Statistical and Mathematical Methods for Data Analysis

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- **Elementary Statistics,** 13th Edition, Mario F. Triola

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco

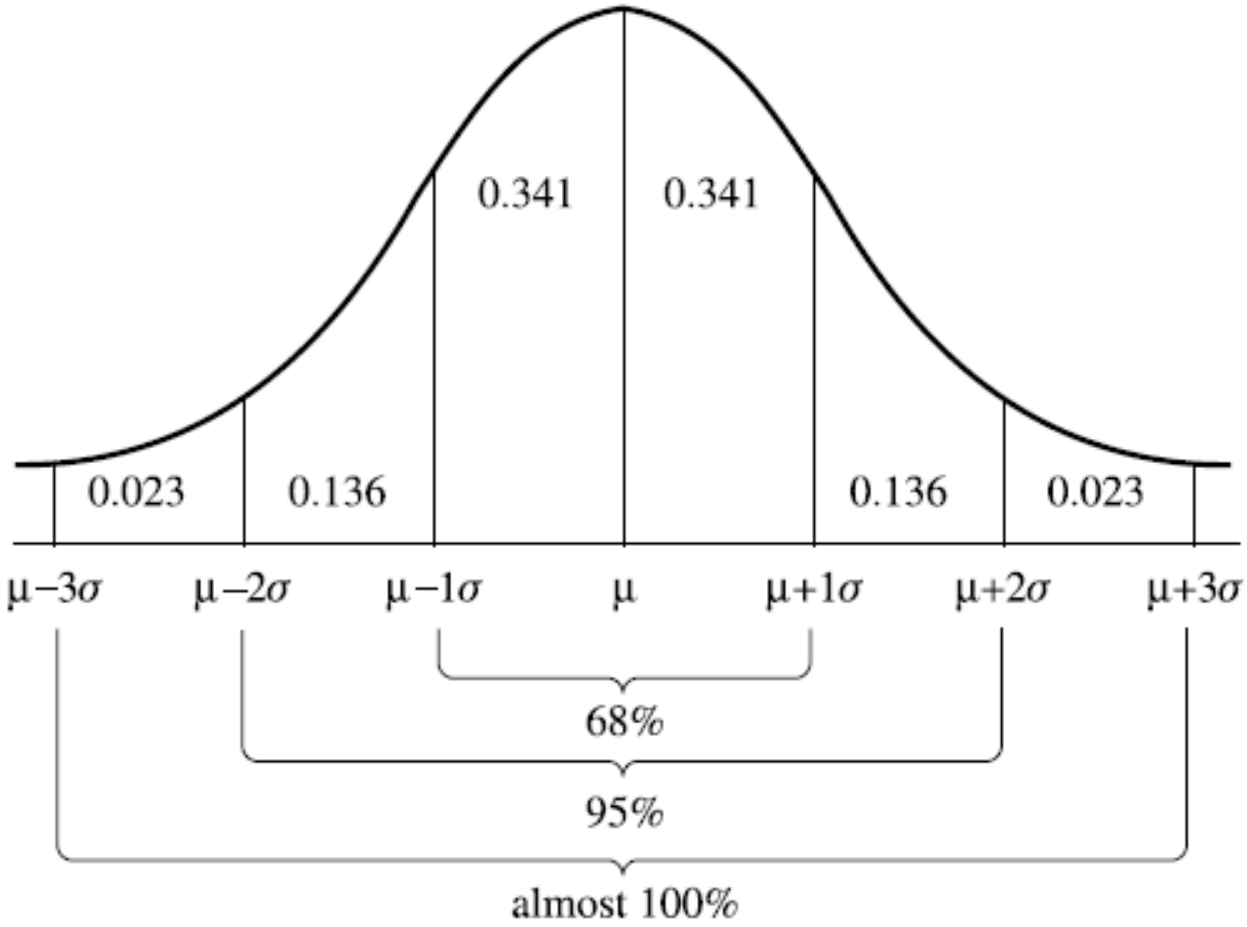
References

Readings for these lecture notes:

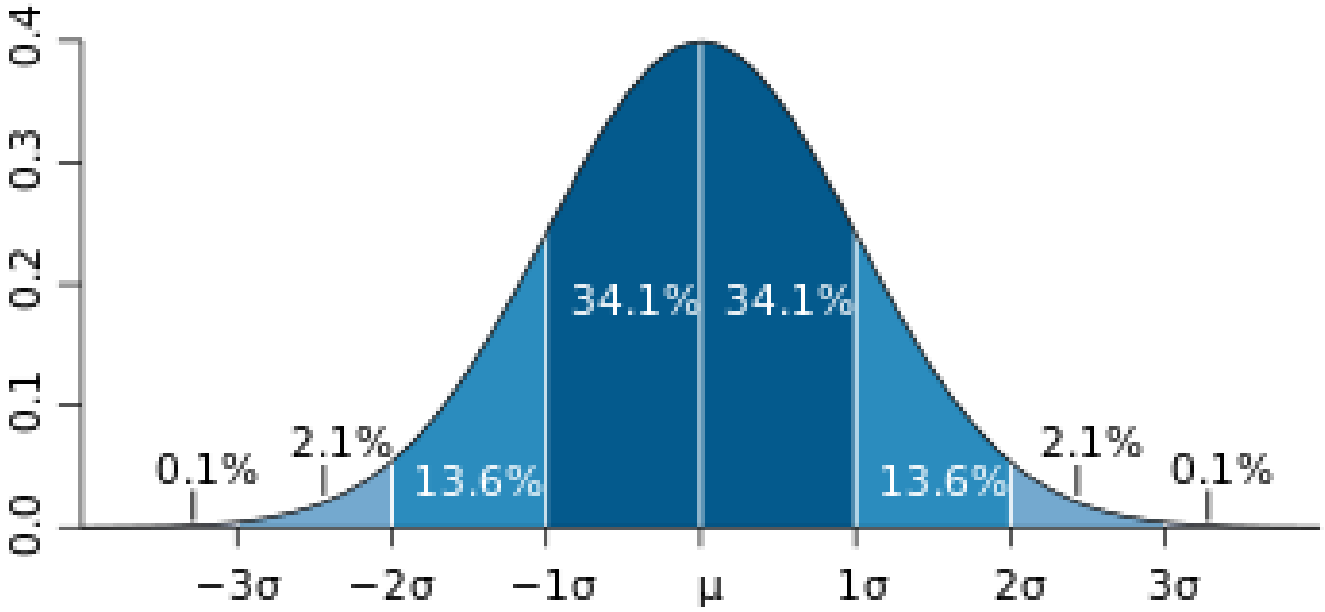
- Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- Probability Demystified, Allan G. Bluman
- Schaum's Outline of Probability, Seymour Lipschutz, Marc Lipson

