**АНАЛИЗ ДАННЫХ**

**в информационных системах**

**(АДвМВ)**

* Математическая статистика.
* Прикладная статистика.
* Планирование вычислительного эксперимента.
* Эконометрика (эконометрия).
* Биометрика или биологическая статистика (Биостатистика).
* Статистика обработки социологических, психологических, геологических данных и др.

***ЛИТЕРАТУРА и интернет источники:***

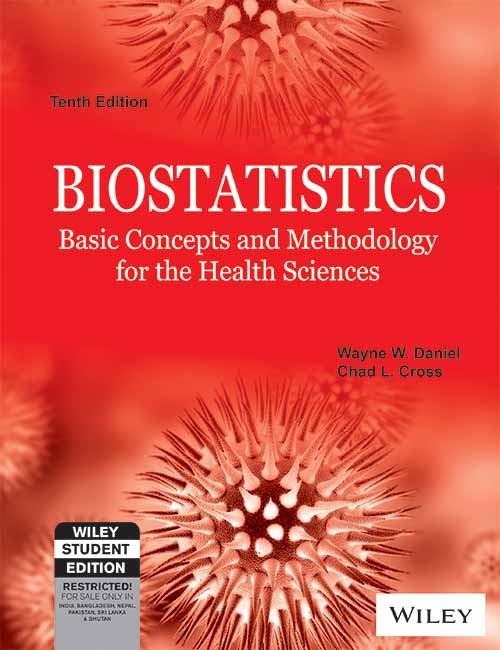
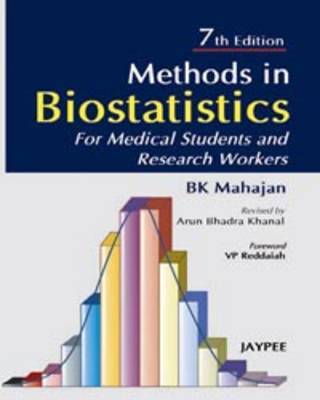
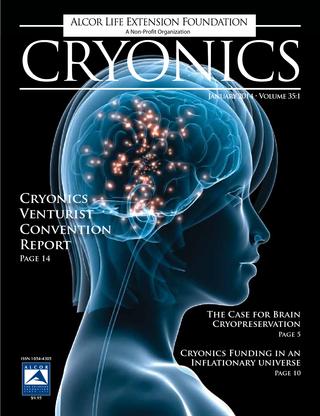
1. Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие.
2. Data sciance. Наука о данных с нуля: пер. с англ. – СПб: БХВ-Петербург. – 2020. -336.с.
3. Орлов А.И. Прикладная статистика. <https://orlov.pp.ru>.
4. Эконометрика/ Под ред. Елисеевой И.И.
5. [Дьяконов В. П.](http://www.biblioclub.ru/author.php?action=book&auth_id=4674) [MATLAB 6.5 SP1/7 + Simulink 5/6 в математике и моделировании](http://www.biblioclub.ru/book/117696/), 2009.
6. Колпак Е.П. Вычисления в MATLAB: учеб. пособие, 2016.
7. Мещеряков В.В. Задачи по статистике и регрессионному анализу с MATLAB, 2009.
8. <https://elibrary.ru> – научная электронная библиотека, российский информационно-аналитический портал в области науки, технологии, медицины и образования

.***ЖУРНАЛЫ:***

1. Биометрика. <http://www.biometrica.tomsk.ru/>
2. Biometrical journal.

<https://www.journals.elsevier.com/biomedical-journal>

1. Вопросы статистики. <https://voprstat.elpub.ru/jour>



**СТАТИСТИЧЕСКИЕ ПРИКЛАДНЫЕ ПАКЕТЫ, ПАКЕТЫ ПРИКЛАДНЫХ ПРОГРАММ (ППП)**

**для анализа статистических данных**:

***MS Excel, MATLAB, MATCAD, MAIPL, STATISTICA, SPSS,* GPSS, *R, ...***

1. Реализация языка программирования Python CPython <https://docs.python.org/3/license.html>
2. <http://matlab.ru/products/matlab> сайт **MATLAB exponenta**
3. <https://www.youtube.com/user/MATLABinRussia> – MathWorks официальный сайт Youtube на русском языке

***Статистические сборники:***

1. Российский статистический ежегодник   
   (Госкомстат России).
2. Регионы России (Госкомстат России).
3. Россия в цифрах. 2020: Кратк.стат.сб./Росстат- М., 2020 - 550 с. <https://istina.msu.ru/collections/332702068/>
4. Тюменская область в цифрах.

