# Statistical Analysis using Mathematical Tool

listenzcc

December 14, 2020

**Abstract**

Random variables are of almost everywhere and everything. Statistical analysis helps to find the ground truth of the variabilities. The article tries to explain the basic concepts.

## Contents

## 1 Concepts

The section will list basic concepts of *random variables* and *statistical analysis.*

### 1.1 Random Variables

A variable is *random* means it is not fixed. It turns out that one can obtain different values every time. The reason behind can be systemic or arbitrary. The aim of statistical analysis is to uncover the reason, however it usually matters little during the calculation. But the analysis can be valid only if the *random variable follows certain rules* instead of being totally unreasonable.

## 1.2 Statistics and Distribution

To understand the rules, the obtained values should be calculated carefully to formulate *new meaningful values.* The new values are called *statistics.* The statistics are asserted to be following some certain *distribution.* The distribution refers to the rule that controls the uncertainty of the random variable. A classic distribution contains two parts:

- Values $x$: The possible values of the statistic
- Probabilities $p(x)$: The probabilities of the values

It is also intrinsic that the sum of the probabilities should be equal to ONE, no more no less.

$$\int p(x) = 1, \{\forall x | p(x) \in (0, 1)\}$$

The function of $p(x)$ is called *probability distribution function (PDF).*

## 1.3 Statistics

There are several commonly used statistics like: *expectation*, *variance*, and *etc.*

- Expectation: The expectation value of every good obtain, expressed as the first-order origin moment
- Variance: The variance of the statistics, expressed as the second-order central moment

$$\begin{aligned} Expectation = \mathcal{E} &= \int x \cdot p(x)dx \\ Variance = \mathcal{V} &= \int (x - \mathcal{E})^2 \cdot p(x)dx \end{aligned} \tag{1}$$

**Lemma 1.1.** *For simplicity, the relationship between expectation and variance can be found as following*

$$\mathcal{V} = \mathcal{E}(X^2) - \mathcal{E}^2(X)$$

Other than one single statistic. The distribution of *two random variables* can be computed using *joint probability* and *conditional probability.*

$$p(x, y) = p(x) \cdot p(y|x) = p(y) \cdot p(x|y)$$

And the second-order moment of the two random variables is

$$\mathcal{E}(X, Y) = \iint x \cdot y \cdot p(x, y)dxdy$$

### 1.3.1 Independent Situation

The simplest situation is the variables of $x$ and $y$ are independent with each other.

**Lemma 1.2.** *If $X$ and $Y$ are independent, then*

$$Cov(X, Y) = \mathcal{E}(X, Y) - \mathcal{E}(X)\mathcal{E}(Y) = 0, \forall X \perp Y$$

### 1.3.2 Un-independent Situation

If the independent situation is not matched, then the covariance is not zero.

$$Cov(X,Y) = \mathcal{E}(X,Y) - \mathcal{E}(X)\mathcal{E}(Y) \neq 0$$

Moreover, in an extreme situation of $X = Y$, the covariance equals to the variance. The second-order moment can be expressed as

$$\begin{aligned}\mathcal{E}(X,Y) &= \iint x \cdot x \cdot p(x) \cdot dxdx \\ \mathcal{E}(X,Y) &= \mathcal{E}(X^2)\end{aligned}$$

where we use the fact of $p(x,y) = p(x)$ since we have $X = Y$ here. Using the definition of variance in (1), we have

$$\begin{aligned}Cov(X,Y) &= \mathcal{E}(X,X) - \mathcal{E}^2(X) \\ Cov(X,Y) &= \mathcal{E}(X^2) - \mathcal{E}^2(X) \\ Cov(X,Y) &= \mathcal{V}\end{aligned}$$

where $X = Y$.

### 1.3.3 Variance of mean value

Without the satisfying independence condition, it can be complicated. Commonly, the covariance matrix $\mathcal{C} \in \mathbb{R}^{ij}$ can be used to express the relationship between variables

$$\mathcal{C}_{ij} = Cov(X_i, X_j)$$

where $i$, $j$ are the indices of the two variables.

**Lemma 1.3.** *The variance of mean value equals to the mean of all covariance terms*

$$\mathcal{V}(\overline{X}) = \frac{1}{n^2}\sum_{i,j} Cov(X_i, X_j), i,j \in 1,2,\ldots,n$$

*no matter the relationship between the statistics of* $X_1, X_2, \ldots, X_n$.

Based on the lemma 1.3, the variance of mean value can be calculated based on the covariance between each two values. Under *independent situation*, since the covariance between each two variables equals to zero, the variance is the mean of diagonal elements of the covariance matrix. Under *un-independent situation*, the variance is the mean of all the elements of the covariance matrix.

## 1.4 Distributions

There are several commonly used distributions like: *Normal distribution*, *Binomial distribution*, *Chi-squared distribution*, *Student's t-distribution* and *etc.*

### 1.4.1 Common Distributions

**The PDF of Normal distribution is**

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{(x-\mu)^2}{2\sigma^2}, x \in (-\infty, \infty) \tag{2}$$

where $\mathcal{E} = \mu$ and $\mathcal{V} = \sigma^2$. The normal distribution is so important that we express it as $p(x) \sim \mathcal{N}(\mu, \sigma^2)$.

**The PDF of Binomial distribution is**

$$p_N(n) = (N, n) \cdot r^n \cdot (1-r)^{N-n}, n \in [0, N] \tag{3}$$

where $\mathcal{E} = N \cdot r$ and $\mathcal{V} = N \cdot r \cdot (1-r)$.

