МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
имени Н.Э. БАУМАНА

С.П.Еркович

ПРИМЕНЕНИЕ РЕГРЕССИОННОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ В ФИЗИЧЕСКОМ ПРАКТИКУМЕ.

Москва, 1994.

Методические указания

Введение

Изучение причинно-следственных связей между физическими явлениями – одна из основных задач экспериментальной физики. На практике часто возникают проблемы анализа причинно-следственных связей в условиях, когда случайные влияния побочных факторов искажают ход эксперимента. В этом случае при повторении опыта значения измеряемых величин изменяются некоторым, заранее непредсказуемым образом, т. е. приобретают характер случайных величин. Вскрыть искомые закономерности в этой ситуации помогает статистическая обработка экспериментальных данных.

При исследовании взаимосвязей между физическими явлениями возникают три основные задачи.

1. Определение на основании экспериментальных данных, как изменялась бы исследуемая физическая величина как функция одного или небольшого числа свих аргументов, если бы другие аргументы, отражающие малые влияния побочных причин не изменялись. При этом задача должна решаться на экспериментальном материале, где эти прочие аргументы на самом деле хаотически изменяются и своей изменчивостью искажают интересующую исследователя зависимость. Эта задача решается методами регрессивного анализа.
2. Определение степени искажающего влияния побочных случайных факторов на исследуемую зависимость. Эта задача решается методами корреляционного анализа.
3. Определение минимального объёма выборки, т. е. минимального числа опытов, которое необходимо выполнить, чтобы погрешность эксперимента, вызванная случайно действующими факторами, не превосходила заранее обусловленной величины. Эта задача решается методами планирования эксперимента.

Данное пособие преследует цель ознакомить студентов с простейшими методами решения этих задач. Подробно данные вопросы освещены в книге: Львовский Е. Н. Статистические методы построения эмпирических формул. М.: Высшая школа, 1988.

Генеральная и выборочная совокупности

Пусть для исследования зависимости между величинами X и Y выполнено $\boldsymbol{n}$ опытов, в результате которых получен ряд наблюдений: $\boldsymbol{x_1, y_1; x_2, y_2; \ldots ; x_i, y_i, \ldots ; x_n, y_n}$.Наиболее полные сведения о случайных величинах X и Y можно получить, проведя бесконечное число измерений.

Генеральной совокупностью называют совокупность всех мыслимых наблюдений, которые могли бы быть сделаны при данном реальном комплексе условий измерений. Число членов, образующих генеральную совокупность, называют объемом генеральной совокупности. Как правила, генеральная совокупность имеет бесконечный объем.

Выборочной совокупностью, или просто выборкой объема $\boldsymbol{n}$, называют совокупность $\boldsymbol{n}$ результатов измерений, отобранных из генеральной совокупности. Представленный ряд наблюдений является примером такой выборки объема $\boldsymbol{n}$.

Вероятность, функция, распределения случайной величины, плотность вероятности.

Вероятность попадания случайной величины в заданную область ее значений можно определить следующим образом:

$$p = \lim_{m \to \infty} n_m . \tag{1}$$

Здесь n_m - число наблюдений случайной величины, оказавшихся в заданной области значений; n - общее число наблюдений.

Аналитическим выражением закона распределения случайной величины являются функция распределения вероятностей и плотность вероятности.

Функция распределения $F(x)$ случайной величины X равна вероятности того, что случайная величина не превышает некоторого заданного или текущего значения x:

$$F(x) = p\{X \le x\} . \tag{2}$$

Плотность вероятностей $f(x)$ для непрерывных случайных величин определяют как производную от функции распределения, т.е.

$$f(x) = F'(x) . \tag{3}$$

Она дает возможность найти вероятность попадания случайной величины X в элементарный интервал dx. Эта вероятность равна $f(x)dx$, а в случае конечного интервала (a_1, a_2) равна интегралу по этому интервалу:

$$p\{a_1 < x < a_2\} = \int_{a_1}^{a_2} f(x)dx . \tag{4}$$

Числовые характеристики случайной величины

Вместо функции распределения или плотности вероятности основные свойства случайной величины могут быть описаны с помощью числовых параметров. Наибольшую роль среди них на практике играют два параметра - генеральное среднее значение случайной величины X, именуемое математическим ожиданием (обозначается $M(X)$ либо m_x), и дисперсия (обозначается $D(X)$ либо σ_x^2). По определению

для непрерывной случайной величины

$$m_x = \int_{a_1}^{a_2} xf(x)dx \tag{5}$$

для дискретной случайной величины

$$m_x = \sum_{i=1}^{\infty} p_i x_i = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1} x_i . \tag{6}$$

Легко показать, что математическое ожидание равно среднему арифметическому значению случайной величины, усредненному по всей генеральной совокупности.

