Introduction to Data Science

Dr. Irfan Yousuf
Department of Computer Science (New Campus)
UET, Lahore

(Week 2; January 22 - 26, 2024)

Normal Distribution

Normal Distribution Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x

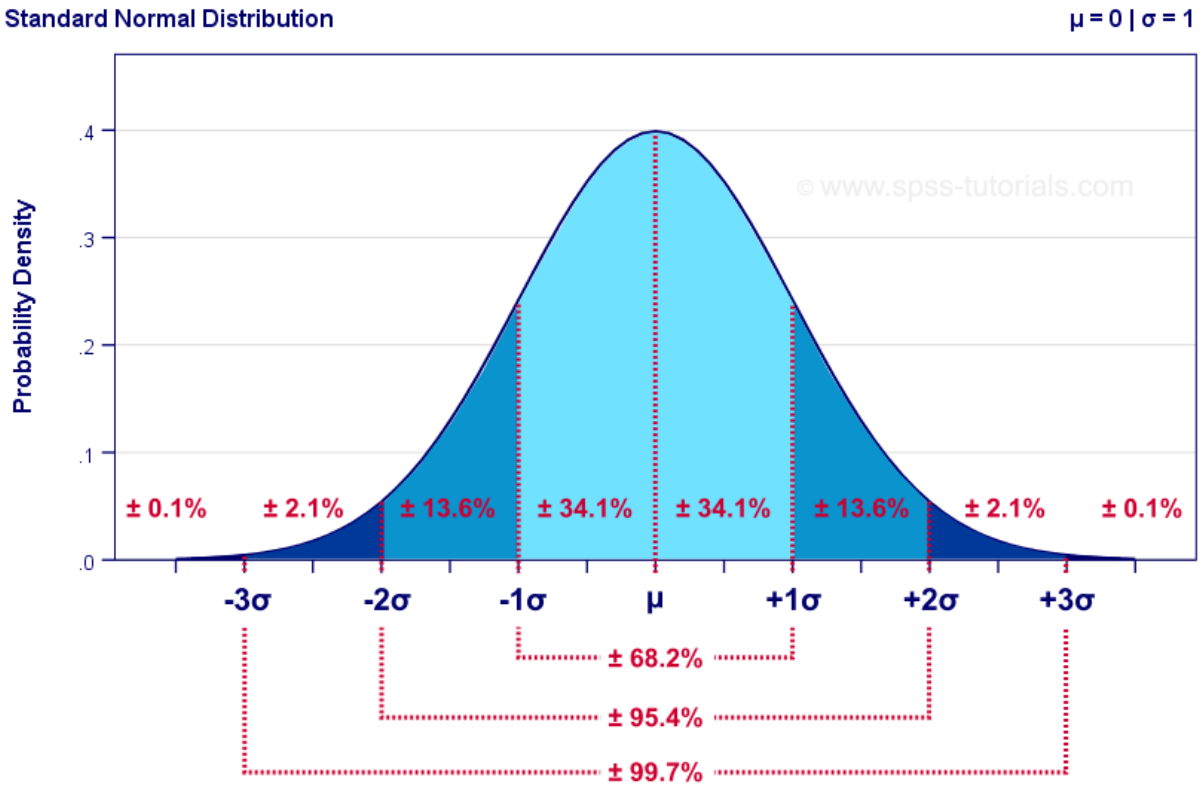
$\sigma =$ standard deviation of x

$\pi \approx 3.14159 \ldots$

$e \approx 2.71828 \ldots$

Normal Distribution

- The mean, median and mode are **exactly the same**.
- The distribution is **symmetric about the mean**—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: **the mean and the standard deviation**.

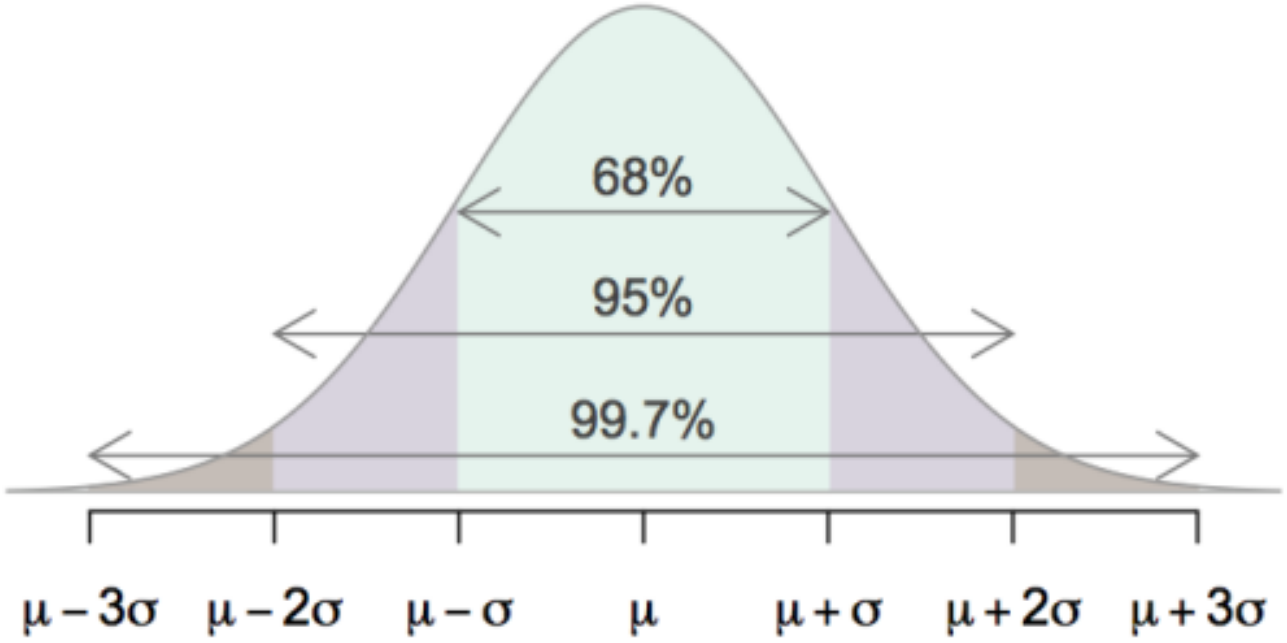


68-95-99.7 Rule

For nearly normally distributed data,

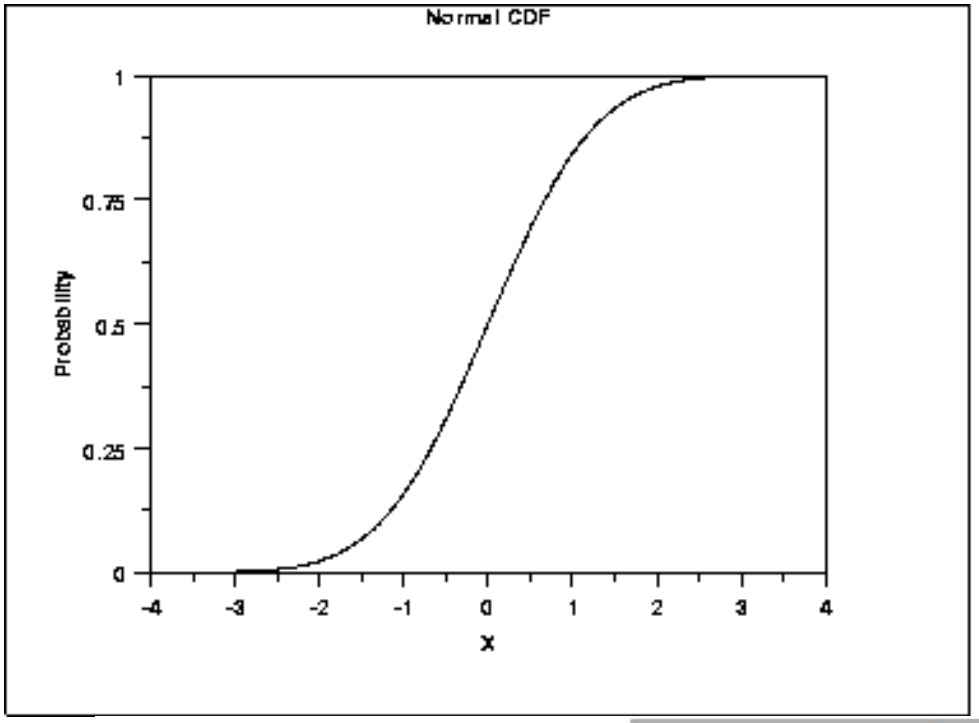
- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



