**§3. Корреляционный анализ**

Как уже указывалось, выявление взаимосвязей между признаками биологических объектов является важным средством исследования механизмов их функционирования.

Различают два вида связи: **функциональную** и **статистическую**.

При функциональной связи каждому значению одного признака соответствует вполне определенное значение (или набор значений) другого. Например, скорость можно выразить уравнением:

$$V=a∙t,$$

где $V$ – скорость;

 $а- $ускорение;

 t – время.

Каждому значению времени t соответствует одно значение скорости V.

При статистической связи такого соответствия нет. Например, если нам сообщили рост человека, то по этой информации точно предсказать вес не возможно. Хотя известно, что вес как случайная величина зависит от роста. Эта зависимость проявляется в том , что при увеличении роста увеличивается среднее значение веса.

**Статистической связью** называется такая зависимость ,при которой изменение одной величины влечет изменение закона распределения другой.

Например, изменение одной величины может влиять на среднее значение или дисперсию другой. Может даже изменяться функция распределения случайной величины (а не только ее числовые характеристики).

**Корреляционной зависимостью** называется такая статистическая связь, при которой изменение одной величины влечет за собой изменение среднего значения другой.

Если изменение одной величины приводит к прямо пропорциональному изменению среднего значения другой, то говорят о **линейной** **корреляции** . В этом случае связь может быть охарактеризована количественно коэффициентом корреляции.

Оценка коэффициента корреляции по выборочным данным вычисляется для малых выборок по формуле:

$$\hat{r}=\frac{D\_{x}+D\_{y}-D\_{d}}{2∙\sqrt{D\_{x}∙D\_{y}}}$$

где$\hat{r}$- выборочное значение коэффициента корреляции.

$$D\_{x}=\sum\_{}^{}x\_{i}^{2}-\frac{\left(\sum\_{}^{}x\_{i}\right)^{2}}{n}$$

$$D\_{y}=\sum\_{}^{}y\_{i}^{2}-\frac{\left(\sum\_{}^{}y\_{i}\right)^{2}}{n}$$

$$D\_{d}=\sum\_{}^{}d\_{i}^{2}-\frac{\left(\sum\_{}^{}d\_{i}\right)^{2}}{n}$$

$$d\_{i}=x\_{i}-y\_{i}$$

Здесь $x\_{i}$ и $y\_{i}$ ***-*** выборочные значения двух признаков, измеренные попарно на выборке объемом $n$.

Для удобства вычислений составляют таблицу.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| $$x\_{i}$$ | $$y\_{i}$$ | $$d\_{i}$$ | $$x\_{i}^{2}$$ | $$y\_{i}^{2}$$ | $$d\_{i}^{2}$$ |
| *х1* | *у1* | *d1* | $$x\_{1}^{2}$$ | $$y\_{1}^{2}$$ | $$d\_{1}^{2}$$ |
| *х2* | *у2* | *d2* | $$x\_{2}^{2}$$ | $$e\_{2}^{2}$$ | $$d\_{2}^{2}$$ |
| *.* | *.* | *.* | *.* | *.* | *.* |
| *.* | *.* | *.* | *.* | *.* | *.* |
| *.* | *.* | *.* | *.* | *.* | *.* |
| *хn* | *уn* | *dn* | $$x\_{n}^{2}$$ | $$y\_{n}^{2}$$ | $$d\_{n}^{2}$$ |
| $\sum\_{}^{}x\_{i}$ | $\sum\_{}^{}y\_{i}$ | $\sum\_{}^{}d\_{i}$ | $\sum\_{}^{}x\_{i}^{2}$ | $\sum\_{}^{}y\_{i}^{2}$ | $\sum\_{}^{}d\_{i}^{2}$ |

Вычислив необходимые суммы, вычисляют Dx, Dy, Dd, подстав­ляют в формулу и находят $\hat{r}$.

Теперь нужно проверить достоверность отличия этого коэффи­циента от нуля. Ведь в силу случайности, даже если между *x* и *y*связи нет, выборочный коэффициент корреляции может оказаться не равным нулю.

Для проверки нулевой гипотезы о равенстве нулю коэффициента корреляции используется критерий Стьюдента, фактическое значение которого вычисляется для малой выборки (n<30) по формуле:

$$t\_{ф}=\frac{Z}{m\_{z}}$$

$$Z=\frac{1}{2}ln\frac{\hat{r}-1}{\hat{r}+1}$$

где $m\_{z}$ – среднеквадратическая ошибка величины Z.

Величину Z находят, обычно, по таблице (Лакин, 1979, с. 286), а

$m\_{z}=\frac{1}{\sqrt{n-3}}$.

Конкурирующая гипотеза имеет вид:

$$H\_{1}:r\ne 0$$

Поэтому критическая область двусторонняя. По таблице Стьюдента находят критическую точку при числе степеней свободы $к=n-2$. Если,$ t\_{ф}>t\_{кр}$ то нулевую гипотезу отвергают, то есть делают вывод о том, что экспериментальные данные свидетельствуют о наличии связи между признаками.

Если $t\_{ф}>t\_{кр}$, то нулевая гипотеза принимается, то есть величины некоррелированы.

Если выборочный коэффициент корреляции оказался по абсолютной величине больше единицы, то это признак ошибки в вычислениях.

