Здравствуйте!

Частный коэффициент корреляции

Пусть теперь имеется **не две**, а **три** случайных величины ξ, η и ζ. Пусть соответствующие парные коэффициенты корреляции есть $r_{\xi\eta}$, $r_{\xi\zeta}$ и $r_{\eta\zeta}$.

В такой ситуации может возникнуть следующий эффект: величина $r_{\xi\eta}$ будет сильно отличаться от нуля, но это происходит совсем не потому, что величины ξ и η зависимы. Просто они каждая по отдельности сильно зависят от величины ζ и эта сильная зависимость и вызывает **кажущуюся** зависимость между величинами ξ и η.

Поэтому хотелось бы иметь такую величину, которая характеризовала бы истинную зависимость между величинами ξ и η **при исключенном влиянии** величины ζ, то есть когда величина ζ поддерживалась бы постоянной. Как показывается в теории вероятностей, для нормальных случайных величин такой характеристикой является так называемый **частный коэффициент корреляции** $r_{\xi\eta\,|\,\zeta}$, который имеет вид

$$r_{\xi\eta|\zeta} = \frac{r_{\xi\eta} - r_{\xi\zeta} r_{\eta\zeta}}{\sqrt{(1 - r_{\xi\zeta}^2)(1 - r_{\eta\zeta}^2)}}.$$

Пусть теперь мы имеем выборку (x_i, y_i, z_i), $i = \overline{1,n}$ объёма n. Вычислим по этой выборке величины $\rho_{\xi\eta}$, $\rho_{\xi\zeta}$ и $\rho_{\eta\zeta}$. Тогда оценка величины $r_{\xi\eta\,|\,\zeta}$ имеет вид

$$\rho_{\xi\eta|\zeta} = \frac{\rho_{\xi\eta} - \rho_{\xi\zeta}\rho_{\eta\zeta}}{\sqrt{(1-\rho_{\xi\zeta}^2)(1-\rho_{\eta\zeta}^2)}}.$$

Относительно неё верно всё сказанное выше относительно величины $\rho_{\xi\eta}$.

Ранговые коэффициенты корреляции

Описанные выше коэффициент корреляции и его оценки имеют тот же недостаток, что и критерий Стьюдента – они «привязаны» к нормальному распределению и все эти формулы и свойства выведены в предположении совместной нормальности случайных величин ξ и η. Если гарантии такой совместной нормальности нет, то значимость полученных результатов и сделанных на их основе выводов может быть поставлена под сомнение.

В таких ситуациях более предпочтительным является использование так называемых ранговых коэффициентов корреляции, «свободных» от вида функции распределения рассматриваемых случайных величин. Рассмотрим два из них.

Итак, пусть снова имеется выборка (x_i, y_i), $i = \overline{1,n}$ объёма n. Упорядочим эти пары данных по возрастанию значений x_i. Тогда порядковый номер r_i пары (x_i, y_i) будет рангом этой пары относительно величин x_i.

А теперь снова упорядочим эти же пары (x_i, y_i) но теперь уже по возрастанию величин y_i. Тогда порядковый номер q_i пары (x_i, y_i) будет рангом этой пары относительно величин y_i.

Таким образом, каждой паре данных (x_i, y_i) будет поставлено в соответствие **два** ранга – r_i и q_i; первый ранг будет соответствовать рангу x_i, а второй – рангу y_i.

Теперь введём ранговые коэффициенты корреляции. Наиболее употребительных – два.

Ранговым коэффициентом корреляции **Спирмена** называется величина

$$\rho = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (r_i - q_i)^2.$$

Значения этой величины лежат в интервале $[-1, 1]$. Их и используют для характеристики степени зависимости случайных величин ξ и η.

Чаще всего его используют для проверки гипотезы о независимости величин ξ и η. В теории показано, что в случае независимости величин ξ и η величина ρ является асимптотически нормальной случайной величиной с математическим ожиданием $M\{\rho\} = 0$ и дисперсией

$$D\{\rho\} = \frac{1}{n-1}.$$

Поэтому при выполнении неравенства

$$|\rho| < \frac{g_\alpha}{\sqrt{n-1}}$$

следует принять гипотезу о независимости случайных величин ξ и η, а при выполнении противоположного неравенства – отвергнуть её.

