The standard deviation

- $SD = \sqrt{\frac{1}{n}\sum(x_j - \overline{x})^2}$
- The SD gives a measure of how the data are *clustered around the mean.*
    - ○ If the SD is large, then the data are more spread out — we are more likely to find data that are relatively far from the mean.
    - ○ If the SD is small, then the data is more tightly clustered around the mean — most of the data are fairly close to the average.

Chebyshev's Inequality: For ***any*** set of data, most of the data lies within several SDs of it's mean. Specifically,

the proportion of the data that lies more than k SDs ***away from the mean*** *is always less than* $\boldsymbol{\frac{100\%}{k^2}}$.

For example...

- less than $\frac{100\%}{4} = 25\%$ of the data is more than 2 SDs away from average...

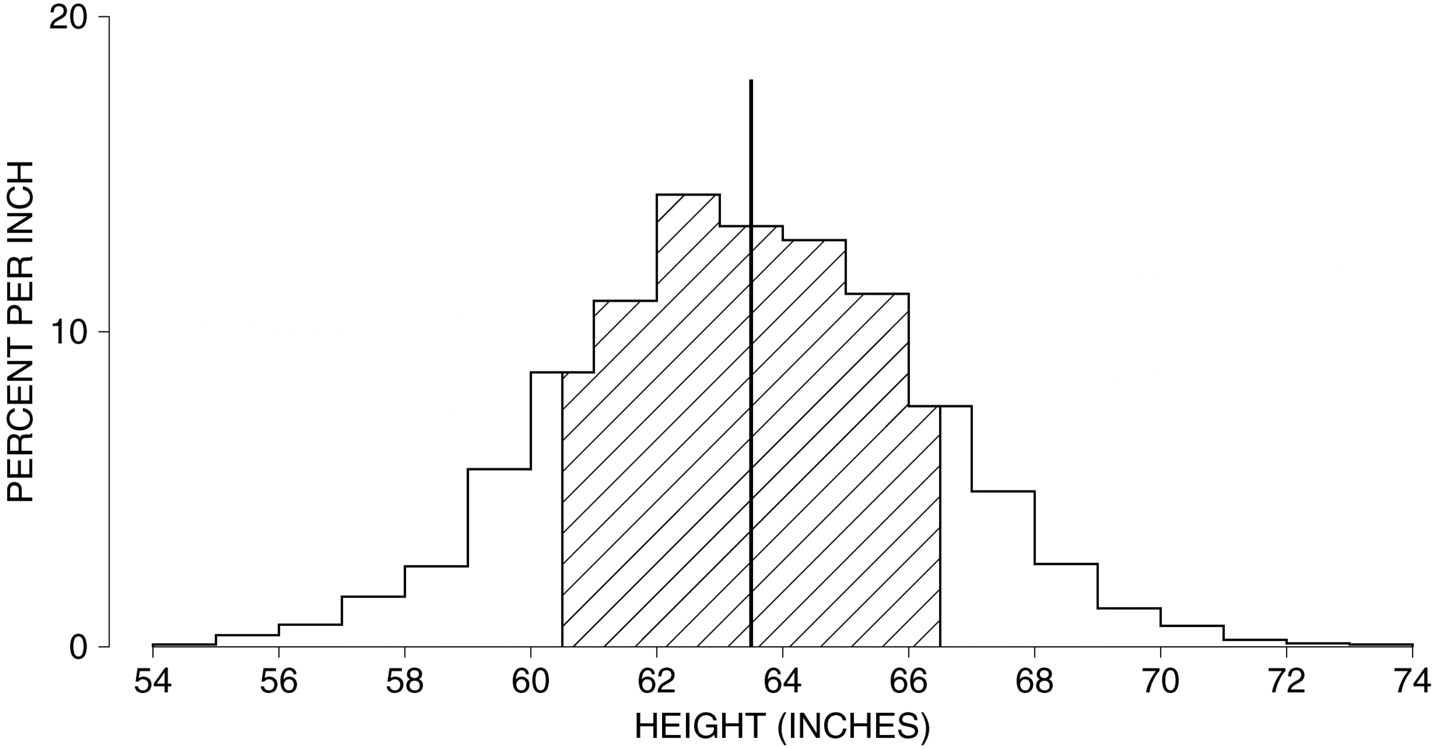
  ... so at least 75% of the data can be found within 2 SDs of the average.

- less than $\frac{100\%}{9} \approx 11.11\%$ of the data is more than 3 SDs away from average...

  ... so at least 88.88% of the data can be found within 3 SDs of the average.

