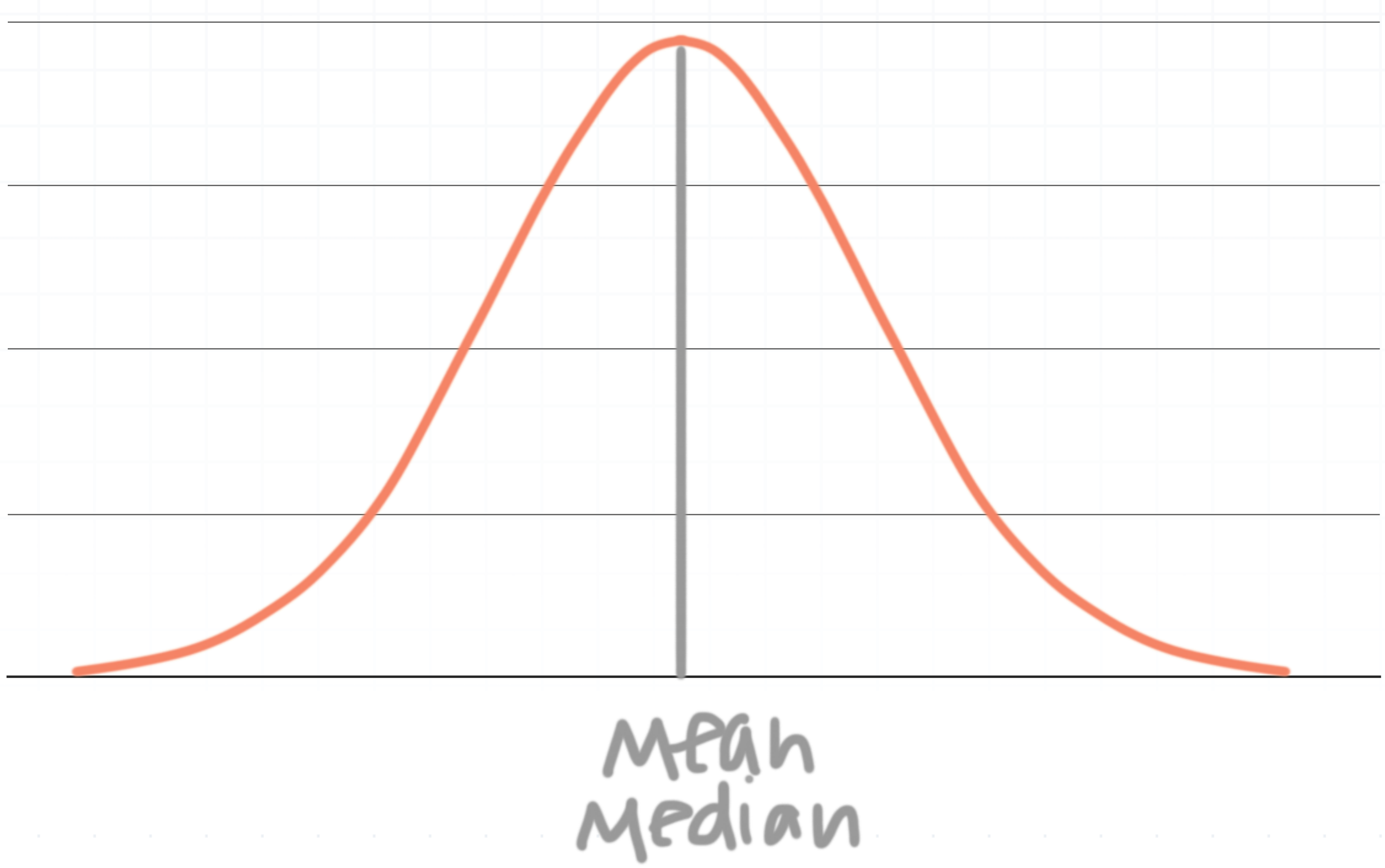
Normal distributions and z-scores

In the last section, we talked about a normal distribution, which is a bell-shaped, symmetric curve for **normally distributed data**, that looks something like this:



We'll spend a lot of time working with distributions like this, so let's talk about some of the most important properties of a normal distribution.

