

# Contents

<b>0</b>	<b>What is a Random Variable?</b>	<b>3</b>
<b>1</b>	<b>Discrete Random Variables</b>	<b>3</b>
1.1	Distributions . . . . .	3
1.1.1	Joint Distribution . . . . .	3
1.1.2	Marginal Distribution . . . . .	3
1.1.3	Conditional Distribution . . . . .	4
1.2	Families of Discrete Random Variables . . . . .	4
1.2.1	Bernoulli Distribution . . . . .	4
1.2.2	Binomial Distribution . . . . .	4
1.2.3	Geometric Distribution . . . . .	4
1.2.4	Poisson Distribution . . . . .	5
1.2.5	Multinomial Distribution . . . . .	5
<b>2</b>	<b>Continuous Random Variables</b>	<b>5</b>
2.1	Distributions . . . . .	5
2.1.1	Joint Distribution . . . . .	6
2.1.2	Marginal Distribution . . . . .	6
2.1.3	Conditional Distribution . . . . .	6
2.1.4	Uniform Distribution . . . . .	6
2.1.5	Normal (Gaussian) Distribution . . . . .	6
2.1.6	Exponential Distribution . . . . .	7
<b>3</b>	<b>Properties and Quantities of Random Variables and Events</b>	<b>7</b>
3.1	Independence . . . . .	7
3.2	Conditional Probability . . . . .	7
3.2.1	Product Rule . . . . .	7
3.2.2	Chain Rule . . . . .	8
3.2.3	Law of Total Probability . . . . .	8
3.2.4	Bayes' Theorem . . . . .	8
3.3	Union and Intersection of Events . . . . .	8
3.4	Expectation . . . . .	9
3.4.1	Law of Total Expectation . . . . .	9
3.4.2	Law of Iterated Expectation . . . . .	9
3.5	Variance . . . . .	9
3.5.1	Law of Total Variance . . . . .	9
3.6	Covariance . . . . .	10
3.7	Correlation . . . . .	10
3.8	Moments . . . . .	10
<b>4</b>	<b>MLE and MAP</b>	<b>10</b>
4.1	Maximum Likelihood Estimation (MLE) . . . . .	10
4.2	Maximum A Posteriori (MAP) Estimation . . . . .	11

<b>5</b>	<b>Limit Theorems</b>	<b>11</b>
5.1	Sample Mean and Variance . . . . .	11
5.2	Weak Law of Large Numbers . . . . .	11
5.3	Strong Law of Large Numbers . . . . .	11
5.4	Central Limit Theorem . . . . .	12
<b>6</b>	<b>Concentration Inequalities</b>	<b>12</b>
6.1	Markov's Inequality . . . . .	12
6.2	Chebyshev's Inequality . . . . .	12
6.3	Hoeffding's Inequality . . . . .	12
6.4	Chernoff's Inequality . . . . .	12
6.5	Normal Distribution Tail Bound . . . . .	13

## 0 What is a Random Variable?

Despite its name, a random variable is not a variable; it is a function. Consider a set of outcomes  $\Omega$  from an experiment. A random variable  $X : \Omega \rightarrow \mathbb{R}$  will take a variable outcome  $\omega \in \Omega$  and map it to a real number. For example, consider the set of outcomes for two coin tosses  $\Omega = \{HH, HT, TH, TT\}$  and let  $X$  be the number of heads from these two coin tosses. Then  $X(HH) = 2$ ,  $X(\{HT, TH\}) = 1$ , and  $X(TT) = 0$ . Now, we can ask the question, “what is the **probability** that I flip two heads?”, or “what is  $\Pr(X(HH) = 2)$ ?” Specifying the outcomes in the argument of  $X$  is redundant, so instead we use shorthand notation  $\Pr(X = 2)$  or  $p_X(2)$ .

## 1 Discrete Random Variables

As stated in the previous section, a random variable  $X$  is a mapping from the outcome space  $\Omega$  to  $\mathbb{R}$ . As a result,  $\text{range}(X) \subset \mathbb{R}$ . A random variable is discrete if  $\text{range}(X)$  is countable. A set is countable if it is either finite or countably infinite ( $\exists f : \Omega \rightarrow \mathbb{N}$ ).

