EMS702P Statistical Thinking and Applied Machine Learning

Week 2.1 – Sampling Distributions & Estimation Statistics

Jun Chen

Introduction to Statistics and Machine Learning

Edition: v1.1

Table of Contents

1 Sampling Distributions

Considering samples from a *distribution* enables us to obtain information about a population where we cannot, for reasons of practicality, economy, or both, inspect the whole of the population. For example, it is impossible to check the complete output of some manufacturing processes. In addition, testing is sometimes destructive - one would not wish to destroy the whole production of a given component. Hence we have to deal with samples taken from a population and estimate those population **parameters** that we need.

1.1 The Law of Large Numbers

We have an intuition that more observations are better. This is the same intuition behind the idea that if we collect more data, our sample of data will be more representative of the problem domain (population). There is a theorem in statistics and probability that supports this intuition, which is called **the law of large numbers**.

The law of large numbers is a theorem from probability and statistics suggesting that the average result from repeating an experiment multiple times will better approximate the true or expected underlying result. It is important to make sure the trial of the experiment is run in an *identical* manner and *does not depend on* the results of any other trial (i.e. *iid*). This is to ensure that the samples are indeed drawn from the same underlying population distribution.

Using the terms from statistics, we can say that as the size of the sample increases, the mean value of the sample will better approximate the mean or expected value in the population. As the sample size goes to infinity, the *sample mean* will converge to the *population mean*.

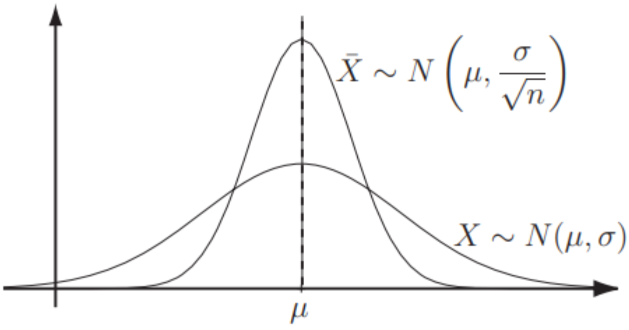
The law of large numbers helps us understand why we cannot trust a single observation from an experiment in isolation. The law reminds us to repeat the experiment in order to develop a large and representative sample of observations before we start making inferences about what the result means. It is why we must be sceptical of inferences from small sample sizes, called small n.

1.2 The Central Limit Theorem

The Central Limit Theorem, or *CLT* for short, is an important finding and pillar in the fields of statistics and probability. The theorem states that as the size of the sample increases, the distribution of the mean across *multiple* samples will approximate a Gaussian distribution. The central limit theorem describes the *shape* or the distribution of sample means as a Gaussian, which is a distribution that statistics knows a lot about.

If we calculate the mean of a sample, it will be an estimate of the mean of the population distribution. But, like any estimate, it will be wrong and will contain some error. If we draw multiple independent samples of size n, where n is *large* enough, and calculate their means $\bar{X}$, the distribution of those means will form a Gaussian distribution, $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. As $\frac{\sigma}{\sqrt{n}}$ indicates the uncertainty in the process of predicting the underlying population mean from the mean of a sample or samples, it is often called the **standard error of the mean** and denoted as σ_n.

It is important that each trial (event) that results in an observation be *iid*, which is to ensure that the sample is drawing from the same underlying population distribution. However, the underlying population **need not be normal**. In the case where the underlying distribution is normal, even n is small, the theorem still holds, which can be illustrated as below.



Example 1

Two-centimetre number 10 woodscrews are manufactured in their millions but packed in boxes of 200 to be sold to the public or trade. If the length of the screws is known to be normally distributed with a mean of $2\ cm$ and variance $0.05\ cm^2$, find the mean and standard deviation of the sample mean of 200 boxed screws. What is the probability that the sample mean length of the screws in a box of 200 is greater than $2.02\ cm$?

Solution

1.3 Implications in Machine Learning

The law of large numbers has the following important implications in applied machine learning.

- **Training data** must be representative of the observations from the domain. This means that it must contain enough information to generalize to the true unknown and underlying distribution of the population.

