Глава вторая

Основные понятия математической статистики и измерение связи

1. Генеральная совокупность и частотное распределение

Фундаментальным понятием математической статистики является понятие группы, или совокупности, которое обычно определяется как генеральная совокупность. Генеральная совокупность—это совокупность, множество элементов, обладающих каким-то одним или многими признаками [1].

Признак является переменной величиной для каждого элемента генеральной совокупности и называется вариантой.

Количественная варианта может быть прерывной (дискретной и непрерывной. Если дана генеральная совокупность N лиц, которые изучаются, например, по своему доходу, то в этом случае варианта (доход) является непрерывной величиной, которая может в определенных пределах принимать любые значения. Если же эти N лиц изучаются по их семейному положению, например, какова величина семьи, в которой живет данный индивид, то в этом случае варианта является величиной прерывной, поскольку она может принимать только целочисленные значения — 1, 2, 3... и т. д.

Рассмотрим случай прерывной варианты.

Предположим, что дана генеральная совокупность объема N, каждый элемент которой характеризуется прерывной вариантой X. Как можно охарактеризовать эту генеральную совокупность по данному признаку (варианте)?

[1] *Blalock H. M.* Social Statistics. N. Y., 1960; *Кендэлл* Д. и М. Теория статистики. М., 1960; Уилкс *С.* Математическая статистика. М., 1967; *Ferguson G. A.* Statistical Analysis in Psychology and Education. N. Y., 1966.

Самый естественный и простой путь — сгруппировать члены генеральной совокупности по всем возможным значениям признака. Сначала группируем элементы генеральной совокупности, имеющие наименьшее значение варианты, а именно значение X_1 Затем члены, имеющие значения X_2, X_3... и т. д. Наконец, отбираем члены, имеющие наибольшее значение варианты— X_k. Количество членов генеральной совокупности в каждой группе, соответствующей определенному значению варианты, называется частотой варианты X и обозначается через n_i

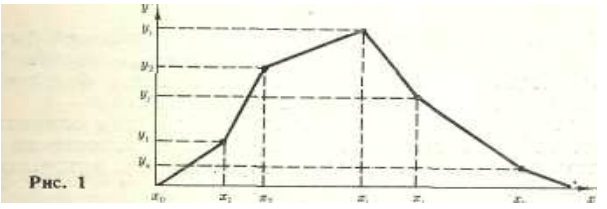
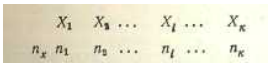


Рис. 1

В результате мы получаем два ряда чисел, которые можно расположить один под другим таким образом:



Получилась таблица, которая дает частное распределение варианта X в генеральной совокупности. Очевидно, что $\sum_{i=1}^{k} n_i = N$.

Иногда частотное распределение представляют графически: на оси X откладывают значение варианты, на оси Y — частоту. Полученные точки соединяют ломаной, которая называется полигоном распределения (рис. 1).

Ломаную принято соединять с осью X в смежных точках оси X, в которых, полагают, частоты равны 0 (в данном случае в точках X_0 и X_{k+1}).

В том случае, если варианта — непрерывная величина, дело несколько усложняется: нельзя непосредственно сгруппировать элементы генеральной совокупности по значениям варианты, поскольку может оказаться, что каждый член имеет свое, отличное от других значение варианты. Тогда поступают следующим образом. Предположим, что все значения варианты находятся на отрезке $[a, b]$. Этот отрезок разбивают на n равных

частей, которые называют разрядами, интервалами, классами или класс-интервалами. Отбирают члены генеральной совокупности, варианты которых попадают в первый класс-интервал, затем элементы, попавшие во второй класс-интервал, и т. д. вплоть до последнего *n-го* класс-интервала. Число элементов генеральной совокупности, попавших в определенный класс-интервал, называется частотой этого класс-интервала. Очевидно, что класс-интервал определяется по формуле

$$\Delta x = \frac{b-a}{n}.$$

Выбор *n* зависит от многих причин и должен быть таким, чтобы класс-интервал был не очень малым (чтобы класс-интервалов было не слишком много) и не очень большим (чтобы не исчезла специфика изменения варианты).

Существует ряд приближенных формул для определения необходимого *Ax,* а также ряд допущений в отношении значений варианты на границах класс-интервалов, за которыми мы отсылаем к соответствующей литературе [2].

Частотное распределение (в случае непрерывной варианты) будет иметь следующий вид:

Класс-интервалы	I	II	III ...
Частоты	n_{I}	n_{II}	n_{III} ...

В класс-интервалы кроме первого и последнего включаются варианты по своему значению больше нижней грани и равные верхней грани или меньше ее и условно принимается, что члены генеральной совокупности, попавшие в данный класс-интервал, имеют одинаковую варианту, равную середине данного класс-интервала.

Частотное распределение в случае непрерывной варианты также может быть изображено графически.

