Муниципальное бюджетное общеобразовательное учреждение
средняя общеобразовательная школа №20

Анализ статистики посещений сайта

Выполнил:
Щербаков Владислав Алексеевич
МБОУ СОШ №20, 11А класс

Руководитель:
Мандрыченко Олег Борисович
учитель математики МБОУ СОШ №20

Пенза 2023

Содержание:

1. Введение.

Одним из самых важных показателей для оценки успешности работы сайта является количество посещений. Многие сайты ведут статистику посещений для анализа поведения пользователей и изменения своего контента в соответствии с их пожеланиями. В частности, достаточно популярным является сервис Яндекс Метрика. Это инструмент веб-аналитики, который помогает получать наглядные отчеты, записи действий посетителей, отслеживать источники трафика и оценивать эффективность онлайн- и офлайн-рекламы.

Поэтому очень важно владеть навыками анализа статистических данных посещения поисковых страниц для построения прогноза количества последующих посещений и систематизации запросов, так как это наиболее востребованный сервис в интернете.

В данной работе для анализа были использованы данные, предоставленные сайтом «Яндекс» о количестве посещений за месяц в течение 2014 – 2021 годов. Для удобства вычисления были взяты данные за 90 месяцев.

Целью данной работы является исследование ежемесячной аудитории поисковой страницы сайта «Яндекс» для составления прогноза количества посетителей в перспективе методами теории вероятностей и математической статистики.

Задачи исследования:

- познакомиться с понятиями, определениями и математическим аппаратом теории вероятностей и математической статистики, требующимися для решения поставленной задачи;
- вычислить выборочные параметры (выборочные средние, выборочные дисперсии, средние квадратические отклонения, корреляционный момент, коэффициент корреляции), требующиеся для выполнения работы;
- построить корреляционную таблицу;
- построить требующиеся графики линейной и параболической регрессий Y на X;
- сделать прогноз посещений на 2022 год.

Объектом исследования являются статистические данные о количестве посещений поисковой страницы сайта «Яндекс» за 90 месяцев.

Субъектом исследования является анализ рассматриваемых значений методами теории вероятностей и математической статистики.

Гипотеза исследования основана на том, что анализ статистики посещений за предыдущее время позволяет построить достоверный прогноз количества таких посещений в будущем, что серьёзно влияет на работу сайта.

В работе применялись аналитический и сравнительный методы исследования.

Актуальность и практическая значимость работы заключается в том, зная статистику посещений сайта и обладая умением её анализировать, возможно корректировать работу так, чтобы количество посещений увеличивалось, что напрямую влияет на получение компанией дополнительного дохода.

2. Основная часть.

2.1. Понятия, определения и формулы.

Теория вероятностей - раздел математики, изучающий закономерности случайныхявлений: случайные события, случайные величины, их свойства и операции над ними.

Выборочная совокупность - часть объектов из генеральной совокупности,отобранных для изучения, с тем чтобы сделать заключение о всей генеральнойсовокупности.

Генеральная совокупность - совокупность всех объектов (единиц), относительно которых учёный намерен делать выводы при изучении конкретной проблемы.

Объемом называют число объектов этой совокупности.

Наблюдаемые значения Xi называются **вариантами**, а последовательность вариант в возрастающем прядке - **вариационным рядом.**

Случайной величиной - называется величина, которая может принимать различные(случайные) значения.

Математическое ожидание - число, вокруг которого сосредоточены значения случайной величины. Математическое ожидание случайной величины X обозначается M(X).

Дисперсия случайной величины́ - мера разброса данной случайной величины, т. е.её отклонения от математического ожидания. Обозначается D(X)

Среднее квадратическое отклонение - показатель рассеивания значений случайной величины относительно её математического ожидания.

Модой случайной дискретной величины называется значение случайной величины,которое имеет максимальную вероятность:

Медианой называется такое значение варьирующего признака, которое приходитсяна середину упорядоченного ряда:

Гистограммой частот называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длинной h, а высотыравны частоте ni.

Корреляция в математической статистике - это вероятностная (статистическая) зависимость между величинами, не имеющая, вообще говоря, строго функционального характера.

Корреляционным моментом двух случайных величин называется математическое ожидание произведения отклонений этих величин.

Коэффициент корреляции - числовая характеристика совместного распределения двух случайных величин, выражает их взаимосвязь. Коэффициент корреляции дает более точную информацию о характеристике и силе связи. Коэффициентом корреляции двух случайных величин называется отношение корреляционного момента к произведениюсреднеквадратических отклонений этих величин.

Функцией распределения называют функцию $F(X)$, определяющую вероятность того, что случайная величина X в результате испытания примет значение, меньшее x. $F(x) = P(X < x)$. Функцией распределения выборки является эмпирическая функция распределения.

Эмпирической функцией распределения называют функцию $F^*(X)$, определяющую для каждого значения x относительную частоту события $X < x$.

