Основные понятия в математической статистике

Д.С. Дружинин

Генеральная совокупность

- *Генеральная совокупность* – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.
- Генеральная совокупность может быть *конечной* или *бесконечной* в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.
- Все что может произвести завод – бесконечная генеральная совокупность, общее число живых людей на планете – конечная генеральная совокупность.

Выборочная совокупность

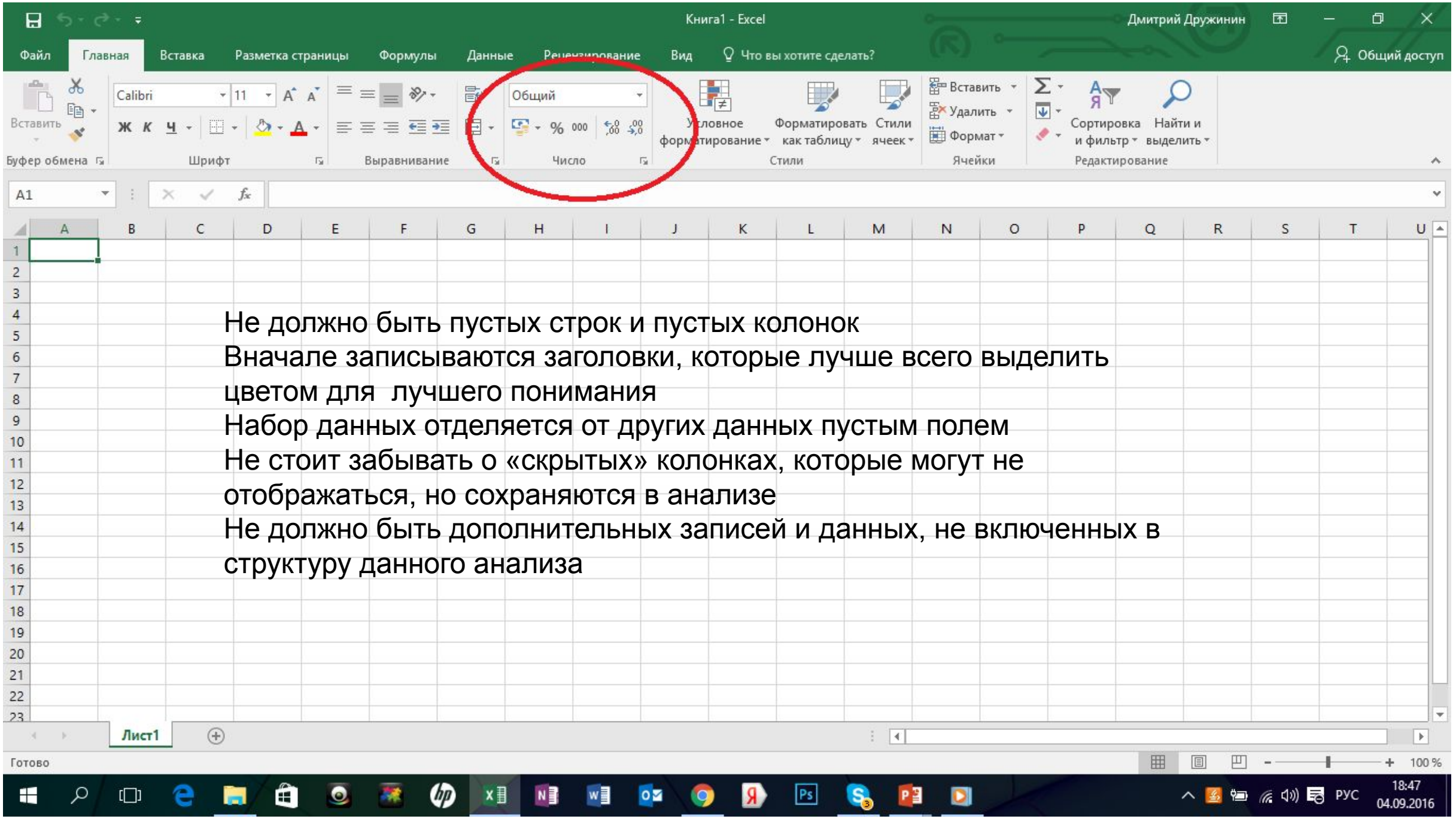
- *Выборкой (выборочной совокупностью)* называется совокупность случайно отобранных объектов из генеральной совокупности.
- Выборка должна быть *репрезентативной (представительной)*, то есть ее объекты должны достаточно хорошо отражать свойства генеральной совокупности.
- Выборка может быть *повторной*, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность, и *бесповторной*, при которой отобранный объект не возвращается в генеральную совокупность.