На оси *X* прямоугольной системы координат отмечаются точки *a* и *b* нижней и верхней грани изменения варианты. Определяется класс-интервал. На интервале *{a, b}* откладываются выбранные класс-интервалы. На каждом класс-интервале как на основании строится прямоугольник с высотой, пропорциональной частоте этого класс-интервала. Верхние основания всех построенных таким образом прямоугольников образуют некоторую ступенчатую линию, называемую гистограммой, ко-

[2] *Венецкий И. Г., Кильдищев Г. С.* Основы теории вероятностей и математической статистики. М., 1968,. с. 55.

торая и является графическим изображением данного частотного распределения (рис.2).

Если соединим середины верхних оснований прямоугольников гистограммы, то получим полигон данного распределения. Тем самым генеральную совокупность непрерывной варианты можно представлять двумя видами графиков — полигоном и гистограммой, а прерывной варианты — только одним видом —

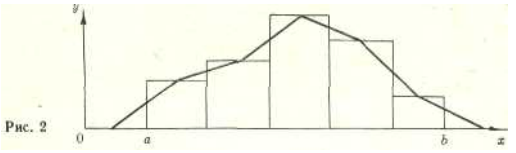


Рис. 2

полигоном. Площадь всей гистограммы пропорциональна объему генеральной совокупности.

Иногда вместо частоты применяют относительную частоту, равную отношению частоты к объему генеральной совокупности.

Если мы исследуем данную генеральную совокупность по варианте *X,* то прежде всего мы получаем частотное распределение, которое м-ожет быть представлено в виде таблицы или графика. Полученное в процессе исследования частотное распределение называется эмпирическим распределением.

Возьмем эмпирический полигон какой-либо непрерывной варианты. При достаточно большом объеме генеральной совокупности *N* будем одновременно увеличивать число и, следовательно, одновременно уменьшать величину класс-интервалов. У полигона будет увеличиваться число все уменьшающихся звеньев, и если продолжать этот процесс до бесконечности, то в пределе полигон перейдет в некоторую гладкую кривую, которая называется кривой распределения.

Каждый полигон эмпирического распределения является некоторым приближением определенной кривой распределения (рис.3).

Эта кривая распределения, являющаяся предельным случаем полигона данного эмпирического распределения, называется по установившейся терминологии функцией плотности распределения и обозначается *f(x).* Интеграл от нее по области

изменения варианты называется функцией распределения и обозначается

$$F(x)=\int_a^x f(x)\,dx.$$

Иногда f(x) и F(x) называют дифференциальным и интегральным законами распределения соответственно.

Возьмем какие-то эмпирические гистограмму и полигон и соответствующую им кривую распределения (рис. 4).

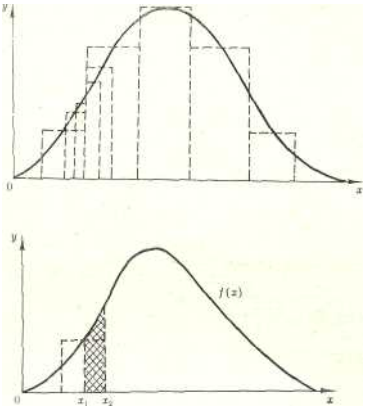


Рис. 3

Гистограмма и полигон в пределе стремятся к кривой распределения f{x). По определению гистограммы, частота значений варианты X равна площади прямоугольников, построенных на класс-интервалах. Частота события по частотному определению вероятности при бесконечном увеличении числа испытаний стремится к вероятности события[3].

Следовательно, для кривой распределения площадь под ней между значениями x^1 и x^2 это вероятность того, что варианта

[3] Гнеденко Б. В. Курс теории вероятностей. М., 1965, с. 37.

примет значения между x_1 и x_2 [4]. Это

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

Частотные распределения обычно характеризуются двумя типами параметров:

I —параметры положения или средние;

II — параметры или меры рассеивания.

Наибольшее значение имеют три вида средних: средняя арифметическая, медиана и мода.

Средняя арифметическая (M) для прерывной варианты генеральной совокупности объема N определяется выражением

$$M = \frac{1}{N}\sum_{i}^{N} x_i$$

или

$$M = \frac{1}{N}\sum_{i}^{k} n_i x_i,$$

где ***k*** —число различных значений варианты X, а x_i — значения варианты.

Для непрерывной варианты X, изменяющейся в интервале {a, b} генеральной совокупности объема N,

$$M = \frac{1}{N}\int_{a}^{b} xf(x)\,dx,$$

где f(x) —функция плотности распределения. Иначе говоря, средняя арифметическая есть абсцисса центра тяжести площади фигуры, образованной кривой распределения и осью абсцисс.

Медиана (Me) —это такое значение варианты, когда половина генеральной совокупности имеет значения меньше его, половина— больше.

Геометрически медиана означает абсциссу прямой, которая делит пополам площадь под кривой распределения.

Мода (Md) — значение варианты, соответствующее наибольшей частоте (вероятности). Графически мода — это значение абсциссы самой высокой точки кривой распределения (рис. 5).

Для симметричного распределения средняя арифметическая медиана и мода совпадают.

[4] Вентцель Е. С. Теория вероятностей. М., 1969, с. 81.

