18
Probability and Statistics

18.1 Probability and Event

18.1.1 Sample Space and Event

Probability is related to the uncertainty and randomness of the world. It measures the likelihood of some outcome or event happening. To formalize this concept, we introduce the following definition:

Definition 18.1.1 (Sample Space and Event)**.** *A set S of all possible outcomes of a random phenomenon in the world is called a **sample space**. A subset $A \subset S$ is called an **event**.*

Example 18.1.2. *The sample space of "the outcome of tossing two dice" is the set $S = \{(1,1),(1,2),\ldots,(6,6)\}$ of 36 elements. The event "the sum of the numbers on the two dice is 5" is the subset $A = \{(1,4),(2,3),(3,2),(4,1)\}$ of 4 elements.*

18.1.2 Probability

Given a sample space S, we define a probability for each event A of that space. This probability measures the likelihood that the outcome of the random phenomenon belongs A.

Definition 18.1.3 (Probability)**.** *A **probability** $\Pr : \mathcal{S} \to \mathbb{R}_{\geq 0}$ is a mapping from each **event** $A \subset S$ to a non-negative real number $\Pr[A] \geq 0$ such that the following properties are satisfied:*

1. $0 \leq \Pr[A] \leq 1$ *for any* $A \subset S$
2. $\Pr[S] = 1$
3. *For any countable collection $\{A_1, A_2, \ldots\}$ of events that are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$ for any $i \neq j$),*

$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i]$$

When the sample space is finite or countably infinite, [1] *the properties above can be simplified into the following condition:*

$$\sum_{x \in S} \Pr[\{x\}] = 1$$

[1] A countably infinite set refers to a set whose elements can be numbered with indices. The set $\mathbb{N}$ of natural numbers or the set $\mathbb{Q}$ of rational numbers are examples of countably infinite sets.

Example 18.1.4. *Consider the sample space of "the outcome of tossing two dice" again. Assuming the two dice are fair, the probability of each outcome can be defined as* $1/36$*. Then the probability of the event "the sum of the numbers on the two dice is* 5*" is* $4/36$*.*

Example 18.1.5. *We are picking a point uniformly at random from the sample space* $[0, 2] \times [0, 2]$ *in the Cartesian coordinate system. The probability of the event that the point is drawn from the bottom left quarter* $[0, 1] \times [0, 1]$ *is* $1/4$*.*

18.1.3 Joint and Conditional Probability

In many cases, we are interested in not just one event, but multiple events, possibly happening in a sequence.

Definition 18.1.6 (Joint Probability)**.** *For any set of events* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *of a sample space* S*, the* ***joint probability*** *of* $\mathcal{A}$ *is the probability* $\Pr[A_1 \cap \ldots \cap A_n]$ *of the intersection of all of the events. The probability* $\Pr[A_i]$ *of each of the events is also known as the* ***marginal probability****.*

Example 18.1.7. *Consider the sample space of "the outcome of tossing two dice" again. Let* A_1 *be the event "the number on the first die is* 1*" and let* A_2 *be the event "the number on the second die is* 4*." The joint probability of* A_1, A_2 *is* $1/36$*. The marginal probability of each of the events is* $1/6$*.*

It is also useful to define the probability of an event A, based on the knowledge that other events $A_1, \ldots, A_n$ have occurred.

Definition 18.1.8 (Conditional Probability)**.** *For any event* A *and any set of events* $\mathcal{A} = \{A_1, \ldots, A_n\}$ *of a sample space* S*, where* $\Pr[A_1 \cap \ldots \cap A_n] > 0$*, the* ***conditional probability of*** A ***given*** $\mathcal{A}$ *is*

$$\Pr[A \mid A_1, \ldots, A_n] = \frac{\Pr[A \cap A_1 \cap \ldots \cap A_n]}{\Pr[A_1 \cap \ldots \cap A_n]}$$

Example 18.1.9. *Consider the sample space of "the outcome of tossing two dice" again. Let* A_1 *be the event "the number on the first die is* 1*" and let* A_2 *be the event "the sum of the numbers on the two dice is* 5*." The conditional probability of* A_1 *given* A_2 *is* $1/4$*. The conditional probability of* A_2 *given* A_1 *is* $1/6$*.*

Using the definition of a conditional probability, we can define a formula to find the joint probability of a set $\mathcal{A}$ of events of a sample space.

