right)-\overline{X}_n\overline{Y}_n}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\frac{1}{n}\sum\limits_{i=1}^{n}\left(Y_i-\overline{Y}_n\right)^2}}$$

- Для доказательства достаточно несколько раз применить теорему Слуцкого, предварительно показав, что:

$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i\xrightarrow{p}E(X_1Y_1),\qquad \overline{X}_n\overline{Y}_n\xrightarrow{p}E(X_1)E(Y_1)$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2}\xrightarrow{p}\sqrt{Var(X_1)},\qquad \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}_n\right)^2}\xrightarrow{p}\sqrt{Var(Y_1)}$$

Пример

Лаврентий оценивает корреляцию между ценой и продажами различных товаров по данным о $n = 100$ фирмах. Средняя цена и средняя выручка оказались равны 50 и 9925, а общий объем продаж всех фирм составил 20000. Наконец, (не исправленные) выборочные дисперсии для цены и объема продаж оказались равны 100 и 225 соответственно.

Лаврентий оценивает корреляцию между ценой и продажами различных товаров по данным о $n = 100$ фирмах. Средняя цена и средняя выручка оказались равны 50 и 9925, а общий объем продаж всех фирм составил 20000. Наконец, (не исправленные) выборочные дисперсии для цены и объема продаж оказались равны 100 и 225 соответственно.

- Через $X = (X_1, ..., X_{100})$ и $Y = (Y_1, ..., Y_{100})$ обозначим выборки из цен и объемов продаж.

Лаврентий оценивает корреляцию между ценой и продажами различных товаров по данным о $n = 100$ фирмах. Средняя цена и средняя выручка оказались равны 50 и 9925, а общий объем продаж всех фирм составил 20000. Наконец, (не исправленные) выборочные дисперсии для цены и объема продаж оказались равны 100 и 225 соответственно.

- Через $X = (X_1, ..., X_{100})$ и $Y = (Y_1, ..., Y_{100})$ обозначим выборки из цен и объемов продаж.
- Из условия известно, что:

$$\overline{x}_{100} = 50, \qquad \sum_{i=1}^{100} y_i = 20000 \implies \overline{y}_{100} = 20000/100 = 200, \qquad \frac{1}{100}\sum_{i=1}^{100} x_i y_i = 9925$$

Выборочная корреляция

Пример

Лаврентий оценивает корреляцию между ценой и продажами различных товаров по данным о $n = 100$ фирмах. Средняя цена и средняя выручка оказались равны 50 и 9925, а общий объем продаж всех фирм составил 20000. Наконец, (не исправленные) выборочные дисперсии для цены и объема продаж оказались равны 100 и 225 соответственно.

- Через $X = (X_1, ..., X_{100})$ и $Y = (Y_1, ..., Y_{100})$ обозначим выборки из цен и объемов продаж.
- Из условия известно, что:

$$\overline{x}_{100} = 50, \qquad \sum_{i=1}^{100} y_i = 20000 \implies \overline{y}_{100} = 20000/100 = 200, \qquad \frac{1}{100}\sum_{i=1}^{100} x_i y_i = 9925$$

$$\frac{1}{100}\sum_{i=1}^{100}(x_i - \overline{x}_{100})^2 = 100, \qquad \frac{1}{100}\sum_{i=1}^{100}(y_i - \overline{y}_{100})^2 = 225$$

Лаврентий оценивает корреляцию между ценой и продажами различных товаров по данным о $n = 100$ фирмах. Средняя цена и средняя выручка оказались равны 50 и 9925, а общий объем продаж всех фирм составил 20000. Наконец, (не исправленные) выборочные дисперсии для цены и объема продаж оказались равны 100 и 225 соответственно.

