Курс «Статистика»

Л.И.Руденко
доцент,к.ф.-м.н., кафедра информационных систем в экономике

Тема 1. Введение.

Статистикой называется отрасль знаний, объединяющая принципы и методы работы с числовыми данными, характеризующими массовые явления. *Статистика* – это также отрасль практической деятельности, направленная на сбор, обработку и анализ статистических данных. Основными понятиями статистики являются:

- *статистическая совокупность*, или массовое явление, – *множество однокачественных варьирующих явлений;*
- *единица совокупности* – неделимый объект исследования, сохраняющий свойства изучаемого процесса;
- *признак* – характеристика или свойство единиц совокупности;
- *статистический показатель* – характеристика группы единиц или совокупности в целом;
- *система статистических показателей* – совокупности статистических показателей, отражающая объективные взаимосвязи характеристик и явлений;
- *статистическая закономерность* – закономерность изменения и развития массовых явлений, которая в массе явлений проявляет себя как необходимость, закон, а для каждой единицы связана со случайностью, индивидуальностью.

Предметом статистики является совокупность варьирующих явлений. ***Цель*** статистического изучения – обработка и анализ данных о статистической совокупности. *Статистический метод* включает сбор данных (статистическое наблюдение), их обобщение, представление, анализ и интерпретацию.

По характеру решаемых задач различают два основных раздела статистических исследований: *описательную статистику* (обобщение и анализ статистических наблюдений) и *статистический вывод* (обобщение выборочных закономерностей на всю совокупность).

Тема 2. Описательная статистика. Статистические показатели

Статистическая совокупность состоит из единиц совокупности, которые обладают характерными свойствами – признаками. *Признаки* единиц совокупности могут быть следующими: по характеру выражения – *описательные* (атрибутивными) и *количественные*; по способу измерения – *первичные* (учитываемые) и *вторичные* (расчетные); по отношению к характеризуемому объекту – *прямые* (непосредственные) и *косвенные* (связанные с другими объектами); по характеру вариации – *альтернативные* (два значения), *дискретные* (конечное множество значений), *непрерывные* (непрерывно изменяющиеся); по отношению ко времени – *моментные* и *интервальные* .

Статистический показатель – характеристика группы единиц или совокупности в целом. Статистические показатели могут быть следующими: по содержанию – *показатели свойств конкретных объектов* (средний возраст работников предприятия, объем реализованной продукции, рождаемость) и *показатели статистических свойств* любых массовых явлений (средние показатели, показатели вариации, показатели связи признаков); по количественной оценке – *абсолютные* (отражают суммарное свойство объекта, измеряются в натуральных единицах) и *относительные* показатели (получены путем сопоставления абсолютных и относительных показателей) .

Основные группы относительных показателей: 1) относительные показатели *структуры* объекта – доли (измеряются в процентах, промилле); 2) относительные показатели *динамики* – темпы роста, темпы прироста (цепные и базисные); 3) относительные показатели *взаимосвязи* – коэффициенты корреляции, детерминации, эластичности; 4) относительные показатели *интенсивности*, например, урожайность, трудоемкость (выражаются в именованных относительных единицах) ; 5) показатели отношения фактических величин признака к нормативным, например, показатели выполнения норм выработки, расхода ресурсов.

Основные функции статистических показателей – информационная, прогностическая, оценочная, рекламная.

Тема 3. Средние величины

Различие между индивидуальными явлениями совокупности называется вариацией. **Средняя величина** – это обобщающая мера варьирующего признака, характеризующая всю совокупность в целом. В статистике используют различные виды средних, в том числе:

Средняя арифметическая
(x_i - значения признаков, n – количество):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Средняя хронологическая
(x_i - значения моментных показателей, n- количество моментов с равными интервалами):

$$\bar{x} = \frac{\frac{x_1 + x_n}{2} + x_2 + \ldots + x_{n-1}}{n-1}$$

Средняя квадратическая
(x_i - значения признаков, n – количество):

$$\bar{x}_{кв} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Средняя геометрическая
(x_i - значения признаков, n – количество):

$$\bar{x}_{геом} = \sqrt[n]{x_1 \cdot \ldots \cdot x_n}$$

Средняя гармоническая
(x_i - значения признаков, n – количество):

$$\bar{x}_{гарм} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Средняя арифметическая взвешенная
(x_i - значения признаков, f_i - веса):

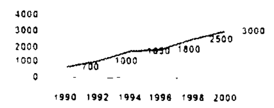
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

Тема 4.. Представление статистических данных

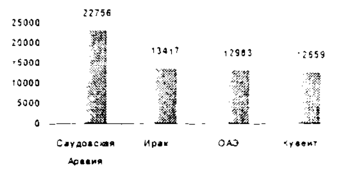
Статистическая таблица – это система строк и столбцов, в которых в определенной последовательности и связи излагается статистическая информация. Различают *подлежащее* и *сказуемое* статистической таблицы. В подлежащем указывается характеризуемый объект (единицы совокупности, группы, совокупность в целом). В сказуемом дается характеристика объекта, обычно количественная. По характеру подлежащего таблицы делятся на *простые, групповые, комбинационные*. Подлежащим простой таблицы является перечень всех единиц совокупности, территориальный или хронологический ряд. В групповой таблице подлежащим является группировка по одному признаку, в комбинационной – по двум и более признакам. *Заголовки* приводятся без сокращений, с указанием единиц измерения. *Итоговая строка*, как правило, завершает таблицу, но иногда может быть первой. Цифровые данные записываются с одинаковой точностью. В таблице не должно быть пустых клеток.

