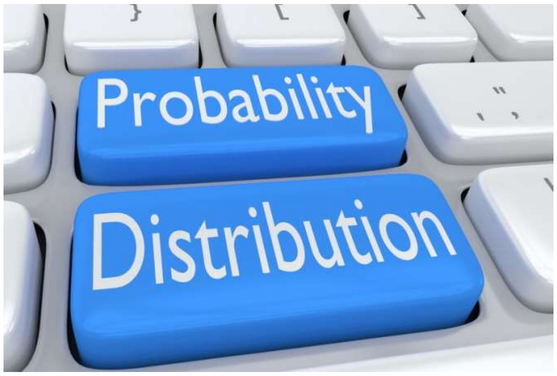
Probability Distributions

Misunderstanding of probability may be the greatest of all impediments to scientific literacy - **Stephen Jay Gould**



To that end, Probability is a foundational topic in Statistics and if it is miss-understood will certainly lead to problems down the road.

To that end, introduce this chapter with a quick review.

Part 1 is an Introduction to Probability Distributions

So we start by answering the question - *What is a Probability Distribution*. Then we discuss how your *data types* impact your probability distributions.

Then we go through a quick example by *creating a probability distribution from scratch*.

Finally, we wrap up this section by introducing the concept of a *cumulative probability distribution*.

Part 2 of this chapter is on the topic of Continuous Probability Distributions.

The includes the Normal Distribution, The Uniform Distribution, The Bivariate Normal Distribution, The Lognormal Distribution, The Exponential Distribution, The Weibull Distribution, The Chi-Squared Distribution, The Student T Test Distribution & the F Test Distribution.

Within this section we will discuss the **common applications** for each distribution along with the general **shape** of each.

We will also discuss how to calculate the **expected value** (Mean) and **variance** for each distribution - where applicable.

Lastly, we will review the **probability calculations** for each distribution & show an **example** of those calculations.

Part 3 of this chapter is on the topic of Discrete Distributions

This includes The Binomial Distribution, The Poisson Distribution, The Hypergeometric Distribution & The Multinomial Distribution.

Similar to part 2, we will review these distributions common applications, shape, expected value calculations, and do example of common probability calculations.

Part 1 - An Introduction to Probability Distributions

What is a Probability Distribution?

Let's start off with a quick definition of a Probability Distribution.

*A **Probability Distribution** is a formula, table or graph that defines the probability associated with each possible outcome in an experiment.*

When performing an experiment, we're generally measuring what statisticians call a **random variable**.

You'll recall from the previous chapter on Probability that all experiments have a **sample space** - which is the combination of all possible outcomes.

Each of these possible outcomes has a **probability of occurrence** which can be defined in a table, graph or formula which makes calculating the probability easier.

So in general, you'll see us referring to "**X**" - this is generally the **random variable** that's being studied.

You'll also see **P(X)**, which is the **probability distribution** and is generally followed by an equation that defines the probability of X.

This is where we find the true value & benefit of a probability distribution. In the ability to quickly & easily analyze a distribution and calculate the probability of occurrence of different scenarios. Some common probability scenarios include:

- the probability of occurrence of all values less than x,
- the probability of occurrence of all values or greater than x,
- the probability that x is between two values (a & b).

Data Types and their Distributions

Recall from chapter 1 (Collecting & Summarizing Data), that there are **2 types of data** - **Discrete & Continuous**.

These 2 different data types lay the foundation for the 2 different types of probability distributions, those for Continuous data and those for Discrete data.

Continuous Data Distributions

Continuous data is data that exists on a continuous scale. Continuous data can take on almost any numerical value and can be broken down into meaningful increments, etc. This includes measurements like weight, or length or temperature, time, cost, etc.

When a random variable is considered **continuous** data, we must use a continuous distribution like the **Normal, Uniform, Bivariate Normal, Exponential, Lognormal, Weibull, Chi Squared, Student T's, & F Distribution.**

Discrete Data Distributions

The 2nd form of data is called **Discrete Data** which can be thought of as Attribute or Counted Data.

This includes things like the number of Defects per lot, Pass/Fail data, True/False Data, the count conforming/non-conforming product, the number of Defectives per Lot, etc.

