A GENTLE INTRODUCTION TO THE CENTRAL LIMIT THEOREM

BOHONG SU

Abstract. Normal distribution, also know Gaussian distribution, is one of the most well-studied distribution function in statistics. Its bell-shape curve is ubiquitous in many statistical reports, from the simple survey analysis to resource allocation. All of these consequences directly come form the central limit theorem, which we will study in this paper.

General Idea: Regardless of the population distribution model, as the sample size increases, the sample mean tends to be normally distributed around the population mean, and its standard deviation shrinks as n increases.

1. The Law of Large Numbers

When we perform statistics, the amount of data must be large, so that the frequency of events in the sample is close to the probability of the event itself. When the sample size tends to infinity, the frequency of random events tends to a stable value, and this stable value is the theoretical probability of occurrence of the event.

This is also the basis of big data. Internet companies often use big data as samples to predict user preferences. When the sample size is large enough, the user preferences of various age groups or men and women inferred from the sample can reflect the preferences of everyone. So if the sample is small, how much error is there in the result we calculated based on this?

Take the coin toss as an example. Suppose I tossed 10 times, 7 times facing up, and 3 times facing down. For these samples, I can estimate the probability of the coin's front and back. How to calculate it? Assuming that the probability of the head of the coin is a, the probability of the back is $1 - a$. Then the probability of occurrence of the above event

$$P = C_{10}^{7} a^{7} (1 - a)^{3}.$$

We ask for a such that the probability of the above event (7 upwards, 3 downwards) is the largest, that is, P is the largest. Similarly, if we toss a coin 1000 times, 620 times face up and 380 times face down. then

$$P = C_{1000}^{620} a^{620} (1 - a)^{380}.$$

It can be seen that the highest point of the curve for coin toss 10 times is approximately at $a = 0.7$, and the curve for coin toss 1000 times is approximately at $a = 0.62$, which is more consistent with our sample data.

In other words, the distribution parameter an estimated from the sample is the highest. This is also called maximum likelihood estimation (MLE). P is called the maximum likelihood function, and the value of a when the maximum likelihood function takes the maximum value is called the maximum likelihood estimation.

So what is the difference between 10 data and 1000 data?

It can be seen from the figure that when there are 10 data, the likelihood function is relatively flat, that is, the probabilities of various possibilities are not much worse. When there are 1000

Date: November 23, 2020.

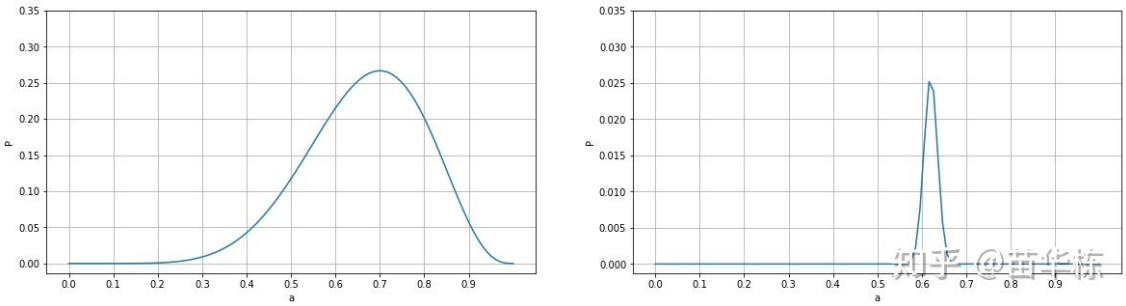

Figure 1. The one on the left is the plot with 10 throws; the one on the right is the plot with 1000 throws.

data, the likelihood function is relatively steep, that is, the probability is relatively concentrated, and the probability near the maximum likelihood estimation drops sharply.

This also gives an explanation mathematically. When we toss the coin 10 times and face up 7 times, we give the probability that the coin is up 0.7. However, when the probability of going up is 0.7, the probability that we toss up 7 times for 10 times is about 0.27 When the probability of a coin going up is 0.5, the probability that we toss 10 times and going up 7 times is about 0.12, and the probability is about half lower. Therefore, when we give the probability that the coin is facing upwards of 0.7, this result is unreliable, because many values near 0.7 may have the phenomenon of tossing 10 times and going upwards 7 times. We can say that the reliability of the result of 0.7 is relatively low.