These notes contain material from the above resources.

Normal Distribution [1]



Normal Distribution [2]



Normal Distribution [3]

Example 1: The mean commuting time between a person's home and office is **24 minutes**. The standard deviation is **2 minutes**. Assume the variable is normally distributed. Find the probability that it takes a person between **24 and 28 minutes** to get to work.

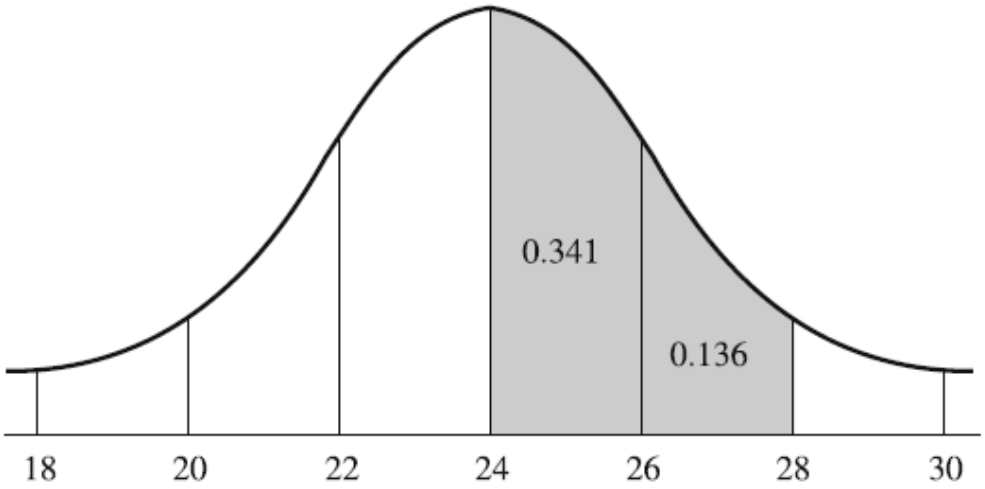
Solution:

Here $\mu = 24$

$\sigma = 2$

$\mu + 1\sigma = 24 + 1(2) = 26$

$\mu + 2\sigma = 24 + 2(2) = 28$



$$P(24 \leq X \leq 28) = P(\mu \leq X \leq \mu + 1\sigma) + P(\mu + 1\sigma \leq X \leq \mu + 2\sigma)$$

$$P(24 \leq X \leq 28) = 0.341 + 0.136 = 0.4777$$

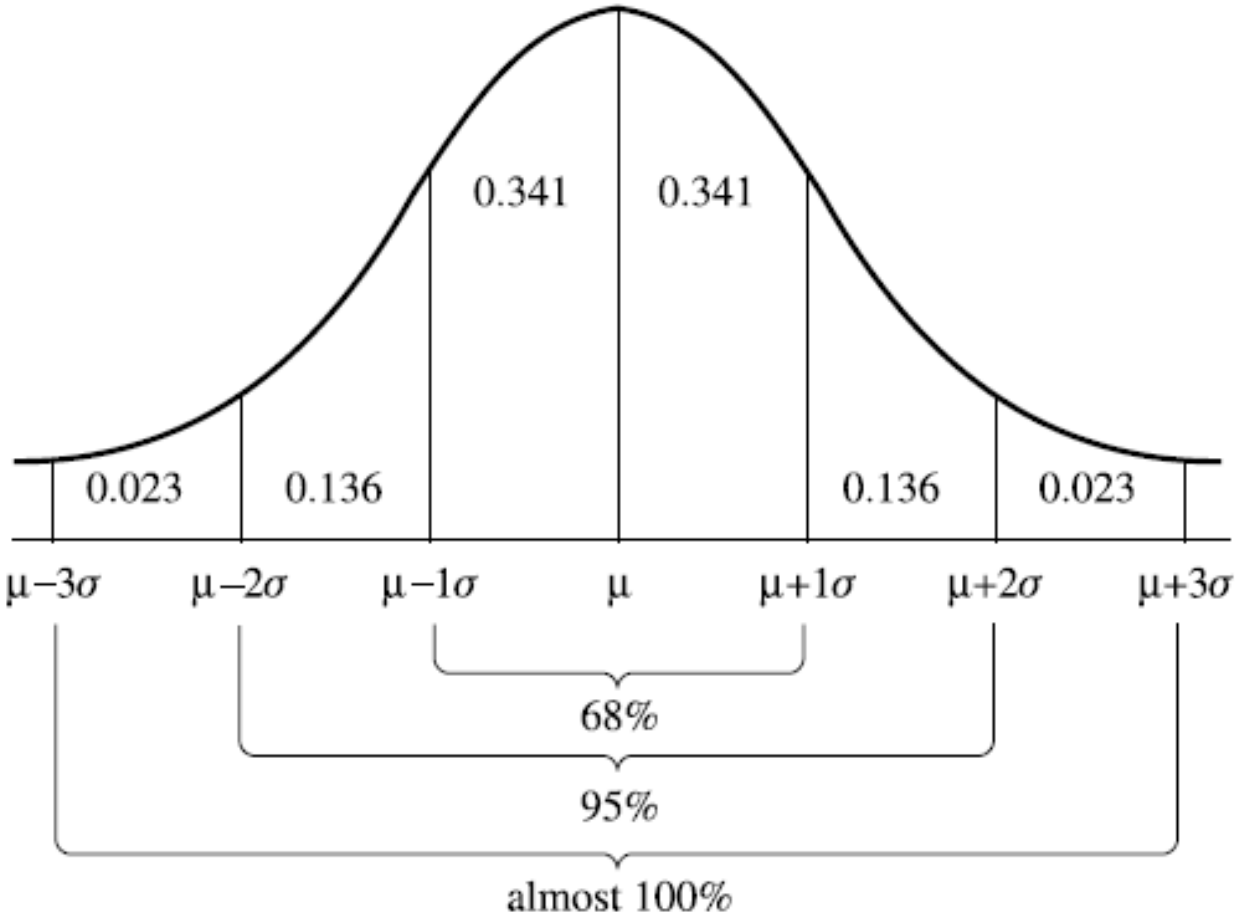
Normal Distribution [4]