***СТАТИСТИЧЕСКИЕ МЕТОДЫ***

***ОБРАБОТКИ ДАННЫХ.***

***Исторические сведения***

♦ Пример раннего применения статистических методов – ***«Книга чисел» (Ветхий Завет).***

Статистические данные собирали в Древнем Египте, Греции, Риме: численность населения, демографические показатели (продолжительность жизни, сравнение числа новорожденных девочек и мальчиков, среднее число детей в семье), количество податей, земель, скота и т.п.

***Известны факты статистического учета:***

* 27– 14 г. до н.э.
* 5 тыс. лет до н.э. Китай
* 527 – 533 г. н.э.
* 1066 г. Англия
* 1245 г. Киевская Русь.

♦ Наука Статистика возникла в 17 веке. Термин «***Статистика***» впервые употребил в 18 веке ученый Готфрид Ахенваль.

«***Статистика»*** имеет латинский корень ***station*** – государство или ***status –*** состояние.

В средние века означало политическое состояние государства (сумма знаний о государстве, необходимая купцам, военным, политикам и ученым).

♦ ***Status*** (Шекспир, «Гамлет») – политическое состояние.

♦ ***2-я половина XVII века***. Формирование статистической науки одновременно в Англии (как политическая арифметика) и в Европе (как государствоведение).

♦ ***XVII век***. Использование вероятностных моделей при обработке статистических данных по числу новорожденных в Париже. Расчеты для страхования жизни, торговли и др.

Задачи теории вероятностей рассмотрены в трудах: ***Б.*** ***Паскаль, П. Ферма, А. Муавра, П. Лапласа, С. Пуассона***.

В 1718 году ***А. Муавр*** выпустил книгу «Учение о случаях», в которой привел данные по росту 1375 случайно отобранных женщин.

Графическое отображение в виде ломаной «Рост-частота» со значениями абсцисс от 140 до 180 см выявило сходство с графиком функции нормального закона распределения со смещением вправо.

♦ ***1794* – *1795гг.*** Разработка ***метода наименьших квадратов*** (МНК) для расчета орбиты астероида Церера, ***К. Гаусс***.

Использование статистических походов в астрономии (средняя ошибка измерения).

♦ ***XIX век.*** Создаются органы государственной статистики в станах Западной Европы, Северной Америки, России (П.Л. Чебышев, А.А.Марков, А.М. Ляпунов, А.Н.Колмогоров).

В 1887 г. создается ***международный статистический институт***.   
Международные статистические конгрессы.

Исследование вопросов статистической устойчивости демографических показателей.

♦ Середина ***XIX века.*** Применение ***методов статистического контроля качества и сертификации продукции*** в России.

***Фрэнсис Гальтон (1822* – *1911).*** **Область интересов**: естествознание, наследственность и теория эволюции. В [1889 году](https://ru.wikipedia.org/wiki/1889_%D0%B3%D0%BE%D0%B4) Ф. Гамильтон ввел понятие ***biometry***; разработал основы [корреляционного анализа](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7).

***Карл Пирсон (1857* – *1936).*** В [1884 году](https://ru.wikipedia.org/wiki/1884_%D0%B3%D0%BE%D0%B4) ***К. Пирсон*** и зоолог [***Уолтер Уэлдон***](https://ru.wikipedia.org/w/index.php?title=%D0%A3%D1%8D%D0%BB%D0%B4%D0%BE%D0%BD,_%D0%A3%D0%BE%D0%BB%D1%82%D0%B5%D1%80&action=edit&redlink=1)разработалианализ зоологических данных.

В 1893 г. К. Пирсон ввёл понятие среднего квадратического отклонения и коэффициента вариации, разработал основы множественной и нелинейной регрессии.

***Рональд Фишер (1890* – *1962).*** В 1919 **–** 1933 годах ***Р. Фишер,*** работая статистиком на опытной сельскохозяйственной станции, разработал теорию выборочных распределений, методы [дисперсионного](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7) анализа, теории планирования экспериментов, основы прикладной статистики и математической генетики.

♦ ***1900 г.*** Основание журнала ***«Биометрика»,*** ***К.*** ***Пирсон***. Основная идея – анализ данных из семейств распределений случайных величин.

♦ ***В 20 веке*** развитие статистики на базе информационных технологий.

♦ ***1926 г.*** Введение термина «***Эконометрика***» для обозначения научного направления, как синтеза экономической теории, математики и статистики, ***Р. Фриш***.