- Через $X = (X_1, ..., X_{100})$ и $Y = (Y_1, ..., Y_{100})$ обозначим выборки из цен и объемов продаж.
- Из условия известно, что:

$$\overline{x}_{100} = 50, \qquad \sum_{i=1}^{100} y_i = 20000 \implies \overline{y}_{100} = 20000/100 = 200, \qquad \frac{1}{100}\sum_{i=1}^{100} x_i y_i = 9925$$

$$\frac{1}{100}\sum_{i=1}^{100}\left(x_i - \overline{x}_{100}\right)^2 = 100, \qquad \frac{1}{100}\sum_{i=1}^{100}\left(y_i - \overline{y}_{100}\right)^2 = 225$$

- Пользуясь найденными реализациями получаем:

$$\hat{\rho}_n(x) = \frac{9925 - 50 \times 200}{\sqrt{100 \times 225}} = -0.5$$

Попробуйте набрать как можно больше очков в игре по ссылке, угадывая значение реализации выборочного коэффициент корреляции по графику:

guessthecorrelation.com

- Имеются выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$ из нормального распределения, такие, что $Corr(X_i, Y_j) = \rho$ при $i = j$ и X_i, Y_j независимы при любых $i \neq j$.

- Имеются выборки $X=(X_1,...,X_n)$ и $Y=(Y_1,...,Y_n)$ из нормального распределения, такие, что $Corr(X_i,Y_j)=\rho$ при $i=j$ и X_i,Y_j независимы при любых $i\neq j$.
- На уровне значимости α гипотезу $H_0:\rho=0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X)=\hat{\rho}_n\sqrt{\frac{n-2}{1-\hat{\rho}_n^2}},\qquad T(X)|H_0\sim t(n-2)$$

Тестирование гипотезы о корреляции
Формулировка

- Имеются выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$ из нормального распределения, такие, что $Corr(X_i, Y_j) = \rho$ при $i = j$ и X_i, Y_j независимы при любых $i \neq j$.
- На уровне значимости α гипотезу $H_0 : \rho = 0$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \hat{\rho}_n \sqrt{\frac{n-2}{1-\hat{\rho}_n^2}}, \qquad T(X)|H_0 \sim t(n-2)$$

- Рассмотрим несколько типов альтернативных гипотез, через $t_{n-2,q}$ обозначая квантиль уровня q распределения $t(n-2)$:

Тип	Левосторонняя	Двухсторонняя	Правосторонняя
Гипотеза	$H_1 : \rho < 0$	$H_1 : \rho \neq 0$	$H_1 : \rho > 0$
$\mathcal{T}_\alpha$	$(-\infty, -t_{n-2,1-\alpha})$	$\left(-\infty, -t_{n-2,1-\alpha/2}\right) \cup \left(t_{n-2,1-\alpha/2}, \infty\right)$	$(t_{n-2,1-\alpha}, \infty)$
p-value	$F_{t(n-2)}(T(x))$	$2\min\left(F_{t(n-2)}(T(x)), 1 - F_{t(n-2)}(T(x))\right)$	$1 - F_{t(n-2)}(T(x))$

Тестирование гипотезы о корреляции

Пример

Ученый кот тестирует гипотезу о наличии корреляции между бюджетом и кассовыми сборами фильмов, предполагая, что они хорошо описываются нормальным распределением. Реализация выборочного коэффициента корреляции, посчитанного по выборке из 227 фильмов, оказалась равна 0.6. На уровне значимости 1% протестируем гипотезу об отсутствии корреляции между бюджетом и кассовыми сборами против двухсторонней альтернативы.

Ученый кот тестирует гипотезу о наличии корреляции между бюджетом и кассовыми сборами фильмов, предполагая, что они хорошо описываются нормальным распределением. Реализация выборочного коэффициента корреляции, посчитанного по выборке из 227 фильмов, оказалась равна 0.6. На уровне значимости 1% протестируем гипотезу об отсутствии корреляции между бюджетом и кассовыми сборами против двухсторонней альтернативы.

- Рассчитаем реализацию тестовой статистики:

$$T(x) = \sqrt{\frac{227-2}{1-0.6^2}}0.6 = 11.25$$

Тестирование гипотезы о корреляции

Пример

Ученый кот тестирует гипотезу о наличии корреляции между бюджетом и кассовыми сборами фильмов, предполагая, что они хорошо описываются нормальным распределением. Реализация выборочного коэффициента корреляции, посчитанного по выборке из 227 фильмов, оказалась равна 0.6. На уровне значимости 1% протестируем гипотезу об отсутствии корреляции между бюджетом и кассовыми сборами против двухсторонней альтернативы.