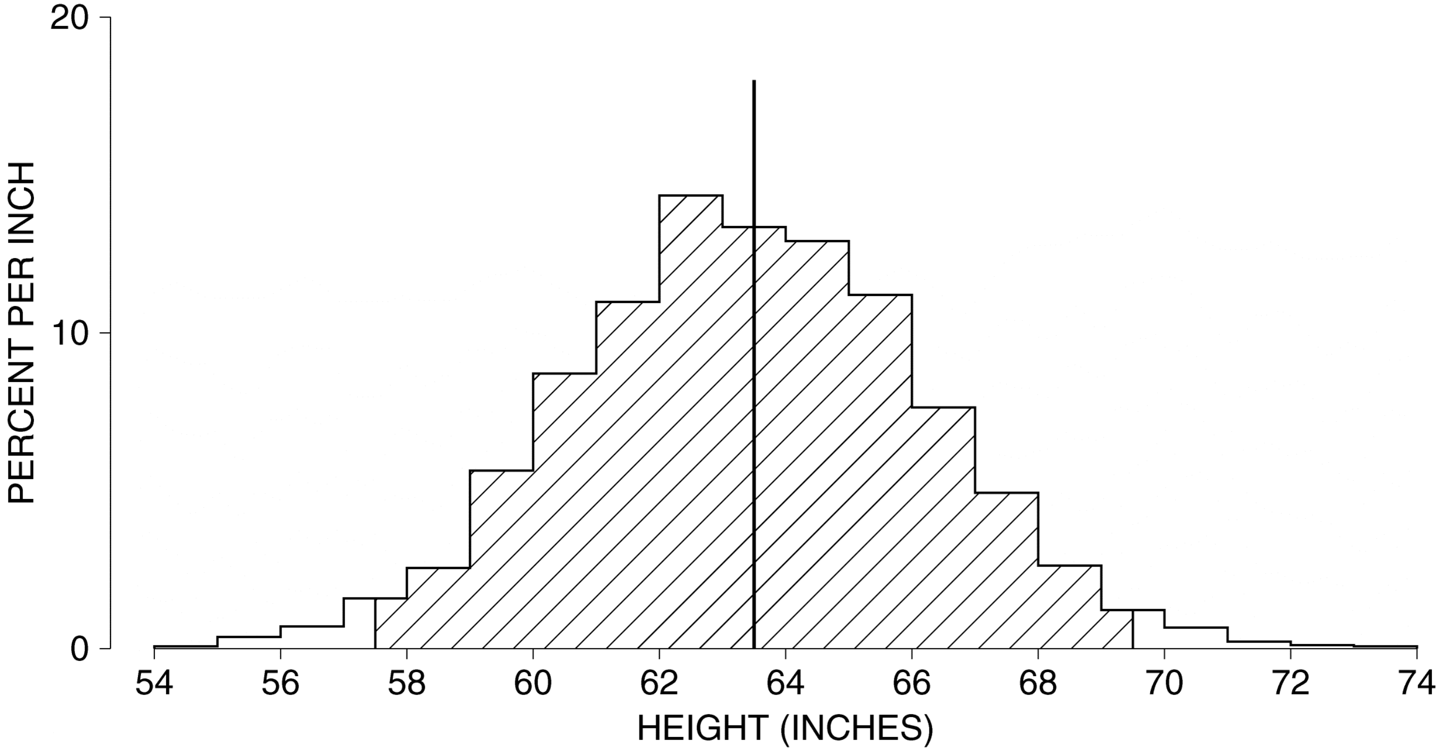
Example (from the book): $\overline{h} = 63.5$ inches and $SD_h \approx 3$ inches...

Figure 8. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within one SD of the average is shaded: 72% of the women differed from average by one SD (3 inches) or less.



Example (continued): $\overline{h} = 63.5$ inches and $SD_h \approx 3$ inches...

Figure 9. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within two SDs of the average is shaded: 97% of the women differed from average by two SDs (6 inches) or less.



Standard units

If we measure distance in multiples of the SD, the estimates for the proportion of the data that lies close to the mean become much more uniform...

(*) Measuring distance from the mean in multiples of the SD leads to the notion of ***standard units***.

If x_j *comes from a distribution with average* $\overline{x}$ *and standard deviation* SD_x*, we convert* x_j *to its standard units,* z_j*, by setting*

$$z_j = \frac{x_j - \overline{x}}{SD_x}.$$

(*) $|z_j|$ tells us how far x_j is from $\overline{x}$, measured in SDs.

(*) z_j tells us whether x_j is above $\overline{x}$ ($z_j > 0$) or below $\overline{x}$ ($z_j < 0$).

(*) Standard units are unit-free numbers.

(*) Using standard units allows us to compare different distributions.