Example 2: According to a study by A.C. Neilson, children between 2 and 5 years of age watch an **average of 25** hours of television per week. Assume the variable is approximately normally distributed with a **standard deviation of 2**. If a child is selected at random, find the probability that the child watched **more than 27 hours** of television per week.

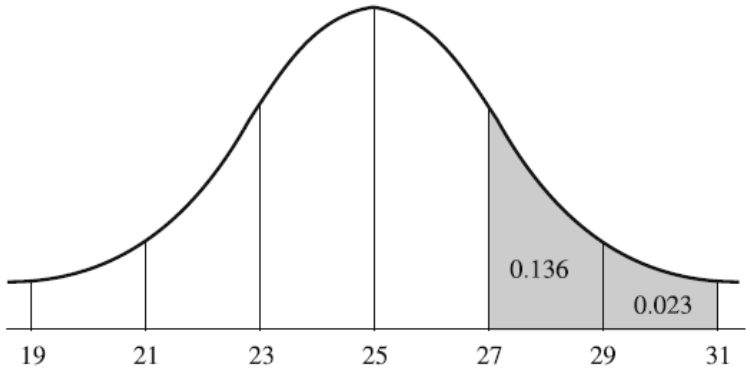
Solution

Draw the normal distribution curve and place 25 at the center; then place 27, 29, and 31 to the right corresponding to one, two, and three standard deviations above the mean, and 23, 21, and 19 to the left corresponding to one, two, and three standard deviations below the mean. Now place the areas (percent) on the graph. See figure in the next slide

Solution cont.



Solution cont.



$\mu + 1\sigma = 25 + 1(2) = 27$

$\mu + 2\sigma = 25 + 2(2) = 29$

$\mu + 3\sigma = 25 + 3(2) = 31$

$P(X > 27) = P(\mu \leq X \leq \mu + 3\sigma) - P(\mu \leq X \leq \mu + 1\sigma)$

$P(X > 27) = \mathbf{0.5} - 0.341$

$= 0.159$

OR

$P(X > 27) = P(\mu + 1\sigma \leq X \leq \mu + 2\sigma) + P(\mu + 2\sigma \leq X \leq \mu + 3\sigma)$

$P(X > 27) = 0.136 + 0.023$

$= 0.159$

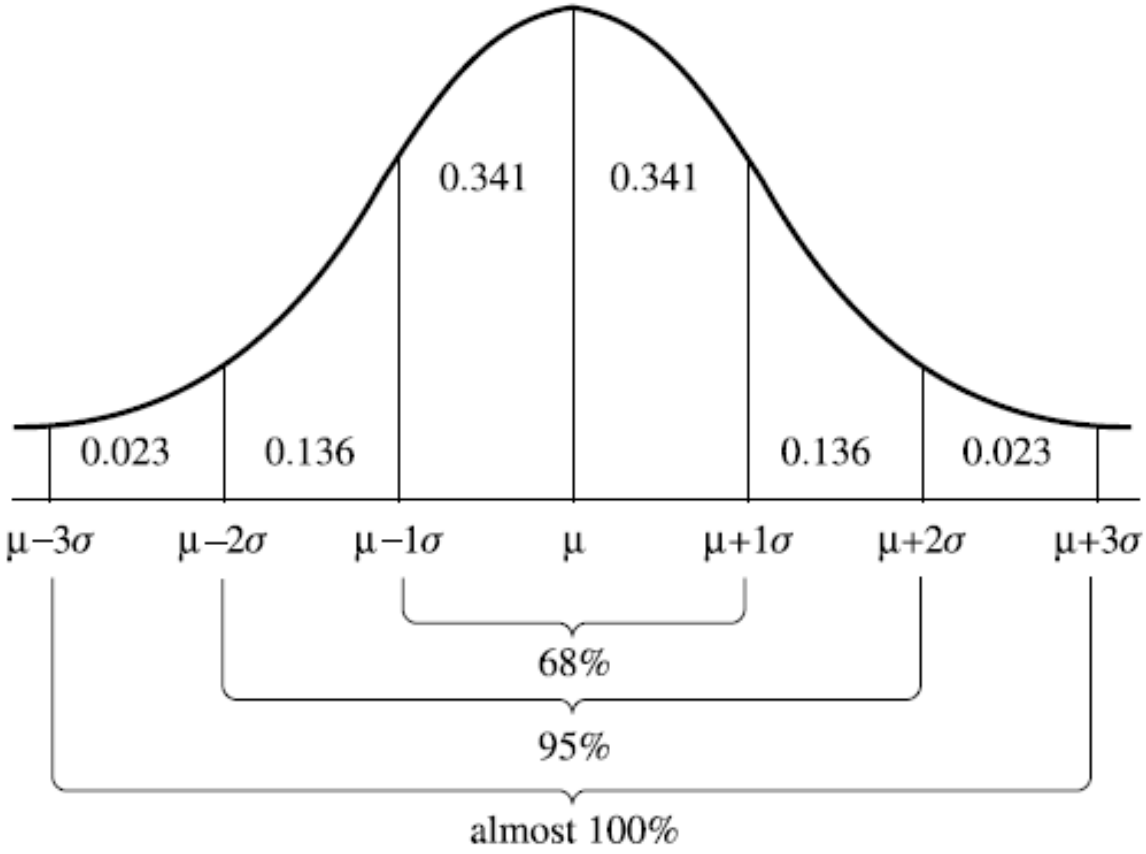
Normal Distribution [5]