CDF of Normal Distribution

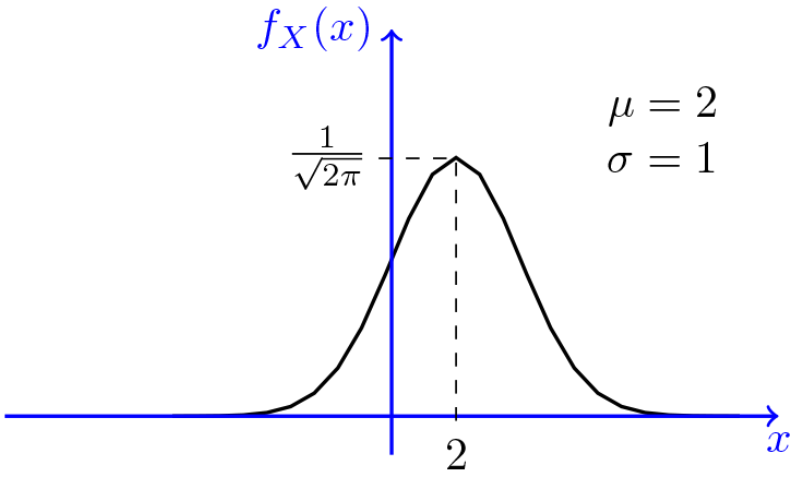
- The cumulative distribution function (cdf) is the probability that the variable X takes a value less than or equal to x.
- (Here in the figure below, Mean=0, SD=1)

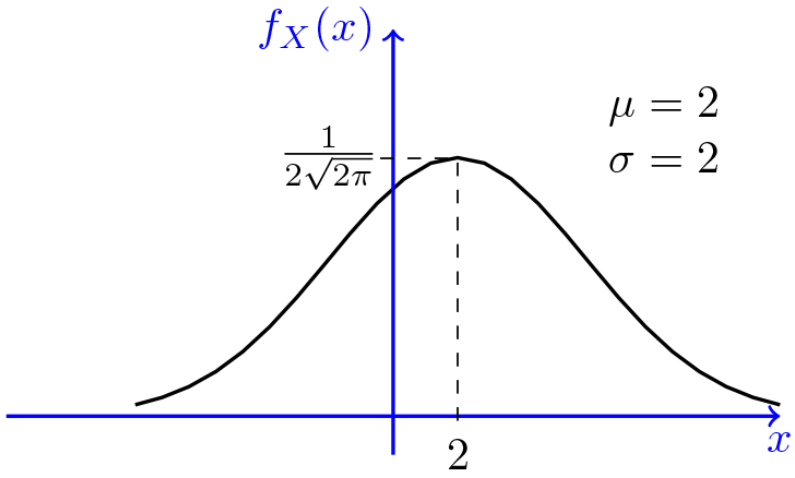


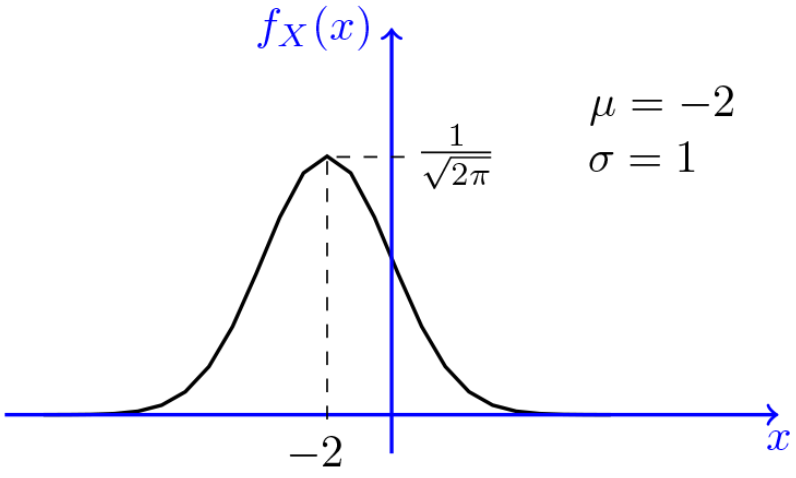
For a normal distribution:

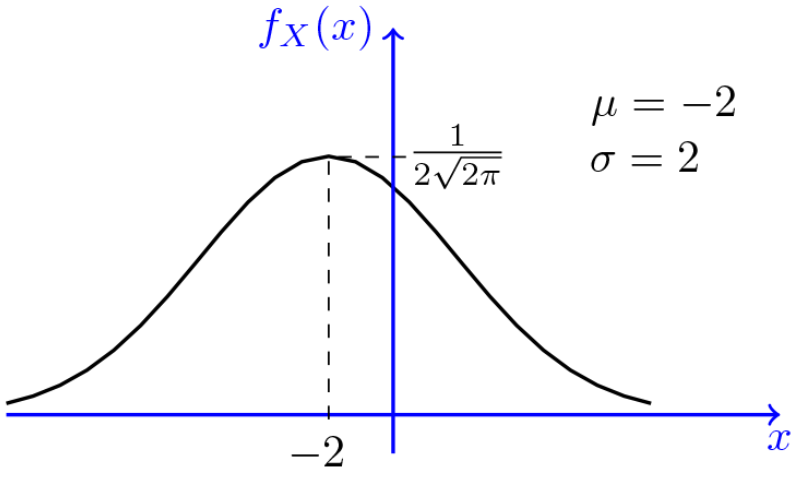
$$CDF(x) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Normal Distribution



