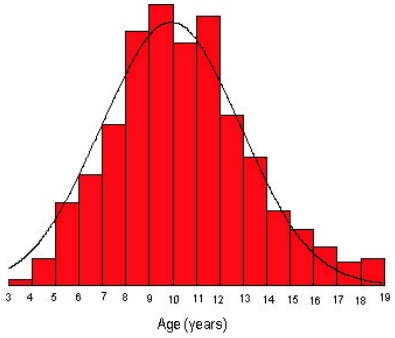
4B: Normal Probability Distributions

Normal density curve

The previous section used the binomial formula to calculate probabilities for binomial random variables. Outcomes were discrete, and probabilities were displayed with probability histograms. We need a different approach for modeling continuous random variables. This approach involves the use of **density curves**.

Think of **density curves** as smoothed probability histograms. A histogram and superimposed Normal density curve for an age distribution might look like this:



Although the fit of the density curve to the histogram in this instance is not perfect, it is a pretty good approximation. Normal curves are the very popular. The reasons for this will soon become evident.

The next curve shows the same distribution with the six leftmost bars shaded. This corresponds to individuals who are less than 9 years of age. There were 215 such individuals, making up 215 ÷ 654 = 0.329 or about 33% of the entries. The probability of randomly selecting someone younger than 9 from this group is 0.329. In notation, Pr($X < 9$) = 0.329.

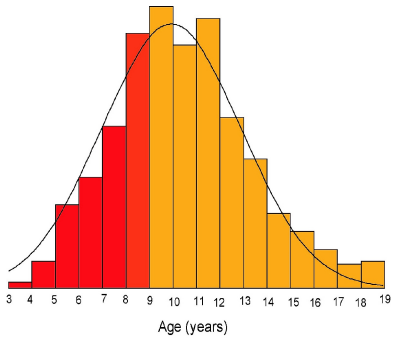


Fig. density-curve1.ai

When working with the continuous *pdf* models, we drop the underlying histogram and look only

at the curve.

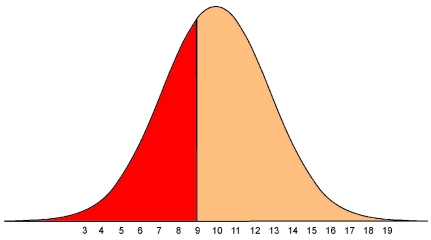


Fig. density-curve2.ai

The area under the curve corresponding to 9 years of age or less is shown in the darker color, making up 37% of the area under the curve. [The discrepancy between the area in the histograms bars (33%) and area in the density curve (37%) is due to the imperfect fit of the curve. The histogram has a slight positive skew, giving it a smaller than expected left tail.]

Normal distributions are characterized by two parameters:

Mean μ (Greek small letter mu)
Standard deviation σ (Greek small letter sigma).

Mean μ locates the center of the distribution. Changing μ shifts the curve along its X axis. Two Normal curves with different means are shown below:

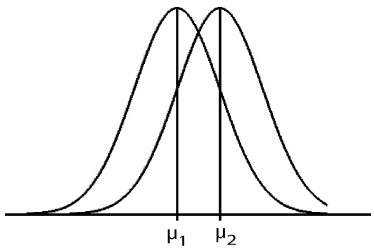


Fig: normal_comp_means.ai

The standard deviation determines the spread of a Normal distribution. The next figure depicts two Normal curves with the same means but different standard deviations.

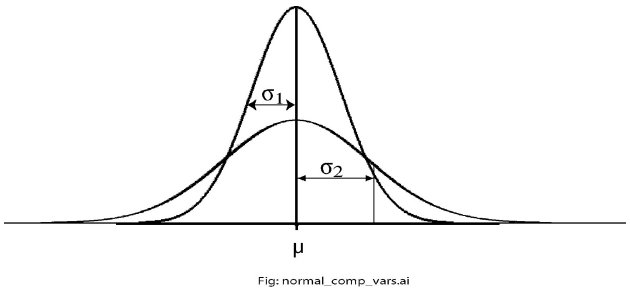


Fig: normal_comp_vars.ai

You can get a rough idea of the size of standard deviation σ by identifying "points of inflection"

on the Normal curve. Points of inflection are where the curve begins to turn and flatten a little. Here's what points of inflection look like:

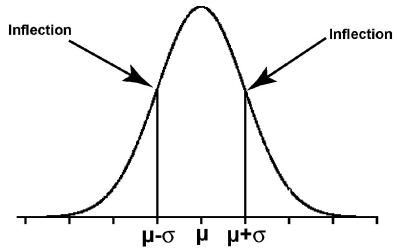


Fig. normal-inflection.ai

You can practice finding point of inflection by tracing a few Normal curves. This is where you feel the line begin to change curvature.

Identifying inflection points allows you to locate standard deviation markers above and below mean μ. These are important landmarks because:

1. Sixty-eight percent (68%) of the area under the Normal curve lies within one standard deviation σ of the mean. This is the region $\mu \pm \sigma$.
2. Ninety-five percent (95%) of the area under the Normal curve lies within one standard deviation σ of the mean. This is the region $\mu \pm 2\sigma$.
3. Ninety-nine-point-seven percent (99.7%) of the area under the Normal curve lies within one standard deviation σ of the mean. This is the region $\mu \pm 3\sigma$.