The empirical rule

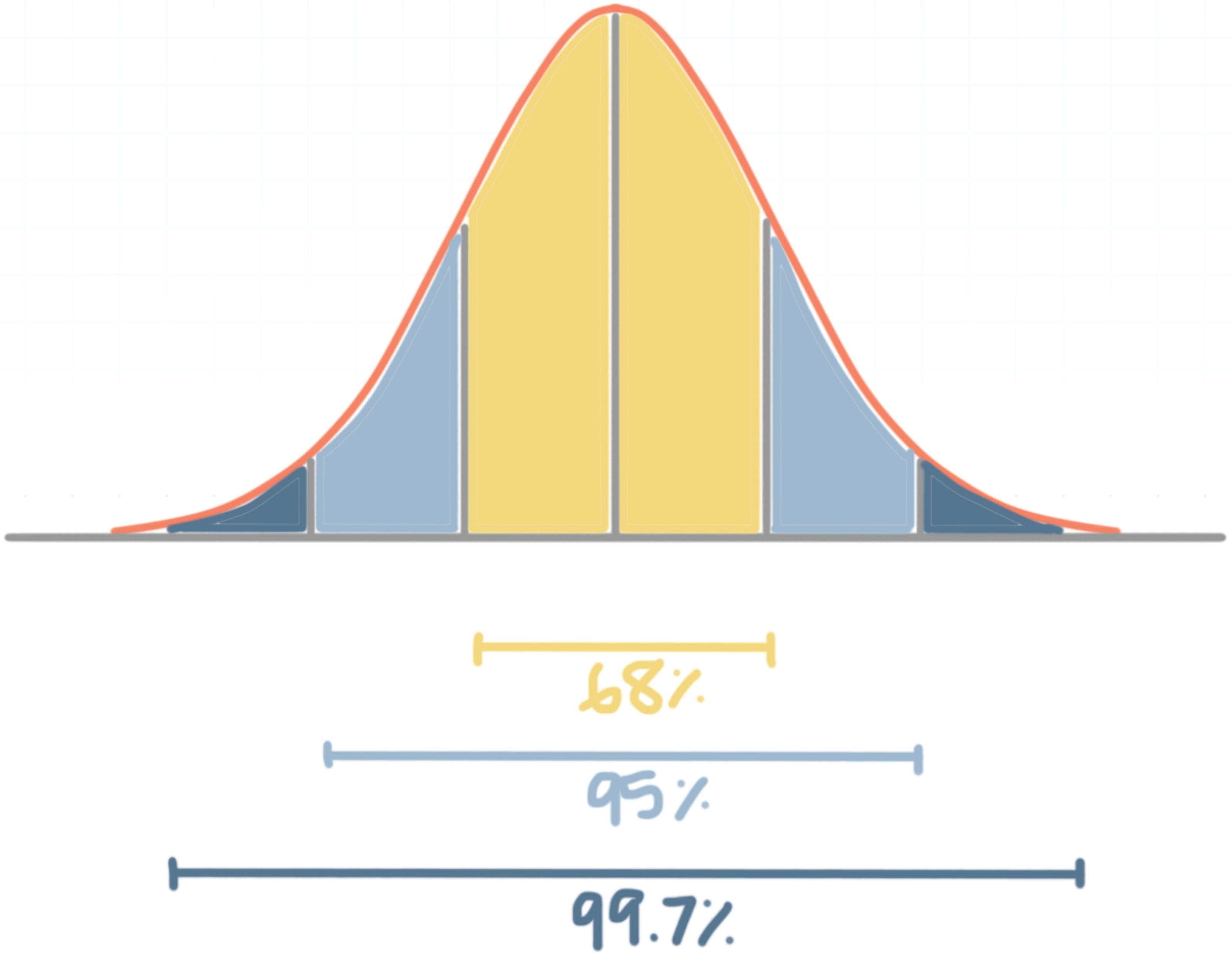
Normal distributions follow the **empirical rule**, also called the **68-95-99.7 rule**. The rule tells us that, for a normal distribution, there's a

