Лекция 9

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Статистические распределения и критерии

Здравствуйте!

В процессе изучения данного материала вы освоите:

- виды статистических распределения и критериев, которые наиболее часто используются при оценке показателей статистики
- нормальное распределение
- распределение Пирсона, Стьюдента и Фишера
- соответствующие критерии

Статистический ряд распределения представляет собой упорядоченное распределение единиц изучаемой совокупности на группы по определенному варьирующему признаку. Он характеризует состав (структуру) изучаемого явления, позволяет судить об однородности совокупности, закономерности распределения и границах варьирования единиц совокупности. Если вы будете изучать дополнительную литературу, независимо от автора, для понимания сути распределения будут необходимы следующие определения. В частности, ряды распределения, построенные по атрибутивным (качественным) признакам, называются атрибутивными (распределение населения по полу, занятости, национальности, профессии и т.д.). Ряды распределения, построенные по количественному признаку, называются вариационными (распределение населения по возрасту, рабочих – по стажу работы, зарплате и т.д.).

Вариационные ряды распределения состоят из двух элементов: вариантов и частот. Числовые значения количественного признака в вариационном ряду распределения называются вариантами. Они могут быть положительными и отрицательными, абсолютными и относительными. Частоты – это численность отдельных вариантов или каждой группы вариационного ряда, т.е. Это числа, показывающие, как часто встречаются те или иные варианты в ряду распределения. Сумма всех частот называется объемом совокупности и определяет число элементов всей совокупности. Частости – это частоты, выраженные в виде относительных величин (долях единиц или %). Сумма частостей равна 1 или 100%. Вариационные ряды в зависимости от характера вариации подразделяются на дискретные и интервальные. Дискретные вариационные ряды основаны на дискретных (прерывных) признаках, имеющих только целые значения, на дискретных признаках, представленных в виде интервалов. Интервальные вариационные ряды основаны на непрерывных признаках (имеющих любые значения, даже дробные).

Ранжирование ряда – расположение всех вариантов в возрастающем (убывающем) порядке. Графически интервальный ряд может изображаться графически в виде гистограммы. При ее построении на оси абсцисс откладывают интервалы ряда, высота которых равна частотам, отложенным на оси ординат. Над осью абсцисс строятся прямоугольники, площадь которых соответствует величинам произведений интервалов на их частоты. В практике также возникает потребность преобразования рядов распределения в кумулятивные ряды, строящиеся по накопленным частотам. Накопленные частоты определяются путем последовательного прибавления к частотам (или частостям) первой группы этих показателей последующих групп ряда распределения. Используя полученные данные, строят график в виде кумуляты (кривой сумм). Рассмотрим распределения, которые достаточно часто используются исследователями при решении практических задач и реализации проектов.

Поговорим дополнительно о критериях

Статистический критерий – строгое математическое правило, по которому принимается или

отвергается та или иная статистическая гипотеза с известным уровнем значимости. Построение критерия представляет собой выбор подходящей функции от результатов наблюдений (ряда эмпирически полученных значений признака), которая служит для выявления меры расхождения между эмпирическими значениями и гипотетическими.

Виды критериев

Статистические критерии подразделяются на следующие категории:
Критерии значимости. Проверка на значимость предполагает проверку гипотезы о численных значениях известного закона распределения:
Критерии согласия. Проверка на согласие подразумевает проверку предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону. Критерии согласия можно также воспринимать, как критерии значимости. Критериями согласия относится Критерий Пирсона о котором мы будем говорить. График нормальности – не столько критерий, сколько графическая иллюстрация: точки специально построенного графика должны лежать почти на одной прямой.

Критерии проверки на однородность. При проверке на однородность случайные величины исследуются на факт значимости различия их законов распределения (т.е. проверки того, подчиняются ли эти величины одному и тому же закону). Используются в факторном (дисперсионном) анализе для определения наличия зависимостей. Это разделение условно, и зачастую один и тот же критерий может быть использован в разных качествах.
Непараметрические критерии. Группа статистических критериев, которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами, к данной категории относится Критерий Пирсона.

Параметрические критерии

Группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии). t-критерий Стьюдента, Критерий Фишера. Основные понятия: Статистическая значимость критерия и мощность критерия.