$F^*(X) = n_x/n$, где n_x – число вариантов, меньших x, n – объём выборки.

Нормальным называют распределение вероятностей случайной величины, плотность которого описывается функцией $f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-m)^2}{2\sigma^2}}$, а функцию распределения $F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} e^{\frac{-(X-m)^2}{2\sigma^2}} dx$.

2.2. Исходные числовые данные.

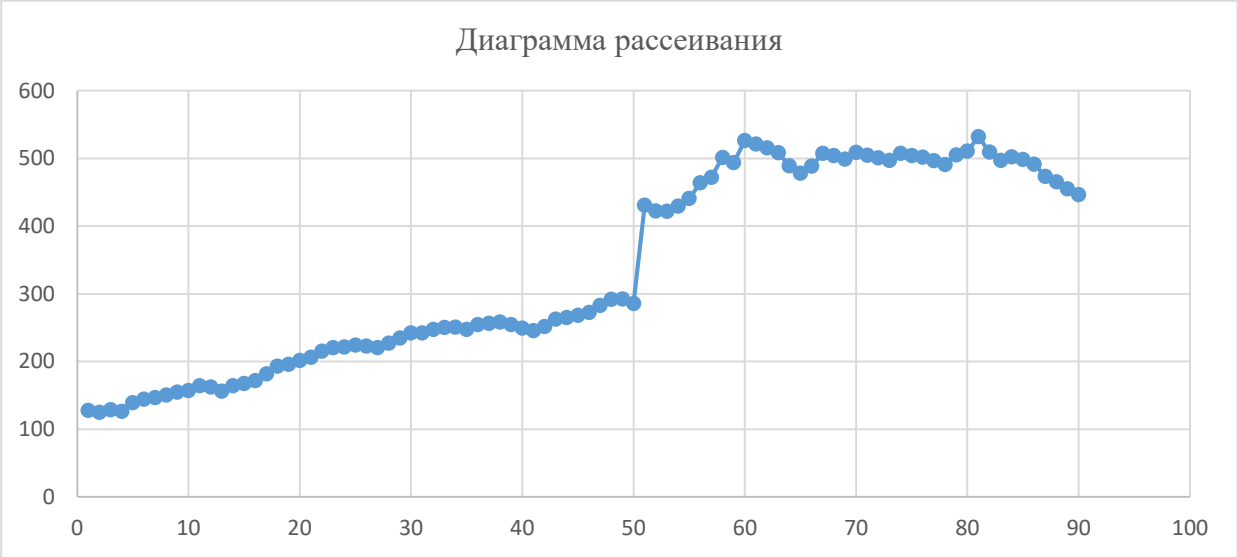
Представлена выборка (объёмом $\boldsymbol{n}$ = 90) зависимости числа Y от числа X.

X – номер месяца; Y – средняя месячная аудитория страницы поиска сайта «Яндекс». В силу большой величины рассматриваемых чисел будем вычисления для Y считать, как $Y\times10^5$.

X	$Y\times10^5$	X	$Y\times10^5$	X	$Y\times10^5$
1	127,9094	31	241,9705	61	521,4348
2	125,0087	32	247,4046	62	515,6451
3	128,5475	33	250,4053	63	508,2334
4	126,4261	34	250,5689	64	489,1397
5	138,9584	35	247,3736	65	477,9481
6	144,5391	36	254,373	66	488,4449
7	146,7449	37	256,3451	67	507,3722
8	150,6279	38	258,4041	68	504,0765
9	154,5571	39	254,308	69	498,7852
10	157,3101	40	249,3241	70	508,9668
11	164,3391	41	245,5744	71	504,5355
12	162,296	42	251,6614	72	501,0904
13	156,1086	43	262,6276	73	497,2581
14	164,1365	44	264,7736	74	507,7316
15	167,7096	45	268,0006	75	504,3579
16	171,7746	46	272,5063	76	501,7373
17	181,666	47	282,8596	77	496,5063
18	193,0372	48	291,6616	78	491,1007
19	195,9025	49	292,2627	79	505,2455
20	201,6512	50	285,9447	80	510,7384
21	206,3331	51	430,9605	81	532,2335
22	215,2265	52	422,2419	82	509,3164
23	220,2904	53	422,0372	83	496,9178
24	221,4521	54	429,2705	84	502,349
25	224,0201	55	440,7198	85	498,4698
26	222,8756	56	464,012	86	491,5623
27	220,6118	57	472,2111	87	473,4554
28	227,1231	58	501,3661	88	465,5044
29	234,8498	59	493,8464	89	455,1961
30	241,9705	60	526,2573	90	446,6343

Статистические данные взяты с сайта *https://stat.yandex,ru/Russia/Search/*

2.3. Диаграмма рассеивания и корреляционная таблица.



Найдем некоторые характеристика для *X* и *Y*.

