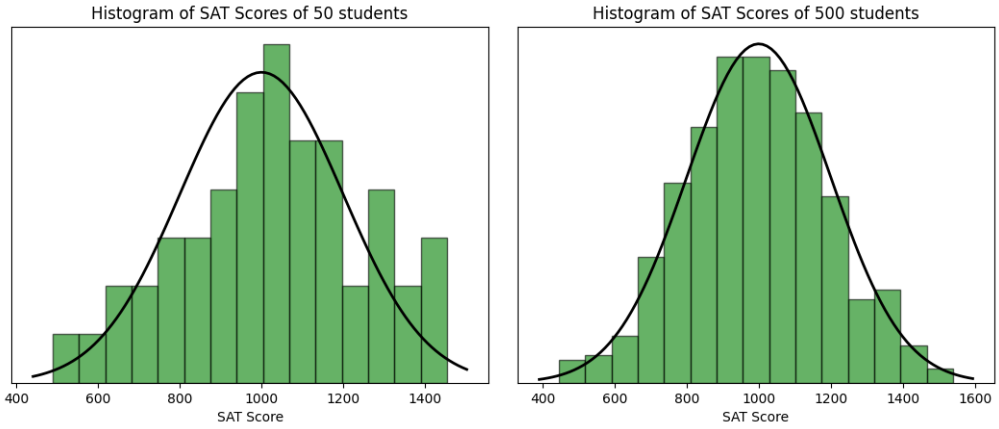
Chapter 8: The Normal Distribution

1 Introduction

The Normal model is the most frequently used probability model for continuous numerical variables. This popularity stems from the fact that many numerical variables of interest to researchers tend to have distributions that closely match the Normal model. Additionally, a significant mathematical theorem known as the Central Limit Theorem (which will be discussed in next chapter) connects the Normal model to key statistical concepts, making it a valuable model to understand.

1.1 Visualizing the Normal Distribution

We always start with unimodal and symmetric dataset to begin with. The following histogram displays the SAT scores for 50 students and 500 students. We have a computer-drawn curve to fit the histogram data. The overlaying of smooth curve on the histograms is to illustrate the unimodal and symmetric shape to you. It's easy to imagine that as more data is collected, the histogram would fill in the curve more precisely, eventually matching the shape almost perfectly. It is also evident that the histogram of SAT score with 500 students fits more closely to the curve than the histogram with 50 students.



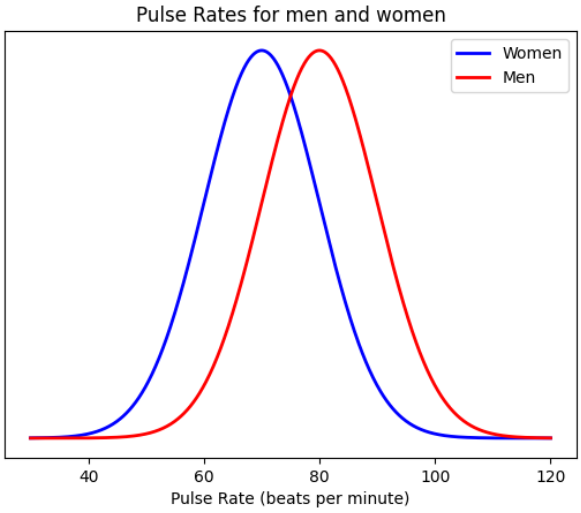
If you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed.

Examples: Height and weight of people, reaction of time of people, length of natural sleep, IQ scores, SAT scores, pulse rate are some examples of normally distributed datasets.