В качестве меры рассеивания наиболее распространены понятия дисперсии и квадратного корня из дисперсии, который называется стандартом или средним квадратическим отклонением.

Дисперсия есть средний квадрат отклонения варианты от ее среднего арифметического; она обозначается σ^2:

$$\sigma^2 = \frac{1}{N}\sum_{i}^{N}(M - x_i)^2,$$

стандарт σ:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}(M - x_i)^2}.$$

На рис. 6 кривая I характеризуется малой дисперсией; кривая II — большой дисперсией.

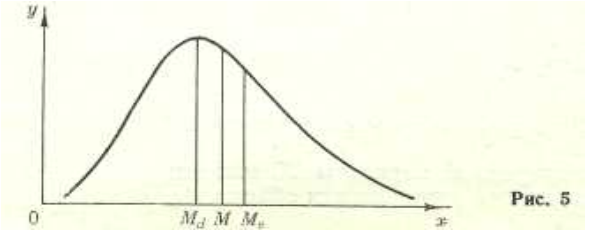


Рис. 5

Наибольшее значение для социологических исследований имеют три теоретических закона распределения.

1) Нормальное распределение, или распределение Гаусса (для непрерывной варианты):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-x)^2}{2\sigma^2}};$$

$$P(x < X < x + dx) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(a-x)^2}{2\sigma^2}} dx;$$

2) Биноминальное распределение, или распределение Бернулли (для прерывной варианты).

Если при каждом испытании вероятность осуществления события есть p, неосуществления — q — l—p, то вероятность того, что при ***n*** испытаниях это событие осуществится т раз, равна

$$P_{m,n} = c_n^m p^m q^{n-m}.$$

3) Распределение Пуассона, или закон малых чисел, представляет собой предельный случай биноминального распределения: когда n →∞, а p~0, то, обозначая *np*= α, имеем

$$P_m = \frac{\alpha^m e^{-\alpha}}{m!}.$$

Это распределение имеет место в случае большого числа испытаний маловероятных событий

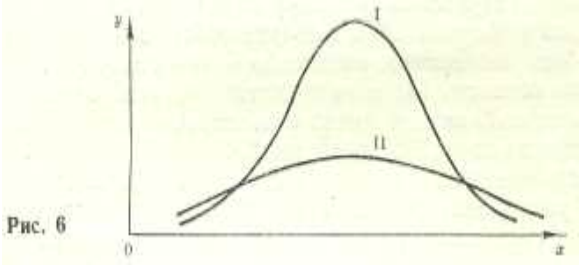


Рис. 6

2. Статистический вывод

Статистика имеет дело с большим числом предметов и явлений, которые образуют генеральную совокупность. Однако исследователь обычно имеет дело с ограниченной частью генеральной совокупности, называемой выборочной совокупностью, или просто выборкой, по изучению которой он делает определенные выводы о генеральной совокупности [5].

Каковы же математические основания этих выводов?

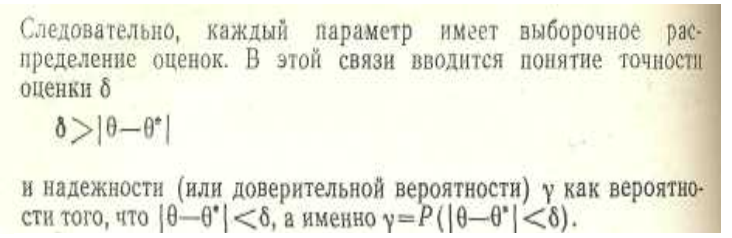
Если F(x)—интегральная функция распределения генеральной совокупности, определяющая вероятность того, что x<X, и если $\overline{F}$ (x) — эмпирическая функция распределения выборки, то по теореме Бернулли при бесконечном увеличении объема выборки эмпирическое распределение по вероятности стремится к распределению теоретическому:

$\overline{F}$ (x)→F(x).

Характеристики распределения генеральной совокупности принято называть параметрами θ, а характеристики выборочного распределения — оценками параметров θ *.

Статистическую выборку можно производить многократно, используя множество способов, и всякий раз будут получаться новые значения оценок параметров.

[5] Гнеденко Б. В. Курс теории вероятностей. М.—Л., 1950, с. 273.


Следовательно, каждый параметр имеет выборочное распределение оценок. В этой связи вводится понятие точности оценки δ

$$\delta > |\theta - \theta^*|$$

и надежности (или доверительной вероятности) γ как вероятности того, что $|\theta - \theta^*| < \delta$, а именно $\gamma = P(|\theta - \theta^*| < \delta)$.

При исследовании генеральной совокупности, подчиняющейся нормальному закону, находят оценки параметров α и σ; в случае распределения Пуассона — оценку параметра *m*.

Результат, полученный в выборке (обычно это среднеарифметическое или дисперсия), еще мало о чем говорит. Необходимо определить точность (δ) и надежность *(γ)* этой оценки. Без этого результат выборки не имеет смысла, поскольку оценка параметра является случайной величиной.