Proposition 18.1.10 (Chain Rule for Conditional Probability)**.** *Given a set $\mathcal{A} = \{A_1, \ldots, A_n\}$ of events of a sample space S, where all appropriate conditional probabilities are defined, we have the following*

$$\begin{aligned}\Pr[A_1 \cap \ldots \cap A_n] &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \cap \ldots \cap A_n] \\ &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \mid A_3 \cap \ldots \cap A_n] \cdot \Pr[A_3 \cap \ldots \cap A_n] \\ &\vdots \\ &= \Pr[A_1 \mid A_2 \cap \ldots \cap A_n] \cdot \Pr[A_2 \mid A_3 \cap \ldots \cap A_n] \cdots \Pr[A_n]\end{aligned}$$

Finally, from the definition of a conditional probability, we see that

$$\Pr[B \mid A] \Pr[A] = \Pr[A \cap B] = \Pr[A \mid B] \Pr[B]$$

This shows that

$$\Pr[B \mid A] = \frac{\Pr[A \mid B] \Pr[B]}{\Pr[A]}$$

This is known as the *Bayes's Rule*.

18.1.4 Independent Events

Definition 18.1.11 (Independent Events)**.** *Two events A, B are **independent** if $\Pr[A], \Pr[B] > 0$ and*

$$\Pr[A] = \Pr[A \mid B]$$

or equivalently

$$\Pr[B] = \Pr[B \mid A]$$

or equivalently

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

Example 18.1.12. *Consider the sample space of "the outcome of tossing two dice" again. Let A_1 be the event "the number on the first die is 1" and let A_2 be the event "the number on the second die is 4." A_1 and A_2 are independent.*

Example 18.1.13. *Consider the sample space of "the outcome of tossing two dice" again. Let A_1 be the event "the number on the first die is 1" and let A_2 be the event "the sum of the numbers on the two dice is 5." A_1 and A_2 are not independent.*

18.2 Random Variable

In the previous section, we only learned how to assign a probability to an event, a subset of the sample space. But in general, we can assign a probability to a broader concept called a *random variable*, associated to the sample space.

Definition 18.2.1 (Random Variable)**.** *Given a sample space S, a mapping $X : S \to \mathbb{R}$ that maps each outcome $x \in S$ to a value $i \in \mathbb{R}$ is called a* ***random variable****.*

Example 18.2.2. *Consider the sample space of "the outcome of tossing two dice" again. Then the random variable $X =$ "sum of the numbers on the two dice" maps the outcome $(1, 4)$ to the value 5.*

Definition 18.2.3 (Sum and Product of Random Variables)**.** *If $X, X_1, \ldots, X_n$ are random variables defined on the same sample space S such that $X(x) = X_1(x) + \ldots + X_n(x)$ for every outcome $x \in S$, then we say that X is the* ***sum*** *of the random variables $X_1, \ldots, X_n$ and denote*

$$X = X_1 + \ldots + X_n$$

If $X(x) = X_1(x) \times \ldots \times X_n(x)$ for every outcome $x \in S$, then we say that X is the ***product*** *of the random variables $X_1, \ldots, X_n$ and denote*

$$X = X_1 \cdots X_n$$

Example 18.2.4. *Consider the sample space of "the outcome of tossing two dice" again. Then the random variable $X =$ "sum of the numbers on the two dice" is the sum of the two random variables $X_1 =$ "the number on the first die" and $X_2 =$ "the number on the second die."*

18.2.1 Probability of Random Variable

There is a natural relationship between the definition of an event and a random variable. Given a sample space S and random variable $X : S \to \mathbb{R}$, the "event that X takes a value in B" is denoted $\Pr[X \in B]$. It is the total probability of all outcomes $x \in S$ such that $X(x) \in B$. In particular, the event that X takes a particular value $i \in \mathbb{R}$ is denoted as $X = i$ and the event that X takes a value in the interval $[a, b]$ is denoted as $a \leq X \leq b$ and so on.