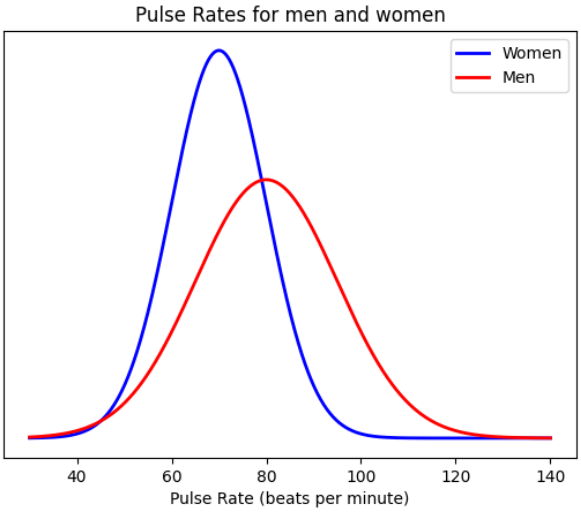
2 Normal Distribution

Colloquially, Normal distribution is also called as normal curve or bell curve. The main characteristic of a normal distribution is that it describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation.

Let's visualize these with the help of examples. We have the dataset for resting pulse rate for men and women in beats per minute. The blue curve is the normally distributed pulse rate for women and it's center is at 70. Hence, the mean value of resting pulse rate in women is 70 beats per minute. Similarly, the red curve is the normally distributed pulse rate for men and it's center is at 80. Hence, the mean value of resting pulse rate in women is 80 beats per minute.



Similarly in the following figure, the standard deviation for women is 10 beats per minute. Hence, the blue curve is narrow in width. However, the standard deviation for women is 25 beats per minute. Thus, we have a wider curve.



3 Standard Normal Distribution and $z-$score

The Normal distribution with mean at 0 and standard deviation of 1 is called a standard normal distribution.

The question, "How unusual is this?" is perhaps the statistician's favorite question. (It is just as popular as "Compared to what?") However answering this question is complicated because the answer depends on the units of measurement. For example, eighty-four is a big value if we are measuring a person's height in inches, but it is a small value if we are measuring their weight in pounds. Unless we know the units of measurement and the objects being measured, we can't judge whether a value is big or small. One way around this problem is to change the units to standard units. Standard units measure a value relative to the sample rather than with respect to some absolute measure. A measurement converted to standard units is called a $z-$score.

The standard score or z-score is equal to first the observation minus the mean, then divided by the standard deviation.

This tells you how many standard deviations from the mean your observation is.

$$z = \frac{x - \bar{x}}{s}$$

where x is the datapoint, $\bar{x}$ is the sample mean and s is the sample standard deviation.

Example: If an observation has a $z-$score of 1.46, then that observation is 1.46 standard deviations above the mean.

Example: If an observation has a $z-$score of -1.46, then that observation is 1.46 standard deviations below the mean.

$z-$ score example: Madison scored 600 on the SAT Math (mean = 500, standard deviation = 100). Her friend Gabriel scored 21 on ACT Math (mean = 18, standard deviation = 6). Assuming scores of both exams are normally distributed, who did better with respect to their test-taking population?

Madison's standard score is:

$$z = \frac{(x - \bar{x})}{s} = \frac{(600 - 500)}{100} = \frac{100}{100} = 1$$

Gabriel's standard score is:

$$z = \frac{(x - \bar{x})}{s} = \frac{(21 - 18)}{6} = \frac{3}{6} = 0.5$$

Because Madison's score is 1 standard deviation above the mean and Gabriel's is only 0.5 standard deviations above the mean, Madison's performance is better.

4 Percentiles of the normal distribution

In working with normal distributions, We must be able to use the normal distribution to compute probabilities, which are areas under a normal curve and above given intervals. Let's see

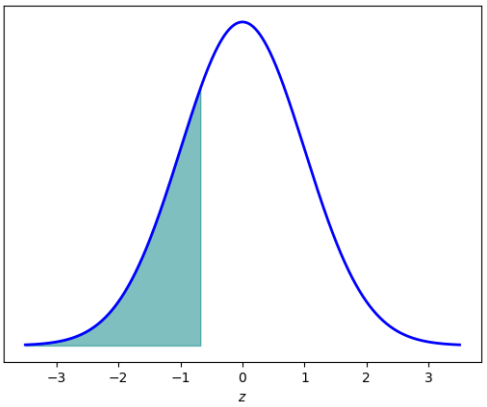
some examples.

Example 1: Assume that the SAT math exam scores are normally distributed. If Ting obtains a score of 432 on the SAT math exam, what percentage of individuals score the same as or lower than Ting?