Точность оценки рассчитывается при определенных предположениях о распределении в генеральной совокупности. Может случиться, что генеральная совокупность отклоняется от предполагаемого теоретического распределения и, следовательно, расхождение эмпирического и теоретического распределения обусловлено не случайностью выборки, а тем, что данная генеральная совокупность характеризуется другим теоретическим распределением.

Всякое предположение о распределении генеральной совокупности называется статистической гипотезой. Встает проблема проверки статистической гипотезы. Гипотеза может касаться общего вопроса соответствия выборочного эмпирического и теоретического распределения. Она может относиться и к сопоставлению тех или иных параметров, например средних или дисперсий.

Обычно, следуя идее Дж. Неймана и Э. Пирсона, принимается начальная, или нулевая, гипотеза об отсутствии различия, которая обозначается H_0 [6].

В каждом отдельном случае определяется характеристика (критерий), по которой идет проверка. Если проверяется какой-либо параметр, а выборочное распределение его при данной гипотезе хорошо известно, то устанавливается предел вероятности, или уровень значимости. Значения характеристики, вероятности которых меньше уровня значимости, образуют так называемую критическую область, а значения, вероятности которых больше уровня значимости — область допустимых значений. Пусть дано выборочное распределение некоторой характеристики *u* (рис. 7).

Возможны два типа ошибок — так называемые ошибки первого и второго рода. Ошибка первого рода состоит в отбрасыва-

[6] *Варден В. дер.* Математическая статистика. М., 1960, с. 307.

нии нулевой гипотезы H_0, когда она верна. Ошибка второго рода связана с принятием нулевой гипотезы, когда она неверна.

Уровень значимости a определяет вероятность ошибки первого рода. Обозначим вероятность ошибки второго рода β. С уменьшением а увеличивается β. Величина $1—\beta$ называется мощностью критерия, с увеличением которой уменьшается вероятность ошибки второго рода[7].

При проверке гипотез приходится находить разумное соотношение уровня значимости и мощности критерия. Нельзя сделать

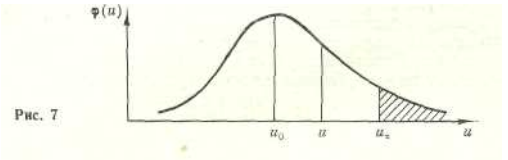


Рис. 7

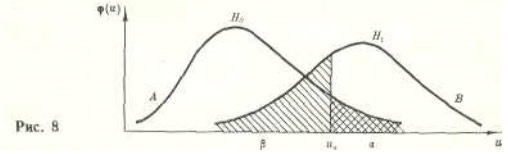


Рис. 8

как угодно малыми одновременно и α, и β. Здесь следует учитывать сложившуюся ситуацию. Это можно представить графически (рис. 8).

Кривая A связана с гипотезой H_0. Кривая B связана с альтернативной гипотезой H_1; u_a — значение критерия, соответствующее уровню значимости α.

Площадь справа от u_a под кривой H_0 дает α — вероятность ошибки первого рода.

Значение u_0 соответствует генеральной характеристике. Точка u_a определяет критическую область в том смысле, что вероятность значений $u>u\alpha$ оказывается меньше уровня значимости α (заштрихованная площадь справа от u_a равна α); α обычно полагают равным 1, 2 и 5%. Для каждого критерия строятся специальные таблицы, в которых имеются значения для каждой величины значения a и объема выборки.

[7] *Дунин-Барковский И., Смирнов Н.* Теория вероятностей и математическая статистика в технике. М., 1955, с. 360.

Если уменьшать а, то, следовательно, будет уменьшаться вероятность отбрасывания верной гипотезы, иначе говоря, станет меньше вероятность ошибки первого рода, но вместе с тем расширится область допустимых значений критерия. Таким образом, если в действительности нулевая гипотеза неверна, то увеличивается вероятность принятия неверной гипотезы.

Когда нулевая гипотеза неверна, то тем самым верна какая-то другая, альтернативная гипотеза H_1. Возможны такие случаи:

1) критерий отвергает H_0, и верна H_0;
2) критерий отвергает H_0, а верна H_1
3) критерий допускает H_0, и верна H_0;
4) критерий допускает H_0, а верна H_1.

Во втором и третьем случаях проверка гипотезы приводит к правильному выводу. Первый случай обусловливает ошибку первого рода, четвертый случай — второго рода.

Площадь слева от u_α под кривой H_1, *определяет* β, вероятность ошибки второго рода, т. е. вероятность принять гипотезу, когда она неверна.

Таковы некоторые положения о статистическом выводе. Использование математического аппарата статистического вывода имеет исключительно большое значение для социологии, так как, во-первых, социолог практически может проанализировать всю генеральную совокупность, а во-вторых, элементы генеральной совокупности в социологии гораздо более сложны и специфичны, чем в других областях науки.

Если ставится задача установить по выборке закон распределения, то используется так называемый критерий X^2 . При сравнении двух выборочных средних используется t-критерий, при сравнении, двух выборочных дисперсий — F-критерий [8].

3. Измерение связи

Из школьного курса математики известно понятие функциональной связи, когда каждому значению независимой переменной (аргумента *x)* ставится определенное значение зависимой переменной (функции *y):* .