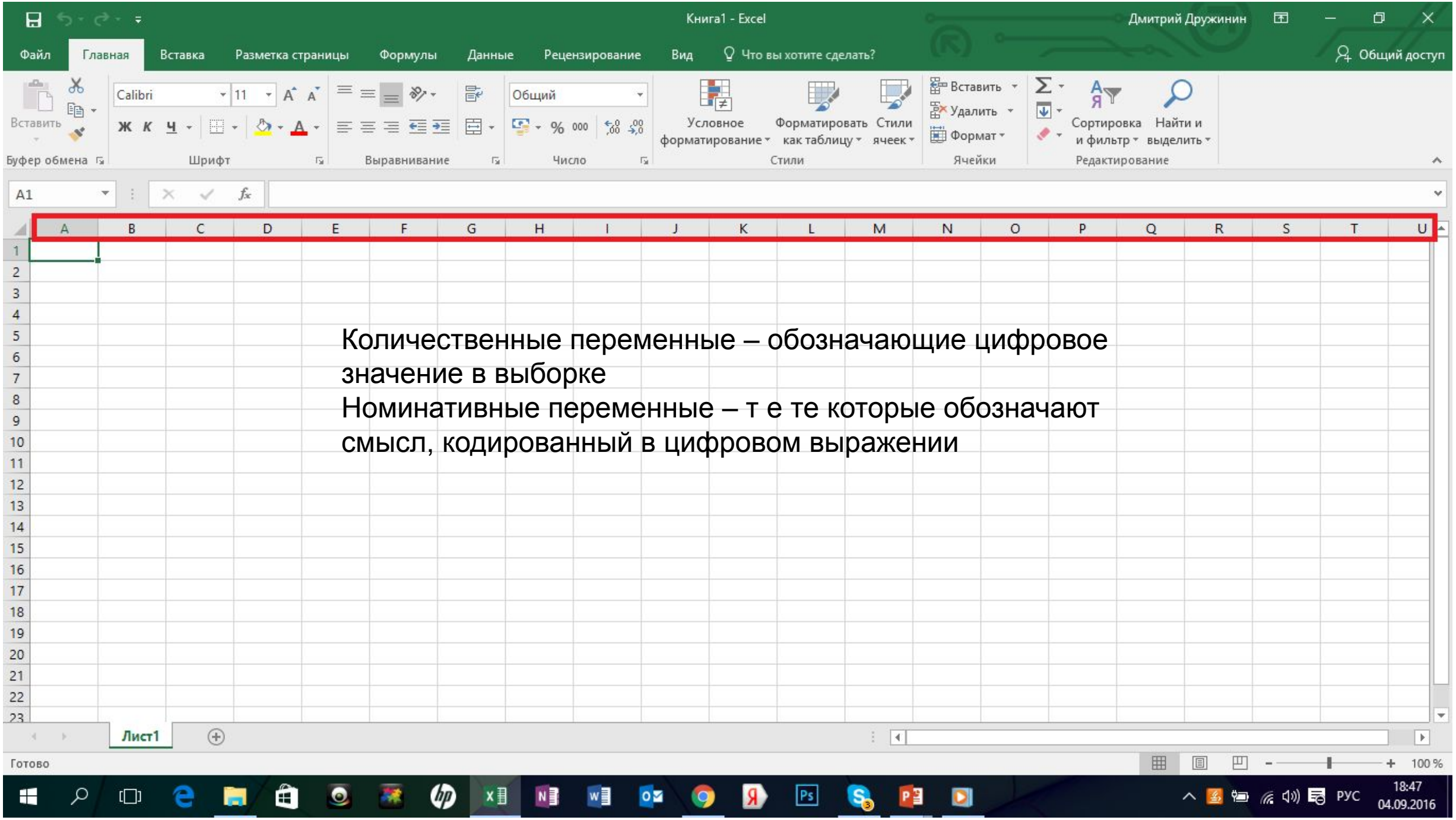
Способ получения выборки

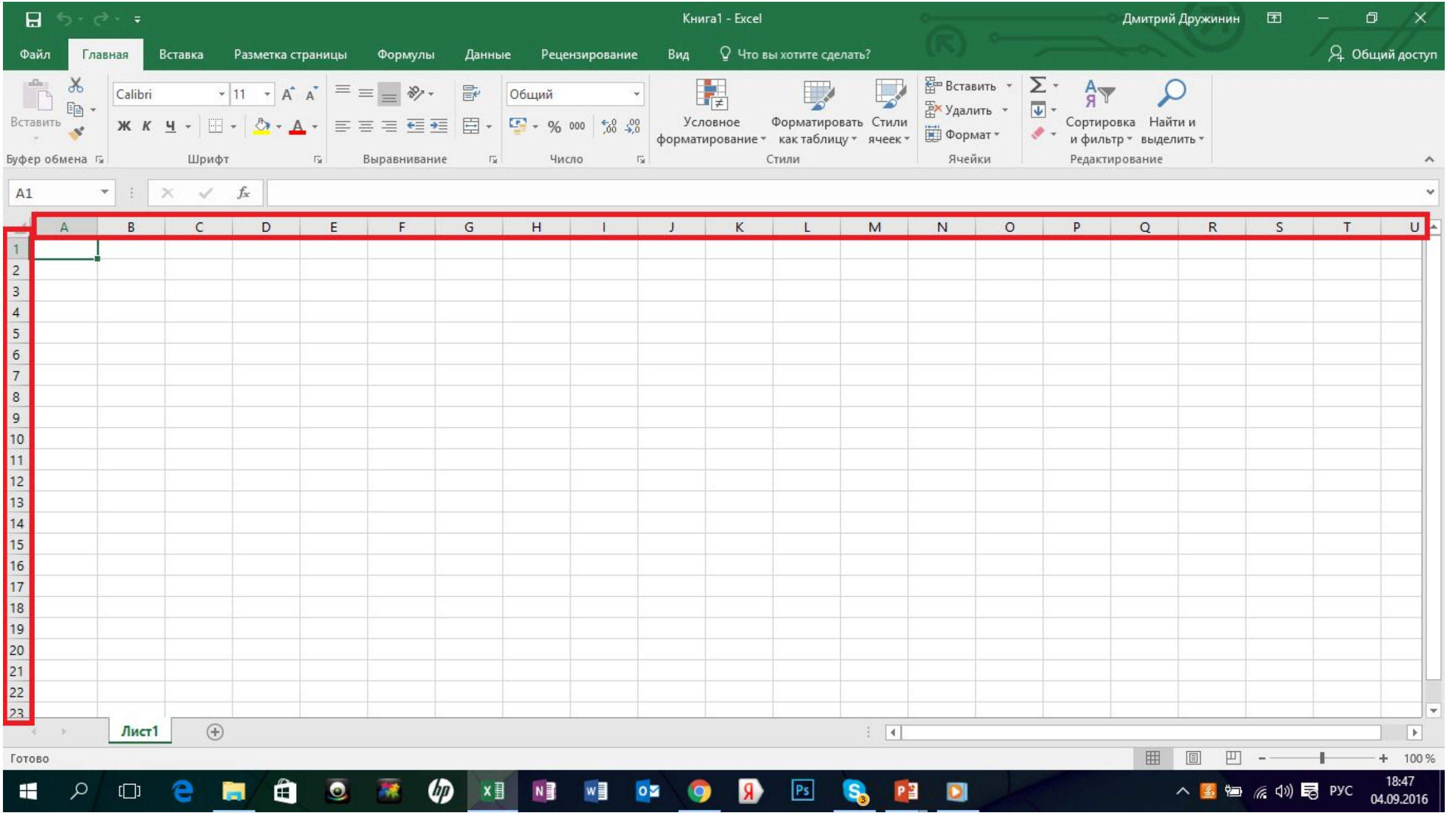
- 1) ***Простой отбор*** – случайное извлечение объектов из генеральной совокупности с возвратом или без возврата.
- 2) ***Типический отбор,*** когда объекты отбираются не из всей генеральной совокупности, а из ее «типической» части.
- 3) ***Серийный отбор*** – объекты отбираются из генеральной совокупности не по одному, а сериями.
- 4) ***Механический отбор*** - генеральная совокупность «механически» делится на столько частей, сколько объектов должно войти в выборку и из каждой части выбирается один объект.