Наглядной формой представления статистических данных являются графики: *диаграммы, картограммы и картодиаграммы.* Наиболее распространены диаграммы, в том числе, *линейные, радиальные, точечные, плоскостные, объемные, фигуральные* диаграммы .

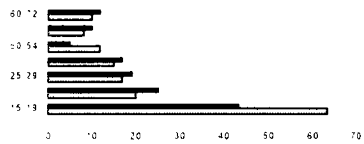
а)линейная диаграмма(график)



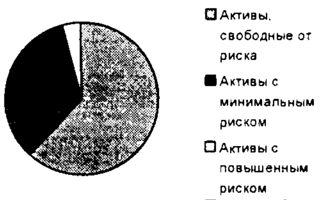
б) столбиковая диаграмма

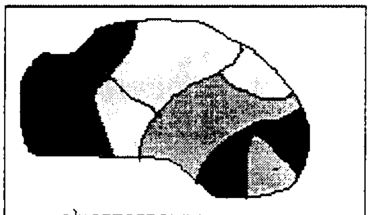


в) линейчатая диаграмма



д) секторная диаграмма





д)картограмма

Тема 5. Вариационный ряд

Вариационный ряд, или ряд распределения, характеризует состав, структуру совокупности по некоторому признаку. Элементами ряда распределения являются значения признака x_j и частоты f_j.

Частотные характеристики

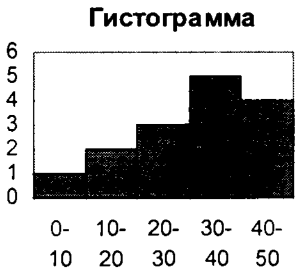
Индивидуальные значения признака $\{x_i\}_{i=1,\dots,n}$, n – количество наблюдений.

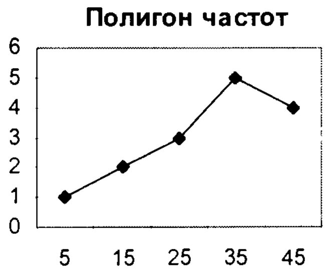
Число интервалов (групп) определяют по формуле: $k=1+3{,}32\ lg\ n = 1+1{,}44\ ln\ n$ (формула Штюргесса).

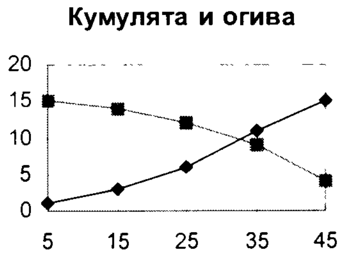
Длина интервала (шаг): $h=(x_{max} - x_{min})/k$. *Середина интервала:* x_j'. *Частота:* f_j ; *частость*: $w_j = \frac{f_j}{n}$; *накопленная частота* $\hat{f}_j = \sum_{\alpha=1}^{j} f_\alpha$. Плотность: абсолютная – $f_j' = \frac{f_j}{h}$; относительная – $w_j' = \frac{f_j}{n}$.

Графическое отображение:

-*гистограмма* (столбцовая диаграмма, на которой по оси абсцисс расположены интервалы значений признака, по оси ординат – интервальные частоты); -*полигон* (ломаная с вершинами $(x_j';f_j)$); *кумулята* (ломаная с вершинами $(x_j';\sum_{\alpha=1}^{j} f_\alpha)$); *огива* (ломаная с вершинами $(x_j;\sum_{\alpha=j}^{k} f_\alpha)$).







Структурные характеристики вариационного ряда

Среднее арифметическое значение $\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$

Медиана (величина признака, делящая совокупность значений на две равные части):

$$Me = x_{Me} + \left(\frac{\sum_{i=1}^{k} f_i}{2} - \hat{f}_{Me-1} \right) * \frac{h}{f_{Me}},$$

где x_{Me} – левая граница медианного интервала,

$f_{Me}, \hat{f}_{Me-1}$ -- соответственно частота медианного и накопленная частота предмедианного интервалов. В медианном интервале накопленная частота больше или равна половине общего числа единиц совокупности.

Мода – величина признака, которая встречается в вариационном ряду наиболее часто. В модальном интервале частота максимальна.