$X, X_1, X_2 \ldots X_i \ldots$

$y, y_1, y_2 \ldots, y_i \ldots$

Функция может быть однозначной и многозначной. Так, длина окружности есть однозначная функция радиуса: $l = 2\pi r$. Квад-

[8] См., например: *Романовский В. И.* Математическая статистика, кн. 1—2. Ташкент, 1961.

ратный корень из действительного числа — двузначная функция этого числа: $y = \pm\sqrt{x}$. Обратные тригонометрические функции угла дают простейший пример многозначных функций.

На графике, в случае однозначной функции, каждой паре значений *x* и *y* соответствует точка плоскости. Множество этих точек на плоскости представляет собой графическое изображение функциональной связи, или график функции *y = f (x).*

Однако в природе существуют не только функциональные связи такого рода.

Рассмотрим обычную игру в карты. При сдаче игрок получает некоторое множество карт. При следующей — другое множество карт и т. д. При каждой сдаче получается новая комбинация. Налицо — связь между сдачей и комбинацией карт игрока.

В этом случае с изменением одной переменной происходит изменение распределения другой переменной. Связь этих переменных называется статистической. Если оказывается, что с изменением одной переменной изменяется среднее значение другой, то говорят, что между этими переменными существует корреляционная связь.

Например, требуется определить зависимость между ростом жены и мужа. Для примера рассмотрим 100 супружеских пар. На плоскости дана прямоугольная система координат, по оси *x* откладывается рост мужа, по оси *y* — рост жены. Точкой на плоскости отмечается каждая супружеская пара. Полученное графическое изображение называется корреляционным полем (рис. 9).

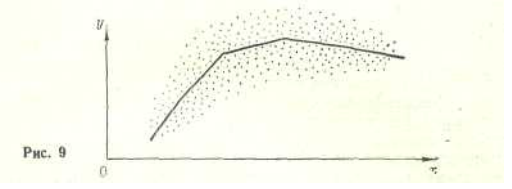


Рис. 9

В нашем случае должно быть 100 точек, которые как-то заполняют плоскость этого корреляционного поля. Для каждого класс-интервала *x* отбираем все соответствующие ему точки. Находим их среднее значение $\bar{y}$. Эту точку наносим на график, обозначая ее крестиком, чтобы выделить среди прочих. Соединяем ломаной все отмеченные крестиком точки. Полученная линия показывает изменение среднего значения роста жены с изменением

роста мужа от одного класс-интервала к другому. Эта линия называется эмпирической линией регрессии.

Если рассмотреть 100 других пар, то получится несколько иная эмпирическая линия регрессии. Если уменьшить величину класс-интервала, то линия покажет увеличение числа звеньев, сохранив в целом контур. Можно убедиться, что все эмпирические линии регрессии каких-либо двух переменных всегда лежат около некоторой плавной линии, называемой теоретической линией регрессии, или просто линией регрессии [9]. Ее уравнение называется уравнением регрессии. Если мы рассматриваем изменение среднего y от x, то получится уравнение регрессии y на x:

$$\bar{y}_x = \varphi(x)$$

Если рассматриваем изменение среднего x от y, то уравнение регрессии x на y;

$$\bar{x}_y = \varphi(y)$$

При $\varphi(x) = ax+b$ говорят о линейной регрессии y на x, т. е. $\bar{y}_x = ax+b$. Аналогично можно ввести уравнение регрессии x на y:

$$\bar{x}_y = cy+d$$

Как найти коэффициенты уравнений регрессии? Предположим, что дано n объектов, характеризующихся двумя переменными: x и y. Для простоты полагаем, что средние x и y равны 0. Выбираем прямоугольную систему координат x y, строим корреляционное поле, устанавливаем класс-интервалы для x и для y, проводим эмпирические линии регрессии, полагая, что искомые линии регрессии— прямые (рис. 10). Символически данная зависимость обозначается так: $y = f(x)$.

Линия $ACDF$ — эмпирическая линия регрессии y на x; PQ — линия регрессии, а ее уравнение $y=a+bx$, коэффициенты которого a и b неизвестны и их надо найти.

Коэффициенты теоретической линии регрессии находят по методу наименьших квадратов: ищут эту линию при том условии, чтобы сумма квадратов расстояний эмпирической линии регрессии от теоретической была бы минимальной. Иначе говоря, теоретическая линия регрессии должна иметь наиближайшее расположение ко всем точкам эмпирической линии регрессии.

Если мы обозначили ординату теоретической линии регрессии $y_т$, эмпирической — $y_э$, то надо найти минимум величины:

$$f = \Sigma(y_э - y_т)^2 = \Sigma(y_э - a - bx)^2.$$

[9] *Романовский В. И.* Элементарный курс математической статистики. М.— Л. 1939, с. 225—228.