  Keep this in mind during data collection, data cleaning, and data preparation, you may choose to exclude sections of the underlying population by setting hard limits on observed values (e.g. for outliers) where you expect data to be too sparse to model effectively.
- The above implication in training data applies to test data as well.
- The theorem highlights the need to develop a sample of **multiple independent (or close to independent) evaluations** of a given model such that the mean reported skill from the sample is an accurate enough estimate of the population mean. It provides a defence for not simply reporting or proceeding with a model based on *a* skill score from *a* single train/test evaluation.

The central limit theorem has the following important implications in applied machine learning.

- **Significance tests** (see Parametric Hypothesis Test in **Week** 2.2) are tools based on the theorem that estimate the likelihood that the two samples of model skill scores were drawn from the same or a different unknown underlying distribution of model skill scores. If it looks like (implying that the *shapes* of the sample distributions need to be considered) the samples were drawn from the same population, then no difference between the models skill is assumed, and any actual differences are due to statistical noise.
- We can develop multiple independent (or close to independent) evaluations of a model accuracy to result in a population of candidate skill

estimates. The mean of these skill estimates will be an estimate (with error) of the true underlying estimate of the model skill on the problem. With the central limit theorem, we can use knowledge of the Gaussian distribution to estimate the likelihood of the sample mean based on the sample size and calculate an interval of desired confidence (i.e. how skilful the model is expected to be in practice) around the skill of the model.

2 Introduction to Estimation Statistics

Estimation statistics are a group of methods to quantify the magnitude of effects and the amount of uncertainty for estimated values. While statistical hypothesis tests (see **Week** 2.2) talk about whether the samples come from the same distribution or not, estimation statistics can describe the size and confidence of the difference. This allows us to comment on how different one method is from another. Among many estimation statistics, we will focus on two basic methods: **point estimation** and **interval estimation**.

2.1 Point Estimation vs. Interval Estimation

The essential difference between the two is that point estimation gives single numbers which, in the sense defined below, are the *best estimates* of population parameters, while interval estimates give a range of values together with a figure called the *confidence* that the true value of a parameter lies within the calculated range. Such ranges are usually called confidence intervals.

In statistics, the word 'estimate' does not mean a guess. It is an agreed precise *procedures* used to find the required values which are 'best values' as they should be:

- **consistent** in the sense that the difference between the true value and the estimate approaches zero as the sample size used to do the calculation increases;

- **unbiased** in the sense that the expected value of the estimator is equal to the true value;
- **efficient** in the sense that the variance of the estimator is small.

The *procedures* to find the 'best values' are called the '**estimator**'. It is not always possible to find a 'best' estimator. You might have to decide (for example) between one which is *consistent, biased and efficient* and one which is *consistent, unbiased and inefficient* (see examples in **Sections** 2.2 and 2.3).

2.2 Estimating the Mean

	Population	**Sample**	**Estimator**
Size	N	n	
Mean	μ or $\mathrm{E}(x)$	$\bar{x}$	$\hat{\mu}$ for μ

It is sensible to use the following as the estimator of the population mean since the difference between the population mean and the sample mean disappears with increasing sample size (i.e. consistent) according to *the law of large numbers* and *the central limit theorem*.

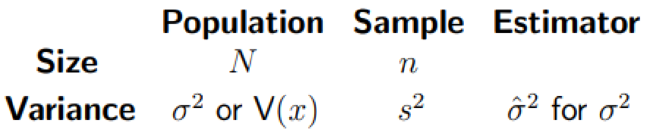
$$\hat{\mu} = \bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

Example 2

Prove the above estimator for the population mean is unbiased.

Solution

2.3 Estimating the Variance



	Population	Sample	Estimator
Size	N	n	
Variance	σ^2 or V(x)	s^2	$\hat{\sigma}^2$ for σ^2

It is *intuitive* to use the following as the estimator of the population variance since the difference between the population variance and the estimated variance disappears with increasing sample size (i.e. consistent) according to *the law of large numbers* and *the central limit theorem*.

$$\hat{\sigma}^2 = \frac{\sum(x-\hat{\mu})^2}{n}$$

Example 3

Prove the above estimator for the population variance is **biased**.