Для *X*: выборочное среднее $M^*(x) = 45{,}5$; выборочная дисперсия $D^*(x) = 682{,}5$; исправленная дисперсия $S(x) = 690{,}17$; среднеквадратичное отклонение $\sigma^*(x) = 26{,}12$; оценка среднеквадратичного отклонения $s^*(x) = 26{,}27$.

Для *Y*: выборочное среднее $M^*(y) = 337{,}15$; выборочная дисперсия $D^*(y) = 20734{,}86$; исправленная дисперсия $S(y) = 20733{,}86$; среднеквадратичное отклонение $\sigma^*(y) = 144$; оценка среднеквадратичного отклонения $s^*(y) = 143{,}99$.

Выборочный коэффициент корреляции $r^*_{XY} = 0{,}93$.

Построим корреляционную таблицу. Для этого разобьем Y на 10 интервалов.

	1	2	3	4	5	6	7	8	9	10
Y_i	125 - 165	165 - 206	206 - 247	247 - 288	288 - 329	329 - 370	370 - 411	411 - 452	452 - 493	493 - 533
Y_i	145	185,5	226,5	267,5	308,5	349,5	390,5	431,5	472,5	513
n_i	14	6	11	16	2	0	0	7	10	24

По корреляционной таблице найдем оценки для *X* и *Y*.

Для X: выборочное среднее $M^*(x) = \frac{1}{n}\sum_{i=1}^{8} x_i n_i$; $M^*(x) = 45{,}5$;

выборочная дисперсия $D^*(x) = M^*(x^2) - \left(M^*(x)\right)^2$; $D^*(x) = 660{,}08$;

исправленная дисперсия $S(x) = \frac{n}{n-1} D^*(x)$; $S(x) = 667{,}5$;

среднеквадратичное отклонение $\sigma^*(x) = \sqrt{D^*(x)}$; $\sigma^*(x) = 25{,}69$;

оценка среднеквадратичного отклонения $s^*(x) = \sqrt{S(x)}$; $s^*(x) = 25{,}84$.

Для Y: : выборочное среднее $M^*(y) = \frac{1}{n}\sum_{i=1}^{8} y_i n_i$; $M^*(y) = 337{,}15$;

выборочная дисперсия $D^*(y) = M^*(y^2) - \left(M^*(y)\right)^2$; $D^*(y) = 20513{,}03$;

исправленная дисперсия $S(y) = \frac{n}{n-1} D^*(y)$; $S(y) = 20512{,}03$;

среднеквадратичное отклонение $\sigma^*(y) = \sqrt{D^*(y)}$; $\sigma^*(y) = 143{,}22$;

оценка среднеквадратичного отклонения $s^*(y) = \sqrt{S(y)}$; $s^*(y) = 143{,}22$.

Выборочный коэффициент корреляции $r^*_{XY} = \frac{\sum n_{xy} xy - nM^*(X)\cdot M^*(Y)}{n\sigma^*(X)\cdot\sigma^*(Y)}$

$r^*_{XY} = 0{,}95$.

Видим, что величины, вычисленные по корреляционной таблице мало отличаются от величин, вычисленных по всей выборке.

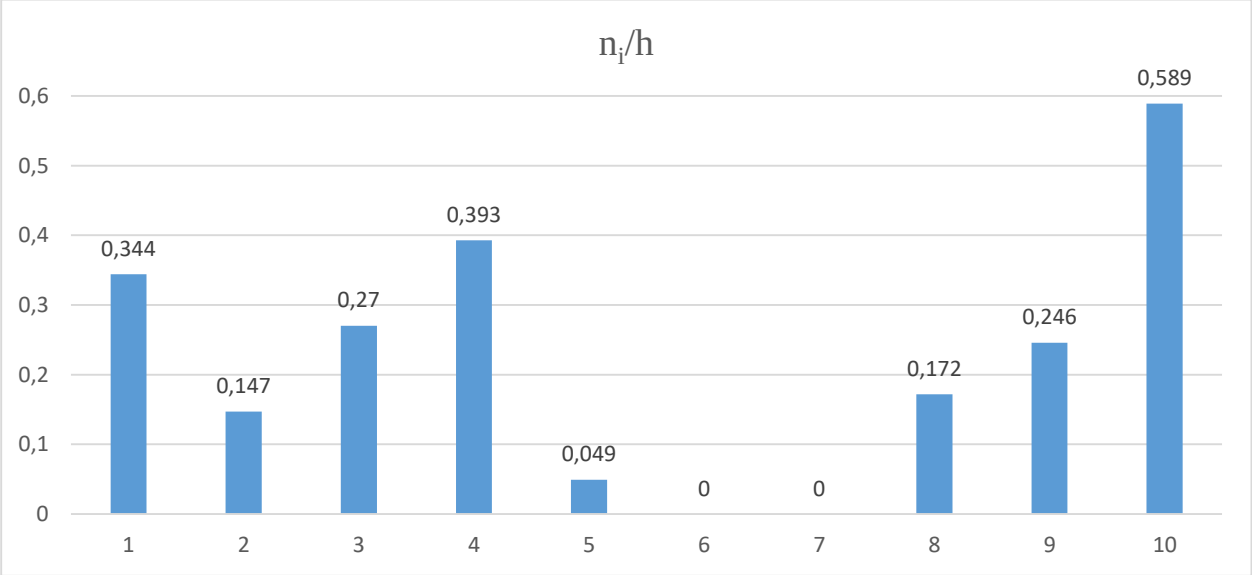
2.4. Графическое представление.