Example 18.2.5. *Consider the sample space of "the outcome of tossing two dice" and the random variable $X =$ "sum of the numbers on the two dice" again. Then*

$$\Pr[X = 5] = \Pr[\{(1, 4), (2, 3), (3, 2), (4, 1)\}] = 4/36$$

Often we are interested in the probability of the events of the form $X \leq x$. Plotting the values of $\Pr[X \leq x]$ with respect to x completely identifies the *distribution* of the values of X.

Definition 18.2.6 (Cumulative Distribution Function)**.** *Given a random variable X, there is an associated* ***cumulative distribution function (cdf)*** *$F_X : \mathbb{R} \to [0, 1]$ defined as*

$$F_X(x) = \Pr[X \leq x]$$

Proposition 18.2.7. *The following properties hold for a cumulative distribution function F_X:*

1. *F_X is increasing*
2. *$\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$*

18.2.2 Discrete Random Variable

If the set of possible values of a random variable X is finite or countably infinite, we call it a *discrete random variable*. For a discrete random variable, the probability $\Pr[X = i]$ for each value i that the random variable can take completely identifies the *distribution* of X. In view of this fact, we denote the *probability mass function (pmf)* by

$$p_X(i) = \Pr[X = i]$$

Proposition 18.2.8. *The following properties hold for a probability mass function p_X:*

1. *$\sum_i p_X(i) = 1$*
2. *$F_X(x) = \sum_{i\leq x} p_X(i)$*

18.2.3 Continuous Random Variable

We now consider the case where the set of all possible values of a random variable X is an interval or a disjoint union of intervals in $\mathbb{R}$. We call such X a *continuous random variable*. In this case, the probability of the event $X = i$ is zero for any $i \in \mathbb{R}$. Instead, we care about the probability of the events of the form $a \leq X \leq b$.

Definition 18.2.9 (Probability Density Function)**.** *Given a continuous random variable X, there is an associated* ***probability density function (pdf)*** *$f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that*

$$\Pr[a \leq X \leq b] = \int_a^b f(x)dx$$

for any $a, b \in \mathbb{R}$.

Proposition 18.2.10. *The following properties hold for a probability density function f_X:*

1. *$\int_{-\infty}^{\infty} f_X(x)dx = 1$*
2. *$F_X(x) = \int_{-\infty}^{x} f_X(y)dy$*

18.2.4 Expectation and Variance

Definition 18.2.11 (Expectation)**.** *The **expectation** or the **expected value** of a discrete random variable X is defined as*

$$\mathbb{E}[X] = \sum_i i \cdot p_X(i) = \sum_i i \cdot \Pr[X = i]$$

where p_X is its associated probability mass function. Similarly, the expectation for a continuous random variable X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

where f_X is the associated probability density function. In either case, it is customary to denote the expected value of X as μ_X or just μ if there is no source of confusion.

Example 18.2.12. *Consider the sample space of "the outcome of tossing one die." Then the expected value of the random variable $X =$ "the number on the first die" can be computed as*

$$\mathbb{E}[X] = 1 \cdot \frac{6}{36} + 2 \cdot \frac{6}{36} + 3 \cdot \frac{6}{36} + 4 \cdot \frac{6}{36} + 5 \cdot \frac{6}{36} + 6 \cdot \frac{6}{36} = 3.5$$

Proposition 18.2.13 (Linearity of Expectation)**.** *If X is the sum of the random variables $X_1, \ldots, X_n$, then the following holds:*

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]$$

Also, if $a, b \in \mathbb{R}$ and X is a random variable, then

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$$

Example 18.2.14. *Consider the sample space of "the outcome of tossing two dice." Then the expected value of the random variable $X =$ "the sum of the numbers of the two dice" can be computed as*

$$\mathbb{E}[X] = 3.5 + 3.5 = 7$$

since the expected value of the number on each die is 3.5*.*

Definition 18.2.15 (Variance)**.** *The **variance** of a random variable X, whose expected value is μ, is defined as*

$$Var[X] = \mathbb{E}[(X - \mu)^2]$$

*Its **standard deviation** is defined as*

$$\sigma_X = \sqrt{Var[X]}$$

It is customary to denote the variance of X as σ_X^2.

