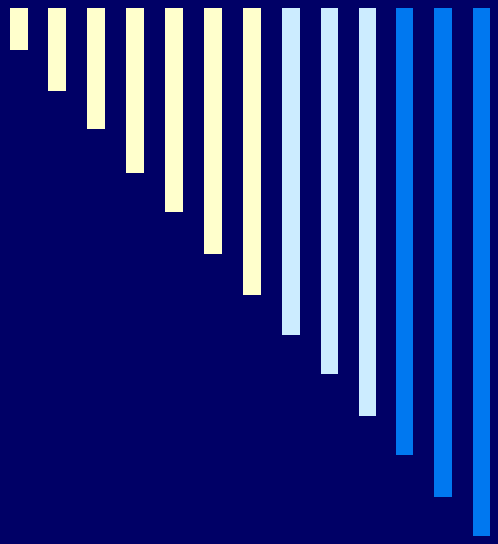


ЛЕКЦИЯ 7

ТЕМА: «КОРРЕЛЯЦИОННЫЙ АНАЛИЗ».

ПЛАН ЛЕКЦИИ:

- Функциональные и статистические связи. Корреляция и причинность.
- Корреляционная зависимость: отрицательная и положительная. Параметрические показатели связи. Коэффициент корреляции. Диаграмма рассеяния. Значение коэффициента корреляции.
- Ковариация. Дисперсия объединенной совокупности. Интерпретация коэффициента корреляции.
- Непараметрические меры связи: ранговый коэффициент корреляции Спирмена. Бисериальный коэффициент корреляции. Множественная корреляция. Частная корреляция.

Случаи, когда определенному значению одной переменной ***X***, называемой ***аргументом***, соответствует определенное значение другой переменной ***Y***, называемой ***функцией***.

Однозначная зависимость между переменными величинами ***X*** и ***Y***, называется ***функциональной***

$$Y = f(X)$$

Зависимость меду переменными величинами называется ***корреляционной*** или ***корреляцией*** от лат ***correlation*** – соотношение или связь.

Впервые термин был применен Ж. Кювье в труде «Лекции по сравнительной анатомии».

Развитие теории корреляции связано с Ф. Гальтоном и К. Пирсеном.

В биометрию ввел понятие Ф. Гальтон (1888г).

коэффициент корреляции -

показатель степени прямолинейной

связи между признаками

$$\operatorname{cov}\{x, y\}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Оценка среднего значения парных произведений центральных отклонений называется ***ковариацией.***

$$\operatorname{cov}\{x, y\}$$

Может рассматриваться как мера совместной вариации величин, как «совместная дисперсия x и y»