Дисперсией σ_x^2 случайной величины X называют математическое ожидание квадрата отклонения значений случайной величины от ее математического ожидания. По определению для непрерывной случайной величины

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - m_x)^2 f(x)dx \tag{7}$$

для дискретной

$$\sigma_x^2 = \sum_{i=1}^{\infty} (x_i - m_x)^2 p_i \tag{8}$$

Можно показать, что дисперсия равна среднему значению от квадрата разности между случайной величиной и ее математическим ожиданием.

Корень квадратный из дисперсии $\sigma_x = \sqrt{\sigma_x^2}$ называют стандартом, или средним квадратическим отклонением.

Системы случайных величин и их числовые характеристики

На практике результат опыта обычно описывается не одной, а двумя и более случайными величинами.

Основные свойства двумерной совокупности случайных величин ***X*** и ***Y*** могут быть охарактеризованы с помощью ряда числовых параметров - математических ожиданий и дисперсий соответствующих случайных величин $m_x, m_y, \sigma_x^2, \sigma_y^2$. Кроме того, для двумерной совокупности случайных величин применяют параметры, характеризующие степень взаимозависимости переменных ***X*** и ***Y***. Ими являются корреляционный момент μ_{xy} и коэффициент корреляции ρ_{xy}.

Корреляционным моментом μ_{xy} системы случайных величин (***X***, ***Y***) называют математическое ожидание произведения отклонений каждой из величин от их математических ожиданий, т.е. μ_{xy} равен математическому ожиданию от произведения $(X-m_x)(Y-m_y)$.

Коэффициентом корреляции величин ***X*** и ***Y*** называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$\rho_{xy} = \mu_{xy} / \sigma_x \sigma_y . \tag{9}$$

Статистическая оценка неизвестных параметров распределения по результатам эксперимента.

Для вычисления параметров распределения необходимо располагать всей генеральной совокупностью величин. В те же время из эксперимента известна лишь выборка конечного объема ***n***.

Точечные оценки параметров распределения - это приближенные оценки их значений по выборке объема ***n***.Оценка математического ожидания (обозначается $\bar{x}$ или $\bar{y}$) носит название выборочного среднего, ее определяют по формулам

$$m_x \to \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \tag{10}$$

$$m_y \to \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i . \tag{11}$$

Выборочную дисперсию, выборочный корреляционной момент и выборочный коэффициент корреляции оценивают по формулам

$$\sigma_x^2 \to S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 , \tag{12}$$

$$\sigma_y^2 \to S_y^2 = \frac{1}{n-1}\sum (\bar{y}_i - \bar{y})^2 , \tag{13}$$

$$\mu_{xy} \to K_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) , \tag{14}$$

$$\rho_{xy} \to r = K_{xy} / S_x S_y . \tag{15}$$

Интервальная оценка математического ожидания

Оценка $\bar{x}$ математического ожидания m_x, полученная по конечной выборке, является приближенной. Мера точности оценки - ширина интервала Δx, который с заданной вероятно-

стью p накрывает искомое значение m_x. Для оценки ширины этого интеграла (погрешности измерения m_x) вводят случайную величину

$$t = (\bar{x} - m_x)/(S/\sqrt{n}). \quad (16)$$

Плотность вероятности $f(t)$ этой случайной величины была получена Стьюдентом, ее называют распределением Стьюдента.

Обозначим $1-\alpha$ вероятность того, что величина t лежит внутри интервала, определенного некоторыми значениями $t=\pm t_\alpha$. Это можно записать в виде

$$p\left\{-t_\alpha < \frac{\bar{x} - m_x}{S/\sqrt{n}} < +t_\alpha\right\} = \int_{-t\alpha}^{t\alpha} f(t)dt = 1-\alpha. \quad (17)$$

Вероятность $p=1-\alpha$, с которой выполняется соотношение (17), называют доверительной вероятностью оценки математического ожидания m_x. С доверительной вероятностью $p=1-\alpha$ математическое ожидание лежит в интервале

$$\bar{x} \pm \Delta x = \bar{x} \pm t_\alpha S/\sqrt{n}. \quad (18)$$

Величину α называют уровнем значимости. Уровень значимости равен вероятности того, что искомая величина окажется за пределами доверительного интервала.