### 1.1 Distributions

Let  $x \in S$ , where  $S$  is a set in  $\mathbb{R}$ . A probability distribution  $\Pr(X = x)$  must satisfy three requirements:

1.  $0 \leq \Pr(X = x) \leq 1 \quad \forall x$
2.  $\sum_{x \in S} \Pr(X = x) = 1$
3. Let  $T \subset S$ , then  $\Pr(X \in T) = \sum_{x \in T} \Pr(X = x) \quad \forall T$

$\Pr(X = x)$  is also called the probability mass function (pmf)  $p_X(x)$  for a discrete random variable  $X$ ; for brevity,  $p_X(x)$  will be used throughout this document to indicate the probability that  $X$  takes on the value  $x$ . Furthermore, the cumulative mass function (cmf) can be determined from the pmf:

$$F(a) = \Pr(X \leq a) = \sum_{x \leq a} p_X(x)$$

#### 1.1.1 Joint Distribution

A joint probability distribution represents the probability of  $X$  and  $Y$  according to a joint distribution  $p_{XY}(x, y)$ . In other words,  $p_{XY}(x, y)$  can be used to find the probability of  $X = x$  **and**  $Y = y$ , i.e.  $p_{XY}(x, y) = \Pr(X = x \cap Y = y)$ . If  $X$  and  $Y$  are independent random variables, then  $p_{XY}(x, y) = p_X(x)p_Y(y)$ .

#### 1.1.2 Marginal Distribution

A marginal distribution only considers the probability distribution of one random variable  $X$  in the presence of other random variables. Let  $X, Y$  be two random variables, and let  $y \in T$  where  $T$  is a set. Then,

$$p_X(x) = \sum_{y \in T} p_{XY}(x, y)$$

### 1.1.3 Conditional Distribution

The conditional distribution of a random variable is the probability distribution of that random variable after observing the outcome of a different random variable. The distribution is given by

$$p_{X|Y}(x | y) = \frac{p_{X|Y}(x, y)}{p_Y(y)}$$

Note that if  $X$  and  $Y$  are independent, then

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

## 1.2 Families of Discrete Random Variables

### 1.2.1 Bernoulli Distribution

Define  $X$  such that  $X = 1$  when an outcome is a success and  $X = 0$  otherwise. Define  $p = \Pr(X = 1)$ , then  $X \sim \text{Bernoulli}(p)$ .

$$p_X(x) = \begin{cases} p & \text{if } X = 1 \\ 1 - p & \text{if } X = 0 \end{cases}$$

$$\mathbb{E}[X] = p$$

$$\text{Var}(X) = p(1 - p)$$

### 1.2.2 Binomial Distribution

Suppose that  $n$  independent experiments are performed with either success  $X = 1$  or failure  $X = 0$ . Define  $p$  to be the probability of success. Then  $X \sim \text{Binomial}(n, p)$ , and we may observe the probability of  $k$  successes.

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\mathbb{E}[X] = np$$

$$\text{Var}(X) = np(1 - p)$$

Additionally, consider  $n$  Bernoulli random variables  $X_1, X_2, \dots, X_n$ . If  $X = \sum_{i=1}^n X_i$ , then  $X$  is a binomial random variable.

### 1.2.3 Geometric Distribution

Define  $X$  such that  $X = 1$  when an outcome is a success and  $X = 0$  otherwise. Define  $p = \Pr(X = 1)$ . The geometric distribution observes the number of Bernoulli trials  $k$  until the first success, and so  $X \sim \text{Geometric}(p)$ .

$$p_X(k) = (1 - p)^{k-1} p$$

$$\mathbb{E}[X] = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1 - p}{p^2}$$

### 1.2.4 Poisson Distribution

Define  $X$  such that  $X = k$  for  $k \in \{0, 1, 2, 3, \dots\}$ . For some  $\lambda > 0$ , we may say that  $X \sim \text{Poisson}(\lambda)$ .

$$p_X(k) = \lambda^k \frac{e^{-\lambda}}{k!}$$

$$\mathbb{E}[X] = \lambda$$

$$\text{Var}(X) = \lambda$$

For large  $n$  and small  $p$ , the Poisson distribution where  $\lambda = n \cdot p$  is a good approximation to the Binomial distribution.