Example. Suppose that the average January temperature in Podunk is 45 °F , with an SD of 2 °F, while in Whoville the average January temperature is 25 °F with an SD of 5 °F. On January 20th, the temperature in Whoville was 16 °F and in Podunk it was 38 °F.

Where was the temperature more *unusual* that day?

We can answer this by converting the temperatures on January 20th in both towns to standard units:

$$z_p = \frac{38 - 45}{2} = -3.5 \quad \text{and} \quad z_w = \frac{16 - 25}{5} = -1.8.$$

(*) Both temperatures were below average.

(*) The *z-score* for Podunk is more negative than the *z-score* for Whoville, so from a statistical point of view the temperature in Podunk was more unusual that day.

Observation. *Converting any set of data, $\{x_1, x_2, \ldots, x_n\}$ with average $\overline{x}$ and standard deviation $SD_x = s$, to standard units produces a set of numbers $\{z_1, z_2, \ldots, z_n\}$ with average $\overline{z} = 0$ and standard deviation $SD_z = 1$.*

Because arithmetic...

$$\begin{aligned}
\overline{z} &= \frac{z_1 + z_2 + \cdots + z_n}{n} \\
&= \frac{\frac{x_1 - \overline{x}}{s} + \frac{x_2 - \overline{x}}{s} + \cdots + \frac{x_n - \overline{x}}{s}}{n} \\
&= \frac{\frac{x_1 - \overline{x}}{n} + \frac{x_2 - \overline{x}}{n} + \cdots + \frac{x_n - \overline{x}}{n}}{s} \\
&= \frac{\frac{x_1 + x_2 + \cdots + x_n}{n} - \frac{\overbrace{\overline{x} + \overline{x} \cdots + \overline{x}}^{n}}{n}}{s} \\
&= \frac{\overline{x} - \overline{x}}{s} = 0
\end{aligned}$$

and more arithmetic

$$
\begin{aligned}
SD_z &= \sqrt{\frac{z_1^2 + z_2^2 + \cdots + z_n^2}{n}} \\
&= \sqrt{\frac{\left(\frac{x_1 - \overline{x}}{s}\right)^2 + \left(\frac{x_2 - \overline{x}}{s}\right)^2 + \cdots + \left(\frac{x_n - \overline{x}}{s}\right)^2}{n}} \\
&= \sqrt{\frac{\frac{(x_1 - \overline{x})^2}{s^2} + \frac{(x_2 - \overline{x})^2}{s^2} + \cdots + \frac{(x_n - \overline{x})^2}{s^2}}{n}} \\
&= \sqrt{\frac{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}}{s^2}} \\
&= \frac{\sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}}}{\sqrt{s^2}} \\
&= \frac{s}{s} = 1
\end{aligned}
$$

The normal approximation, I

- Different sets of data may be seen to have very similar distributions, *once they have been converted to standard units.*

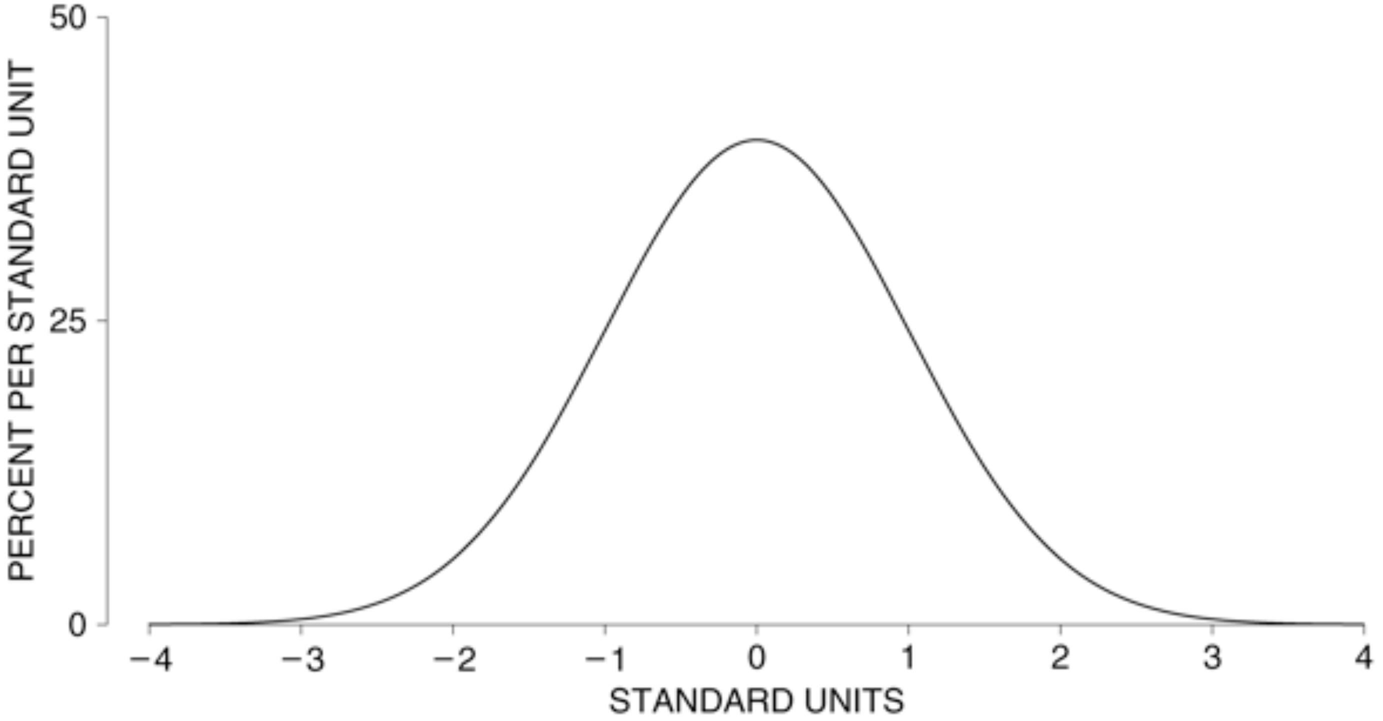
- Converting to standard units moves the center of the histogram (the average of the data) to 0, and scales the data as a whole so that one SD is converted to 1 unit.

