6 The Normal Distribution

The normal distribution is very important in Probability and Statistics, as it is typically the probability distribution that is used when modeling measurements of a quantity in physics, engineering or social sciences.

6.1 The Probability Density Function of the normal distribution

The PDF of a normally distributed random variable is one of the most important special funcitons in Ma

Definition 1. We say that the random variable X is normally distributed with mean of μ and standard deviation σ if it has the PDF

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

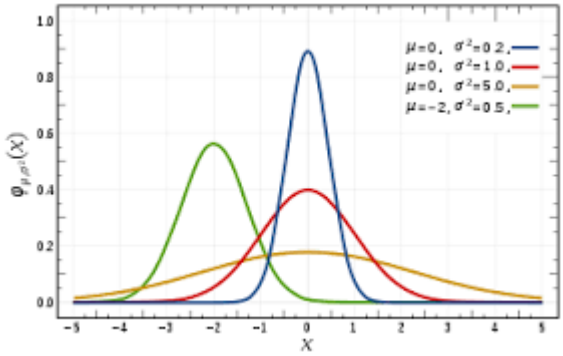
We denote $X \sim \mathcal{N}(\mu, \sigma)$.

We observe that the parameters of the normal PDF are the mean μ and standard deviation σ of the random variable, i.e.,

$$EV[X] = \mu$$

$$VAR[X] = \sigma^2$$

The figure illustrates normal PDFs with different μ and σ.



Remark 1. The constant $\frac{1}{\sqrt{2\pi}\sigma}$ in the expression of the PDF is chosen such that $f(x)$ is a valid PDF, i.e.,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1.$$

Indeed, it can be proved that $\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma$. We will skip the proof as it is outside the scope of the class.

Example 1. Battery life is normally distributed with mean $\mu = 500$ days and standard deviation $\sigma = 61$ days. The probability that the battery lasts between 439 and 561 days can be calculated using the PDF

$$P(439 \leq X \leq 561) = \frac{1}{\sqrt{2\pi}61^2} \int_{439}^{561} e^{-\frac{(x-500)^2}{2\cdot 61^2}} dx = 0.6826$$

The integrals involved are calculated using numerical methods. There are several software packages that can be used to calculate these integrals. Also, most graphing calculators have them implemented.

6.2 The standard normal distribution

The normal distribution with mean of 0 and standard deviation of 1 is called the standard normal distribution. Its PDF is given by

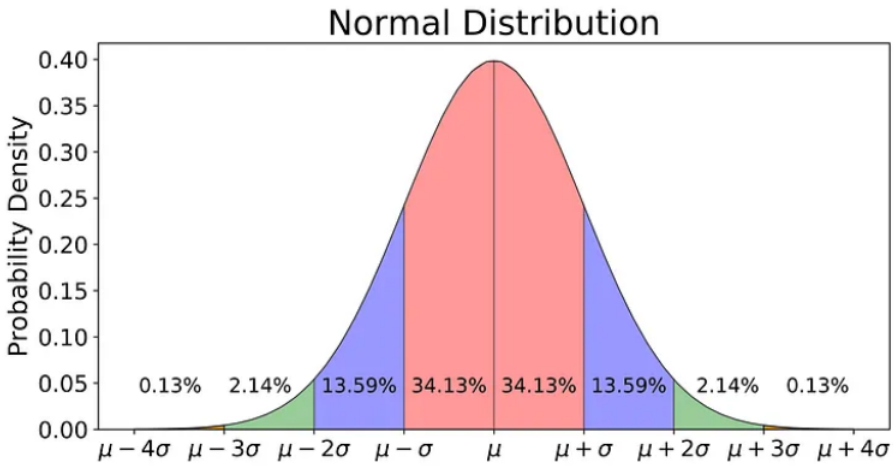
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

The transformation

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

transforms a normal distribution with mean μ and standard deviation σ to the standard normal distribution. This transformation is used a lot in Machine Learning as many ML models require normalization of the data.

The value z associated to a raw score x is called the $z-score$. The z-score measures how many standard deviations is the datapoint from the mean. So $z = 1$ means that the corresponding score is one standard deviations greater than the mean, while z-score $z = -2$ means two standard deviations less than the mean. The figure illustrates the connection between the z-score and probabilities.



Example 2. Consider the battery life example above with $\mu = 500$ and $\sigma = 61$.

a) What are the z-scores corresponding to $x = 439$ and $x = 561$ respectively.

b) Use the z-scores to find $P(439 < X < 561)$.

c) What battery life has a z-score of 3 and what is the probability that a battery will last beyond this life span?