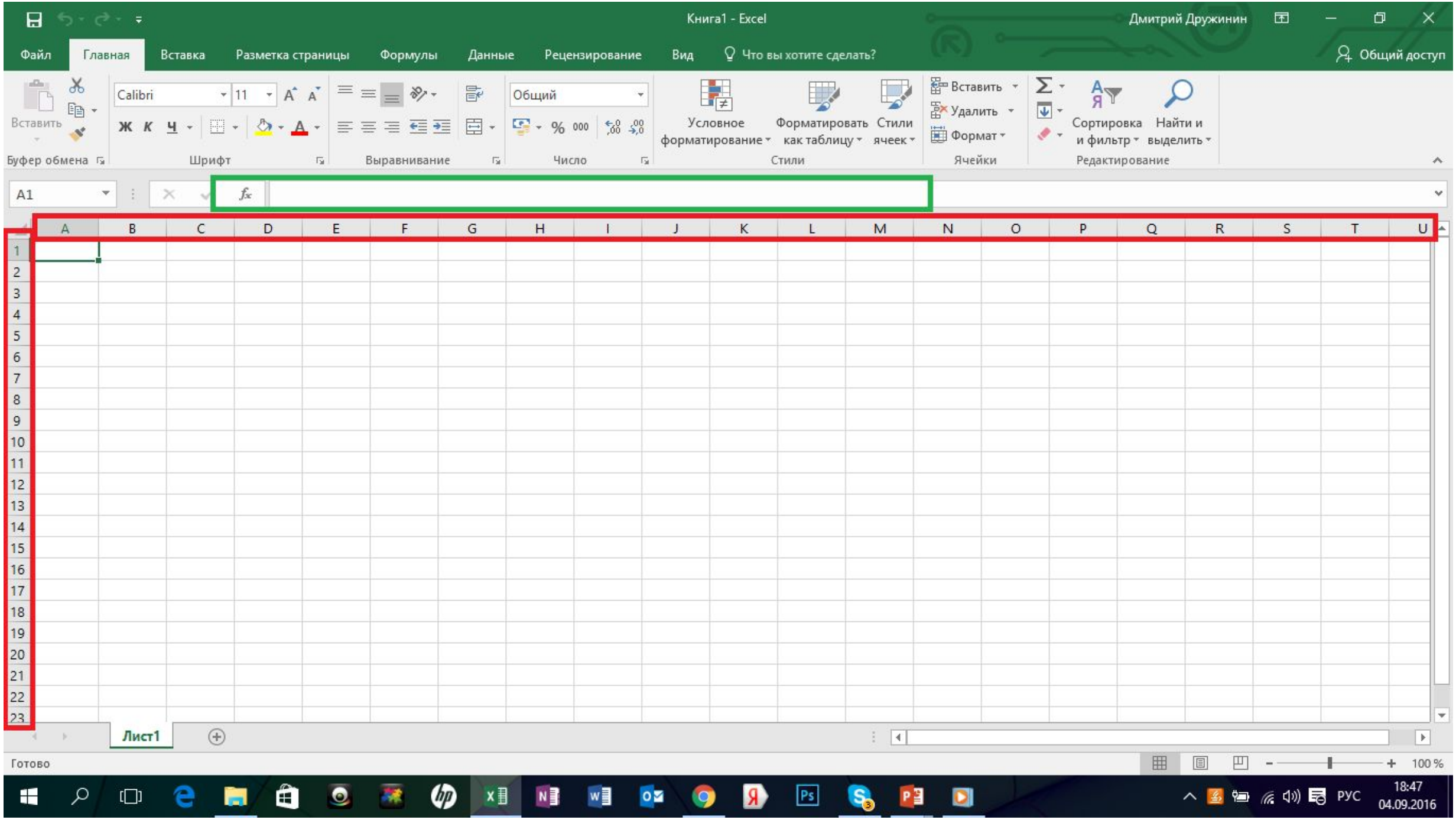
Основные понятия

- Цифровое значение, имеющие соответствующее смысловое значение называется **вариантом.**
- Последовательный алгоритм представляющий варианты в порядке их возрастания или убывания – **ранжирование.**
- Последовательность вариантов, записанных в возрастающем порядке, называется ***вариационным рядом*.**
- Число, которое показывает, сколько раз встречаются соответствующие значения вариантов в ряде наблюдений, называется ***частотой*** или **весом варианта.**
- Отношение частоты данного варианта к общей сумме частот называется относительной частотой или *частостью* (*долей*) соответствующего варианта


Книга1 - Excel
Дмитрий Дружинин
Файл
Главная
Вставка
Разметка страницы
Формулы
Данные
Рецензирование
Вид
Что вы хотите сделать?
Общий доступ
Вставить
Calibri
11
Общий
Условное форматирование
Форматировать как таблицу
Стили ячеек
Вставить
Удалить
Формат
Сортировка и фильтр
Найти и выделить
Буфер обмена
Шрифт
Выравнивание
Число
Стили
Ячейки
Редактирование
A1
Не должно быть пустых строк и пустых колонок
Вначале записываются заголовки, которые лучше всего выделить цветом для лучшего понимания
Набор данных отделяется от других данных пустым полем
Не стоит забывать о «скрытых» колонках, которые могут не отображаться, но сохраняются в анализе
Не должно быть дополнительных записей и данных, не включенных в структуру данного анализа
Лист1
Готово
100 %
РУС
18:47
04.09.2016


Книга1 - Excel
Дмитрий Дружинин
Файл
Главная
Вставка
Разметка страницы
Формулы
Данные
Рецензирование
Вид
Что вы хотите сделать?
Общий доступ
Вставить
Calibri
11
Общий
Условное форматирование
Форматировать как таблицу
Стили ячеек
Вставить
Удалить
Формат
Сортировка и фильтр
Найти и выделить
Буфер обмена
Шрифт
Выравнивание
Число
Стили
Ячейки
Редактирование
A1
Количественные переменные – обозначающие цифровое значение в выборке
Номинативные переменные – т е те которые обозначают смысл, кодированный в цифровом выражении
Лист1
Готово
100 %
РУС
18:47
04.09.2016


Книга1 - Excel
Дмитрий Дружинин
Файл
Главная
Вставка
Разметка страницы
Формулы
Данные
Рецензирование
Вид
Что вы хотите сделать?
Общий доступ
Вставить
Буфер обмена
Calibri
11
Шрифт
Выравнивание
Общий
Число
Условное форматирование
Форматировать как таблицу
Стили ячеек
Стили
Вставить
Удалить
Формат
Ячейки
Сортировка и фильтр
Найти и выделить
Редактирование
A1
Лист1
Готово
100 %
РУС
18:47
04.09.2016