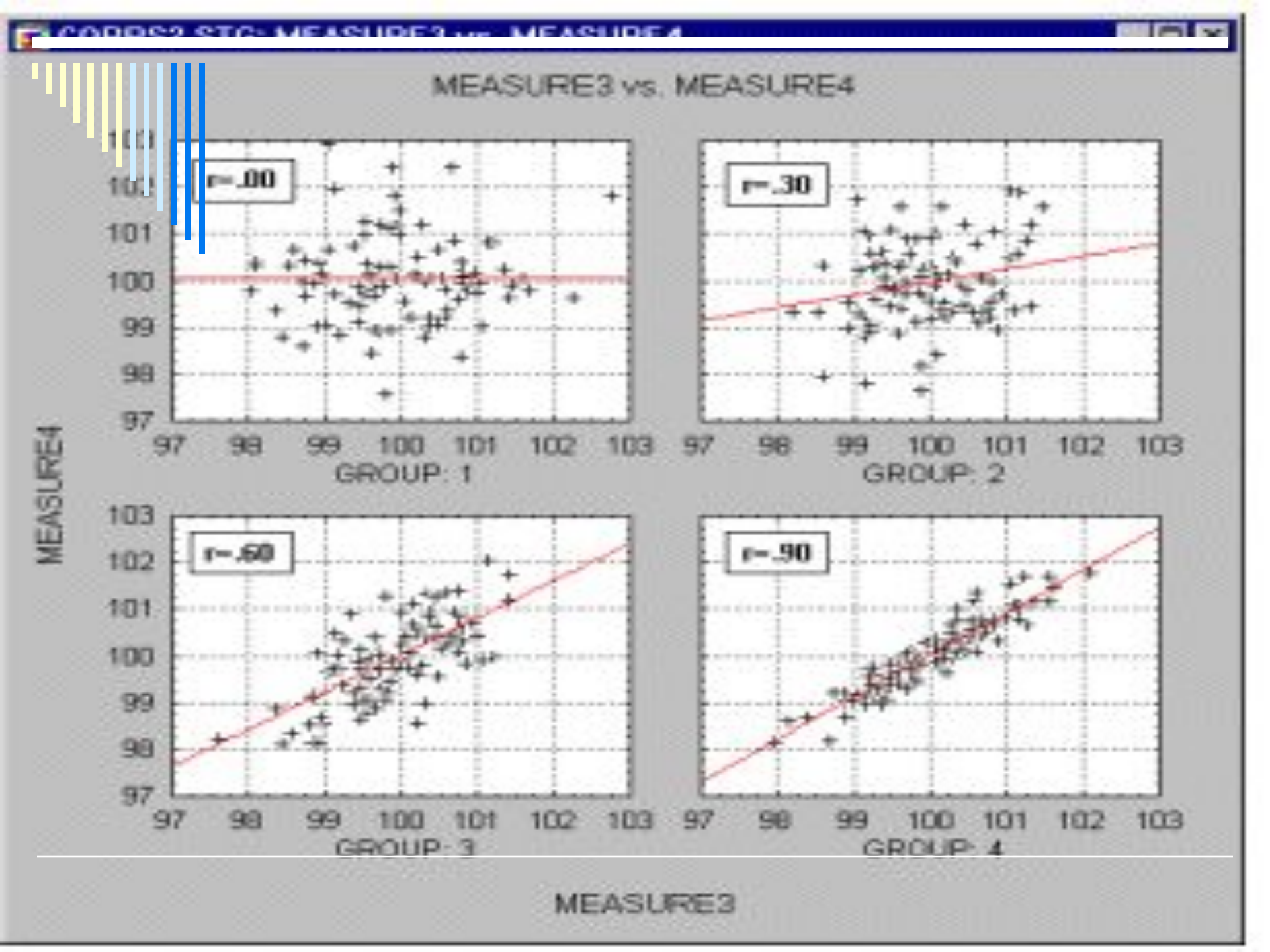
Оценку ошибки коэффициента корреляции вычисляют по формулам:

