**Мой день 1** 10.05.2020

**Уроки: 1.1-1.5**

* **Генеральная** **совокупность** (от лат. generis — общий, родовой) — **совокупность** всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи. Далее ***ГС***
* **Репрезентативная** **выборка** – это такая **выборка**, в которой все основные признаки генеральной совокупности, из которой извлечена данная **выборка**, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности.
* **Унимодальное** - это распределение имеющее только одну моду (т.е. один "пик"). Типичный пример - это нормальное распределение.

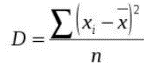
**Способы репрезентативной выборки:**

* Простая случайная выборка (simple random sample)
* Стратифицированная выборка (stratified sample) – разделение *ГС* на страты (группы) а оттуда уже делается случайная выборка.
* Групповая выборка (cluster sample) – похожие группы выбираются из выборки и далее делается случайная выборка (например, районы одного города)

**Типы переменных:**

* Количественные – измеряемое (например, рост):
  + *Непрерывные* – переменная принимает любое значение на опр. промежутке;
  + *Дискретные* – только определенные значения (3.5 ребенка в семье **не** будет).
* Номинативные (= качественные) – разделение испытуемых на группы, цифры как маркеры (например: 1 -женщины, 2 – мужчины). Цифры как имена групп, не для расчетов.
* Ранговые – похоже на номинативные, только возможны сравнения (быстрее/медленнее и т.п.)

**Описательная статистика, типы:**

* Меры центральной тенденции – узкий диапазон, высокие значения:
  + ***Мода*** *(mode)* – значение измеряемого признака, которое встречается максимально часто. График [Dot Plot (точечный график)](https://www.onlinemath4all.com/images/analyzingdotplots3.png) поможет;
  + ***Медиана*** *(median)* – значение признака, которое делит упорядоченное (по возрастанию/убыванию) множество данных пополам. Если в середине 2 значения оказываются в случае с как с четными числами, лорроллгорвемцаб***Среднее значение*** – сумма всех значений измеренного признака делится на количество измеренных значений. (переменная = «Х» с черточкой «-» вверху используется для среднего значения из выборки, а для *генеральной совокупности* латинская буква М). [**Свойства.**](https://stepik.org/lesson/8075/step/12?discussion=1181787&unit=1355)
* Меры изменчивости – вариативность, широкий диапазон, значения менее высокие как правило, чем в первом случае:
  + *Размах (range)* – разность максимального и минимального значения (R = X max – X min), недостаток способа в том, что при добавлении сильно отличающегося значения данные меняются сильно и могут быть некорректные;
  + [*Дисперсия (variance)*](https://univer-nn.ru/statistika/dispersiya/)– средний квадрат отклонений индивидуальных значений признака от их средней величины также n-1. И выведем ***среднее (стандартное) отклонение* sd** = квадратный корень дисперсии. [Свойства дисперсии](https://ucarecdn.com/fef17799-e6e3-4fbd-ba1e-9a4ea5df7817/2568_720_746e6e7e2d8955290de1daf72a5f1b74.mp4): если добавить новое значение после расчетов, то дисперсия (D) и среднее отклонение (sd) не изменяться. А если умножить, то изменяться на умноженное кол-во раз. 

**Мой день 2** 11.05.2020

**Уроки: 1.6-1.9**

* **Квартили распределения**
  + *Квартили* – три точки (значения признака), которые делят упорядоченное множество данных на четыре равные части ( медиана (или 2-й квартиль) на 2 части делит одной точкой, а тут на 4 части 3-мя точками):
    - [*Box plot*](https://edwvb.blogspot.com/2018/02/kak-chitat-diagrammu-boxplot-yashchik-s-usami-na-konkretnom-primere.html) - такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.
* **Нормальное распределение, это:**
  + *Унимодальное*
  + *Симметричное*

[*Стандартизация или z-преобразование*](https://ru.wikihow.com/%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D1%8C-Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D1%83)- преобразование полученных данных в стандартную Z-шкалу (Z-scores) со средним значением = 0 и дисперсией = 1. Чтобы привести к такому виду из каждого наблюдения нужно отнять среднее значение и разделитьна стандартное отклонение.