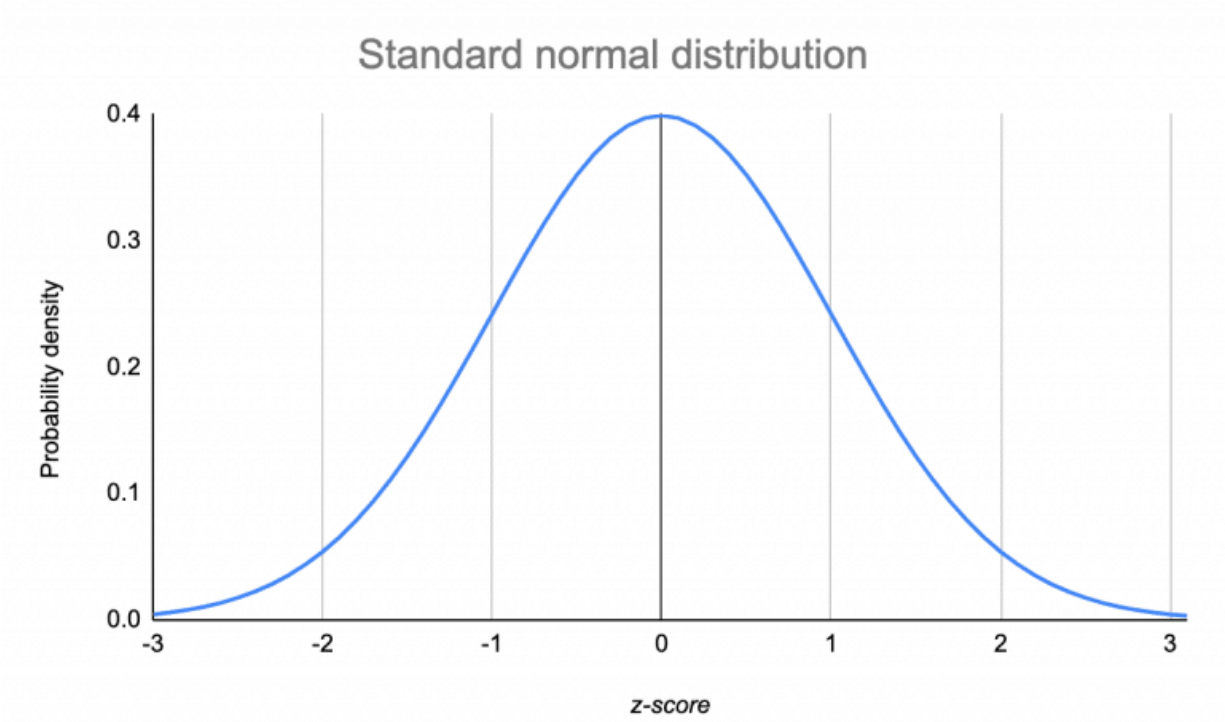




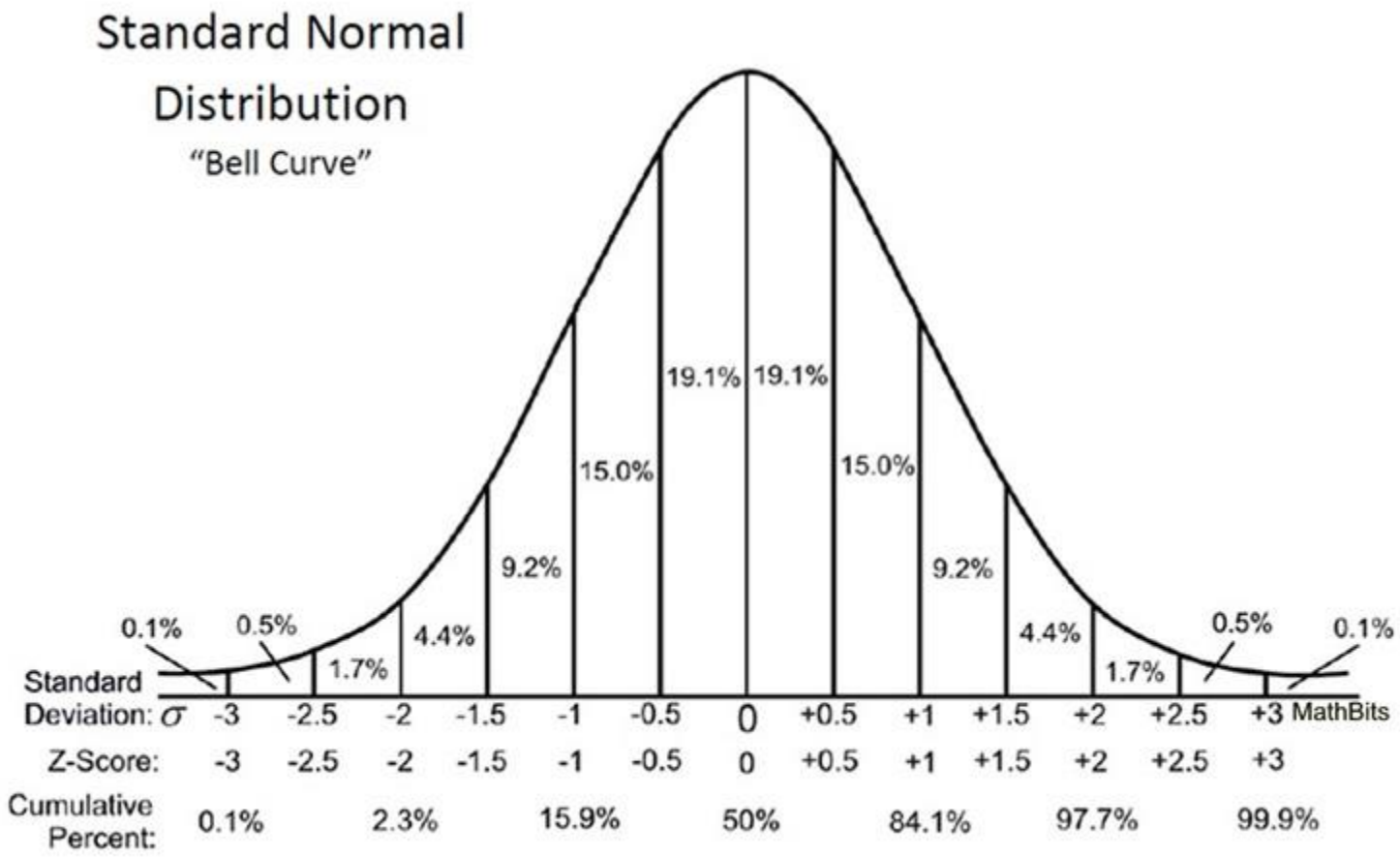


Z-Distribution

- The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.
- Z-scores tell you how many standard deviations away from the mean each value lies.



Z-Distribution



Z-Score

$$z = \frac{x - \mu}{\sigma}$$

x = raw score

μ = mean

σ = standard deviation (std)

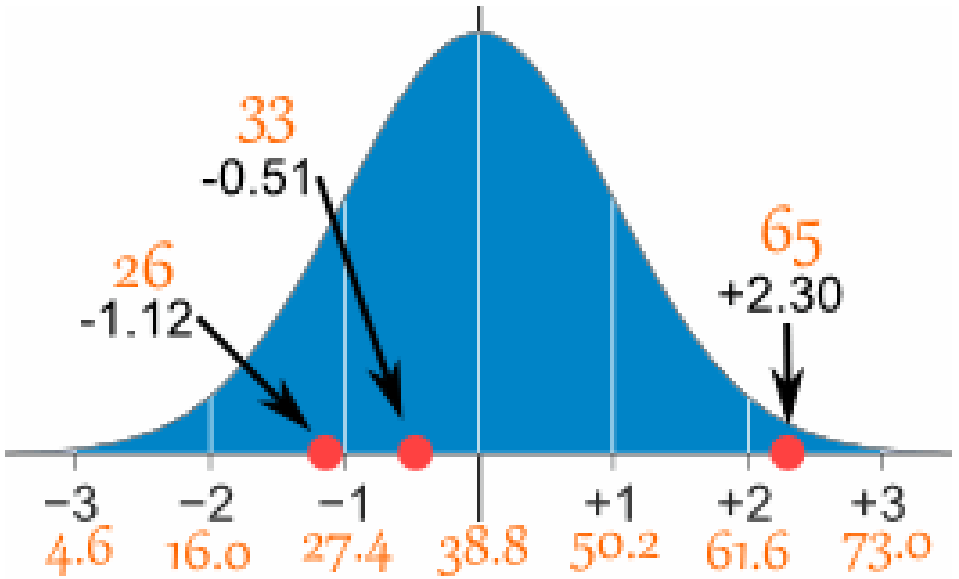
As the formula shows, the **z-score** is simply the raw score minus the population mean, divided by the population standard deviation.

Z-Distribution

Day	Time
1	26
2	33
3	65
4	28
5	34
6	55
7	25
8	44
9	50
10	36
11	26
12	37
13	43
14	62
15	35
16	38
17	45
18	32
19	28
20	34

Mean is 38.8 minutes

Standard Deviation is 11.4 minutes



Z-Table

- A z-table, also known as the standard normal table, provides the area under the curve to the left of a z-score.

- This area represents the probability that z-values will fall within a region of the standard normal distribution.

- Use a z-table to find probabilities corresponding to ranges of z-scores.

Positive Z-Score Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Negative Z-Score Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
-0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
-0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
-0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
-0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
-0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414

Example

Example: Finding a *z* score

You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150. You want to find the probability that SAT scores in your sample exceed 1380.

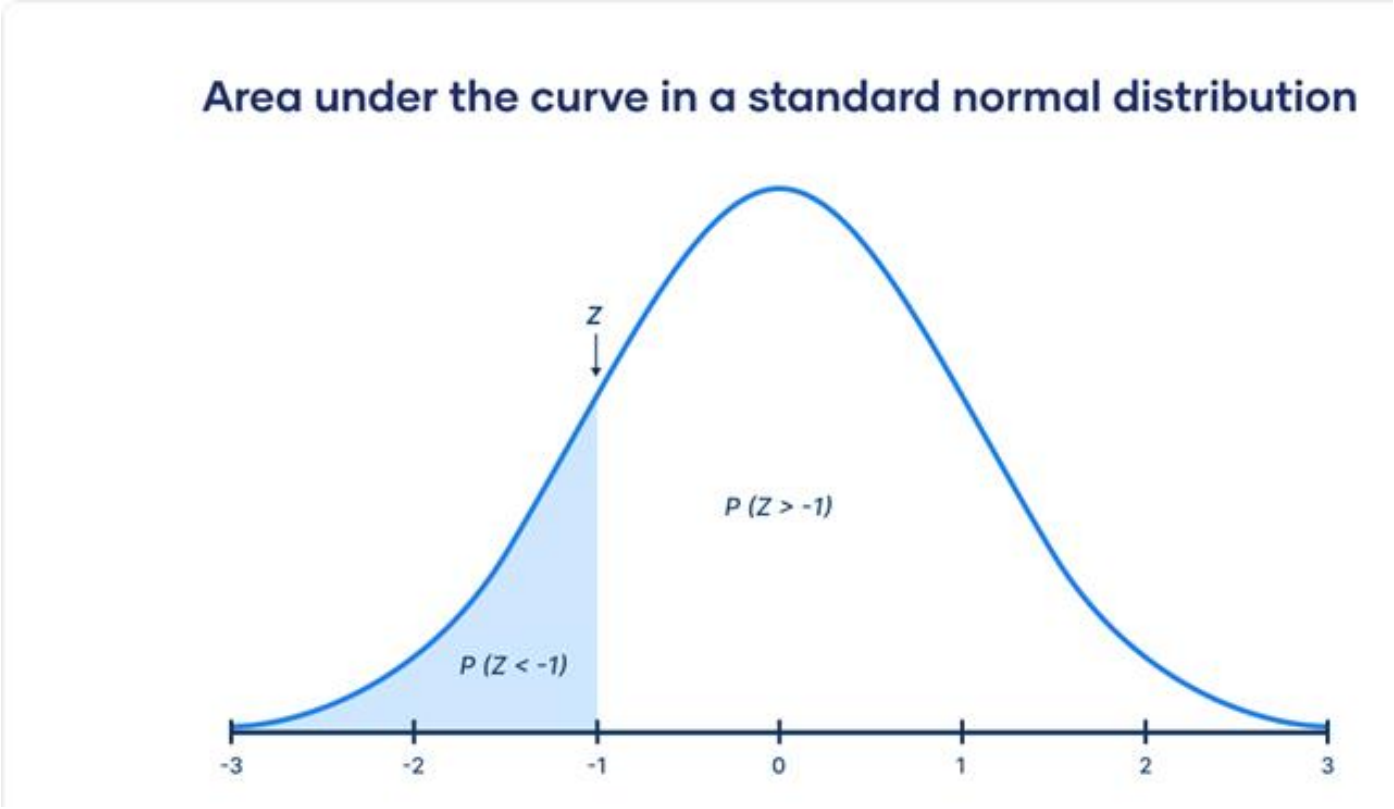
To standardize your data, you first find the *z* score for 1380. The *z* score tells you how many standard deviations away 1380 is from the mean.

