Normal Distribution

listenzcc

February 5, 2021

Abstract

The hand book of Normal distribution. How it is formulated, what it can predict and why we use it.

Contents

1 Normal Distribution in Classic View

1.1 Binominal Distribution

Perform experiment for n times, we assume the trials are independent and follow the same distribution. The output of the experiment is noted as 1 or 0, with no vague. The probability of observing the 1 output is noted as p. Then, we have

$$P(n, m) = (n, m) \cdot p^m \cdot (1 - p)^{n-m} \tag{1}$$

where m refers the fact that we observe 1 output for m times.

It is easy to see that the $P(n, m)$ produces a distribution since

$$\sum_{i=0}^{n} P(n, i) = 1, i \in \mathcal{N}$$

Use the computation of **expectation and Variation** of the random variable, we have

$$\mathcal{E}(m) = n \cdot p$$
$$\mathcal{V}(m) = n \cdot p \cdot (1 - p)$$

1.2 Poisson Distribution

The poisson distribution is the infinity binominal distribution (see (1)), when p is **small** and n is **large**. It is defined as

$$P(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \lambda = np \tag{2}$$

Proof. Use the equation of $\lambda = np$, we can rewrite (1) as

$$P(n, k) = (n, k)(\frac{\lambda}{n})^k(1 - \frac{\lambda}{n})^{n-k}$$

Since p is small and n is large, we have k is relatively small compared to n. As a result, in infinity case,

$$\lim_{n\to\infty} \frac{(n, k)}{n^k} = \frac{1}{k!}$$
$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^{n-k} = e^{-\lambda}$$

Hence proved. □

In practice, we require $\lambda < 1$ to produce a valid approximation.

To compute the expectation and variation of the poisson distribution, we use the taylor series of Exp function

$$e^{\lambda t} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}, t_0 = 0$$

it naturally guarantees the property of PDF that

$$\sum_{k=0}^{\infty} P(k) = 1$$

Use the definition of the expectation, we have

$$\mathcal{E}(k) = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \cdot e^{-\lambda}$$
$$\mathcal{E}(k) = \lambda$$

Use the definition of the variation, we have

$$\mathcal{E}(k^2) = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} \cdot e^{-\lambda}$$
$$\mathcal{E}(k^2) = \lambda + \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} \cdot e^{-\lambda}$$
$$\mathcal{E}(k^2) = \lambda + \lambda^2$$

where we used the idea of $k = 1 + (k - 1)$. Thus, the variation is

$$\mathcal{V}(k) = \lambda$$

1.3 Normal Distribution

When n is large and p is not so small, the poisson distribution fails on approximate the binominal distribution. The Normal distribution is used as a more general replacement.

Basically, when n, np and nq are large, the binominal distribution is well approximated by the Normal distribution

$$p(x) = (n, x)p^x q^{n-x} \approx \frac{1}{\sqrt{2\pi npq}} e^{-(x-np)^2/2npq}$$

where $p + q = 1$. See the website [1] for detail.

And they are linked based on Sterling's formula,

$$n! = n^n e^{-n}\sqrt{2\pi n}[1 + \mathcal{O}(1/n)] \tag{3}$$

See the website [2] for detail.

Formally, the normal distribution is expressed as

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

where expectation $\mu = \lim_{n\to\infty} np$ and variation $\sigma^2 = \lim_{n\to\infty} npq$.

1.3.1 The Pdf of Normal Distribution

The normal distribution is usually expressed as $p(x) \sim \mathcal{N}(\mu, \sigma^2)$.

Proof. The Pdf of normal distribution is a pdf.

$$\int_{-\infty}^{\infty} p(x) = 1, p(x) > 0$$
$$p(x) \sim \mathcal{N}(\mu, \sigma^2)$$

Firstly, the $p(x) > 0$ is obvious. Secondly, using the equation of $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, the integral is 1.

Hence proved. □

The expectation and variance are

$$\mathcal{E}(x) = \mu$$
$$\mathcal{V}(x) = \sigma^2$$

Proof. The expectation and variance of normal distribution are μ and σ^2.

Use the definition of expectation, and variable change with $y = x - \mu$, one gets

$$\mathcal{E}(x) = \int_{-\infty}^{\infty} (y + \mu)p(y + \mu)dy$$

Since integrand function of $y \cdot p(y + \mu)$ is odd function, and use the PDF property of normal distribution, one gets $\mathcal{E}(x) = \mu$.