Example 3: The scores on a national achievement exam are normally distributed with a **mean of 500** and a **standard deviation of 100**. If a student who took the exam is randomly selected, find the probability that the student scored **below 600**.

Solution

Draw the normal distribution curve and place 500 at the center. Place 600, 700, and 800 to the right and 400, 300, and 200 to the left, corresponding to one, two, and three standard deviations above and below the mean respectively. Fill in the corresponding areas. See Figure 9-4.

Solution cont.



Solution cont.

$P(X < 600)$ = **$P(\mu - 3\sigma \leq X \leq \mu) + P(\mu \leq X \leq \mu + 1\sigma)$**

= 0.50 + 0.341

= 0.841

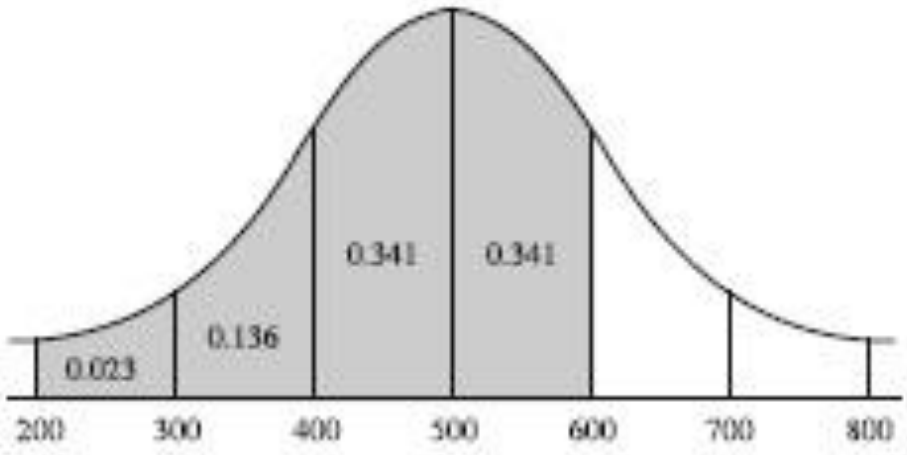


Fig. 9-4.

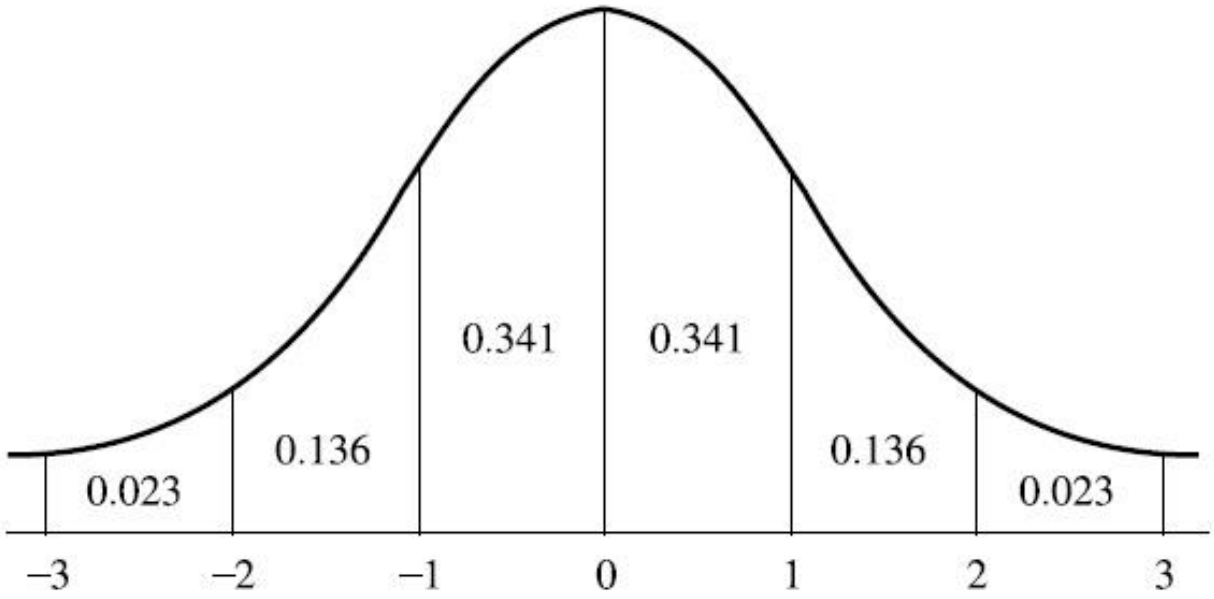
The Standard Normal Distribution [1]

- The normal distribution can be used as a model to solve many problems about variables that are approximately normally distributed.