Для нахождения параметров случайной величины, таких как частота, вероятность, функция накопления, удобно использовать статистические таблицы, на основе которых будут рисоваться графики частот.

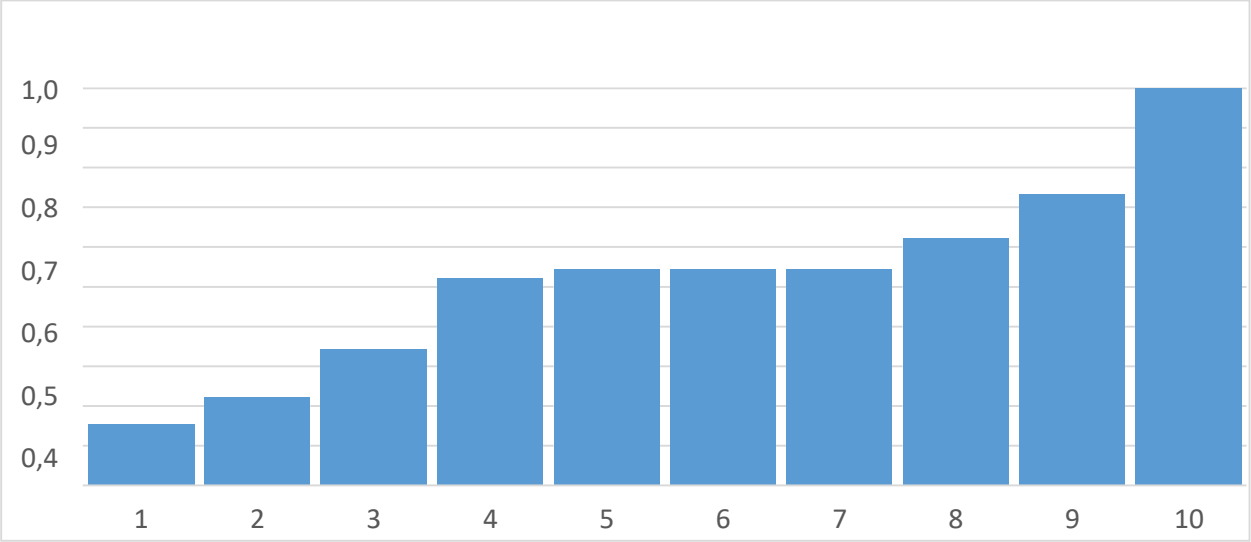
	1	2	3	4	5	6	7	8	9	10
Y_i	125 - 165	165- 206	206 - 247	247 - 288	288 - 329	329 - 370	370 - 411	411 - 452	452 - 493	493 - 533
Y_i	145	185,5	226,5	267,5	308,5	349,5	390,5	431,5	472,5	513
n_i	14	6	11	16	2	0	0	7	10	24
$P_i^* = n_i/n$	0,156	0,067	0,122	0,178	0,022	0,000	0,000	0,078	0,111	0,267
n_i/h	0,344	0,147	0,270	0,393	0,049	0,000	0,000	0,172	0,246	0,589
$n_i/(h*n)$	0,004	0,002	0,003	0,004	0,001	0,000	0,000	0,002	0,003	0,007

Построим полигоны частот и гистограммы для исследуемых признаков.

Гистограмма частот Y.



Эмпирическая функция по Y.



Изображать диаграммы и полигоны абсолютных и относительных частот, а также эмпирическую функцию по X не нужно, так как значения частот будут равны межу собой, и никакой зависимости наблюдаться не будет. Это следствие того, что мы берем значение даты для X и месячной аудитории для Y.

2.5. Линейная регрессия.

Регрессия – это зависимость среднего значения какой-либо величины *Y* от другой величины *X*. Понятие регрессии в некотором смысле обобщает понятие функциональной зависимости *y = f(x)*. Только в случае регрессии одному и тому же значению *x* в различныхслучаях соответствуют различные значения *y*.

По форме зависимости различают два вида регрессий: линейную регрессию, которая выражается уравнением прямой $Y = aX + b$ и параболическую (нелинейную): $Y = pX^2 + qX + r$.