Proposition 18.2.16. *If $a \in \mathbb{R}$ and X is a random variable, then*

$$Var[aX] = a^2 Var[X] \quad \sigma_{aX} = |a|\,\sigma_X$$

Problem 18.2.17. *Prove **Chebyshev's inequality**:*

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

for any $k > 0$. (Hint: Suppose the probability was greater than $1/k^2$. What could you conclude about $\mathbb{E}[(X-\mu)^2]$?)

18.2.5 Joint and Conditional Distribution of Random Variables

Just as in events, we are interested in multiple random variables defined on the sample space.

Definition 18.2.18 (Joint Distribution)**.** *If X, Y are discrete random variables defined on the same sample space S, the **joint probability mass function** $p_{X,Y}$ is defined as*

$$p_{X,Y}(i,j) = \Pr[X = i, Y = j]$$

where the event $X = i, Y = j$ refers to the intersection $(X = i) \cap (Y = j)$.

*If X, Y are continuous random variables defined on S, there is an associated **joint probability density function** $f_{X,Y} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that*

$$\Pr[a \leq X \leq b, c \leq Y \leq d] = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

*The joint probability mass/density function defines the **joint distribution** of the two random variables.*

Definition 18.2.19 (Marginal Distribution)**.** *Given a joint distribution $p_{X,Y}$ or $f_{X,Y}$ of two random variables X, Y, the **marginal distribution** of X can be found as*

$$p_X(i) = \sum_j p_{X,Y}(i,j)$$

if X, Y are discrete and

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

if continuous. We can equivalently define the marginal distribution of Y.

Definition 18.2.20 (Conditional Distribution)**.** *Given a joint distribution $p_{X,Y}$ or $f_{X,Y}$ of two random variables X, Y, we define the **conditional distribution of X given Y** as*

$$p_{X \mid Y}(i \mid j) = \frac{p_{X,Y}(i,j)}{p_Y(j)}$$

if X, Y are discrete and

$$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

if continuous. We can equivalently define the marginal distribution of Y given X.

18.2.6 Bayes' Rule for Random Variables

Sometimes it is easy to calculate the conditional distribution of X given Y, but not the other way around. In this case, we can apply the *Bayes' Rule* to compute the conditional distribution of Y given X. Here, we assume that X, Y are discrete random variables. By a simple application of Bayes' Rule, we have

$$\Pr[Y = j \mid X = i] = \frac{\Pr[X = i \mid Y = j] \Pr[Y = j]}{\Pr[X = i]}$$

Now by the definition of a marginal distribution, we have

$$\Pr[X = i] = \sum_{j'} \Pr[X = i, Y = j] = \sum_{j'} \Pr[X = i \mid Y = j'] \Pr[Y = j']$$

for all possible values j' that Y can take. If we plug this into the denominator above,

$$\Pr[Y = j \mid X = i] = \frac{\Pr[X = i \mid Y = j] \Pr[Y = j]}{\sum_{j'} \Pr[X = i \mid Y = j'] \Pr[Y = j']}$$

Example 18.2.21. *There is a coin, where the probability of **Heads** is unknown and is denoted as θ. You are told that there is a 50% chance that it is a fair coin (i.e., $\theta = 0.5$) and 50% chance that it is biased to be $\theta = 0.7$. To find out if the coin is biased, you decide to flip the coin. Let D be the result of a coin flip. Then it is easy to calculate the conditional distribution of D given θ. For example,*

$$\Pr[D = H \mid \theta = 0.5] = 0.5$$

But we are more interested in the probability that the coin is fair/biased based on the observation of the coin flip. Therefore, we can apply the Bayes' Rule.

$$\Pr[\theta = 0.7 \mid D = H] = \frac{\Pr[D = H \mid \theta = 0.7] \Pr[\theta = 0.7]}{\Pr[D = H]}$$

which can be calculated as

$$\frac{\Pr[D = H \mid \theta = 0.7] \Pr[\theta = 0.7]}{\Pr[D = H \mid \theta = 0.7] \Pr[\theta = 0.7] + \Pr[D = H \mid \theta = 0.5] \Pr[\theta = 0.5]}$$
$$= \frac{0.7 \cdot 0.5}{0.7 \cdot 0.5 + 0.5 \cdot 0.5} \simeq 0.58$$

*So if we observe one **Heads**, there is a* 58% *chance that the coin was biased and a* 42% *chance that it was fair.*

Problem 18.2.22. *Consider Example 18.2.21 again. This time, we decide to throw the coin* 10 *times in a row. Let N be the number of observed **Heads**. What is the probability that the coin is biased if $N = 7$?*

18.2.7 Independent Random Variables

Analogous to events, we can define the independence of two random variables.