- Since each variable has its **own mean** and **standard deviation**, statisticians use what is called the **standard normal distribution** to solve the problems.

The Standard Normal Distribution [2]

The **standard normal distribution** has all the properties of a normal distribution, but the **mean is zero** and the **standard deviation is one**.



Areas under the Normal Curve [1]

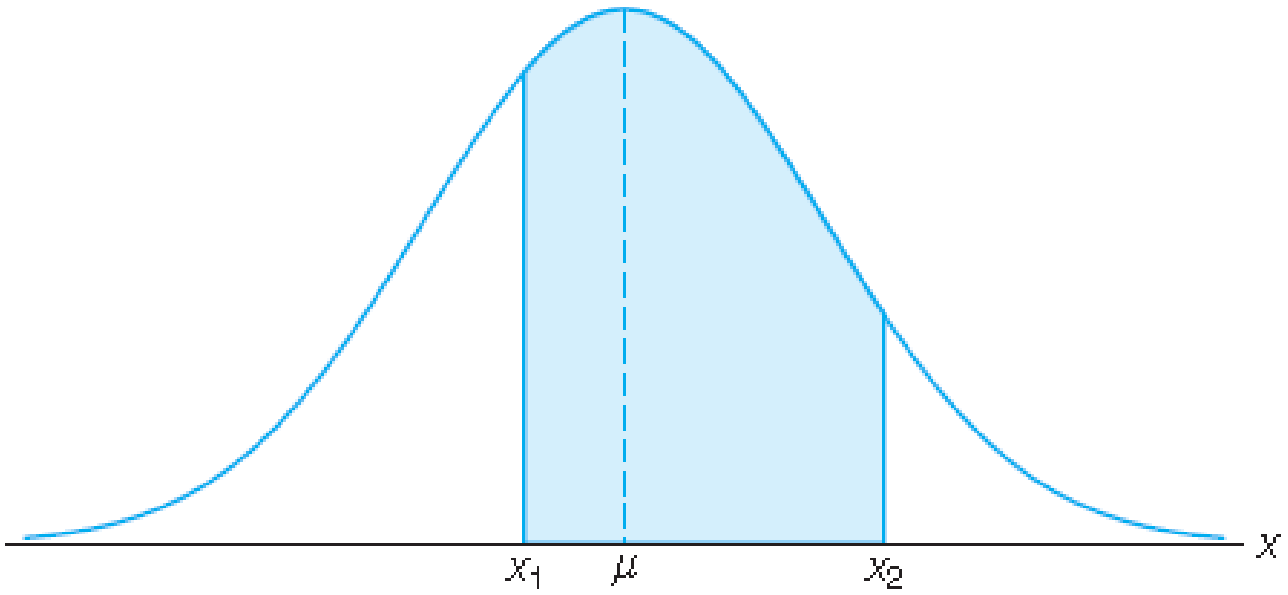
$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad , -\infty \leq x \leq +\infty$$

Here $\Pi = 3.1416$, $e = 2.7183$

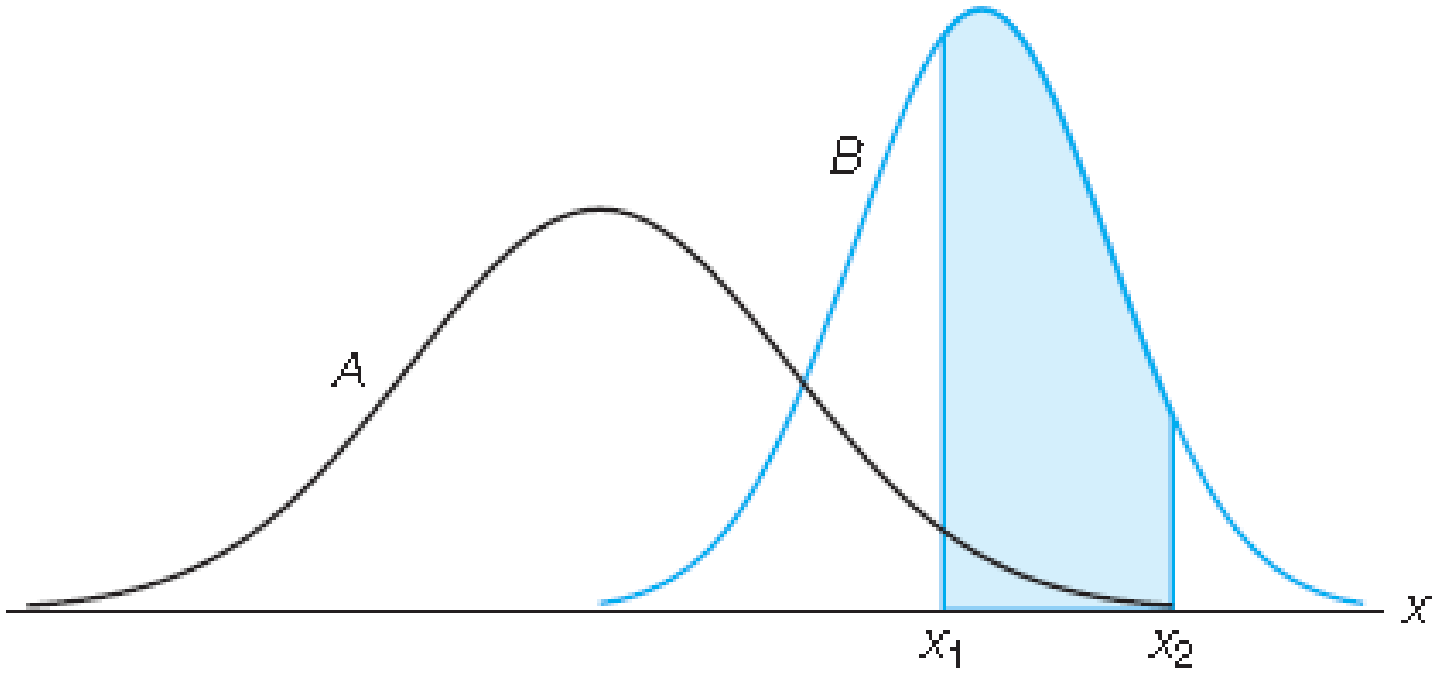
Areas under the Normal Curve [2]

