KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 2 урок > СТАТИСТИКА

> Оглавление

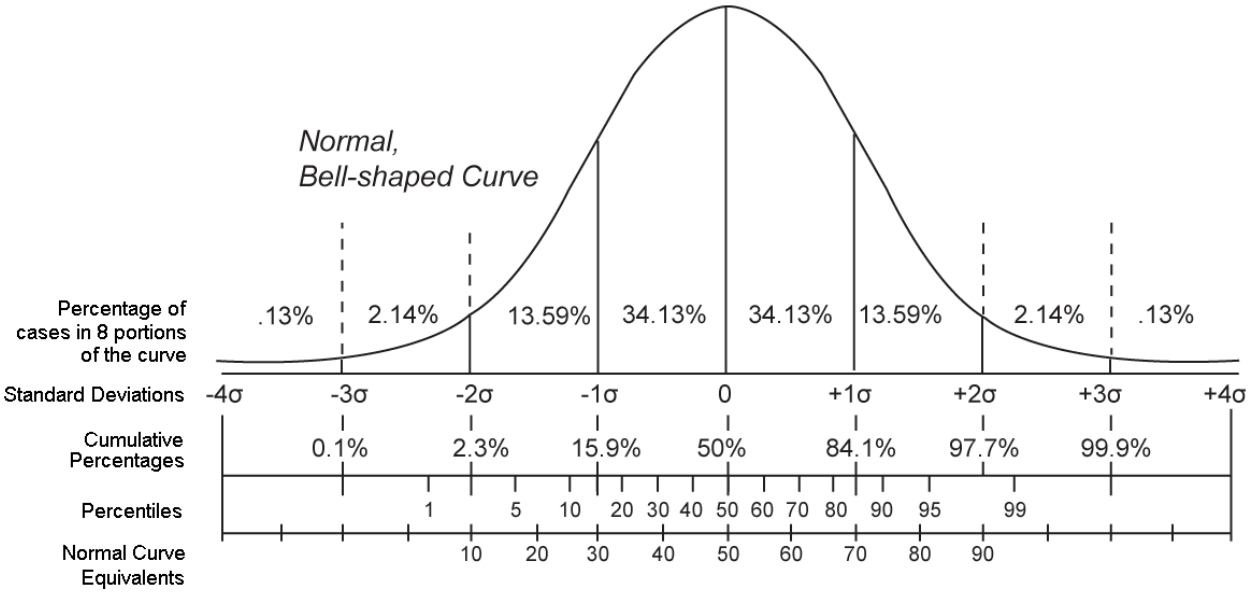
> Нормальное распределение

1. Унимодально
2. Симметрично
3. Отклонения подчиняются закону

Например:

- В диапазоне от среднего до 1σ (одного стандартного отклонения) будет находиться примерно 34.1% всех наблюдений
- В диапазоне от 1σ до 2σ – примерно 13.6%
- Очень маловероятно встретить наблюдение, которое бы превосходило среднее значение больше чем на 3 стандартных отклонения (3σ)

Также мы можем заметить, что отклонение от среднего равновероятно как в большую, так и в меньшую стороны.



> Стандартизация

Стандартизация (Z-преобразование) – преобразование, которое позволяет любую шкалу перевести в стандартную Z-шкалу (Z-scores), где среднее значение будет равно нулю, а стандартное отклонение – равняться 1 ($M_z = 0$, $D_z = 1$).Форма распределения при этом не изменится.

Таким образом, если мы из каждого наблюдения в нашей выборке отнимем среднее значение и разделим выражение на стандартное отклонение, то получим Z-шкалу, где новое среднее станет равно нулю, а дисперсия – единице.

$$Z_i = \frac{x_i - \overline{X}}{\sigma_x}$$

Как посчитать в python

```
from scipy.stats import zscore
zscore(df.A)
```

> Правило "двух" и "трех" сигм

- $M_x \pm \sigma \approx 68\%$ наблюдений находятся в этом интервале
- $M_x \pm 2\sigma \approx 95\%$ наблюдений находятся в этом интервале
- $M_x \pm 3\sigma \approx 100\%$ наблюдений находятся в этом интервале

Пример: Среднее значение равняется 150, а стандартное отклонение равно 8. Какой процент наблюдений превосходит значение, равное 154?

Для этого нужно сделать Z-преобразование. Как найти интересующее нас Z-значение? Из 154 нужно вычесть среднее значение по нашей выборке и разделить на стандартное отклонение (8). В результате:

$$\frac{154 - 150}{8} = \frac{4}{8} = 0.5$$

Воспользуемся специальной таблицей, которая предоставит нам ответ. Как читать эту таблицу?

- По вертикали находятся **целые и десятичные доли** z-значения
- По горизонтали - **сотые доли**
- Нужный процент находится на пересечении этих элементов z-значения. Например, если у нас получилось z-значение, равное 0.93, то нужный процент будет в строчке 0.9 и столбце 0.03 (.17619)
- Так как нормальное распределение симметрично, то знак z-значения не принципиален. Таблица ниже даёт одинаковые результаты как для отрицательных, так и для положительных z-значений.

В нашем случае видим, что в диапазоне превышающем 154 (или 0.5 в z-шкале), находится примерно 30% наших наблюдений. Иными словами, вероятность встретить значение, превосходящее 0.5 в z-шкале, составляет $\approx$ три десятых.

