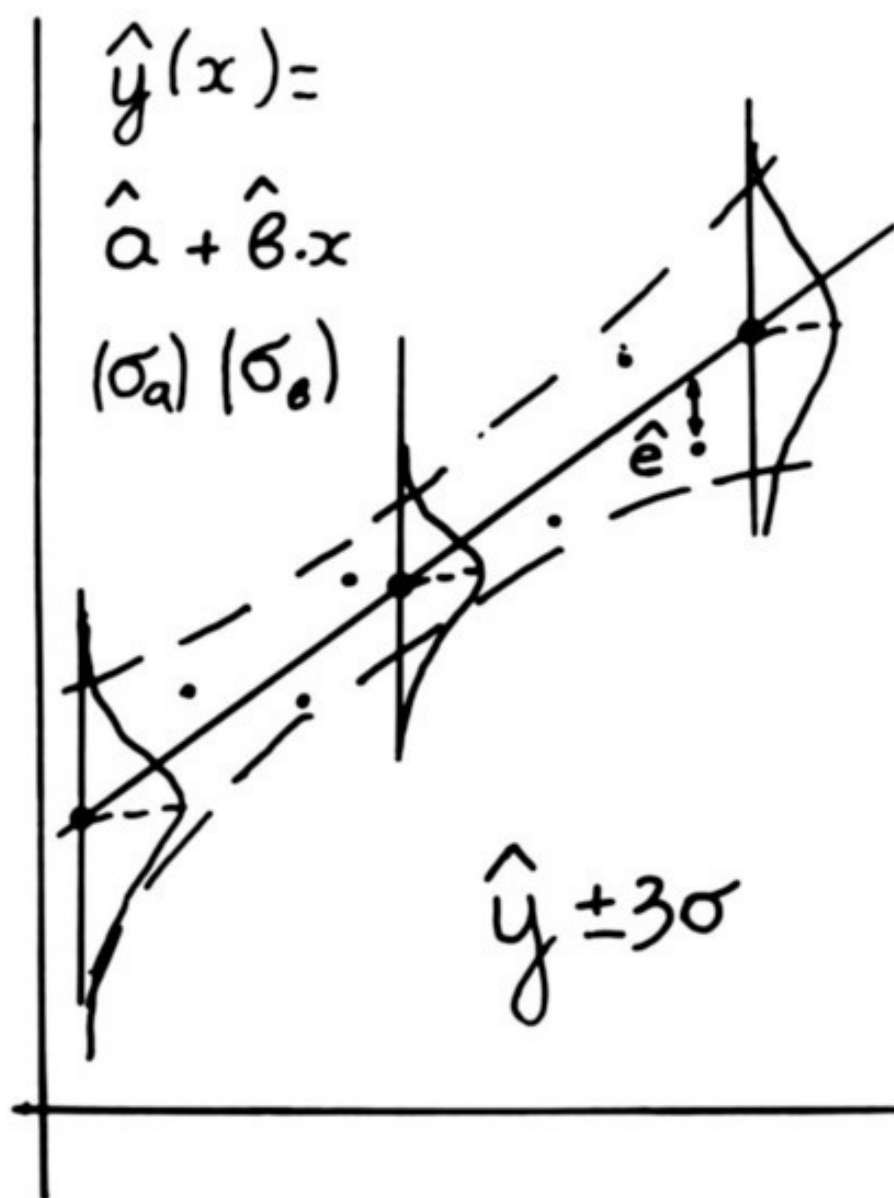


Валентин Юльевич Арьков

Эконометрические оценки

Учебное пособие



Валентин Арьков

**Эконометрические
оценки. Учебное пособие**

«Издательские решения»

Арьков В. Ю.

Эконометрические оценки. Учебное пособие / В. Ю. Арьков —
«Издательские решения»,

ISBN 978-5-00-553064-6

Каждое новое название грозит заказчику повышением стоимости услуг: статистика, математическая экономика, эконометрика, бизнес-аналитика, наука о данных, машинное обучение... Все перечисленные технологии используют метод наименьших квадратов (классический регрессионный анализ), который мы и будем рассматривать — в самых разных видах. Нас ожидает парная и множественная, линейная и нелинейная регрессия, разное количество входов и выходов модели, учёт качественных и количественных признаков.

ISBN 978-5-00-553064-6

© Арьков В. Ю.
© Издательские решения

Содержание

Предисловие	6
1. Введение	7
Регрессия	8
Электронная таблица	9
Демонстрация	10
Интерпретация	12
Предмет эконометрики	13
2. Инструменты	16
Демонстрация Jupyter LAb	22
Демонстрация Google Colab	23
3. Вероятность	25
Excel – надстройка	29
Excel – функция	32
Libre Office Calc – надстройка	35
Libre Office Calc – функция	38
Jupyter Lab	40
Google Colab	41
4. Распределение	43
Демонстрация. Excel	50
Конец ознакомительного фрагмента.	51

Эконометрические оценки

Учебное пособие

Валентин Юльевич Арьков

© Валентин Юльевич Арьков, 2021

ISBN 978-5-0055-3064-6

Создано в интеллектуальной издательской системе Ridero

Предисловие

Данное учебное пособие основано на серии лекций, прочитанных автором в дистанционном режиме. Видеозаписи лекций доступны на канале автора на Youtube. Ссылки на видеоролики имеются в конце данного пособия.

1. Введение

Наш предмет называется ЭКОНОМЕТРИКА. Существует эконометрика достаточно давно. На Западе – уже десятки лет преподаётся как самостоятельная дисциплина. В России эконометрика «появилась» лет двадцать-тридцать тому назад – как следование западной моде.

Смысл в том, что 30 лет назад это была какая-то абстрактная теория, непонятное отвлечение статистики. На сегодняшний день это работающие программы и технологии. Когда мы слышим про загадочные «большие данные» – Big Data, про анализ больших данных или про науку о данных – Data Science – за этим скрывается то, что как раз и изучается в курсе «Эконометрика».

Мы с вами будем разбирать очень практические вещи. Соответственно, и в лекциях, и на лабораторных работах у вас будет именно практическая сторона материала.

Регрессия

Весь предмет эконометрики сводится к задаче регрессии.

У нас имеется некоторое количество данных, которые условно можно назвать «иксы» и «игреки». Их можно изобразить в виде точек. Это могут быть люди, станки, квартиры, предприятия, страны. Имеем массив числовых значений и точек на графике. По этим точкам нужно В СРЕДНЕМ провести какую-нибудь линию, см. рис.

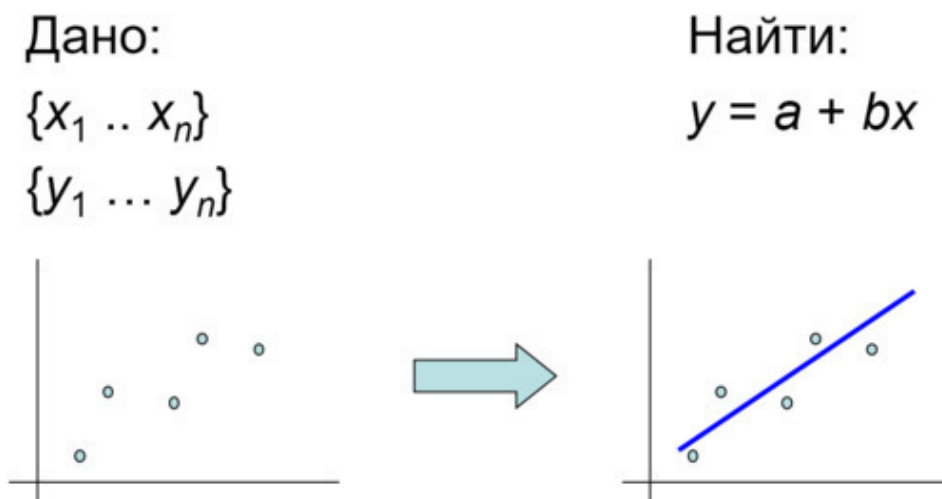


Рис. Регрессия – линия в среднем по точкам

Это может быть прямая, или кривая, или даже ломаная.

Если нанести точки на плоском графике и нужно построить прямую линию, то достаточно приложить линейку и провести линию. Главное, что это должно быть в среднем. Это значит, что линия может вообще не пройти ни по одной точке. Она пройдет там, где густо. Там, где много точек.

Такая технология называется РЕГРЕССИЯ.

На занятиях по бизнес-аналитике и статистике мы с разных сторон рассмотрим историю и разберём данный раздел.

Буквально слово «регрессия» означает «движение назад». В большинстве случаев это действительно возвращение или противоположность прогрессу и т. п. – кроме нашей ситуации. В эконометрике и в статистике слово «регрессия» означает «провести по точкам подходящую линию». По сути это попытка получить очень упрощенную закономерность из большого количества данных.

Вспомним школьный курс математики и геометрии. Прямую линию можно провести только через две точки. Это математика.

В статистике и в эконометрике мы говорим по-другому. Чтобы провести прямую линию, нужно 100 точек, а лучше 1000, а ещё лучше 1000000. Наша линия пройдет по ним в среднем. Эта идея в самых разных видах и будет рассматриваться.

Электронная таблица

Самый простой инструмент, с которым мы будем работать, – это табличный редактор, или электронная таблица. Это может быть Microsoft Excel или любой другой подобный программный продукт. Существует множество программ, которые совместимы с Excel на уровне формата файла. Есть бесплатный продукт – Libre Office Calc. Это могут быть облачные средства, например, Таблицы Гугл – Google Sheets.

На экране мы видим столбец иксов и столбец игреков. На графике нанесены эти точки. Можно себе представить, что это люди разного роста и разного веса. Каждого человека измерили и взвесили.

Теперь мы хотим получить среднюю, общую закономерность. Если мы выберем людей с очень красивой фигурой и с очень хорошим здоровьем, тогда можно будет говорить про «формулу идеального веса», см. рис.

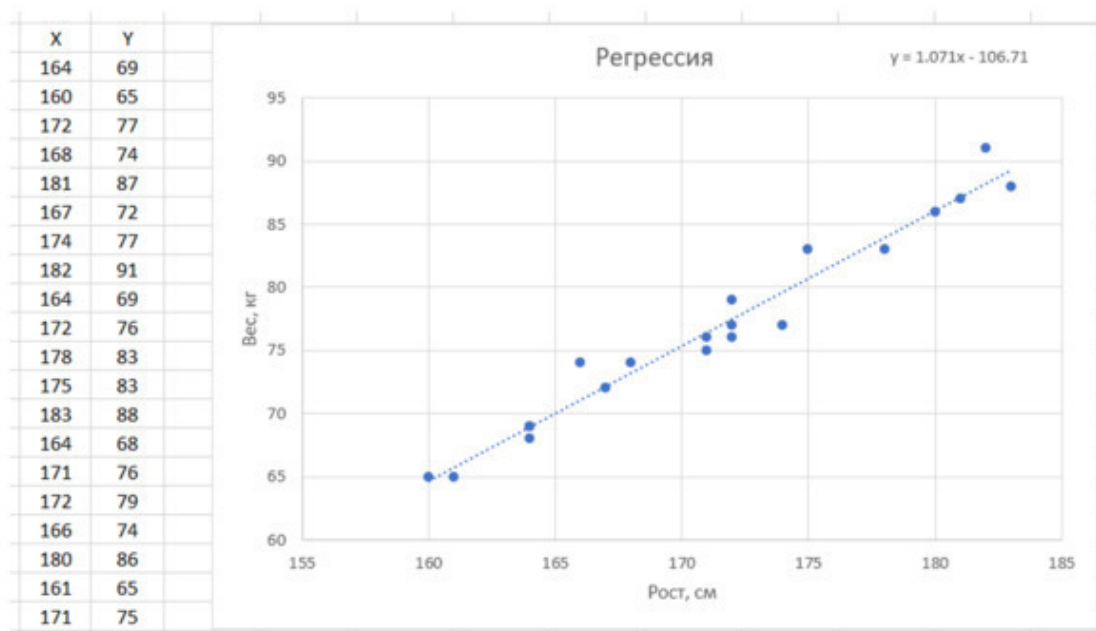


Рис. Линия регрессии в Excel

Предмет «Эконометрика» имеет некоторое отношение к экономике, к услугам, к товарам. Можно использовать знание таких закономерностей и для мониторинга здоровья сотрудников компании. Многие предприятия отправляют работников на регулярный медосмотр. При этом можно выяснить, у кого излишний вес, у кого нормальный, у кого недостаточный. Это может влиять на здоровье сотрудников, а значит, и на работу предприятия. Лучше потратить средства на поддержание здоровья, чем на лечение или срочную замену в самый неподходящий момент. Конечно, это будет важно для тех, кого интересует долгосрочное существование компании.

Другой пример. Пускай в качестве иксов у нас будет площадь квартиры, а по игреку – стоимость квартиры. Тогда мы можем рассматривать такую финансовую закономерность: сколько стоит квадратный метр в среднем и сколько может стоить конкретная квартира?

Есть множество таких ситуаций, где можно найти и использовать подобную зависимость. Хорошо, если есть всего один икс и один игрек. Это можно представить на плоскости. А если будет 20 иксов и 30 игреков? А если 120 иксов? Это на плоскости сложно изобразить – и от руки, и на компьютере.

Демонстрация

Разберем простую демонстрацию того, как выглядит регрессионный анализ.

Запускаем Excel.

Перед нами столбец иксов и столбец игреков.

Мы выделяем эти столбцы, выбираем **Вставка – График — ...** Находим в серединке кнопку «**Точечная диаграмма**». По-английски это называется **Scatter Plot**, по-русски – диаграмма разброса. У нас появилась картинка, см. рис.

