

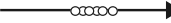

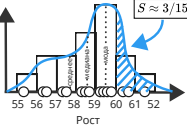


# Занятие 1

## Элементы статистики и визуализации данных.

<p><b>Выборка</b></p>  <p><i>Случайная выборка ростов из <math>n = 6</math> людей</i></p>	<p><b>Генеральная совокупность</b></p>  <p><i>Рост всех <math>N</math> людей на Земле</i></p>
<p><b>Выборочное среднее</b></p> $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$	<p><b>Генеральное среднее</b></p> $\overline{x_{general}} = \frac{x_1 + x_2 + \dots + x_N}{N}$
<p><b>Выборочное стандартное отклонение</b></p> <p><b>Смещённая оценка</b> (показывает разброс выборочных данных от выборочного среднего)</p> $\sigma_{biased} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ <p><b>Несмещённая оценка</b> (оценивает разброс выборочных данных от генерального среднего).</p> $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \approx \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x_{general}})^2}$	<p><b>Генеральное стандартное отклонение</b></p> $\sigma_{general} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \overline{x_{general}})^2}$ <div style="border: 1px dashed black; padding: 5px; margin-top: 10px;"> <p><i>Низкое стандартное отклонение</i></p>  <p><i>Высокое стандартное отклонение</i></p>  </div>
<p><b>Стандартная ошибка среднего</b></p> $S = \frac{\sigma_{general}}{\sqrt{n}} \approx \frac{\sigma}{\sqrt{n}}$ <div style="display: flex; align-items: center; margin-top: 10px;"> <div style="border: 1px solid black; padding: 5px; margin-right: 10px;">Допущение, необходимое для расчёта доверительного интервала.</div> <div style="text-align: center;"> <math>\sigma_{general}</math> мы почти никогда не знаем         </div> </div>	
<p><b>Распределение</b> — это расположение данных на числовой прямой.</p> <p><b>Функция плотности распределения</b> (синяя прямая) показывает кривую, площадь под любым участком которой равна вероятности того, что случайно выбранный элемент выборки будет из этого промежутка.</p> <p><b>Медиана</b> — число, которое делит упорядоченную последовательность чисел пополам.</p> <p><b>Мода</b> — число, соответствующее пику распределения.</p>	 <p><i>На картинке показано соотношение между средним, медианой и модой, характерное для скошенного влево распределения.</i></p>
<p><b>Центральная предельная теорема</b></p> <p>Если сделать много выборок размером <math>n</math>, то распределение выборочных средних будет нормальным со средним <math>\overline{x_{general}}</math> и стандартным отклонением <math>S</math>.</p>	
<p><b>Доверительный интервал по нормальному распределению:</b></p> <p>Среднее значение признака <math>x</math> в генеральной совокупности с <math>(1 - \alpha)</math> вероятностью входит в интервал:</p> $\overline{x_{general}} = \bar{x} \pm z(\alpha) \cdot \frac{\sigma}{\sqrt{n}}$ <div style="display: flex; justify-content: flex-end; margin-top: 10px;"> <div style="margin-right: 20px;"><math>z(0.05) = 1.96</math></div> <div><math>z(0.01) = 2.58</math></div> </div>	
<p><b>Доверительного интервал по распределению Стьюдента:</b></p> <p>Среднее значение признака в генеральной совокупности с <math>(1 - \alpha)</math> вероятностью входит в интервал:</p> $\overline{x_{general}} = \bar{x} \pm t(\alpha, n) \cdot \frac{\sigma}{\sqrt{n}}$ <div style="display: flex; justify-content: flex-end; margin-top: 10px;"> <div style="margin-right: 20px;"><math>t(0.05, 5) = 2.77</math></div> <div style="margin-right: 20px;"><math>t(0.05, 10) = 2.23</math></div> <div style="margin-right: 20px;"><math>t(0.05, 30) = 2.04</math></div> <div><math>t(0.05, \infty) = 1.96</math></div> </div>	

## Занятие 1

### Элементы статистики и визуализации данных.

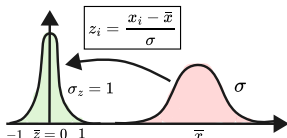
#### Стандартизация численных данных

##### 1) StandardScaler()

(Трансформирует данные так, чтобы среднее стало 0, а стандартное отклонение 1)

	$x$	$z$
1	22.5	$\frac{22.5 - \bar{x}}{\sigma}$
...	...	...
n	45.6	$\frac{45.6 - \bar{x}}{\sigma}$

$\bar{x}$  и  $\sigma$



Статья о причинах, по которым необходимо проводить Стандартизацию



**Физический смысл:** значение  $z_i$  представляет собой отклонение  $x_i$  от  $\bar{x}$ , выраженное в стандартных отклонениях величины  $x$ .

#### Зачем проводить стандартизацию?

1. Чтобы значения переменной приобрели физический смысл и позволили сравнивать отдельные наблюдения между собой по удалению от среднего.
2. Чтобы коэффициенты линейных моделей (см. занятия 2 и 3) приобрели физический смысл и позволили сравнивать признаки по силе влияния на целевую переменную.
3. Линейные модели и нейронные сети быстрее обучаются и равномернее сходятся, если фичи стандартизованы. Более того, есть [подтверждения](#) факта, что стандартизация признаков увеличивает качество популярного метода понижения размерности PCA (см. QR-код), следовательно, и качество классификации и регрессии, использующих этот метод.