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)\,dx$$

$$= \frac{1}{\sigma\sqrt{2\Pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\,dx$$



$P(x_1 \leq x \leq x_2)$ = Area of shaded region



$P(x_1 \leq x \leq x_2)$ for different normal curves

Standard Normal Value [1]

We saw how the normal curve is **dependent** on **the mean and the standard deviation** of the distribution under investigation. The area under the curve between any two ordinates must then also depend on the values μ and σ.

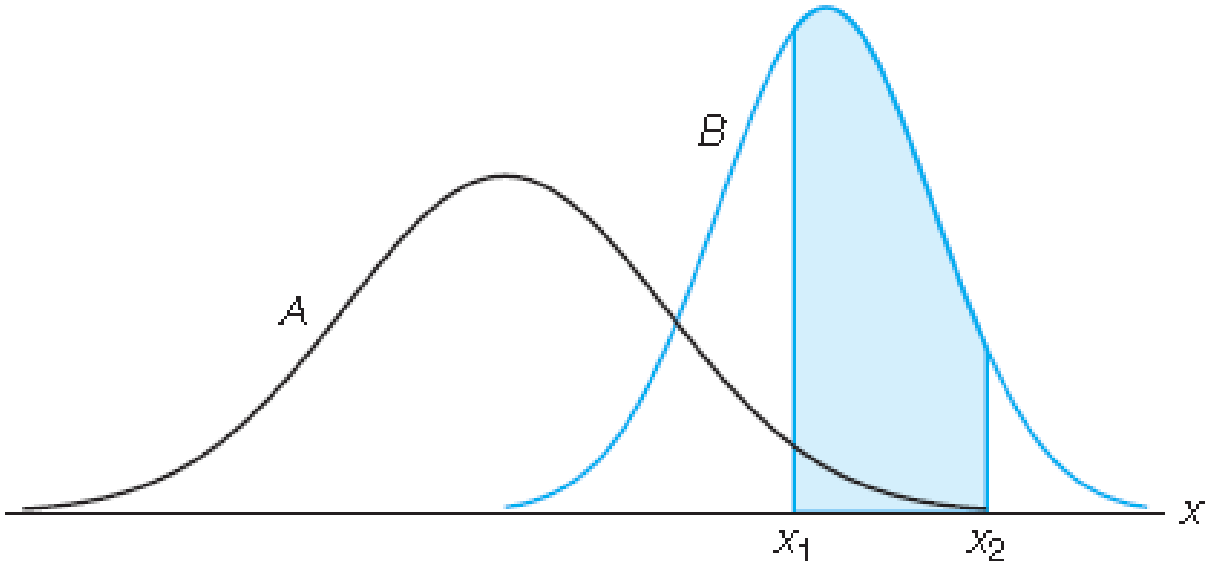
This is evident in from pervious figure, where we have shaded regions corresponding to $P(x_1 < X < x_2)$ for two curves with different means and variances.

Standard Normal Value [2]

There are many types of statistical software that can be used in calculating areas under the normal curve. The difficulty encountered in solving integrals of normal density functions necessitates the tabulation of normal curve areas for quick reference. However, **it would be a hopeless task to attempt to set up separate tables for every conceivable value of μ and σ**.

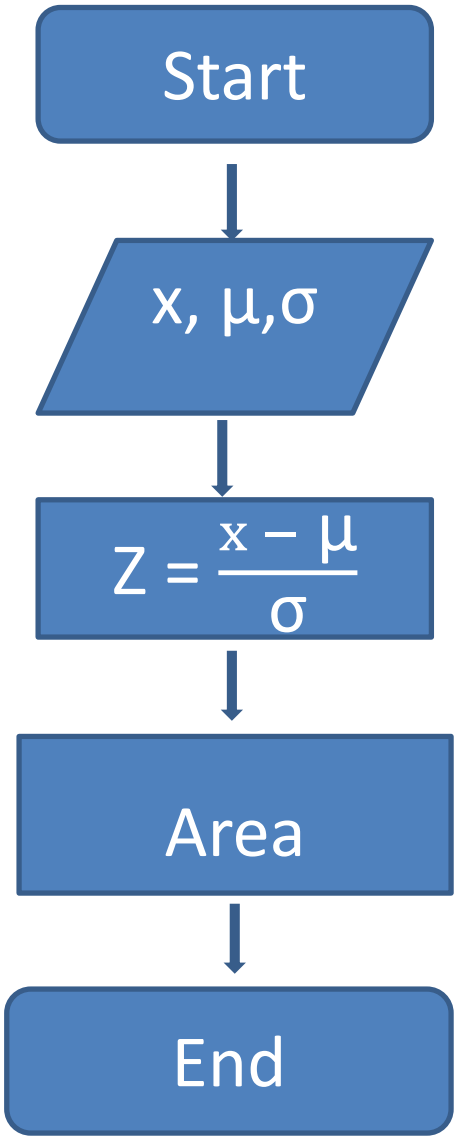
Standard Normal Value [3]

Fortunately, we are able to transform all the observations of any normal random variable *X* into a new set of observations of a normal random variable *Z* with mean 0 and variance 1.