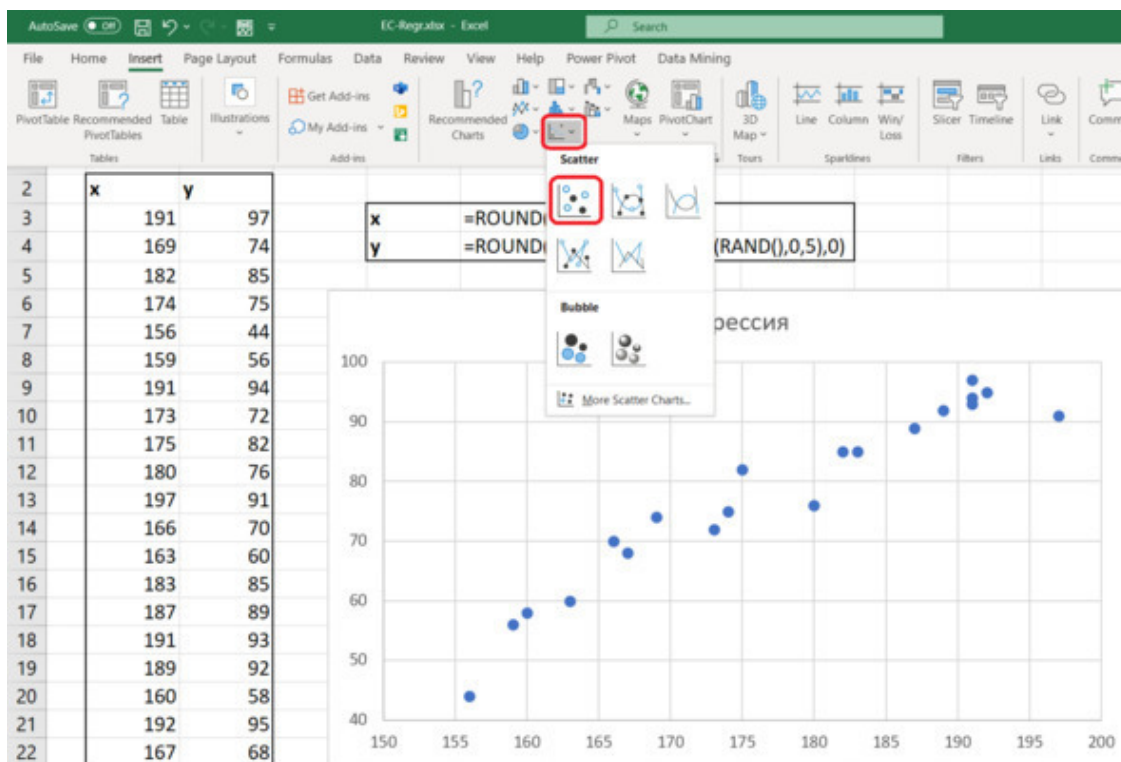


Рис. Исходные данные и диаграмма разброса

Наши игреки начинаются не от нуля, поэтому мы настроим ось координат. Задаём минимальное значение 40, максимальное 110.

Нажимаем плюсики справа от диаграммы и добавляем «линию тренда». Вообще-то слово «тренд» означает «тенденция изменений, общее направление развития». Но в данном случае это просто общая закономерность. Мы говорим, что нас интересует прямая линия. На графике появляется прямая.

Мы хотели бы вывести уравнение этой линии на экран. Нажимаем **More Options** – Дополнительные параметры. Затем **Display equation on chart** – Вывести уравнение на график. Получаем уравнение нашей линии, см. рис.

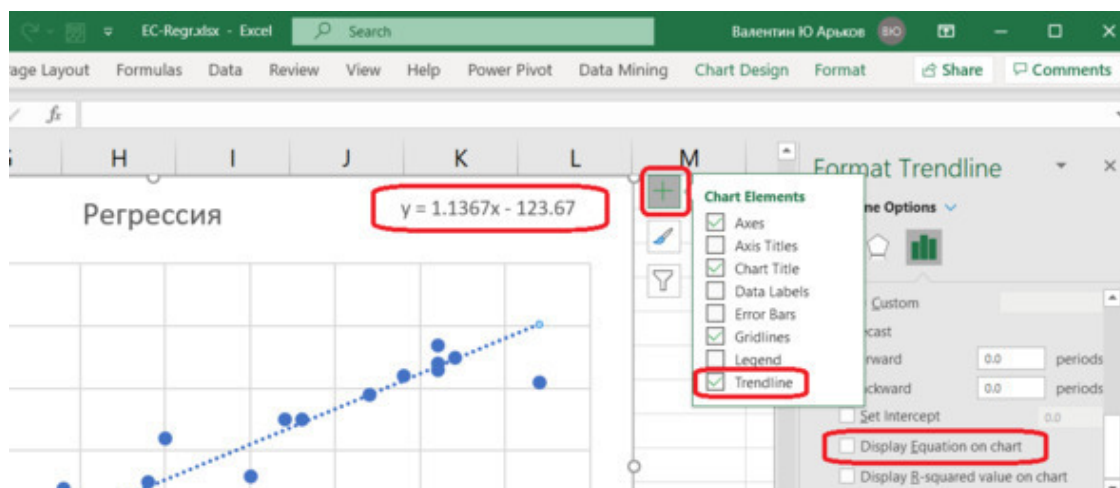


Рис. Уравнение и линия регрессии

Это самый простой вариант: как мы можем в среднем по точкам провести линию. Естественно, здесь есть масса нюансов, масса тонкостей. Это самая общая идея того, что нам предстоит освоить.

У нас были условные иксы – рост человека в сантиметрах – и условные игреки – вес в килограммах. Мы получили формулу для среднего веса: нужно взять икс с коэффициентом 1,14 и отнять от него число 123. Для именно этого набора данных это будет средняя закономерность.

Примерно таким образом получают полезные советы для отрывных календарей и для книг по фитнесу. Хотите узнать свой идеальный вес? Возьмите свой рост и отнимите 100.

Чтобы получить такую формулу, нужно собрать людей с правильной фигурой и с хорошим здоровьем. Вот тогда можно будет вывести уравнение.

Итак, мы рассмотрели общую идею: нужно взять большое количество данных и по ним найти общую закономерность. Всё это называется словом «регрессия», и это слово просто означает «линия в среднем по точкам».

Мы строим линию – она называется «линия регрессии». Мы получаем уравнение этой линии – оно называется «уравнение регрессии».

В качестве упражнения вам предстоит повторить то, что было в этой демонстрации. Чтобы сгенерировать исходные данные, используйте формулы, показанные на рис.

	A	B	C	D	E	F	G	H	I	J
1										
2		x	y							
3			191	97	x	=ROUND(150+50*RAND(),0)				
4			169	74	y	=ROUND(B3-100+NORMINV(RAND(),0,5),0)				
5			182	85						

Рис. Генерируем исходные данные

Вам нужно будет построить диаграмму разброса, потом добавить линию регрессии под названием «линия тренда». Затем нажать пару кнопочек, чтобы на экране появилось уравнение связи.

Интерпретация

Теперь по поводу полученного уравнения. Мы берем эту формулу и переводим ее на русский язык – выражаем её смысл словами. Чтобы узнать свой «идеальный» вес, возьмите рост и сделайте с ним то-то и то-то. Нужно умножить рост на что-то, а потом отнять что-то.

Это особое действие, которое выполняют в эконометрике. Это попытка перевести найденную закономерность на обычный разговорный язык. Называется он ИНТЕРПРЕТАЦИЯ уравнения регрессии. Слово «интерпретация» буквально означает «перевод с одного языка на другой». Здесь мы переводим с математического языка формул на человеческий, разговорный язык.

Предмет эконометрики

Наш предмет эконометрики находится на стыке нескольких дисциплин, нескольких областей деятельности. Можно обнаружить раздел «Эконометрика» в некоторых учебниках по экономической теории – обычно в самом конце. Кроме того, мы обнаруживаем очень похожий материал в курсе статистики. В математике тоже рассматриваются похожие задачи. Фактически, эконометрика находится на стыке трёх областей деятельности, трёх областей знания. На рисунке мы приводим такую схему:

- экономика;
- математика;
- статистика.



Рис. Место предмета эконометрики

На схеме мы показали другие пограничные области, которые находятся на стыке дисциплин.

Между экономикой и математикой находится математическая экономика – там изучаются математические модели экономических систем и явлений. Но они изучаются по большей части с точки зрения теории: как, в принципе, должна зависеть цена от спроса.

На стыке экономики и статистики находится так называемая экономическая статистика. Можно найти учебники, которые так и называются. Здесь обсуждается применение статистики для решения самых разных экономических задач.

Есть предмет статистики, который иногда называют «Общая теория статистики». Его тоже можно отдельно изучать.

Наконец, на границе между математикой и статистикой есть математическая статистика. Там будет много формул, много интегралов —очень тяжелый материал для тех, кто сдавал ЕГЭ.

Мы будем разбирать материал попроще и попонятней, потому что наша аудитория – специалисты по информатике. Такие работники используют готовые инструменты. Они внедряют готовые информационные системы (ИС) или, в лучшем случае, их настраивают.

Не все «информатики» занимаются разработкой и написанием программ. Чаще всего, речь идет о настройке, установке, обслуживании, сопровождении ИС. Это целая профессия.

В плане эконометрики у нас тоже есть готовые инструменты и требуется грамотно ими пользоваться. А если не нужно писать свою программу, которая будет строить линию по точкам, то работа с интегралами нам не особо потребуется. Но нам потребуется научиться грамотно нажимать некоторые кнопки и получать правильные графики. И, конечно, объяснять смысл полученных закономерностей.

Предмет эконометрики

- ЭКОНО+МЕТРИКА = экономика + измерение
- построение *количественных* моделей экономических явлений по фактическим данным
 - статистические модели экономики
- в отличие от экономической теории (модели на *качественном* уровне)
- на стыке трех дисциплин: экономика + статистика + математика
- используется для принятия управленческих решений

Рис. Предмет эконометрики

Слово «эконометрика» состоит из нескольких частей, см. рис.

Первая часть «эконо-» намекает на экономику. Напомним, что экономика – это производство, обмен и потребление товаров и услуг. Конечно, есть и наука, которая изучает эту сторону нашей жизни. Это самое простое определение, которое можно обнаружить в любом словаре или энциклопедии. Это наша жизнь. Всё, что мы делаем. 90 процентов нашей жизни связано или с производством, или с потреблением чего-нибудь, или с посещением мест, где это можно купить или продать. даже если мы лежим на диване и смотрим телевизор, мы что-то потребляем – товар (диван) и услугу (телевизионную передачу, «контент»). Есть и другие ситуации, где не так много экономики, а просто общественные отношения.

Вторая часть – «метр» – в данном случае означает «измерить, выразить числами, количественно». Это модели с какими-то числовыми коэффициентами. Модели, в которых есть конкретные числа. Модели строятся по реальным данным.

Эти модели нужны для того, чтобы выяснить, какие взаимосвязи существуют и почему происходит то или иное событие. Почему люди посещают магазин в определённое время? Почему люди переходят из одной сотовой компании в другую? В какой момент клиенты собираются «сбежать»? Это описание и анализ текущей ситуации.

Кроме анализа, существует задача прогнозирования. Что будет дальше, если верить существующим данным? Какие прогнозы по развитию ситуации? Исходя из прогнозов, можно давать рекомендации. Вот эти клиенты собираются перейти к другому оператору. Чтобы их удержать, обычно помогает вот такой приём, такое особое предложение – для такой категории клиентов.

Другой пример. Обычно в такое-то время такая категория покупателей приобретает такой набор товаров. Эти товары часто находятся в одном чеке. Поэтому давайте эти вроде бы разные товары на соседних полочках поставим.

Есть много задач, где обработка больших массивов данных помогает что-то обнаружить и что-то улучшить, причём с небольшими затратами.

В отличие от экономической теории, здесь рассматриваются конкретные данные, числа, факты. В экономической теории описывают общие закономерности, а здесь мы смотрим на реальные события, причём в большом количестве. Всё это нужно для того, чтобы чем-то управлять, чтобы что-то улучшать и корректировать. Эта деятельность называется управление, или принятие решений.

Окончание слова «-ика» означает «наука или дисциплина, изучение чего-либо».

Эконометрика занимается построением моделей. Мы рассмотрели пример такой модели в виде уравнения, но в среднем.

Есть большое количество книг с названием эконометрика. Каждый год появляются новые книги. Есть они на русском языке, есть и на других языках. На английском языке издаются в огромном количестве. Каждый университет с большим удовольствием издаёт свой вариант этого материала.

Задание

Предлагаем выполнить одно несложное задание, чтобы сориентироваться в общей картине. Посмотрите, какие существуют на сегодняшний день онлайн-курсы – платные и бесплатные – на тему «Эконометрика», «Большие данные», «Наука о данных», «Data Science», «Программирование на языке Python».

Есть платформы для массовых открытых онлайн-курсов – МООК. Английское название: Massive Open Online Courses (МООС). Есть и отдельные компании, которые на чём-то специализируются.

Посмотрите сколько предлагается таких курсов и сколько они стоят.

2. Инструменты

Мы переходим к следующей теме. Это инструменты анализа данных, см. рис.

Инструменты

- Электронная таблица
 - Excel, Libre Office Calc, Google Sheets...
- Язык программирования
 - Python, R, Matlab, C#...
- Варианты приложений
 - Настольная, облачная, мобильная

Рис. Инструменты анализа данных

Следует отметить несколько вариантов.

Первый способ обработки данных – простой наглядный и не всегда удобный. Это электронные таблицы. Они существуют в самых разных видах, но мы будем, в основном, ориентироваться на Excel.

Существуют разные системы, языки программирования или среды того или иного уровня для обработки данных. Очень часто используются Python или язык R. Есть более специализированные инструменты, например Matlab, но это уже коммерческий, дорогой продукт.

Естественно, есть обычные языки программирования для работы с данными. Мы увидим, что в системах типа Python несколько удобнее работать с данными.

Следующий момент – это варианты реализации программы, с которой мы работаем.

Чаще всего и привычней настольные, локальные варианты – Desktop Version.

При этом становится доступным всё больше облачных вариантов, когда мы можем работать через браузер. Здесь не всегда доступен полный функционал.

Наконец, есть варианты для мобильных устройств. Но чаще всего мобильные устройства – смартфоны и планшеты – не такие удобные, если требуется много печатать. Нужно подключать внешнюю клавиатуру. Кроме того, мобильные варианты приложений чаще всего ограничены по своим возможностям. Они подходят для определенных целей, например, чтобы посмотреть готовый график / отчёт или отсканировать QR-код.