Метод наименьших квадратов - один из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.Метод основан на минимизации суммы квадратов остатков регрессии.

Если статистические оценки наблюдений независимы и подчиняются нормальному распределению, то МНК дает оценки неизвестных с наименьшей средней квадратичной ошибкой. В этом смысле МНК является самым лучшим среди других способов, позволяющих находить линейные несмещенные оценки.

Задача статистического и корреляционного анализа состоит в нахождении параметров, от которых зависит вид функции. Для этого и используется метод наименьших квадратов.

Рассмотрим случайную двумерную величину (*X, Y*), где *X* и *Y* - зависимые случайные величины. Представим одну из величин как функцию другой. Ограничимся приближенным представлением величины Y в виде линейной функции величины X: $Y \cong g(x) = aX + b$, где *a* и *b* - параметры, подлежащие определению.

Функцию g(x) называют среднеквадратической регрессией Y на X:

$F(a,b) = \sum_{1}^{100} \left(y_i - (ax_i + b) \right)^2$, где F – суммарное квадратичное отклонение.

Подберем *a* и *b* так, чтобы сумма квадратов отклонений была минимальна. Для того,чтобы найти коэффициенты *a* и *b,* при которых *F* достигает минимального значения,необходимо приравнять частные производные к нулю:

$$\begin{cases} -2\sum_{i=1}^{100} \left(y_i - (ax_i + b) \right) x_i = 0 \\ -2\sum_{i=1}^{100} \left(y_i - (ax_i + b) \right) = 0 \end{cases}.$$

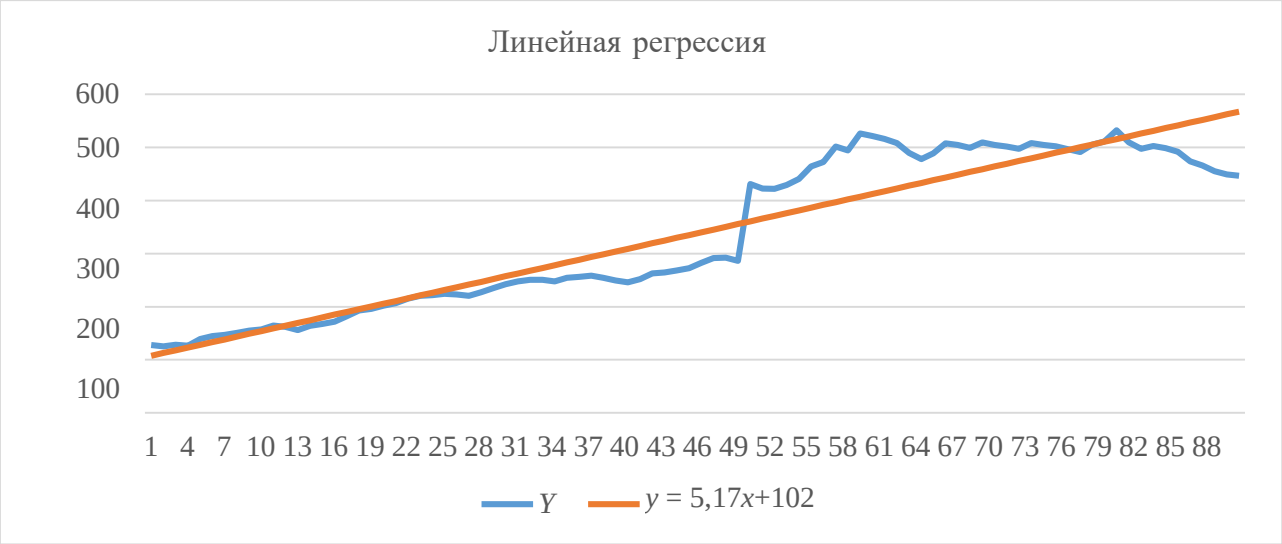
Выполнив требуемые преобразования, получим систему двух линейных уравнений относительно *a* и *b*:

$$\begin{cases} \left(\sum_{i=1}^{100} X_i^2 \right) \cdot a + \left(\sum_{i=1}^{100} X_i \right) \cdot b = \sum_{i=1}^{100} X_i Y_i \\ \left(\sum_{i=1}^{100} X_i \right) \cdot a + n \cdot b = \sum_{i=1}^{100} Y_i \end{cases}; \qquad a = \sum_{i=1}^{100} X_i^2, \qquad b = \sum_{i=1}^{100} X_i.$$

Получим $a = 5{,}17$; $b = 102$.

Следовательно, уравнение линейной регрессии имеет вид: $y = 5{,}17x + 102$.

Построим её график:



2.6. Нелинейная регрессия.

По данным наблюдений найдём выборочное уравнение кривой линии среднеквадратичной регрессии: $Y = pX^2 + qX + r$, где p, q, и r - параметры, подлежащие определению.