Similarly, after we toss 1000 times, the probability that the coin will face up is 0.62, which is the highest, which is equal to about 0.025. If the probability of the coin facing upward is 0.6, the probability is about 0.01. If the probability of the coin going up is 0.5, it can be seen from the curve that the probability of going up 1000 times and 620 times is basically 0. That is to say, the data of 0.62 is still very reliable, even if it is bad, it is not much different. 0.6 is possible, but if the probability is 0.5, it is basically impossible.

If thrown infinitely, the peak of the curve will approach a constant value, and the shape is similar to a pulse, which is a vertical line. After an infinite number of trials, except for the pulse point with the highest probability, other values are unreliable, because the probability of other points will quickly become zero. In other words, the frequency measured by the sample is the same as the expectation of the event itself. This is also a graphical interpretation of the law of large numbers.

2. Central Limit Theorem

The normal distribution is very common in our lives. Why are there so many normal distributions in the real world? Before, let us introduce the principle of maximum entropy.

The entropy here:

- Thermodynamics is the degree of chaos of objects.
- The degree of uncertainty is used to describe information in information theory.

In thermodynamics, the principle of entropy increase is that the entropy of an isolated thermodynamic system does not decrease, but always increases or does not change. Used to give the evolution direction of an isolated system. It shows that an isolated system cannot develop towards a low-entropy state or become orderly.

For example, if you pour half a cup of hot water into a half cup of cold water, after a long time, the water in the cup must be evenly hot and cold. And if a cup of warm water is left there, it will not become hot on the top and cold on the bottom. Because the top and the bottom are orderly,

that is, the degree of chaos is low, that is, the entropy is low. An isolated system can only change in the direction of entropy, that is, mixing hot and cold water together.

In information theory, entropy is used to describe the degree of uncertainty of information, which is actually the degree of confusion of information. The principle of maximum entropy is to keep the uncertainty of the information to the maximum when the maximum entropy is satisfied under the given constraints, that is, the most reasonable estimation is made, and the cost is the least when the estimation is wrong. For example, if we roll a dice, when we don't know any information, we will predict that the probability of any point is $\frac{1}{6}$. Before we know it, we use the principle of maximum entropy, that is, when we don't know any information, equal probability is the best assumption. What is the best assumption is that the cost function is the smallest.

What is the cost function? For example, I roll the dice to gamble, and I lose the least money for the same amount of money for each point. Otherwise, if I throw more money at 1 point, the final result must be more losses than the average throw (assuming the dice are even). From the above description of entropy, you can roughly experience the feeling of entropy, which represents the degree of uncertainty and confusion.

The universe must be more and more chaotic, and finally silent. If nature does not interfere, it must eventually exist in the form of maximum entropy. After talking for a long time, what is the relationship between maximum entropy and the central limit theorem?

In real life, the distribution of various events is unknown, but generally the mean μ and variance σ are fixed. Then, under the premise that these two values are determined, what distribution is the largest? The answer is the *normal distribution*.

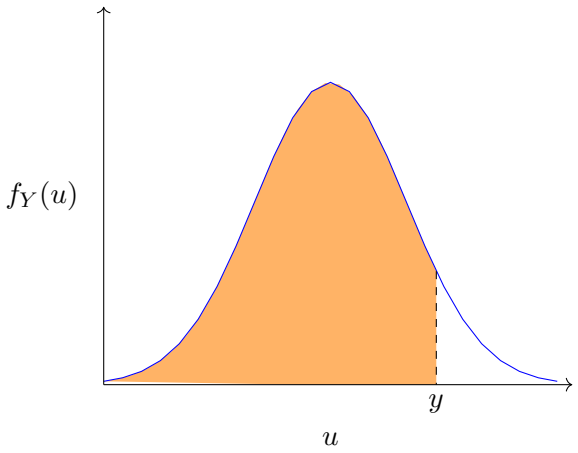

Figure 2. This is an illustration of the normal curve.

