Муниципальное бюджетное общеобразовательное учреждение
"Сеяхинская школа-интернат" с. Сеяха Ямальского р-на, ЯНАО

ТЕМА

«Обработка статистических данных с использованием элементов статистического и математического моделирования на уроках информатики в 11 классе»

учитель информатики и ИКТ
высшей квалификационной категории,
методист по ИТ
Куртлацкова ОльгаАнатольевна

2018 г.

Сегодня мы поговорим об **обработке статистических данных с использованием элементов статистического и математического моделирования на уроках информатики в 11 классе**.

Сразу скажу, что работаю я в 11 классе по УМК Семакина И.Г.: учебник плюс практикум. Тема в учебнике называется «Технологии компьютерного моделирования».

На первом уроке этой темы я кратко рассказываю учащимся о статистике - отрасле знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых количественных данных. Слово «статистика» происходит от латинского status — состояние дел.

В науку термин «статистика» ввёл немецкий ученый Готфрид Ахенвалль в 1746 году, предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины. Несмотря на это, статистический учёт вёлся намного раньше: проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, вёлся учёт имущества граждан в Древнем Риме и т.д. Например, чтобы подсчитать численность своей армии, персидский царь Дарий {522-486 гг. до н. э.) обязал каждого воина принести и положить в назначенное место камень. В Персии человека, занимающегося учетом, называли «глаза и уши царя».

По свидетельству греческого историка Геродота (484-420 гг. до н. э.), скифский царь Арианта, желая знать число своих подданных, приказал каждому скифу под страхом смертной казни принести медный наконечник стрелы (Геродот. История в девяти книгах. Т. I)

Статистика является мультидисциплиной, так как она использует методы и принципы, заимствованные из других дисциплин. Так, в качестве теоретической базы для формирования статистической науки служат знания в области социологии и экономической теории. В рамках этих дисциплин происходит изучение законов общественных явлений. Статистика помогает произвести оценку масштаба того или иного явления, а также разработать систему методов для анализа и изучения. Статистика, несомненно, связана с математикой, так как для выявления закономерностей, оценки и анализа объекта исследования требуется ряд математических операций, методов и законов, а систематизация результатов находит отражения в виде графиков и таблиц.

Итак, каково состояние дел в вашем классе по успеваемости? Каково состояние дел в школе после сдачи ЕГЭ учащимися 11 классов в этом учебном году? А в городе? А в области? А в России? А в прошлом году лучше или хуже? Это все статистика. Статистикой (в общем смысле этого понятия) мы занимаемся, даже, не задумываемся об этом.

Например, пиццерия. Моделируем ситуацию: человек, работающий там, получает сдельную зарплату. Ему важно знать, когда он больше продаст пицц: днем или вечером, в рабочий день или в выходные дни, когда тепло или когда холодно, зависит ли это от сезона года? Вольно или невольно, но он будет вести статистику посещаемости клиентов, в зависимости от тех или иных условий. Предлагаю учащимся рассмотреть модель бизнес-плана реальной пиццерии. Здесь они знакомятся с ценообразованием продукта.

Рассматриваем основные моменты бизнес-плана и далее предлагаю учащимся выдвинуть свои гипотезы, какие зависимости с их точки зрения могут иметь место. Как известно, социализация учащегося предполагает включение обучающегося в систему общественных взаимоотношений. Данная тема, как никакая другая предоставляет такую возможность для учителя.

Рассматриваем различные гипотезы, попутно объясняя учащимся, что имеют место разные, уходя от привычных догматов.

Выбираем 1 наиболее поддерживаемую гипотезу (вариативность ваших учащихся неограниченна). Например, продажа пицц зависит от времени года (зимой продается больше, чем летом или наоборот).

Как подтвердить или опровергнуть нашу рабочую гипотезу? Нужен сбор данных. Посмотрим отчетность за 2014 и 2015 годы реально существующей сети пиццерий "Додо Пицца".

Зависимость между величинами представлена в табличной форме. Рассматриваем и видим, что проанализировать достаточно сложно.

А в каком виде еще можно представить эти зависимости? (В графическом виде и виде математической модели - вспоминаем, что является математической моделью). А какой из этих видов наиболее наглядный? (Графический).

Строим точечную диаграмму и анализируем ее. Видим, что в целом идет рост продаж, но есть некоторые спады.