Применим метод наименьших квадратов. Подберем параметры p, q и r так, чтобы сумма квадратов отклонений была минимальной. Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих параметров:

$$F(p,q,r) = \sum_{i=1}^{100} \left(Y_i - pX_i^2 - qX_i - r\right)^2 .$$

Для отыскания минимума приравняем к нулю соответствующие частные производные:

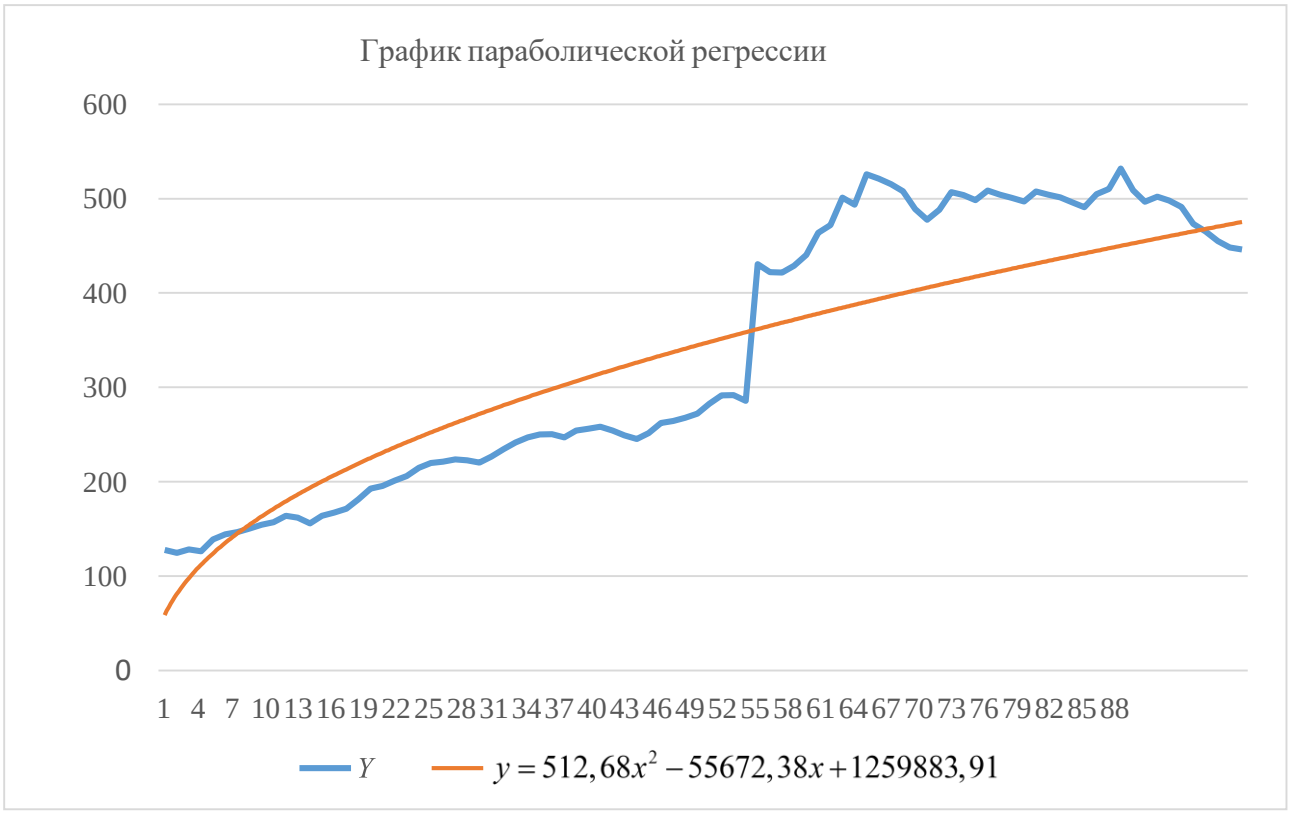
$$\begin{cases} F_p' = -2\sum_{i=1}^{100} X_i^2 (Y_i - pX_i^2 - qX_i - r) = 0 \\ F_q' = -2\sum_{i=1}^{100} X_i^2 (Y_i - pX_i^2 - qX_i - r) = 0 \\ F_r' = -2\sum_{i=1}^{100} X_i^2 (Y_i - pX_i^2 - qX_i - r) = 0 \end{cases}.$$

Выполнив преобразования, получим систему линейных уравнений относительно p, q и r:

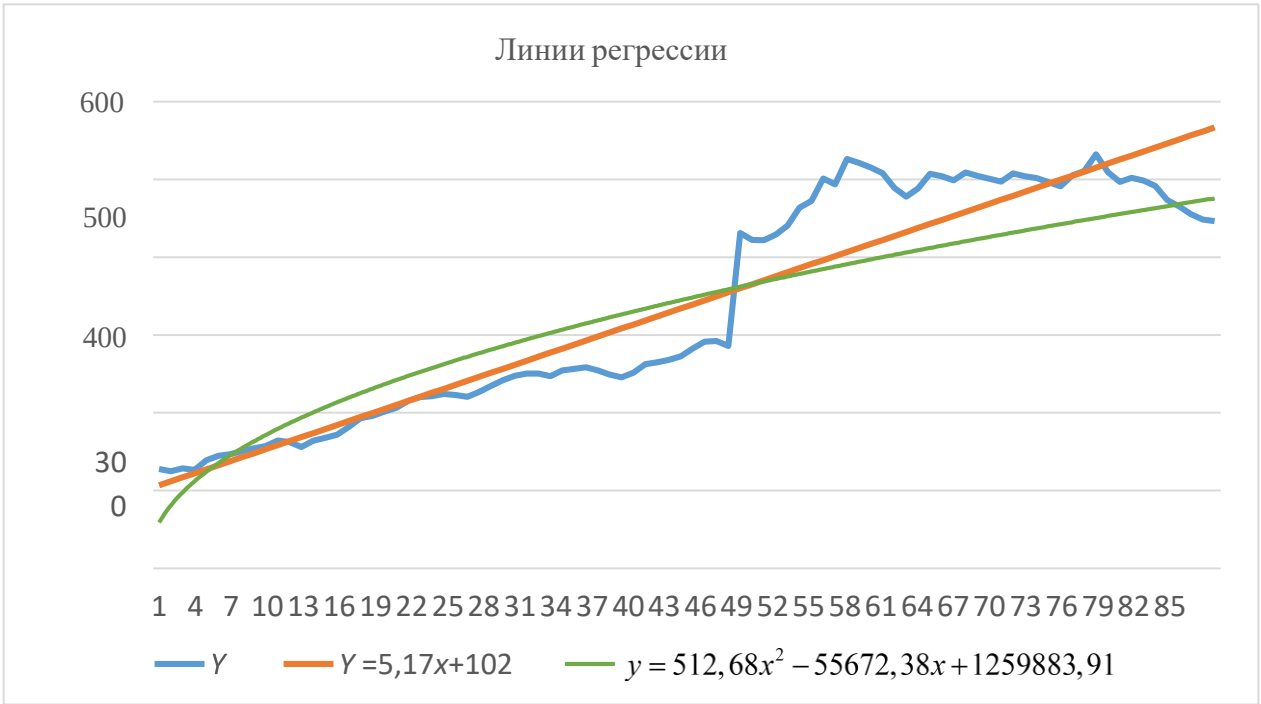
$$\begin{cases} \left(\sum_{i=1}^{100} X_i^4\right) \cdot p + \left(\sum_{i=1}^{100} X_i^3\right) \cdot q + \left(\sum_{i=1}^{100} X_i^2\right) \cdot r = \sum_{i=1}^{100} X_i^2 Y_i \\ \left(\sum_{i=1}^{100} X_i^3\right) \cdot p + \left(\sum_{i=1}^{100} X_i^2\right) \cdot q + \left(\sum_{i=1}^{100} X_i\right) \cdot r = \sum_{i=1}^{100} X_i Y_i \\ \left(\sum_{i=1}^{100} X_i^2\right) \cdot p + \left(\sum_{i=1}^{100} X_i\right) \cdot q + n \cdot r = \sum_{i=1}^{100} Y_i \end{cases}.$$

Решая эту систему, получим: p = 512,68; q = -55672,38; r =1259883,91. Следовательно, уравнение параболической регрессии имеет вид: $y = 512,68x^2 - 55672,38x + 1259883,91$.

Построим её график:



Теперь изобразим линии параболической и линейной регрессии на одном графике.