Книга1 - Excel
Дмитрий Дружинин
Файл
Главная
Вставка
Разметка страницы
Формулы
Данные
Рецензирование
Вид
Что вы хотите сделать?
Общий доступ
Вставить
Calibri
11
Общий
Условное форматирование
Форматировать как таблицу
Стили ячеек
Вставить
Удалить
Формат
Сортировка и фильтр
Найти и выделить
Буфер обмена
Шрифт
Выравнивание
Число
Стили
Ячейки
Редактирование
A1
Лист1
Готово
100 %
РУС
18:47
04.09.2016

Описательная статистика

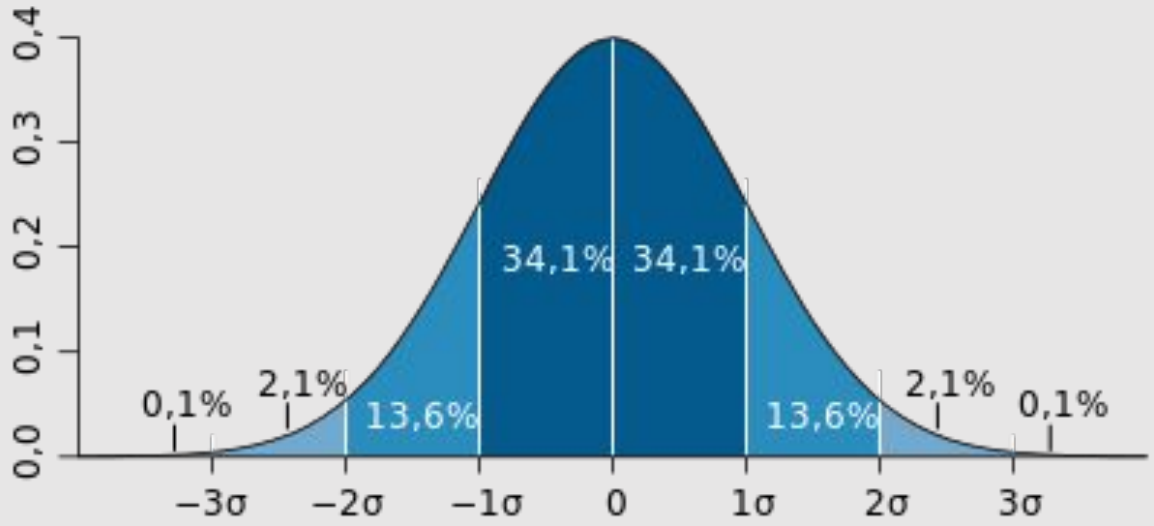
- **Среднее значение** – среднее арифметическое из группы чисел
- **Стандартная ошибка** –теоретическое стандартное отклонение всех средних выборки n, извлекаемых из генеральной совокупности N.