**The PDF of Chi-squared distribution is**

$$p_r(x) = \frac{x^{r/2-1} e^{-x/2}}{\Gamma(r/2) 2^{r/2}}, x \in (0, \infty) \tag{4}$$

where $\mathcal{E} = r$ and $\mathcal{V} = 2r$.

The statistic follows Chi-squared distribution refers

$$p_r(x) \sim \mathcal{X}^2(r) = \sum_{i=1}^{r} Y_i^2$$

where $Y_i \sim \mathcal{N}(0, 1)$, and $Y_i$s are independent with each other.

### 1.4.2 Gamma and Related Functions

**Gamma function** The function of generalized integral is defined as $\Gamma$ function,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \tag{5}$$

There are several useful properties,

$$\begin{aligned}
\Gamma(0) &= 1 \\
\Gamma(1/2) &= \sqrt{\pi} \\
\Gamma(z) &= z \cdot \Gamma(z-1) \\
\Gamma(n) &= n!, n \in 1, 2, 3, \ldots \\
\Gamma(z) &= 2 \cdot \int_0^\infty t^{2z-1} e^{-t^2} dt
\end{aligned} \tag{6}$$

The $\Gamma$ function can be used to calculate the integral functional of distributions.

**Beta Function** The $\Gamma$ function is also useful in compute binominal-like integral functions. The Beta function is defined as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{7}$$

## 1.5 Parameter Estimation

One goal of statistical analysis is to *determine the parameters of the distribution.* There are several methods of the estimation:

- MLE: Maximum Likelihood Estimation
- MAP: Maximum A posteriori Probability estimate

# A Proofs of Concepts Section

## A.1 Proof of Lemma1.1

Prove the relationship between expectation and variance can be found as following

$$\mathcal{V} = \mathcal{E}(X^2) - \mathcal{E}^2(X)$$

*Proof.* Compute the square in (1), we have

$$\begin{aligned}\mathcal{V} &= \int (x^2 - 2x\mathcal{E} + \mathcal{E}^2)p(x)dx \\ &= \mathcal{E}(X^2) - \mathcal{E}^2(X)\end{aligned}$$

where $\mathcal{E}$ refers $\mathcal{E}(X)$. And, the equation uses the condition that the $\mathcal{E}$ is constant in the integral. □

## A.2 Proof of Lemma1.2

Prove that, if the statistics of $X$ and $Y$ are independent, then

$$Cov(X, Y) = \mathcal{E}(X, Y) - \mathcal{E}(X)\mathcal{E}(Y) = 0, \forall X \perp Y$$

*Proof.* The independency guarantees

$$\begin{aligned}p(x|y) &= p(x) \\ p(y|x) &= p(y) \\ p(x, y) &= p(x) \cdot p(y)\end{aligned}$$

Using the definition of expectation in (1), we have $\mathcal{E}(X, Y) = \mathcal{E}(X) \cdot \mathcal{E}(Y)$. □

## A.3 Proof of Lemma1.3

Prove that, the variance of mean value equals to the mean of all covariance terms

$$\mathcal{V}(\overline{X}) = \frac{1}{n^2} \sum_{i,j} Cov(X_i, X_j), i, j \in 1, 2, \ldots, n$$

no matter the relationship between the statistics of $X_1, X_2, \ldots, X_n$.

*Proof.* The mean value of $n$ statistics can be expressed as

$$\overline{X} = \frac{1}{n} \sum_i X_i, i \in 1, 2, \ldots, n$$

Based on the definition, the variance can be expressed as

$$\begin{aligned}Var(\overline{X}) = &\int_{X_1, X_2, \ldots, X_n} \overline{x}^2 p(x_1, x_2, \ldots, x_n) dx_1 x_2 \ldots x_n \\ &- (\int_{X_1, X_2, \ldots, X_n} \overline{x} p(x_1, x_2, \ldots, x_n) dx_1 x_2 \ldots x_n)^2 \\ &, i, j \in 1, 2, \ldots, n\end{aligned}$$

Use the property of full probability rules, we have

$$\begin{aligned}
&\int_{X_k,\dots} f(x_k)\cdot p(x_k,\dots)dx_k\dots \\
&= \int_{X_k} f(x_k)\cdot p(x_k)dx_k \\
&\int_{X_i,X_j,\dots} f(x_i,x_j)\cdot p(x_i,x_j,\dots)dx_ix_j\dots \\
&= \int_{X_i,X_j} f(x_i,x_j)\cdot p(x_i,x_j)dx_ix_j
\end{aligned}$$

Thus the variance of the mean value can be formulated as

$$\begin{aligned}
n^2\cdot Var(\overline{X}) &= \sum_i \int x_i^2 p(x_i)dx_i + \sum_{i\neq j}\int x_ix_jp(x_i,x_j)dx_ix_j \\
&- (\sum_i \int x_ip(x_i)dx_i)^2
\end{aligned}$$

since the positive and negative terms are both of $n^2$ terms. By pairing them one-by-one, we come to

$$\begin{aligned}
n^2\cdot Var(\overline{X}) &= \sum_{i,j}\mathcal{E}(X_iX_j) - \mathcal{E}(X_i)\mathcal{E}(X_j) \\
Var(\overline{X}) &= \frac{1}{n^2}\sum_{i,j} Cov(X_i,X_j)
\end{aligned}$$

Hence proved. □