3. The Central Limit Theorem

Before we introduce the central limit theorem, we. will talk about the MGF.

Definition 3.1. (Moment Generating Function). The moment generating function M(t) of a random variable X is defined for all real values of t by

$$M(t) = E\left[e^{tX}\right] = \begin{cases} \sum_x e^{tX} p(x) & \text{if X is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tX} f(x)dx & \text{if X is continuous with density } f(x) \end{cases}$$

Theorem 3.2. *(The Central limit Theorem)*

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then the distribution of $\frac{X_1+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$ tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$

$$P\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \quad as \quad n \to \infty.$$

Proof. We begin the proof with the assumption that $\mu = 0, \sigma^2 = 1$ and that the MGF of the X_i exists and is finite. We already know the MGF of a normal random variable, but we still need to compute the MGF of the sequence of random variables we are interested in: $\sum_{i=0}^{n} X_i/\sqrt{n}$. By definition of MGF, we can see that $M\left(\frac{t}{\sqrt{n}}\right) = E\left[\exp\left(\frac{tX_i}{\sqrt{n}}\right)\right]$. However, we are interested in the MGF of $E\left[\exp\left(t\sum_{i=1}^{n}\frac{X_i}{\sqrt{n}}\right)\right]$. Here is how we find the MGF:

$$\begin{aligned}
E\left[\exp\left(t\sum_{i=1}^{n}\frac{X_i}{\sqrt{n}}\right)\right] &= E\left[\exp\left(\sum_{i=1}^{n} X_i \cdot \frac{t}{\sqrt{n}}\right)\right] \\
&= E\left[\prod_{i=1}^{n}\exp\left(X_i \cdot \frac{t}{\sqrt{n}}\right)\right] \\
&= \prod_{i=1}^{n} E\left[\exp\left(X_i \cdot \frac{t}{\sqrt{n}}\right)\right] \\
&= \prod_{i=1}^{n} M\left(\frac{t}{\sqrt{n}}\right) \\
&= \left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n.
\end{aligned}$$

Now we define $L(t) = \log \mathrm{M(t)}$ and evaluate $L(0), L'(0), L''(0)$

$$L(0) = \log M(0) = \log E\left[e^{0\cdot X_i}\right] = \log E[1] = \log 1 = 0.$$

$$L'(0) = \frac{M'(0)}{M(0)} = M'(0) = E\left[Xe^{0\cdot X_i}\right] = E[X] = \mu = 0.$$

$$L''(0) = \frac{M(0)M''(0) - [M'(0)]^2}{[M(0)]^2} = \frac{1 \cdot E\left[X^2\right] - 0^2}{1^2} = E\left[X^2\right] = \sigma^2 = 1.$$

Now we are ready to prove the Central limit theorem by showing that $\left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n \to e^{t^2/2}$ as $n \to \infty$. By taking the log of both sides, we can see that this is equivalent to showing $nL(t/\sqrt{n}) \to t^2/2$ as $n \to \infty$. Hence, we compute:

$$\begin{aligned}
\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \\
&= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})t}{-2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \\
&= \lim_{n\to\infty} L''\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{2} \\
&= \frac{t^2}{2}.
\end{aligned}$$

■

4. Application of the Central Limit Theorem in Baseball

Before we talk about the application of the CLT, let us introduce some terms related to the CLT.

Definition 4.1. A conjecture concerning one or more populations is known as a statistical hypothesis.

Definition 4.2. A null hypothesis is a hypothesis that we wish to test and is denoted H_0.

Definition 4.3. An alternative hypothesis represents the question to be answered in the hypothesis test and is denoted by H_1.

REMARK. The null hypothesis H_0 opposes the alternative hypothesis H_1.H_0 is commonly seen as the complement of H_1. Concerning our problem, the null hypothesis and the alternative hypothesis are:

$$H_0 : \text{There is no home-field advantage,}$$
$$H_1 : \text{There is a home-field advantage.}$$

When we do a hypothesis test, the goal is to determine if we will reject the null hypothesis or if we fail to reject the null hypothesis. If we reject H_0, we are in favour of H_1 because of sufficient evidence in the data. If we fail to reject H_0, then we have insufficient evidence in the data.