When the random variable is considered **discrete** data, we can use a variety of discrete distributions to describe that data, like the **Binomial, Poisson, Hypergeometric & Multinomial distributions**.

Probability Distribution Example

To show you how this all works, let's work through an example to demonstrate the relationship between a **random variable** X and its **probability distribution**.

Suppose you perform an experiment by flipping a coin 4 times. Below is the sample space for this experiment where we can see there are 16 possible outcomes:

HHHH	**HHHT**	**HTHH**	**HHTH**
HTTH	**HHTT**	**HTHT**	**THHH**
THHT	**THTH**	**TTHH**	**TTTH**
HTTT	**THTT**	**TTHT**	**TTTT**

Let's define our random variable X as the number of heads that occur during this experiment. As you can see from the sample space above, this random variable X can take on the values 0, 1, 2, 3 or 4 which I've color coded.

The table below shows each potential outcome of the random variable (1 Heads, 2 Heads, etc.) along with its probability of occurrence.

We can calculate the probability of occurrence by simply counting the number of outcomes and dividing by the total number of outcomes. The table below is an example the probability distribution of that random variable.

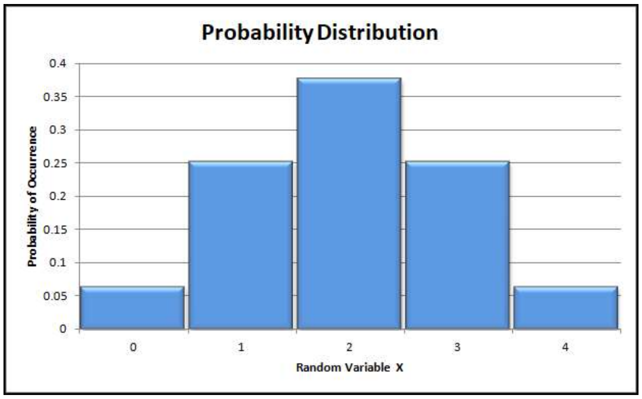
x	Outcomes	P(x)
(# of Heads)		Probability of x
0	**TTTT**	1 / 16
1	**TTTH, THTT, TTHT, HTTT**	4 / 16
2	**HTTH, HHTT, HTHT, THHT, THTH, TTHH**	6 / 16
3	**HHHT, HTHH, HHTH, HHHT**	4 / 16
4	**HHHH**	1 / 16

I'd also like to take this opportunity to cover some of the common notation you'll see below.

For example, throughout the chapter you'll see this common notation P(X = x). This is read as the probability that the random variable X is equal to a particular value, denoted by x.

As an example, $P(X = 1)$ refers to the probability that the random variable X is equal to 1. If we use the table above, the $P(X=1) = 4/16$ (25%).

This probability distribution table can be turned into a graph to help you visualize the probability of occurrence, which is equivalent to the area under of curve of the distribution.



Cumulative Probability Distributions

Before jumping into the actual distributions, there's another topic worth discussing - the Cumulative Probability Distribution.

Cumulative probability refers to the probability that the value of a random variable falls within a specified range or above a specific value or below a specific value.

Let's return to the coin flip experiment from above where we could ask a question like:

What is the probability that the coin flips would result in three or fewer heads?

The answer would be found in the concept of **cumulative probability**, where the probability of three or fewer heads can be translated the following equation: $\mathbf{P(X \leq 3)}$

Where $\boldsymbol{P(X \leq 3)} = P(X = 3) + P(X = 2) + P(X=1) + P(X=0).$

To put that into words - the probability that X is equal to or less than three is the cumulative probability of $X = 3$, plus the probability of $X = 2$, plus the probability of $X = 1$, plus the probability of $X = 0$.