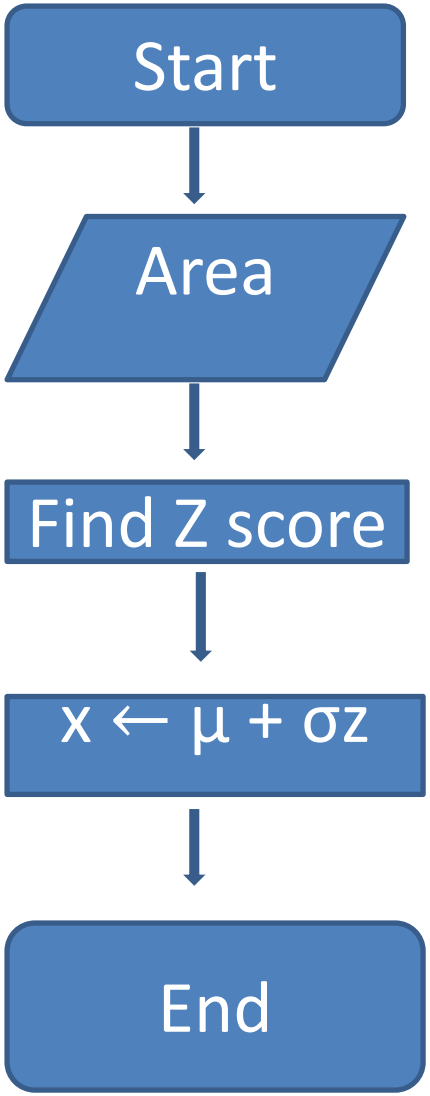


$P(x_1 \leq x \leq x_2)$ for different normal curves

Flowchart for Computing Probability



Flowchart for Computing x (Inverse Problem)



Standard Normal Value [3]

The distribution of a normal random variable with **mean 0** and **variance 1** is called a **standard normal distribution**.

A value for any variable that is approximately normally distributed can be transformed into a standard normal value by using the following formula:

$$Z = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}}$$ **OR**

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

The standard normal values are called **z values or z scores**.

Z values or Z scores [1]

Example: Find the corresponding z value for a value of 18 if the mean of a variable is 12 and the standard deviation is 4.

Solution:

$$Z = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}}$$

OR

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

$$Z = \frac{18 - 12}{4} = 1.5$$

Hence the z value of 1.5 corresponds to a value of 18 for an approximately normal distribution which has a mean of 12 and a standard deviation of 4. Z values are negative for values of variables that are below the mean Z

Z values or Z scores [2]

Example: Find the corresponding z value for a value of **9** if the **mean** of a variable is **12** and the standard deviation is 4.

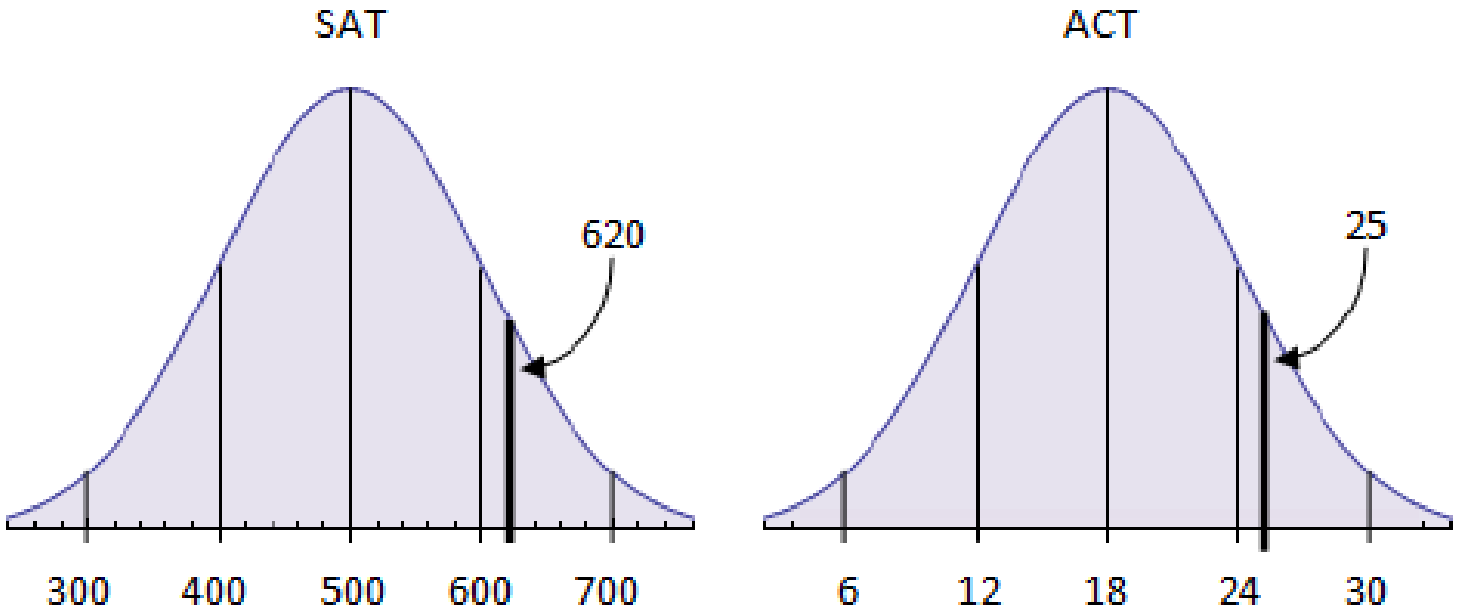
Solution:

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

$$Z = \frac{9 - 12}{4} = -0.75$$

Example Suppose a student can either submit only their **SAT (Scholastic Aptitude Test)** score or their **ACT (American College Testing)** score to a particular college. Suppose their **SAT** score was **620** and that the **SAT** has a **mean** of **500** and a **standard deviation** of **100**. Suppose also that the same student scored a **25** of their **ACT exam** and that the **ACT** exam has a **mean** of **18** and a **standard deviation** of **6**. Which score should the student submit?