Definition 18.2.23 (Independent Random Variables)**.** *Two discrete random variables X, Y are **independent** if for every i, j, we have*

$$p_X(i) = p_{X \mid Y}(i \mid j)$$

or equivalently,

$$p_Y(j) = p_{Y \mid X}(j \mid i)$$

or equivalently

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$$

*Two continuous random variables X, Y are **independent** if the analogous conditions hold for the probability density functions.*

Definition 18.2.24 (Mutually Independent Random Variables)**.** *If any pair of n random variables $X_1, X_2, \ldots, X_n$ are independent of each other, then the random variables are **mutually independent**.*

Proposition 18.2.25. *If $X_1, \ldots, X_n$ are mutually independent random variables, the following properties are satisfied:*

1. $\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n]$
2. $Var[X_1 + \ldots + X_n] = Var(X_1) + \ldots + Var(X_n)$

We are particularly interested in independent random variables that have the same probability distribution. This is because if we repeat the same random process multiple times and define a random variable for each iteration, the random variables will be *independent and identically distributed*.

Definition 18.2.26. *If $X_1, \ldots, X_n$ are mutually independent random variables that have the same probability distribution, we call them **independent, identically distributed** random variables, which is more commonly denoted as **iid** or **i.i.d.** random variables.*

18.3 Central Limit Theorem and Confidence Intervals

Now we turn our attention to two very important topics in statistics: *Central Limit Theorem* and *confidence intervals*.

You may have seen *confidence intervals* or *margin of error* in the context of election polls. The pollster usually attaches a caveat to the prediction, saying that there is some probability that the true opinion of the public is $\pm\epsilon$ of the pollster's estimate, where ϵ is typically a few percent. This section is about the most basic form of confidence intervals, calculated using the famous Gaussian distribution. It also explains why the Gaussian pops up unexpectedly in so many settings.

A running example in this chapter is estimating the bias of a coin we have been given. Specifically, $\Pr[\textbf{Heads}] = p$ where p is unknown and may not be $1/2$. We wish to estimate p by repeatedly tossing the coin. If we toss the coin n times, we expect to see around np Heads. Confidence intervals ask the converse question: after having seen the number of heads in n tosses, how "confidently" can we estimate p?

18.3.1 Coin Tossing

Suppose we toss the same coin n times. For each $i = 1, 2, \ldots, n$, define the random variable X_i as an *indicator random variable* such that

$$X_i = \begin{cases} 1 & i\text{-th toss was } \textbf{Heads} \\ 0 & \textit{otherwise} \end{cases}$$

It is easily checked that $X_1, \ldots, X_n$ are iid random variables, each with $\mathbb{E}[X_i] = p$ and $Var[X_i] = p(1-p)$. Also if we have another random variable $X =$ "number of heads," notice that X is the sum of $X_1, \ldots, X_n$. Therefore, $\mathbb{E}[X] = np$ and $Var[X] = np(1-p)$.

Problem 18.3.1. *Show that if* $\Pr[\textbf{Heads}] = p$ *then* $\mathbb{E}[X] = np$ *and* $Var[X] = np(1-p)$. *(Hint: use linearity of expectation and the fact that* X_i*'s are mutually independent.)*

Suppose $p = 0.8$. What is the distribution of X? Figure 18.1 gives the distribution of X for different n's.

Let's make some observations about Figure 18.1.

Expected value may not happen too often. For $n = 10$, the expected number of Heads is 8, but that is seen only with probability 0.3. In other words, with probability 0.7, the number of Heads is different from the expectation. [2]

The highly likely values fall in a smaller and smaller band around the expected value, as n increases.

For $n = 10$, there is a good chance that the number of Heads is

[2] In such cases, *expectation* can be a misleading term. It may in fact be *never* seen. For instance, the expected number of eyes in an individual drawn from the human population is somewhere between 1 and 2 but no individual has a non-integral number of eyes. Thus *mean value* is a more intuitive term.