Для дальнейшей работы нам понадобится так называемая надстройка «Анализ данных». Чтобы включить настройку, нужно перейти в меню Файл Настройки – Надстройки: **File – Options – Add-ins – Excel Add-ins – Go – Analysis ToolPak**, см. рис.

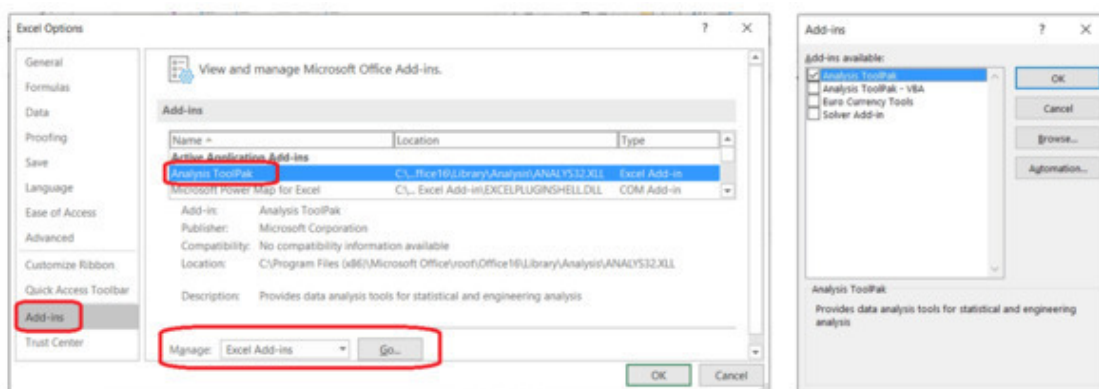


Рис. Включение надстройки

Теперь в верхнем меню появится надстройка – в разделе «Данные» – кнопка «Анализ данных», см. рис.

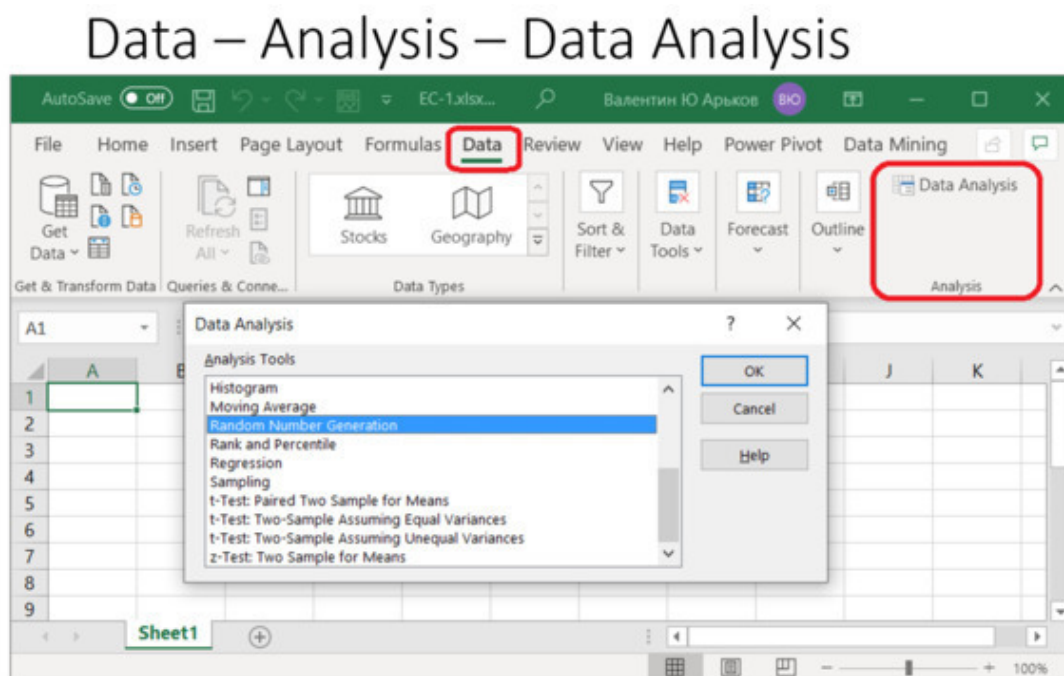


Рис. Вызов надстройки

Другой инструмент, который имеет очень похожие возможности – это Libre Office. Это бесплатный свободно распространяемый инструмент с открытым исходным кодом (Open Source). Как видим, у нас есть варианты для MS Windows, Linux и MacOS, см. рис.

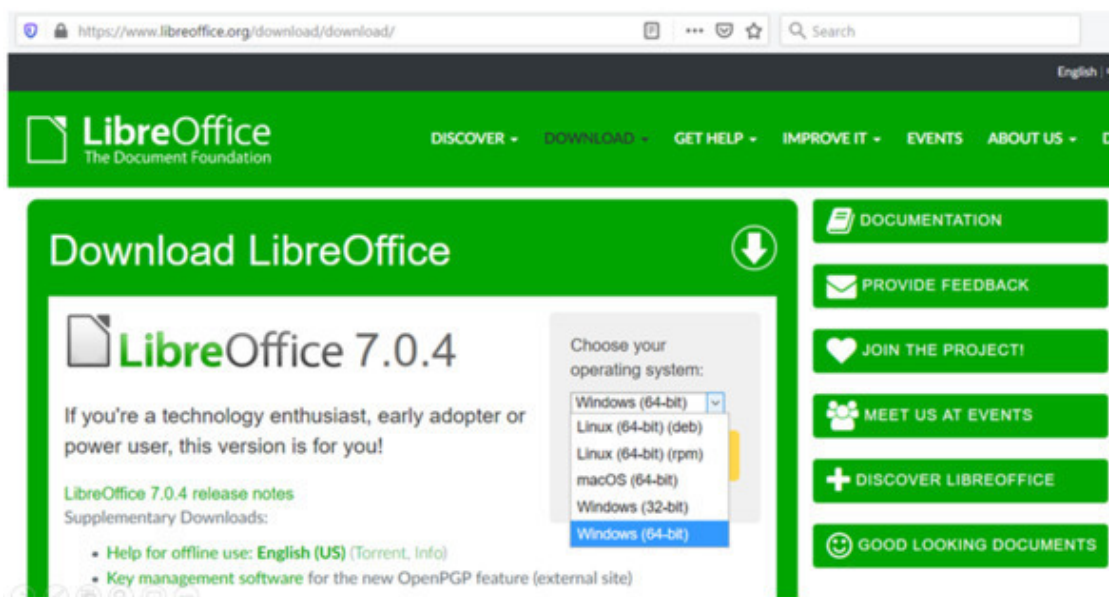


Рис. Варианты Libre Office

Мы рекомендуем использовать английский вариант программных пакетов. К сожалению, перевод интерфейса часто оставляет желать лучшего и выполняется по остаточному принципу. В ряде случаев можно обнаружить половину интерфейса на русском, а половину на английском. Разработчики не успевают перевести – слишком быстро выходит новая версия программы. Перевод интерфейса и справочной системы очень часто делается чуть не автоматическим, машинным, программным путём, что отрицательно отражается на качестве.

В ряде случаев неудачный перевод может сбивать с толку и запутывать. Мы обнаруживаем английские слова, написанные русскими буквами. Конечно, это не помогает в изучении программ. Гораздо проще запомнить несколько английских слов, но при этом желательно узнать, как эти понятия грамотно называются по-русски. А для этого нужно знать теорию и читать учебники.

По поводу Libre Office надо отметить наличие так называемой Portable Version, см. рис. Это «переносимая» версия программы, которая не требует установки. Мы просто скачиваем архивный файл, разворачиваем его в отдельном каталоге и оттуда его запускаем на выполнение. Если он нас больше не интересует, мы удаляем этот каталог – и он не засоряет операционную систему.

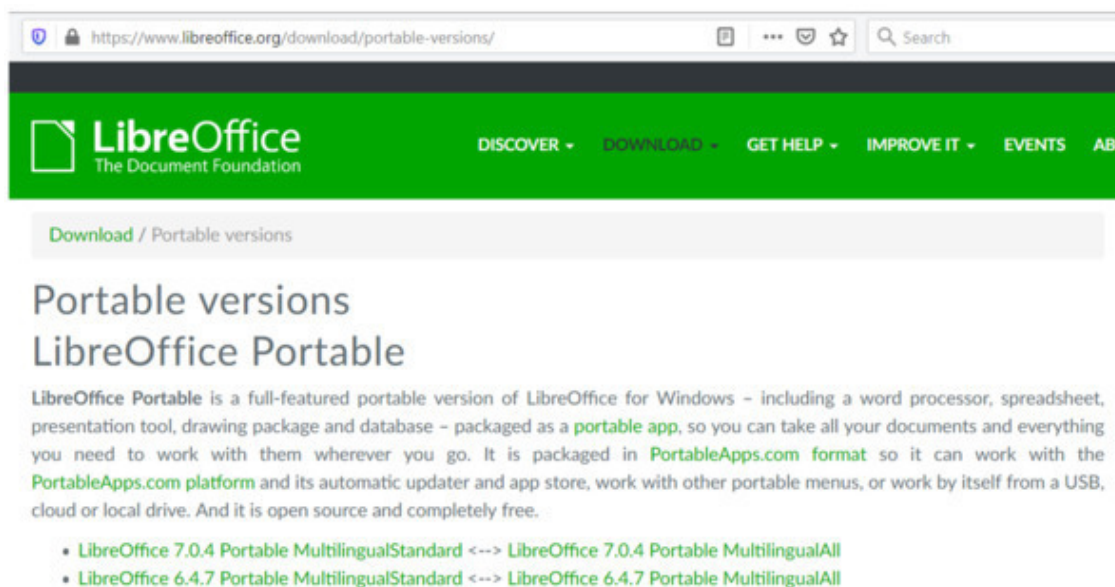


Рис. Переносимая версия

Электронная таблица Libre Office Calc очень похожа на MS Excel – внешне и по возможностям. Есть совместимость на уровне формата файлов *.XLSX. По сути, мы встречаем здесь стандартный вид электронной таблицы. Может немного отличаться верхнее меню, но разобраться совсем несложно, см. рис.

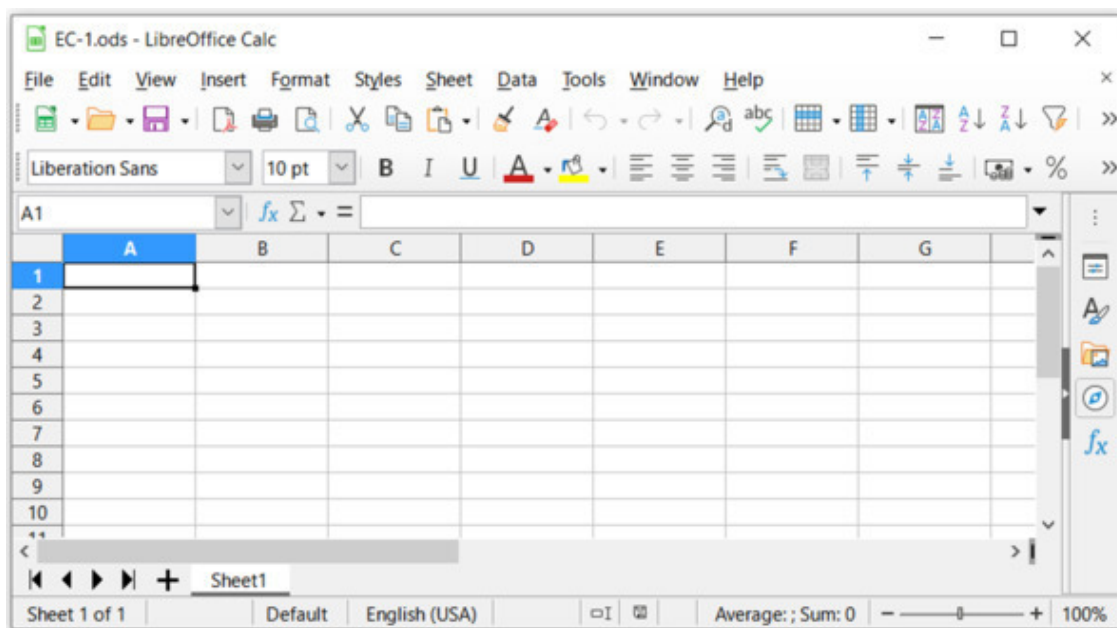


Рис. Интерфейс Libre Office Calc

Что касается Python, у нас есть тоже два варианта. Первый вариант – скачать какой-нибудь пакет, например, Анаконда. Анаконда – это целый набор инструментов, см. рис.

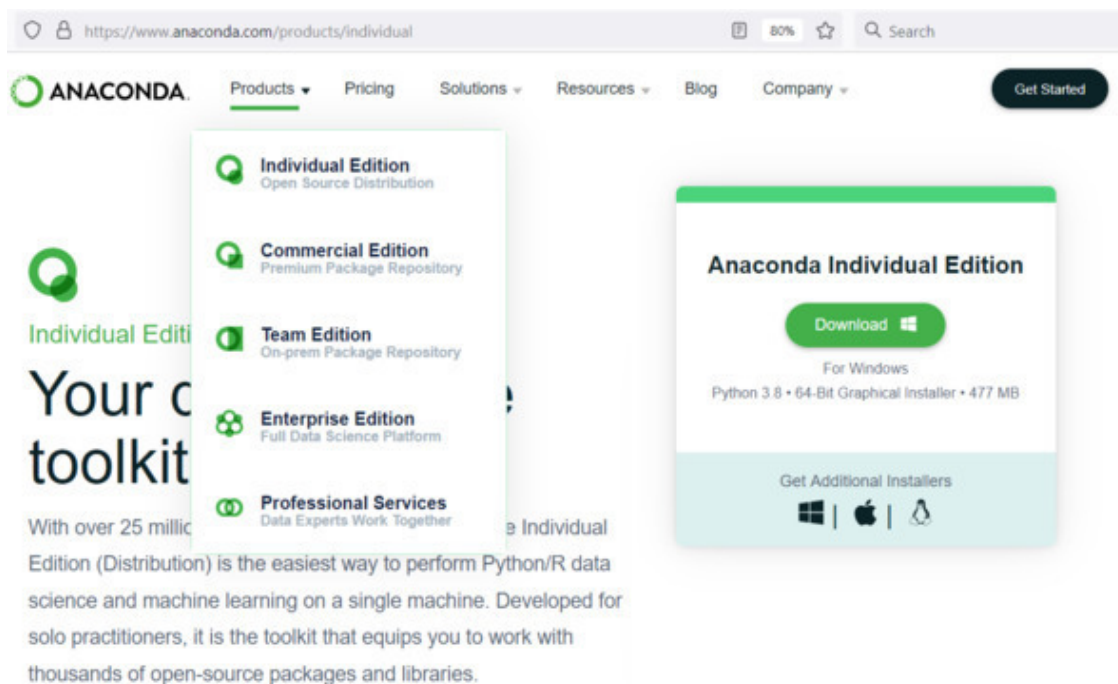


Рис. Варианты пакета Anaconda

На сайте www.anaconda.com нас будет интересовать бесплатная версия Individual Edition для личного пользования.