- Рассчитаем реализацию тестовой статистики:

$$T(x) = \sqrt{\frac{227-2}{1-0.6^2}}0.6 = 11.25$$

- Вычислим p-value:

$$\text{p-value} = 2\min(F_{t(227-2)}(11.25), 1 - F_{t(227-2)}(11.25)) \approx 0$$

В результате нулевая гипотеза отвергается на любом разумном уровне значимости, в том числе на 1%-м.

Тест Хи-квадрат Пирсона о независимости

Формулировка

- Рассмотрим выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$ из дискретных распределений (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения выборки $Z \in \{X, Y\}$ принимают значения $v_{Z,1}, ..., v_{Z,m_Z}$ с вероятностями $p_{Z,1}, ..., p_{Z,m_Z}$ соответственно, где $p_{Z,1} + ... + p_{Z,m_Z} = 1$. Предположим, что X_i и Y_j независимы при $i \neq j$, причем случайные векторы (X_i, Y_i) одинаково распределены при любом $i \in \{1, ..., n\}$.

Тест Хи-квадрат Пирсона о независимости

Формулировка

- Рассмотрим выборки $X=(X_1,...,X_n)$ и $Y=(Y_1,...,Y_n)$ из дискретных распределений (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения выборки $Z\in\{X,Y\}$ принимают значения $v_{Z,1},...,v_{Z,m_Z}$ с вероятностями $p_{Z,1},...,p_{Z,m_Z}$ соответственно, где $p_{Z,1}+...+p_{Z,m_Z}=1$. Предположим, что X_i и Y_j независимы при $i\neq j$, причем случайные векторы (X_i,Y_i) одинаково распределены при любом $i\in\{1,...,n\}$.
- При $n\geq 50$ на уровне значимости α гипотезу $H_0:(X_1$ и Y_1 независимы) можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X)=\sum_{i=1}^{m_X}\sum_{j=1}^{m_Y}\frac{\left(V_{i,j}-n\hat{p}_{X,i}\hat{p}_{Y,j}\right)^2}{n\hat{p}_{X,i}\hat{p}_{Y,j}},\qquad T(X)|H_0\xrightarrow{d}\chi^2((m_X-1)(m_Y-1))$$

$$\hat{p}_{X,i}=\frac{1}{n}\sum_{j=1}^{n}I(X_j=v_{X,i}),\quad \hat{p}_{Y,i}=\frac{1}{n}\sum_{j=1}^{n}I(Y_j=v_{Y,i}),\quad V_{i,j}=\sum_{k=1}^{n}I(X_k=v_{X,i})I(Y_k=v_{Y,j})$$

Тест Хи-квадрат Пирсона о независимости

Формулировка

- Рассмотрим выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$ из дискретных распределений (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения выборки $Z \in \{X, Y\}$ принимают значения $v_{Z,1}, ..., v_{Z,m_Z}$ с вероятностями $p_{Z,1}, ..., p_{Z,m_Z}$ соответственно, где $p_{Z,1} + ... + p_{Z,m_Z} = 1$. Предположим, что X_i и Y_j независимы при $i \neq j$, причем случайные векторы (X_i, Y_i) одинаково распределены при любом $i \in \{1, ..., n\}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : (X_1$ и Y_1 независимы) можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} \frac{\left(V_{i,j} - n\hat{p}_{X,i}\hat{p}_{Y,j}\right)^2}{n\hat{p}_{X,i}\hat{p}_{Y,j}}, \qquad T(X)|H_0 \xrightarrow{d} \chi^2((m_X - 1)(m_Y - 1))$$

$$\hat{p}_{X,i} = \frac{1}{n}\sum_{j=1}^{n} I(X_j = v_{X,i}), \quad \hat{p}_{Y,i} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j = v_{Y,i}), \quad V_{i,j} = \sum_{k=1}^{n} I(X_k = v_{X,i})I(Y_k = v_{Y,j})$$

Где $V_{i,j}$ - частота, с которой встречается пара $(v_{X,i}, v_{Z,j})$. $\hat{p}_{X,i}\hat{p}_{Y,j}$ - оценка вероятности получить пару $(v_{X,i}, v_{Z,j})$ при верной нулевой гипотезе о независимости. $n\hat{p}_{X,i}\hat{p}_{Y,j}$ – оценка ожидаемой частоты появления пары $(v_{X,i}, v_{Z,j})$ при верной нулевой гипотезе.