- $68\,\%$ chance a data point falls within 1 standard deviation of the mean

- $95\,\%$ chance a data point falls within 2 standard deviations of the mean

- $99.7\,\%$ chance a data point falls within 3 standard deviations of the mean

In other words, if we want to show this graphically,

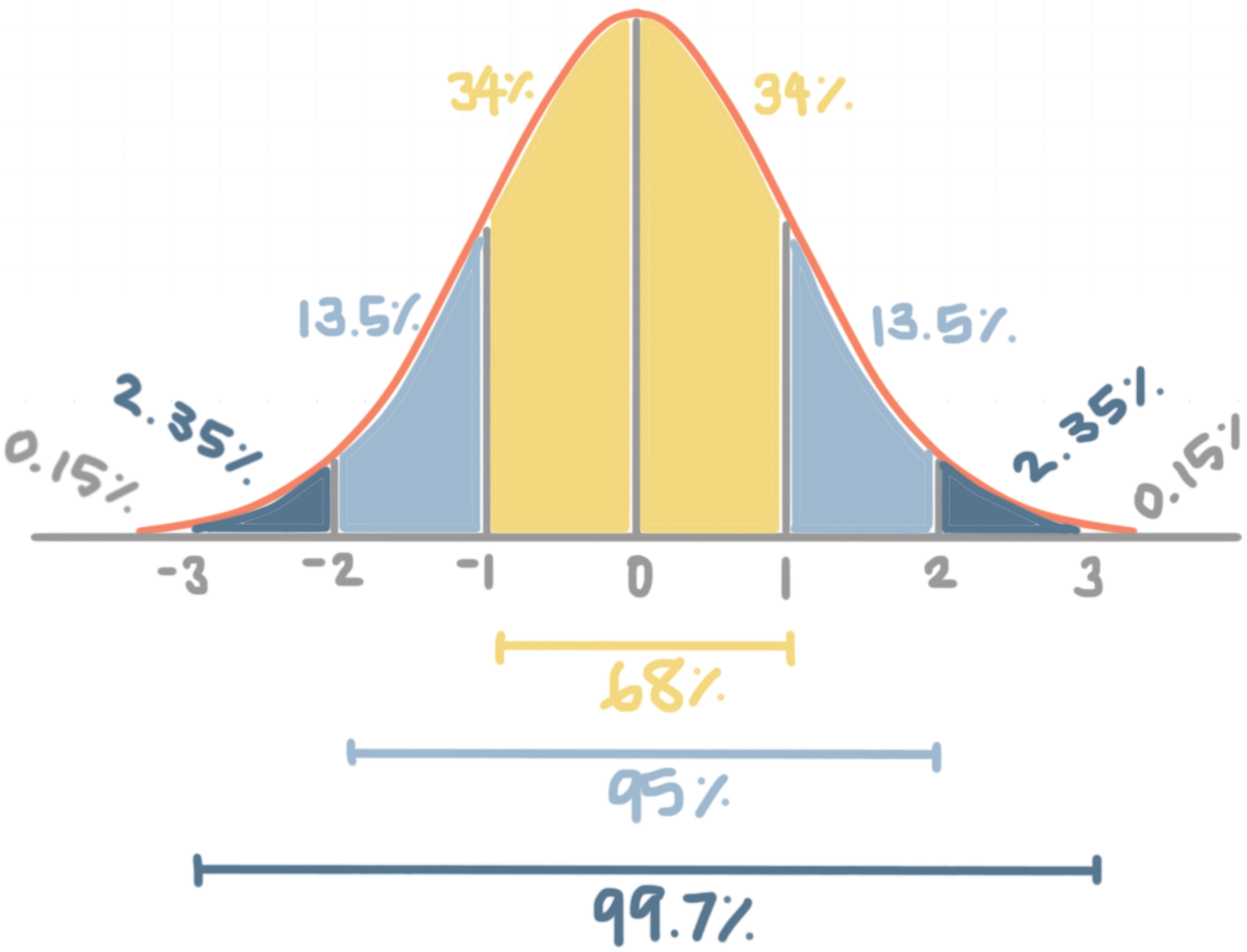


we can show that $68\,\%$ of the data will fall within 1 standard deviation of the mean, that within 2 full standard deviations of the mean we'll have $95\,\%$

of the data, and that within 3 full standard deviations from the mean we'll have $99.7\,\%$ of the data.

And we can draw all kinds of conclusions based on this information, and the fact that all the area under the graph represents $100\,\%$ of the data. For example, since total area is $100\,\%$, and the data within three standard deviations is $99.7\,\%$, that means that we'll always have $0.3\,\%$ of the data in a normal distribution that lies outside three standard deviations from the mean. Or if we wanted to know how much of our data will lie between one and two standard deviations from the mean, we can say that it's

$95\% - 68\,\% = 27\,\%$.