После установки мы получаем первое окно – так называемый навигатор Anaconda Navigator. В рамках этого окна нас будет интересовать Jupyter Lab – лаборатория Юпитер. Здесь есть ещё Jupyter Notebook – Блокнот Юпитер, см. рис. Слово Юпитер здесь написано почти как название планеты, но с намёком на язык Питон.

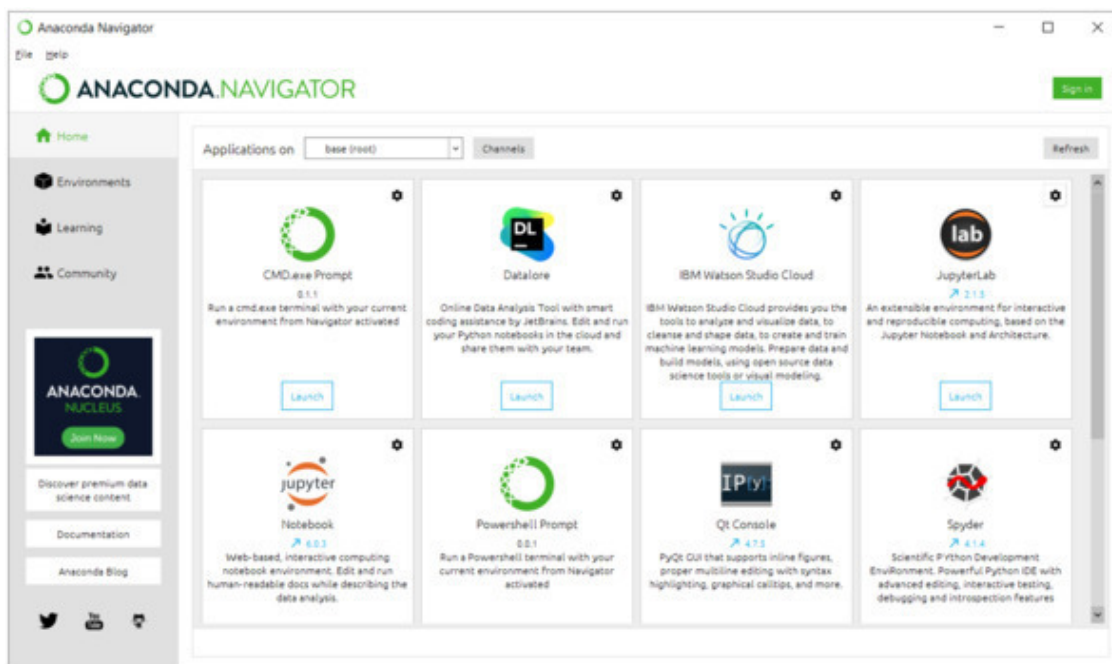


Рис. Интерфейс Anaconda Navigator

Мы запускаем Jupyter Lab получаем новое окно браузера. При этом через браузер мы получаем доступ к своему локальному компьютеру. В строке адреса указан локальный компьютер localhost и номер порта, см. рис.

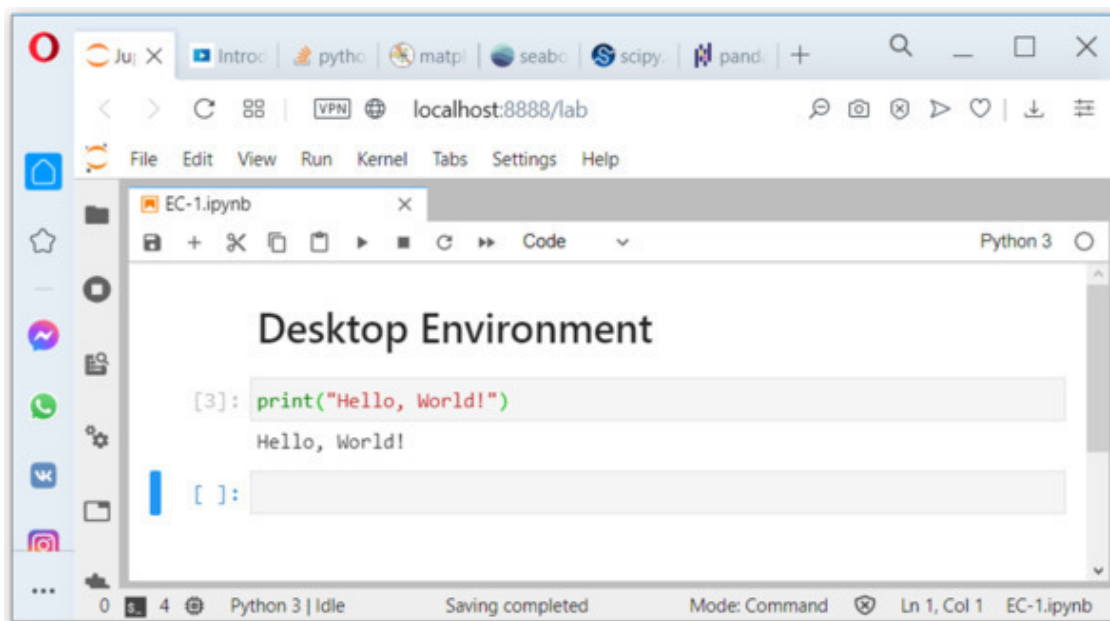


Рис. Среда Jupyter Lab

Мы получаем в окне браузера возможность работать с Jupyter Notebook – Блокнотом Юпитер. Нас интересует возможность работать в диалоге. Каждая ячейка блокнота – это одна или несколько строк кода. Мы можем запустить на выполнение любую ячейку блокнота в любом порядке и любое количество раз – и сразу видим результат выполнения.

Задание

Просмотрите статьи в Википедии про Anaconda и Jupyter Notebook на русском и английском языках. Обратите внимание, насколько различаются русский и английский варианты этих статей. Если есть трудности с английским, можно использовать Переводчик Google или включить автоматический переводчик, встроенный в браузер Google Chrome. Ознакомьтесь с расшифровкой названия Jupyter.

Задание

Скачайте и установите пакет Anaconda. Запустите Jupyter Lab. Создайте новый блокнот. Запустите программу «Hello, World».

Демонстрация Jupyter Lab

Запускаем Jupyter Lab. Создаём новый блокнот. Notebook – это страница с любым количеством строк. Появляется пустое окно. Есть разные варианты ячеек. Нас будет интересовать два вида. В кодовой ячейке Code пишем команды. В текстовой ячейке Markdown пишем текст. В нашем случае это будут просто заголовки. Markdown – это язык разметки текста, который позволяет украшать текст и вставлять формулы в формате LaTeX.

Вставляем символ «решётки» # и вводим название нашего первого упражнения. Напомним, что выражение «Hello, World» вообще-то переводится как «Всем привет». Перевод «Привет, мир» – это слишком дословно.

Нажимаем комбинацию клавиш **[Shift+Enter]**. Ячейка выполняется. Теперь у нас просто текст. Для редактирования можно дважды щёлкнуть по ячейке и внести исправления.

Обратим внимание, что первая ячейка имеет тип **Markdown**. Следующая ячейка автоматически получает тип кодовой **Code**.

Наша простая программа напечатает приветствие. Мы вводим команду **print**. Обратите внимание, что команда пишется маленькими буквами. Python различает заглавные и строчные буквы (большие и маленькие). Открываем круглую скобочку – автоматически появляется закрывающая скобка. Нажимаем кавычки – появляется закрывающая кавычка. Пишем сообщение. Нажимаем комбинацию клавиш **[Ctrl+Enter]**. В этом случае выполняется текущая ячейка, а новая ячейка не создаётся.

Система работает в режиме диалога. Можно загружать файлы и обрабатывать данные. Можно строить разнообразные графики.

Демонстрация Google Colab

Посмотрим в действии облачный сервис – очень простой и очень бесплатный. Называется он Google Colab – «ко-лаборатория», то есть совместная работа. Адрес в интернете:

<https://colab.research.google.com>

Единственное ограничение: нужно зайти со своей учётной записью Google. Для этого нужно зарегистрироваться на сайте Google. Подойдёт и учётная запись почты Gmail.

Интерфейс пользователя очень похож на предыдущий вариант, см. рис.

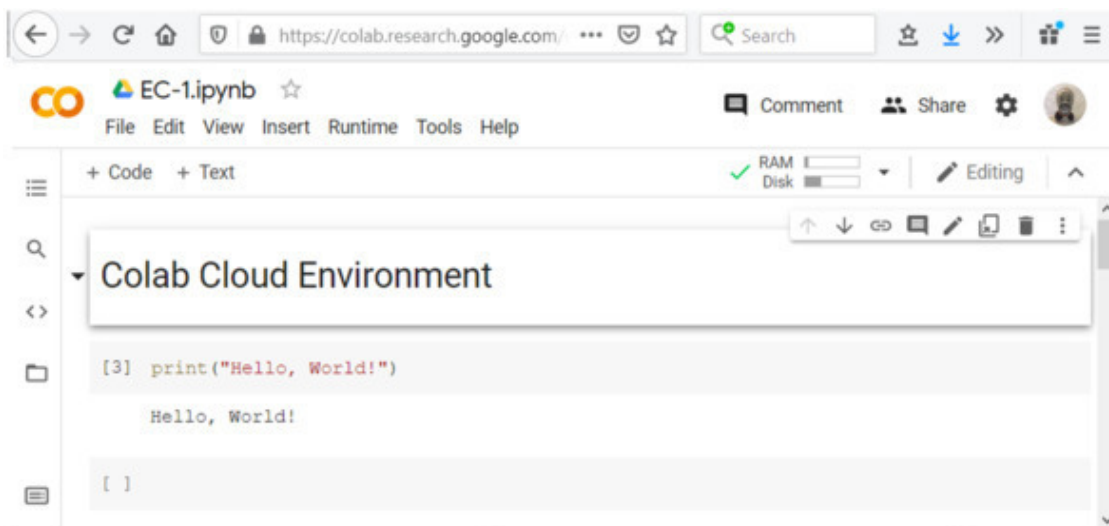


Рис. Интерфейс Google Colaboratory

Ко-лаборатория – это облачный сервис. Можно сказать, что это облачный блокнот Юпитер. Сам блокнот хранится на облачном диске Google Drive. Сюда можно закачивать файлы или скачивать их оттуда.

В верхней части окна отображается название файла. Расширение *.ipynb говорит о том, что в нём сохраняются не только команды, но и результаты выполнения, в том числе, графики. Такой файл можно посмотреть, не запуская его на выполнение, и увидеть, какие результаты были в прошлый раз. Можно также скачать только саму программу – файл с расширением *.py.

Интерфейс очень похож – тот же самый блокнот Юпитер. Точно также создаём ячейки с текстом или кодом, можем нажимать **[Shift+Enter]** и **[Ctrl+Enter]**.

Мы запускаем этот самый инструмент под названием Колаб. Создаем новый блокнот и запускаем нашу «игрушечную» программу, которая говорит: «Всем привет!»

Мы можем при желании переименовать блокнот. Двойным щелчком начинаем редактирование названия. В начале работы, при первом запуске первой ячейки происходит соединение с виртуальной машиной, выделяются ресурсы на сайте Google. Для добавления текстовой ячейки нажимаем кнопку **[+Text]**. Значок «плюс» означает «добавить ячейку в блокнот». Для организации заголовка ставим символ «решетки» #. Одиночный знак решетки означает, что это главный заголовок – первого, верхнего уровня. Для удаления ячейки щёлкаем по ней и справа над ячейкой нажимаем кнопку с иконкой мусорного ведра.

Начиная вводить команду, можно заметить всплывающую подсказку. Можно выбрать подходящую команду или аргумент.

Когда мы открываем скобки, автоматически появляются закрывающие. Сразу появляется краткое описание вводимой функции.

Обычно принято считать, что облачный сервис работает медленнее, чем программа на локальном компьютере. Однако, компания Google предоставляет свои дополнительные вычислительные возможности – причем бесплатно – для общего знакомства с продуктом. Есть специальные настройки для использования графических и тензорных аппаратных ускорителей. С другой стороны, если на вашей машине имеется мощная графическая плата с поддержкой технологии CUDA, тогда можно ускорить работу локальной версии Python. Это может быть полезно, если проводить сложные расчеты.

3. Вероятность

Приступим к новой теме. В каком-то смысле это будет повторение, потому что многие студенты изучают – в том или ином виде – курс теории вероятностей и математической статистики. Судя по школьным программам, в некоторых школах даже в старших классах затрагивают тему вероятности. Кроме того, данный материал частично обсуждается в курсе «Бизнес-аналитика и статистика».