А как построить математическую модель полученных данных? Очевидно, что нужно получить формулу зависимости доходов от месяцев года. Вид такой функции неизвестен, потому что, глядя на график этой функции ученики не могут определить какой она будет. К тому же, напоминаю, что статистические данные всегда являются приближенными усредненными, поэтому они носят оценочный характер.

И тут приходит время рассказать о **методе наименьших квадратов,** предложенным в 18 веке Карлом Фридрихом Гауссом. Суть его заключается в следующем: *искомая функция должна быть построена так, чтобы сумма квадратов отклонений y-координат всех точек от y-координат графика функции была бы минимальной*. Не вдаваясь в подробности говорю о том, что этот метод широко используется в статистической обработке данных и встроен во многие математические пакеты программ.

По методу НК можно построить любую функцию. А вот будет ли она нам подходить, это уже другой вопрос.

Чаще всего выбор производится среди следующих функций:

y=ax+b - линейная функция;
y=ax2+bx+c - квадратичная функция;
y=a ln(x)+b - логарифмическая функция;
$y=ae^{bx}$ - экспоненциальная функция;
$y=ax^{b}$ - степенная функция

Квадратичная функция называется **полиномом второй степени**. Могут использоваться и полиномы более высоких степеней.

Показываю учащимся как строится график функции с использованием МНК, попутно объясняя, что полученную функцию в статистике принято называть **регрессионной моделью**, график регрессионной модели - **трендом**.

Термину регрессионная модель, используемому в регрессионном анализе, можно сопоставить синонимы: «теория», «гипотеза».

Эти термины пришли из статистики, в частности из раздела «проверка статистических гипотез».

Регрессионная модель есть прежде всего гипотеза, которая должна быть подвергнута статистической проверке, после чего она принимается или отвергается.

Обязательно показываем **уравнение на диаграмме** и **величину достоверности аппроксимации**, пока, не объясняя для чего она нужна.

Предлагаю выбрать наиболее подходящую для нашего запроса регрессионную модель. Обычно учащиеся выбирают между экспоненциальной и полиномиальной 2-ой степени.

И вот тут я объясняю для чего нужен параметр R^2, который в статистике называется **коэффициентом детерминированности**, а при выборе линии тренда мы ставим галочку на **"Поместить на диаграмму величину достоверности аппроксимации"** (это одна и та же величина).

Аппроксимация – приближенное решение сложной функции с помощью более простых, что резко ускоряет и упрощает решение задач.

Именно коэффициент аппроксимации и определяет, насколько удачной является полученная регрессионная модель. Он всегда заключен в диапазоне от 0 до 1. Если он равен 1, то имеет место полная корреляция с моделью, т. е. нет различия между фактическим и оценочным значениями y, и функция точно проходит через табличные значения. В противоположном случае, если коэффициент детерминированности равен 0, то уравнение регрессии неудачно для предсказания значений y.

Таким образом, гладя на наши модели и их коэффициент детерминированности, выбираем ту модель, у которой коэффициент наиболее приближен к 1. У нас – это полиномиальная модель 2-ой степени. Предлагаю сравнить с полиномиальной 5-ой степени (сравнить их коэффициенты).

На этом 1-ый урок закончен. Домашнее задание состоит из §36 и 37 и вопросов к параграфам.

Следующий урок начинается с проверки домашнего задания в виде теста (Тест_1). Далее переходим к выполнению практической работы. В Практикуме для 10-11 классов это Работа 3.16. «Получение регрессионных моделей в MS Excel». Кратко объясняю этапы выполнения работы, показываю как добавлять линии тренда. Обращаю внимание, что выполнение работы на оценку в 5 баллов подразумевает выполнение следующих моментов:

1. Правильно составленная таблица
2. Наличие точечной диаграммы. Для диаграммы обязательно: название, подписи осей (без легенды).
3. Наличие 3-х регрессионных моделей с теми же требованиями, что и к выполнению диаграммы.

4. Вывод, где обязательно должна быть выбрана наилучшая регрессионная модель и аргументируем почему.

С этого урока я начинаю обучать детей правильно формулировать гипотезу и вывод. Это достаточно трудоемкий процесс. Гипотеза для данного задании уже прописана в §37 (Пример из области медицинской статистики, где предположили, что концентрация угарного газа в воздухе оказывает сильное влияние на бронхо-легочные заболевания). На последующих уроках в качестве разминки предлагаю выдвинуть свою гипотезу, которую теоретически можно проверить (максимально приближенную к реальной жизни).

Третий урок по теме: прогнозирование по регрессионной модели.