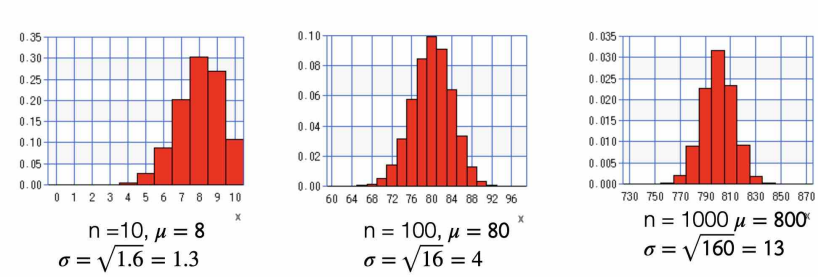

Figure 18.1: Distribution of X when we toss a coin n times, and $p = 0.8$. The plots were generated using a calculator.

quite far from the the expectation. For $n = 100$, the number of Heads lies in $[68, 90]$ with quite high probability. For $n = 1000$ it lies in $[770, 830]$ with high probability.

The probability curve becomes more symmetrical around the mean. Contrast between the case where $n = 10$ and the case where $n = 100$.

Probability curve starts resembling the famous Gaussian distribution . Also called *Normal Distribution* and in popular math, the *Bell curve*, due to its bell-like shape.

18.3.2 Gaussian Distribution

We say that a real-valued random variable X is distributed according to $\mathcal{N}(\mu, \sigma^2)$, the Gaussian distribution with mean μ and variance σ^2, if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{18.1}$$

It is hard to make an intuitive sense of this expression. The following figure gives us a better handle.

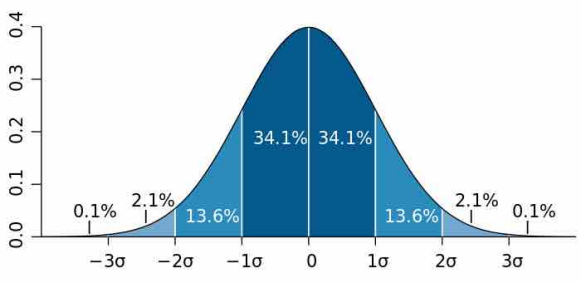

Figure 18.2: Cheatsheet for the Gaussian distribution with mean μ and variance σ^2. It is tightly concentrated in the interval $[\mu - k\sigma, \mu + k\sigma]$ for even $k = 1$ and certainly for $k = 2, 3$. *Source:* https://en.wikipedia.org/wiki/Normal_distribution

Figure 18.2 shows that X concentrates very strongly around the mean μ. The probability that X lies in various intervals around μ of the type $[\mu - k\sigma, \mu + k\sigma]$ are as follows: (i) For $k = 1$ it is 68.2%; (ii) For $k = 2$ it is 95.4%; (iii) For $k = 3$ it is 99.6%.

18.3.3 Central Limit Theorem (CLT)

This fundamental result explains our observations in Subsection 18.3.1.

Theorem 18.3.2 (Central Limit Theorem, informal statement)**.** *Suppose $X_1, X_2, \ldots,$ is a sequence of random variables that are mutually independent and each of whose variance is upper bounded by some constant C. Then as $n \to \infty$, the sum $X_1 + X_2 + \ldots + X_n$ tends to $\mathcal{N}(\mu, \sigma^2)$ where $\mu = \sum_i \mathbb{E}[X_i]$ and $\sigma^2 = \sum_i Var(X_i)$.*

We won't prove this theorem. We will use it primarily via the "cheatsheet" of Figure 18.2.

18.3.4 Confidence Intervals

We return to the problem of estimating the bias of a coin, namely $p = \Pr[\textbf{Heads}]$. Suppose we toss it n times and observe X heads. Then $X = \sum_i X_i$ where X_i is the indicator random variable that signifies if i-th toss is Heads.