Вторым ранговым коэффициентом корреляции, который мы рассмотрим, является так называемый ранговый коэффициент корреляции **Кендалла**. Он имеет вид

$$\tau = \frac{1}{n(n-1)} \sum_{i \neq j} \operatorname{sign}(r_i - r_j) \operatorname{sign}(q_i - q_j).$$

Значения этой величины также лежат в интервале [–1, 1]. Их и используют для характеристики степени зависимости случайных величин ξ и η.

Как и ранговый коэффициент корреляции Спирмена, ранговый коэффициент корреляции Кендалла чаще всего используют для проверки гипотезы о независимости величин ξ и η. В теории показано, что в случае независимости величин ξ и η величина τ является асимптотически нормальной случайной величиной с математическим ожиданием $M\{\tau\} = 0$ и дисперсией

$$D\{\tau\} = \frac{2(2n+5)}{9n(n-1)}.$$

Поэтому при выполнении неравенства

$$|\tau| < g_\alpha \sqrt{\frac{2(2n+5)}{9n(n-1)}}$$

следует принять гипотезу о независимости случайных величин ξ и η, а при выполнении противоположного неравенства – отвергнуть её.

Связь ранговых и обычного коэффициентов корреляции

Пусть имеется двумерная нормальная случайная величина (ξ, η) коэффициент корреляции которых равен $r_{\xi\eta}$.

Если мы имеем выборку (x_i, y_i), $i = \overline{1,n}$, то по ней можно построить ранговые коэффициенты корреляции Спирмена и Кендалла. Какова их связь с $r_{\xi\eta}$?

Оказывается, что эта связь даётся следующими формулами

$$\mathrm{M}\{\tau\} = \bar{\tau} = \frac{2}{\pi}\arcsin r_{\xi\eta},$$

$$\mathrm{M}\{\rho\} = \bar{\rho} = \frac{6}{\pi}\arcsin\frac{r_{\xi\eta}}{2},$$

или, наоборот,

$$r_{\xi\eta} = \sin\left(\frac{\pi}{2}\bar{\tau}\right),$$

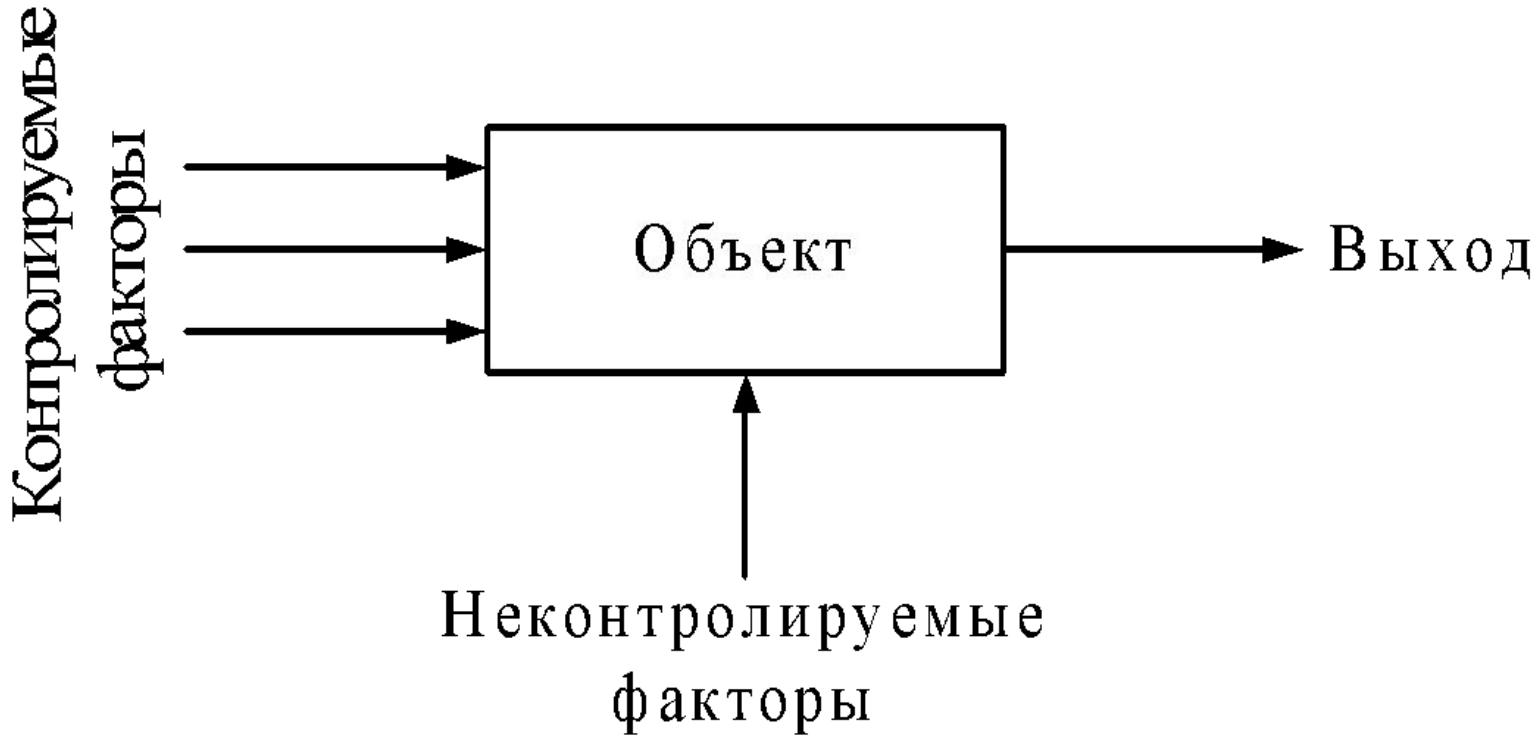
$$r_{\xi\eta} = 2\sin\left(\frac{\pi}{6}\bar{\rho}\right),$$