Точечная мода

$$Mo = x_{Mo} + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} \cdot h,$$

где $f_{Mo}, f_{Mo-1}, f_{Mo+1}$ -- частоты модального, предмодального и послемодального интервалов.

Характеристики вариации		
	для сгруппированных данных	*для несгруппированных данных*
Среднее арифметическое значение	$\bar{x} = \frac{\sum_{i=1}^{k} x'_i f_i}{\sum_{i=1}^{k} f_i}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Размах вариации		$R = x_{max} - x_{min}$
Средний модуль отклонений (среднее линейное отклонение)	$a = \frac{\sum_{i=1}^{k} \lvert x'_i - \bar{x} \rvert f_i}{\sum_{i=1}^{k} f_i}$	$a = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert$
Среднее квадратическое отклонение	$s = \sqrt{\frac{\sum_{i=1}^{k} (x'_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i}}$	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$

Выборочная дисперсия	$s^2 = \frac{\sum_{j=1}^{k}(x'_j - \bar{x})^2 f}{\sum_{j=1}^{k} f_j}$	$s^2 = \frac{\sum_{1}^{n}(x - \bar{x})^2}{n}$
Относительный размах вариации (коэффициент осцилляции)	$\rho = \frac{R}{\bar{x}}$	$\rho = \frac{R}{\bar{x}}$
Относительное отклонение по модулю (линейный коэффициент вариации)	$m = \frac{a}{\bar{x}}$	$m = \frac{a}{\bar{x}}$
Коэффициент вариации (квадратичный)	$v = \frac{s}{\bar{x}}$	$v = \frac{s}{\bar{x}}$

Моменты распределения

Центральные моменты	*для сгруппированных данных*	*для несгруппированных данных*
первый m_1	$\frac{\sum_{j=1}^{k}(x'_j - \bar{x}) f_j}{\sum_{j=1}^{k} f_j} = 0$	$\frac{\sum_{=1}^{n}(x - \bar{x})}{n} = 0$
второй m_2	$\frac{\sum_{j=1}^{k}(x'_j - \bar{x})^2 f_j}{\sum_{j=1}^{k} f_j} = s^2$	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = s^2$
третий m_3	$\frac{\sum_{j=1}^{k}(x'_j - \bar{x})^3 f_j}{\sum_{j=1}^{k} f_j}$	$\frac{\sum_{i=1}^{n}(x - \bar{x})^3}{n}$
четвертый m_4	$\frac{\sum_{j=1}^{k}(x'_j - \bar{x})^4 f_j}{\sum_{j=1}^{k} f_j}$	$s^2 = \frac{\sum_{=1}^{n}(x - \bar{x})^4}{n}$
Показатель асимметрии	$As = \frac{m_3}{s^3}$	$As = \frac{m_3}{s^3}$
Показатель асимметрии Пирсона	$As_{II} = \frac{\bar{x} - Mo}{s}$	$As_{II} = \frac{\bar{x} - Mo}{s}$
Показатель эксцесса	$Ex = \frac{m_4}{s^4} - 3$	$Ex = \frac{m_4}{s^4} - 3$

Нормальное распределение

$$\phi(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-v)^2}{2\sigma^2}};$$

$$F(x) = \int_{-\infty}^{+\infty} \phi(t)dt.$$

Тема 6. Группировка по признакам

Группировка – это разбиение совокупности на группы по какому-либо признаку. Группировка производится на основе *группировочного признака* и его значений, определяющих *интервалы группировки.* Группировка по одному признаку называется *простой,* по нескольким признакам – *сложной, комбинационной.. Виды группировок*: 1)*типологическая* группировка служит для выявления социально-экономических типов; 2)*структурная* группировка характеризует структуру совокупности по какому-либо признаку: 3)*аналитическая* (факторная) группировка характеризует взаимосвязь между двумя и более признаками. Примеры представления группировок таблицами:

а) простые структурные группировки;

Количество членов домохозяйства	Количество домохозяйств
2	4
3	10
4	6
Всего	20

Общий денежный доход, д.е.	Количество домохозяйств
Менее 200	3
200-400	10
400-600	5
600 и более	2
Всего	20

б) комбинационная группировка;

Количество членов домохозяйства	Общий денежный доход домохозяйства, д.е.				Всего
	Менее 200	200-400	400-600	600 и более	
2	2	2	0	0	4
3	1	5	3	1	10
4	0	3	2	1	6
Всего	3	10	5	2	20

в) аналитическая группировка.

