

Normal Distributions

The Normal Curve

Normal Probabilities

Percentiles and the Normal Curve

The Central Limit Theorem

The Normal Curve

In a **normal distribution**, data values are more frequent closer to the mean. Exact frequencies of data values can be calculated based on number of standard deviations above the mean, called **z**.

68% of a normal curve is within one standard deviation of the mean: $P(-1 < z < 1) \approx 68\%$.

95% of a normal curve is within two standard deviations of the mean: $P(-2 < z < 2) \approx 95\%$.

Almost all of a normal curve is within three standard deviations of the mean: $P(-3 < z < 3) \approx 99.7\%$.

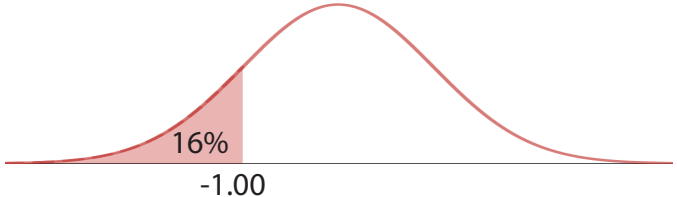
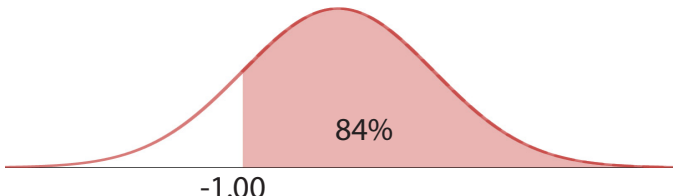
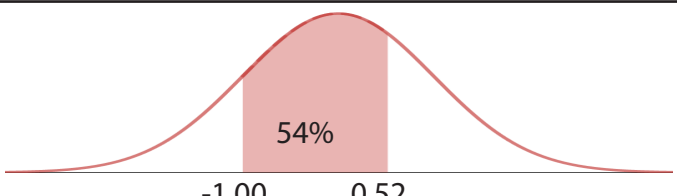
Real-world data sets typically are approximately normal, allowing for calculations of frequency or probability if the mean and standard deviation are known. Some examples are shown below.

Variable	Parameters	68% of values	Curve
IQ scores	$\mu = 100, \sigma = 15$	$85 < x < 115$	
PreCalc final	$\mu = 158, \sigma = 29$	$129 < x < 187$	
Newborn girls (kg)	$\mu = 3.4, \sigma = 0.5$	$2.9 < x < 3.9$	

Z Scores

The **z score** of a data value x is the number of standard deviations it is above the mean: $z = \frac{x - \mu}{\sigma}$.

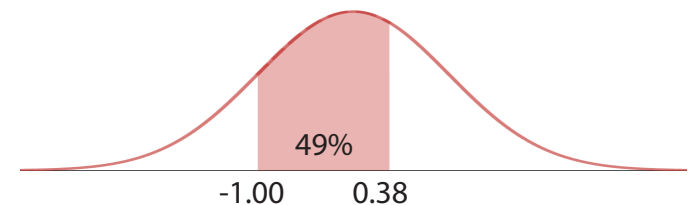
The area under a normal curve below a given z score can be found in a z table or by a graphing calculator or spreadsheet function. Since this is a probability, it is expressed with probability notation: $P(z < z_1)$.

Range	Area	Explanation	Example
Below z_1	$P(z < z_1)$	given in table or as function result	 <p>$P(z < -1.00) \approx \mathbf{16\%}$</p>
Above z_1	$1 - P(z < z_1)$	complement of below z_1	 <p>$P(z > -1.00) \approx 100\% - 16\% = \mathbf{84\%}$</p>
Between z_1 and z_2	$P(z < z_2) - P(z < z_1)$	everything below z_2 except what is also below z_1	 <p>$P(-1.00 < z < 0.52) \approx 70\% - 16\% = \mathbf{54\%}$</p>

Calculating Normal Percentages From Raw Scores

Percentages of the normal curve can be calculated by converting the raw scores to z scores so they can be looked up: $X \rightarrow Z \rightarrow \%$.

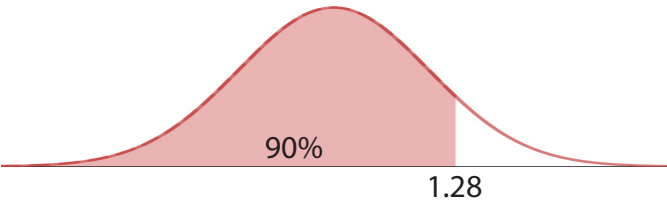
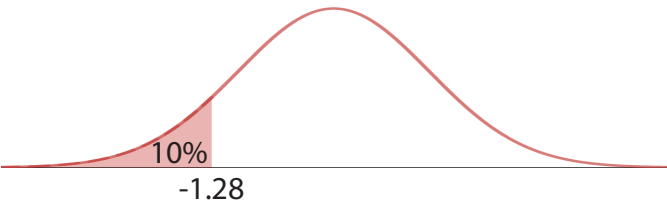
The example shown finds the percentage of fish weighing between 600 grams and 710 grams, given a normal distribution with mean 680 grams and standard deviation 80 grams.