Вероятность: p – Probability

- Число от 0 до 1 $\approx$ частота событий (уверенность)
- Бросаем монетку – орёл или решка

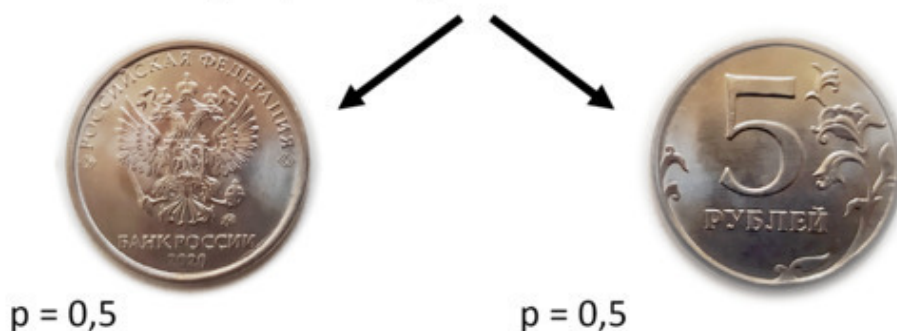


Рис. Вероятность и бросание монеты

Мы работаем с большим количеством исходных данных, реальных данных. В них всегда есть случайность, а случайность измеряется с помощью ВЕРОЯТНОСТИ.

Разберём классический пример – бросание монетки, см. рис. Мы бросаем монетку. Она может упасть с одной стороной или другой стороной. Это два ИСХОДА (результата) нашего опыта. Другие варианты пока не учитываем (например, монетка закатилась под стол и её не нашли). Стороны монеты обычно называют «орел» и «решка» («решётка»).

Если шансы абсолютно равны, то говорят, что вероятность – одна вторая.

Вероятность – это число от нуля до единицы.

Чем больше это число, чем ближе вероятность к единице, тем больше у нас уверенность, что произойдет именно это событие.

Если шансы одинаковы, то единица делится на количество этих вариантов. В нашем случае единичка делится на два варианта. Каждый вариант получает вероятность 0,5. Это традиционное объяснение из теории вероятностей.

Теория объясняет общие законы природы. Там, где есть случайность, можно говорить о вероятности события.

Вероятность обозначается латинской буквой p . Это первая буква английского слова *probability*. Происходит от латинского слова *proba* – «пробовать, проверять». Получается, что один раз попробовал что-то сделать – получилось, в другой раз попробовал – не получилось. Когда мы что-то пробуем, появляется случайность, неопределённость, непредсказуемость. И вероятность – это частота события, насколько часто происходит то, что нас интересует.

Русское слово «вероятность» имеет отношение к слову «верить». И это вторая сторона вероятности: насколько мы доверяем какому-то сообщению, утверждению, прогнозу.

Вероятность

- Теория вероятностей

- «Точное» теоретическое значение (предполагаем равные вероятности)

$$p(A) = \frac{1}{2} = 0,5$$

- Математическая статистика

- Оценка – «приблизительное» значение (по фактическим данным)

$$p(A) \approx \hat{p}(A) = \frac{N_A}{N} = \frac{4\,995}{10\,000} = 0,4995$$

Рис. Вероятность: теория и оценка

Теория – это всегда красиво и всегда «точно». Теория вероятности говорит нам: если шансы равны, то вероятность будет ровно 0,5 для орла и ровно 0,5 для решки. Это абсолютно «точное» значение – никаких сомнений. Но с одной оговоркой: если шансы равны, см. рис.

Статистика – это фактические данные. Когда мы переходим к обработке данных, то никогда не видим идеальные числа и идеальные условия. Мы каждый раз видим реальные наборы данных. Например, тысячу раз бросили монетку. Посчитали, сколько раз выпал орёл. Представим, что из десяти тысяч раз почти в половине случаев монетка упала орлом вверх. Делим на общее число бросаний и получаем число, которое будет приближаться вот к теоретическому значению.

Это число называется **ОЦЕНКА**. Оценки часто обозначают символом «крышечка» или «крышка». Настоящая вероятность и оценка по реальным данным приблизительно соответствуют друг другу, но только **ПРИБЛИЗИТЕЛЬНО**. Оценку мы можем посчитать: берём реальные данные и считаем. Такие опыты были в истории. В книгах можно найти упоминание о том, как ещё до появления компьютеров математики решили проверить свои теории. Они 10 тысяч раз бросали монетку. Можете представить, какая это работа. Бросать монетку и каждый раз записывать результат на бумаге, а потом вычислить оценку вероятности.

Обработка данных

- Реальные, фактические данные
 - Бросать монетку
- Имитационное моделирование
 - Генератор случайных чисел
- Результат
 - Оценка

Рис. Эксперименты и оценки

Вот в чём разница между теорией и обработкой данных. Теория говорит нам, как всё должно быть в принципе, при идеальных условиях. На практике мы берём реальные данные и обсчитываем то, что получили. Если у нас были разумные предположение, то результаты будут соответствовать теории.

Когда мы говорим о реальных данных, опыт с монеткой можно провести «физически» – действительно бросать монетку много раз. Есть и другой вариант – использовать программный генератор случайных чисел. В этом случае мы имеем дело с тем, что называется ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ. Это целая технология, но мы с вами затронем эту область только для того, чтобы увидеть, как это происходит. Мы будем моделировать идеальные, красивые данные, на которых можно тренироваться. Потом, на лабораторных работах можно будет взять реальные, настоящие данные. Они уже не такие «красивые». Они могут быть «корявыми», «неправильными», какими угодно. Но они чем-то будут напоминать нашу теорию и наши результаты.

В любом случае, мы можем взять реальные данные или результаты моделирования. То, что мы по ним посчитаем, каждый раз будет называться ОЦЕНКА. Оценка – это очень приблизительное значение, которое может соответствовать настоящему, теоретическому.

Есть разные способы моделирования, генерирования, создания случайных чисел. Пока в качестве инструмента для работы мы обсуждаем табличный редактор – Excel или его аналоги. Первый вариант – генератор случайных чисел в настройке «Анализ данных», см. рис.

Генератор - Надстройка Оценка вероятности ($n = 10000$)

D2 ✕ ✓ <i>fx</i> =AVERAGE(B2:B10001)								
	A	B	C	D	E	F	G	H
1	x	A		p				
2	0.124149	0		0.522				
3	0.0065	0						
4	0.389447	0						
5	0.267281	0						
6	0.703635	1						
7	0.235511	0						

Рис. Моделирование с помощью надстройки

Мы просто генерируем столбец чисел и определяем вероятность того, что выпадет орёл. В нашем случае получилось значение 0,522. Мы как бы «бросали монетку» заданное количество раз. Генератор случайных чисел создаёт число от нуля до единицы. Эти числа имеют равномерное распределение, то есть у нас одинаковые шансы (вероятность) появления любого числа от нуля до единицы. Далее мы берём полученное число и используем функцию округления. Как вы знаете, обычно округляют по стандартному правилу. Если число меньше 0,5 – округляем в меньшую сторону и пишем 0. Если число больше 0,5 – округляем в большую сторону и пишем 1. Таким образом, мы половина чисел округляется до нуля, а другая половина чисел округляется до единицы. Если ровно 0,5 – то в большую сторону.

Напомним, что есть как минимум шесть методов округления компьютера. Это обсуждается в курсе «Средства вычислительной техники». Можете этот материал ещё раз посмотреть. В нашем примере мы используем для округления функцию ROUND / ОКРУГЛ.

Excel – надстройка

Запускаем табличный редактор Excel.

Делаем заголовок – это наши «иксы».

Вызываем надстройку: Данные – Анализ данных. В диалоговом окне выбираем генератор случайных чисел.

Демонстрация

- Excel
 - Генератор случайных чисел
 - Надстройка «Анализ данных»
 - Равномерное распределение (Uniform)
 - Округление
 - Функция ROUND()
 - Оценка вероятности
 - Функция AVERAGE()

Рис. Общий план демонстрации

Задаём число переменных. Имеется ввиду число столбцов. В электронных таблицах каждая переменная – это, прежде всего, столбец. Это же касается и баз данных – там тоже есть таблицы, и данные в таблицах тоже расположены по столбцам. Нас интересует одна колонка случайных чисел. Пускай у нас будет 10 тысяч значений.

В разделе «Распределение» выбираем **Uniform** / Равномерное.

Диапазон значений от 0 до 1.

Параметр **Random seed** в русском переводе звучит как «Случайное рассеивание». На самом деле, это НАЧАЛЬНОЕ СОСТОЯНИЕ ГЕНЕРАТОРА. Этот параметр задаётся целым числом. Мы в качестве примера напомним 1234.

Диапазон для вывода результатов – **Output range**. Мы говорим, что надо выводить случайные числа, начиная с указанной ячейки.

Машина немного подумает и сгенерирует нам заданное количество случайных чисел.

После этого мы их округлим и посчитаем вероятность нашего события.

Событие А заключается в том, что у нас выпала единица. Мы пишем ROUND, даём ссылку на соседнюю ячейку и говорим: ноль знаков после запятой.

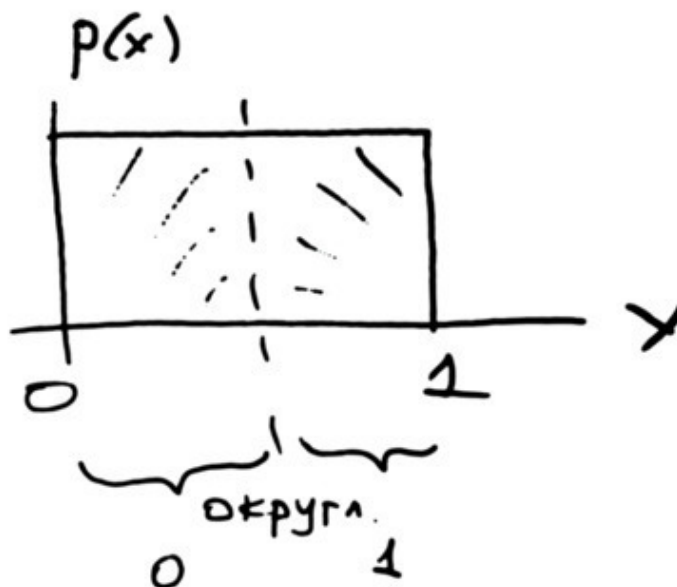


Рис. Моделирование нулей и единиц

Конечно, есть некоторые сомнения по поводу округления числа, равного точно 0,5. Проведем опыт: напишем 0,5 и округлим.

Сделаем откат и возвращаемся к исходным данным.

Теперь нам предстоит заполнить второй столбец. Берём правый нижний уголок ячейки – это маркер автозаполнения. Дважды щёлкаем по нему – и вся колонка заполняется нулями и единицами. Чтобы это проверить, нажимаем **[Ctrl+Down]**. Перемещаемся самый низ нашего столбца – в последнюю заполненную ячейку. Мы видим 10001 строку. Это последняя строка. Возвращаемся в начало: **[Ctrl+Home]**.

Теперь подсчитаем вероятность. Я напомним, что это только оценка вероятности. Поскольку наши данные – это нули и единицы, оценка вероятности – это среднее значение по столбцу А. Среднее значение в английском варианте Excel называется **AVERAGE**, в русском варианте – **СРЗНАЧ**. Указываем диапазон ячеек: щёлкаем первую ячейку, затем нажимаем комбинацию клавиш **[Ctrl+Shift+Down]**. Мы отметили все ячейки. Нажимаем клавишу **[Enter]** или **[Ввод]**.

Вот наш результат – мы получили число, которое очень близко к 0,5. Но это не ровно 0,5. Просто достаточно близко к теоретическому значению.

В чём особенность этого метода? Мы генерируем числа с помощью надстройки. Если мы хотим сгенерировать новый набор данных, нам нужно будет снова вызвать «Анализ данных». Мы вызываем генератор, задаём новое состояние генератора, нажимаем **[OK]**. Нас предупреждают, что ячейки вообще-то заняты предыдущими числами. Машин подумает и заполнит наши 10 тысяч ячеек. Обратим внимание на новую оценку вероятности.

Такой эксперимент вам предстоит проделать несколько раз – в рамках знакомства с данным материалом.

Подведём итоги. Мы вызвали генератор случайных чисел с помощью надстройки, установили равномерное распределение, округлили и нашли оценку вероятности – как среднее из нулей и единиц. Что это значит применительно к нашей формуле? В числителе находится количество единиц – это число событий, которые нас интересуют. В данном случае сумма единиц – это количество событий, когда выпал орёл. В знаменателе – общее количество событий. В нашем примере это 10000. Так что если у нас нули и единицы и мы считаем среднее из этого количества данных, мы автоматом получаем оценку нашей вероятности. Она называется **ЧАСТОТА**.

Если оценка вероятности получена по реальным данным, её называют частотой. Конечно, это просто количество – в отличие от физики. В физике частота – это число событий в единицу времени или число оборотов в единицу времени. В физике частоту измеряют в Герцах, а в статистике – в «единицах», в «штуках», в количестве объектов или событий. Так что это частота в разном понимании.

Excel – функция

Есть и другой способ генерирования случайных чисел. Это готовая функция **RAND** / **СЛЧИС**. Мы его тоже можем разобрать. Внешне всё будет выглядеть точно так же, как в предыдущем примере: столбец «иксов» и столбец нулей и единиц, затем оценка частоты, то есть вероятность, см. рис.

Генератор - Функция Оценка вероятности ($n = 10000$)

D2 X ✓ fx =AVERAGE(B2:B10001)								
	A	B	C	D	E	F	G	H
1	x	A		p				
2	0.94415	1		0.4930				
3	0.095111	0						
4	0.663017	1						
5	0.247932	0						
6	0.240889	0						
7	0.544861	1						

Рис. Демонстрация генератора-функции

Здесь есть некоторые тонкости.

Чтобы взять содержимое одной ячейки и заполнить диапазон ячеек, можно просто взяться за уголок маркера и протянуть вниз. Но чтобы дотянуть до ячейки под номером 10001, придётся потратить пару минут. А теперь представьте, что нужно скопировать формулу в миллион ячеек. Такое заполнение – надолго. Да ещё можно промахнуться.

Но не всё так плохо. У нас в запасе есть другой приём: мы вводим формулу в одну ячейку, затем указываем в окошке адреса ячейки нужный диапазон **A2:A10001**. В результате мы выделяем этот диапазон. Затем в верхнем меню выбираем **Home – Editing – Fill – Down**. Это означает «заполнить вниз от текущей ячейки». Так мы заполняем столбец нужного размера, см. рис.