так что для нормальных случайных величин всегда можно перейти от обычного коэффициента корреляции к ранговому, и наоборот.

Элементы дисперсионного анализа

Постановка задачи

Пусть нам дан некоторый объект (см. рис.) выход которого мы можем измерять.



На этот объект действуют какие-то внешние факторы, которые можно разбить на две группы.

1. Контролируемые факторы, то есть те факторы, которые мы можем задавать по своему усмотрению, или хотя бы измерять.

2. Неконтролируемые факторы, природу которых мы не знаем, и измерить их не можем.

Дисперсионный анализ позволяет ответить на следующие вопросы:

1. Какова степень влияния контролируемых и неконтролируемых факторов на выход объекта?

2. Можно ли считать влияние контролируемых факторов доказанным?

3. В ситуации, когда контролируемых факторов несколько, может быть так, что влияют не каждый фактор в отдельности, а основное влияние оказывает некоторая комбинация факторов. Тогда для каждой комбинации факторов надо ответить на те же вопросы, которые сформулированы выше.

Название дисперсионного анализа зависит от числа контролируемых факторов. Если контролируемый фактор один, то мы имеем дело с однофакторным дисперсионным анализом, если факторов два – с двухфакторным дисперсионным анализом и т.д.

Однофакторный дисперсионный анализ

Оценка степени влияния контролируемого фактора

Пусть у нас имеется один контролируемый фактор, значения которого мы можем фиксировать на каких-то уровнях с номерами 1, 2, 3, …, s.

Пусть мы зафиксировали значение контролируемого фактора на уровне i и при этом уровне производим измерения выхода нашего объекта. В силу того, что на наш объект действуют также неконтролируемые факторы, измеренные значения будут, вообще говоря, различными. Обозначим их как $x_i^{(j)}$, $j = \overline{1, n_i}$, где нижний индекс относится к уровню контролируемого фактора, а верхний индекс указывает номер измерения. Совокупность данных $x_i^{(j)}$, $j = \overline{1, n_i}$ называется **классом** или **группой.**

Введём следующие величины

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^{(j)},$$

которые мы будем называть средними внутри i-й группы и величину

$$m = \frac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} x_i^{(j)} = \frac{1}{n} \sum_{i=1}^{s} n_i m_i,$$

которую мы будем называть общим средним. Здесь $n = \sum_{i=1}^{s} n_i$ есть общее число сделанных измерений.

Введём теперь основную величину

$$D_{\text{общ}} = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_i^{(j)} - m)^2,$$

которую называют общей или полной дисперсией или изменчивостью наших измерений.