Step	This example
Convert the raw scores to z score(s).	$z_1 = \frac{600 - 680}{80} = -1.00$ $z_2 = \frac{710 - 680}{80} \approx 0.38$
Find the area under the normal curve below the z score(s).	$P(z < -1.00) \approx 16\%$ $P(z < 0.38) \approx 65\%$
If above a score, take the complement. If between scores, subtract.	$P(-1.00 < z < 0.38) \approx 65\% - 16\% = 49\%$
State the answer with respect to the original question.	$P(600 < x < 710) \approx 49\%$

Percentiles

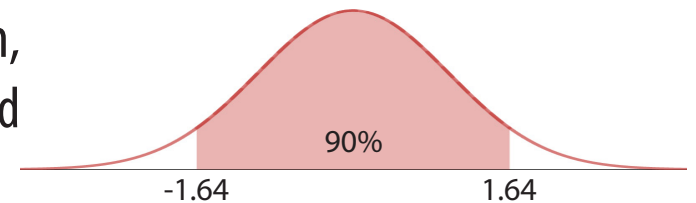
A percentile can be found in a normal distribution by identifying the z score needed for that percentile and converting the z score to a raw score: $x = \mu + z\sigma$. The examples below are for SAT I math tests, which are designed to have a mean of 500 and a standard deviation of 100.

Percentile	Z score	Raw score	Curve
90 th	$z \approx 1.28$	$x \approx 500 + 1.28(100) = 628$	
10 th	$z \approx -1.28$	$x \approx 500 - 1.28(100) = 372$	

Calculating Raw Scores From Normal Percentages

Raw scores (with units) can be calculated by looking up z scores from percentages and converting the z scores to raw scores: $X \leftarrow Z \leftarrow \%$.

Find the range of weights that includes the middle 90% of fish, given a normal distribution with mean 680 grams and standard deviation 80 grams.



Step	This example
Identify the percentile(s).	The middle 90% goes from the 5 th percentile to the 95 th percentile.
Find the z score(s) that have this much area below them under the normal curve.	$P(z < -1.64) = 5\%$ $P(z < 1.64) = 95\%$
Convert the z score(s) to raw score(s).	$x_1 = 680 - 1.64(80) \approx 549$ $x_2 = 680 + 1.64(80) \approx 811$
State the answer with respect to the original question.	$P(549 < x < 811) = 90\%$

The Purpose of This Course

This course involves a lot of calculations by hand. However, outside of a school setting, these values are usually calculated by statistical software rather than by hand or calculator. The purpose of this course is not to teach you how to calculate statistics, but rather to understand what the calculations represent and how the different components affect the result. Learning the processes initially by hand, instead of simply knowing a series of steps on a calculator, can provide this understanding, allowing for:

- Figuring out how to carry out the process in any context, such as in new software programs.
- Understanding the notation and terminology, enabling meaningful conversations and communication of ideas.
- Recognizing when a result is unreasonable or remarkable, and being able to analyze what aspects of the inputs may have led to this.

To help achieve these objectives, the following steps are required for normal curve calculations in this course:

Step
Shade the percentage given in the problem or the percentage representing the answer.
Label each boundary of the shaded area with its z score.
Show every step in the work, including values from the normal table.
Other than on the curve, only write equations.
Include x , \bar{x} , or z in every equation.
End with an equation showing a probability as equal to the percentage of the curve that is shaded.

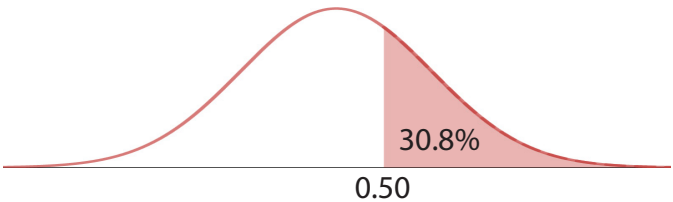
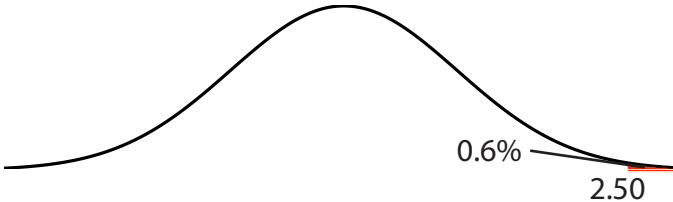
The Law of Large Numbers

The **law of large numbers** states that the larger a sample is, the closer its statistics tend to be to the actual population parameters. As a result, as sample size increases, a sample statistic becomes more likely to fall within a given range that includes the population parameter.

Scenario example	Range Considered	Population Parameter	Effect of sample size
Percentage of heads out of n coins	More than 30% are heads.	$p = .50$	$\hat{p} > .30$ includes $p = .50$, so it becomes more likely as the sample size increases.
Average weight of freshmen boys	The average weight is between 80 and 100 pounds.	$\mu = 112$ pounds	$80 < \bar{x} < 100$ does not include $\mu = 112$, so it becomes less likely as the sample size increases.

Sampling Distributions

A **sampling distribution** is a distribution of a statistic, such as \bar{x} , for a given sample size and population. Based on the law of large numbers, the **central limit theorem** states that the larger the sample size is, the closer a sampling distribution will be to a normal distribution, and the less variation from the mean there will be in the samples. The standard deviation of the sampling distribution of means, called the **standard error** of the mean, can be estimated as $\sigma_{\bar{x}} \approx \frac{\sigma}{\sqrt{n}}$.

Event	Z score	Probability	Curve
A random student has an SAT math score above 550.	$z = \frac{550 - 500}{100} = 0.5$	$P(x > 550)$ $= P(z > 0.5)$ $\approx 1 - .692$ $= 30.8\%$	
The average of 25 random students' SAT math scores is above 550.	$z = \frac{550 - 500}{100/\sqrt{25}} = 2.5$	$P(\bar{x} > 550)$ $= P(z > 2.5)$ $\approx 1 - .994$ $= 0.6\%$	

As suggested by the law of large numbers, it would not be unusual for an individual to have a score above 550, but a group of 25 individuals is very likely to have a mean score around $\mu = 500$, and thus very unlikely to have a mean above 550.