Тест Хи-квадрат Пирсона о независимости

Формулировка

- Рассмотрим выборки $X = (X_1, ..., X_n)$ и $Y = (Y_1, ..., Y_n)$ из дискретных распределений (но можно обобщить и на случай непрерывного) с конечным носителем. То есть наблюдения выборки $Z \in \{X, Y\}$ принимают значения $v_{Z,1}, ..., v_{Z,m_Z}$ с вероятностями $p_{Z,1}, ..., p_{Z,m_Z}$ соответственно, где $p_{Z,1} + ... + p_{Z,m_Z} = 1$. Предположим, что X_i и Y_j независимы при $i \ne j$, причем случайные векторы (X_i, Y_i) одинаково распределены при любом $i \in \{1, ..., n\}$.
- При $n \geq 50$ на уровне значимости α гипотезу $H_0 : (X_1$ и Y_1 независимы$)$ можно протестировать с помощью следующей тестовой статистики с критической областью $\mathcal{T}_\alpha$:

$$T(X) = \sum_{i=1}^{m_X}\sum_{j=1}^{m_Y}\frac{\left(V_{i,j} - n\hat{p}_{X,i}\hat{p}_{Y,j}\right)^2}{n\hat{p}_{X,i}\hat{p}_{Y,j}}, \qquad T(X)|H_0 \xrightarrow{d} \chi^2((m_X - 1)(m_Y - 1))$$

$$\hat{p}_{X,i} = \frac{1}{n}\sum_{j=1}^{n} I(X_j = v_{X,i}), \quad \hat{p}_{Y,i} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j = v_{Y,i}), \quad V_{i,j} = \sum_{k=1}^{n} I(X_k = v_{X,i})I(Y_k = v_{Y,j})$$

Где $V_{i,j}$ - частота, с которой встречается пара $(v_{X,i}, v_{Z,j})$. $\hat{p}_{X,i}\hat{p}_{Y,j}$ - оценка вероятности получить пару $(v_{X,i}, v_{Z,j})$ при верной нулевой гипотезе о независимости. $n\hat{p}_{X,i}\hat{p}_{Y,j}$ – оценка ожидаемой частоты появления пары $(v_{X,i}, v_{Z,j})$ при верной нулевой гипотезе.

- Из асимптотического распределения тестовой статистики при условии верной нулевой гипотезы получаем, что p-value $= 1 - F_{\chi^2((m_X-1)(m_Y-1))}(T(x))$ и $\mathcal{T}_\alpha = \left(\chi^2_{(m_X-1)(m_Y-1),1-\alpha}, \infty\right)$.

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

- Через X и Y обозначим выборки из числа пересдач и форм обучения соответственно. Из таблицы следует, что $m_X = 3$ и $m_Y = 2$. Кроме того, в (i, j)-й ячейке таблицы фактически располагается $V_{i,j}(x, y)$, например, $V_{3,2}(x, y) = 15$. Наконец, найдем реализации оценок вероятностей:

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

- Через X и Y обозначим выборки из числа пересдач и форм обучения соответственно. Из таблицы следует, что $m_X = 3$ и $m_Y = 2$. Кроме того, в (i, j)-й ячейке таблицы фактически располагается $V_{i,j}(x, y)$, например, $V_{3,2}(x, y) = 15$. Наконец, найдем реализации оценок вероятностей:

$$\hat{p}_{X,1}(x) = (200 + 50)/350 = 10/14, \quad \hat{p}_{X,2}(x) = (40 + 35)/350 = 3/14, \quad \hat{p}_{X,3}(x) = (10 + 15)/350 = 1/14$$