Step 1: Subtract the mean from the *x* value.	$x = 1380$ $M = 1150$ $x - M = 1380 - 1150 = 230$
Step 2: Divide the difference by the standard deviation.	$SD = 150$ $z = 230 \div 150 = 1.53$

The *z* score for a value of 1380 is **1.53**. That means 1380 is 1.53 standard deviations from the mean of your distribution.

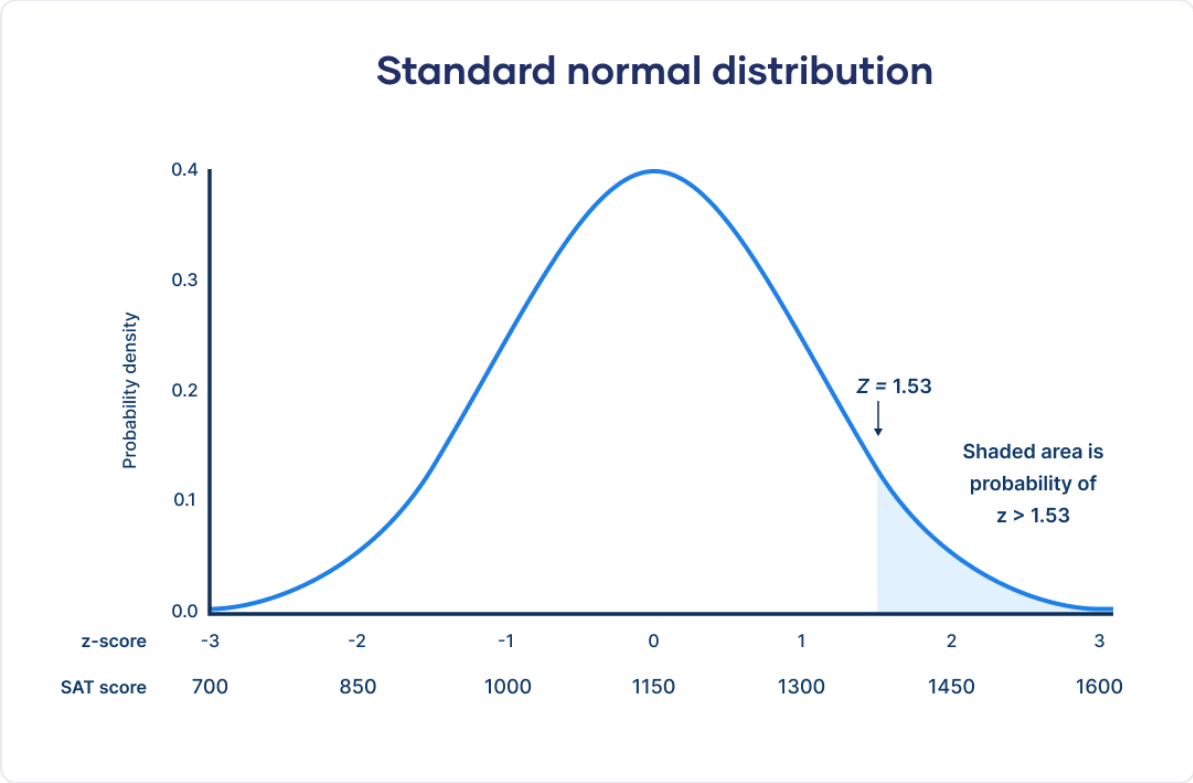
Example

Every z score has an associated p value that tells you the probability of all values below or above that z score occuring. This is the area under the curve left or right of that z score.



Example

We've calculated that a SAT score of 1380 has a z score of 1.53. Using the full z table, we find that for a z score of 1.53, the p value is 0.937.



Positive Z-Score Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Example 1

The scores on a certain college entrance exam are normally distributed with mean $\mu = 82$ and standard deviation $\sigma = 8$. Approximately what percentage of students score less than 84 on the exam?

First, we will find the z-score associated with an exam score of 84:

$z\text{-score} = (x - \mu) / \sigma = (84 - 82) / 8 = 2 / 8 = \mathbf{0.25}$

Approximately **59.87%** of students score less than 84 on this exam.

Example 2

The height of plants in a certain garden are normally distributed with a mean of $\mu = 26.5$ inches and a standard deviation of $\sigma = 2.5$ inches. Approximately what percentage of plants are greater than 26 inches tall?

First, we will find the z-score associated with a height of 26 inches.

$z\text{-score} = (x - \mu) / \sigma = (26 - 26.5) / 2.5 = -0.5 / 2.5 = \mathbf{-0.2}$

We see that 42.07% of values fall below a z-score of -0.2. However, in this example we want to know what percentage of values are *greater* than -0.2, which we can find by using the formula 100% – 42.07% = 57.93%.

Example 3

The weight of a certain species of dolphin is normally distributed with a mean of $\mu = 400$ pounds and a standard deviation of $\sigma = 25$ pounds. Approximately what percentage of dolphins weigh between 410 and 425 pounds?

First, we will find the z-scores associated with 410 pounds and 425 pounds

z-score of 410 = $(x - \mu) / \sigma = (410 - 400) / 25 = 10 / 25 =$ **0.4**

z-score of 425 = $(x - \mu) / \sigma = (425 - 400) / 25 = 25 / 25 =$ **1**

Lastly, we will subtract the smaller value from the larger value: **0.8413 – 0.6554 = 0.1859**.

Thus, approximately **18.59%** of dolphins weigh between 410 and 425 pounds.

Hypothesis

- A hypothesis is an educated guess about something in the world around you.
- It should be testable, either by experiment or observation.

  - A new medicine you think might work.
  - A possible location of aliens in the universe.
  - The average CGPA of your class is 3.2

Hypothesis Statement

- If you are going to propose a hypothesis, it's customary to write a statement. Your statement will look like this:
- "If I…(do this to an independent variable)….then (this will happen to the dependent variable)."

  - If I (increase the amount of sugar in tea) then (the tea will be sweeter)
  - If I (study from the textbooks) then (my CGPA will improve)
  - If I (go to the main campus via motorway) then (I will reach earlier)

Hypothesis Testing

- Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.

- You're basically **testing whether your results are valid** by figuring out the **odds that your results** have happened by chance.

- If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Hypothesis Testing

- A hypothesis is a statement regarding what you believe might happen.
- **State your null hypothesis.** The null hypothesis is a commonly accepted fact. It's the default, or what we'd believe if the experiment was never conducted.
- **State an alternative hypothesis.** You'll want to prove an alternative hypothesis. This is the opposite of the null hypothesis, demonstrating or supporting a statistically significant result.
- **Determine a significance level.** This is the determiner, also known as the alpha (α). It defines the probability that the null hypothesis will be rejected. A typical significance level is set at 0.05 (or 5%).
- **Calculate the Z-score**

Z-Test

- This is a statistical hypothesis test where the distribution of the **statistic** we are measuring, for example the mean, is part of the normal distribution.

- There are multiple types of Z-Tests, we will learn only **one sample mean test.**

- This is used to determine if the difference between the **mean of a sample** and **the mean of a population** is **statistically significant**.

- This is a measure of **how many standard deviations** away a sample statistics is from the populations' mean.

Z-Test: Requirements

- Sample size is greater than 30. This is because we want to ensure our sample mean comes from a distribution that is normal.

- The standard deviation and mean of the population is known.

- The sample data is collected/acquired randomly.