- In many cases, the histogram of the data in standard units takes on a somehwat *bell-shaped* form — the form of the ***normal curve.***

- The normal curve is the graph of the function

$$y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

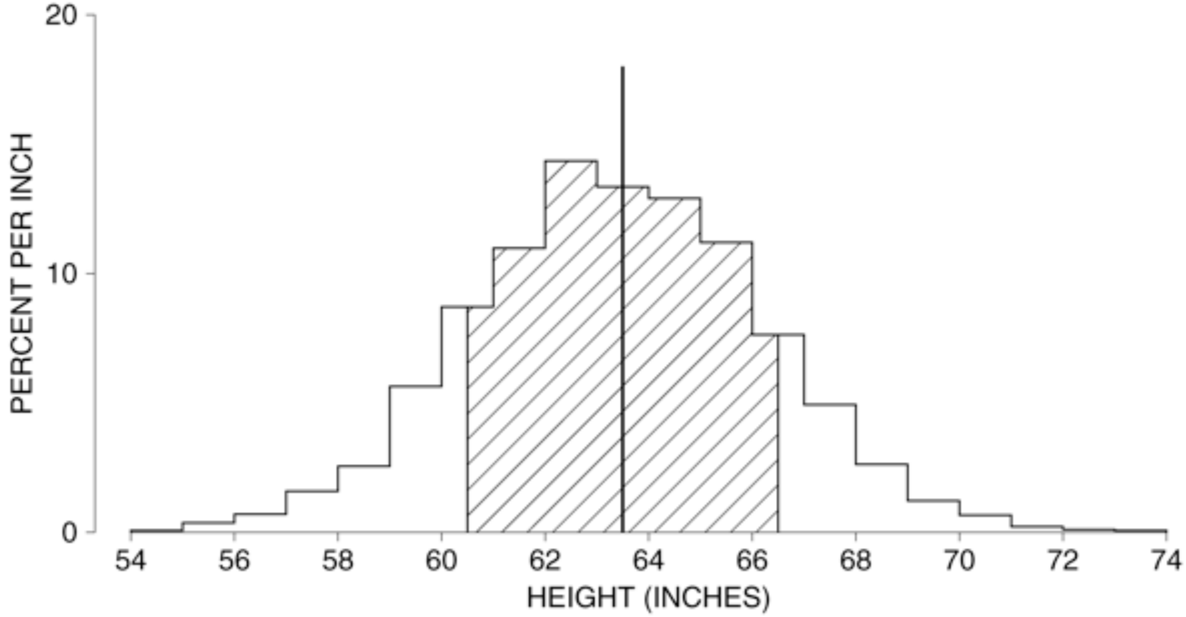
(where $e = 2.7182818\ldots$).



Statistics, Fourth Edition

The normal curve is symmetric around the line $x = 0$, and the total area under the curve is equal to 1 (or 100%, if you prefer).

Example: The distribution of heights of women age 18 and over in HANES5 (Health and Nutrition Examination Study, '03 - '04) appears in the histogram below (from page 81 in chapter 5 of FPP).