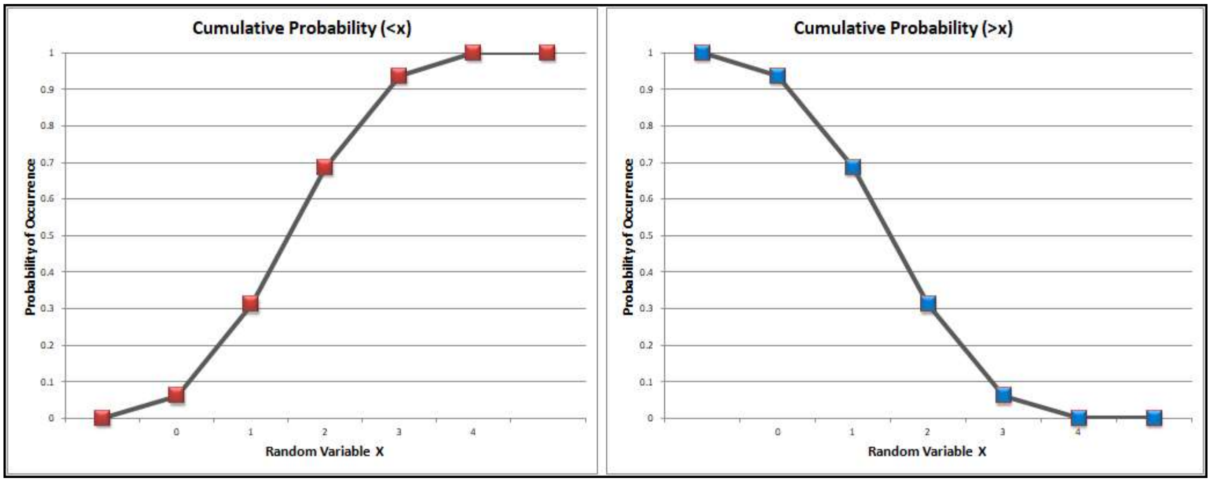
Look at the equation above - it honestly boggled my mind the first time I saw this equation because there are so many equal signs mixed in with plus signs. . . but all we're doing here is adding up the individual probabilities for 3, 2, 1 & 0.

Like the probability distribution above, a cumulative probability distribution can be represented by a graph, equation or table.

So I've modified the table from above to show the cumulative probability of $X > x$; along with the probability of $X \leq x$.

x (# of Heads)	P(X) Probability of X	P(X ≤ x) Cumulative Probability of X	P(X > x) Cum. Probability of X
0	1 / 16	1 / 16	15 / 16
1	4 / 16	5 / 16	11 / 16
2	6 / 16	11 / 16	5 / 16
3	4 / 16	15 / 16	1 / 16
4	1 / 16	16 / 16	0 / 16

Below is a graph of the Cumulative Probability Distribution for the probability distribution above. I've shown both the P(X ≤ x) & the P(X > x).



Part 1 - Continuous Distributions

Alright, it's time now to jump into the details of the most common distributions that can be used with **continuous data**.

Just to review quickly. . .

There are 9 Distributions that we're going to review:

The Normal, Uniform, Bivariate Normal, Exponential, Lognormal, Weibull, Chi Squared, Student T's, & F Distribution.

Within each section we will discuss the **common applications** for each distribution, and review the general **shape** for each.

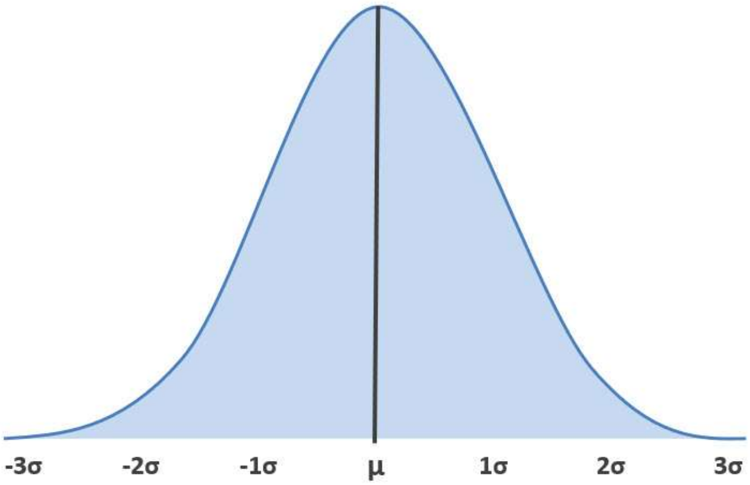
We will also be discussing how to calculate the **expected value** (Mean) and **variance** for each distribution - where it's applicable.

Lastly, and probably most importantly we will review the **probability calculations** for each distribution & show an **example** of those calculations.

Let's start with one of the most common distributions - the Normal Distribution.