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

- Через X и Y обозначим выборки из числа пересдач и форм обучения соответственно. Из таблицы следует, что $m_X = 3$ и $m_Y = 2$. Кроме того, в (i, j)-й ячейке таблицы фактически располагается $V_{i,j}(x, y)$, например, $V_{3,2}(x, y) = 15$. Наконец, найдем реализации оценок вероятностей:

$$\hat{p}_{X,1}(x) = (200 + 50)/350 = 10/14, \quad \hat{p}_{X,2}(x) = (40 + 35)/350 = 3/14, \quad \hat{p}_{X,3}(x) = (10 + 15)/350 = 1/14$$

$$\hat{p}_{Y,1}(y) = (200 + 40 + 10)/350 = 10/14, \quad \hat{p}_{Y,2}(y) = (50 + 35 + 15)/350 = 4/14$$

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

- Через X и Y обозначим выборки из числа пересдач и форм обучения соответственно. Из таблицы следует, что $m_X = 3$ и $m_Y = 2$. Кроме того, в (i, j)-й ячейке таблицы фактически располагается $V_{i,j}(x, y)$, например, $V_{3,2}(x, y) = 15$. Наконец, найдем реализации оценок вероятностей:

$$\hat{p}_{X,1}(x) = (200 + 50)/350 = 10/14, \quad \hat{p}_{X,2}(x) = (40 + 35)/350 = 3/14, \quad \hat{p}_{X,3}(x) = (10 + 15)/350 = 1/14$$

$$\hat{p}_{Y,1}(y) = (200 + 40 + 10)/350 = 10/14, \quad \hat{p}_{Y,2}(y) = (50 + 35 + 15)/350 = 4/14$$

$$T(x) = \frac{(200 - 350 \times (10/14)(10/14))^2}{350 \times (10/14)(10/14)} + ... + \frac{(15 - 350 \times (1/14)(4/14))^2}{350 \times (1/14)(4/14)} \approx 33.133$$

Тест Хи-квадрат Пирсона о независимости

Пример

Изучается зависимость между числом пересдач у студентов и формой обучения: онлайн или оффлайн. Результаты опроса $n = 350$ случайно отобранных студентов были собраны в таблицу:

Форма обучения \ Число пересдач	0	1	2
0 (оффлайн)	200	40	10
1 (онлайн)	50	35	15

Например, из таблицы следует, что всего в онлайне обучалось $50 + 35 + 15 = 100$ студентов и у 35 из них имеется ровно одна пересдача.

- Через X и Y обозначим выборки из числа пересдач и форм обучения соответственно. Из таблицы следует, что $m_X = 3$ и $m_Y = 2$. Кроме того, в (i, j)-й ячейке таблицы фактически располагается $V_{i,j}(x, y)$, например, $V_{3,2}(x, y) = 15$. Наконец, найдем реализации оценок вероятностей:

$$\hat{p}_{X,1}(x) = (200 + 50)/350 = 10/14, \quad \hat{p}_{X,2}(x) = (40 + 35)/350 = 3/14, \quad \hat{p}_{X,3}(x) = (10 + 15)/350 = 1/14$$

$$\hat{p}_{Y,1}(y) = (200 + 40 + 10)/350 = 10/14, \quad \hat{p}_{Y,2}(y) = (50 + 35 + 15)/350 = 4/14$$

$$T(x) = \frac{(200 - 350 \times (10/14)(10/14))^2}{350 \times (10/14)(10/14)} + ... + \frac{(15 - 350 \times (1/14)(4/14))^2}{350 \times (1/14)(4/14)} \approx 33.133$$

- Поскольку p-value $= 1 - F_{\chi^2((3-1)(2-1))}(33.133) = 1 - F_{\chi^2(2)}(33.133) \approx 0$, то нулевая гипотеза отвергается на любом разумном уровне значимости.

Коэффициент ранговой корреляции Спирмена

Формулировка

- Классический коэффициент корреляции, оценивавшийся ранее, отражает меру **линейной** связи между переменными. Однако, связь может быть и более сложной.