Firstly, let's the SAT Score to standard score (z-Score):

$$z = \frac{(x - \bar{x})}{s} = \frac{(432 - 500)}{100} = \frac{-68}{100} = -0.68$$

We want the shaded area as in the figure below.



To do so, we look up the percentile corresponding to $z = -0.68$ on $z-$table.

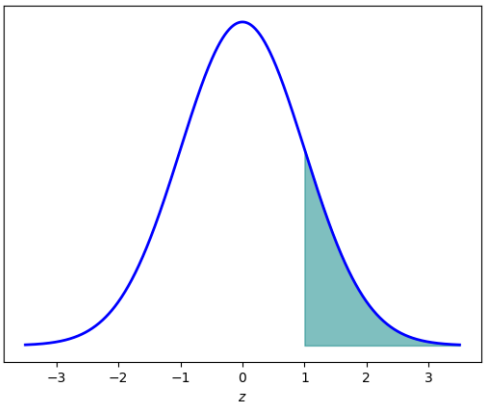
So,$P(z < -0.68) = 0.2483$
Hence, 24.83% of students scored the same as or lower than Ting in SAT math section.

Example 2: In a certain city, the average daily temperature in July follows a normal distribution with a mean of 85°F and a standard deviation of 5°F. What is the probability that on a randomly chosen day in July, the temperature will be higher than 90°F?

Firstly, let's the temperature to standard score (z-Score):

$$z = \frac{(x - \bar{x})}{s} = \frac{(90 - 85)}{5} = \frac{5}{5} = 1$$

We want the shaded area as in the figure below.



To do so, we look up the percentile corresponding to $z = 1$ on $z-$table. However remember, **In a standard $z-$ table, the area corresponding to a $z-$score is always the area to the left.**

So, $P(z > 1) = 1 - P(z < 1) = 1 - 0.8413 = 0.1587$

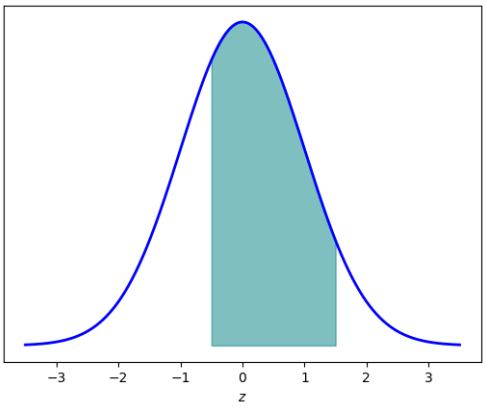
Hence, the probability that on a randomly chosen day in July, the temperature will be higher than 90°F is 15.87%

Example 3: Suppose that resting pulse rates for healthy adults are found to follow a Normal distribution, with a mean of 70 beats per minute and a standard deviation of 10 beats per minute. What percentage of the pulse rates fall between 65 and 85 beats per minute?

Solution: Firstly, we calculate the z-scores for each beats.

$$z = \frac{x - \bar{x}}{s} = \frac{65 - 70}{10} = \frac{-5}{10} = -0.5$$
$$z = \frac{x - \bar{x}}{s} = \frac{85 - 70}{10} = \frac{15}{10} = 1.5$$

We want the shaded area as in the figure below.



To do so, we calculate the area under each $z-$score and subtract the highest from lowest.

$$P(-0.5 \leq z \leq 1.5) = P(z \leq 1.5) - P(z \leq -0.5) = 0.9332 - 0.3085 = 0.6247$$

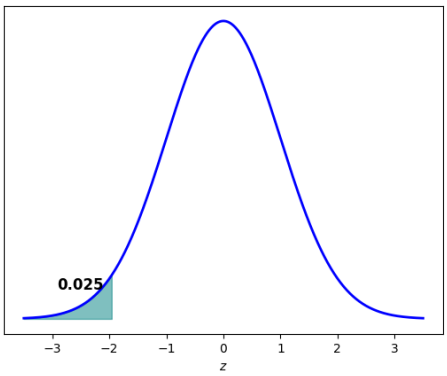
Therefore, around 62.47% of pulse rates are between 55 and 85 beats per minute.