Количество членов домохозяйства	Количество домохозяйств	Суммарное количество членов домохозяйств	Доход за месяц, д.е.		
			Общий денежный	В среднем	
				на одно домохозяйство	на одного члена домохозяйства
2	**4**	8	1096	274,0	**137,0**
3	**10**	30	3752	375,2	**125,1**
4	**6**	24	2652	442,0	**110,5**
По совокупности в целом	20	62	7500	375,0	121,0

Многомерная группировка

Многомерная группировка (*классификация)* основана на использовании *меры сходства* между объектами. Выделяют три типа мер сходства: 1)*коэффициенты подобия*, используемые для измерения степени близости между парами объектов, признаки которых принимают значения 0 и 1 (бинарные);

2)*коэффициенты связи*, в частности, коэффициенты корреляции (линейной корреляции для количественных признаков, ранговой корреляции для атрибутивных признаков); 3) *функции расстояния:*

а) хеммингово расстояние $d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$ для бинарных признаков; б) евклидово расстояние $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$ для количественных признаков и другие. Для выполнения многомерных классификаций применяют: метод дендритов, метод шаров, метод многомерной средней, метод корреляционных плеяд. Одной из распространенных процедур классификации является основанная на вычислении расстояний *процедура кластерного анализа*. Эта процедура предполагает последовательное оценивание расстояний между объектами и группами и объединение в группы (классы) объектов с наименьшими расстояниями.

Тема 7. Статистический вывод1. Выборочный метод

Метод статистического вывода позволяет по данным выборки делать заключения о свойствах генеральной совокупности – при условии репрезентативности выборки.

Выборочные характеристики: среднее значение, выборочная дисперсия, доля – являются *оценками параметров* **генеральной совокупности**: генеральная средняя, генеральная дисперсия, генеральная доля и др. Оценки могут быть состоятельными, несмещенными, эффективными. Выборочные характеристики позволяют сделать предположение о характере теоретического распределения.

Метод статистического вывода предполагает формулировку *статистических гипотез* и их проверку на основе *статистических критериев*.

Тема 8.Статистические гипотезы

Статистическая гипотеза – это некоторое допущение о свойствах генеральной совокупности, которое можно проверить по данным выборочного исследования. Гипотеза, которую следует проверить (*нулевая гипотеза*, H_0) формулируется как отсутствие расхождений между параметром генеральной совокупности А и его выборочной оценкой *a* (H_0: $A=a$), *альтернативная гипотеза*, противоположная нулевой, может иметь вид: H_1: $A>a$; H_1: $A<a$; H_1: $A\neq a$. При проверке гипотезы возможны две ошибки: *ошибка первого рода* – отвергнуть правильную гипотезу; *ошибка второго рода* – принять неправильную гипотезу. Вероятность α ошибки первого рода называют уровнем значимости. *Статистическим критерием* называют случайную величину *K* с известным законом распределения. В качестве статистических критериев используются стандартное нормальное распределение, распределения Стьюдента, χ^2, Фишера. *Критической областью* называют совокупность значений критерия, при которой гипотезу отвергают. Точки, отделяющие критическую область от области принятия гипотезы, называют *критическими точками*. Основной принцип проверки статистических гипотез: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

Тема 9. Репрезентативность выборки

Пусть $\{x_i\}$ – выборка размера *n*, для которой вычислены среднее значение $\bar{x}$, выборочная дисперсия s^2, исправленная выборочная дисперсия S^2. Пусть γ - доверительная вероятность (надежность), $\alpha=1-\gamma$ -- уровень значимости.

1) *Ошибка выборки* (точность оценки) для среднего значения.

При известной генеральной дисперсии σ: $\Delta_x = t\frac{\sigma}{\sqrt{n}}$: доверительное число t определяется по таблице функции Лапласа из соотношения $\Phi(t) = \frac{\gamma}{2}$.

При неизвестной генеральной дисперсии: $\Delta_x = t\frac{S}{\sqrt{n}}$ (для больших выборок), $\Delta_x = t\frac{S}{\sqrt{n-1}}$ (для малых выборок); доверительное число t определяется по таблице распределения Стьюдента: $t = t_{(1-\gamma;n-1)}$.

Доверительный интервал: $(\bar{x}-\Delta_x; \bar{x}+\Delta_x)$ с надежностью γ покрывает генеральную среднюю.

2) *Ошибка доли* $\Delta_p = t\sqrt{\frac{p(1-p)}{n}}$; *доверительный интервал* для доли $(p-\Delta_p; p+\Delta_p)$; доверительное число t определяется по таблице распределения Стьюдента: $t = t_{(1-\gamma,n-1)}$.

3)*Ошибка дисперсии* $\Delta_S = Sq$, *доверительный интервал* $(S-\Delta_S; S+\Delta_S)$; q – табличное значение при заданных n и γ (см. табл. приложения 4 [1]).

Тема 10. Проверка гипотезы о нормальном законе распределения

Используются следующие критерии согласия:

а) критерий САО (средних абсолютных отклонений): $\left|\frac{\sum|x_i-\bar{x}|}{nS} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{n}}$;

б) R/S-критерий: $(R/S)_н < R/S < (R/S)_в$, где R/S – отношение размаха вариации к среднеквадратическому отклонению, $(R/S)_н$, $(R/S)_в$,-- его табличные значения при заданном γ.