Z-Test: Steps

- State the null hypothesis H_0
- State the alternate hypothesis, H_1
- Choose your critical value, α, which determines whether you accept or reject the null hypothesis. Typically for the Z-Test we would use a statistical significance of 5%.
- Calculate Z-statistic

Z-Test: Tests

Z-TEST

- Formula to find the value of Z (z-test) Is:

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

- $\bar{x}$ = mean of sample
- μ_0 = mean of population
- σ = standard deviation of population
- n = no. of observations

Z-Test: Example

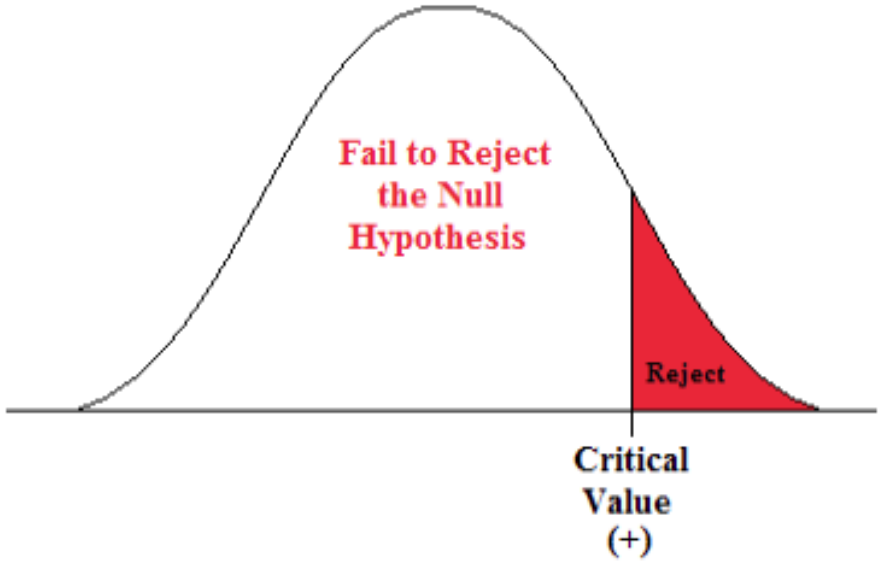
A school says that its students are on average smarter than other schools. It takes a sample of 50 students whose average IQ measures to be 110. The population, the rest of the schools, have an average IQ of 100 and standard deviation of 20. Is the schools claim of its students correct?

Null Hypothesis: The mean IQ of students is 100
Alternate Hypothesis: The mean IQ of students is greater than 100.

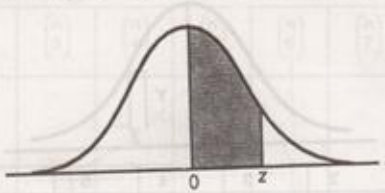
Z-Test: Example

Since Sample mean is greater than population mean so this is called a right sided one-tailed test.

So, choosing a critical value of 5%, which equals a Z-Score of **1.645**, we can only reject the null hypothesis if our Z-Test Statistic is greater than 1.645.



VII. AREA UNDER STANDARD NORMAL CURVE



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.01	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.05	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.09	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.13	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1519	.1628	.1664	.17	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.20	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.23	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.27	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.29	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.32	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.35	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.37	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.39	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.40	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.42	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.43	.4394	.4406	.4418	.4429	4441
1.6					.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

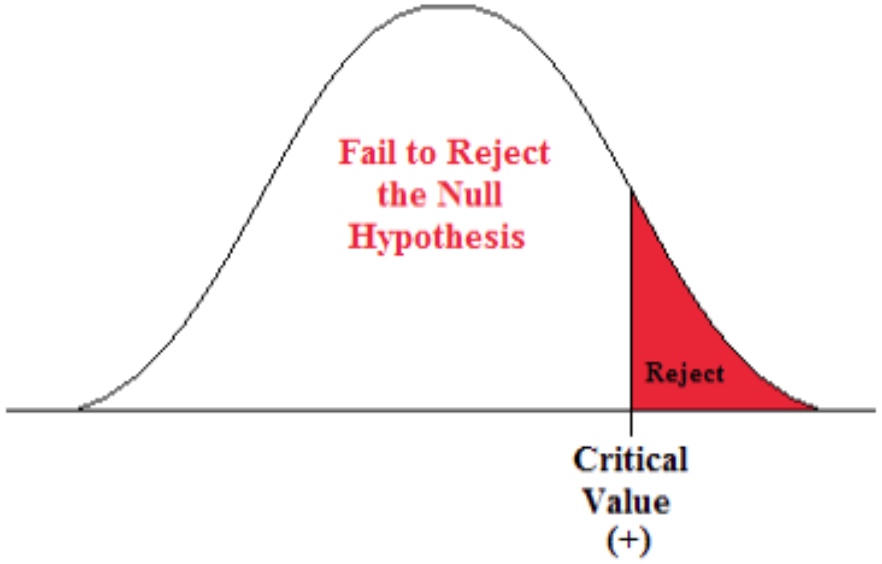
Z-Test: Critical Value

Cumulative Standardized Normal Distribution Table (Z-Table)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Z-Test: Example

$$z = \frac{110 - 100}{20/\sqrt{50}} \approx 3.5$$

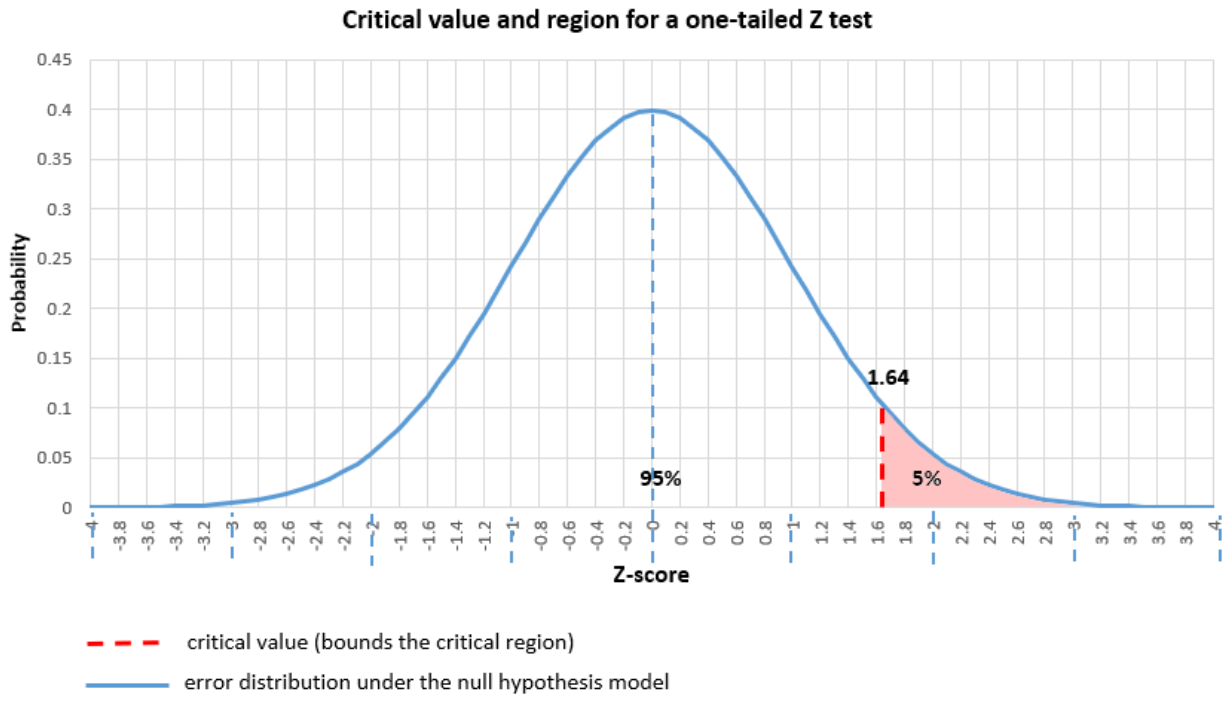


Therefore, we reject the null hypothesis and the schools' claim is right!

Z-Test: Critical Value

A critical value is a line on a graph that splits the graph into sections.

One of the sections is the “rejection region“; if your test value **falls into that region**, then you **reject the null hypothesis**.