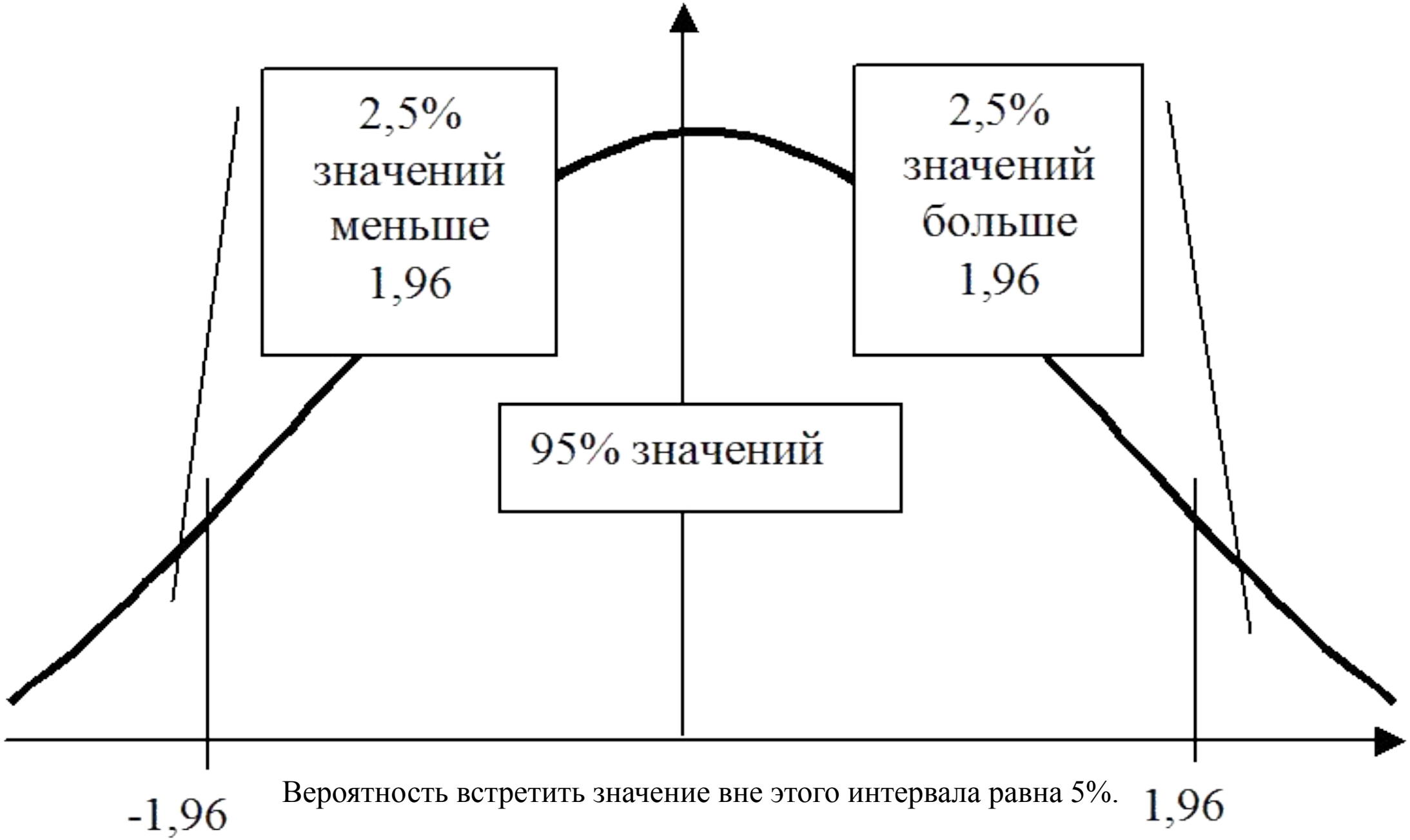
$$SEM = \frac{s}{\sqrt{n}},$$

- **Медиана** - это значение, которое разбивает выборку на две равные части. Половина наблюдений лежит ниже медианы, и половина наблюдений лежит выше медианы.
- **Мода** - описательная статистика, соответствующая значению признака, наиболее часто встречающемуся в исследуемой выборке. Подходит для описания дискретных, порядковых, номинальных данных.
  Не подходит для описания непрерывных данных. Мода может не существовать или быть не единственной.

Стандартное отклонение