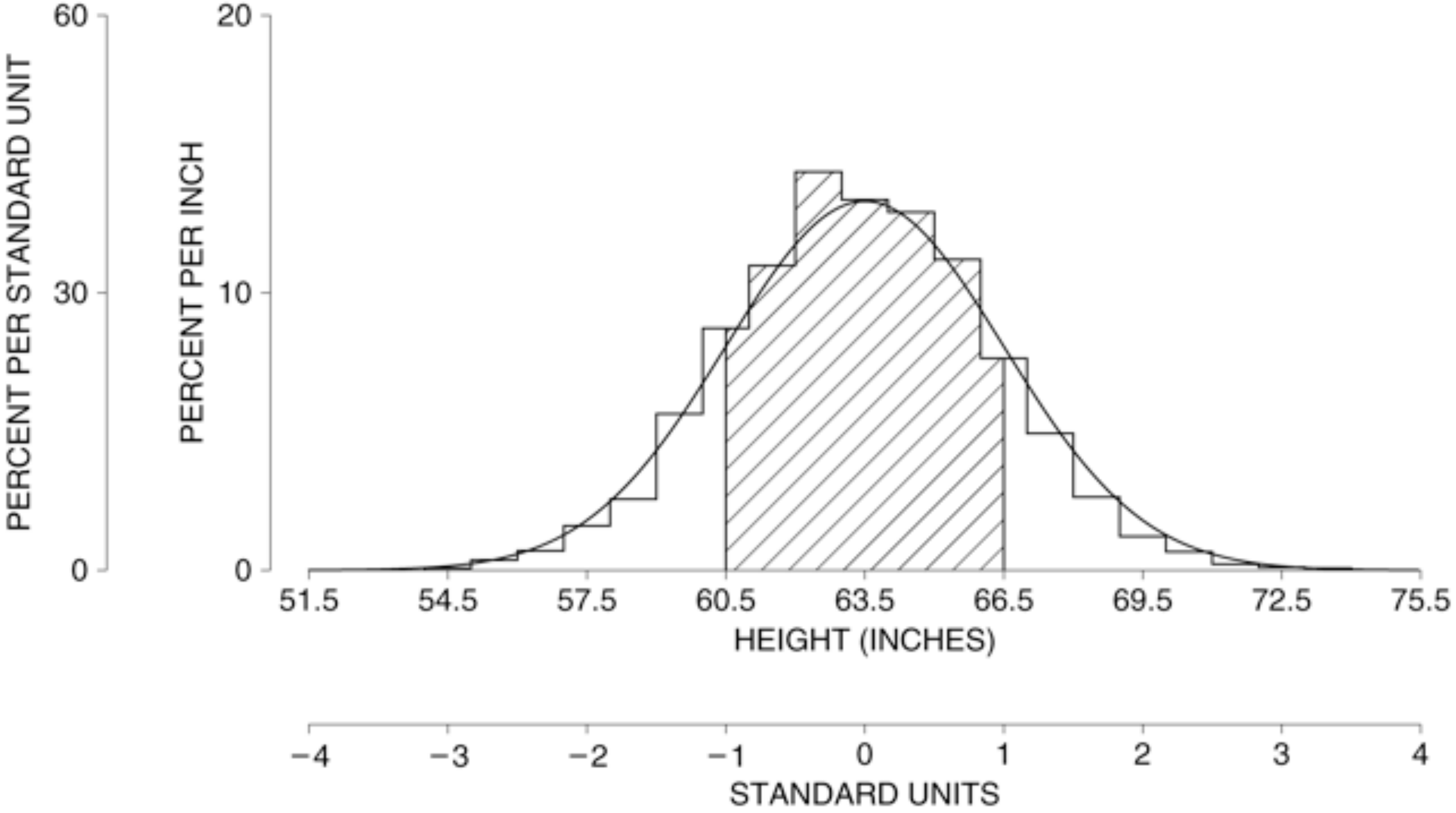


Statistics, Fourth Edition

The average height is 63.5 and the SD is about 3. The shaded region represents the heights that fall within one SD of average.

To see how well the distribution of the height data is approximated by the normal curve, we must convert the data to standard units and sketch the histogram for the ‘standardized’ (or normalized) data.