Sol: a) We can see that $x = 439$ corresponds to the z-score of

$$z = \frac{439 - 500}{61} = -1$$

while $x = 561$ corresponds to

$$z = \frac{561 - 500}{61} = 1$$

b) Using the figure above we can also conclude that

$$P(439 < X < 561) = P(-1 < z < 1) = 0.6826.$$

c) If $z = 3$ then we can solve

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

as an equation for x. Indeed,

$$3 = \frac{x - 500}{61}$$

gives $x = 683$. We can see that $P(X > 683) = 0.13\%$,i.e., only 0.13% of batteries have their life span longer than 683 days.

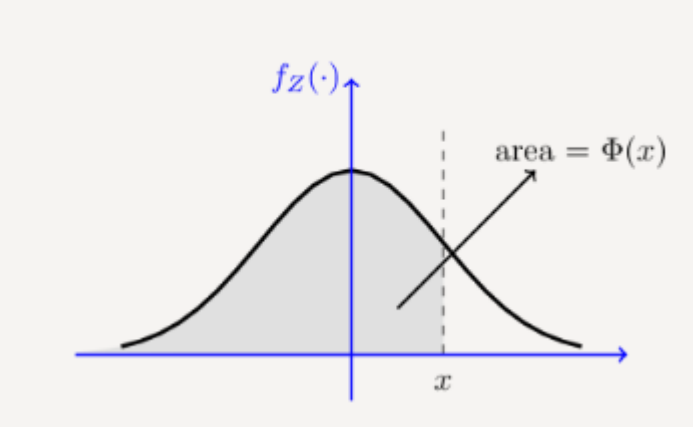
Remark 2. Z-scores have been important historically as the z-scores and corresponding probabilities have been published in tables and used for statistical calculations, before computers became ubiquitous.

6.3 The Cummulative Density Function of the Normal Distribution.

In most applications we will be using the CDF of the normal distribution. It is defined as

$$\Phi(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

and it is the area to the left of x under the graph of the PDF as illustrated in the figure. We will use the notation $\Phi(x)$ for the CDF of the normal distribution as shown in the figure.



Example 3. SAT scores of students in 2012 was approximately normally distributed with mean $\mu = 496$ and standard deviation of $\sigma = 114$.

a) What is the probability that a randomly selected student scored below 400 points.

b) What is the probability that someone scored more than 600 points.

c) What is the probability that someone scored between 400 and 600 points?

Sol: We use the CDF $\Phi(x)$ with $\mu = 496$,$\sigma = 114$. a) $P(X < 400) = \Phi(400) = 0.1998$

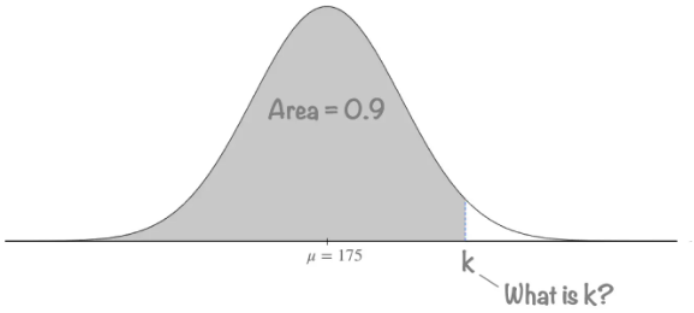
b) $P(X > 600) = 1 - \Phi(600) = 0.1808$
c) $P(400 \leq X \leq 600) = \Phi(600) - \Phi(400) = 0.8192 - 0.1998 = 0.6194$

Another important function that we will use in this chapter is the inverse of the normal CDF. It is easy to see that the CDF $\Phi(x)$ is increasing and continuous, therefore it is invertible. Its inverse function is denoted by $\Phi^{-1}(p)$. By the definition of the inverse function we can solve the equation

$$\Phi(x) = p$$

and its solution is

$$x = \Phi^{-1}(p).$$



Example 4. Using the SAT scores example above find the score of students who scored more than 90% of their peers.

Sol: This question asks for the 90^{th} percentile. I.e., the solution of the equation $\Phi(x) = 0.9$. The solution can be calculated using the inverse CDF as $x = \Phi^{-1}(0.9) = 642$.