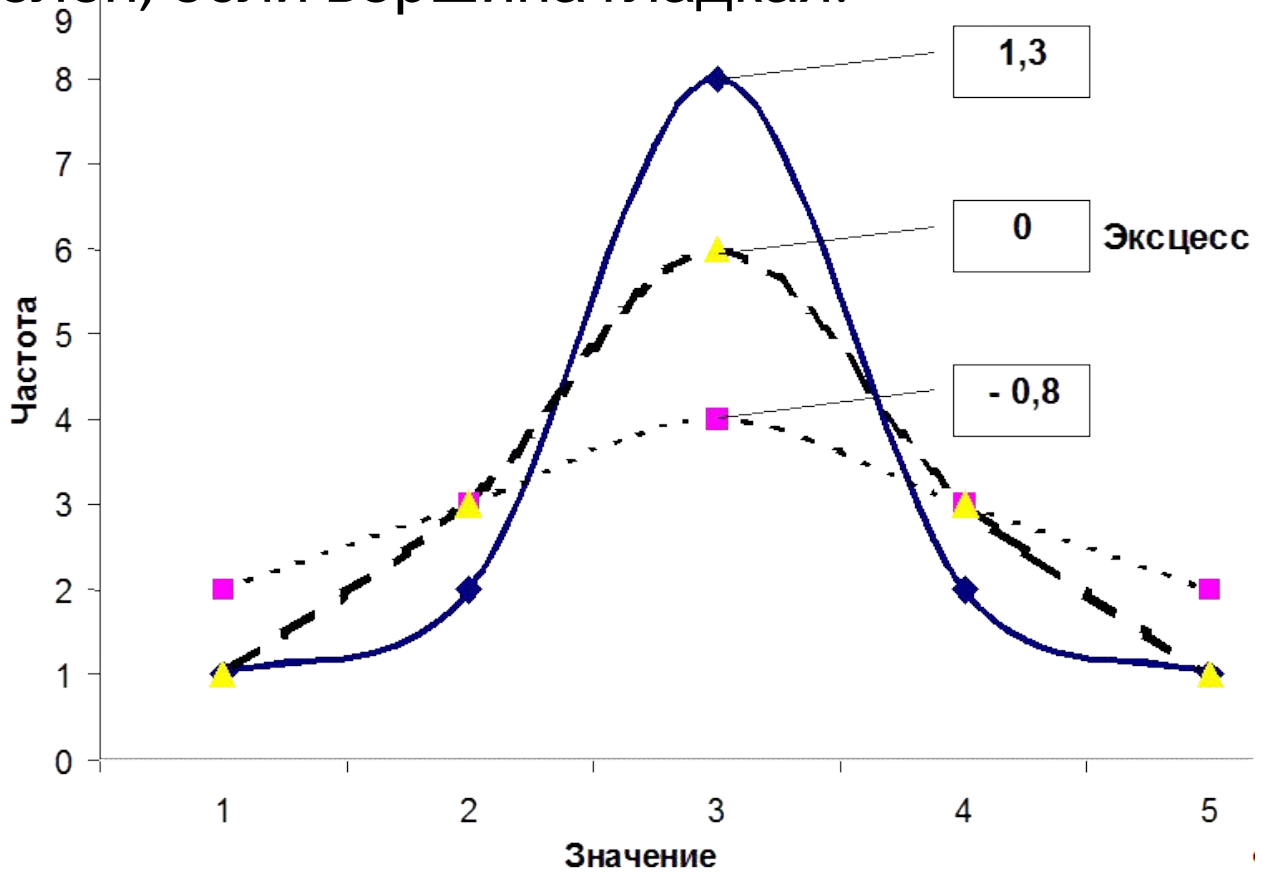
- Стандартное отклонение - в теории вероятностей и статистике наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания. При ограниченных массивах выборок значений вместо математического ожидания используется среднее арифметическое совокупности выборок.
- Большее значение среднеквадратического отклонения показывает больший разброс значений в представленном множестве со средней величиной множества; меньшее значение, соответственно, показывает, что значения в множестве сгруппированы вокруг среднего значения.