$при \quad n > 100$

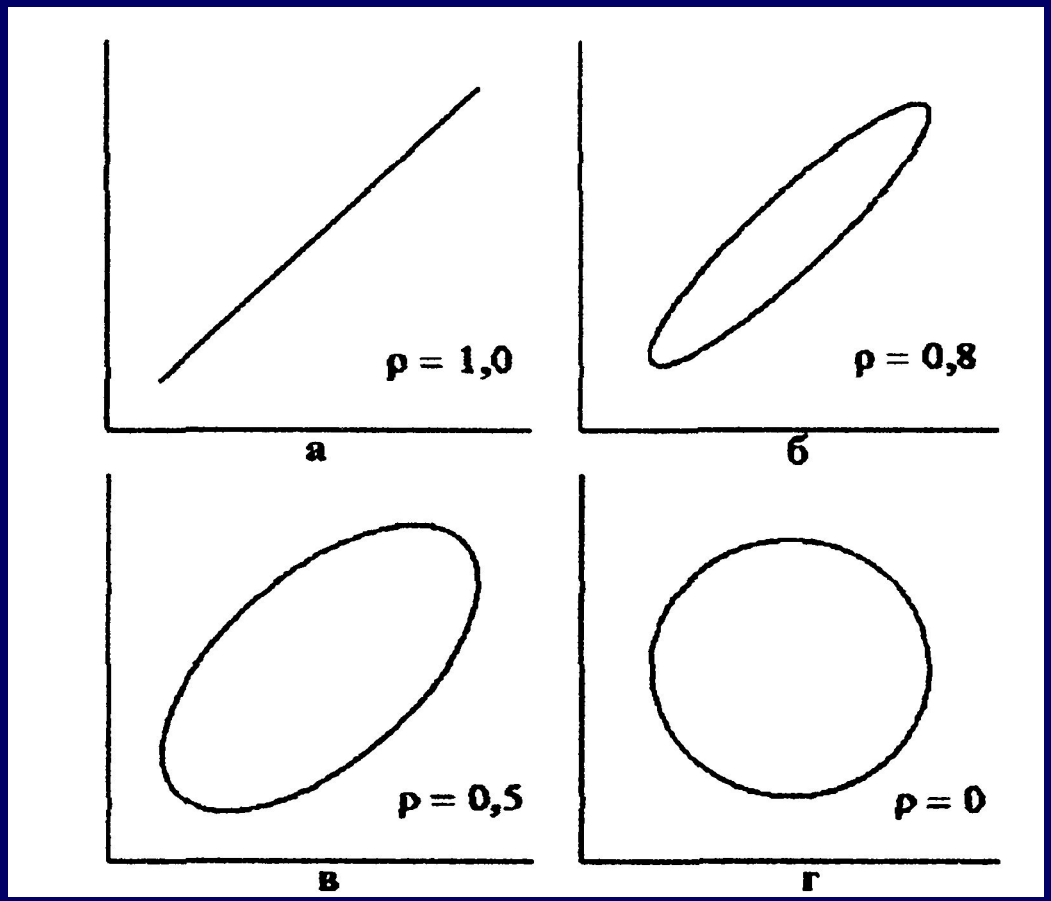
$$S_r = \frac{1 - r^2}{\sqrt{n}}$$

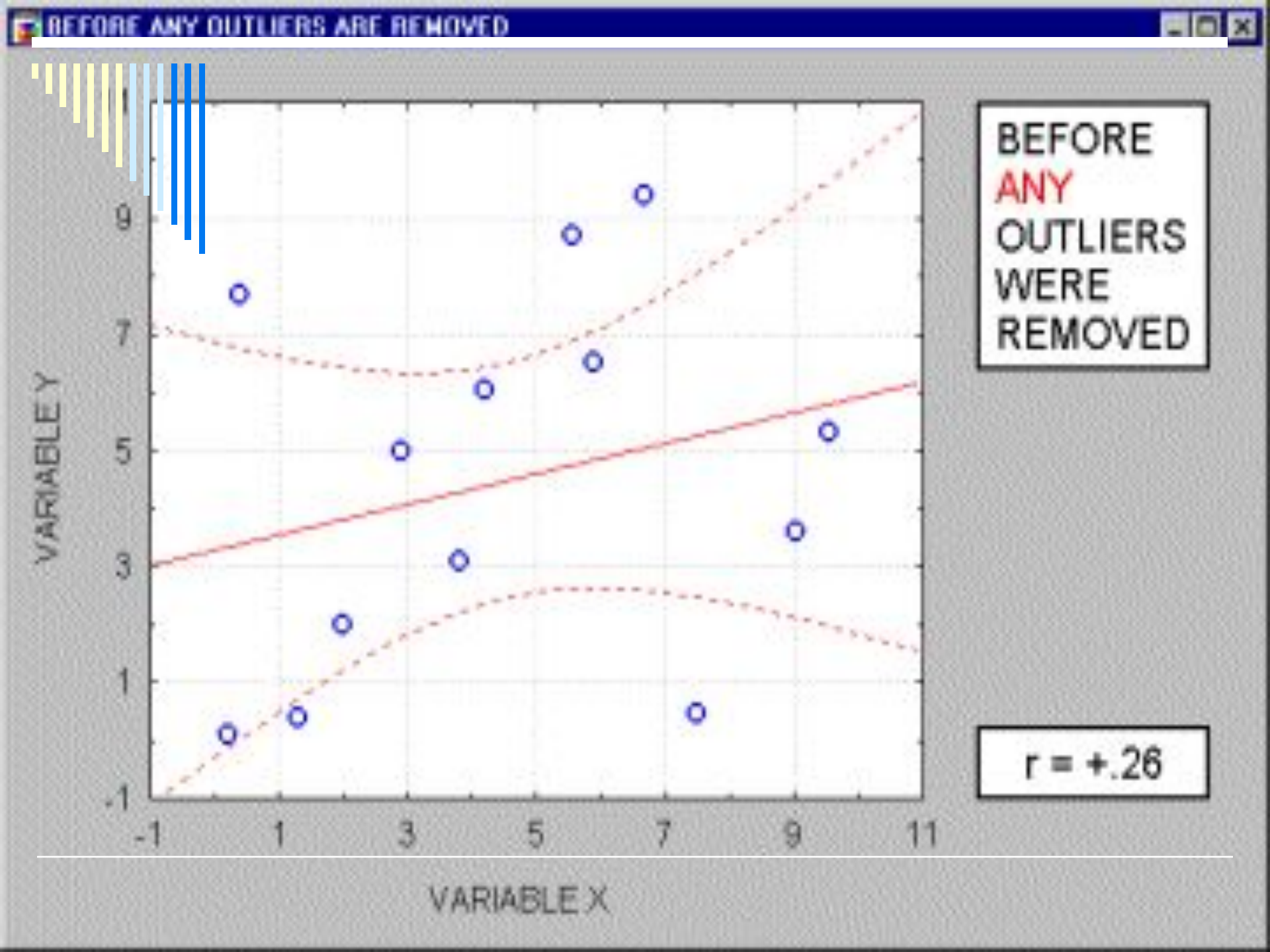
$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$при \quad n < 100$