Home – Editing – Fill – Fill Down

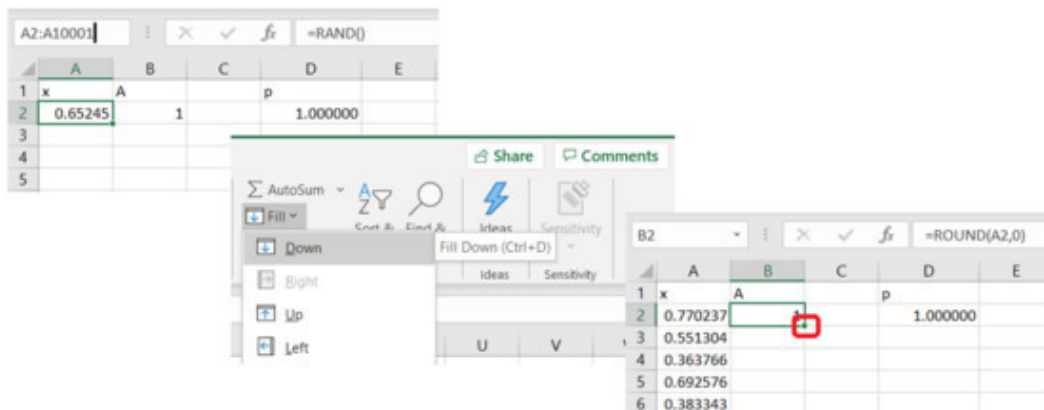


Рис. Заполнение столбца

Вторую колонку мы заполняем с помощью маркера автозаполнения – как в первом примере.

Демонстрация

- Excel
 - Генератор случайных чисел
 - Функция **RAND()**
 - Округление
 - Функция **ROUND()**
 - Оценка вероятности
 - Функция **AVERAGE()**
 - Изменение оценки

Рис. План демонстрации

Начнём демонстрацию. Переходим в Excel. Создаём новую страничку. В первой колонке у нас находится «икс». Мы пишем **RAND** и пустые скобки. Это функция без параметров. Она создает нам одно случайное число. Щёлкаем по ячейке. В окне адреса укажем диапазон адресов от **A2** до **A10001**. Видим, что у нас выделился диапазон. Далее нужно его заполнить: **Home – Editing – Fill – Down**. Мы заполнили весь диапазон. Можем проверить [**Ctrl+Down**]. Переходим в последнюю ячейку **A10001**, дальше ничего нет.

Нас интересует событие A. Мы пишем **ROUND**. Обратите внимание, что первый столбец автоматически пересчитывается. Округляем ячейку **A2**, ставим запятую и ноль. В русском варианте аргументы функции разделяются другим значком (точка с запятой), потому что в русском варианте символ запятой – это десятичный разделитель. Он разделяет целую и дробную части числа. Двойной щелчок по маркеру автозаполнения – получили столбец нулей и единиц.

Находим нашу частоту, то есть оценку вероятности случайного числа 0,5. Вызываем функцию **AVERAGE**.

Сохраняем файл – нажимаем [**CTRL+S**] несколько раз. Каждый раз при сохранении файла значения пересчитываются.

Как видим, для того чтобы вызвать генерацию нового набора данных, достаточно внести любые изменения или просто обновить / сохранить таблицу. При этом мы явно не указываем начальное состояние генератора. Просто при каждом новом обращении в функции **RAND** текущее состояние генератора меняется.

В одних случаях именно это интересно. В других случаях, наоборот, требуется работать с одним набором данных и не изменять его. Это первая разница между генератором в надстройке и генератором-функцией.

Есть и другое отличие – это количество данных, которые можно сгенерировать. У надстройки есть ограничения. Попробуйте до него дойти.

Надо понимать, что у самого Excel есть ограничение по количеству строк на одном листе электронной таблицы. Их может быть чуть больше 1 миллиона. То есть полтора или два миллиона строк вы уже не сможете создать на одном листе.

Но если вы захотите сгенерировать свой «разрешённый» миллион случайных чисел с помощью надстройки, то может и не получиться. Попробуйте выяснить, до какого предела можно генерировать с помощью надстройки. Это тоже интересный эксперимент.

Итак, вам предстоит повторить показанные опыты двумя способами – как мы с вами разбирали – с помощью надстройки Анализ данных и с помощью функции **RAND**, см. рис.

Задание

- Excel
- Сгенерировать 10000 случайных чисел
 - Надстройка
 - Функция
- Повторить опыт 10 раз
- Записать оценки вероятности

Рис. План задания

Вам предстоит сгенерировать 10000 случайных чисел и повторить этот опыт 10 раз. Затем записать ваши оценки вероятности. Можете записать от руки в конспект или приложить копию экрана, чтобы показать, что вы это проделали.

Вот таким образом мы с вами знакомимся с понятием вероятности. Вероятность существует в теории. Мы на неё можем посмотреть через обработку данных, но каждый раз это число будет немножко отличаться. Она будет близкой к настоящему, правильному, теоретическому значению. Но каждый раз она будет отличаться. Это явление природы называется словом **ОЦЕНКА**. Как видим, оценка тоже содержит себе некоторую случайность, некоторую непредсказуемость, маленькую случайную ошибку. Насколько она маленькая, что с этим делать и как дальше работать? Эти тонкости и будут темой следующего занятия.

В данном занятии мы посмотрели на так называемые псевдослучайные числа – то есть они на самом деле не совсем случайные. Нам только кажется что они случайные. Генерируется последовательность, которая зависит от начального состояния. Мы увидим этот инструмент и в Excel, и в любых других программных генераторах, в том числе, и в питоне. То есть в Python.

Libre Office Calc – надстройка

Проведём ещё один эксперимент. Нам предстоит как сгенерировать случайные числа в электронной таблице Libre Office Calc. Calc – это сокращение от слова «Калькулятор».

Основная идея всё та же – вызвать генератор через надстройку. Немного отличается верхнее меню, но разобраться будет несложно.

Демонстрация

- Libre Office Calc
- Генератор случайных чисел
 - Заполнение ячеек
- Округление
 - Функция ROUND()
- Оценка вероятности
 - Функция AVERAGE()

Рис. План демонстрации

В верхнем меню выбираем **Sheet – Fill Cells – Fill Random Number**, см. рис.

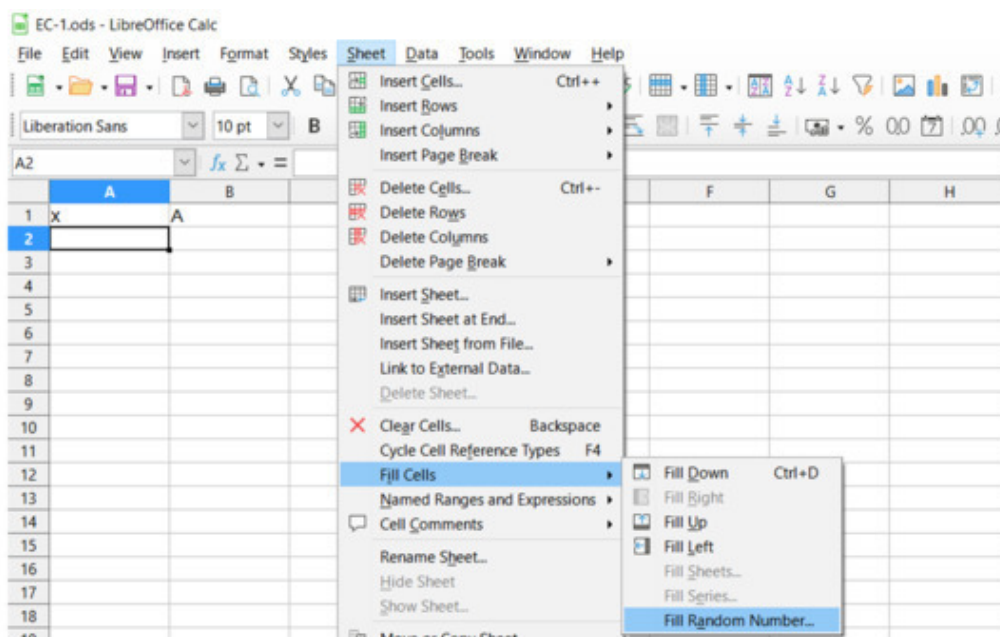


Рис. Вызов генератора случайных чисел

Появляется меню настройки генератора случайных чисел, см. рис. Выбираем равномерное распределение: **Distribution – Uniform**. Значения **Minimum / Maximum**: от нуля до единицы. Диапазон адресов ячеек **Cell range – A2:A10001**. Как видим, здесь у нас имеется дополнительная галочка **Enable custom seed**, если мы вручную задаём начальное состояние генератора.

Есть дополнительная возможность сразу же и округлять полученные случайные числа – **Enable rounding**. Но мы просто повторим предыдущие шаги, чтобы увидеть, насколько похожи все электронные таблицы.

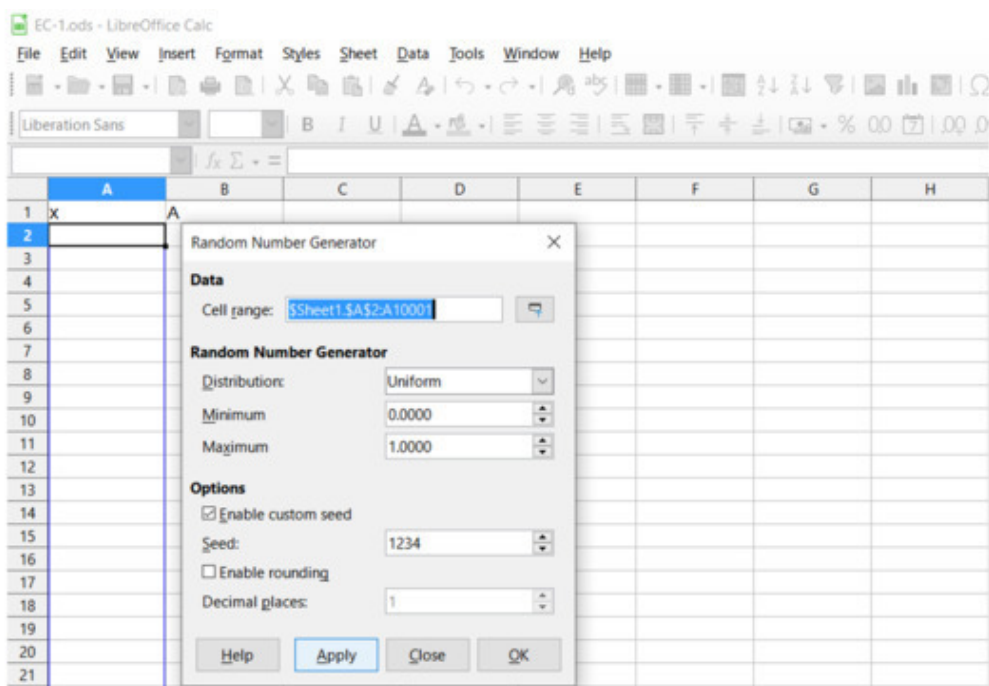


Рис. Настройка генератора случайных чисел

Далее мы округляем эти числа и находим среднее значение, см. рис. Функции такие же, как в Excel. Это обеспечивает почти полную совместимость на уровне файлов, включая названия функций.

Генератор – Fill Random Number Оценка вероятности ($n = 10000$)

	A	B	C	D	E	F	G
1	x	A		p			
2	0.497663666	0		0.4958			
3	0.817838443	1					
4	0.612111894	1					
5	0.771359919	1					
6	0.860669773	1					
7	0.150636966	0					

Рис. Результаты моделирования

Как можно видеть, и в этом примере оценка вероятности приблизительно соответствует точному, теоретическому значению 0,5.

Libre Office Calc – функция

Проведём ещё один эксперимент – по примеру того, что мы проделали в Excel с вызовом функции **RAND**, см. рис.

Демонстрация

- Libre Office Calc
- Генератор случайных чисел
 - Функция RAND()
- Округление
 - Функция ROUND()
- Оценка вероятности
 - Функция AVERAGE()
- Изменение оценки

Рис. План демонстрации

Вводим функцию **RAND** в ячейку **A2**.

Затем вызываем заполнение диапазона нашей формулой: **Sheet – Fill Cells – Fill Down**, см. рис.

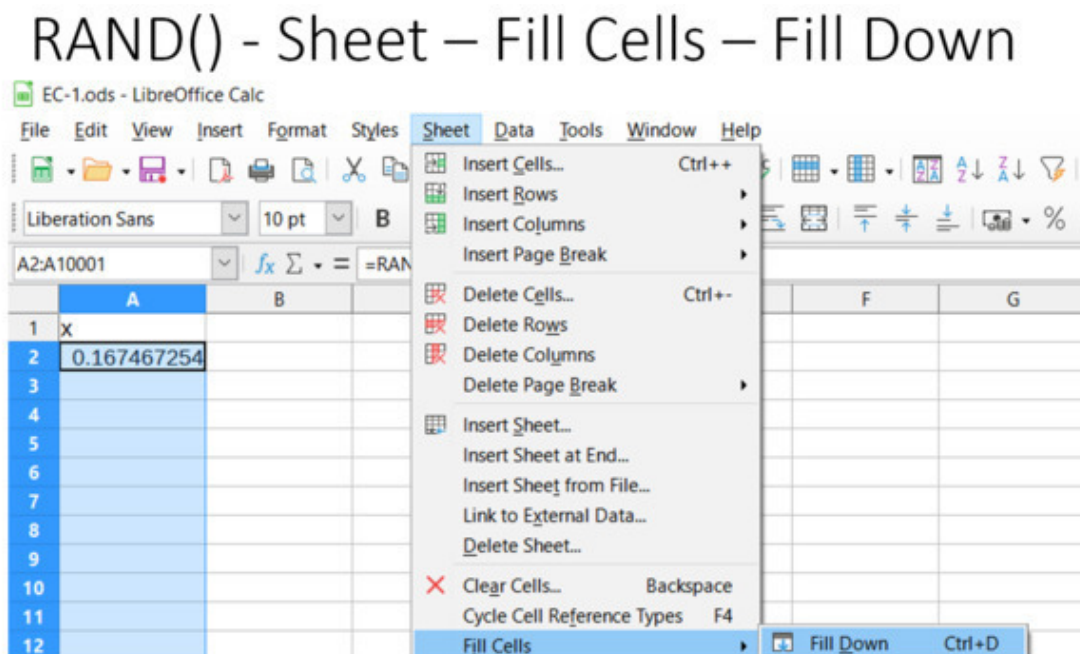


Рис. Заполнение ячеек

Генератор - Функция Оценка вероятности ($n = 10000$)

	A	B	C	D	E	F	G
1	x	A		p			
2	0.716929299	1		0.473			
3	0.593258634	1					
4	0.164923766	0					
5	0.211267995	0					
6	0.206067317	0					
7	0.852706934	1					

Рис. Результаты моделирования

Как видим, мы можем заполнять ячейки почти так же, как мы это проделали в Excel. Названия функций тоже совпадают. Во многом это объясняется тем, что пользователи ожидают совместимость на уровне файлов. А в файлах могут быть не только числа, но и вызовы функций. В нашем примере это функции **RAND**, **ROUND** и **AVERAGE**.