в) χ^2-критерий: $\chi^2 < \chi^2(1-\gamma;n-1)$, где $\chi^2(1-\gamma;n-1)$ – табличное значение (критическая точка); $\chi^2 = \sum\frac{(f_i - z_i)^2}{z_i}$-наблюдаемое значение, вычисленное для выборочных f_j и теоретических частот z_j.

Тема 11. Основы дисперсионного анализа

Модель дисперсионного анализа используется для изучения влияния качественного фактора F на изучаемый признак X. При этом сравниваются «факторная дисперсия», порожденная влиянием фактора и «остаточная дисперсия», обусловленная случайными причинами.

Пусть фактор F изменяется на p уровнях, x_{ij} – значения признака X в i-м испытании на j-м уровне фактора $(i = \overline{1,q}; j = \overline{1,p}; pq = n)$. Вводятся следующие суммы квадратов отклонений:

$SS_{общ} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x})^2$; $SS_{факт} = q\sum_{j=1}^{p}(\bar{x}_j - \bar{x})^2$; $SS_{ост} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x}_j)^2$ и

соответствующие дисперсии:

$s^2_{общ} = \frac{SS_{общ}}{pq-1}$ – общая; $s^2_{факт} = \frac{SS_{факт}}{p-1}$ – факторная; $s^2_{ост} = \frac{SS_{ост}}{p(q-1)}$ – остаточная дисперсии.

F-отношение $F = \frac{s^2_{факт}}{s^2_{ост}}$ имеет распределение Фишера со степенями свободы $p-1$ и $p(q-1)=n-p$. Гиптеза H_0, состоящая в том , что фактор F не оказывает влияния принимается, если $F \geq F_{(\alpha, p-1; n-p)}$, где $F_{(\alpha, p-1, n-p)}$ – критическая точка (процентная точка распределения Фишера).

Тема 12. Корреляционный и регрессионный анализ Задачи корреляционно-регрессионного анализа и моделирования

Различают два типа связей явлений: *функциональную* и *статистическую.* При функциональной связи величина y однозначно зависит от величины x и ни от чего более. При статистической связи разным значениям одной величины соответствуют разные *распределения* значений другой величины. Важнейший случай статистической связи – *корреляционная связь*, при которой разным значениям одной величины соответствуют различные *средние* значения другой величины. Предполагается, что результат и факторы являются количественными признаками.

Основные задачи корреляционно-регрессионного анализа: 1) вычисление параметров уравнения связи (*уравнения регрессии*) средних величин результативного признака с значениями одного или нескольких факторов: 2) оценка тесноты связи двух или более признаков между собой. Эти задач и решаются и исследуются на основе двух моделей: аналитической группировки и регрессионного анализа.

Модель аналитической группировки предполагает, что признак-фактор x изменяется на p уровнях (разбить на p групп), а индивидуальные значения результата $\{y_i\}_N$ преобразованы в интервальный ряди $\{y_l', f_l\}$, где f_l – частота в l-й группе по значению результата, f_j – частота в j-й группе по значению фактора, f_{jl} – частота результата l –й группы при значении фактора из j-й группы. После вычисления средних значений результата в j-й группе $\bar{y}_j = \frac{\sum_l y_l' f_{jl}}{\sum_l f_{jl}}$ и общей средней

$\bar{y} = \frac{\sum_j y_j f_j}{\sum_j f_j}$ вычисляются *общая* дисперсия, *межгрупповая* (*факторная*) дисперсия, *внутригрупповые* дисперсии и средняя из межгрупповых (*остаточная*) дисперсия:

$$s^2_{общ} = \frac{\sum_l (y_l^* - \bar{y})^2 f_l}{\sum_l f_l}; s^2_{факт} = \frac{\sum_j (\bar{y}_j - \bar{y})^2}{\sum_j f_j}; s_j^2 = \frac{\sum_l (y_l^* - \bar{y}_j)^2 f_{jl}}{\sum_l f_{jl}}; s^2_{ост} = \frac{\sum_j s_j^2 f_j}{\sum_j f_j}.$$

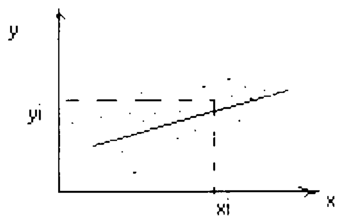
Теснота связи результативного признака y с признаком-фактором x оценивается величинами **коэффициента детерминации** $\eta^2 = \frac{s^2_{факт}}{s^2_{общ}}$ и **эмпирического корреляционного отношения** $\eta = \sqrt{\frac{s^2_{факт}}{s^2_{общ}}}$. Последняя величина принадлежит интервалу [0,1]: чем ближе к 1, тем теснее связь.