To save a lot of drawing time, we observe that the conversion to standard units is just a *rescaling.* This means that instead of actually converting all of the heights to their standard units and then drawing a new histogram, we can simply *change the horizontal and vertical scales on the original histogram.*



Statistics, Fourth Edition

- If the (rescaled) histogram is similar to the normal curve, then area of regions under the histogram will be approximately equal to areas under the normal curve for the same range of standard units.
- I.e., the percentage of the data that lies within 1 SD of the average will be approximately equal to the area under the normal curve between -1 and 1; the percentage of the data lying within 2 SDs of the average will be approximately equal to the area under the normal curve between -2 and 2; and so forth.
- This is useful, because the distribution of the area under the normal curve is well understood.
- In particular:
  - The area under the normal curve between -1 and 1 is approximately $0.68 = 68\%$;
  - The area under the normal curve between -2 and 2 is approximately $0.95 = 95\%$;
  - The area under the normal curve between -3 and 3 is approximately $0.99 = 99\%$;

‘Rule of thumb’:

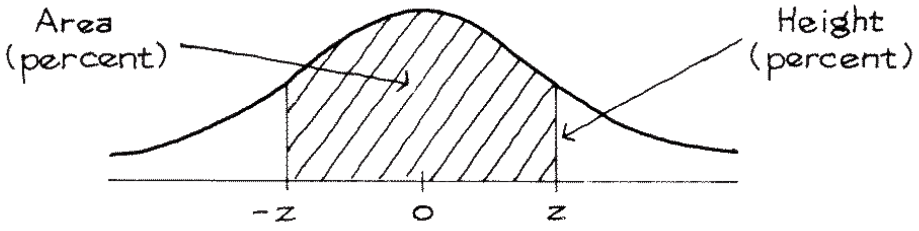
If a set of data has an approximately normal distribution, then:

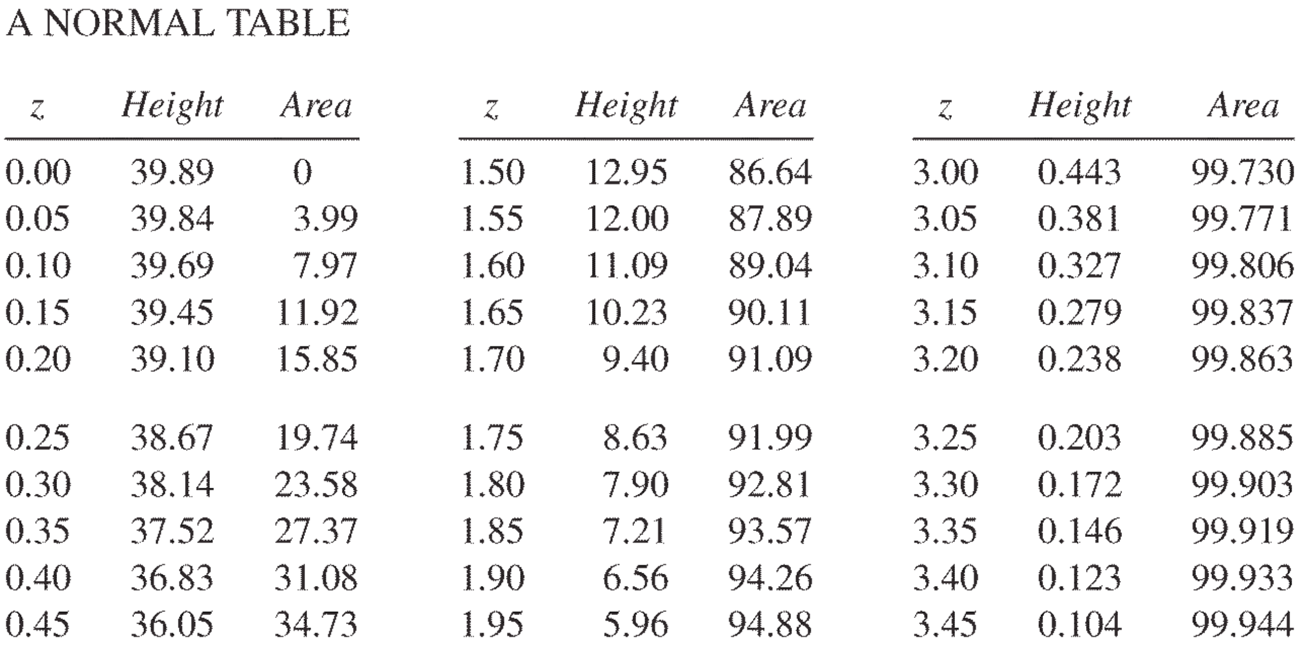
- ○ ***About*** 68% ***of the data lies within one SD of average;***
- ○ ***About*** 95% ***of the data lies within two SDs of average;***
- ○ ***About*** 99% ***of the data lies within three SDs of average;***

To calculate areas under the normal curve for regions other than those above (-1 to 1, -2 to 2 and -3 to 3), we use a ***normal table***, like the one found in the back of the textbook.