Статистической гипотезой называется любое предположение относительно функции распределения наблюдаемых случайных величин. Если статистическая гипотеза полностью определяет функцию распределения наблюдаемых случайных величин, она называется простой статистической гипотезой. Если статистическая гипотеза не является простой, она является сложной. Сложная гипотеза указывает некоторое множество распределений. Обычно это множество распределений обладает определенными свойствами. Правило, согласно которому проверяемая гипотеза принимается или отвергается, называется статистическим критерием.

Нормальный закон распределения и его параметры

Нормальный закон распределения (часто называемый законом гаусса) играет исключительно важную роль в статистике среди других законов распределения особое положение. Это – наиболее часто встречающийся на практике закон распределения. Главная особенность,

выделяющая нормальный закон среди других законов, состоит в том, что он является предельным законом, к которому приближаются другие законы распределения при весьма часто встречающихся типичных условиях. Нормальное распределение распределение вероятностей случайной величины x, возникающее обычно, когда x представляет собой сумму большого числа независимых случайных величин, каждая из которых играет в образовании всей суммы незначительную роль.

Нормальное распределение унимодально, описывается колоколообразной (симметричной) кривой, что вы можете увидеть на рисунке и его средняя (математическое ожидание) совпадает с модой, чрезвычайно широко используется в статистике. В частности, в моделях регрессии, о которых мы будем с вами говорить позже, ошибка принимается распределенной по этому закону.

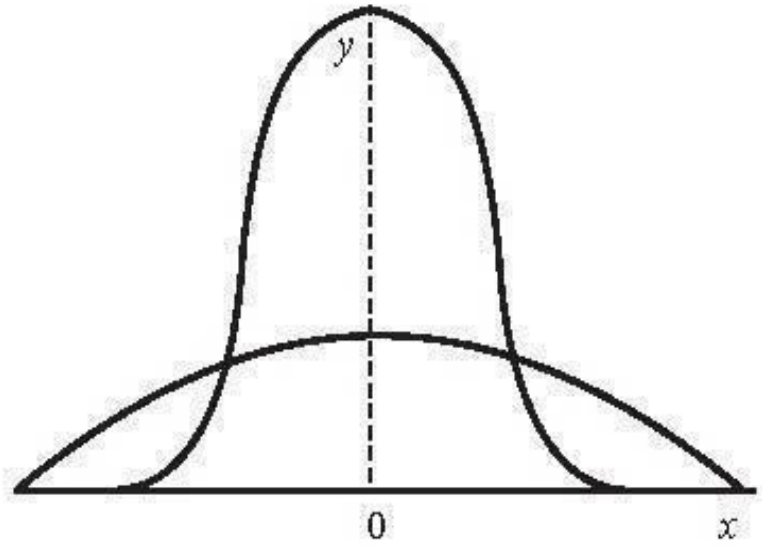

Рис. 9.1. Нормальное распределение. Представлены два нормальных распределения с разными дисперсиями.

Хи-квадрат или распределение Пирсона

Распределение (хи-квадрат) с n степенями свободы –это распределение суммы квадратов n независимых стандартных нормальных случайных величин.

Критерий хи-квадрат Пирсон, критерий $\chi 2$ Пирсона – это непараметрический метод, который позволяет оценить значимость различий между фактическим (выявленным в результате исследования) количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах. Выражаясь проще, метод позволяет оценить статистическую значимость различий двух или нескольких относительных показателей (частот, долей). Критерий хи-квадрат может применяться при анализе таблиц сопряженности, содержащих сведения о частоте исходов в зависимости от наличия фактора риска. Например, четырехпольная таблица сопряженности выглядит следующим образом:

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	A	B	A + b
Фактор риска отсутствует (0)	C	D	C + d
Всего	A + c	B + d	A + b + c + d