Solution

This result is **biased**, for an unbiased estimator the result should be σ^2 not $\frac{n-1}{n}\sigma^2$. The Bessel's correction is used as the remedy in this case as below.

$$\hat{\sigma}^2 = \frac{n}{n-1}\frac{\sum(x-\hat{\mu})^2}{n} = \frac{\sum(x-\hat{\mu})^2}{n-1}$$

Two points to note here:

- You should not take samples of size 1 since the variance cannot be estimated from such samples.
- This is also the reason why we used $n-1$ to calculate the sample variance in **Week** 1.3.

3 Interval Estimation

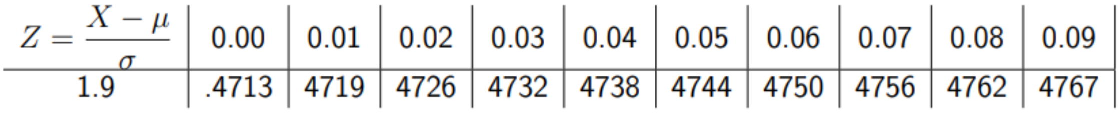
Much of machine learning involves estimating the performance of a machine learning algorithm on unseen data. **Confidence intervals** are a way of quantifying the uncertainty of an estimate. They can be used to add a bounds or likelihood on a population parameter, such as a mean, estimated from a sample of independent observations from the population. A confidence interval is different from a *tolerance interval* that describes the bounds of data sampled from the distribution. It is also different from a *prediction interval* that describes the bounds on a single observation. The latter two are not covered in this lecture but can be found in further readings.

Confidence intervals could be used in presenting the skill of a predictive model, e.g. *the accuracy of the model is* $a +/- b$ *at the* 95% *confidence level*. The value of a confidence interval is its ability to quantify the uncertainty of the estimate. It provides both a lower and upper bound and a likelihood.

3.1 Interval Estimation for the mean

The sample mean $\bar{x} = \frac{x_1+x_2+\cdots x_n}{n}$ is a good estimator of the population mean μ. However, $\bar{x}$ is unlikely to be exactly equal to μ given a sample of size n. It is sensible that we construct an interval around $\bar{x}$ in such a way that we can quantify the confidence that the interval actually contains the population mean μ.

From the *central limit theorem*, we have $\bar{x}$ follows a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ if n is *large*. Looking at the following extract from the normal probability tables,



$Z = \frac{X-\mu}{\sigma}$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.9	.4713	4719	4726	4732	4738	4744	4750	4756	4762	4767

Therefore, we have

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

If we know the population, we say with 95% confidence that

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

which is equivalent to

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

This interval is called a 95% confidence interval for the mean μ. Since we often don't know the population variance in practice, we use the best estimate of the population variance from a sample of size n given in **Section** 2.3.

$$\bar{x} - 1.96\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}$$

Note that while the 95% level is very commonly used. However, we can go through the same procedure to demand we need to be 99% certain that μ lies within the confidence interval developed and we obtain the interval below.

$$\bar{x} - 2.58\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.58\frac{\hat{\sigma}}{\sqrt{n}}$$

Example 4

After 1000 hours of use, the weight loss, in gm, due to wear in certain rollers in machines, is normally distributed with mean μ and variance σ^2 . Fifty independent observations are taken. (This may be regarded as a "large" sample.) If observation i is y_i , then $\sum_{i=1}^{50} y_i = 497.2$ and $\sum_{i=1}^{50} {y_i}^2 = 5473.58$.

Estimate μ and σ^2 and give a 95% confidence interval for μ.

Solution

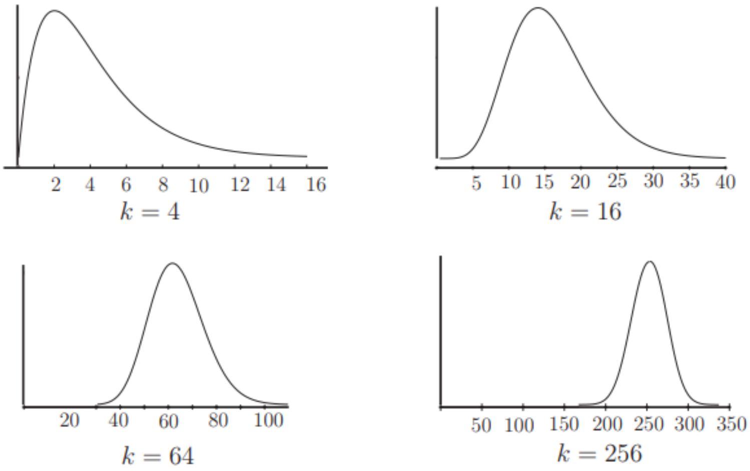
3.2 Interval Estimation for the Variance

Just as the sample means form a distribution, the **sample variances** also form a distribution, called the **chi-squared** (usually written as χ_k^2) distribution.