6.4 Normal approximation of the binomial distribution

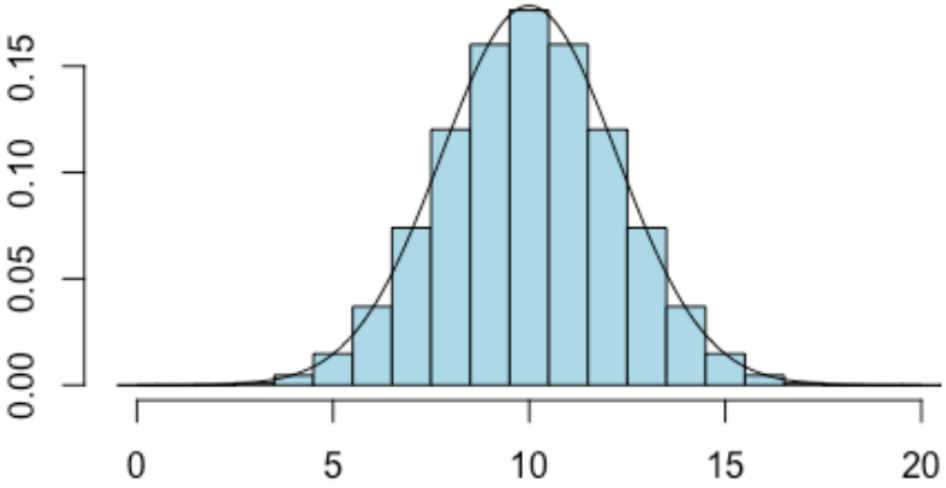
Let us consider the following example. We toss a biased coin with probability of heads 0.6, 100 times. What is the probability to get less than 50 heads? The problem can be modeled using the binomial distribution $X \sim B(100, 0.6)$. We have

$$P(X = k) = \begin{pmatrix} 100 \\ k \end{pmatrix} 0.6^k \cdot 0.4^{100-k}.$$

To get less than 40 heads we can calculate

$$P(X < 50) = \sum_{k=0}^{49} \left(\begin{array}{c} 100 \\ k \end{array} \right) 0.6^k \cdot 0.4^{100-k}.$$

An alternative approximate solution method for the problem uses the normal approximation of the binomial distribution. The figure illustrates this approximation.



For large values of n we can approximate the binomial random variable $X \sim B(n, p)$ by a normally distributed random variable $N(\mu, \sigma)$ with $\mu = np$ and $\sigma = \sqrt{npq}$ where $q = 1 - p$. We can write

$$B(n, p) \approx N(np, \sqrt{npq})$$

Example 5. We toss a biased coin with probability of heads 0.6, 100 times.

a) What is the probability to get less than 50 heads?

b) 90% of the time if we repeat the experiment of tossing the coin 100 times, we will get less than how many heads?

Sol. The binomial distribution $X \sim B(100, 0.6)$ can be approximated by

$$B(100, 0.6) \approx N(100 \cdot 0.6, \sqrt{100 \cdot 0.6 \cdot 0.4}) = N(60, 4.9).$$

a) $P(X < 50) = \Phi(50) = 0.02$.

b) The question asks for the 90th percentile of the number of heads, i.e., $\Phi^{-1}(0.9) = 66.28$.

If we reconsider part c of the previous problem we observe that we might as well use the following reasoning: $P(X < 50) = P(X \leq 49) = \Phi(49)$. The issue is due to the discrepancy between a discrete and a continuous random variable. To resolve the problem we use the so called continuity correction to the normal approximation of the binomial. In this case we consider $P(X < 50) = \Phi(49.5)$. More generally, when using the normal approximation of the binomial with continuity correction we will have

$$P(X < k) = \Phi\left(k - \frac{1}{2}\right)$$

and

$$P(X \leq k) = \Phi\left(k + \frac{1}{2}\right).$$

6.5 Limit Theorems

Laplace observed that by repeating an experiment multiple times, and averaging the outcomes, the average is usually normally distributed.

6.5.1 The Central Limit Theorem for the sample mean

We roll a fair dye 5 time and we get 2, 2, 3, 4, 6, and the average is $\overline{X} = 3.4$. If we repeat the experiment 10 times, we will get an average that is closer and closer to 3.5. The more we repeat the experiment the closer we get to 3.5. If we repeat the experiment and this time we record the standard deviations, we will get a decreasing sequence with respect to the number of experiments. We address the following question: Given a random variable X, what is the mean and standard deviation of $\overline{X}$, i.e., the average of n samples of X.