Use the value of $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$, and variable change with $y = \frac{x-\mu}{\sqrt{2\sigma^2}}$, one gets

$$\mathcal{V}(x) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx = \sigma^2$$

Hence proved. □

The normal distribution of $\mu = 0$ and $\sigma^2 = 1$ is known as the *standard normal distribution*.

[1] http://scipp.ucsc.edu/~haber/ph116C/NormalApprox.pdf

[2] https://www.researchgate.net/publication/237571154_A_Very_Short_Proof_of_Stirling's_Formula

1.4 Multivariate Normal Distribution

In this section, multivariate normal distributions is described, it mainly answers the question of the density function of the linear combination of normal distribution variables.

Standard Multivariate Normal Distribution

The *standard multivariate normal distribution* is the joint distribution of several standard normal distribution variables, and the variables are independent with each other. It says the X has a standard multivariate normal distribution if its joint probability density function is

$$f_X(x) = (2\pi)^{-K/2} e^{-x^T x/2} \tag{5}$$

where X be a $K \times 1$ continuous random vector, $X \in \mathbb{R}^K$, and the elements of X are *independent* with each other.

Multivariate Normal Distribution in General

The *multivariate normal distribution in general* is the joint distribution of several variables of normal distribution, the difference is the variables are not under certain constraints, like being independent with each other. In fact, the density function is

$$f_X(x) = (2\pi)^{-K/2} |det(V)|^{-1/2} e^{-1/2(x-\mu)^T V^{-1}(x-\mu)} \tag{6}$$

where $X \in \mathbb{R}^K$, μ and V is the expectation vector and variance matrix of X. It demonstrates the fact the standard normal distribution is the special case of multivariate normal distribution.

Proposition 1.1. *Let $X \in \mathbb{R}^K$, the μ and V are the expectation vector and variance matrix. The density can be formulated as*

$$X = \mu + \Sigma Z$$

where $Z \in \mathbb{R}^K$ is a standard normal distribution *vector, and $V = \Sigma\Sigma^T$.*

Proof. Use the inverse of Σ

$$f_X(x) = \frac{1}{|det(\Sigma)|} f_Z(\Sigma^{-1}(x-\mu))$$

where $Z = \Sigma^{-1}(x-\mu)$ is a *standard normal distribution* vector.

It is convenience to use (5).

And the variance is computed as $V = \Sigma\Sigma^T$

Hence proved. □

2 Family Members

There are several common distributions in the family of Normal distribution.

2.1 Chi-squared Distribution

If $Y_i \sim \mathcal{N}(0,1)$, and Y_is are independent with each other, then

$$\chi^2 \equiv \sum_{i=1}^{r} Y_i^2 \tag{7}$$

is distributed as Chi-squared χ^2 distribution with r degrees of freedom. The symbolic notion is $p_r(x) \sim \chi^2(r)$.

The pdf of Chi-squared distribution is

$$p_r(x) = \frac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}, 0 < x < \infty \tag{8}$$

2.1.1 Mean and Variance

The mean and variance of the chi-squared distribution is

$$Mean \triangleq E(x) = r$$
$$Variance \triangleq E(x^2) - E^2(x) = 2r$$

2.1.2 The Pdf of Chi-squared distribution

Lemma 2.1. *To get the pdf of a Chi-squared distribution, we have to prove that*

$$p_n(x) \propto x^{n/2-1} \cdot e^{-x/2}$$

in which, $x = \sum_{i=1}^{n} y_i^2$ *and* $y_i \sim \mathcal{N}(0,1)$*. Each* y_i *are independent.*

Proof. The joint probability of $\{y_1, y_2, \ldots, y_n\}$ is

$$p_{joint} = exp(\sum_{i=1}^{n} -y_i^2/2)$$

Thus, the cumulative sum of $p_n(x)$ can be computed using surface integral

$$P_n(r < \sqrt{x}) \propto \int_S p_{joint} ds$$
$$P_n(r < \sqrt{x}) \propto \int_S e^{-r^2/2} ds$$

in which, S refers the volume of a sphere with radius of x.