Получим теперь основное соотношение однофакторного дисперсионного анализа. Имеем

$$D_{\text{общ}} = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_i^{(j)} - m)^2 = \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_i^{(j)} - m_i + m_i - m)^2 =$$

$$= \sum_{i=1}^{s} \sum_{j=1}^{n_i} (x_i^{(j)} - m_i)^2 - 2\sum_{i=1}^{s} (m_i - m) \sum_{j=1}^{n_i} (x_i^{(j)} - m_i) + \sum_{i=1}^{s} n_i (m_i - m)^2.$$

Но

$$\sum_{j=1}^{n_i}(x_i^{(j)} - m_i) = n_i m_i - n_i m_i = 0,$$

так что окончательно

$$D_{\text{общ}} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_i^{(j)} - m_i)^2 + \sum_{i=1}^{s} n_i (m_i - m)^2.$$

Величина

$$D_{\text{нф}} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_i^{(j)} - m_i)^2$$

называется дисперсией или изменчивостью внутри групп. Так как в *i*-й группе значение контролируемого фактора фиксировано, то отличия $x_i^{(j)}$ от m_i обусловлены действием неконтролируемых факторов. Поэтому считается, что эта величина характеризует влияние неконтролируемых факторов.

Величина

$$D_{\text{кф}} = \sum_{i=1}^{s} n_i (m_i - m)^2$$

называется дисперсией или изменчивостью между группами. Так как группы отличаются между собой значением контролируемого фактора, то считается, что эта величина характеризует влияние контролируемого фактора. Таким образом $D_{\text{общ}} = D_{\text{нф}} + D_{\text{кф}}$.

Теперь можно определить и степень влияния контролируемого фактора. Она определяется величиной

$$\eta_{\text{кф}} = \frac{D_{\text{кф}}}{D_{\text{общ}}} \cdot 100\%,$$

и обычно указывается в процентах. Аналогично, степень влияния неконтролируемых факторов измеряется величиной

$$\eta_{\text{нф}} = \frac{D_{\text{нф}}}{D_{\text{общ}}} \cdot 100\%,$$

которая также указывается в процентах.

Можно ли считать доказанным влияние контролируемого фактора?

Рассмотрим теперь вопрос о том, можно ли считать доказанным влияние контролируемого фактора на наш объект.

Решение этого вопроса связано со следующей математической моделью рассматриваемой ситуации. Считается, что величины $x_i^{(j)}$ можно представить в виде

$$x_i^{(j)} = \mu_i + n_i^{(j)},$$

где $n_i^{(j)}$ – независимые одинаково распределённые случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Величина μ_i описывает влияние контролируемого фактора.

Проверка того, влияет ли контролируемый фактор на изучаемый объект, сводится к проверке следующей статистической гипотезы

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_s.$$

Принятие этой гипотезы означает, что влияние контролируемого фактора не доказано (действительно, ведь в этом случае μ_i не меняются!); если же эта гипотеза будет отвергнута, то влияние контролируемого фактора можно считать доказанным.

Мы не будем приводить вывод следующих соотношений

$$\mathrm{M}\{D_{\text{нф}}\} = (n-s)\sigma^2,$$

$$\mathrm{M}\{D_{\text{кф}}\} = (s-1)\sigma^2 + \sum_{i=1}^{s} n_i(\mu_i - \overline{\mu})^2,$$

где $\overline{\mu} = \dfrac{1}{n}\sum_{i=1}^{s} n_i\mu_i$.

Введём величины

$$s^2_{\text{нф}} = \frac{D_{\text{нф}}}{n-s}, \quad s^2_{\text{кф}} = \frac{D_{\text{кф}}}{s-1}.$$

Тогда проверка нашей гипотезы сведётся к проверке гипотезы вида

$$H_0\colon \mathrm{M}\{s^2_{\text{кф}}\} = \mathrm{M}\{s^2_{\text{нф}}\},$$

при альтернативе

$$H_1\colon \mathrm{M}\{s^2_{\text{кф}}\} > \mathrm{M}\{s^2_{\text{нф}}\}.$$

Для проверки этой гипотезы используют так называемый F-критерий. Согласно нему, следует вычислить величину

$$F = \frac{s^2_{\text{кф}}}{s^2_{\text{нф}}}$$

и сравнить это значение с пороговым значением F_α. Если окажется, что

$$F \le F_\alpha,$$

то следует принять гипотезу H_0, то есть влияние контролируемого фактора нельзя считать доказанным. При выполнении противоположного неравенства $F > F_\alpha$ влияние контролируемого фактора можно считать установленным по уровню значимости α_0.

Пороговое значение F_α находится из таблиц F-критерия по выбранному уровню значимости α_0, числу степеней свободы числителя $f_1 = s - 1$ и числу степеней свободы знаменателя $f_2 = n - s$.