Числовые значения параметра t_α, полученные в результате решения уравнения (I7) при заданном уровне значимости α, называют квантилями распределения Стьюдента. Они определяются не только уровнем значимости α, но и числом независимых результатов измерений, называемым числом степеней, свободы распределения (обозначается f). В данном случае число степеней свободы на единицу меньше числа опытов n. Так как по уравнению

$$\bar{x} = \sum x_i / n$$

один из результатов можно выразить с помощью остальных.

Квантили распределения Стьюдента для различных α и f табулированы. Так как доверительная вероятность $1-\alpha$ должна быть близка к единице, то уровень значимости α измеряется малыми долями единицы. Наиболее часто полагают α=0,1; 0,05; 0,02; 0,01; 0,005; 0,001.

Условный закон распределения. Условное математическое ожидание

Случайную величину *Y* называют независимой от случайной величины *X*, если закон распределения величины *Y* не зависит от того, какое значение приняла величина *X*. Если между *X* и *Y* имеется взаимосвязь, то зависимость между ними можно охарактеризовать с помощью законов распределения. Условным законом распределения величины *Y* называют ее закон распределения при условии, что другой случайной величине *X* приписано определенное фиксированное значение. Условную функцию распределения означают $F(y/x)$, условную плотность распределения $f(y/x)$.

На практике важное значение имеет условное математическое ожидание случайной величины. Для ее выборочной оценки, обозначаемой $\bar{y}_x$, выполняют n опытов по измерению величины *Y* при фиксированном значении $X=x$. Полученный ряд значений $y_1; y_2; \ldots; y_i; \ldots; y_n$ обрабатывают по следующей формуле:

$$\bar{y}_x = \sum_{i=1}^{n} y_i / n \quad (19)$$

Эмпирическое уравнение регрессии

Одной из основных задач экспериментальной физики является исследование функцио-

нальных зависимостей между двумя или несколькими физическими величинами. Каждый экспериментатор знает, что понятие функциональной зависимости является лишь абстракцией, так как неизбежен разброс экспериментальных значений. Поэтому в практических приложениях, когда разброс велик, ограничиваются исследованием зависимости между x и условным математическим ожиданием $\bar{y}_x$. Зависимости такого рода называют регрессионными.

Зависимость $\bar{y}_x$ от x полученная на основании выборочных данных по намерению n пар величин $x_1, y_1; x_2, y_2; \ldots x_i, y_i; \ldots ; x_n, y_n$, называется эмпирическим уравнением регрессии Y на X.

Линейная регрессия

Часто на практике встречаются случаи, когда условное математическое ожидание является линейной функцией, т.е.

$$\bar{y}_x = a_0 + a_1 x . \tag{20}$$

Соотношение (20) называют эмпирическим уравнением линейной регрессии. Коэффициенты a_0 и a_1 (коэффициенты линейной регрессии) оцениваются по формулам:

$$a_1 = K_{xy} / S_x^2 = rS_y / S_x , \tag{21}$$

$$a_0 = \bar{y} - K_{xy}\bar{x} / S_x^2 \tag{22}$$

Линейный корреляционный анализ

Корреляционный анализ - совокупность методов оценки коэффициентов, характеризующих корреляцию между случайными величинами, и методов проверки гипотез об их значениях, основанных на выборочных данных. В случае линейной корреляции между двумя случайными величинами корреляционный анализ сводится к анализу коэффициента корреляции. Он позволяет ответить на вопрос: существует ли линейная зависимость между математическими ожиданиями случайных величин X и Y.

Коэффициент корреляции ρ лежит в пределах $-1 \le \rho \le +1$. Если $\rho=-1$ и $\rho=+1$, то все экспериментальные точки лежат на одной прямой (на линии регрессии) и существует строгая пропорциональность между y и x. Знак «+» характеризует положительную корреляцию ($\bar{y}_x$ растет с увеличением x). Знак «-» характеризует отрицательную корреляцию ($\bar{y}_x$ уменьшается с ростом x). Если $\rho=0$, то между величинами X и Y нет линейной связи. Чтобы ответить на вопрос, находятся ли случайные величины X и Y в линейной корреляционной связи, необходимо по выборке объема n найти выборочный коэффициент корреляции и вычислить величину

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{23}$$

Линейная корреляционная зависимость имеет место с вероятностью $1-\alpha$, если $|t| \ge t_{\alpha,n-2}$, где $t_{\alpha,n-2}$ - квантиль распределения Стьюдента для уровня значимости α и числа степеней свободы $f=n-2$, который находят по таблицам распределения Стьюдента, приведенным в данном пособии.