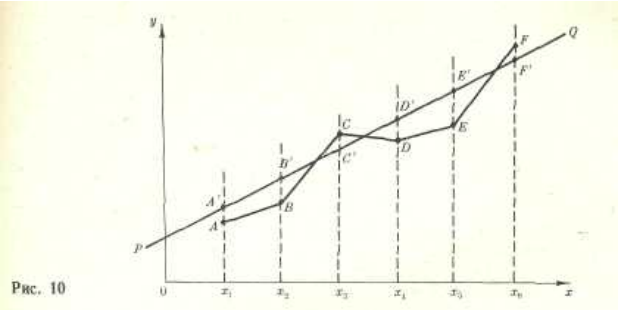


Рис. 10

Это означает, что

$$\frac{\partial f}{\partial a}=2\Sigma(y-a-bx)=0; \qquad \frac{\partial f}{\partial b}=2\Sigma(y-a-bx)x=0.$$

Получаем нормальные уравнения для определения коэффициентов линии регрессии:

$$\Sigma[y-(a+bx)]=0; \quad \Sigma[y-(a+bx)]x=0,$$

или

$$\Sigma y=na+b\Sigma x; \qquad \Sigma xy=a\Sigma x+b\Sigma x^2.$$

Поскольку средние x и y, как мы предположили, равны нулю то $\Sigma x=0$; $\Sigma y=0$ и, следовательно, $a=0$; $b=\frac{\Sigma xy}{\Sigma x^2}=\frac{\Sigma xy}{n\sigma_x^2}$,

где σ_x^2 — дисперсия x.

Найденный коэффициент b называется коэффициентом регрессии y на x и обозначается p_{yx}.

Аналогично можно построить линию регрессии x на y и соответственно найти коэффициент:

$$\rho_{xy}=\frac{\Sigma xy}{n\sigma_y^2},$$

где σ_y^2 — дисперсия y.

Коэффициент корреляции определяют как среднее геометрическое из коэффициентов регрессии [10]:

$$r_{xy} = \sqrt{\rho_{yx}\rho_{xy}} = \frac{\Sigma xy}{n\sigma_x\sigma_y}.$$

Дадим геометрическую интерпретацию коэффициенту корреляции [11] (рис. 11).

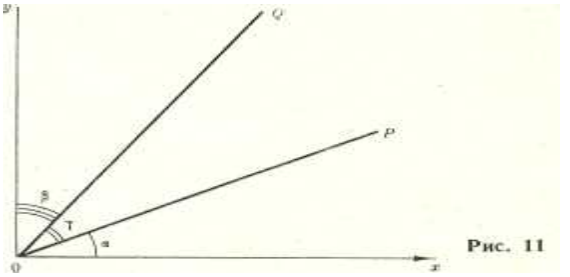


Рис. 11

OP — это линия регрессии *y* на *x;*

OQ — линия регрессии *x* на *y;* $p_{yx} = tg\ \alpha$; $p_{yx}, = tg\ \beta$ по определению коэффициентов регрессии

$$r_{xy} = \sqrt{\operatorname{tg}\alpha \operatorname{tg}\beta}.$$

Если корреляции нет, то или линия *OP,* или *OQ* или обе вместе совпадают с осями координат, так как*: a —0 или β — 0* и, следовательно, *r = 0.*

Если корреляционная связь переходит в функциональную, то обе линии регрессии совпадают. Тогда $\alpha+\beta=90°$, $r=\sqrt{\operatorname{tg}\alpha \operatorname{tg}\beta}=\sqrt{\operatorname{tg}\alpha \operatorname{tg}(90°-\alpha)}=\sqrt{\operatorname{tg}\alpha \operatorname{ctg}\alpha}=1$, т. е. коэффициент корреляции равен 1.

Чем теснее связь между переменными, тем меньше угол между обеими линиями регрессии.

Рассмотренный коэффициент корреляции измеряет линейную связь между двумя количественными переменными. Этим, одна-

[10] *Романовский В. И.* Элементарный курс математической статистики, с. 234

[11] *Слуцкий Е. Е.* Теория корреляции и элементы учения о кривых распределения. Киев, 1912, с. 85.

ко, не исчерпывается Все Возможное многообразие связей в социологии.

Во-первых, переменные могут иметь криволинейную регрессию: линия регрессии может быть параболой, кубической параболой, экспонентой и т. п. В каждом случае надо находить пути измерения связи между данными переменными.

Во-вторых, возможно наличие связи между более чем двумя переменными. Это проблема множественной корреляции, или многофакторного корреляционного анализа.

В-третьих, возможно существование связи между не только количественными переменными. В этом случае в статистике и социологии используются специальные показатели связи.

В случае криволинейной регрессии вместо коэффициента корреляции (иногда говорят «коэффициента линейной, или парной, корреляции») вводится корреляционное отношение[12].

$$\eta_{yx} = \frac{\sigma(\bar{y}_x)}{\sigma_y},$$

где $\sigma(\bar{y}_x)$ —среднее квадратичное отклонение условных средних $\bar{y}_x$ от их средней, σ_y — среднее квадратичное отклонение y (аналогично для $a(\bar{x}_y)$ и σ_x) и корреляционное отношение

$$\eta_{xy} = \frac{\sigma(\bar{x}_y)}{\sigma_x}.$$

Следовательно, прежде чем определять связь между количественными переменными социального объекта, необходимо сначала построить их линии регрессии и оценить характер регрессии. В том случае, если эмпирическая линия регрессии находится близко от некоторой прямой, можно вычислить коэффициент линейной корреляции Пирсона. Если же эмпирическая линия регрессии — явный изгиб, то надо использовать корреляционное отношение. При наличии более двух количественных переменных применяют частные коэффициенты корреляции [13].