Definition 4.4. A test statistic is a sample that is used to determine whether or not a hypothesis is rejected or not.

Definition 4.5. A critical value is a cut off value that is compared to the test statistic to determine whether or not the null hypothesis is rejected.

Definition 4.6. The level of significance of a test statistic is the probability that H_0 is rejected, although it is true.

Definition 4.7. A z-score or z-value is a number that indicates how many standard deviations an element is away from the mean.

Definition 4.8. A confidence interval is an interval that contains an estimated range of values in which an unknown population parameter is likely to fall into.

Definition 4.9. A p-value is the lowest level of significance in which the test statistic is significant.

Now let us talk about the example problem: is there such thing as a home-field advantage? How can we test this notion? In the 2013 Major League Baseball season, there were 2431 games played, and of those games, 1308 of them were won at home. This indicates that approximately 53.81% of the games played were won at home. We will let our observed value be this value, so $\hat{p} = 0.5381$. It seems as though there is such thing as a home-field advantage, but we must test this notion to be certain. To do this, we will test the hypothesis that there is no such thing as a home-field advantage, so our null hypothesis will be

$$H_0 : p = 0.50.$$

That is, 50% of the Major League Baseball games are won at home, hence, there is no home-field advantage. Our alternative hypothesis will be $H_1 : p > 0.50$. If there is no home-field advantage, then we would expect our proportion to be 0.50 , since half of the games would be won at home and the other half on the road.

Before we begin to compute if there is such thing as a home-field advantage we must first satisfy four conditions; independence assumption, random condition, 10% condition, and the success/failure condition. These conditions will assure that we can test our hypothesis.

Each game is independent of one another and one game does not effect how another game is played. Although in some cases when a key batter or pitcher is injured, the team may not do as well in the immediate upcoming games, but roughly speaking, the games played are generally independent of one another, and so our independence condition holds.

Since there have been many games played over the years, each year having roughly 2430 games, it can be seen that taking just one year to observe the data will account for our randomization condition.

Also, as stated above, we can see that the 2431 games played in the 2013 season, are less than 10% of the total games played over the years that Major League Baseball has been around, so our 10% condition also holds true, that is, the sample size is no more than 10% of the population.

Finally we must check that the number of games multiplied by our proportion of 0.50 , is larger than 10. So we have

$$np = 2431(0.50) = 1215.5.$$

which is larger than 10, so our success / failure condition holds as well. since all of these conditions are met, we are now able to use the Normal Distribution model to help us test our hypothesis. We will test our hypothesis using two different methods: the first by using a confidence interval, and the second using a p-value. First, we will test our hypothesis using a confidence interval. For testing

$$H_0 : p = 0.50 \text{ vs. } H_1 : p > 0.50.$$

at the 0.05 level of significance, we may construct a right-sided 95% confidence interval for p. If our test statistic of $p = 0.50$ is in the interval, then we fail to reject H_0 at the 0.05 level of significance. If $p = 0.50$ is not in the interval, we reject H_0. The right-sided $100(1-\alpha)\%$ confidence interval for p for a large sample is given by

$$\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p \leq 1.$$

where α is the level of significance. since $n = 2431, \hat{p} = 0.5381$, and $\alpha = 0.05$, we see from the Normal Distribution table in the Appendix that $z_{0.05} = 1.645$. So a right-sided 95% confidence interval for p is

$$0.5381 - 1.645\sqrt{\frac{(0.5381)(1-0.5381)}{2431}} < p \leq 1.$$
$$0.5381 - 1.645(0.0101114) < p \leq 1.$$
$$0.5215 < p \leq 1.$$

since $0.50 \notin (0.5215, 1]$, we reject $H_0 : p = 0.50$ in favour of $H_1 : p > 0.50$ at the 0.05 level of significance, that is, we have enough evidence to support that there is a home-field advantage, and the home team wins more than 50% of the games played at home.