Transfer the integral into sphere coordinates, we have

$$P_n(r < \sqrt{x}) \propto \int_{r=0}^{\sqrt{x}} e^{-r^2/2} r^{(n-1)} dr$$

Derivate to x, we have

$$\frac{\partial}{\partial x} P_n(r < \sqrt{x}) \propto e^{-r^2/2} r^{(n-1)} x^{-1/2}$$
$$\frac{\partial}{\partial x} P_n(r < \sqrt{x}) \propto x^{n/2-1} \cdot e^{-x/2}$$

The first step is because of the Newton's integral rule, the second step is based on the replacement of $r = \sqrt{x}$.

Hence proved.

□

Lemma 2.2. *Next, we have to prove that the integral of* $p_n(x)$ *with* $p_n(x) \sim \chi^2(n)$ *is*

$$\int_0^\infty p_n(x) dx = \Gamma(n/2) \cdot 2^{r/2}$$

Proof. Use the definition of Γ function

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx$$

Use variable replacement of $z = 2x$, we have

$$\Gamma(n) = 2^{-n} \int_0^\infty z^{n-1} e^{-z/2} dz$$

Then, use substitution of $n = n/2$, we have

$$\Gamma(n/2) \cdot 2^{n/2} = \int_0^\infty z^{n/2-1} e^{-z/2} dz$$

Hence proved.

□

2.2 Student's T Distribution

The student's t distribution describes a random variable T of the form

$$T = \frac{\bar{x} - m}{s/\sqrt{N}} \tag{9}$$

where $\bar{x}$ is the sample mean value of all N samples, m is the population mean value and s is the population standard deviation.

Or, in a more formal one

$$T = \frac{X}{\sqrt{Y/r}} \tag{10}$$

where $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi_r^2$.

The pdf of Student's t-distribution is

$$t_r(x) = \frac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})\sqrt{r\pi}}(1 + \frac{x^2}{r})^{-\frac{r+1}{2}}, -\infty < x < \infty \tag{11}$$

2.2.1 Mean and Variance

The mean and variance of the Student's t-distribution is

$$Mean \triangleq E(x) = 0$$
$$Variance \triangleq E(x^2) - E^2(x) = \frac{r}{r-2}$$

2.2.2 Relationship with Normal Distribution

It is easy to see that $\lim_{r\to\infty} t_r(x) \sim \mathcal{N}(0,1)$. It demonstrates that when r is large enough, the Student's t-distribution is equalize to Normal Distribution. It should to be noted that the calculation of $\frac{\Gamma(\frac{r+1}{2})}{\Gamma\frac{r}{2}\sqrt{r\pi}}$ is somehow difficult, however, and fortunately, the value is constant with x when $r \to \infty$. Since the other factor can be formulated as the form of $e^{\frac{x^2}{2}}$, the constant can be calculated using the property of Normal distribution. Thus, the equation is also an useful approximation to the constant.

2.2.3 The pdf of Student's T Distribution

Here, we provide a simple computation of the pdf of the Student's t-distribution.

$$T = \frac{X}{\sqrt{Y/r}}$$

in which $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2(r)$, and they are independent. Thus, we have

$$p(x) \propto e^{-x^2/2}$$
$$p(y) \propto y^{r/2-1} \cdot e^{-y/2}$$

The random variable t follows the equation $t = \frac{x}{\sqrt{y/r}}$.

Lemma 2.3. *Since then we want to prove that*

$$p(t) \propto (1 + \frac{t^2}{r})^{-\frac{r+1}{2}} \tag{12}$$

Proof. The joint probability of $p(x, y)$ matches

$$p(x, y) \propto e^{-x^2/2} \cdot y^{r/2-1} \cdot e^{-y/2}$$

And the divergence of $p(x, y)$ is $p(x, y)dxdy$. We can use the variable replacement of

$$y = \frac{x^2}{t^2} \cdot r$$
$$\frac{dy}{dt} \propto \frac{x^2}{t^3}$$

Thus we have the joint probability of $p(x, t)$ matches

$$p(x, t) \propto e^{-x^2/2} \cdot (\frac{x^2}{t^2})^{r/2-1} \cdot e^{-\frac{x^2}{2t^2}r} \cdot \frac{x^2}{t^3}$$