Если, например, рассматривают три переменные *x, y, z,* то вводят частные коэффициенты корреляции $r_{xy,\,z}$; $r_{xz,\,y}$; $r_{zy,\,x}$; $r_{xz,\,y}$ определяет связь между x и z при исключении влияния переменной y:

$$r_{xz,y} = \frac{r_{xz} - r_{yz}r_{yx}}{\sqrt{(1-r_{xy}^2)(1-r_{yz}^2)}},$$

[12] *Романовский В. И.* Элементарный курс математической статистики, с. 272-281.

[13] Там же, с. 276

где справа находятся обычные коэффициенты парной корреляции. Выражения для двух других коэффициентов получаются простой круговой перестановкой индексов в правой части.

Можно также ввести понятия частых коэффициентов корреляции и сводный коэффициент для *n* переменных.

Для измерения связи между качественными (номинальными) переменными Используется таблица сопряженности.

	B_1	B_2	...	Bj	
A_2	n_{11}	n_{12}	...	n_{1j}	$n_{2.}$
A_2	n_{21}	n_{22}	...	n_{2j}	n_2
...	...	...	...	...	...
A_i	n_{1j}	n_{i2}	...	n_{ij}	n_i
	$n_{.1}$	$n_{.2}$	...	$n_{.i}$	n

Имеются два номинальных признака (переменные) *A и B,* которые принимают соответственно значения A_1, A_2.. . A_n и B_1, B_2 ... B_m. Это могут быть, например, образование (начальное, неполное среднее, среднее и высшее), социальное положение (рабочий, крестьянин, служащий, военный), возрастная группа (ученики, молодые рабочие, средний возраст, пожилые кадровые рабочие), участие в общественной жизни (не участвуют, слабо участвуют, участвуют, активно участвуют) и др.

Рассмотрим *N* лиц и их распределение по признакам *A* и *B*.

В каждой клетке первой строки пишется число лиц, которые одновременно обладают значением *A,* признака *A* и соответствующими значениями признака *B,* т. е. в левой клетке первой строки стоит n_{11} число лиц, обладающих признаками A_1 и B_1 одновременно, во второй клетке — n_{12} число лиц, обладающих признаками A_1 и B_2, и т. д. Вообще в клетке на пересечении *i*-й строки и /-го столбца находится число n_{ij}, обозначающее число лиц, обладающих признаками A_i и B_j.

Таблица сопряженности в данном случае очень сходна с корреляционной таблицей с той лишь разницей, что первая дает со-

вместные частоты качественных значений признаков, а вторая — совместные частоты класс-интервалов количественных признаков.

Вместо n_{ij} введем относительную частоту P_{ij}.

Пирсон предложил следующий коэффициент связи признаков A и B:

$$\varphi^2 = \sum \frac{(P_{ij} - P_{\cdot i} P_{\cdot j})^2}{P_{\cdot i} P_{\cdot j}},$$

который сконструирован так, что квадраты отклонений взвешены по отношению к ожидаемым частотам и нейтрализованно влияние знаков (как :в случае дисперсии).

При полной независимости переменных $\varphi^2 = 0$, при полной зависимости число строк равняется t — числу столбцов, и в таком случае $\varphi^2 > = t—1$.

Иногда используют так называемый коэффициент сопряженности в виде

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}},$$

где φ^2 — только что рассмотренный коэффициент. Коэффициент C дает более прямое непосредственное указание на связь между признаками.

Для определения связи между ранжированными переменными можно использовать так называемый ранговый коэффициент Спирмена:

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2-1)},$$

где n — число объектов; D_i— разности между значениями переменных для i-го объекта.

Рассмотрим числовой пример: даны 13 лиц, проранжированных по двум признакам. Результаты таковы:

Лица	Ранги		Разности		Лица	Ранги		Разности	
	I	II	D_i	D^2_i		I	II	D_i	D^2_i
А	3	1	2	4	Ж	1	3	—2	4
Б	4	2	2	4	З	2	7	—5	25
В					И				
Г	8	4,5	3,5	12,25	К	11	4,5	6,5	42,25
Д	5	6	—1	1		7	8,5	—1,5	2,25
Е	9	11	—2	4		6	8,2	—2,5	6,25

В первом столбце — лица; во втором — их ранги по первой переменной (признаку); в третьем — их ранги по второй переменной (признаку); в четвертом — разность рангов этих лиц. В последнем столбце — квадраты разностей, которые используются в формуле.