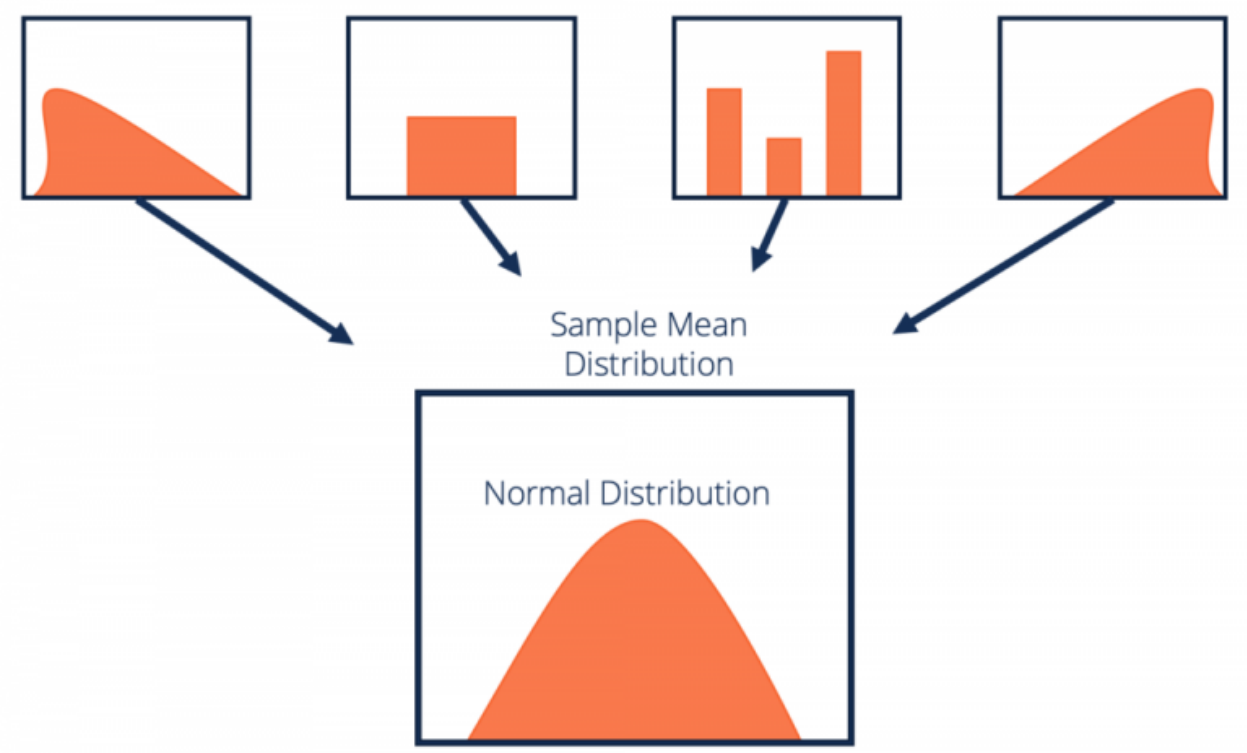
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

> Центральная предельная теорема

Предположим, что некоторый признак нормально распределен в генеральной совокупности (ГС), среднее = 0, стандартное отклонение = 15. Если мы будем

многократно извлекать выборки по N наблюдений из генеральной совокупности и внутри каждой выборки рассчитывать среднее значение и стандартное отклонение, то заметим, что распределение признака будет изменяться от выборки к выборке, при этом значения средних также будет варьироваться в положительную или отрицательную сторону.

Далее мы строим распределение выборочных средних значений. Если в каждой выборке оценка среднего не является точной, то как раз среднее всех средних будет очень близко к реальному среднему в генеральной совокупности. Большинство всех средних будет лежать рядом с нулем, а какие-то – отклоняться.



Стандартное отклонение этого распределения называется **стандартной ошибкой среднего**. Она показывает, насколько выборочные средние отклоняются от среднего ГС.

Если мы увеличим объем каждой из выборок, то распределение признака внутри каждой из групп станет больше похоже на распределение в ГС. Оценки также станут более точными, при этом стандартная ошибка тоже уменьшится.

Иными словами: при увеличении числа выборок и их размера уменьшается изменчивость выборочного распределения средних, и средние выборок будут

находиться ближе к реальному среднему ГС (закон больших чисел).

NB! Выше мы предположили, что ГС распределена нормально - что далеко не всегда так. Однако при достаточных размерах выборок и повторных их извлечений выборочное распределение средних всё равно будет нормальным! На этом факте базируются многие статистические тесты, о которых речь пойдёт далее.

Поиграться с ЦПТ и посмотреть, как из ненормальной ГС получается нормальное выборочное распределение, можно тут.

> Стандартная ошибка среднего

Стандартная ошибка среднего (SE) показывает, насколько выборочные средние "разбросаны" вокруг среднего генеральной совокупности. SE при увеличении размера выборки будет стремиться к нулю.

$$se = \frac{\sigma}{\sqrt{n}}$$

Если выборка репрезентативна и число наблюдений n ≥ 30, то в качестве стандартного отклонения ГС мы можем использовать стандартное отклонение нашей выборки:

$$se = \frac{sd_x}{\sqrt{n}}$$

Как посчитать:

```
import pandas as pd
df.A.sem()
```

```
from scipy import stats
stats.sem(df.A)
```