Theorem 1. *Central Limit Theorem (CLT). Let X be a random variable with known or unknown distribution. Let μ_X and σ_X be the mean and standard deviation of X. We consider n samples of X. When n is large, then the sample mean $\overline{X}$ is normally distributed with mean*

$$\mu_{\overline{X}} = \mu_X$$

and standard deviation

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}.$$

It is interesting to note that X does not need to be normally distributed, however the average is always normally distributed. Also, let us observe that the statement n is large can be interpreted mathematically as $n \to \infty$. For practical purposes, we can consider the above normal distribution as an approximation of the distribution of the sample mean.

Example 6. An unknown distribution has a mean of 90 and standard deviation of 15. $n = 25$ samples are drawn and the average calculated.

a) Find the distribution of the average/mean $\overline{X}$.

b) What is the probability that the mean is between 85 and 92?

c) Find the value that is two standard deviations above the sample mean.

d) What is the 90th percetile of values of $\overline{X}$?

Sol: a) $\overline{X} \sim N(90, \frac{15}{\sqrt{25}}) = N(90, 3)$

b) $P(85 < \overline{X} < 92) = \Phi(92) - \Phi(85) = 0.6997$

c) We are looking for the value that has a z-score of 2,

$$2 = \frac{x - 90}{3},$$

with $x = 96$.

d) To find the 90th percentile we solve the equation

$$\Phi(k) = 0.9,$$

with solution $k = \Phi^{-1}(0.9) = 93.85$.

6.5.2 Central Limit Theorem for sums

We consider a random variable X with mean μ_X and standard deviation σ_X and we take the sum of n samples of the distribution, denoted ΣX. We can formulate the following variant of the CLT for sums.

Theorem 2. *Let X be a random variable with known or unknown distribution. Let μ_X and σ_X be the mean and standard deviation of X. We consider the sum of n samples of X. When n is a large number, the sum ΣX is normally distributed with mean*

$$\mu_{\Sigma X} = n \cdot \mu_X$$

and standard deviation

$$\sigma_{\Sigma X} = \sqrt{n} \cdot \sigma_X.$$

Example 7. Let X be a uniformly distributed random variable within the interval $[0, 10]$. We take 20 samples and calculate their sum.

a) Calculate the mean and standard deviation of the sum ΣX.

b) Find $P(\Sigma X > 120)$.

c) What sum values are within one standard deviation from the mean and what does this mean in terms of probabilities?

Sol: a) For the uniform random variable X we have $\mu_X = 5$, $\sigma_X = \sqrt{\frac{10^2}{12}} = \frac{5}{\sqrt{3}} \approx 2.887$. For the sum, using the CLT we have $\mu_{\Sigma X} = 100$,$\sigma_{\Sigma X} = \sqrt{20} \cdot \sqrt{2.887} = 10.91$. As a connclusion we have $\Sigma X \sim N(100, 10.91)$.

b) $P(\Sigma X > 120) = 1 - \Phi(120) = 0.0334$.

c) z-scores of -1 and 1 correspond to $100-10.91 = 89.09$ and $100+10.91 = 110.91$, respectively. As a probability interpretation we can say

$$P(|\Sigma X - 100| < 10.91) = 0.682$$

6.6 Exercises

1. The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ = 125 and standard deviation σ = 14. Systolic blood pressure for males follows a normal distribution.

   (a) Calculate the z-scores for the male systolic blood pressures 100 and 150 millimeters.

   (b) If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, what would be his blood pressure?

2. Kyle's doctor told him that the z-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean μ = 125 and standard deviation σ = 14. If X = a systolic blood pressure score then X ~ N (125, 14).

   (a) Which answer(s) is/are correct?

      i. Kyle's systolic blood pressure is 175.

ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.

iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.

iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.

(b) Calculate Kyle's blood pressure.

3. IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let X = IQ of an individual.

   (a) Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.

   (b) MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.

   (c) The middle 50% of IQs fall between what two values?

4. Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

   (a) If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.

   (b) Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

5. Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K through 5. A simple random sample of 300 is surveyed.

   (a) Find the probability that at least 150 favor a charter school.

   (b) Find the probability that at most 160 favor a charter school.

(c) Find the probability that more than 155 favor a charter school.

(d) Find the probability that fewer than 147 favor a charter school.

6. Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races.

   (a) Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.

   (b) Find the 80th percentile for the average of these 49 marathons.

7. The length of songs in a collector's iTunes album collection is uniformly distributed from two to 3.5 minutes. Suppose we randomly pick five albums from the collection. There are a total of 43 songs on the five albums.

   (a) What distribution does the average length of songs follow.

   (b) What is the probability that a song is longer than 6 min?