Remember: This rule ***only applies to data that is (approximately) normally distributed!*** Absent that condition, or any other assumption about how the data is distributed, we have to rely on weaker estimates, like those given by Chebyshev’s inequality.

A normal table





A NORMAL TABLE

z	*Height*	*Area*	*z*	*Height*	*Area*	*z*	*Height*	*Area*
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863
0.25	38.67	19.74	1.75	8.63	91.99	3.25	0.203	99.885
0.30	38.14	23.58	1.80	7.90	92.81	3.30	0.172	99.903
0.35	37.52	27.37	1.85	7.21	93.57	3.35	0.146	99.919
0.40	36.83	31.08	1.90	6.56	94.26	3.40	0.123	99.933
0.45	36.05	34.73	1.95	5.96	94.88	3.45	0.104	99.944

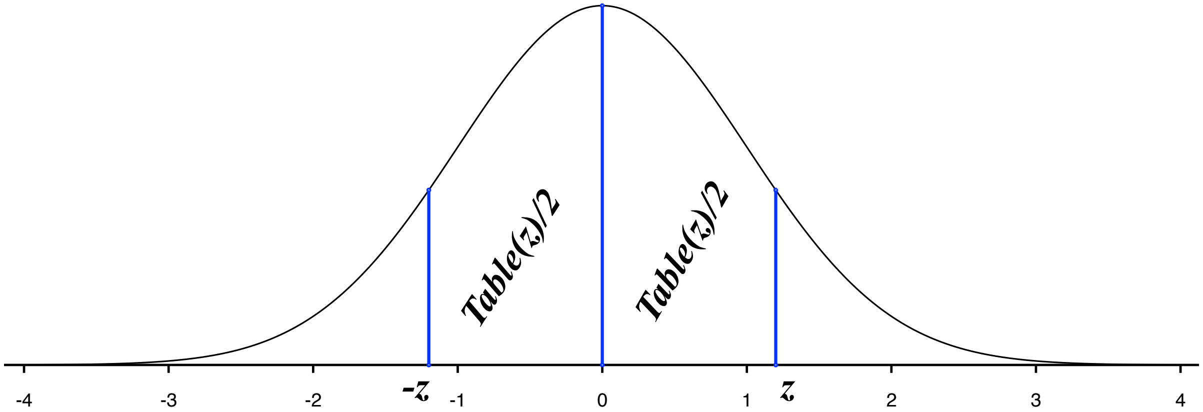
0.50	35.21	38.29	2.00	5.40	95.45	3.50	0.087	99.953
0.55	34.29	41.77	2.05	4.88	95.96	3.55	0.073	99.961
0.60	33.32	45.15	2.10	4.40	96.43	3.60	0.061	99.968
0.65	32.30	48.43	2.15	3.96	96.84	3.65	0.051	99.974
0.70	31.23	51.61	2.20	3.55	97.22	3.70	0.042	99.978
0.75	30.11	54.67	2.25	3.17	97.56	3.75	0.035	99.982
0.80	28.97	57.63	2.30	2.83	97.86	3.80	0.029	99.986
0.85	27.80	60.47	2.35	2.52	98.12	3.85	0.024	99.988
0.90	26.61	63.19	2.40	2.24	98.36	3.90	0.020	99.990
0.95	25.41	65.79	2.45	1.98	98.57	3.95	0.016	99.992
1.00	24.20	68.27	2.50	1.75	98.76	4.00	0.013	99.9937
1.05	22.99	70.63	2.55	1.54	98.92	4.05	0.011	99.9949
1.10	21.79	72.87	2.60	1.36	99.07	4.10	0.009	99.9959
1.15	20.59	74.99	2.65	1.19	99.20	4.15	0.007	99.9967
1.20	19.42	76.99	2.70	1.04	99.31	4.20	0.006	99.9973
1.25	18.26	78.87	2.75	0.91	99.40	4.25	0.005	99.9979
1.30	17.14	80.64	2.80	0.79	99.49	4.30	0.004	99.9983
1.35	16.04	82.30	2.85	9.69	99.56	4.35	0.003	99.9986
1.40	14.97	83.85	2.90	0.60	99.63	4.40	0.002	99.9989
1.45	13.94	85.29	2.95	0.51	99.68	4.45	0.002	99.9991

Reading a normal table

We first need to learn how to read the normal table and use it to calculate areas of regions under the normal curve.