Оценка точности коэффициентов линейного уравнения регрессии

Соотношения (22) и (23) позволяют получить лишь оценки коэффициентов эмпирического уравнения регрессии. Эмпирические дисперсии коэффициентов регрессии определяют по формулам

$$S_{a_0}^2 = S_{yx}^2 \left(\frac{1}{n} + \frac{(x)^2}{(n-1)S_x^2} \right) \tag{24}$$

$$S_{a_1}^2 = S_{yx}^2 \frac{1}{(n-1)S_x^2} \tag{25}$$

Здесь S_{yx} - выборочное стандартное отклонение экспериментальных значений y_i от условных математических ожиданий $\overline{y}_x$, характеризующее рассеяние эмпирических точек относительно линии регрессии и вычисляемое по соотношению

$$S_{yx} = \left[\frac{n-1}{n-2} S_y^2 (1 - r^2) \right]^{\frac{1}{2}} \tag{26}$$

Доверительные интервалы для коэффициентов a_0 и a_1 вычисляют по формулам:

$$a_0 \pm \Delta a_0 = a_0 \pm t_{\alpha,n-2} S_{a_0} \tag{27}$$

$$a_1 \pm \Delta a_1 = a_1 \pm t_{\alpha,n-2} S_{a_1} \tag{28}$$

где $t_{\alpha,n-2}$ - квантиль распределения Стьюдента для уровня значимости α.

Полуширина доверительного интервала (Δa_0 и соответственно Δa_1) является погрешностью соответствующего коэффициента регрессии.

Доверительный интервал для условных математических ожиданий

Полуширина доверительного интервала ε, накрывающего с доверительной вероятностью 1-α условнее математическое ожидание $\overline{y}_x$, вычисляют по формуле

$$\varepsilon = t_{\alpha,n-2} S_{yx} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)S_x^2}} \tag{29}$$

Эта величина является погрешностью, возникающей при вычислении $\overline{y}_x$ по заданному x с помощью эмпирического уравнения регрессии.

Планирование эксперимента

Простейшая задача при планировании эксперимента состоят в определении минимального объема выборки n^*, при которой погрешность условного математического ожидания, вычисляемого по регрессионной формуле, не превышает заранее обусловленной величин. Вычисления выполняются по соотношению (29) на основании предварительной серии опытов объемом n, Для этого, обозначая предельно допустимую погрешность ε^*, вычисляют по (29) объём n_1^* выборки в первом приближении:

$$n_1^* = \frac{1}{(\varepsilon^*)^2} \tag{30}$$

Здесь $t_{\alpha,n-2}$ - квантиль распределения Стьюдента, который выбирают по таблице при уровне значимости α и числе степеней свободы $f=n-2$, а величины $\overline{x}, S_x^2$ и S_{yx}^2 вычисляют по формулам (10) , (12) и (26) на основании предварительной серии опытов объема n. Более точный результат получают во втором и третьем приближениях.

Второе и третье приближения n_2^* и n_3^* рассчитывают по формулам

$$n_2^* = \frac{1}{(\varepsilon^*)^2} \left[t_{\alpha,n_1^*-2} S_{yx}^2 \left(1 + \frac{(x - \overline{x})^2}{S_x^2} \right) \right] \tag{31}$$

$$n_3^* = \frac{1}{(\varepsilon^*)^2}\left[t_{\alpha,n_2^*-2} S_{yx}\left(1+\frac{(x-\bar{x})^2}{S_x^2}\right)\right] \qquad (32)$$

Квантили распределения Стьюдента $t_{\alpha,f}$ для различных α и f

$t_{\alpha,f}$	α					
	0,10	0,05	0,020	0,010	0,002	0,001
1	6,314	12,706	31,821	63,657	318,3	636,6
2	2,920	4,303	6,965	9,925	22,327	31,600
3	2,353	3,182	4,541	5,841	10,214	12,922
4	2,132	2,776	3,747	4,604	7,137	8,610
5	2,015	2,571	3,365	4,032	5,893	6,869
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,408
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,974	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	3,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,610	3,922
19	1,792	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,522	3,849
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,305	3,792
23	1,714	2,069	2,500	2,807	3,485	3,767
23	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,386	3,646
40	1,684	2,021	2,423	2,704	3,307	3,551
50	1,676	2,009	2,403	2,678	3,262	3,495
60	1,671	2,000	2,390	2,660	3,232	3,460
80	1,664	1,990	2,374	2,639	3,185	3,415
100	1,660	1,984	2,365	2,626	3,174	3,389
200	1,653	1,972	2,345	2,601	3,131	3,339