Условия и ограничения применения критерия хи-квадрат Пирсона

Сопоставляемые показатели должны быть измерены в номинальной или в порядковой шкале.
Данный метод позволяет проводить анализ не только четырехпольных таблиц, когда и фактор, и исход являются бинарными переменными, то есть имеют только два возможных значения (например, мужской или женский пол к примеру). Критерий хи-квадрат Пирсона может применяться и в случае анализа многопольных таблиц, когда фактор и (или) исход принимают три и более значений.
Сопоставляемые группы должны быть независимыми, то есть критерий хи-квадрат не должен применяться при сравнении наблюдений «до-»после». При анализе четырехпольных таблиц ожидаемые значения в каждой из ячеек должны быть не менее 10. В том случае, если хотя бы в одной ячейке ожидаемое явление принимает значение от 5 до 9, критерий хи-квадрат должен рассчитываться с поправкой. Если хотя бы в одной ячейке ожидаемое явление меньше 5, то для анализа должен использоваться точный критерий Фишера. В случае анализа многопольных таблиц ожидаемое число наблюдений не должно принимать значения менее 5 более чем в 20% ячеек.

Для расчета критерия хи-квадрат необходимо:
Рассчитываем ожидаемое количество наблюдений для каждой из ячеек таблицы сопряженности (при условии справедливости нулевой гипотезы об отсутствии взаимосвязи) путем перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений. Общий вид таблицы ожидаемых значений представлен ниже:

	Исход есть (1)	Исхода нет (0)	Всего
Фактор риска есть (1)	(A+B)*(A+C) / (A+B+C+D)	(A+B)*(B+D)/ (A+B+C+D)	A + B
Фактор риска отсутствует (0)	(C+D)*(A+C)/ (A+B+C+D)	(C+D)*(B+D)/ (A+B+C+D)	C + D
Всего	A + C	B + D	A+B+C+ D

Находим значение критерия χ2 по следующей формуле:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Где i – номер строки (от 1 до r), j – номер столбца (от 1 до c), Oij – фактическое количество наблюдений в ячейке ij, Eij – ожидаемое число наблюдений в ячейке ij.
Определяем число степеней свободы.
Сравниваем значение критерия χ2 с критическим значением при числе степеней свободы f (по таблице).
Распределение Стьюдента – это однопараметрическое семейство абсолютно непрерывных распределений.
t-критерий Стьюдента – общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента. Наиболее частые случаи применения t-критерия связаны с проверкой равенства средних значений в двух выборках. Т-критерий Стьюдента используется для определения статистической значимости различий средних величин. Может применяться как в случаях сравнения независимых выборок, так и при сравнении связанных совокупностей. Для применения t-критерия Стьюдента необходимо, чтобы исходные данные имели нормальное распределение. В случае применения двухвыборочного критерия для независимых выборок также необходимо соблюдение условия равенства дисперсий.
Как рассчитать t-критерий Стьюдента? Для сравнения средних величин t-критерий Стьюдента рассчитывается по следующей формуле:

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}$$

Где M1 – средняя арифметическая первой сравниваемой совокупности (группы), M2 – средняя арифметическая второй сравниваемой совокупности (группы), m1 – средняя ошибка первой средней арифметической, m2 – средняя ошибка второй средней арифметической. Полученное значение t-критерия Стьюдента необходимо правильно интерпретировать. Для этого нам необходимо знать количество исследуемых в каждой группе (n1 и n2). Находим число степеней свободы f по следующей формуле:

$$F = (n_1 + n_2) - 2$$

После этого определяем критическое значение t-критерия Стьюдента для требуемого уровня значимости (например, p=0,05) и при данном числе степеней свободы f по таблице.

Сравниваем критическое и рассчитанное значения критерия:
Если рассчитанное значение t-критерия Стьюдента равно или больше критического, найденного по таблице, делаем вывод о статистической значимости различий между сравниваемыми величинами.
Если значение рассчитанного t-критерия Стьюдента меньше табличного, значит различия сравниваемых величин статистически не значимы.

Критерий Фишера позволяет сравнивать величины выборочных дисперсий двух независимых выборок. Для вычисления Fэмп нужно найти отношение дисперсий двух выборок, причем так, чтобы большая по величине дисперсия находилась бы в числителе, а меньшая – в знаменателе. Формула вычисления критерия Фишера такова:

$$F_{эмп} = \frac{\sigma_x^2}{\sigma_y^2},$$

Где σ_x^2, $\sigma^2 y$ – дисперсии первой и второй выборки соответственно.
Так как, согласно условию критерия, величина числителя должна быть больше или равна величине знаменателя, то значение Fэмп всегда будет больше или равно единице.