[Помогает выявить](https://gallery.shinyapps.io/dist_calc/) какой процент наблюдений лежит в абсолютно любом диапазоне, на python: import scipy.stats as st

print(1 - st.norm.cdf(0.5))

**Правило «двух» и «трех» сигм:**

* +-1 сигма = примерно 68% выборки
* +-2 сигмы = примерно 95% выборки
* +-3 сигмы = примерно 100% выборки

**Центральная предельная теорема –** [понятное видео](https://youtu.be/uNDBo3K0ILY) объяснение понятия

[Сайт с симуляцией данных для центральной предельной теоремы](https://gallery.shinyapps.io/CLT_mean/)

* *Стандартная ошибка среднего* – стандартное отклонение признаков ***ГС*** деленное (среднее выборочное) на корень числа наблюдений.

[**Доверительные интервалы для среднего**](https://stepik.org/lesson/8079/step/3?unit=1359)– 95% средних значений имеют интервал 1.98 отклонением который включает в себя *ГС.* Подробное объяснение с формулами и на примерах [тут](https://ru.wikihow.com/%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D1%8C-%D0%B4%D0%BE%D0%B2%D0%B5%D1%80%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9-%D0%B8%D0%BD%D1%82%D0%B5%D1%80%D0%B2%D0%B0%D0%BB).

**Мой день 3** 12.05.2020

**Уроки: 1.10**

**Идея статистического вывода**

Имеем: М (среднее ГС), N (кол-во выборки, уч астников), Х (предполагаемое среднее значение выборки), sd (среднее статистическое отклонение).

М = 20, N = 64, Х = 18.5, sd = 4

H0 = 20 (нулевая гипотеза, о том, что изменений не было, никаких различий или взаимосвязи с ГС нет)

H1 != 20 (альтернативная гипотеза, что истинна)

Рассчитываем среднюю ошибку se = sd / корень n = 4/корень 64 = 0.5

Z = x – m / se = 18.5 – 20 / 0.5 = -3

[Расчёт р - уровня](https://gallery.shinyapps.io/dist_calc/) (р = вероятность события) рассчитываем на сайте, указывая параметры выше и получаем уровень значимости **p = 0.003** при гипотезе Н0, принято, что если значение ниже 0.05 то нулевая гипотеза ложная. Значение p-уровня значимости, которое выбирается, в качестве порога обозначается буквой *α* (альфа) и при строгости и для точности может быть 0.001. [Подробнее.](https://habr.com/ru/company/stepic/blog/250527/)

Ошибка первого и второго рода – см [тут](https://ucarecdn.com/7fbb0394-b60e-410d-85ff-2aa0fd2f7e60/).

**Рекомендовано к ознакомлению + полезные ссылки.**

1. С. Гланц. Медико-биологическая статистика. (<http://medstatistic.ru/articles/glantz.pdf>)  
   Хороший учебник по введению в статистику на русском языке со множеством примеров из биологии и медицины.
2. Jerrold H. Zar.  Biostatistical Analysis.  
   Более подробный учебник, в котором очень тщательно разбирается большинство статистических методов для анализа биологических данных.
3. OpenIntro Statistics (<https://www.openintro.org/stat/textbook.php>)  
   Замечательный учебник с массой примеров из различных областей, а также видео лекции.
4. ﻿Отличный учебник для первых шагов в области анализа данных: [Статистика и котики](https://www.litres.ru/vladimir-savelev-10569666/statistika-i-kotiki/). Несмотря на несерьезное название в учебнике очень понятно и доступно объясняются основные темы, необходимые для уверенного использования статистики. ﻿
5. Группa в контакте (<https://vk.com/introstats>), в которой можно общаться на тему курса и вообще статистики.
6. <https://vk.com/meaning_data> - группа ВК по аналитике
7. [Лекция ТЕД к уроку 1.4 шаг 10](https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen)
8. [Сайт программа для статистики](https://www.gapminder.org/tools/?from=world#$chart-type=bubbles), можно ознакомиться или создать самой используя встроенные инструменты
9. [Специальный сайт](https://gallery.shinyapps.io/dist_calc/), позволяющим вычислить процент наблюдений в интересующем нас диапазоне (по умолчанию выставлено стандартное нормальное распределение M=0, sd=1)
10. [Сайт](https://gallery.shinyapps.io/dist_calc/) для расчета p – уровня