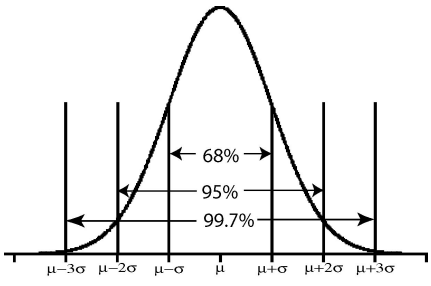


Fig: normal-68-95.ai

This characteristic of Normal curves is referred to as the **68–96–99.7 rule**.

Illustrative example: ***Wechsler IQ Scores.*** Wechsler IQ scores compare people of the same age using a score in which values are scalled to be Normal with a mean of 100 and standard deviation of 15. Let X represent Wechsler IQ scores. Using our notation, $X \sim N(100, 15)$. Based on the 68 – 95 – 99.7 rule, we know that 68% of the Wechler IQ scores lie in the range $100 \pm 15 = 85$ to 115, 95% lie in the range $100 \pm (2)(15) = 70$ to 130, and 99.7% lie in the range $100 \pm (3)(15) = 55$ to 145.

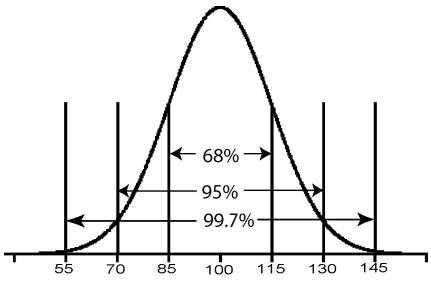


Fig: normal-IQ.ai

In the Wechsler IQ distribution (just above), 95% of scores are between 70 and 130. The other 5% of scores lie outside this range. Because the curve is symmetrical, half of the scores outside this range are below 70 and the other half are above 130. This means that the lowest 2.5% of scores are below 70 and the highest 2.5% of scores are above 130 (Figure).

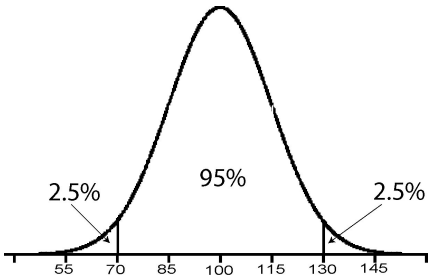


Fig: normal-IQ2sd.ai

Standardizing a value

Before determining Normal probabilities, you need to standardize your values. You do this by subtracting the mean and dividing by its standard deviation. This is called a **z-score**:

$$z = \frac{x - \mu}{\sigma} \qquad \textbf{(4.1)}$$

If the initial variable is Normal, making it into a *z* score will create a Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.This process is called **standardization**. Standardization merely re-scales the variable so that it has mean 0 and standard deviation 1. Data points that are larger than the mean will have positive *z* scores. Data points that are smaller than the mean have negative *z* scores. For example, a *z* score of +1 tells you that the individual is one standard deviation *above* the mean. A *z* score −2 tells you that the individual is two standard deviations *below* the mean.

Illustrative example: *Weschler IQ*. Weschler IQ scores vary Normally with mean $\mu = 100$ and standard deviation sigma = 15. An individual with an IQ of 115 has a $z = (115 - 100) / 15 = 1.00$ or 1.00 standard deviations above the mean IQ. An individual with a Weschler IQ of 95 has $z = (95 - 100) / 15 = -0.33$ or 0.33 standard deviations below the mean IQ.

Illustrative example: *Pregnancy length*. Pregnancy lengths measured from the last menstrual period to birth vary according to a Normal distribution with mean $\mu = 39$ weeks and standard deviation $\sigma = 2$ weeks. A woman whose pregnancy lasts 41 weeks has $z = (41 - 39) / 2 = 1$ or 1 standard deviations above the mean. A woman whose pregnancy is 36 weeks has $z = (36 - 39) / 2 = -1.5$ or 1.5 standard deviations below the mean.

Standardization moves the initial measurement to a "standard deviation scale," making any Normal random variable into a *Z* variable. This allows use of Z table to calculate Normal probabilities.

Calculating Normal probabilities

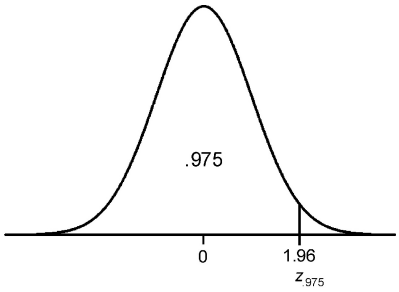
Probabilities for Normal random variables can be found by determining areas under Normal curves. Although there are an infinite number of different Normal curves (each with unique μ and σ), all Normal random variables will have a Standard Normal (Z) distribution once they are standardized. This allows use of a single table to look up probabilities. The table is called a "Standard Normal table" or "Z table."