Число степеней свободы определяется также просто:
K1 = nl -1 для первой выборки (т.е. для той выборки, величина дисперсии которой больше) и k2=n2 - 1 для второй выборки.

В таблицах распределения Фишера критические значения критерия Фишера находятся по величинам k1 (верхняя строчка таблицы) и k2 (левый столбец таблицы).

Прикладную сторону представленных определений мы рассмотрим в следующих лекциях.

Хотелось бы еще добавить, так как данная лекция носит информационный характер поэтому, рассмотрим перечень основных функция распределения в MS Excel. Поскольку данные функции достаточно сложно воспринимаются слушателями, и возникает ряд вопросов по поводу синтаксиса функция. Остановимся на них более детально. Практическую строну данных функций рассмотрим в последующих лекциях, где они будут представлены параллельно с функциями описательной статистики.

А пока вернемся к сложному.
ФункцияF.РАСП (функция F.РАСП) – возвращает F-распределение вероятности. Эта функция позволяет определить, имеют ли два множества данных различные степени разброса результатов.
Синтаксис имеет вид
Опишем аргументы функции X – обязательный аргумент. Значение, для которого вычисляется функция.

Степени_свободы1 – обязательный аргумент. Числитель степеней свободы.
Степени_свободы2 – обязательный аргумент. Знаменатель степеней свободы.
Интегральная – обязательный аргумент. Логическое значение, определяющее форму функции. Если аргумент «интегральная» имеет значение ИСТИНА, функция F.РАСП возвращает интегральную функцию распределения; если этот аргумент имеет значение ЛОЖЬ, возвращается функция плотности распределения.

НОРМ.СТ.РАСП (функция НОРМ.СТ.РАСП) Возвращает стандартное нормальное интегральное распределение. Это распределение имеет среднее, равное нулю, и стандартное отклонение, равное единице.

Данная функция используется вместо таблицы площадей стандартной нормальной кривой.

Синтаксис – стандартное нормальное распределение имеет вид
Опишем аргументы функции Z Обязательный. Значение, для которого строится распределение.

Интегральная Обязательный. Логическое значение, определяющее форму функции. Если аргумент «интегральная» имеет значение ИСТИНА, функция НОРМ.СТ.РАСП возвращает интегральную функцию распределения; если этот аргумент имеет значение ЛОЖЬ, возвращается весовая функция распределения.

СТЬЮДЕНТ.РАСП (функция СТЬЮДЕНТ.РАСП) Возвращает левостороннее t-распределение Стьюдента. Т-распределение используется для проверки гипотез при малом объеме выборки. Данную функцию можно использовать вместо таблицы критических значений t-распределения.

Опишем аргументы функции.
X Обязательный. Числовое значение, для которого требуется вычислить распределение. Степени_ свободы. Обязательный. Целое, указывающее число степеней свободы.

Интегральная Обязательный. Логическое значение, определяющее форму функции. Если аргумент «интегральная» имеет значение ИСТИНА, функция СТЬЮДЕНТ.РАСП возвращает интегральную функцию распределения; если этот аргумент имеет значение ЛОЖЬ, возвращается функция плотности распределения.

ХИ2.РАСП (функция ХИ2.РАСП) Возвращает распределение хи-квадрат. Функция распределения хи-квадрат обычно используется для изучения вариации в процентах какой-либо величины между выборками – например, части дня, которую люди проводят у телевизора.

Синтаксис ХИ2.РАСП (x;степени_свободы;интегральная).
Опишем аргументы функции. x – обязательный аргумент. Значение, для которого требуется вычислить распределение.
Степени свободы – обязательный аргумент. Число степеней свободы.
Интегральная – обязательный аргумент. Логическое значение, определяющее форму функции. Если аргумент «интегральная» имеет значение ИСТИНА, функция ХИ2.РАСП возвращает интегральную функцию распределения; если этот аргумент имеет значение ЛОЖЬ, возвращается функция плотности распределения.