Можнопроизвести вычисления. Получим

$$\rho = 1 - \frac{6 \cdot 105}{10 \cdot (100 - 1)} = 0{,}36.$$

Большое значение для социологических исследований имеет бисериальный коэффициент корреляции, определяющий связь между количественной переменной и дихотомической качественной переменной. Он вычисляется по формуле

$$r_{pb} = \frac{\bar{y} - \bar{y}_0}{\sigma_y} \sqrt{\frac{N_0 N}{N_1 (N-1)}},$$

где N_1 — число индивидов с положительным ответом по признаку; N_0 — число индивидов с отрицательным ответом по признаку; N — общее число индивидов; $\bar{y}_1$ — средняя в N_1; $\bar{y}_0$ — средняя в N_0; y — средняя всей группы;

$$\sigma = \sqrt{\left[\Sigma \frac{y_i^2}{N} - \left(\Sigma \frac{y_i}{N} \right)^2 \right] \frac{N}{N-1}}.$$

Пример [14].

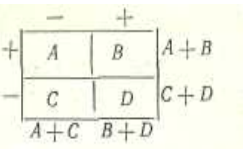
Лица	Количественная переменная	Качественная переменная		Лица	Количественная переменная	Качественная переменная	
		+	–			+	–
А	58	1	1	**Ж**	49	1	0
Б	45	0	0	**З**	52	1	1
В	56	0	1	**И**	54	0	0
Г	55	1	1	**К**	50	0	0
Д	51	0	1	**Л**	49	0	0
Е	44	1	0	**М**	59	0	1

Вычисления дают: $r_{pb} = 0{,}57$.

Представляет интерес для социологии группа коэффициентов для измерения корреляции в четырехклеточной таблице, которые измеряют связь между дихотомическими переменными.

[14] *Бернстайн А.* Справочник статистических решений. М., 1968, с. 92.

Для таблицы в виде



имеют место коэффициенты

$$\Phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}};$$

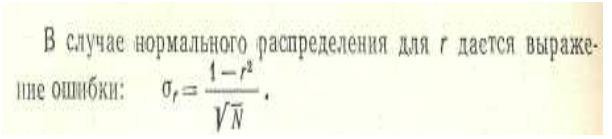
$$Q = \frac{BC - AD}{AD + BC}.$$

При измерении связи в конкретном социологическом исследовании мы вычисляем коэффициент корреляции по выборке. По сути дела, мы всегда располагаем только некоторой оценкой коэффициента корреляции генеральной совокупности. Любая выборочная оценка, как мы уже отмечали, требует проверки. Без указания точности расчета и проверки статистической гипотезы выборочная оценка не имеет смысла, к ней неизвестно как подступиться.

Остановимся па статистических оценках коэффициента парной корреляции r и рангового коэффициента корреляции Спирмена ρ.

Критические величины коэффициента корреляции Спирмена p

№	Уровень существенности		№	Уровень существенности	
	0,05	**0,01**		**0,05**	**0,01**
4	1,000		12	0,506	0,712
5	0,900	1,000	14	0,456	0,645
6	0,828	0,943	16	0,425	0,601
7	0,714	0,893	18	0,399	0,564
8	0,643	0,833	20	0,377	0,534
9	0,600	0,783	22	0,359	0,508
10	0,564	0,746	24	0,343	0,485
			26	0,329	0,465
			28	0,317	0,448
			30	0,306	0,432



Это означает, что коэффициент корреляции r_0 генеральной совокупности вероятностью 0,997 находится *в* интервале

$r - 3\,\sigma_r < r_0 < + 3\,\sigma_r$

Если использовать специальные таблицы, то можно построить доверительный интервал при данной доверительной вероятности и проверить нулевую гипотезу равенства выборочного и генерального коэффициентов корреляции.

Для проверки ρ Спирмена используют таблицу критических величин. Если $\rho_{табл} > \rho$ при данном количестве объектов и данном уровне существенности, то считается, что связь между ранговыми переменными существует. В нашем примере, вычисляя коэффициент Спирмена, мы получим значение p = 0,36. Используя таблицу, мы получим для n = 10 и уровня существенности 0,05 $\rho_{табл.}$= = 0,564. Следовательно, $\rho\ табл > \rho$, и мы полагаем, что ранжирование коррелировано.

В случае нормального распределения для r дается выражение ошибки: $\sigma_r = \frac{1-r^2}{\sqrt{N}}$.

Это означает, что коэффициент корреляции r_0 генеральной совокупности с вероятностью 0,997 находится *в* интервале $r-3\sigma_r<r_0<r+3\sigma_r$.

Если использовать специальные таблицы, то можно построить доверительный интервал при данной доверительной вероятности и проверить нулевую гипотезу равенства выборочного и генерального коэффициентов корреляции.

Для проверки p Спирмена используют таблицу критических величин. Если $p_{табл}>p$ при данном количестве объектов и данном уровне существенности, то считается, что связь между ранговыми переменными существует. В нашем примере, вычисляя коэффициент Спирмена, мы получим значение p = 0,36. Используя таблицу, мы получим для я = 10 и уровня существенности 0,05 $p_{табл.}$= = 0,564. Следовательно, $p_{табл}> p$, и мы полагаем, что ранжирование коррелировано.