5 Some special examples:

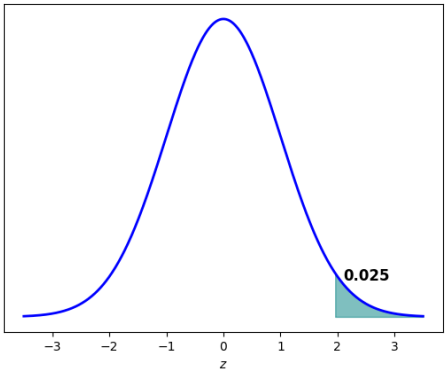
The following are some of the examples, which can be used as a further practice to calculate the area under a normal curve. Additionally, they will be of great use in the following two chapters.

Example 1: Calculate the area to the left of $z = -1.96$ and to the right of $z = 1.96$ separately. Then calculate the area in between those two $z-$scores.

We see from the normal table that the area to the left of $z = -1.96$ is 0.025 or 2.5%

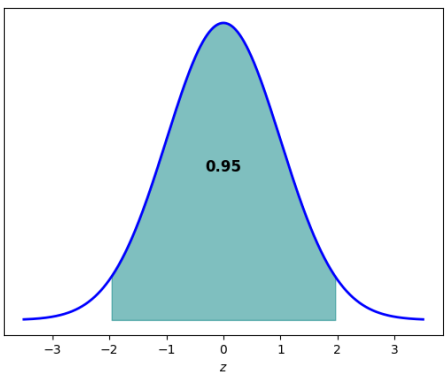


We see from the normal table that the area to the left of $z = 1.96$ is 0.9750. So the are to the right of $z = 1.96$ is also 0.025 or 2.5%.



This could have also been inferred from the symmetry of normal distribution.

Now we can calculate that the area between $z = -1.96$ and $z = 1.96$ is 95%

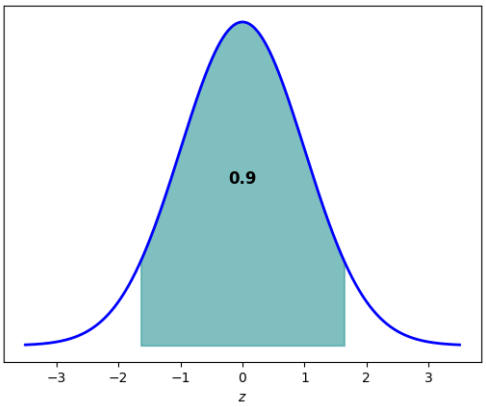


Example 2: Calculate the area in between $z = -1.64$ and $z = 1.64$.

We see from the normal table that the area to the left of $z = -1.64$ is 0.0505.

We see from the normal table that the area to the left of $z = 1.64$ is 0.9495.

Now we can calculate that the area between $z = -1.96$ and $z = 1.96$ as $0.9495 - 0.0505 = 0.899$.



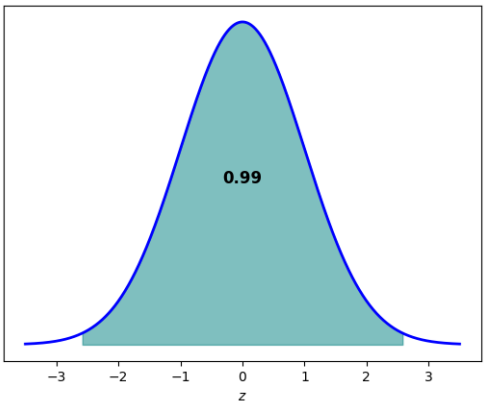
This result will come in handy in the following two chapters and we will round it up to 0.9 in those two chapters.

Example 3: Calculate the area in between $z = -2.58$ and $z = 2.58$.

We see from the normal table that the area to the left of $z = -2.58$ is 0.0049.

We see from the normal table that the area to the left of $z = 2.58$ is 0.9951.

Now we can calculate that the area between $z = -2.58$ and $z = 2.58$ as $0.9951 - 0.0049 = 0.9902$.



This result will come in handy in the following two chapters and we will round it up to down to 0.99 in those two chapters.