The probability of $p(t)$ can be expressed as

$$p(t) \propto \int_x p(x, t)dx$$

Analysis the expression, we have

$$p(t) \propto t^{-r-1} \int_x x^r \cdot e^{-\frac{1}{2}(1+\frac{r}{t^2})x^2} dx$$
$$p(t) \propto t^{-r-1} \cdot (1 + \frac{r}{t^2})^{\frac{-r-1}{2}} \int_z z^r \cdot e^{z^2} dz$$
$$p(t) \propto (t^2 + r)^{-\frac{r+1}{2}}$$
$$p(t) \propto (1 + \frac{t^2}{r})^{-\frac{r+1}{2}}$$

The process uses the integral of Γ function is constant, and r is constant. □

After that, combining with the following, we should finally have the pdf function.

Lemma 2.4. *The values of $t_r(x)$ is positive and the integral is* 1.

$$\int_{-\infty}^{\infty} t_r(x)\,\mathrm{d}x = 1$$

Proof. Consider the variable part of Student's t-distribution

$$f(x) = (1 + \frac{x^2}{r})^{-\frac{r+1}{2}}, -\infty < x < \infty$$

use a replacement as following

$$x^2 = \frac{y}{1-y}$$

it is easy to see that $\lim_{y\to 0} x = 0$ and $\lim_{y\to 1} x = \infty$. Additionally, the x^2 is even function. Thus we can write the integral of $f(x)$

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 2\sqrt{r}\int_0^1 (\frac{1}{1-y})^{-\frac{r+1}{2}}\,\mathrm{d}(\frac{y}{1-y})^{\frac{1}{2}}$$

it is not hard to find out that the integral may end up with

$$\sqrt{r}\int_0^1 (1-y)^{\frac{r}{2}-1} y^{\frac{1}{2}-1}\,\mathrm{d}y = \sqrt{r}B(\frac{r}{2},\frac{1}{2})$$

Finally the Normalization factor has to be

$$\frac{\Gamma(\frac{r+1}{2})}{\sqrt{r}\Gamma(\frac{r}{2})\Gamma(\frac{1}{2})}$$

which makes the integral of $t_r(x)$ is 1.

□

3 Examples

One important usage of normal distribution is to valuate the output of the experiment. Since a large number of observations can be fitted into a normal distribution.

In an experiment, the results can be variance on different experiment trails. In the standard analysis pipeline, the *random variables* which can be fitted into the normal distribution, are formulated based on the obtained outputs.

The statistical analysis is commonly used to obtain the underlying ground truth value of the outputs. Particularly, we are interested in the *expectation* and *variance* of the random variables. The expectation is used to estimate the ideal value, often refers noise-free situation, of one trial. The variance is used to estimate the reasonable range of the output of one trial. Additionally, the distribution of the selected random variable is also used to *valuate* or *compare* the output of the experiment.

3.1 Paired and Un-paired T-test

The t-test method is the useful method to compare the effects of *two* factors. A common situation is to compare the difference of two methods doing the same job. We assume the outputs can be evaluated as the random variable fitting in *student's t distribution.* Then the sum of the

variables is also fitting in *student's t distribution*, but the only question is the parameters being left as unknown.

To perform valid comparison of the variables of two methods, the *generation* of the random variables and the *estimation* of the parameters should be investigated. Back to the question of binary comparison. If for each experiment trial, we can apply the two methods to it in parallel, the *paired t-test* method will be used. If we can only apply the method on a trial once, the *un-paired t-test* method will be used.

Paired T-test

Thinking of an experiment of several trials, the aim is to test whether there is significant difference between two methods. The methods are applied to every trial in parallel. The outputs of the two methods are subtracted to generate the obtained random variables.

We assume the random variables are in the student's t distribution. And the parameters are defined as the prior knowledge. Naturally, the average of the obtained value refers to a position in the distribution. As the result, the p-value of the position refers the possibility of the obtained value happens according to the distribution with the prior knowledge.

If the p-value is *small enough*, we can reject the prior assumption reasonable. In practice, a certain *threshold of p-value* is used to determine whether we can call a rejection, like the well known threshold of 0.05.

In one word, we assume the substitution of the obtained values are of student's t distribution. If the obtained values occur at a very low possibility, we can reject the prior assumption for sure. That is also the main idea of *hypothesis testing*.