Now we will use the p-value approach to test our hypothesis. We must find the z-value for testing our observed value. We use the following equation to do so;

$$z = \frac{(\hat{p} - p_o)}{\sqrt{\frac{poq_o}{n}}}.$$

Now, with $p = 0.50, \hat{p} = 0.5381$, and $n = 2431$, we have

$$z = \frac{(\hat{p} - p_o)}{\sqrt{\frac{poq_o}{n}}} = \frac{0.5381 - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{2431}}} = \frac{0.0381}{0.010140923} = 3.76.$$

This results in a p-value < 0.0001. So we can conclude, since the p-value < 0.0001 is less than 0.05, we reject H_0. That is, the data seems to support that the home field team wins more than 50% of the time, and hence there is such thing as a home-field advantage in Major League Baseball.

We have shown that taking all of the games played in the 2013 Major League Baseball season, that there is a home-field advantage, but is there a difference between the American League and

the National League? Do both leagues have a home-field advantage? We will test this notion using a $100(1-\alpha)\%$ confidence interval at the 0.01 level of significance. This will allow us to be 99% confident of our results.

In the 2013 season, the National League played 1211 games, and won 660 of those games at home. So this indicates that approximately 54.5% of the games were won at home. As we calculated above, we will let the observed value be $\hat{p} = 0.545$ and we will test the same hypothesis, that is,

$$H_0 : p = 0.50 \text{ vs. } H_1 : p > 0.50.$$

since $n = 1211, \hat{p} = 0.545$ and $\alpha = 0.01$, we can see from the Normal Distribution table in the Appendix that $z_{0.01} = 2.33$. So a right-sided 99% confidence interval for p is

$$\begin{gathered} 0.545 - 2.33\sqrt{\tfrac{(0.545)(1-0.545)}{1211}} < p \leq 1. \\ 0.545 - 2.33(0.014309744) < p \leq 1. \\ 0.5117 < p \leq 1. \end{gathered}$$

since $0.50 \notin (0.5117, 1]$, we reject $H_0 : p = 0.50$ in favour of $H_1 : p > 0.50$. So we can conclude that the National League has a home-field advantage. Will the same be true for the American League? We will again test the same hypothesis, using a 99% confidence interval for the American League.

In the 2013 season, the American League played slightly more games than the National League. They played 1220 games and of those games, 648 of them were won at home. So this indicates that approximately 53.11% of the games played were won at home. Once again, let our observed value be $\hat{p} = 0.5311$, and testing the same hypothesis above, we see that a 99% confidence interval for p is

$$\begin{gathered} 0.5311 - 2.33\sqrt{\tfrac{(0.5311)(1-0.5311)}{1220}} < p \leq 1. \\ 0.5311 - 2.33(0.01428724) < p \leq 1. \\ 0.4978 < p \leq 1. \end{gathered}$$

since $0.50 \in (0.4978, 1]$, we fail to reject $H_0 : p = 0.50$. That is, we do not have enough evidence to support that there is a home-field advantage in the American League.

We can see that by testing these hypotheses for the National League and the American League, that we can confidently state that there is a home-field advantage in the National League, but we cannot say the same thing for the American League based on the 2013 Major League Baseball season.

5. Standard Normal Distribution Table.

It might sounds old fashion, but this is very useful when you don't have a computer with you. The following table gives the proportion of the standard normal distribution to the left of a z-score.

	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

References

[And14] Nicole Anderson, 2014. URL: https://www.lakeheadu.ca/sites/default/files/uploads/77/docs/Anderson_Project.pdf.

[Gou11] Kevin Goulding. Tikz diagrams for economists: A normal pdf with shaded area., Jun 2011. URL: https://thetarzan.wordpress.com/2011/06/17/tikz-diagrams-for-economists-a-normal-pdf-with-shaded-area/.

[Hua20] Miao Huadong. How to understand the law of large numbers intuitively, Feb 2020. URL: https://zhuanlan.zhihu.com/p/104687806.

[Kro11] Vlad Krokhmal. Introductory probability and the central limit theorem. page 10–11, Jul 2011.

[Mia20] Huadong Miao, Feb 2020. URL: https://zhuanlan.zhihu.com/p/104687806.