Здесь использую **Работу 3.17**. На примере задания 1 разбираем прогнозирование количественных характеристик системы по регрессионной модели **путем восстановления значений** (по формуле лучшей регрессионной модели) и **экстраполяции** (с помощью с использованием «Прогноз» на вкладке «Добавить линию тренда»).

Практическую работу выполняем по заданию для самостоятельного выполнения в этой же работе, но с небольшим добавлением. В работе предлагается рассчитать прогноз методом экстраполяции. Поясняю, что с экстраполяцией надо быть осторожным, так как применение модели ограничено, так как всякая экстраполяция держится на гипотезе, что выявленная закономерность, за пределами экспериментальной области сохраняется. А если нет? Особенно, если это касается сложных экологических или биологических систем.

Я добавляю эту работу прогнозированием способом восстановления значений. Для этого предлагается, используя данные таблицы, рассчитать прогноз средней температуры для еще 4-х городов, с указанием их широты.

Для этого учащимся необходимо построить несколько регрессионных моделей и выбрать из них наилучшую, и используя ее данные произвести прогноз.

Экстраполяция производится с использованием формулы лучшей регрессионной модели. Далее учащиеся. При помощи маркера заполнения копируют формулу в остальные ячейки. После получения результатов в ячейках, необходимо проанализировать эти данные, чтобы температурные

показатели были реальными. Если таковыми они не являются, необходимо искать ошибку.

Четвертый урок по теме посвящен корреляционным зависимостям.

Регрессионные модели строятся в тех случаях, когда известно, что эта зависимость между двумя факторами существует. А если мы не знаем есть ли зависимость? Например, взаимосвязь между ростом и весом детей, взаимосвязь между успеваемостью и результатами выполнения теста IQ, между стажем работы и производительностью труда?

Рассматриваем в §38 пример: зависит ли успеваемость учащихся от финансовых расходов на нужды школы.

Корреляция (от лат. correlatio), корреляционная зависимость — взаимозависимость двух или нескольких случайных величин. Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшению или увеличению) другой(-их) переменной(-ых).

При расчете корреляций пытаются определить, существует ли статистически достоверная связь между двумя или несколькими переменными в одной или нескольких выборках. Исследование таких зависимостей называется **корреляционным анализом** - метод обработки статистических данных, заключающийся в изучении коэффициентов (корреляции). Данный метод обработки статистических данных весьма популярен в социальных науках (в частности в психологии), хотя сфера применения **коэффициентов корреляции** обширна: контроль качества промышленной продукции, металловедение, агрохимия и проч.

Популярность метода обусловлена двумя моментами: коэффициенты корреляции относительно просты в подсчете, их применение не требует специальной математической подготовки. В сочетании с простотой интерпретации (принятие гипотезы о наличии корреляции означает что изменение переменной А, произойдет одновременно с изменением значения Б), простота применения коэффициента привела к его широкому распространению в сфере анализа статистических данных. *К недостаткам корреляционного анализа относится априорное предположение о линейной зависимости наблюдаемых переменных.*

Практическая часть урока состоит из выполнения **Работы 3.18** (Практикум), задание 2 (выполнить расчеты корреляционных зависимостей

успеваемости учащихся от обеспеченности учебниками и от обеспеченности компьютерами). Опять акцентирую внимание на уже готовой гипотезе, а вот вывод ученики пишут сами. *Но вначале учащиеся должны проанализировать полученные результаты.*

Коэффициент корреляции – число из диапазона -1 до +1. Если число по модулю близко к 1, то имеет место сильная корреляция, если к 0, то слабая. Близость к +1 означает, что возрастанию одного набора данных соответствует возрастание другого набора данных (прямо пропорциональная зависимость). Близость к -1 означает, что возрастанию одного набора данных соответствует убывание другого набора.

То есть, в выводе учащиеся должны подтвердить или опровергнуть предложенную гипотезу, используя данные коэффициентов корреляции (анализируют по модулю и по знаку). В помощь предлагается таблица «Интерпретация коэффициента корреляции». В нашем случае обеспеченность учебниками больше влияет на успеваемость, чем обеспеченность компьютерами.

Контрольная работа.

Задание. Результаты.

Практическая работа

Расчет корреляционных зависимостей в MS Excel

Требуется выполнить расчеты корреляционной зависимости успеваемости учащихся от хозяйственных расходов школы.