В результате мы тоже получили оценку вероятности около 0,5. И тоже с небольшой погрешностью. И эта случайность тоже заметна при многократном повторении опыта.

Когда мы вызываем генератор через функцию, электронная таблица пересчитывает все значения при любых изменениях, при обновлении таблицы и при сохранении файла. Все оценки будут вокруг теоретического значения 0,5, но все будут немного разными – плюс-минус.

Вам предстоит проделать показанные эксперименты. Повторите этот опыт несколько раз, чтобы убедиться, что оценка вероятности немного меняется. Но в среднем оценка «крутится» вокруг точного значения, см. рис.

Задание

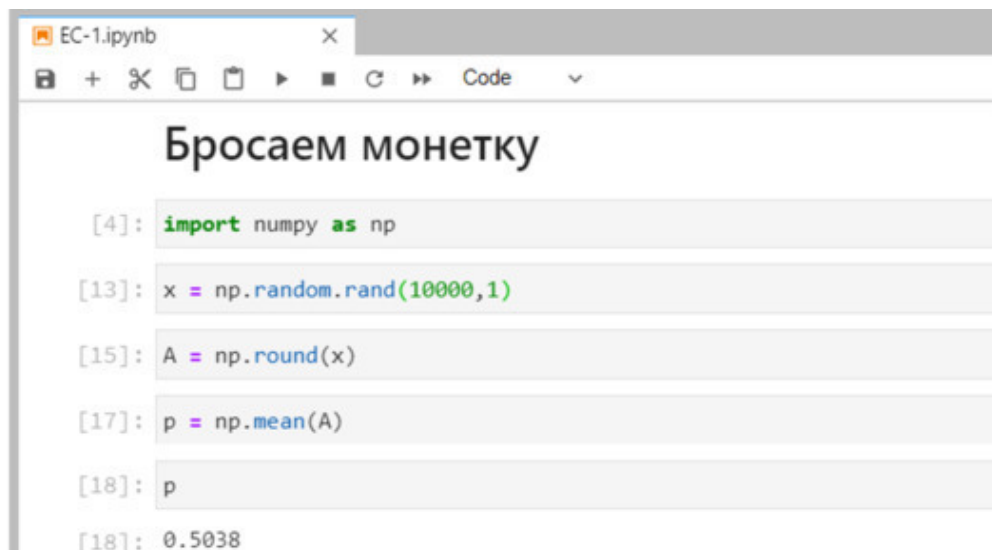
- Libre Office Calc
- Сгенерировать 10000 случайных чисел
 - Заполнение диапазона
 - Функция
- Повторить опыт 10 раз
- Записать частоты (оценки вероятности)

Рис. План задания

Jupyter Lab

Следующий эксперимент мы сделаем в питоне. Или в Python – если больше нравятся английские названия.

Здесь мы с вами познакомимся с некоторыми приемами работы в диалоговой среде Anaconda / Jupyter Lab и некоторыми командами Python.



```
EC-1.ipynb
[4]: import numpy as np
[13]: x = np.random.rand(10000,1)
[15]: A = np.round(x)
[17]: p = np.mean(A)
[18]: p
[18]: 0.5038
```

Рис. Программа в Jupyter Lab

В первой строчке мы импортируем библиотеку **numpy** и назначаем ей псевдоним **np** – для краткости. Это библиотека для работы с числовыми массивами. В обычном, базовом питоне мы тоже можем создавать различные объекты. Однако, **numpy** позволяет работать с матрицами, то есть с массивами / таблицами чисел. Это могут быть столбцы, или строки, или таблички чисел. Все они условно называются массивами.

Дальше мы будем обращаться к функциям из этой библиотеки **np**.

Вторая строка – вызов генератора случайных чисел с равномерным распределением. В аргументах функции **rand** указываем размеры массива, который хотим получить: 10000 строк и 1 столбец.

Следующим шагом мы округляем эти числа с помощью функции **round**.

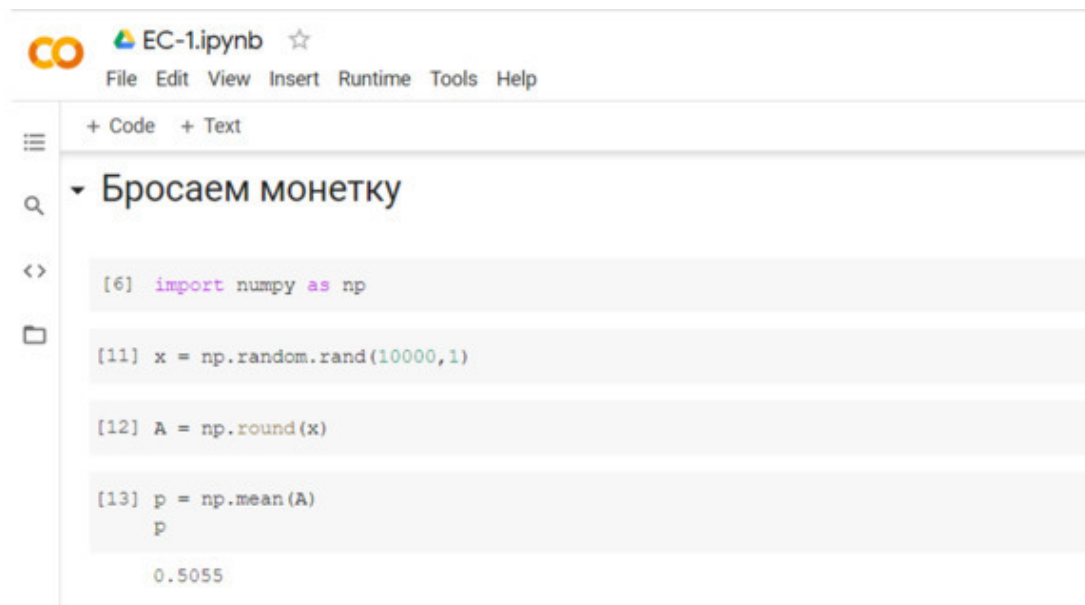
Далее находим среднее значение для всего этого массива чисел. Это делает функция **mean**. Полученную оценку вероятности выводим на экран.

Здесь надо отметить один любопытный момент. Функции для вычисления среднего значения могут называться **MEAN** и **AVERAGE**. Могут быть и другие названия. Причём это происходит в рамках одного пакета программ – если это делали в разное время и разные команды разработчиков. За этим приходится следить. Конечно же, мы всегда можем посмотреть справку под названием **help**. Там же обычно даются примеры использования команды.

Запускаем нашу программу несколько раз и наблюдаем, что выводится на экран в качестве оценки вероятности.

Google Colab

Мы можем проделать наши опыты с теми же результатами и в облачном сервисе Google Colaboratory, см. рис.



```
[6] import numpy as np

[11] x = np.random.rand(10000,1)

[12] A = np.round(x)

[13] p = np.mean(A)
p
0.5055
```

Рис. Программа в Google Colab

Внешне всё очень похоже на Anaconda / Jupyter Lab. Мы получаем практически стопроцентную совместимость с программой, отлаженной на локальном компьютере. Тем не менее, иногда бывает тонкости. Colab работает с самыми последними версиями библиотек. И нам не требуется постоянно их обновлять вручную.

Итак, мы запускаем Colab. Нас спрашивают, хотим ли мы открыть существующий файл или создать новый. Создаем новый блокнот.

Вводим первую команду, нажимаем комбинацию клавиш **[Shift+Enter]**. Проходит некоторое время на запуск – в правой верхней части экрана выводится сообщение про соединение с виртуальной машиной: **Connect**.

Обратим внимание, что при вводе функцию сразу появляется всплывающая подсказка.

Когда ячейка выполнена, слева от этой ячейки блокнота видим комментарий в квадратных скобках – выводится какое-то число. Это порядковый номер выполненной ячейки. Ячейки можно запускать в разном порядке, и это будет отображаться в квадратных скобках.

И ещё один момент: при вычислении среднего значения не уточняется метод расчёта. В описании функции говорится: *arithmetic mean*, то есть среднее арифметическое. На занятиях по статистике вы можете узнать, что среднее можно считать десятью разными способами. Но средняя арифметическая простая используется чаще всего.

Чтобы вывести полученную оценку на экран мы просто вводим имя переменной.

Запускаем несколько раз: **Runtime – Run All**.

Для вывода на экран можно также использовать команду **print**. Это обеспечит побольше знаков после запятой. Здесь можно задать любой формат вывода.

Подведём итоги. Неважно, какими средствами анализа мы пользуемся. Результаты обработки данных каждый раз представляют собой случайные числа. Они будут приближаться к точному, правильному значению. Но оценка содержит внутри себя случайность.

Ваша задача – потренироваться и убедиться в следующем. Оценки – это результат обработки реальных данных. Исходные данные содержат случайность. Поэтому оценки тоже являются случайными числами. Нужно проделать этот опыт на локальном компьютере и в облаке, см. рис.

Задание

- Jupyter Lab
- Colab
- Сгенерировать 10000 случайных чисел
- Повторить опыт 10 раз
- Записать частоты (оценки вероятности)

Рис. План задания

4. Распределение

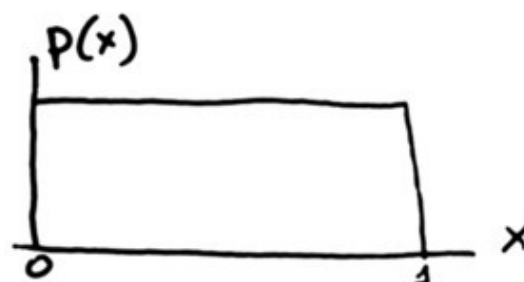
Наша следующая тема – распределение. А точнее, распределение вероятностей. Это понятие из теории вероятностей.

Чтобы всё запутать, у нас есть ещё одно понятие распределения – в экономике. Это касается дистрибуции, когда оптом берут крупную партию товара и развозят по магазинам мелкими партиями. Конечно, это не имеет никакого отношения к распределению в статистике.

Нас будет интересовать статистика, эконометрика, теория вероятности. Здесь распределение – это зависимость, показанная на рис.

Распределение

- Равномерное



- Нормальное

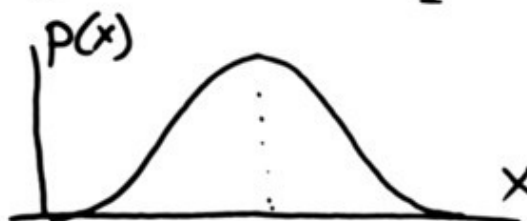


Рис. Примеры стандартных распределений

Итак, распределение – это вероятность появления разных значений какой-то случайной величины. На рисунке приведены два примера – равномерное и нормальное распределение. Мы их подробно исследовали на лабораторных работах по статистике.

При использовании программного генератора достаточно указать название распределения и его параметры.

Нормальное распределение имеет один пик. В целом, такая форма кривой называется колоколообразной. То есть она похожа по форме на колокол.

Соответствующее английское название – **Probability Distribution**. Probability – это вероятность, Distribution – распределение.

Распределение вероятностей – это вероятность появления разных значений случайной величины. Когда мы обрабатываем реальные данные, эту вероятность мы можем найти только приблизительно с помощью оценок. На практике распределение – это частота появления разных значений. Что-то бывает чаще, что-то бывает реже.

Чтобы сгенерировать случайные числа, мы используем программный генератор.

Генератор случайных чисел

- Data – Analysis – Data Analysis – Random Number Generation

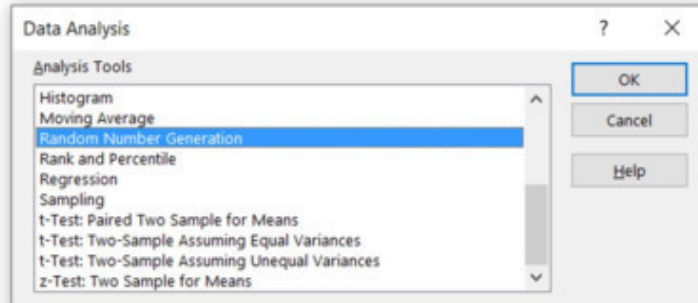


Рис. Запуск генератора

Всё начинается с равномерного распределения. Случайное число от нуля до единицы. Это считается своеобразным стандартом, строительным «кирпичиком» для реализации любого другого распределения.

Равномерное распределение

- Uniform Distribution
- Числа от 0 до 1

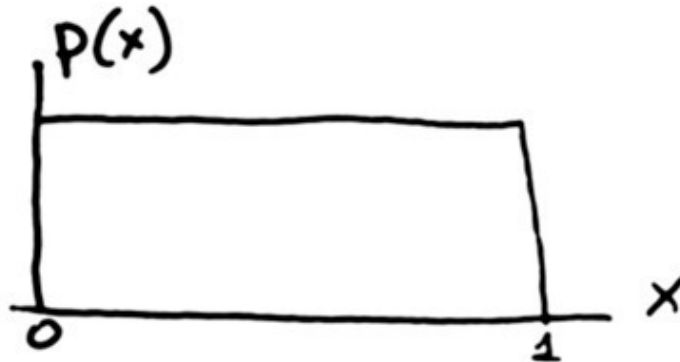


Рис. Стандартное нормальное распределение

В некоторых случаях мы можем вручную указать тот диапазон значений, который нас интересует. Стандартные параметры – это диапазон значений от нуля до единицы.