### 1.2.5 Multinomial Distribution

The multinomial distribution generalizes the binomial distribution. Instead of a binary success or failure, there are  $k$  outcomes.  $n$  is the number of independent experiments. The pmf  $p_X(\mathbf{x})$  accepts an  $n$ -length vector  $\mathbf{x}$  of possible outcomes, and each element in  $\mathbf{x}$ ,  $x_i$ , has probability  $p_i$  of occurring (we require that  $\sum_{i=1}^k p_i = 1$ ).

$$p(\mathbf{x}) = \frac{n!}{\prod_{i=1}^n (x_i!)} \prod_{i=1}^n p_i^{x_i}$$

$$\mathbb{E}[X_i] = np_i$$

$$\text{Var}(X_i) = np_i(1 - p_i)$$

## 2 Continuous Random Variables

A random variable  $X$  is continuous if its cumulative distribution function (cdf)  $F_X(x)$  is continuous for all  $x \in \mathbb{R}$ .

### 2.1 Distributions

Let  $A \subset \mathbb{R}$  be some set, then  $\Pr(X \in A) = \int_A f_X(x)dx$ , where  $f_X(x)$  is the probability density function (pdf) of  $X$ . The pdf has three properties (assuming  $A \in \mathbb{R}$ ):

1.  $f_X(x) \geq 0 \ \forall x$
2.  $\int_{-\infty}^{\infty} f_X(x)dx = 1$
3.  $\Pr(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f(x)dx$

The third property is the definition of the cumulative density function (cdf), given by  $F(x)$ , and has the following relationship to the pdf:

$$f(x) = \frac{dF_X(x)}{dx}$$

One thing to note about the pdf:  $\Pr(X = a) = \Pr(a \leq X \leq a) = \int_a^a f_X(x)dx = 0$ . In other words, the probability that a continuous random variable takes the value of a single real number is 0.

**2.1.1 Joint Distribution**

A joint probability distribution represents the probability of  $X$  and  $Y$  according to a joint distribution  $f_{XY}(x, y)$ . In other words,  $f_{XY}(x, y)$  can be used to find the probability of  $X = x$  **and**  $Y = y$ , i.e.  $f_{XY}(x, y) = \Pr(X = x \cap Y = y)$ . If  $X$  and  $Y$  are independent random variables, then  $f_{XY}(x, y) = f_X(x)f_Y(y)$ .

**2.1.2 Marginal Distribution**

A marginal distribution only considers the probability distribution of one random variable  $X$  in the presence of other random variables. Let  $X, Y$  be two random variables. Then,

$$f_X(x) = \int_y f_{XY}(x, y)$$

**2.1.3 Conditional Distribution**

The conditional distribution of a random variable is the probability distribution of that random variable after observing the outcome of a different random variable. The distribution is given by

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

Note that if  $X$  and  $Y$  are independent, then

$$f_{X|Y}(x | y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

**2.1.4 Uniform Distribution**

$X \sim \text{Uniform}(\alpha, \beta)$  if, on the interval  $[\alpha, \beta]$ ,

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & x < \alpha, x > \beta \end{cases}$$

$$\mathbb{E}[X] = \frac{\alpha + \beta}{2}$$

$$\text{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

$$F_X(x) = \frac{x - \alpha}{\beta - \alpha}$$

**2.1.5 Normal (Gaussian) Distribution**

If  $X$  is distributed normally, then  $X \sim \mathcal{N}(\mu, \sigma^2)$ . If  $\mu = 0$  and  $\sigma^2 = 1$ , then  $X$  is distributed according to the standard normal distribution.