Тема 13.. Парная линейная корреляция

Уравнение парной линейной корреляционной связи называется *уравнением парной регрессии* и имеет вид: $\hat{y} = a + bx + \varepsilon$, где $\hat{y}$ - среднее значение результативного признака, вычисляемое по уравнению связи, a – свободный член уравнения, b – коэффициент регрессии, ε - ошибка. Параметры уравнения a, b определяются на основе метода наименьших квадратов (МНК) из условия минимизации суммы квадратов отклонений $\sum_{i=1}^{n}(y_i - (a + bx_i))^2 \to \min$, которое позволяет получить систему нормальных уравнений и найти a, b:

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n}(x_i)^2 = \sum_{i=1}^{n} x_i y_i; \end{cases} \Rightarrow$$

$$a = \bar{y} - b\bar{x};\ b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$



Графически линия регрессии есть прямая, проходящая через «облако точек» -- эмпирических данных.

Коэффициент регресси b имеет смысл показателя *силы связи* между вариацией факторного признака и вариацией результата: он измеряет среднее по совокупности отклонение y от его средней величины при отклонении признака x от своей средней величины на единицу измерения. Теснота связи измеряется коэффициентом детерминации $\eta^2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ и корреляционным отношением $\eta = \sqrt{\eta^2}$. Кроме того при линейной форме связи применяется стандартизованный коэффициент регрессии , или *коэффициент корреляции* $r = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$.

При этом $|r| \le 1$. Знак коэффициента определяет направление связи (при положительном r связь прямая).

Тема 14. Оценка надежности

1)Для коэффициента регрессии вычисляется средняя ошибка $s_b = \sqrt{\dfrac{\sum(y_i - \hat{y}_i)^2}{(n-2)\sum(x_i - \bar{x})^2}}$ и доверительное число $t = \dfrac{|b|}{s_b}$. Если $t > t_{(\alpha,n-2)}$, где $t_{(\alpha,n-2)}$ -- процентная точка распределения Стьюдента, то коэффициент b является значимым (надежным) с вероятностью 1-α.

2)Для уравнения регрессии вычисляется отношение $F=\frac{\frac{1}{n-1}\sum(y_i-\bar{y})^2}{\frac{1}{n-2}\sum(y_i-\hat{y}_i)^2}$ и критическое значение $F_{(n-1,n-2,\alpha)}$ по таблице распределения Фишера. Если $F> F_{(n-1;n-2;\alpha)}$, то уравнение описывает линейную связь с надежностью 1-α.

3)Для коэффициента корреляции вычисляется средняя ошибка $s_r=\sqrt{\frac{1-r^2}{n-2}}$ и доверительное число $t=\frac{|r|}{s_r}$. Если $t>t_{(\alpha;n-2)}$, то коэффициент корреляции значим с вероятностью 1-α.

Наряду с уравнением линейной регрессии для описания связи признаков используются следующие нелинейные модели, параметры которых также оцениваются на основе метода наименьших квадратов:

1)параболическая корреляция (регрессия) $\hat{y}=a+bx+cx^2$;

2)кубическая корреляция (регрессия) $\hat{y}=a+bx+cx^2+dx^3$:

3)гиперболическая корреляция (регрессия) $\hat{y}=a+\frac{b}{x}$,

а также другие модели, линейные по параметрам. Выбор оптимальной формы связи предполагает последовательной построение и оценку значимости различных уравнений связи, из которых предпочтение отдается тому, которое обеспечивает наименьшую остаточную дисперсию $s^2_{ост}=\frac{1}{n-2}\sum(y_i-\hat{y}_i)^2$.

Тема 15. Множественный корреляционный анализ

Если требуется изучить влияние на вариацию результативного признака y факторов x_1, …, x_p, то строится модель множественного линейного корреляционного анализа: $\hat{y}=b_0+\sum_{j=1}^{p}b_jy_j+\varepsilon$. Обозначим

$$X=\begin{pmatrix}1 & x_{11} & \dots & x_{1p}\\ \dots & & \dots & \\ 1 & x_{i1} & \dots & x_{ip}\\ \dots & & \dots & \\ 1 & x_{n1} & \dots & x_{np}\end{pmatrix};X^T=\begin{pmatrix}1 & \dots & 1 & \dots & 1\\ x_{11} & \dots & x_{i1} & \dots & x_{n1}\\ & \dots & & \dots & \\ x_{1p} & \dots & x_{ip} & \dots & x_{np}\end{pmatrix};B=\begin{pmatrix}b_0\\ b_1\\ \dots\\ b_p\end{pmatrix};Y=\begin{pmatrix}y_1\\ \dots\\ y_n\end{pmatrix}.$$

можно представить систему нормальных уравнений (метода наименьших квадратов) для определения параметров уравнения регрессии в матричном виде: $(X^TX)B=X^TY$, откуда $B=(X^TX)^{-1}X^TY$. Оценка надежности коэффициентов регрессии b_j основана на вычислении величин $s_{bj}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2c_{jj}}{n-2}}$, $t=\frac{|b_j|}{s_{bj}}$ (где c_{jj} -- диагональные элементы матрицы X^TX) и сравнении t с критическим значением $t_{(\alpha,n-p-1)}$.