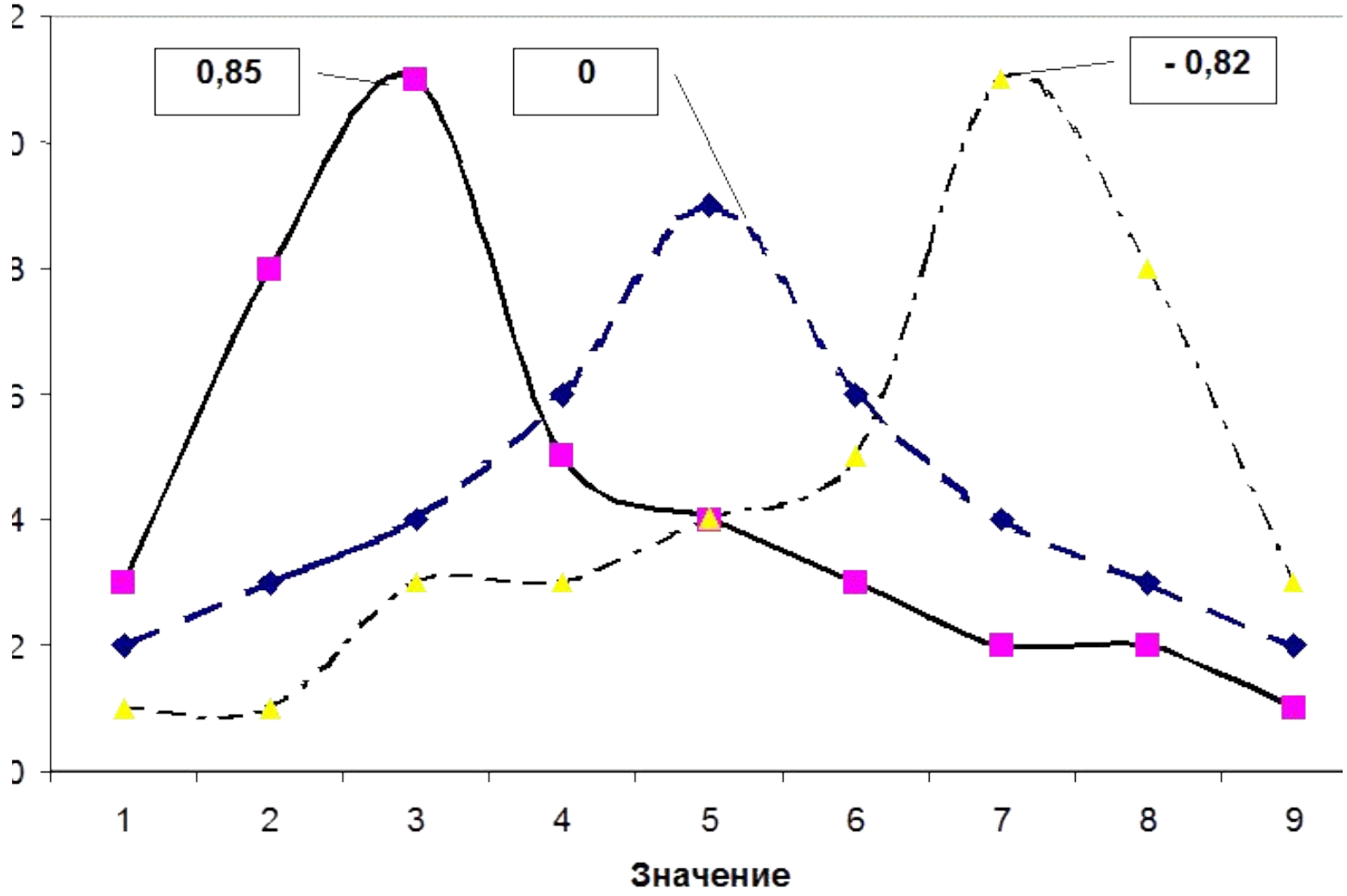


Вероятность встретить значение вне этого интервала равна 5%.