Solution:



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

$$Z_{SAT} = \frac{620 - 500}{100}$$

$$= 1.2$$

$$Z_{ACT} = \frac{25 - 18}{6}$$

$$= 1.17$$

Since the **z-score** is higher on the **SAT**, the student should submit the **SAT exam score.**

Area u

Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Area under the Normal Curve [2]

Table A.3 (continued) Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

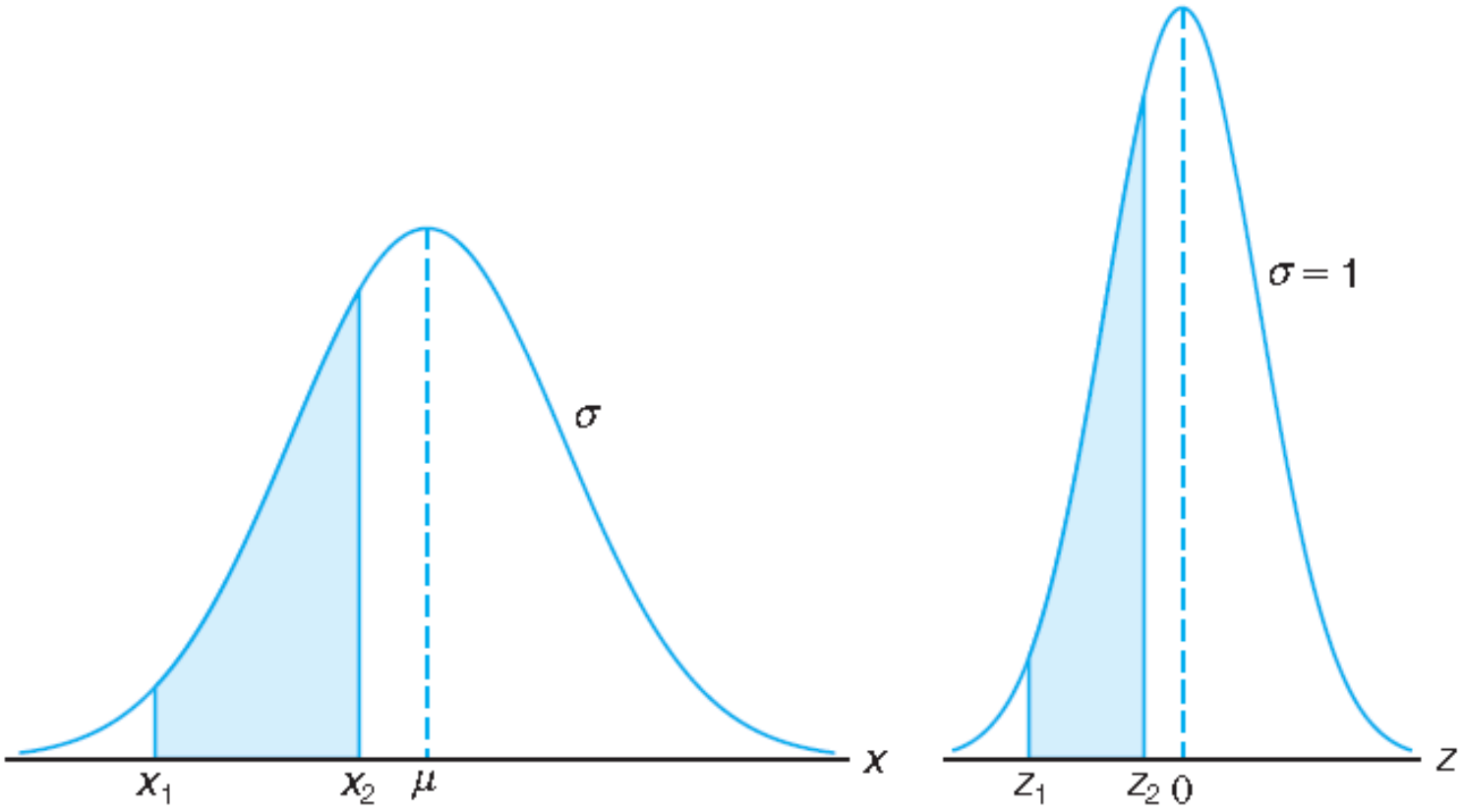
Area under the Normal Curve [3]

Table A.3 indicates the area under the standard normal curve corresponding to **P(Z < z)** for values of *z* ranging from **-3.49 to 3.49.**

To illustrate the use of this table, let us find the probability **that *Z* is less than 1.74**. First, we locate a value of *z* equal to **1.7** in the left column, then move across the row to the column under 0.04, where we read 0.9591. Therefore, **P(Z < 1.74) = 0.9591**.

To find a *z* value corresponding to a given probability, the process is **reversed**. For example, the *z* value leaving an x area of **0.2148** under the curve to the left of *z* is seen to be **-0.79**.

The Original and Transformed Normal Distributions [1]



The original and transformed normal distributions