Тема 16. Связь атрибутивных признаков

В случае изучения связи неколичественных (атрибутивных) признаков используются следующие меры связи:

1)коэффициент корреляции рангов Спирмена $r_s = \dfrac{\sum(p_{xi} - \bar{p}_x)(p_{yi} - \bar{p}_y)}{\sqrt{\sum(p_{xi} - \bar{p}_x)^2 \sum(p_{yi} - \bar{p}_y)^2}}$,

где p_{xi}, p_{yi} -- ранги признаков, т.е. их порядковые номера. Если обозначить $d_i^2 = (p_{xi} - p_{yi})^2$, то , с учетом того, что $\bar{p}_x = \bar{p}_y = \dfrac{n+1}{2}, \sum(p_{xi} - \bar{p}_x)^2 = \sum(p_{yi} - \bar{p}_y)^2 = \dfrac{n^3 - n}{12}$, получим: $r_s = 1 - \dfrac{6\sum d_i^2}{n^3 - n}$.

2)коэффициент ранговой корреляции Кендалла $r_k = \dfrac{4R}{n(n-1)}$, где $R = \sum_{j=1}^{n-1} R_j$, R_j -- количество рангов, больших, чем $p_{yj}, j = \overline{1, n-1}$,

а также коэффициент ассоциации (Пирсона) и коэффициент контингенции (Кендалла).

Тем 17. Статистическое изучение динамики. Динамический ряд

Динамический ряд – это размещенные в хронологической последовательности значения некоторого статистического показателя. В *моментном* динамическом ряду уровни показателя y_t фиксируют состояние явления на некоторые моменты времени t, в *интервальном* ряду – за некоторые промежутки. С течением времени значения уровней y_t варьируют. Если эта вариация в среднем монотонна, то говорят о *тенденции* (роста, снижения). Но в отдельные периоды уровни отклоняются от основной тенденции – испытывают *колебания*. Изучение тенденции и ее уравнения (тренда) и колеблемости – главные задачи динамического анализа.

Показатели тенденции динамики:

1)*абсолютное изменение* (асолютный рост) – цепное $\Delta_t = y_t - y_{t-1}$, базисное $\Delta_{t0} = y_t - y_0$;

2)*темп* изменения (темп роста) – цепной $k_t = \dfrac{y_t}{y_{t-1}}$, базисный $k_{t0} = \dfrac{y_t}{y_0}$; 3) *ускорение* (только цепное) $\Delta_t' = \Delta_t - \Delta_{t-1}$.

Средние показатели динамики:

1)*средний уровень* – для интервального ряда $\bar{y} = \dfrac{\sum_{t=1}^{n} y_t}{n}$; для моментного ряда $\bar{y} = (\dfrac{y_1 + y_n}{2} + \sum_{t=2}^{n-1} y_t)/(n-1)$; 2) *среднее абсолютное изменение* (*рост*) $\bar{\Delta} = \dfrac{\sum_{t=1}^{n} \Delta_t}{n} = \dfrac{y_n - y_0}{n}$;

3) *средний темп изменений* $\bar{k} = \sqrt[n]{\prod_{t=1}^{n} k_t} = \sqrt[n]{\dfrac{y_n}{y_0}}$.

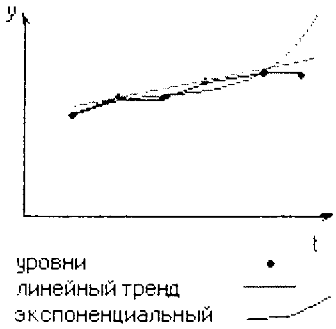
Тренд (трендовое уравнение) есть уравнение тенденции $y_t = f(t)$. При относительно стабильных приростах используют **линейный тренд** $\hat{y}_t = a + bt$, при при стабильных темпах прироста – показательную функцию $\hat{y}_t = ak^t$. Расчет

$$\begin{cases} na + b\sum_{t=1}^{n} t = \sum_{t=1}^{n} y; \\ a\sum_{t=1}^{n} xt + b\sum_{t=1}^{n} (xt)^2 = \sum_{t=1}^{n} ty; \end{cases} \Rightarrow$$

параметров производится на основе метода наименьших квадратов из системы нормальных уравнений:

(при переносе начала отсчета t=0 в середину ряда)

$$a = \bar{y}; b = \frac{\sum_{t=1}^{n} yt}{\sum_{t=1}^{n} t^2}.$$



Вычисленное по уравнению тренда значение y называется *точечным прогнозом*. Ошибка прогноза рассчитывается по формуле $s_p = s\sqrt{\frac{n+1}{n} + \frac{3(n+2v-1)}{n(n^2-1)}}$, где $s_{ост}^2 = \frac{1}{n-2}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$, v – период прогноза..