1. Заполнить электронную таблицу следующими данными:

A	B	C
№ п/п	Затраты (руб./чел.)	Успеваемость (средний балл)
1	50	3,81
2	345	4,13
3	79	4,30
4	100	3,96
5	203	3,87
6	420	4,33
7	210	4
8	137	4,21
9	463	4,40
10	231	3,99
11	134	3,90
12	100	4,07
13	294	4,15
14	396	4,10
15	77	3,76
16	480	4,25
17	450	3,88
18	496	4,50
19	102	4,12
20	150	4,32

2. Построить точечную диаграмму зависимости величин.
3. Ответить на вопрос (письменно), можно ли на основании этой точечной диаграммы выдвинуть гипотезу о наличии линейной корреляции между величинами.
4. Выполнить статистическую функцию КОРРЕЛ, указав в диалоговом окне диапазоны значений: B2:B21 и C2:C21.
5. Написать вывод, указав значение коэффициента корреляции, и подтвердить или опровергнуть свою гипотезу.

Контрольная работа
«Обработка статистических данных с помощью электронной таблицы»

Таблица 1

Доля населения Казахстана, не доживающая до 60 лет, в 1998 и 1999 годах

Области	1998	1999	Области	1998	1999
Акмолинская	35,8	33,2	Кызылординская	30,8	28,2
Актюбинская	34,0	31,4	Костанайская	32,8	28,2
Алматинская	29,2	26,5	Мангистауская	34,0	31,4
Атырауская	34,1	31,5	Павлодарская	35,0	32,4
В-Казахстанская	36,8	34,2	С-Казахстанская	35,0	32,4
Жамбылская	30,5	27,9	Ю-Казахстанская	29,2	26,5
З-Казахстанская	33,7	31,1	Астана	31,1	28,4
Карагандинская	37,8	35,3	Казахстан	33,0	30,3

Таблица 2

Региональные различия доли населения, имеющего уровень потребления ниже прожиточного минимума, в Казахстане в 1998 и 1999 годах

Области	1998	1999	Области	1998	1999
Акмолинская	19	35	Кызылординская	45	22
Актюбинская	53	24	Костанайская	31	55
Алматинская	61	44	Мангистауская	18	38
Атырауская	60	50	Павлодарская	31	48
В-Казахстанская	27	17	С-Казахстанская	41	27
Жамбылская	49	46	Ю-Казахстанская	87	56
З-Казахстанская	26	29	Астана	19	15
Карагандинская	35	18	Казахстан	43	39

Таблица 3

Индекс бедности населения Казахстана в региональном разрезе в 1998 и 1999 годах

Области	1998	1999	Области	1998	1999
Акмолинская	24,2	27,6	Кызылординская	31,1	20,2
Актюбинская	36,4	22,5	Костанайская	25,9	36,5
Алматинская	39,9	29,8	Мангистауская	22,7	28,0
Атырауская	39,7	34,0	Павлодарская	26,4	33,1
В-Казахстанская	26,1	22,5	С-Казахстанская	30,7	23,9
Жамбылская	33,0	31,1	Ю-Казахстанская	55,6	36,2
З-Казахстанская	24,4	23,9	Астана	21,1	18,9
Карагандинская	29,1	23,3	Казахстан	31,0	28,1

Задание к контрольной работе

Цель работы: на основании имеющихся статистических данных определить наличие существенной зависимости между различными вариантами этих данных.

1. Используя данные, постройте с помощью MS Excel таблицу, содержащую информацию из всех трех таблиц с разбивкой по годам.
2. Построить **6 точечных диаграмм**, визуально отображающих табличные зависимости.

Примечание:

Зависимости, которые должны быть отражены (по годам):

- *зависимость индекса бедности от доли населения, не доживающих до 60-лет*
- *зависимость, не доживающих до 60 лет, от уровня жизни ниже прожиточного минимума*
- *зависимость индекса бедности от доли населения, имеющих уровень потребления ниже прожиточного минимума.*

3. Выдвинуть гипотезы, на основании точечных диаграмм, о наличии зависимости между величинами (или отсутствии).
4. Найти коэффициенты корреляции (3 шт.) и определить, оказывает ли один фактор влияние на другой. Выбрать наиболее существенный показатель и проанализировать его.
5. Выявив наиболее существенную зависимость, построить несколько регрессионных моделей к этой зависимости. Выбрать наиболее подходящую функцию.
6. Сделать вывод о проведенной работе и полученных результатах. Описать в выводе выявленную существенную зависимость, опираясь на полученные результаты (коэффициент корреляции и регрессионную модель).