If $x_1, x_2, \cdots, x_n$ is a random sample taken from a normal population with mean μ and variance σ^2, then if the sample variance is denoted by S^2, the random variable

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_k^2$$

Where, $k = n - 1$ is the degrees of freedom of the sample. The mathematical formulation of the chi-squared distribution is a little involved, hence omitted here. However, the following graphs of the chi-squared distribution can be quite Intuitive for you to understand the essential characteristics of the distribution.

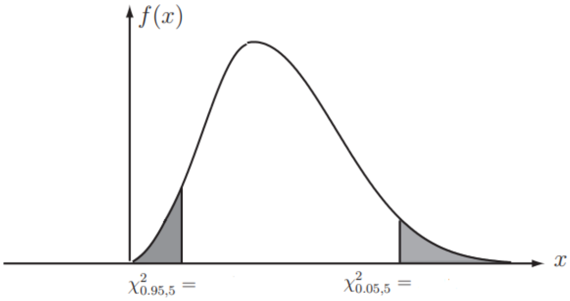


As k increases, the peak of each curve occurs at values closer to k. As k increases, the shape of the curve appears to become more symmetrical. As $k \rightarrow \infty$ the χ_k^2 distribution becomes normal. One further fact, not obvious from the above graphs, is that the **mean** and **variance** of the χ_k^2 distribution are k and $2k$ respectively.

Similar to the normal distribution, the values of the chi-squared distribution are tabulated for ease of use. The table in the Appendix shows the values of $\chi_{\alpha,\upsilon}^2$ for a variety of values of the area under the curve α (i.e. the probability) and the number of degrees of freedom $\upsilon = k$. Notice that the table gives the area values corresponding to the right-hand tail of the distribution. For the left-hand area values we need to use $1 - \alpha$.

Example 5

Find the corresponding values of $\chi_{0.95,5}^2$ and $\chi_{0.05,5}^2$.



Solution

We can now construct a confidence interval for the sample variance. If the level of confidence is 95%, which means that we need two 2.5% tails. Therefore, we have:

$$P\left(\chi^2_{0.975,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{0.025,n-1}\right) = 0.95$$

Hence

$$\frac{1}{\chi^2_{0.025,n-1}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{0.975,n-1}}$$

Therefore, $\frac{(n-1)S^2}{\chi^2_{0.025,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{0.975,n-1}}$ is the confidence interval for the sample variance. The corresponding confidence interval for the standard deviation is found by taking square roots. In general, we have the following for any confidence level that we are interested in.

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{(1-\alpha)/2,n-1}}$$

Example 6

A random sample of 20 nominally measured $2mm$ diameter steel ball bearings is taken and the diameters are measured precisely. The measurements, in $mm,$ are as follows:

2.02 1.94 2.09 1.95 1.98 2.00 2.03 2.04 2.08 2.07

1.99 1.96 1.99 1.95 1.99 1.99 2.03 2.05 2.01 2.03

Assuming that the diameters are normally distributed with unknown mean, μ, and unknown variance σ^2.

a) find a two-sided 95% confidence interval for the variance, σ^2;

b) find a two-sided confidence interval for the standard deviation, σ.

Solution

As a final note on the confidence intervals for the sample mean and variance, *often*, the larger the sample from which the estimate was drawn, the more precise the estimate and the smaller (better) the confidence interval.

- **Smaller Confidence Interval:** A more precise estimate.
- **Larger Confidence Interval:** A less precise estimate.