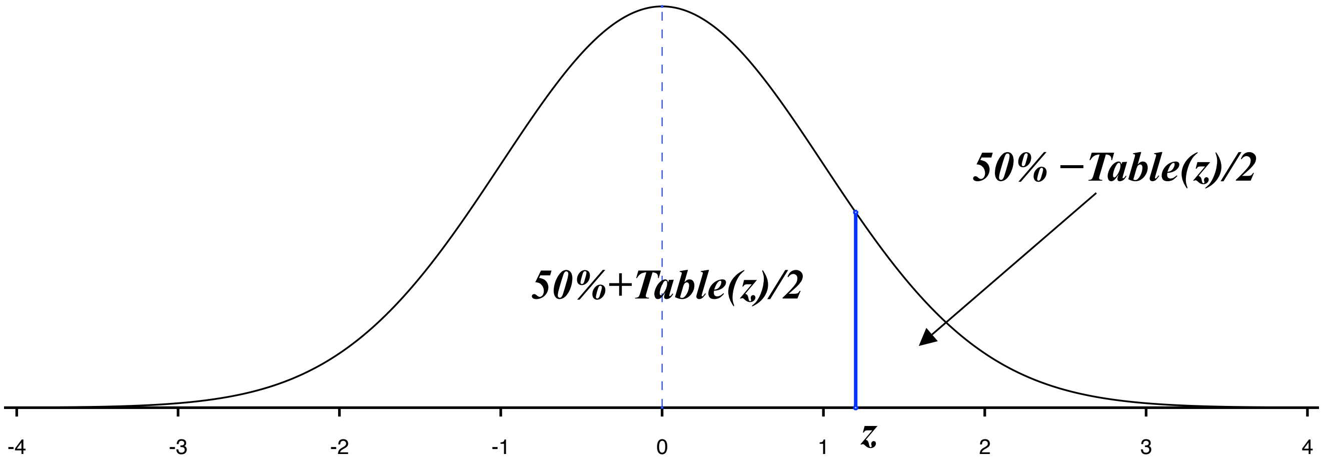
i. The table in the appendix gives the areas for regions of the form $-z \leq t \leq z$ (as *percentages*), where $0 \leq z \leq 4.45$. If $z \geq 4.50$, you can assume that the corresponding area is 99.9999%.

ii. The normal curve is ***symmetric around the vertical axis*** so the area under the curve between 0 and z is equal to the area under the curve between $-z$ and 0, and both are equal to exactly one half the table entry for z.



In particular the areas corresponding to $0 \leq t$ and $t \leq 0$ are both exactly 50%.

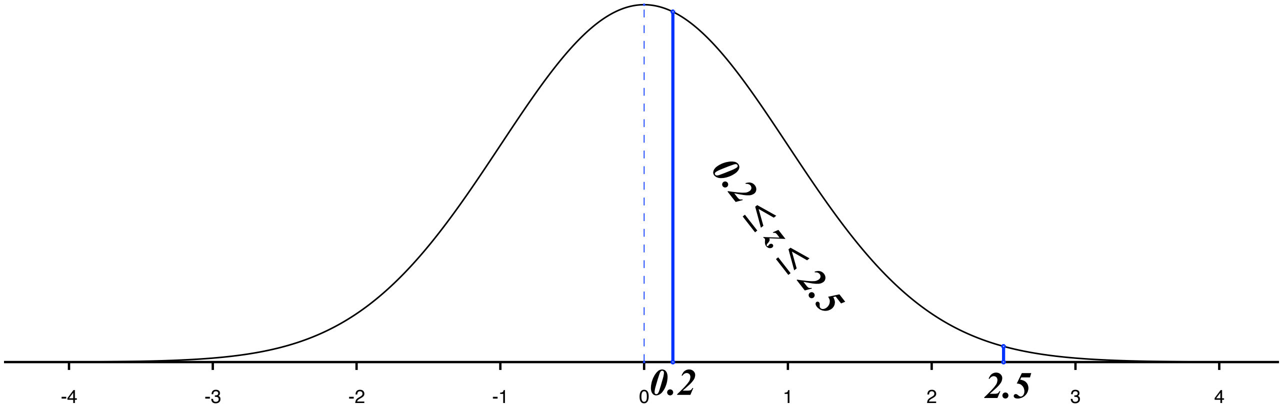
iii. If $0 < z$, then the area under the curve corresponding to the region $t \leq z$ is $50\% + Table(z)/2$.



The area corresponding to $z < t$ is $50\% - Table(z)/2$.

The areas of other regions can be calculated using these rules and similar ones, derived by sketching the appropriate region under the normal curve.

Example. Find the area of the region under the normal curve corresponding to $0.2 \le z \le 2.5$. This region appears between the two vertical lines in the figure below.



The area of this region is equal to the area of the region

$$0 \le z \le 2.5$$

minus the area of the region $0 \le z \le 0.2$, which is

$$(98.76\%)/2 - (15.85\%)/2 = 41.455\%.$$