Normal Distribution

The Normal Distribution is also commonly referred to as the Gaussian Curve or the Bell Curve due to its symmetric bell shape.



There are **two Parameters that fully define this distribution** - the **Mean** (µ) & **Standard Deviation** (σ).

The **Mean** value is a measure of the central tendency of the distribution & often exists at the peak & centerline of the distribution.

The **standard deviation** is a measure of the variation or spread associated with the distribution. The shape of the curve is governed mostly by the standard deviation.

The smaller the standard deviation the more data is centered around the mean. When the standard deviation gets bigger, the tails get longer and the data is more dispersed.

Skewness & Kurtosis of the Normal Distribution

When the normal distribution is not perfectly symmetric we use the word skewed; and we can measure **skewness**.

***Skewness** is a measure of the location of the mode (most frequently occurring data point) in relationship to the mean.* If the distribution is perfectly symmetrical, the skewness is zero.

Another characteristic of the normal distribution that's often discussed is **Kurtosis**.

***Kurtosis** provides a measure of the peakness or flatness of the distribution.*

The kurtosis of a standard normal distribution is 3.

If the distribution has a higher kurtosis, then the distribution has a higher and more narrow peak. If the kurtosis is low, the distribution is flatter and wider, with more data in the tails of the distribution.

The Z-Transformation of the Normal Distribution

Similar to other probability distributions, the area under the normal curve represents the probability of occurrence of X.

To more quickly calculate the area under the normal distribution curve statisticians have given us the Z-transformation, along with the Z-tables.

To perform the Z-transformation, you can use the following equation. This will transform your random variable X, into a Z-value based on the distributions mean & standard deviation.

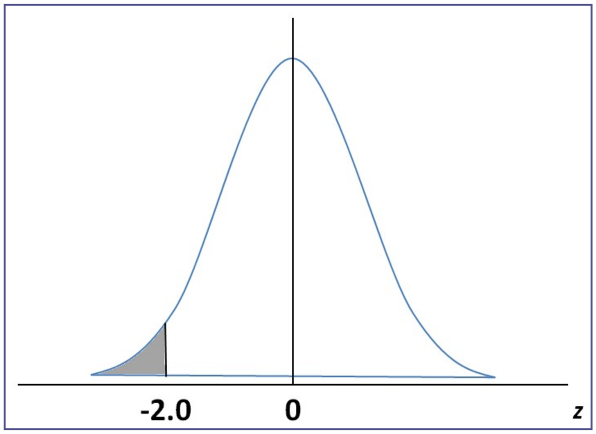
$$Z = \frac{X - \mu}{\sigma}$$

For example, let's say you've got a variable X (Grades on the CQE Exam) that follows the normal distribution with a mean value μ = 82 and a standard deviation σ = 6. The Z-score for an exam grade of 70 can be calculated as:

$$Z = \frac{70 - 82}{6} = -2.0$$

We can interpret this result by saying that the exam score of 70 is 2.0 standard deviations below the mean.

If you wanted to calculate the proportion of the population which scored less than 70% on the exam, it would look like the gray shaded area below on the distribution:

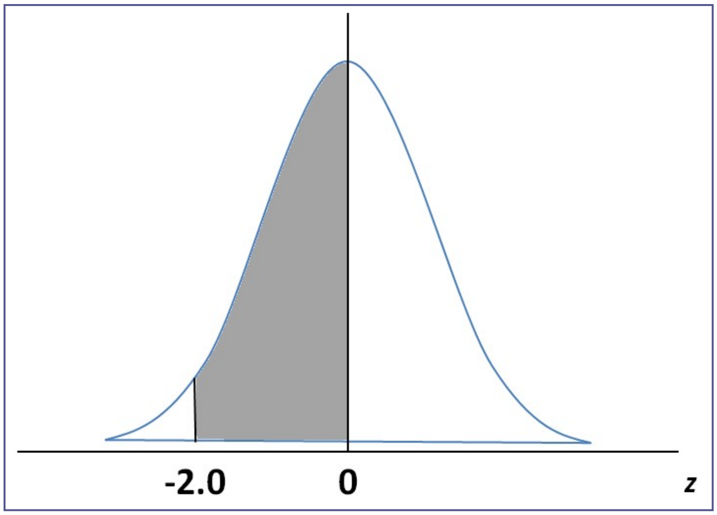


Notice this distribution is not a reflection of the exam score (centered at z=0), but it's a reflection of the transformed z-score associated with the exam.