Our *Z* table, which is in the back of the book, lists one and tens places in the left column with hundredths in the top row. Areas under the curve to the left of Normal *z* scores are listed in the body of the table. For example, the area under the curve to the left of $z = 1.96$ is found at the intersection of row 1.9 and column 0.06:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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

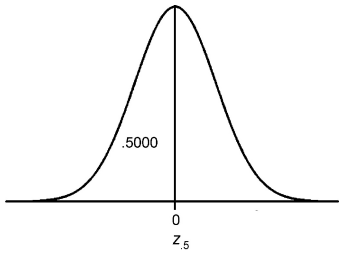
$Z_{.975} = 1.96$

Visually:



Let z_p denote a Normal *z*-score with a left tail area of *p*. The left tail of the distribution corresponds to the "cumulative probability" for that point. This is the probability of observing a value of that point or less. The figure above shows $z_{.975} = 1.96$. This is verbalized as"97.5[t]% of the area under the curve is to the left of 1.96" or "the cumulative probability of a Normal z score of 1.96 is 97.5%."

Here's an important landmark to keep in mind: $z_{.5} = 0$. We need no table for this landmark since we know the 50% of the area under the curve is to the left of 0. The cumulative probability of a Normal *z*-score of 0 is 50%.



Preliminary exercise: ***Notation for Normal z-score***. How would you verbalize $z_{.48} = -0.05$? **ANS:** This means "48 percent of the area under the curve on a standard Normal curve is to the left of –0.05." Alternatively, we could say "a Normal *z* score of –0.05 is larger than 48% of the individuals in the population."

Illustrative example: ***Wechsler IQ***. Recall that Wechsler IQ scores vary according to a Normal distribution with mean = 100 and standard deviation = 15. What proportion of IQ scores are less than 129.4? This is that same as asking what is the probability of selecting someone at random from the population who has an IQ of 129.4 or less. This probability is the area under $N(100, 15)$ to the left of the point 129.4. The standardized IQ score of 129.4 has a *z*-score of $z = (129.4 - 100) / 15 = 1.96$. The probability of seeing this score is the area under the Standard Normal curve to the left of 1.96. As noted, the area under the curve to the left of this point is .9750.

We can find the probability of any range of values for a Normal random variable by standardizing the values and determining the area under the curve that corresponds to the range of interest using Table B. Here is a step-by-step approach to the process:

1. State the probability that is being requested in terms of original variable *X*.
2. Turn the value of *X* into a *z*-score.
3. Draw a Normal curve that is to scale and then shade the area under the curve corresponding to the probability you want to know.
4. Find the area under the curve with the help of Table B.

Illustrative example: ***Pregnancy length*.** The lengths of human pregnancies from conception to birth varies according to a Normal distribution with mean = 39 weeks and standard deviation = 2 weeks. What proportion of pregnancies lasts less than 41 weeks?

1. **State the problem**: Let *X* represent gestational length: X~N(39, 2). The probability of seeing a value less than 41 is Pr(*X* <= 41).
2. **Standardize the value**: z = (41 – 39) / 2 = 1. This value is one standard deviation

above average.

3. **Draw the curve**: The figure below shows this the landmarks on the *X*-axis and the area under the curve.
4. **Use the Z table**: Pr(*X* <= 41) = Pr(*Z* <= 1) = .8413. This is about .84 or 84%. About 84% of pregnancies last 41 weeks or less.

Probabilities for observations above a certain value (right tail). Our *Z* table provide probabilities that are less than or equal to a given value ("cumulative probabilities"). These cumulative probabilities correspond to the area under the curve to the *left* of a point.

When you need to know the probability of values greater than or equal to a particular point (corresponding to the right tail of a distribuion, use the fact:

(Area under the curve in the right tail) = 1 – (Area under the curve in the left tail)

For example, in the "Pregnancy length" illustration, the probability of a gestation that is greater than or equal to 41 weeks is 1 – Pr(*X* <= 41) = 1 – .8413 = .1587, or about 16%. Approximately 16% of pregnancies last more than 41 weeks .

Probabilities for observations between certain values. You can calculate probabilities for Z between any two value *a* and *b* by subtracting left tail areas obtained from Table B according to the formula Pr($a \le Z \le b$) = Pr($Z \le b$) – Pr($Z \le a$).

Illustrative example: *Pregnancy length*. By a standard definition, gestations of less than 35 weeks are considered premature and those more than 40 weeks are considered post-date (Durham, 2002, p. 3, adjusted for gestation from LMP statements). What proportion of pregnancy lengths fell between these values?

1. **State the problem:** Let *X* represent gestational length: X~N(39, 2). The probability of seeing value between 35 and 40 weeks is denoted Pr(35 <= *X* <= 40).
2. **Standardize the values:** For the lower bound of 35 weeks, $z = (35 - 39) / 2 = -2$. For the upper bound of 40 weeks, $z = (40 - 39) / 2 = 0.5$.
3. **Draw the curve**: Figure 8.13 shows landmarks for this problem.
4. **Use the Z table**: Figure 8.13 also shows that the area under the curve in this range is .6687, or nearly 67%. About two-thirds of the pregnancies are in this range.