Since the X_i's are mutually independent, we can apply the CLT and conclude that X will approximately follow a Gaussian distribution as n grows. This is clear from Figure 18.1, where the probability histogram (which is a discrete approximation to the probability density) looks quite Gaussian-like for $n = 1000$. In this course we will assume for simplicity that CLT applies exactly. Using the mean and variance calculations from Problem 18.3.1, X is distributed like $\mathcal{N}(\mu, \sigma^2)$ where $\mu = np$, $\sigma^2 = np(1-p)$. Using the cheatsheet of Figure 18.2, we can conclude that

$$\Pr[X \notin [np - 2\sigma, np + 2\sigma]] \leq 4.6\%$$

Since $X \in [np - 2\sigma, np + 2\sigma]$ if and only if $np \in [X - 2\sigma, X + 2\sigma]$, some students have the following misconception:

> *Given the observation of X heads in n coin tosses, the probability that $np \notin [X - 2\sigma, X + 2\sigma]$ is at most 4.6%.*

But there is no *a priori distribution* on p. It is simply some (unknown) *constant* of nature that we're trying to estimate. So the correct inference should be:

> *If $np \notin [X - 2\sigma, X + 2\sigma]$, then the probability (over the n coin tosses) that we would have seen X heads is at most 4.6%.*

The above is an example of confidence bounds. Of course, you may note that σ also depends on p, so the above conclusion doesn't give us a clean confidence interval. In this course we use a simplifying assumption: to do the calculation we estimate σ^2 as $np'(1-p')$ where $p' = X/n$. (The intuitive justification is that we expect p to be close to X/n.)

Example 18.3.3. *Suppose $X = 0.8n$. Using our simplified calculation, $\sigma^2 \approx n(0.8)(0.2)$, implying $\sigma = 0.4\sqrt{n}$. Thus we conclude that if $p \notin [0.8 - 0.4/\sqrt{n}, 0.8 + 0.4/\sqrt{n}]$, then the probability of observing this many Heads in n tosses would have been less than $100 - 68.2\%$, that is, less than 31.8%.*

The concept of confidence intervals is also relevant to ML models.

Example 18.3.4. *A deep neural network model was trained to predict cancer patients' chances of staying in remission a year after chemotherapy, and we are interested in finding out its accuracy p. When the model is tested on $n = 1000$ held-out data points, this problem is equivalent to the coin flipping problem. For each of the held-out data point, the probability that the model makes the correct prediction is p. By observing the number of correct predictions on the held-out data, we can construct a confidence interval for p. Say the test accuracy was $p' = 70\%$. Then the 68% confidence interval can be written as*

$$np \in [np' - \sigma, np' + \sigma]$$

Substituting $p' = 0.7, \sigma \approx \sqrt{np'(1-p')}, n = 1000$, we get

$$1000p \in [685.5, 714.5]$$

or equivalently,

$$p \in [0.6855, 0.7145]$$

18.3.5 Confidence Intervals for Vectors

In the above settings, sampling was being used to estimate a real number, namely, $\Pr[\textbf{Heads}]$ for a coin. How about estimating a vector? For instance, in an opinion poll, respondents are being asked for opinions on multiple questions. Similarly, in *stochastic gradient descent* (Chapter 3), the gradient vector is being estimated by sampling a small number of data points. How can we develop confidence bounds for estimating a vector in $\mathbb{R}^k$ from n samples?

The confidence intervals for the coin toss setting can be easily extended to this case using the so called *Union Bound*:

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_k] \leq Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_k] \qquad (18.2)$$

This leads to the simplest confidence bound for estimating a vector in $\mathbb{R}^k$. Suppose the probability of the estimate being off by δ_i in the i-th coordinate is at most q_i. Then

$$\Pr[\text{estimate is off by } \vec{\delta}] \leq q_1 + q_2 + \cdots + q_k$$

where $\vec{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)$

18.4 Final Remarks

The CLT applies to many settings, but it doesn't apply everywhere. It is useful to clear up a couple of frequent misconceptions that students have:

1. Not every distribution involving a large number of samples is Gaussian. For example, scores on the final exam are usually not distributed like a Gaussian. Similarly, human heights are not really distributed like Gaussians.

2. Not everything that looks Gaussian-like is a result of the Central Limit Theorem. For instance, we saw that the distribution of weights in the sentiment model in Chapter 1 looked vaguely Gaussian-like, but they are not the sum of independent random variables as far as we can tell.