Термин «***Эконометрика***» **–** комбинация греческих слов «экономика» и «метрика». Буквальный перевод – измерения в экономике.

♦ ***1928 г. Ч. Кобб и П. Дуглас*** на основании данных по обрабатывающей промышленности США в области сельского хозяйства за период **1899 – 1922** гг. получили первую эмпирическую производственную функцию: ***Р = 1,01М 0,75К0,25.***

Где ***Р*** – производственная функция,

***М*** – индекс занятости,

***К*** – индекс основного капитала.

♦ ***1931 г.*** Создание ***международного эконометрического общества***.

♦ С ***1981 г.*** используют термин «***Прикладная статистика».***

♦ ***В 1989 г***. организован ***Институт высоких статистических технологий и эконометрики*** на базе университета имени Н.Э. Баумана.

♦ ***В 1990 г***. ***Всесоюзная статистическая ассоциация***.

**Ученые – номинанты Нобелевской премии:**

***Я. Тильберген, Р. Фриш, Л. Клейн, Т. Хаавельдо, Д. Хекман, К. Риглз и др.***

**Периоды использования ППП**

**Период 1970 – 1980 гг. *Исследования в области компьютерного моделирования.***

***Т. Нейлор***, «Машинные имитационные эксперименты с моделями экономических систем».

**Разработка моделирующих систем. GPSS** **–** General Purpose Simulation System.

**Период 1980 – 1990 гг.** GASP-IV, SIMULA-67, GPSS-V, SLAM-II (фортран, алгол-60, ассемблер, PL/1).

Недостатки: сложность применения неподготовленным пользователем, требования высокого уровня математической подготовки.

**Период 1990 – 2000 гг.** Process Charter -1.0.2, **Ithink-3.0.61**, **Extend+BPR -3.1**, ReThink (США); Powersim-2.1 (Норвегия).

Недостатки: использование для дискретного или непрерывного моделирования, высокая стоимость компьютерной техники.

**Pilgrim** (Россия). Возможность пространственной, временной, финансовой динамики. Создание дискретно-непрерывных моделей. Высокое быстродействие.

**Пакет РДО – Ресурсы-Действия-Операции** (МГТУ им. Н.Э. Баумана), **Система СИМПАС – Система-Моделирования-на-ПАСкале** (МГТУ им. Н.Э. Баумана), **MATLAB (MATrixLABoratory**) (корпорация The Math Wors, Inc).

**Область применения**: решение задач линейной и векторной алгебры, численные методы, решение систем дифференциальных уравнений, имитационное моделирование, анализ данных, разработка приложений на языках С++, Fortran

♦ ***21 век***. Анализ данных (***Data sciance***), модели и методы искусственного интеллекта, машинное обучение.

**Анализ данных** (наука о данных, англ. ***data science, DS***) **–** объединение ряда научных областей (статистика, машинное обучение, большие данные, визуализация), занимающихся построением систем анализа и обработки данных.

Специалисты по анализу данных **–** ***data scientist***.

Аналитиков данных называют «самой востребованной профессией 21 века». Через лет 10 потребность в специалистах возрастет в миллионы раз.

МиС

АДД

ПО

АиП

Рисунок. ***Диаграмма Венна-Эйлера.***

Где МиС – математика и статистика,

АиП – алгоритмы и программирование,

ПО – профессиональный опыт,

АД – анализ данных

Есть шуточное определение таких специалистов:

# **“Data scientist – это такой человек, который умеет программировать лучше, чем статистик, и знает статистику лучше, чем программист”**

**Примеры применения анализа данных**

1. ***Ввод нововведения*** – переход на выпуск новой продукции с использованием новых технологий.

2. ***Прогнозирование поведения временного ряда*** ***(тренда)*** с помощью регрессионной модели.

3. ***Теория надежности***. Методы контроля качества продукции.

***ВОПРОСЫ ДЛЯ ПОВТОРЕНИЯ (ИЗУЧЕНИЯ):***

1. Решение и исследование СЛАУ.
2. Основы дифференциального исчисления функции нескольких переменных.
3. Понятие вероятности события. Классическое и статистическое определения.
4. Способы задания дискретных и непрерывных случайных величин (СВ).
5. Числовые характеристики СВ.
6. Законы распределения случайных величин.
7. Статистическая обработка данных.
8. Построение регрессионных моделей.
9. Основы корреляционного анализа.
10. Получение прогнозных оценок.