Percentile

We look a lot at percentiles within a normal distribution. The nth **percentile** is the value such that n percent of the values lie below it. In other words, a value in the 95th percentile is greater than 95% of the data. The 50th percentile in a normal distribution always gives the median, or Q_2, and the IQR is always found using the 75th percentile, Q_3, minus the 25th percentile, Q_1.

Z-scores

A z**-score** tells us the number of standard deviations a point is from the mean. To calculate a z-score for normally distributed data (normal distributions) we use the formula

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

where x is the data point, μ is the mean, and σ is the standard deviation.

The z-score for a data point is how far it is from the mean, and we always want to give the z-score in terms of standard deviations. Therefore, to find the z-score at a certain point in the distribution, we use the formula above, taking the data point, subtracting the mean, and then dividing that result by the standard deviation. That gives us a value for z.

We'll look up the z-score in a z-table, which is a table that takes the number of standard deviations and tells us the percentage of the area under the curve up to that point.

Data points that are less than the mean will be to the left of the mean and will have a negative z-score. They should be looked up in the table of negative z-scores:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Data points that are greater than the mean will be to the right of the mean and will have a positive z-score. They should be looked up in the table of positive z-scores:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

A z-score is unusual if it's further than three standard deviations from the mean. Essentially the z-score tells us the percentile rank of the data point that we started with. If the z-score for our data point is 0.7123, it means that the data point is greater than $71.23\,\%$ of the data, meaning that our data point is in the 71.23 percentile. Remember, the z-table always gives us the percentage of data that's below our data point. Therefore, to find the percentage of data above our data point, we have to take 1 minus the value from the table.

Thresholds

Sometimes we want to know the **threshold**, or cutoff, in our data set. In other words, we might want to know "What's the minimum value needed in order to be in the "top $30\,\%$" of the data?

In order to figure this out, we need to work backwards starting from the z-table. For example, if we want to find the top $30\,\%$ of the data, we'd use the z-table to find the first z-score that's just barely above $70\,\%$, or 0.7000. Then we'll look at the row and column headers that correspond with a z-table value of 0.7000. The decimal number given by the row and column headers tells us how many standard deviations above the mean we need to be in order to be above $70\,\%$, or, in the top $30\,\%$.

If we multiply that decimal number by the standard deviation, and then add the result to the mean, that will tell us the value that's at the bottom of the top $30\,\%$. If instead we were looking up the "bottom $40\,\%$" in the z-table, we'd need to look for the z-table value that's just under 0.4000.

Example

Let's say the mean finishing time for male speed skaters at the winter Olympics on the 500 meter track is 70.42 seconds, with a standard deviation of 0.34 seconds (the data is normally distributed). What is the maximum time a skater can post if he wants to skate faster than 95% of his competitors?

We know that $\mu = 70.42$ and $\sigma = 0.34$. This athlete wants to be faster than 95% of the event's participants, which means he wants his time to be in the fastest 5%. Keep in mind here that, if he's finishing in the fastest 5%, that means his finishing time is in the lowest 5% of times.

In other words, he wants his time to be in the top 5% of finishers, which is equivalent to having a finishing time in the fastest 5% of all finishing times, which is equivalent to having a value in the bottom 5% of the normal distribution, so we need to look in a z-table for the negative z-score that will keep us under 0.05.

We look for the largest value in the body of the negative z-table that's still below 0.05. That value is 0.0495, since the next smallest value of 0.0505 surpasses our 0.05 threshold.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	**.0495**	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

The z-score that keeps us within the top 5% of participants is -1.65. A z-score of -1.64 would push us into the bottom 95% of competitors, over the 5% threshold.

If our standard deviation is $\sigma = 0.34$, and our z-score is -1.65, then we can calculate the maximum skater's time to be in top 5%.

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

$$-1.65 = \frac{x - 70.42}{0.34}$$

$$-1.65(0.34) = x - 70.42$$

$$x = 70.42 - 1.65(0.34)$$

$x = 69.859$ seconds

Therefore, if the skater's time is 69.859 seconds or faster, he'll be within the fastest 5% of finishers in the event.