Для больших выборок (n>100) фактическое значение критерия Стьюдента вычисляется по формуле:

$$t\_{ф=}=\frac{\hat{r}}{\sqrt{\frac{1-\hat{r}}{n-2}}}$$

и дальнейшая проверка нулевой гипотезы проводится, как указано выше.

При нелинейной связи между признаками коэффициент линейной корреляции неадекватно отражает связь и поэтому не используется.

Пример. Измерение массы овцы (X) и массы шерсти (У), настригаемой с нее дало следующие результаты.

X: 58.0, 62.0, 62.8, 52.4, 56.2, 62.2, 2.0 72.0

У: 7.4, 7.3, 7.1, 6.1, 6.6, 7.1, 7.2

Вычислить коэффициент корреляции и его достоверность.

Решение. Составим таблицу:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *х* | *у* | *d* | *x2* | *y2* | *d2* |
| 58.0 | 7.4 | 50.6 | 3364.0 | 64.76 | 2650.36 |
| 62.0 | 7.3 | 54.7 | 3844.0 | 53.39 | 2992.09 |
| 52.4 | 7.1 | 42.3 | 2745.76 | 50.41 | 2052.09 |
| 56.2 | 6.1 | 50.1 | 3158.44 | 37.21 | 2510.01 |
| 62.2 | 7.1 | 55.1 | 3868.84 | 50.41 | 3036.01 |
| 72.0 | 7.2 | 64.8 | 5184.00 | 51.84 | 5199.04 |
| 62.8 | 7.1 | 55.7 | 3943.84 | 50.41 | 3102.49 |
| 425.6 | 49.3 | 376.3 | 26108.88 | 348.33 | 20452.09 |

$$D\_{x}=26108.68-\frac{\left(425.6\right)^{2}}{7}=232.40$$

 $D\_{y}=348.33-\frac{\left(49.3\right)^{2}}{7}=1.12$

$$D\_{d}=20452.09-\frac{\left(376.3\right)^{2}}{7}=223.28$$

$$\hat{r}=\frac{232.40+1.12-223.28}{2∙232.40∙1.12}=\frac{10.24}{32.26}=0.317$$

Полученное значение выборочного коэффициента корреляции указывает на слабую положительную связь между весом овцы и настригаемой с нее шерсти. Но окончательный вывод о наличии такой связи пока делать рано. Нужно проверить достоверность коэффициента корреляции. По таблице находим значение Z. Оно равно 0.332. Вычислим значение критерия Стьюдента:

$$t\_{ф}=\frac{Z}{m\_{z}}=\frac{0.332}{\frac{1}{\sqrt{7-3}}}=0.332∙\sqrt{4}=0.664$$

Полученное значение сравним с критическим, по таблице Стьюдента при β = 0.95 и К = 7-2 = 5. Оно равно 2.57. Мы видим, что $t\_{ф}<t\_{кр}$. Значит коэффициент корреляции статистически недостоверен, то есть принимается нулевая гипотеза, согласно которой r = 0; другими словами, нельзя утверждать, что между весом овцы и настригаемой шерстью есть связь. Возможно, что выборка слишком мала (всего 7 особей) и поэтому необходимо провести более обширное исследование, чтобы проверить нулевую гипотезу с большей точностью.

**Контрольные вопросы и задания**

I. Вычислить коэффициент корреляции и его достоверность:

1. Высота растений *(****X*** см) и количество язычковых цветков *(У)* нивяника обыкновенного:

Х: 48 48 44 45 66 60 54 69 33

У: 18 11 12 14 18 20 21 22 12

1. Интенсивность миграции ***(X)*** и средняя жирность *(У)* зябликов на Куршской косе:

У: 360 280 210 220 190 240 170

***X*** 4875 4103 3038 1307 840 506 351

1. Рост ***(X)*** и вес *(У)* у мужчин:

X: 166 187 170 171 182 188 175 183 179

У: 63 98 82 80 86 85 84 79 75

1. Рост (***X)*** и вес *(У)* у женщин:

X: 161 158 165 154 166 170 167 177

***У:*** 60 61 68 52 67 68 69 80

1. Рост отцов *(X)* и их старших сыновей *(У)* в дюймах:

X: 67 63 70 66 68 71 65 62 69

У: 68 66 68 64 72 69 68 67 71

1. длины окружности грудной клетки (***X***, дюйм) и объем легких (*У*, кубические дюймы) у первокурсников:

X: 38.9 35.0 31.3 30.3 38.0 31.5 30.8 33.7

У: 311 305 330 210 269 238 305 219

1. Возраст женщин (***X***, годы) и систолическое давление (У, мм рт.

ст.):

X: 36 47 49 42 38 68 56 42 63

У: 118 128 145 125 115 152 147 140 149

1. Длина лепестка (X, см) и ширина лепестка (У, см):

X: 4.7 4.5 4.9 4.0 4.5 4.7 3.3 4.6

У: 1.4 1.5 1.5 1.3 1.5 1.3 1.0 1.3

1. Коэффициент интеллектуальности родителей ***(X)*** и их детей (У):

X: 125 120 110 105 95 95 90 90 80 75

У: 110 105 95 125 120 105 75 95 90 80

1. Рост (X, дюймы) и вес (фунты) первокурсников:

X: 65.8 68.3 72.7 66.1 73.1 71.8 73.1 66.5

У: 166.0 115.2 157.8 152.5 149.3 181.0 173.2 120.4