- **Дисперсия** – мера разброса случайной величины, т е ее отклонения от математического ожидания. Вычисляют как среднее арифметическое квадратов отклонения наблюдаемых значений.
- **Коэффицие́нт эксце́сса** (коэффициент островершинности) в теории вероятностей — мера остроты пика распределения случайной величины. Он положителен, если пик распределения около математического ожидания острый, и отрицателен, если вершина гладкая.



- ***Ассиметричность*** - также называют «скос» или «асимметрия». Статистика указывает на сдвиг вершины распределения влево или вправо от среднего значения. Если распределение строго симметрично, то асимметрия равна 0.



Интервал

- **Интервал** — это значения варьирующего признака, лежащие в определенных границах. Каждый интервал имеет верхнюю и нижнюю границы или одну из них. ***Нижней границей*** называется наименьшее значение признака в интервале. ***Верхней границей*** выступает наибольшее значение признака в интервале. ***Величина интервала*** представляет собой разность между верхней и нижней границами интервала.

Проверка гипотезы о виде распределения. (Критерий согласия)

- При получении выборки, закон распределения значений параметра заранее не известен, но есть основания предположить, что он имеет определенный вид (назовем его вид *A*).

- В таких случаях используют критерий согласия, в котором формулируют следующую нулевую гипотезу:

H_0 – *параметр генеральной совокупности распределен по закону A.*

- Для проверки гипотезы используют критерий Колмогорова-Смирнова или Пирсона

1.Формулировка гипотез

Задача критерия согласия - проверить, согласуются ли имеющиеся данные с тем или иным видом распределения (чаще, с нормальным).

Основная гипотеза (H_0)

Различия между имеющимися данными и теоретическим распределением случайны

Альтернативная гипотеза (H_1)

Различия между имеющимися данными и теоретическим распределением не случайны

2. Определение уровня значимости

Пусть уровень значимости равен 0,05 (5%)

Критерии согласия для нормального распределения

- **Критерий согласия Колмогорова** предназначен для проверки гипотезы о принадлежности выборки некоторому закону распределения, то есть проверки того, что эмпирическое распределение соответствует предполагаемой модели.
- Назначение критерия заключается в том, что он определяет, относятся ли сравниваемые вами два распределения к одному и тому же типу. Если мы будем сравнивать экспериментально полученное распределение с нормальным распределением, то с помощью критерия сможем получить ответ о том, нормально ли наше распределение.

Тест Колмогорова Смирнова

- Полученные результаты включают:
- среднее значение и стандартное отклонение
- промежуточные результаты, полученные в результате теста Колмогорова-Смирнова
- вероятность ошибки p.
- Отклонение от нормального распределения считается существенным при значении $p < 0,05$; в этом случае для соответствующих переменных следует применять **непараметрические тесты**. В рассматриваемом примере (значение $p = 0,616$), то есть вероятность ошибки не является значимой; поэтому значения переменной достаточно хорошо подчиняются нормальному распределению и можно применять **параметрические тесты**.

Критерий Лиллиефорса

- **Критерий Лиллиефорса** — статистический критерий, названный по имени Хьюберта Лиллиефорса, профессора статистики Университета Джорджа Вашингтона, являющийся модификацией критерия Колмогорова–Смирнова.
- Используется для проверки нулевой гипотезы о том, что выборка распределена по нормальному закону для случая, когда параметры нормального распределения (математическое ожидание и дисперсия) априори неизвестны.

Критические значения для Z-критерия

Критические значения $\lambda_{\alpha 0}$ для распределений:	Уровень значимости α			
	0,20	0,10	0,05	0,01
Распределение Колмогорова	1,073	1,224	1,358	1,627
Распределение Лиллиефорса	0,736	0,805	0,886	1,031

Понятие статистического критерия

Параметрические

Применимы к выборкам, извлеченным из нормально распределенных ГС.

Необходимо доказать с помощью критериев согласия.

Непараметрические

Применимы к выборкам, извлеченным из ГС, имеющих распределения отличные от нормального.

Статистический критерий.

Статистический критерий – правило, в соответствии с которым принимается или отклоняется нулевая гипотеза.

Статистика критерия – специально выработанная случайная величина, функция распределения которой известна (Стьюдента, Фишера, Пирсона Гаусса).

Критическая область – совокупность значений критерия, при которых нулевую гипотезу отвергают.

Область принятия гипотезы – совокупность значений критерия, при которых нулевую гипотезу принимают.

Критические значения критерия – это точки, отделяющие критическую область от области принятия гипотезы.

Наблюдаемое значение критерия - значение критерия, вычисленное по данным выборки.