#### Джойны

**Inner Join** — совокупность поставленных рядом строк левой и правой таблицы, в которых совпадает некоторый идентификатор.

**Left Join** — то же самое, что и Inner Join + те записи из левой таблицы, для которой в правой по указанному идентификатору ничего не нашлось.

T1

id	Name	Age
5236	John	25
2362	William	19
7425	Jack	38

T2

id	medical test results
5236	positive
6436	positive
5125	negative

T1 inner join T2

id	Name	Age	medical test results
2362	William	19	positive

T1 left join T2

id	Name	Age	medical test results
5236	John	25	NaN
2362	William	19	positive
7425	Jack	38	NaN

#### Кодировка номинативных признаков

One-Hot-Encoding позволяет привести номинативную переменную с множеством значений в совокупность бинарных признаков.

	Цвет
0	синий
1	белый
2	зелёный



	синий	белый	зелёный
0	1	0	0
1	0	1	0
2	0	0	1

Колонку с зелёным цветом необходимо удалить, поскольку этот признак не привносит в данные новую информацию

Поскольку мы знаем, что цветов может быть всего 3, и каждый элемент обязательно либо синий, либо белый, либо зелёный, то, если элемент не синий и не белый, то он зелёный.

# Занятие 1

## Приложение

### 1. Почему в стандартном отклонении выборки принято делить на (n-1)?

Деление на (n-1) вместо n в формуле несмещённого стандартного отклонения происходит по той причине, что данная формула оценивает стандартное отклонение выборки относительно генерального среднего. Слово «смещение» означает смещение относительно генерального среднего.

То есть, **среднее значение множества выборочных «несмещённых стандартных отклонений» будет равно генеральному стандартному отклонению.**

Путём кропотливых математических преобразований (см. QR-код справа), можно показать, что дисперсия выборочных значений относительно генерального среднего состоит из дисперсии выборочных значений относительно выборочного среднего (biased) + разброс выборочного среднего относительно генерального:

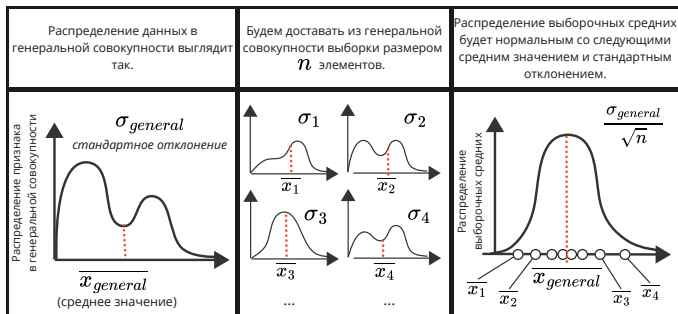
$$\sigma^2 = \sigma_{biased}^2 + \frac{\sigma^2}{n} \quad \sigma_{biased} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\sigma = \frac{n}{n-1} \cdot \sigma_{biased} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Вы можете ознакомиться с более подробным обсуждением этого вопроса в треде на Quora (необходимо подключение к VPN)



### 2. Центральная предельная теорема. Визуализация.



### 3. Как через доверительный интервал для выборочного среднего, вычисленный на основе центральной предельной теоремы, выразить доверительный для среднего генеральной совокупности?

Поскольку характеристики нормального распределения известны, то мы можем посчитать, что с  $(1 - \alpha)$  вероятностью выборочное среднее попадёт в интервал.

$$\bar{x}_i = \bar{x}_{general} \pm z(\alpha) \cdot \frac{\sigma_{general}}{\sqrt{n}}$$

$z(\alpha)$  — постоянная величина

Поменяем  $\bar{x}_{general}$  и  $\bar{x}_i$  местами, а также учтём, что  $\sigma_{general} \approx \sigma_i$ , получим, что с  $(1 - \alpha)$  вероятностью генеральное среднее попадёт в интервал:

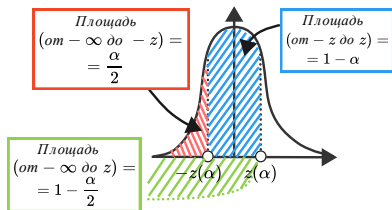
$$\bar{x}_{general} = \bar{x}_i \pm z(\alpha) \cdot \frac{\sigma_i}{\sqrt{n}}$$

# Занятие 1

## Приложение

### 4. Расчёт $z$ и $t$ статистик через квантили распределения.

$z(\alpha)$  — значение под графиком функции плотности вероятности стандартного нормального распределения, соответствующее уровню значимости  $\alpha$  (см. рисунок)



Таким образом, мы можем посчитать статистику распределений для вычисления доверительного интервала через квантили этих распределений, т.к. квантиль распределения по функции плотности вероятности распределения считается так:

$$q(P) = \text{такое число, что Площадь(от } -\infty \text{ до этого числа) } = P$$

Получим:

$$z(\alpha) = q_{normal}\left(1 - \frac{\alpha}{2}\right)$$

$$t(\alpha, n) = q_{student}\left(1 - \frac{\alpha}{2}, n\right)$$

Вычисление квантилей забито во многие статистические пакеты. Для Python это пакет stats, внутри которого для распределений есть метод .ppf (percentile point function).

`scipy.stats.norm.ppf(0.95)`

`scipy.stats.t.ppf(0.95, n=30)`