Un-paired T-test

In a different situation, we can not apply the two methods on each experiment trials in parallel. Then the method of substitution can not be used. To overcome the problem, we *assume* the values of two methods are of the student's t distribution with the *same* parameters.

Two ideas can be used to solve the hypothesis testing problem,

- Estimate the parameters using the values of one method, testing the p-value of the other method.
- Compute the substitution of the averaged values of the two methods, testing the p-value of the substitution.

however, the computing method is the same although the ideas are different. The key is to compare the substitution of the averaged values of the two methods, instead of compute the p-value of the substitution of every trials.

Moreover, the variance of the student's t distribution is unknown in a large number of circumstances. The parameters can be estimated by the sample values. It is needed to be noted that the sample variance is the *biased* estimation of the overall variance. The factor of $\frac{n}{n-1}$ is used to time the sample variance to estimate the overall variance. It also suggests that the sample variance should be smaller than the overall variance when sample size is small. It matches the extreme situation of sample size is 0, when the sample variance is 0 (which means it is too small), and the overall variance is positive.

3.2 How to Determine Sample Count

There are a number of experiments are performed to determine the ground truth value of something. Normally, we can denote the value as a random

variable of X. The population size is N. The mean and variance of the population are stable, and known as μ and σ^2. The experiment trials are performed as randomly sampling n times from the population. It can be different from the sampling methods.

3.2.1 Variance of Different Sampling Methods

One method is randomly *sampling with replacement*, the sample variance is

$$\mathcal{V}(\bar{X}) = \frac{1}{n}\sigma^2 \tag{13}$$

the other is randomly *sampling without replacement*, the sample variance is

$$\mathcal{V}(\bar{X}) = \frac{1}{n}\sigma^2\frac{N-n}{N-1} \tag{14}$$

it shows that the without-replacement provides lower variance as n increasing. The decreasing is monotone, from $\frac{1}{n}\sigma^2$ to 0.

Additionally, no matter the sampling methods we use, the expectation of the mean value is the same as μ.

Based on *Central Limit Theorem*, a common functional is satisfied

$$P(|\hat{\mu} - \mu| > d) < \alpha$$
$$P(\frac{|\hat{\mu} - \mu|}{\hat{\sigma}} > z) < \alpha$$

where α refers the confidential level, d refers the margin error, and z refers the z-score of the α value to the standard normal distribution. And d is pre-defined value, refers the reasonable (tolerable) error range of the experiment. And $\hat{\sigma} = \frac{d}{z}$. One can also see the fact that the z controls the value of α.

In a simple word, we would like to restrict the error of the mean value's estimation lower than d with the confidential level of α. To achieve the goal, the change-able variable is $\hat{\sigma}$, the smaller it is, the better. It is because the smaller the $\hat{\sigma}$ is, the larger the z is, and thus the smaller the α is to match the pre-defined level.

3.2.2 Lower Bound of Samples Count

As a result, because of the monotone of the sample variance, the way is simply to increase the sampling size of n, until it satisfies

$$\mathcal{V}(\bar{X}) = \frac{d^2}{z^2}$$

Finally, one can conclude the sample size has the lower bound, which is computed as

$$With - Replacement : \frac{1}{n}\sigma^2 = \frac{d^2}{z^2} \tag{15}$$

$$Without - Replacement : \frac{N-n}{N-1}\frac{1}{n}\sigma^2 = \frac{d^2}{z^2} \tag{16}$$

3.2.3 Application in Population Polling

For example, the people can either agree or disagree with a proposal. The aim of the polling is to uncover the ratio of agreement, p, in the population. Then it is reasonable to believe the p is following binominal distribution,

whose variance is population variance which equals to $p(1-p)$. And the sampling process is without-replacement sampling, since we only ask one person for one time. As a result, the minimum sampling size n to lower the error less than d at confidential level of z is

$$n = N\frac{z^2\sigma^2}{d^2(N-1)+z^2\sigma^2}$$

where $\sigma^2 = p(1-p)$. It should be noted that the larger that σ^2, the more samples are needed. Thus, it is relatively easier (less samples are needed) to determine a large or small value of p, since its variance is smaller.

Following is the brief description of how we come to the conclusion.