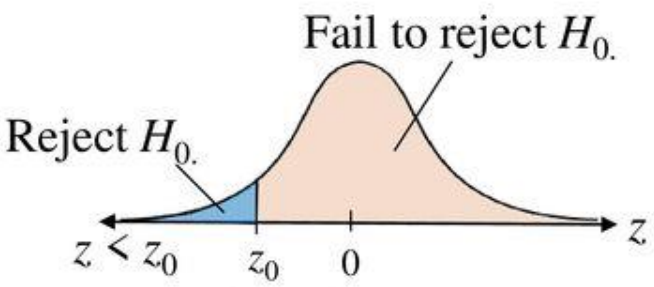
Z-Test Critical Value

CL	α	z_α for one-tailed test	$z_{\alpha/2}$ for two-tailed test
90%	10%	-1.28 or 1.28	-1.645 and 1.645
95%	5%	-1.645 or 1.,645	-1.96 and 1.96
99%	1%	-2.33 or 2.33	-2.58 and 2.58
99.9%	0.1%	-3.09 or 3.09	-3.295 and 3.295

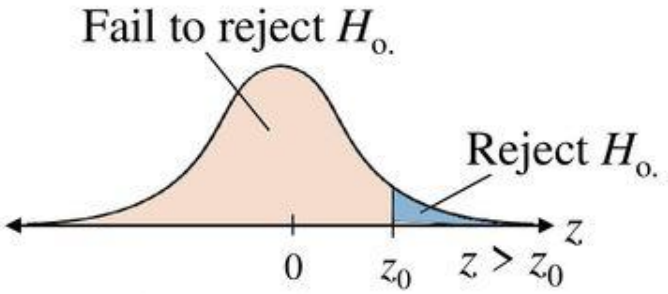
Decision Rule Based on Rejection Region

To use a rejection region to conduct a hypothesis test, calculate the standardized test statistic, z. If the standardized test statistic

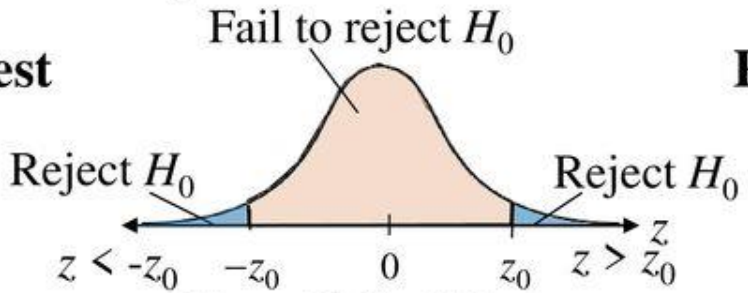
1. is in the rejection region, then reject H_0.
2. is *not* in the rejection region, then fail to reject H_0.



Left-Tailed Test



Right-Tailed Test



Two-Tailed Test

Right- , Left-, or Two-tailed Test

There are three basic types of 'tails' that hypothesis tests can have:

- One-tailed test
  - Right-tailed test: where the alternative hypothesis includes a '>' symbol.
  - Left-tailed test: where the alternative hypothesis includes a '<' symbol.
- Two-tailed test: where the alternative hypothesis includes a ≠.

Right Tailed Test

The mean lifetime E[X] of the light bulbs produced by Lighting Systems Corporation is 1570 hours with a standard deviation of 120 hours. The president of the company claims that a new production process has led to an increase in the mean lifetimes of the light bulbs. If a worker tested 100 light bulbs made from the new production process and found that their mean lifetime is 1600 hours, test the hypothesis that E[X] is greater than 1570 hours using a level of significance 0.05.

Null Hypothesis: Mean lifetime of a bulb is 1570 hours
Alternate Hypothesis: Mean lifetime of a bulb is more than 1570 hours

Right Tailed Test

The average CGPA of a class of 100 students is 3.1. A teacher claims that his teaching method has improved the mean CGPA to 3.2 when 36 students were randomly sampled. If the standard deviation of CGPA of the class is 0.4 then check if teacher’s claim is acceptable or not?

Left tail Z-test

- The average weight of an iron bar population is 90kg. Supervisor believes that the average weight might be lower. Random samples of 36 iron bars are measured, and the average weight is 82kg and a standard deviation of 18kg. With a 95% confidence level, is there enough evidence to suggest the average weight is lower?

Two tail Z-test

- The average score for the mean population is 80, with a standard deviation of 10. With a new training method, the professor believes that the score might change. Professor tested randomly 36 students' scores. The average score of the sample is 88. With a 95% confidence level, is there enough evidence to suggest the average score changed?

Confidence Interval

- In inferential statistics, a primary goal is to estimate population parameters.
- These parameter values are not only unknown but almost always unknowable.
- Confidence intervals incorporate the uncertainty and sample error to create a range of values the actual population value is likely to fall within.

Confidence Interval

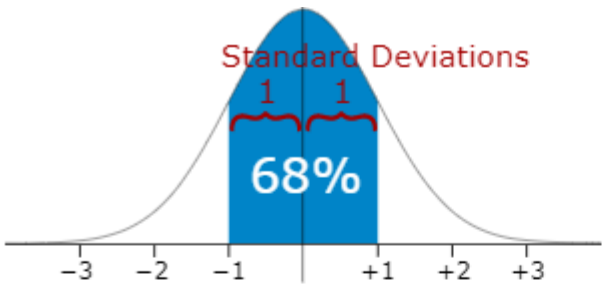
- A confidence interval is the **mean of your estimate plus and minus the variation in that estimate**.

- Confidence, in statistics, is another way to describe probability.

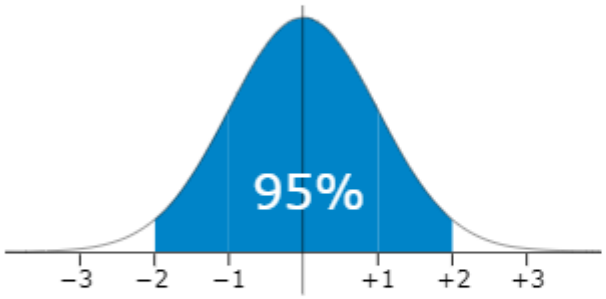
- For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Confidence Interval

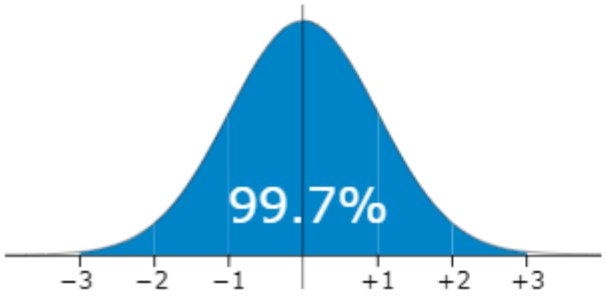
Day	Travel Time (min)
1	25
2	26
3	26
4	28
5	28
6	32
7	33
8	34
9	34
10	35
11	36
12	37
13	38
14	43
15	44
16	45
17	50
18	55
19	62
20	65
Average	38.8
SD	11.70



68% of values are within **1 standard deviation** of the mean



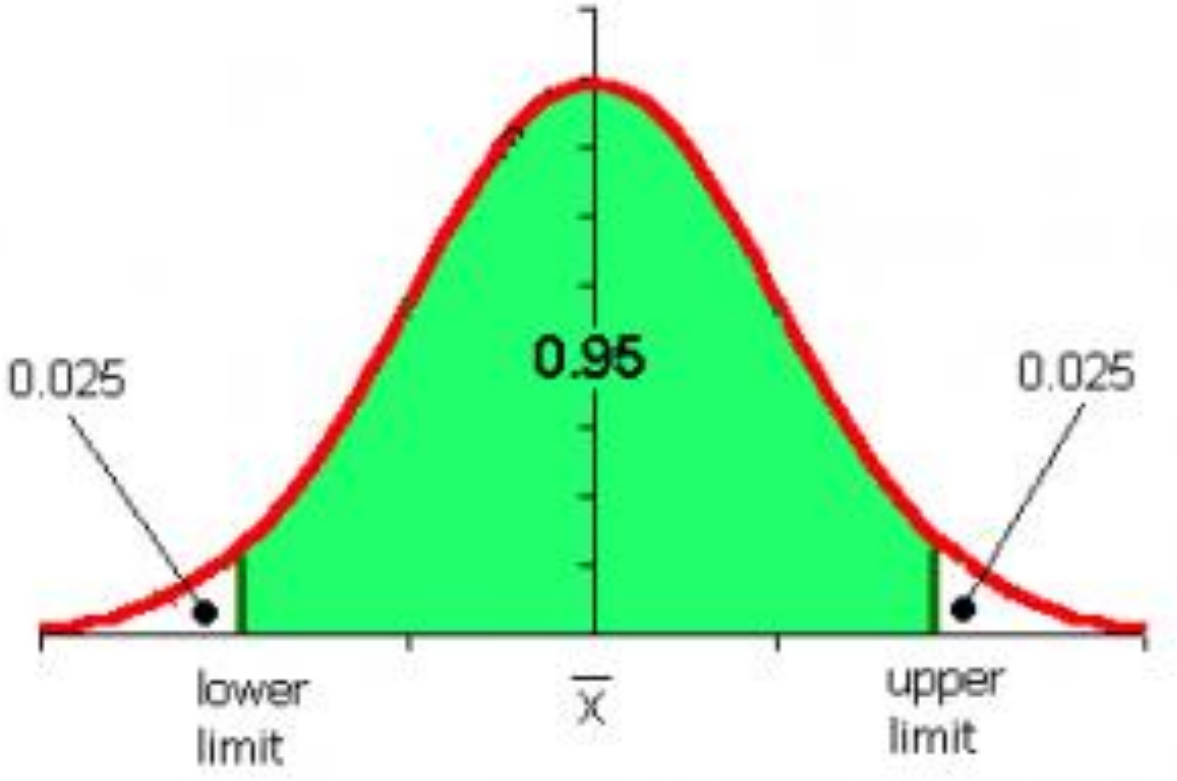
95% of values are within **2 standard deviations** of the mean



99.7% of values are within **3 standard deviations** of the mean

The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score".