Коэффициент ранговой корреляции Спирмена

Формулировка

- Классический коэффициент корреляции, оценивавшийся ранее, отражает меру **линейной** связи между переменными. Однако, связь может быть и более сложной.
- Для измерения **монотонной** связи между случайными величинами можно ориентироваться на линейную связь между значениями их функций распределения:

$$Corr_{\text{Spearman}}(X_i, Y_i) = Corr(F_{X_i}(X_i), F_{Y_i}(Y_i))$$

- Классический коэффициент корреляции, оценивавшийся ранее, отражает меру **линейной** связи между переменными. Однако, связь может быть и более сложной.
- Для измерения **монотонной** связи между случайными величинами можно ориентироваться на линейную связь между значениями их функций распределения:

$$Corr_{\text{Spearman}}(X_i, Y_i) = Corr(F_{X_i}(X_i), F_{Y_i}(Y_i))$$

- Состоятельная оценка данной корреляции именуется **ранговым коэффициентом корреляция Спирмена** и рассчитываем как выборочная корреляция между рангами наблюдений:

$$Corr_{\text{Spearman}}(X_i, Y_i) = \widehat{Corr}(R(X_i), R(Y_i))$$

$$R(X_i) = \sum_{j=1}^{n} I(X_i \leq X_j) \qquad R(Y_i) = \sum_{j=1}^{n} I(Y_i \leq Y_j)$$

Коэффициент ранговой корреляции Спирмена

Формулировка

- Классический коэффициент корреляции, оценивавшийся ранее, отражает меру **линейной** связи между переменными. Однако, связь может быть и более сложной.
- Для измерения **монотонной** связи между случайными величинами можно ориентироваться на линейную связь между значениями их функций распределения:

$$Corr_{\text{Spearman}}(X_i, Y_i) = Corr(F_{X_i}(X_i), F_{Y_i}(Y_i))$$

- Состоятельная оценка данной корреляции именуется **ранговым коэффициентом корреляция Спирмена** и рассчитываем как выборочная корреляция между рангами наблюдений:

$$Corr_{\text{Spearman}}(X_i, Y_i) = \widehat{Corr}(R(X_i), R(Y_i))$$

$$R(X_i) = \sum_{j=1}^{n} I(X_i \leq X_j) \qquad R(Y_i) = \sum_{j=1}^{n} I(Y_i \leq Y_j)$$

- Если все ранги различаются, то формула упрощается до:

$$Corr_{\text{Spearman}}(X_i, Y_i) = 1 - \frac{6\sum_{i=1}^{n}(R(X_i) - R(Y_i))^2}{n(n^2-1)}$$

- Имеются реализации выборок цен акций:

$$x = (51, 32, 43, 14, 25) \qquad y = (80, 90, 70, 100, 60)$$

Коэффициент ранговой корреляции Спирмена

Пример

- Имеются реализации выборок цен акций:

$$x = (51, 32, 43, 14, 25) \qquad y = (80, 90, 70, 100, 60)$$

- Нетрудно посчитать, что выборочный коэффициент корреляции между ценами акций составит $\hat{\rho} \approx -0.358$.

Коэффициент ранговой корреляции Спирмена

Пример

- Имеются реализации выборок цен акций:

$$x = (51, 32, 43, 14, 25) \qquad y = (80, 90, 70, 100, 60)$$

- Нетрудно посчитать, что выборочный коэффициент корреляции между ценами акций составит $\hat{\rho} \approx -0.358$.
- Для расчета рангового коэффициента корреляции спирмена сперва запишем ранги:

$$R(x) = (5, 3, 4, 1, 2) \qquad R(y) = (3, 4, 2, 5, 1)$$

Коэффициент ранговой корреляции Спирмена

Пример

- Имеются реализации выборок цен акций:

$$x = (51, 32, 43, 14, 25) \qquad y = (80, 90, 70, 100, 60)$$

- Нетрудно посчитать, что выборочный коэффициент корреляции между ценами акций составит $\hat{\rho} \approx -0.358$.
- Для расчета рангового коэффициента корреляции спирмена сперва запишем ранги:

$$R(x) = (5, 3, 4, 1, 2) \qquad R(y) = (3, 4, 2, 5, 1)$$

- Далее, можно либо рассчитать выборочный коэффициент корреляции между рангами, либо воспользоваться приведенной ранее упрощенной формулой:

$$Corr_{\text{Spearman}}(X_i, Y_i) = 1 - 6 \times \frac{(5-3)^2 + (3-4)^2 + (4-2)^2 + (1-5)^2 + (2-1)^2}{5 \times (5^2 - 1)} = -0.3$$