We can then use the Z-score tables to answer any probability question associated with this value without having to use a calculator.

Z-Transformation Example

For example, the graph above shows all exam scores less than z = -2.0; however, you could also use the z-table to find the probability of a Z-score between -2 and 0, which graphically looks like this:



Once you've performed the Z-transformation, you can now calculate the probability associated with your Z-value using the table below.

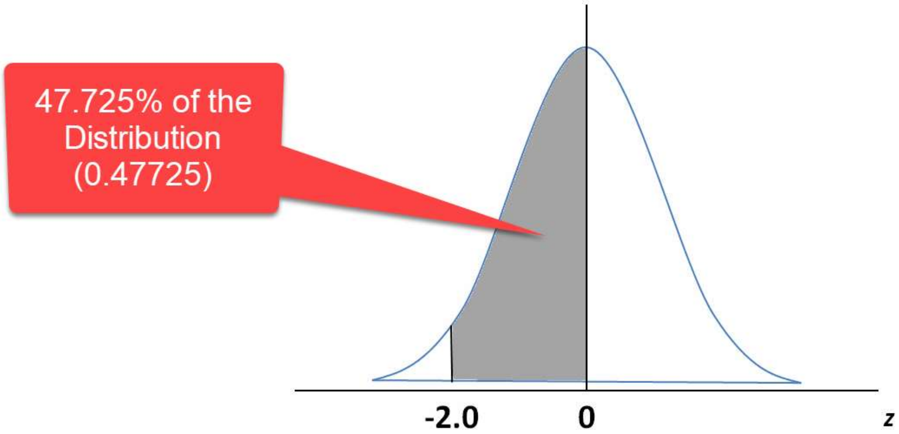
This Probability Table can be used to take your Z-value and convert it into the probability.

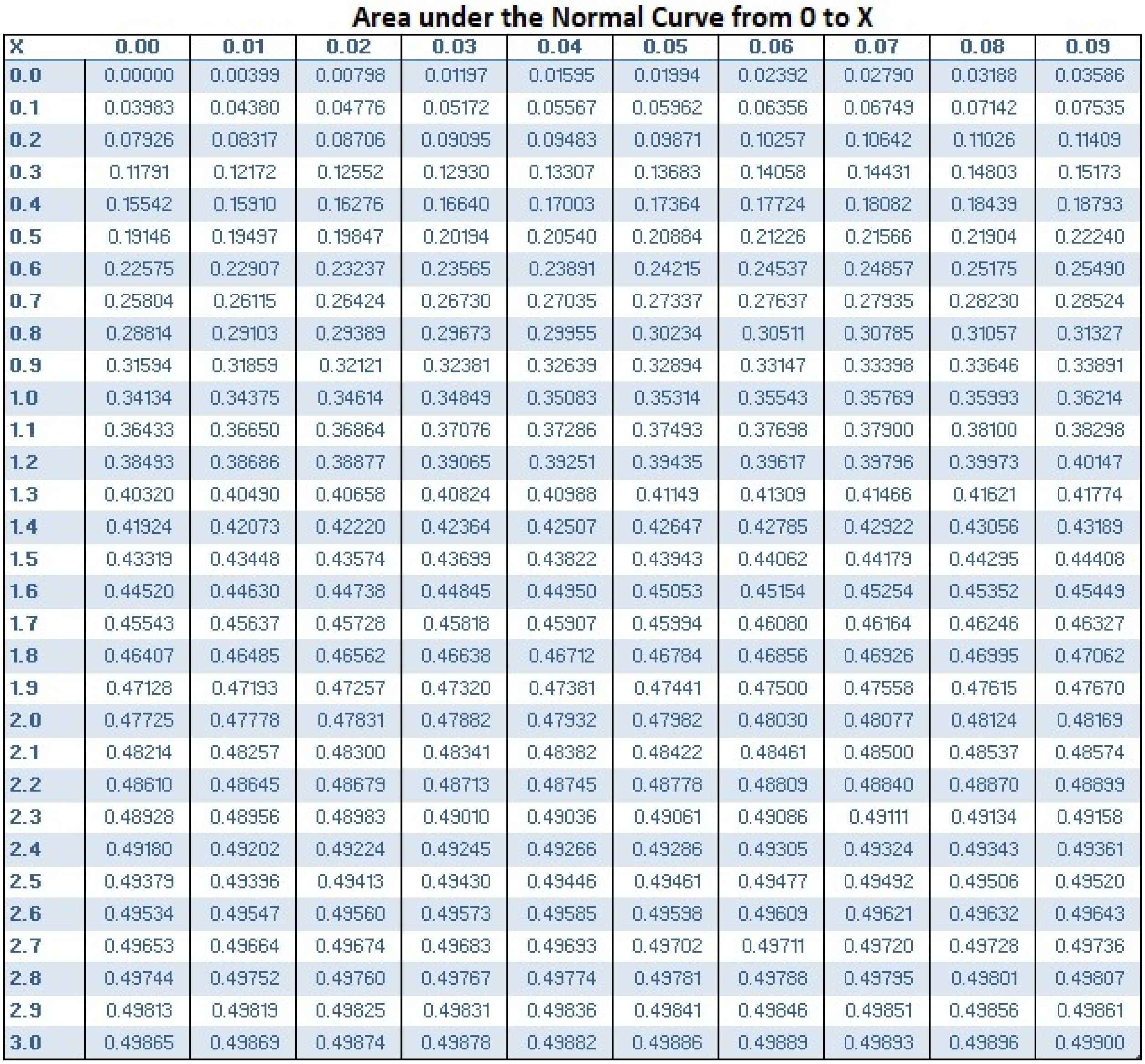
This table is potentially different from other Z-Probability tables in that it only provides the probability of positive Z values.