Confidence Interval



Confidence Interval

We measure the heights of 40 randomly chosen men and get a mean height of 175cm. We also know the standard deviation of men's heights is 20cm. Find the 95% Confidence Interval.

- number of observations **n = 40**
- mean $\overline{X}$ **= 175**
- standard deviation **s = 20**

$$\overline{X} \pm Z\frac{s}{\sqrt{n}}$$

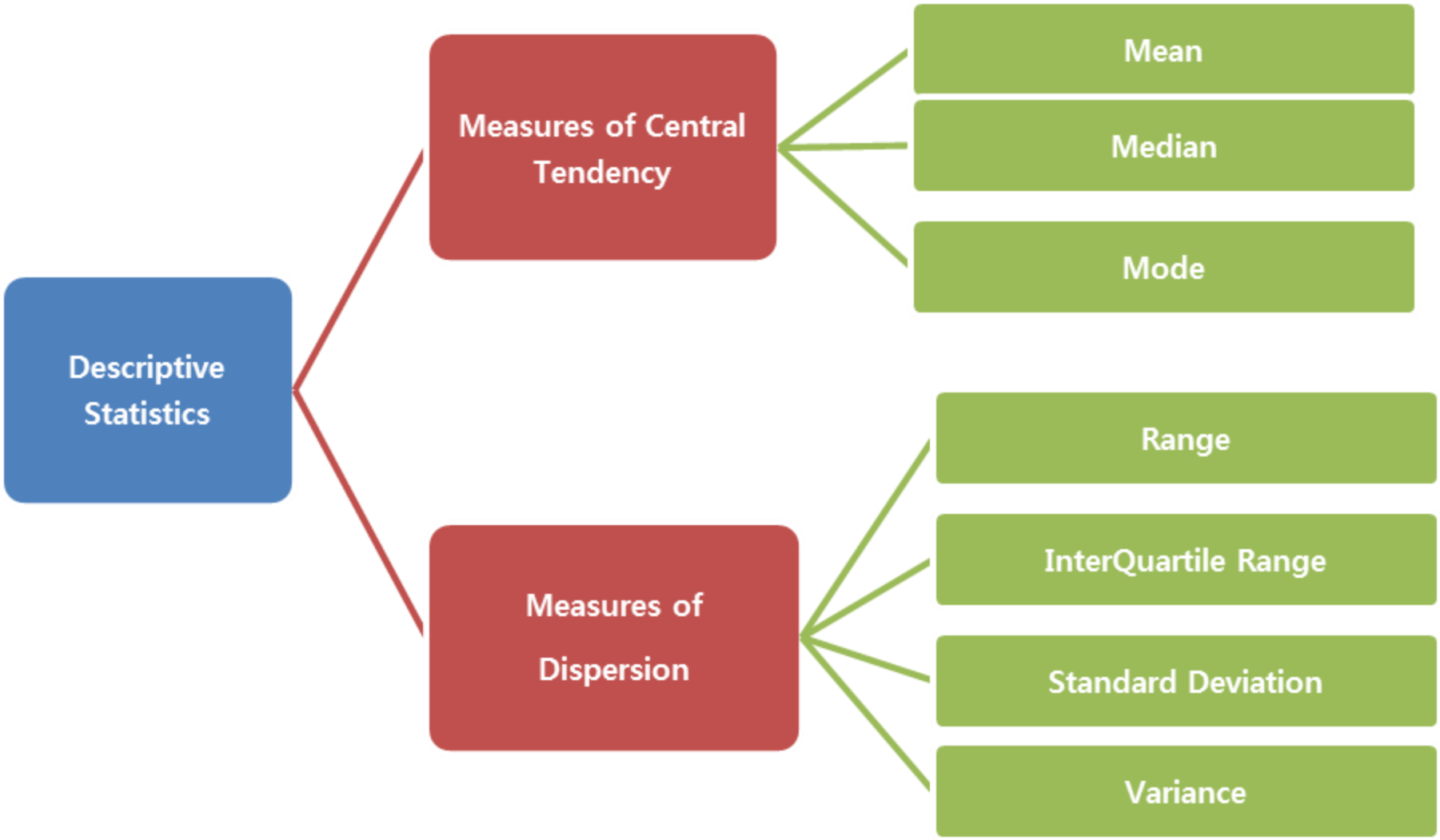
- $\overline{X}$ is the mean
- **Z** is the chosen Z-value from the table above
- **s** is the standard deviation
- **n** is the number of observations

Confidence Level	Alpha	Alpha/2	z alpha/2
90%	10%	5.0%	1.645
95%	5%	2.5%	1.96
98%	2%	1.0%	2.326
99%	1%	0.5%	2.576

$$175 \pm 1.960 \times \frac{20}{\sqrt{40}}$$

175cm ± 6.20cm

Descriptive Statistics



Measures of Central Tendency

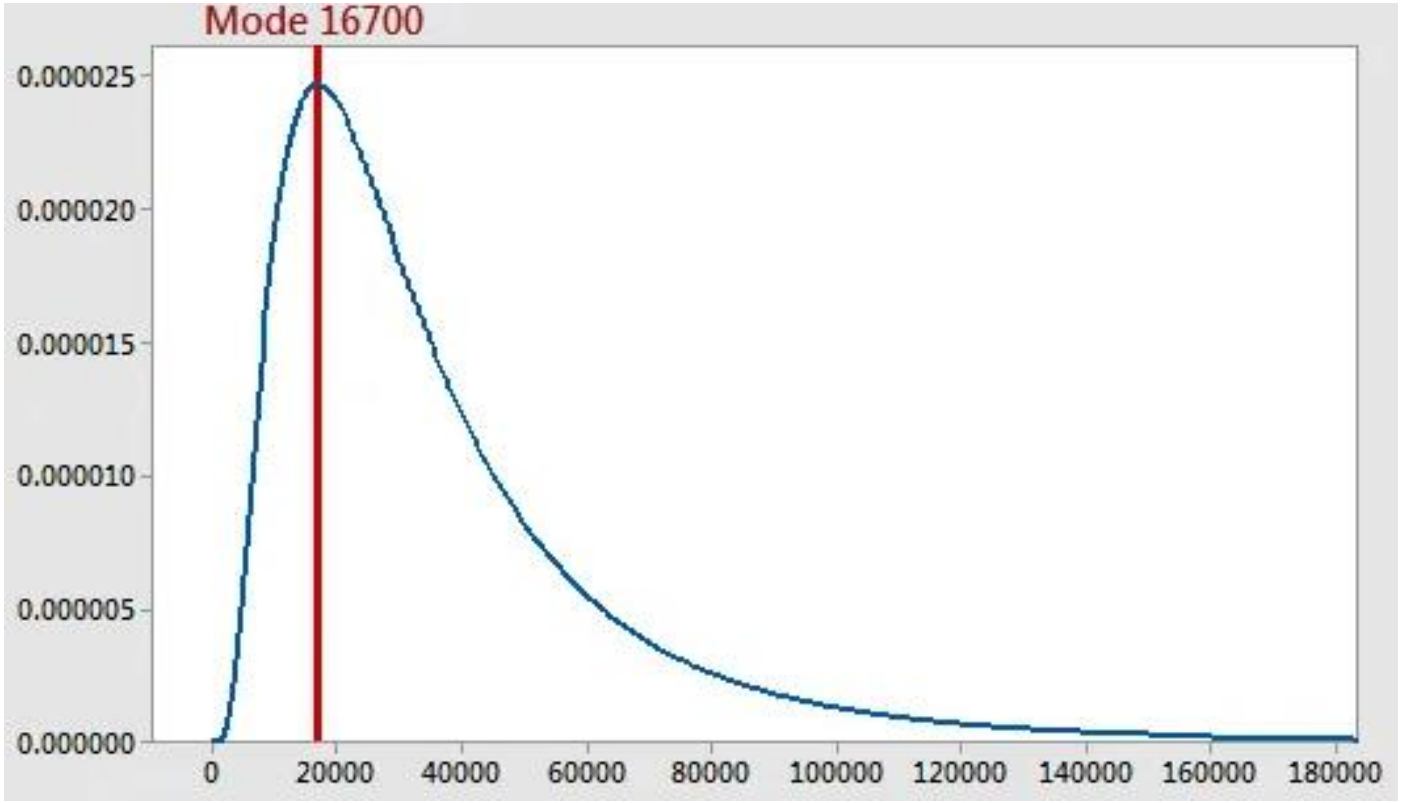
Central Tendency Measures		
Measure	**Formula**	**Description**
Mean	$\sum x/n$	Balance Point
Median	n+1/2 Position	Middle Value when ordered
Mode	None	Most frequent

Measures of Central Tendency

- When you are working with the continuous data, don’t be surprised if there is no mode.
- How to find the mode for continuous data ?