3.2.4 Explaining

Start with the two situations: One is the *independent situation*, it means the sampling *with-replacement* is performed during the experiment; The Other is the *dependent situation*, it refers *without-replacement* sampling experiment is performed. The dependency refers the relationship between one random variable and another. For example, under with-replacement situation, the sampling process for every trial are the same since the whole set is the same. But, under without-replacement situation, the whole set is decreased as sampling process going on. As a result, the output of latter trials are affected by the pervious.

The following description is inspired by the website paper [3].

To have a overlook of the population as a whole. The population mean μ and variance σ^2 is given as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{m} \xi_i n_i$$
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{m} {\xi_i}^2 n_i - \mu^2$$

where x_i refers the sampling variables, ξ_i and n_i refers possible value of x and its count. We can also denote the frequency of ξ_i by the probability of $p_i = n_i/N$.

Random sampling with replacement

Let's start with the simpler case of *with-replacement sampling*. We let $X_1, X_2, \ldots, X_n$ to be the n obtained random variables. We accept the idea that X_is are independent. The mean and variance estimation can be easily computed based on the observations

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\sigma^2} = \frac{1}{n-1}\mathcal{E}(\sum_{i=1}^{n}(X_i-\bar{X})^2)$$

Use the definition of mean and variance, one can conclude the $\hat{\mu}$ and $\hat{\sigma^2}$ are the unbiased estimation of the mean and variance.

$$\mathcal{E}(\hat{\mu}) = \mu$$
$$\mathcal{E}(\hat{\sigma^2}) = \sigma^2$$

[3] `http://dept.stat.lsa.umich.edu/~moulib/sampling.pdf`

Random sampling without replacement

Under sampling without-replacement setting, the output of latter trials are affected by the previous. It leads to dependency between samples. To estimate the mean value, it is the same as the with-replacement setting,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$$

To estimate the variance, it is a computation of the 2nd order moment of the random variable. The joint probability between two random variables, like X_i and X_j, is required

$$P(X_i = \xi_s, X_j = \xi_r) = \frac{n_s}{N}\frac{n_r}{N-1}, s \neq r$$
$$P(X_i = \xi_s, X_j = \xi_s) = \frac{n_s}{N}\frac{n_s - 1}{N-1}$$

It is also important to mention that the equations of the joint probability are always satisfied for every i and j, as long as the random sampling process is selecting samples randomly one-by-one, without applying other restrictions.

$$Cov(X_i, X_j) = Cov(X_1, X_2)$$

where 1 and 2 are simply two constant foot note, in fact they can be other values, as long as they do not equal.

The variance of averaged value of $\bar{X}$ is expressed as

$$\mathcal{V}(\bar{X}) = \mathcal{E}(Cov(X_i, X_j)) \tag{17}$$
$$\mathcal{V}(X_i) = Cov(X_i, X_i) \tag{18}$$

Then the variance of the averaged random variable is computed as

$$\begin{aligned}\mathcal{V}(\bar{X}) &= \frac{1}{n^2}\sum_{i,j} Cov(X_i, X_j) \\ &= \frac{1}{n^2}(\sum_{i=1}^{n}\mathcal{V}(X_i) + \sum_{i\neq j} Cov(X_i, X_j)) \\ &= \frac{1}{n}\sigma^2 + \frac{n-1}{n}Cov(X_1, X_2)\end{aligned}$$

where $Cov(X_1, X_2) = E(X_1, X_2) - \mu^2$. Additionally, on boundary situation of $n = 1$ or $Cov(X_i, X_j) = 0$, the variance is as the same as the single variable situation or independent situation. Another interesting point is the more samples we get, the less variance of single variable goes into the system, and the more covariance is taken into account.

Use the joint probability of $X_1 = \xi_i$ and $X_2 = \xi_j$, we have

$$\begin{aligned}E(X_1, X_2) &= \sum_{i,j}\xi_i\xi_j p_{ij} \\ &= \mu^2 - \frac{1}{N-1}\sigma^2\end{aligned}$$

Thus,

$$Cov(X_1, X_2) = -\frac{1}{N-1}\sigma^2$$

Finally,

$$\mathcal{V}(\bar{X}) = \frac{1}{n}\sigma^2(1 - \frac{n-1}{N-1})$$

under boundary situation of $n = N$, the variance equals to the population variance.