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

$$\mathbb{E}[X] = \mu$$

$$\text{Var}(X) = \sigma^2$$

The cdf of a normal distribution is not in closed form:

$$F_X(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x - \mu}{\sigma\sqrt{2}} \right) \right]$$

where, for some constant  $t$ ,

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

### 2.1.6 Exponential Distribution

$X \sim \text{Exponential}(\lambda)$  if, given parameter  $\lambda > 0$ ,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

## 3 Properties and Quantities of Random Variables and Events

This section will present other important properties of random variables. The notation for discrete random variables (for example, summation instead of integration,  $p_X(x)$  instead of  $f_X(x)$ , etc.) will be used here, but the same properties apply to continuous random variables.

### 3.1 Independence

If either of the following two criteria are met  $\forall x, y, z$  such that  $X = x$ ,  $Y = y$ , and  $Z = z$ , then the two random variables  $X$  and  $Y$  are independent.

1.  $p_{XY}(x, y) = p_X(x)p_Y(y)$
2.  $p_{XY|Z}(x, y | z) = p_{X|Z}(x | z)p_{Y|Z}(y | z)$  (conditional independence between  $X$  and  $Y$ )

### 3.2 Conditional Probability

#### 3.2.1 Product Rule

Let  $X$  and  $Y$  be two random variables. The definition of conditional probability is

$$p_{X|Y}(x | y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

By rearranging, we obtain the product rule for conditional probability:

$$p_{XY}(x, y) = p_{X|Y}(x | y)p_Y(y)$$

### 3.2.2 Chain Rule

The chain rule is essentially a generalized form of the product rule. Let  $X_1, X_2, \dots, X_n$  be random variables. Then,

$$\begin{aligned} & p_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n) \\ &= p_{X_1}(x_1) p_{X_2|X_1}(x_2, x_1) p_{X_3|X_2, X_1}(x_3 | x_2, x_1) \cdots p_{X_n|X_{n-1}, X_{n-2}, \dots, X_1}(x_n | x_{n-1}, x_{n-2}, \dots, x_1) \\ &= \prod_{i=1}^n p_{X_i|X_{i-1}, X_{i-2}, \dots, X_1}(x_i | x_{i-1}, x_{i-2}, \dots, x_1) \end{aligned}$$

### 3.2.3 Law of Total Probability

Let  $A$  be some event in a sample space, and let  $B_1, B_2, \dots, B_n$  be mutually exclusive events that partition the entire sample space. Then The Law of Total Probability states that

$$\Pr(A) = \sum_{i=1}^n \Pr(A \cap B_i) \Pr(B_i)$$

We can obtain something that looks like the Law of Total Probability for random variables as well. Let  $X$  and  $Y$  be random variables with a joint probability distribution  $p_{XY}(x, y)$ . Using the definition of conditional probability, the Law of Total Probability can be obtained.

$$p_X(x) = \sum_y p_{XY}(x, y) = \sum_y p_{Y|X}(y | x) p_Y(y)$$

### 3.2.4 Bayes' Theorem

Bayes' Theorem can be determined by equating both sides of the product rule:

$$p_{X|Y}(x | y) p_Y(y) = p_{XY}(x, y) = p_{Y|X}(y | x) p_X(x)$$

$$p_{X|Y}(x | y) p_Y(y) = p_{Y|X}(y | x) p_X(x)$$

$$p_{X|Y}(x | y) = \frac{p_{Y|X}(y | x) p_X(x)}{p_Y(y)}$$

In this case,  $p_{X|Y}(x | y)$  is the *posterior*,  $p_{Y|X}(y | x)$  is the *likelihood*,  $p_X(x)$  is the *prior*, and  $p_Y(y)$  is the *normalization* term. Using the Law of Total Probability, the normalization term can be substituted for:

$$p_{X|Y}(x | y) = \frac{p_{Y|X}(y | x) p_X(x)}{\sum_x p_{Y|X}(y | x) p_X(x)}$$

## 3.3 Union and Intersection of Events

Let  $A$  and  $B$  be two events in the same sample space. Then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