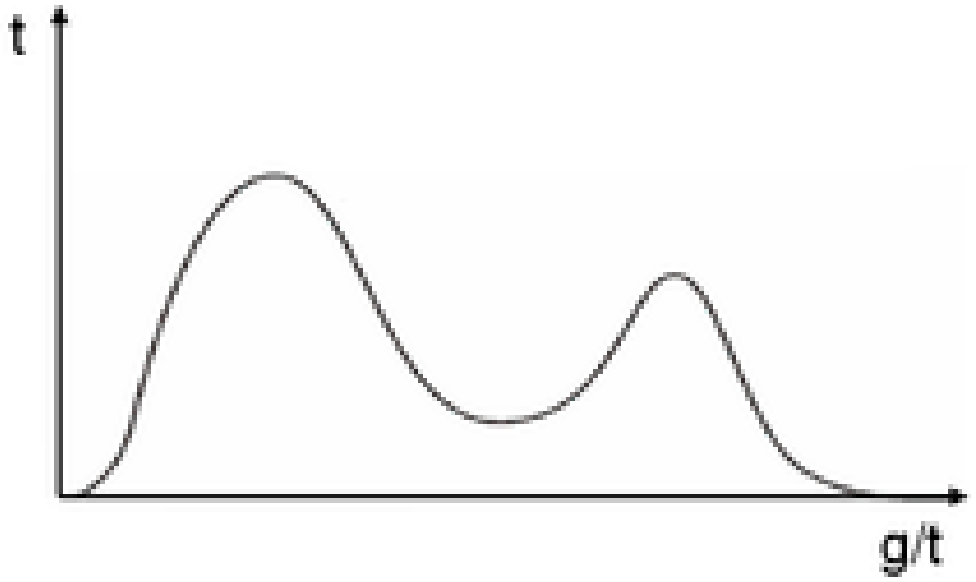
Measures of Central Tendency

- We can find the mode for continuous data by locating the maximum value on a probability distribution plot



Measures of Central Tendency

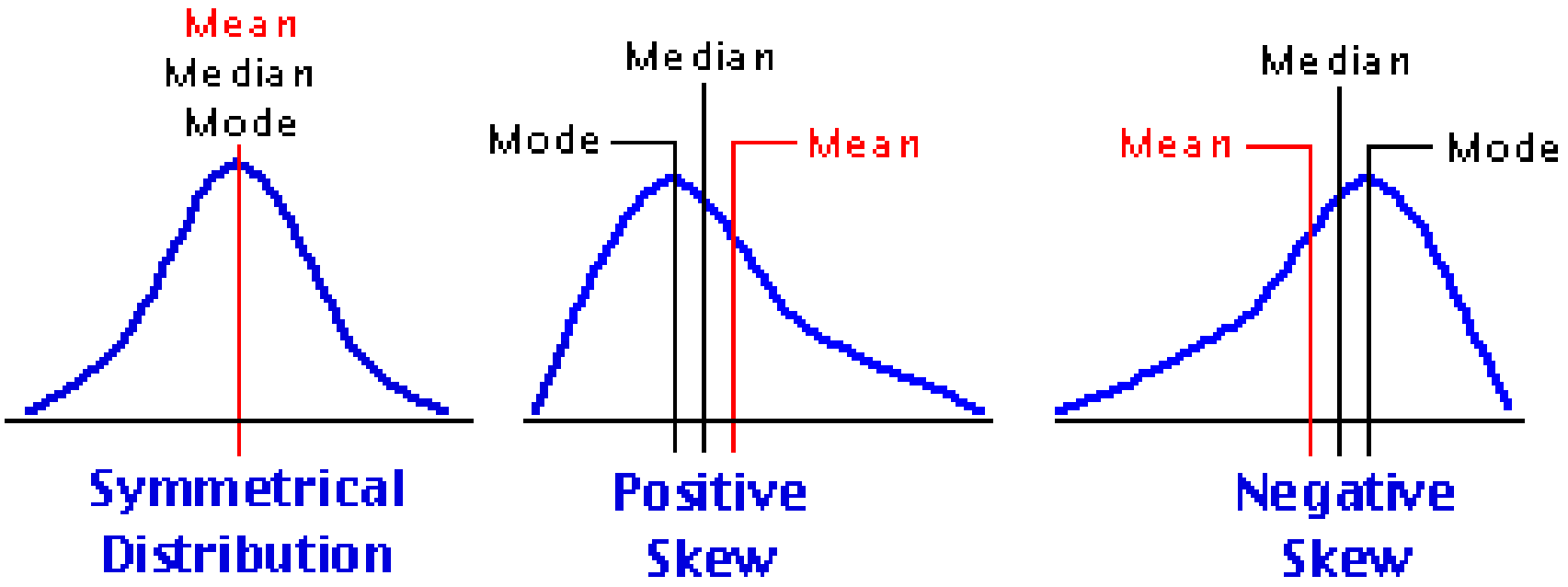
- If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution



Skewness

- Symmetrical Distribution
- Positive Skew
- Negative Skew

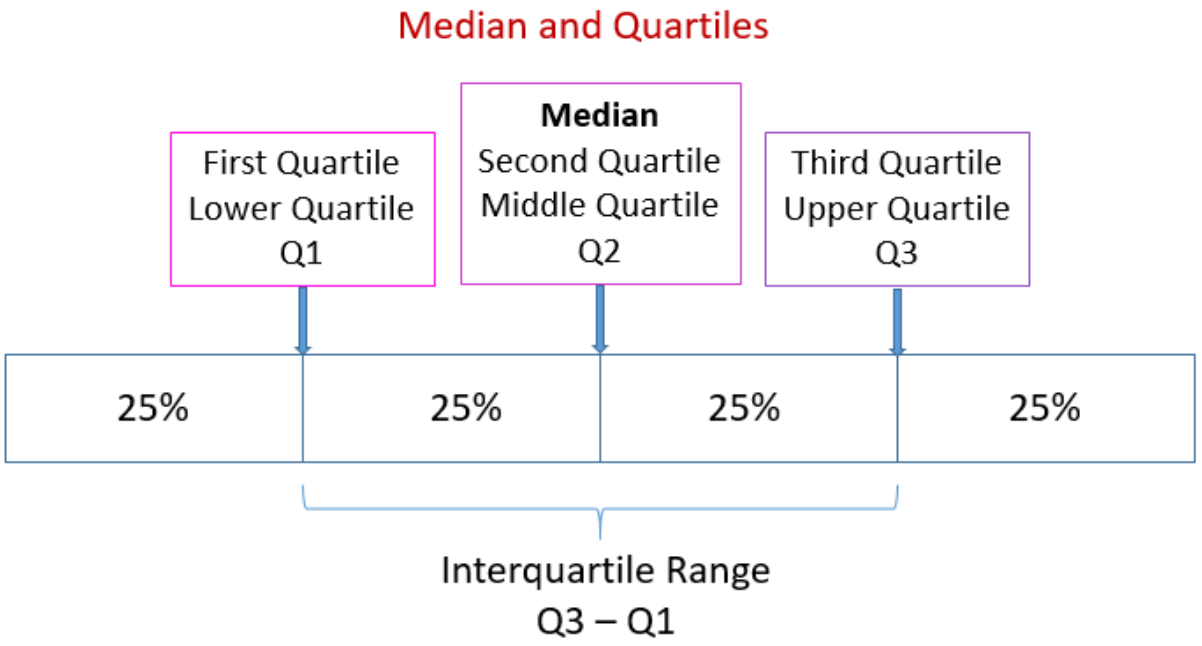
Skewness


Mean
Median
Mode
Symmetrical Distribution
Median
Mode
Mean
Positive Skew
Median
Mean
Mode
Negative Skew

Measures of Dispersion

Quartiles are also quantiles; they divide the distribution into **four equal** parts.

Percentiles are quantiles that divide a distribution into **100 equal parts.**



Measures of Dispersion

A **standard deviation** is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Measures of Dispersion

The **variance** is a measure of variability.
It is calculated by taking the average of squared deviations from the mean.

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

- σ^2 = population variance
- Σ = sum of...
- X = each value
- μ = population mean
- N = number of values in the population

Summary

- Introduction to Data Science