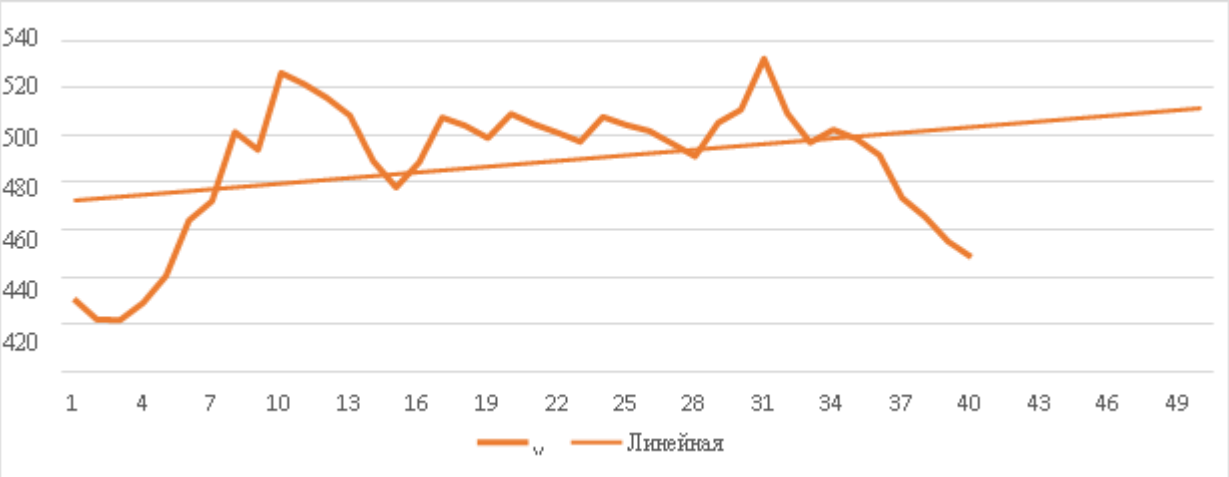


2.7. Статистический прогноз.

Для составления прогноза используем последнюю часть данных (от 50 до 90) и продлим график, построив линейную регрессию для этого интервала. Также, можно взять часть графика на определенном интервале и построить линию тренда. С построением линии тренда автоматически справляется программа Excel.

X	$Y\times10^5$	X	$Y\times10^5$
1	430,9605	21	504,5355
2	422,2419	22	501,0904
3	422,0372	23	497,2581
4	429,2705	24	507,7316
5	440,7198	25	504,3579
6	464,012	26	501,7373
7	472,2111	27	496,5063
8	501,3661	28	491,1007
9	493,8464	29	505,2455
10	526,2573	30	510,7384
11	521,4348	31	532,2335
12	515,6451	32	509,3164
13	508,2334	33	496,9178
14	489,1397	34	502,349
15	477,9481	35	498,4698
16	488,4449	36	491,5623
17	507,3722	37	473,4554
18	504,0765	38	465,5044
19	498,7852	39	455,1961
20	508,9668	40	448,6426

Построим график и линию тренда.



Можно дать прогноз посещений на 2022 год. Количество посетителей сайта «Яндекс» на середину 2022 года (соответствует номеру 49 по оси X) примерно равно 514×10^5, что является хорошим результатом с экономической точки зрения.

3. Заключение.

Статистические данные интересны сами по себе. Однако умение анализировать представленные цифры, делать выводы и прогнозы поднимает простые статистические данные на совсем другой уровень использования.

Мы увидели, что при наличии соответствующих статистических данных можно сделать прогноз результатов на следующий период, что подтверждает гипотезу данной работы.

Но кроме прогнозов, при внимательном изучении статистики, можно сделать и очень серьёзные выводы, связанные с экономическим развитием предприятия.

4. Литература и источники.

1. Кремер Н.Ш. Теория вероятностей и математическая статистика. Учебник для вузов. — М.: ЮНИТИ-ДАНА, 2002.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. — М., «Высшая школа», 2001

3. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и её инженерные приложения. – М. : Наука, 1988.

4. Родионов М.А., Яремко Н.Н. Краткий курс комбинаторики, теории вероятностей и математической статистики.: учебное пособие для старшеклассников и студентов. – Пенза : ПГПУ им. В.Г. Белинского, 2007.

Интернет-ресурсы:

https://stat.yandex.ru/Russia/Search

Рецензия.

Содержание работы ученика 11А класса МБОУ СОШ №20 Щербакова Владислава полностью соответствует избранной теме: «Анализ статистики посещений сайта». В работе тема раскрыта полностью, проведена большая исследовательская работа по изучению и применению способов и методов статистического анализа с использованием построения графиков линейной и нелинейной регрессии, видна хорошая математическая подготовка учащегося.

Грамотно обоснованы причины и актуальность выбора темы. В частности, обращено внимание на экономическую составляющую данного исследования.

В работе рассмотрена и проанализирована достаточно большая выборка, представляющая собой количество посещений сайта «Яндекс» за определённое время. Уточнено, что в подобных исследованиях заинтересовано большинство сайтов и отмечено, что многие пользуются сервисом Яндекс Метрика.

Требуется отметить, что для решения поставленной задачи применялся блок знаний, выходящих за рамки школьного курса математики, из чего следует творческая направленность данного исследования.

При проведении работы применялись аналитический и сравнительный методы исследования поставленной задачи.

Признавая необходимость практической значимости задач, решаемых с использованием методов статистического анализа, можно сделать вывод об актуальности данной работы.

Рецензент: Мандрыченко Олег Борисович
Место работы: МБОУ СОШ №20
Должность: Учитель математики, высшая квалификационная категория
Научная степень: ----------
Домашний адрес: г.Пенза, ул. Калинина, д.9, кв. 78
Телефон: 8 927 364 19 84

Подпись: _____ /О.Б. Мандрыченко/

Директор МБОУ СОШ №20 _____ /И.А. Николаева/

М. П.