MEASURE3 vs. MEASURE4
r=.00
r=.30
r=.60
r=.90
GROUP: 1
GROUP: 2
GROUP: 3
GROUP: 4
MEASURE4
MEASURE3
97
98
99
100
101
102
103

Вид корреляционного эллипса при различной степени связи




BEFORE ANY OUTLIERS ARE REMOVED
BEFORE
ANY
OUTLIERS
WERE
REMOVED
r = +.26
VARIABLE Y
VARIABLE X
-1
1
3
5
7
9
11

Величина и смысл коэффициента корреляции

при ***r > 0,85*** (при этом варьирование признаков взаимосвязано приблизительно на 75% и более) - весьма тесная связь,

при ***0,85 > r > 0,7*** (при этом взаимосвязанная вариация признаков лежит в пределах 75-50%) - тесная связь,

если ***r≤ 0,7*** (при этом варьирование одного признака менее чем на 50% связано с варьированием другого признака) - связь можно считать слабой.

Коэффициент детерминации (R^2) - величина квадрата коэффициента корреляции.

Величина R^2 показывает долю (%) части варьирования одного из признаков, связанную с варьированием другого.

Может иметь самостоятельный интерес, поэтому ее иногда выделяют в качестве особого параметра.

Минимальная повторность, которая может обеспечить значимость коэффициента корреляции при r = 0,70, есть $n_{0,05}=9$, что следует иметь в виду, если опыт планируется повторить.

Минимальное число наблюдений для планируемой точности

$$n = \frac{t^2}{z^2} + 3$$

Где:

n- искомый объем выборки;

t - величина, заданная по принятому уровню значимости;

z - преобразованная (по Фишеру) величина эмпирического коэффициента корреляции.

Вычисление коэффициента корреляции

Вычисление коэффициента корреляции

- Корреляционная решетка
- Способ условных средних

ОЦЕНКИ И ЗНАЧИМОСТЬ КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Для проверки нулевой гипотезы

$H_0: p = 0$

против альтернативы

$H_0: p \neq 0$

прибегают к вычислению статистики

t - Стьюдента

И если $t \geq t_a$ (t_a берется при $n < 100$ для $k = n - 1$, при $n > 100$ для $k = \infty$), то Но отвергается с соответствующим уровнем значимости делается утверждение о наличии линейной связи ($\rho \neq 0$).