4 Further Readings

[1] HELM Workbook 40 Sampling Distributions and Estimation

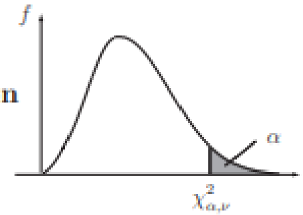
https://nucinkis-lab.cc.ic.ac.uk/HELM/HELM_Workbooks_36-40/WB40-all.pdf

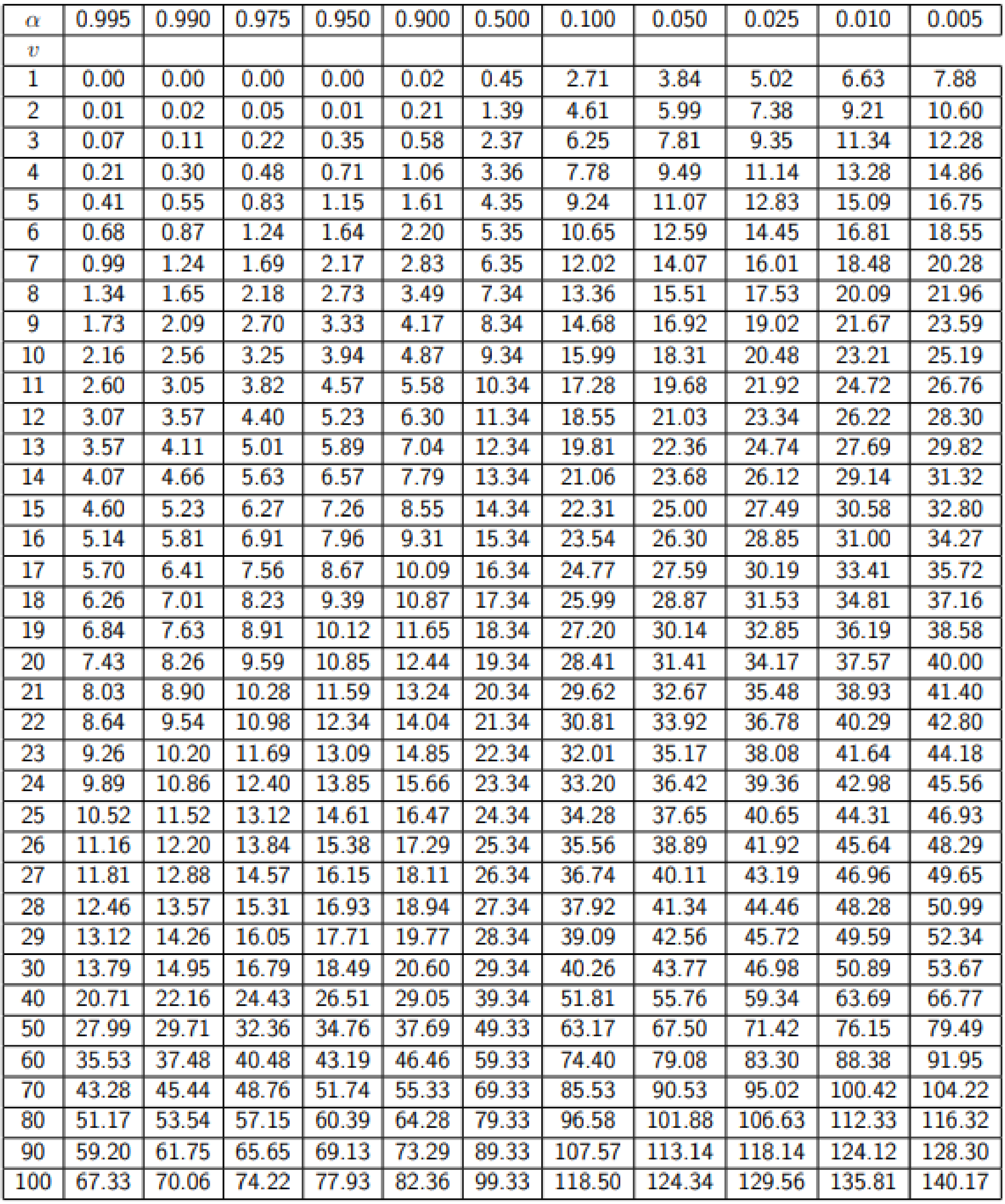
[2] The distinction between confidence intervals, prediction intervals and tolerance intervals

https://www.graphpad.com/support/faq/the-distinction-between-confidence-intervals-prediction-intervals-and-tolerance-intervals/

5 Appendix

Percentage Points $\chi^2_{\alpha,\nu}$ of the χ^2 distribution





α	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
v											
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.01	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.28
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	31.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17