and

$$\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B)$$



### 3.4 Expectation

The expected value, or mean, of a random variable is given by

$$\mathbb{E}[X] = \sum_x xp_X(x)$$

Or, more generally, for some function  $g(X)$ ,

$$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$$

#### 3.4.1 Law of Total Expectation

The Law of Total Expectation is similar to the Law of Total Probability, but is used to determine the expected value of a random variable. Let  $X$  be a random variable and let  $A_1, A_2, \dots, A_n$  be mutually exclusive events that partition the entire sample space. Then by the Law of Total Expectation,

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X | A_i] \Pr(A_i)$$

#### 3.4.2 Law of Iterated Expectation

For two random variables  $X$  and  $Y$ ,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]]$$

### 3.5 Variance

Let  $\mu = \mathbb{E}[X]$ . The variance  $\text{Var}(X)$  of  $X$  is given by

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

Alternatively,

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \left( \sum_x x^2 p_X(x) \right) - \left( \sum_x x p_X(x) \right)^2$$

The standard deviation is  $\sqrt{\text{Var}(X)}$ .

#### 3.5.1 Law of Total Variance

For two random variables  $X$  and  $Y$ ,

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y])$$

### 3.6 Covariance

Let  $\mu_X = \mathbb{E}[X]$  and  $\mu_Y = \mathbb{E}[Y]$ , and let  $n$  be the number of outcomes for  $X$  and  $Y$ . The covariance  $\text{Cov}(X, Y)$  is given by

$$\text{Cov}(X, Y) = \mathbb{E}[X - \mu_X][Y - \mu_Y] = \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)p_{X_i Y_i}(x_i, y_i)$$

Note that  $\text{Cov}(X, X) = \text{Var}(X)$

### 3.7 Correlation

The correlation of  $X$  and  $Y$  is a value between -1 and 1. A correlation of -1 indicates that the random variables are perfectly inversely related, 0 indicates no relationship, and 1 indicates that the two variables vary together perfectly. The correlation coefficient  $\rho_{XY}$  of  $X$  and  $Y$  is given by

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

### 3.8 Moments

The moment generating function of a random variable  $X$  is defined as

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_x e^{tx} p_X(x)$$

for all  $t \in \mathbb{R}$ . Let  $k$  be an integer greater than 0, then the  $k$ th moment of  $X$  is  $\mathbb{E}[X^k]$  and the  $k$ th central moment of  $X$  is  $\mathbb{E}[(X - \mathbb{E}[X])^k]$ . The  $k$ th moment may be calculated in the following way:

$$\mathbb{E}[X^k] = \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0}$$

## 4 MLE and MAP

### 4.1 Maximum Likelihood Estimation (MLE)

Given a probability distribution  $p_{X_i|Y}(x_i | y)$ , it's often useful to estimate random variable  $Y$  in order to maximize the probability the joint distribution  $p_{X_1, X_2, \dots, X_n|Y}(x_1, x_2, \dots, x_n | y)$ , where all  $X_i$  are independent. The likelihood function is defined as

$$L(y | x_1, x_2, \dots, x_n) = p_{X_1, X_2, \dots, X_n|Y}(x_1, x_2, \dots, x_n | y) = \prod_{i=1}^n p_{X_i|Y}(x_i | y)$$

Since  $p_{X_i|Y}(x_i | y)$  is always between 0 and 1, this product often results in a very small number. As a result, it is more convenient to consider the loglikelihood:

$$l(y | x_1, x_2, \dots, x_n) = \log(L(y | x_1, x_2, \dots, x_n)) = \log\left(\prod_{i=1}^n p_{X_i|Y}(x_i | y)\right) = \sum_{i=1}^n \log(p_{X_i|Y}(x_i | y))$$

Using these equations,  $y$  may be computed

$$y = \arg \max_y l(y | x_1, x_2, \dots, x_n) = \arg \max_y \sum_{i=1}^n \log(p_{X_i|Y}(x_i | y))$$