Recall though that the Z-value is symmetric around the mean value, so if you were looking for the probability from -1.04 to 0, it would be the same probability as that from +1.04 to 0.

Also, if you wanted to look up the probability between -1 to +1, then you'd double probability of Z=1.0 (0.34134).

The Z-Transformation of the Normal Distribution





Area under the Normal Curve from 0 to X

X	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900

Reliability Example Using the Normal Distribution

Let's do another example of the Z-transformation in a real-life situation.

Within the world of **Reliability**, the normal distribution curve can be used to model the reliability of a system over time.

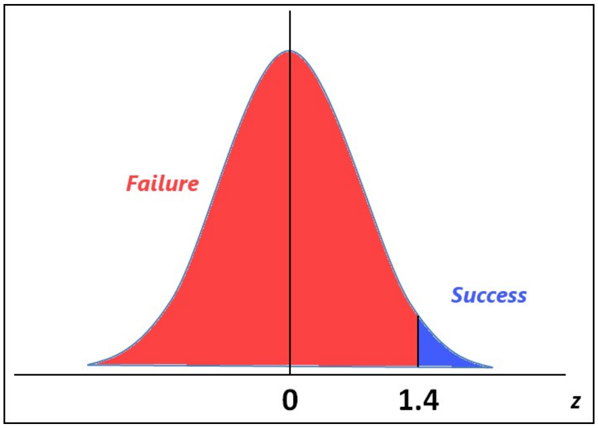
Let's say we're dealing with a motor and we've modeled the motors failure over time and it fits the normal distribution.

Your test data indicates that the mean and standard deviation associated with the motor is 6,500 hours and 500 hours respectively.

What is the **reliability** (*the probability that the motor is still operational*) of the motor at 7,200 hours?

$$\boldsymbol{Z = \frac{(X - \mu)}{\sigma} = \frac{(7,200 - 6500)}{500} = 1.4}$$

Graphically, this looks like:



Using the Z-Tables, *the area under the curve at Z = 1.4 is .4192*, and we add to that the 0.500 that represents the left half of the normal distribution curve which add up to 0.9192.

Remember that the Z-Score and the resulting probability represent the area to the left of the time value (7,200 hours).

So the **reliability** is the area to the **right of the curve**, which is 1 - .9192 = 0.0808.

Therefore, there is an 8% probability that the motor has not yet failed after 7,200 hours.

Or, said differently, 8% of the original population of motors are likely still operational after this amount of time.