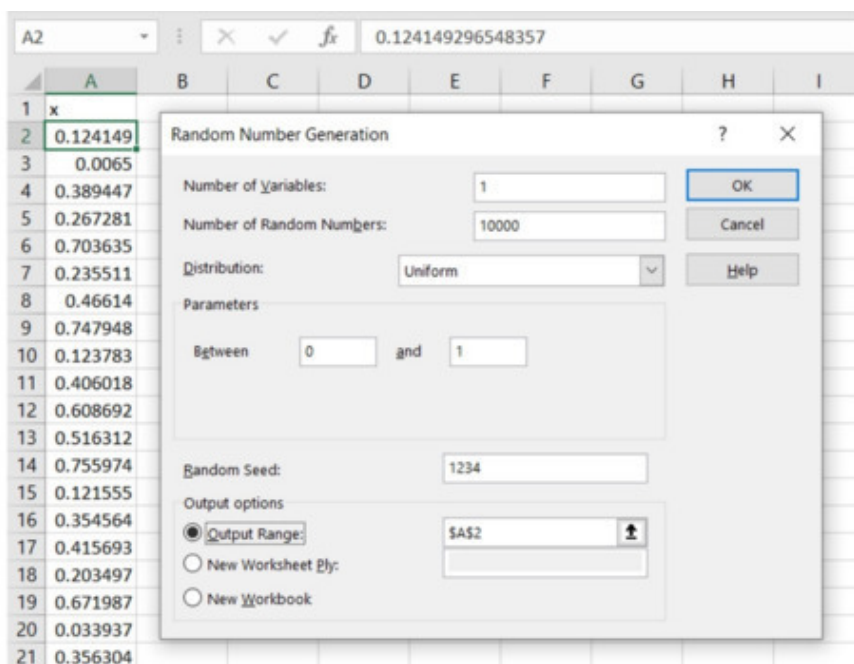


Рис. Настройка генератора

Запускаем генератор случайных чисел. В диалоговом окне указываем число переменных, см. рис. Напомним, что переменные в электронных таблицах и во многих других случаях располагаются по столбцам. Это имеет отношение к истории. Традиционно, задолго до появления компьютеров числа записывали в колонку. Внизу столбца подсчитать сумму. Вручную числа удобно складывать столбиком. Соответственно, и в компьютерах используется традиционное расположение данных. Оно интуитивно понятно.

Это касается электронных таблиц, баз данных, обработки данных в Python. И это касается настройки генератора случайных чисел. Количество случайных значений задаётся как количество строк. Количество переменных – это число столбцов.

Далее указываем форму распределения – равномерное.

Параметры равномерного распределения – минимальное и максимальное значения. По умолчанию от нуля до единицы.

Начальное значение генератора: 1234.

Диапазон ячеек для вывода.

Вставка гистограммы

- Выделяем диапазон ячеек
- Insert – Charts – Insert Statistical Chart – Histogram

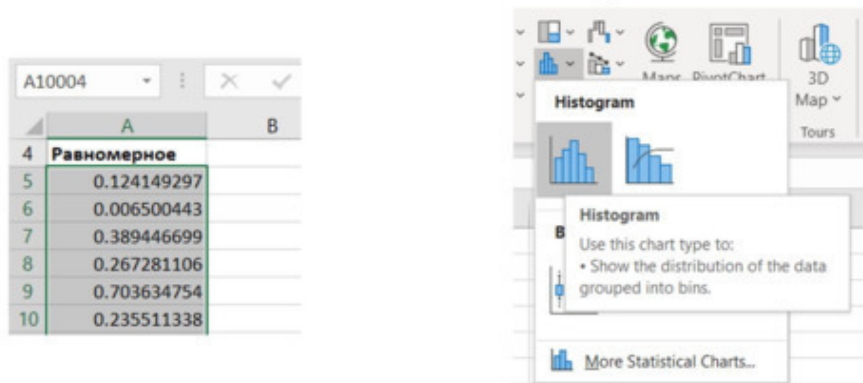


Рис. Вставка гистограммы как статистической диаграммы

Чтобы рассмотреть полученное распределение, используем график под названием гистограмма.

Мы рассматриваем пример в Excel.

Выделяем диапазон ячеек. Вызываем вставку гистограммы, см. рис.

В меню **Insert – Chart** есть две похожие кнопки: **Bar Chart** и **Histogram**.

Bar Chart – это простая столбиковая диаграмма, причём для каждого значения входного диапазона строится свой отдельный столбик. В некоторых вариантах перевода эта кнопка обозначена как «Гистограмма». Конечно, это запутывает пользователей. Если построить столбиковую диаграмму по 10000 значений, мы получим 10000 столбиков. Можете попробовать – чтобы больше не «попадаться».

Histogram – это гистограмма, то есть столбиковая диаграмма частот. Частоты считают для сгруппированных данных. Это частота попадания чисел в интервалы (диапазоны значений), см. рис.

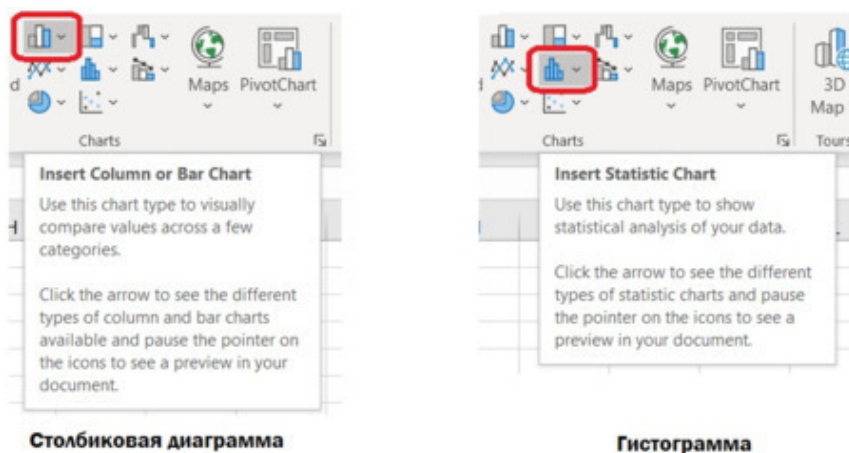


Рис. Меню Вставка – Диаграмма

Этот новый инструмент – «Гистограмма как статистическая диаграмма» – работает только в последней версии Excel. Поэтому рекомендуем при возможности обновить версию MS Office.

Построим гистограмму и посмотрим, чем она отличается от других графиков. Всплывающая подсказка сообщает, что этот график позволяет изучить распределение данных, сгруппированных по интервалам значений. В английском варианте интервал группировки называется **bin**. Буквально слово **bin** означает «корзина». Иногда его переводят словом «карман». Более грамотно было бы назвать его «интервал группировки данных».

Таким образом, наши данные – от нуля до единицы – сгруппированы по нескольким интервалам. Excel сам автоматически подсчитывает, сколько чисел в какой интервал попадает. Затем строится столбиковый график полученных частот. Так что это особый график со встроенной автоматической обработкой данных.

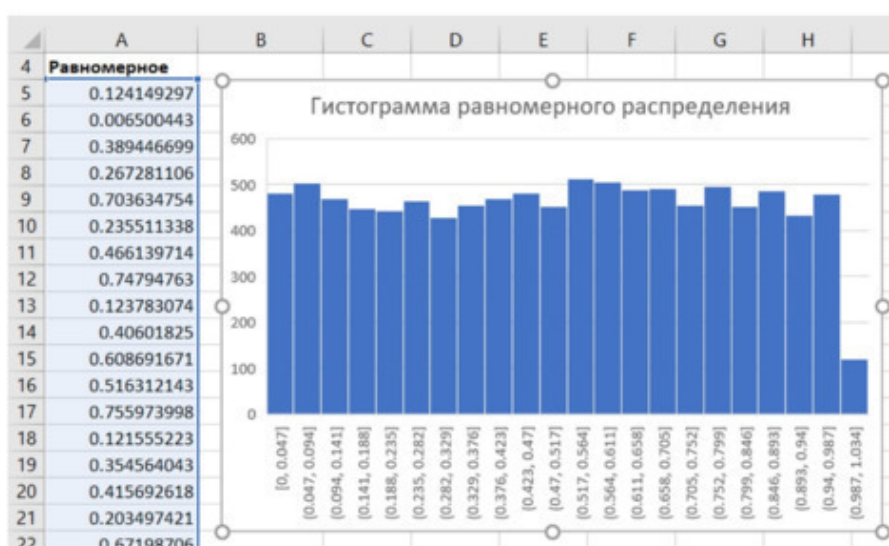


Рис. Гистограмма с автоматическими настройками

Мы получаем график распределения – «автоматическую гистограмму». В теории все столбики для гистограммы равномерного распределения должны быть одной высоты. Гистограмма по ограниченному набору данных не идеальная. Высота каждого столбика – частота – содержит в себе случайность, небольшую случайную ошибку.

Внизу, в качестве подписи под столбиками указаны те самые интервалы группирования данных. Нижняя граница обозначена круглой скобкой – не включается в расчёты. Верхняя граница – квадратная скобка – включается. Таки образом, число на границе будет относиться к нижнему интервалу. Что, ка и почему здесь происходит – более подробно обсуждается в курсе «Статистика».

Гистограмма

histo – «столбик»

+

gram – «график, рисунок»

- Столбики – частота появления разных значений
- Частота $\approx$ вероятность

Рис. Происхождение слова «Гистограмма»

Само слово «гистограмма» буквально означает «столбиковая диаграмма», то есть «график в виде столбиков». Первая часть – *histo* – означает «столбик», вторая часть – *gram* – «график, диаграмма».

Наши столбики изображают частоту появления разных значений случайной величины. Напомним, что частота примерно соответствует вероятности.

Далее нас будет интересовать другой диапазон значений. Для этого мы задаём другие параметры генератора и получаем желаемую форму распределения, см. рис.

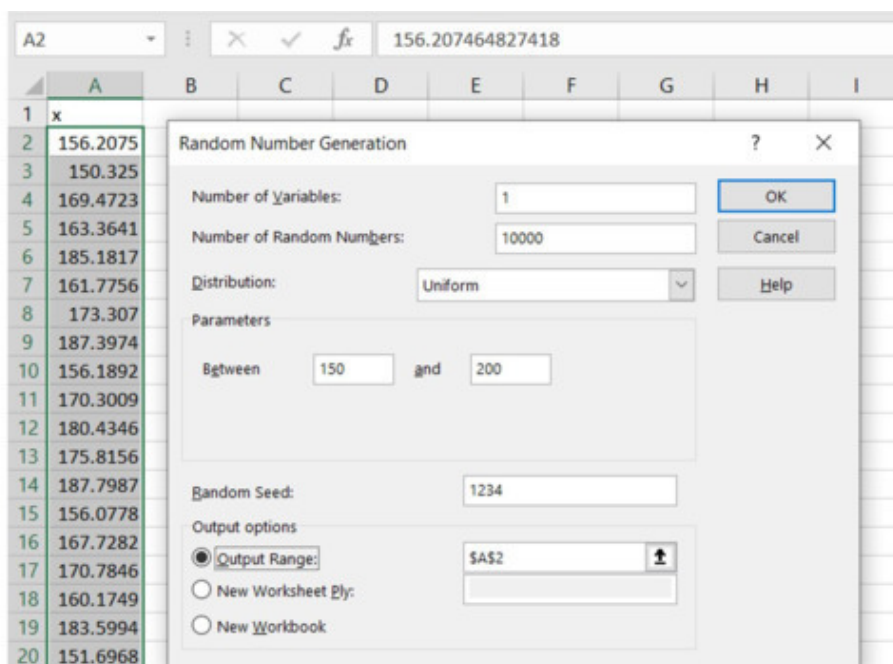


Рис. Заданные границы генератора

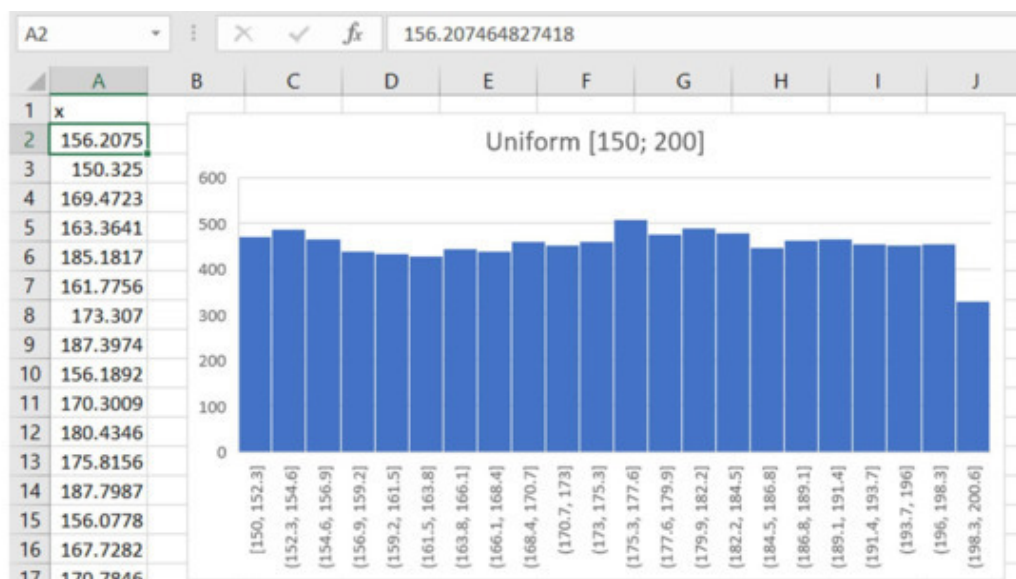


Рис. Распределение с заданными границами

Демонстрация. Excel

Конец ознакомительного фрагмента.

Текст предоставлен ООО «ЛитРес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на ЛитРес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.