## 4.2 Maximum A Posteriori (MAP) Estimation

MAP Estimation is nearly identical to MLE, except that it utilizes the prior from Bayes' Theorem.

$$L(y \mid x_1, x_2, \dots, x_n) = \prod_{i=1}^n \frac{p_{X_i|Y}(x_i \mid y)p_Y(y)}{p_{X_i}(x_i)}$$

Since we are maximizing with respect to  $y$ ,  $p_{X_i}(x_i)$  is just a constant and can be ignored when finding the argmax.

$$y = \arg \max_y \log \left( \prod_{i=1}^n p_{X_i|Y}(x_i \mid y)p_Y(y) \right) = \arg \max_y \sum_{i=1}^n \log(p_{X_i|Y}(x_i \mid y)p_Y(y))$$

## 5 Limit Theorems

### 5.1 Sample Mean and Variance

Let  $X_1, X_2, \dots, X_n$  be independent random variables and have mean  $\mathbb{E}[X_i] = \mu$  and variance  $\text{Var}(X_i) = \sigma^2$ . Then,

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu$$

and

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$$

### 5.2 Weak Law of Large Numbers

The Weak Law of Large Numbers states that the sample mean of a collection of random variables will converge in probability to the expected value as the number of samples  $n$  increases. For any  $\epsilon > 0$  and for i.i.d. random variables  $X_1, X_2, \dots, X_n$ ,

$$\lim_{n \rightarrow \infty} \Pr \left( \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \epsilon \right) = 0$$

### 5.3 Strong Law of Large Numbers

As the name implies, the Strong Law of Large Numbers is a stronger condition than the weak law of large numbers. It says that the sample mean of a collection of random variables will converge almost surely (with a probability of 1) to the expected value as the number of samples  $n$  increases. For i.i.d. random variables  $X_1, X_2, \dots, X_n$ ,

$$\Pr \left( \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = \mu \right) = 1$$

## 5.4 Central Limit Theorem

For i.i.d. random variables  $X_1, X_2, \dots, X_n$  with the same mean  $\mu$  and variance  $\sigma^2$ . Let  $X = \sum_{i=1}^n X_i$ . Normalize  $X$  by subtracting its mean and standard deviation and let this new random variable be  $Z$ :

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}} = \frac{\sum_{i=1}^n X_i - \mu}{\sigma\sqrt{n}}$$

As  $n \rightarrow \infty$ ,  $Z \rightarrow \mathcal{N}(0, 1)$ .

## 6 Concentration Inequalities

Concentration inequalities are bounds that show the deviation of a random variable from the expected value or some other value. Thus, they provide information about where the mass (or density) of the random variables probability distribution is concentrated.

### 6.1 Markov's Inequality

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

### 6.2 Chebyshev's Inequality

This inequality is derived directly from Markov's inequality; it provides a tighter bound. Given the variance  $\sigma^2$  of a random variable  $X$ ,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

### 6.3 Hoeffding's Inequality

Let there be  $n$  random variables  $X_1, X_2, \dots, X_n$  where each  $X_i \in [a_i, b_i]$ , for  $a_i, b_i \in \mathbb{R}$ . Then,

$$\Pr\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(\frac{-2nt^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

### 6.4 Chernoff's Inequality

Chernoff's inequality is obtained by applying Markov's inequality to the moment generating function:

$$\Pr\left(\sum_{i=1}^n X_i \geq r\right) \leq e^{-tr} \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right]$$

Recall that the moment generating function  $e^{tX}$  depends on a free parameter  $t > 0$ . Thus, we can choose  $t$  such that the bound is minimized to ensure the tightest bound possible:

$$\Pr\left(\sum_{i=1}^n X_i \geq r\right) \leq \min_{t>0} e^{-tr} \mathbb{E}\left[\prod_{i=1}^n e^{tX_i}\right]$$

## 6.5 Normal Distribution Tail Bound

Let  $X \sim \mathcal{N}(0, 1)$ , then

$$\Pr(X \geq t) \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$