Многомерный анализ корреляционных связей

Множественная корреляция

Используется когда корреляция измеряется одновременно между несколькими варьирующими признаками

$$R = \sqrt{\frac{r^2_{xz} + r^2_{yz} - 2 r_{xy} r_{xz} r_{yz}}{1 - r^2_{xy}}}$$

$$0 \le R \le 1$$

$r_{xy} r_{xz} r_{yz}$ – парные коэф. корреляции

Частный коэффициент корреляции

В случае, когда между любой парой признаков из X, У и Z связь не очень сильно отличается от прямолинейной и степень связи оценивается парными коэффициентами корреляции r_{xy}, r_{xz} и r_{yz}, то частный коэффициент корреляции $r_{xy(z)}$ между признаками X и У при исключенном влиянии Z может быть вычислен по формуле:

✔*Частный коэффициент корреляции имеет тот же смысл, что и обыкновенный парный коэффициент корреляции*

$$t = \frac{r_{частн}\sqrt{n-m}}{\sqrt{1-r^2_{частн}}}$$

$$r_{xy}(z) = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r^2_{xz})(1-r^2_{yz})}}$$

Коэффициент корреляции Спирмена

$$r_s = 1 - \frac{\sum d^2}{n(n^2 - 1)}$$

$\sum$ – знак суммирования

d – разность между рангами сопряженных значений признаков

$d = x_i - y_i$

n – число парных наблюдений

Используют в тех случаях, когда о законе распределений ничего не известно, а тем более, когда есть серьезные основания думать, что одна обе случайные величины имеют распределения заметно отличные от нормального или "засоренные" сильно отклоняющимися от основной массы значениями

Бисеральный коэффициент корреляции

$$r_{bs} = \frac{\overline{x_1} - \overline{x_2}}{S_x} \sqrt{\frac{n_1 n_2}{N(N-1)}}$$

$\overline{x_1}$ и $\overline{x_2}$ – средние арифметические альтернативных групп

$n_1 n_2$ – объемы этих групп

$N = (n_1 + n_2)$ – общее число наблюдений или объем выборки

S_x – среднее квадратическое отклонение для всей выборки

Применяется при измерении тесноты связи между качественными признаками, группируемыми в альтернативные группы (+ и -) и непрерывно варьирующими количественными признаками

Тест по теме:

«Корреляционный анализ»

Эмперический коэффициент (r), характеризующий степень связи между признаками в генеральных совокупностях

а) коэффициент пропорциональности	
б) коэффициент водопрочности	
в) коэффициент вариации	
г) коэффициент корреляции	

Значение коэффициента корреляции лежит в пределах

а) от 1 до ∞	
б) от -1 до +1	
в) от 1 до 10	
г) от – 10 до +10	

Что показывает коэффициент детерминации?

а) долю (%) части варьирования одного из признаков, связанную с варьированием другого	
б) отклонение вариант от средней арифметической	
в) разницу между максимальным и минимальным значением	
г) среди предложенных вариантов нет правильного ответа	

Какие задачи можно решать с помощью коэффициента Спирмена? Выберите неправильный вариант ответа

а) оценка закономерности изменения переменной в пространстве и во времени	
б) оценка степени связи, когда один или оба признака имеют нормальное распределение	
в) оценка степени связи когда один или оба признака измерены на порядковой шкале	
г) оценка степени связи, когда один или оба признака имеют распределение заметно отличное от нормального	

Укажите константу, не являющуюся характеристикой среднего уровня случайной величины

а) мода	
б) медиана	
в) дисперсия	
г) коэффициент корреляции	

С каким знаком может быть коэффициент корреляции между переменными X и Y?

а) только «+»	
б) только «-»	
в) «+» или «-»	
г) среди вариантов нет правильного	

Формула описывающая коэффициент корреляции в генеральных совокупностях X и Y

а) $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) S_x S_y}$	
б) $\rho = \frac{\operatorname{cov}\{x, y\}}{\sigma_x \sigma_y}$	
в) $-\infty \le r \le +\infty$	
г) другой вариант ответа	

Формула для расчета оценки непараметрического коэффициента корреляции Спирмена

а) $$r_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

б) $$r = \frac{Cxy}{\sqrt{C_{2x}C_{2y}}}$$

в) **другой вариант ответа**

г) $$t_б = r\sqrt{\frac{N-2}{1-r^2}}$$

Как оценивается статистическая значимость коэффициента корреляции?

а) $H_0 : \rho \leq 0$	
б) $H_0 : \rho \neq 0$	
в) $H_0 : \rho = 0$	
г) $H_0 : \rho \geq 0$	

Спасибо за внимание!