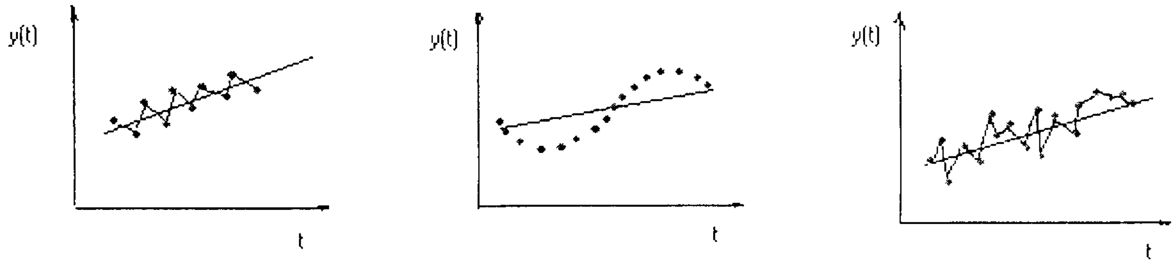
Показатели колеблемости:

среднее абсолютное отклонение $a(t) = \frac{\sum |y_t - \hat{y}_t|}{n-p}$, ($p$ – число параметров в уравнении тренда);

среднее квадратическое отклонение $s(t) = \sqrt{\frac{\sum (y_t - \hat{y}_t)^2}{n-p}}$;

амплитуда колебаний $R(t) = y_{\max} - y_{\min}$;

коэффициент колеблемости $v(t) = \frac{s(t)}{\bar{y}}$. Типы колеблемости:



1)пилообразная (маятниковая); 2)циклическая; 3)случайная (хаотичная).

Индексы

Индекс – это показатель сравнения двух состояний одного и того же явления. Индекс включает два вида данных – текущие(1) и базисные(0). Индексы выполняют две функции: *синтетическую* – это обобщающая характеристика изменения явления; *аналитическую* – изучение влияния отдельных факторов.

Индексы делятся на *динамические* (характеризуют изменение во времени), *территориальные* (изменение явления в пространстве, по регионам), *межгрупповые* (характеризуют отклонение от стандарта или среднего уровня). По степени агрегированности информации индексы делят на **индивидуальные** (i) и

сводные (I). Система обозначений индексируемых величин: p – цена, q – количество, c – себестоимость, t – трудоемкость.

Индивидуальный индекс динамики цен $i_p = \frac{p_{1_i}}{p_{0_i}}$, сводный индекс является средним из них. Для его построения необходимо условие созмеримости цен на разнородные товары, поэтому изпользуется вес – удельный вес товара в общем объеме в базисном периоде: $d_0 = \frac{q_0 p_o}{\sum q_0 p_0}$. Тогда сводный индекс $I_p = \frac{\sum i_p d_0}{\sum d_{0_i}} = \frac{\sum q_0 p_1}{\sum q_0 p_0}$ (индекс Ласпейреса). Если строить веса по отчетному периоду, то получают индекс $I_p = \frac{\sum q_1 p_1}{\sum q_1 p_0}$ (индекс Пааше).

Агрегатные индексы

Индексы, представленные в виде суммы произведений (агрегатов), называют *агрегатными*. Приведенные сводные индексы цен – агрегатные. Аналогично можно построить сводный индекс физического объема (количества): $I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0}$ (в форме Ласпейреса) или $I_q = \frac{\sum p_1 q_1}{\sum p_1 q_0}$ (в форме Пааше). В агрегатных индексах признаки различают как *индексируемые* и *весовые*. Так, в I_p индексируемй признак p, весовой – q. Значение индексируемого признака меняется: отчетное значение сопоставляется с базисным.

Агрегат в целом $\sum pq$ (товарооборот) оценивается общим индексом товарооборота: $I_{pq} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$.

Если индексы рассматриваются в системе, то должна выполняться взаимосвязь между ними, например, $I_{pq} = I_p I_q$. Эта связь обеспечивается, только если индексы строятся с весами *разных* периодов (один с базисным весом, другой – с текущим).

Литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1999. – 479 с.
2. Елисеева И.И., Юзбашев М.М. Общая теория статистики. – М.: Финансы и статистика, 1999. – 480 с.
3. Єріна А.М., Пальян З.О. Тоерія статистики. Практикум. – К.:Товариство "Знання", КОО,1997. – 325 с.
4. Харченко Л.П. Долженкова В.Г., Ионин В.Г. Статистика: Курс лекций. – М.: ИНФРА-М, 1998. – 310 с.
5. Афифи А., Эйзен С. Статистический анализ. – М.: Мир, 1982. – 478 с.