On the application of t-distribution in the estimation of population mean interval

Bin Song

School of Finance and economics, Harbin Vocational College of science and technology, Harbin, China.

* Corresponding author: 875458215@qq.com

Abstract. This article introduces the origin of the t distribution, the demonstration of the t distribution on small samples, the advantages of the t distribution, and the reasoning of the application of the t distribution to the estimation of the population mean. The specific application of the t distribution is given through a case, and the difference between the t distribution and the normal distribution is compared.

Keywords: t distribution; Overall mean; Interval estimation; application

1 The origin of the t distribution

More than 100 years ago, Gossett proposed a method to correct the normal distribution on the basis of trial and error, which is the t distribution. As a representative of small sample distribution, t distribution improves the test conditions of some experiments, and promotes the development of statistics and quantitative economics to a large extent. [1]

2 Arguments for the t distribution

2.1 The era of normal distribution

Before Gossett proposed the t distribution, people used the normal distribution to solve problems. For a sample, if the amount of data is large enough, for the sake of convenience, people will standardize this group of samples, that is, u transformation.This transformation includes a major premise, that is, when the sample size is large enough, its standard deviation s can represent the overall standard deviation. However, many fields do not allow large-scale experiments, and sometimes repeated experiments cannot be carried out, so large samples The distribution appears weak. [2]

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_54

2.2 Small Sample Era

Gossett explored the small sample problem, which was inspired by the "moment" idea in the Pearson curve distribution theoretical system, and used ingenious conversion to overcome the corresponding problem between the sample standard deviation and the overall standard deviation. [3]The general roadmap for establishing the small sample theory is as follows:

That is, if the population obeys a normal distribution $N(0,\sigma^2)$, the sample is,

$$x_1, x_2, \cdots, x_n, \quad s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1},$$

Then $\frac{\bar{x}\sqrt{n}}{s}$ obey the t-distribution with n degrees of freedom, denoted as $\frac{\bar{x}\sqrt{n}}{s} \sim t_n$.

These situations boil down to comparisons of deviations between the true and estimated values of the variables present in a normal population. Similarly, an estimate of the standard deviation of a sample whose population is an independent normal distribution can be treated as the standard deviation from a normal distribution homogeneous to its population. For the calculation of this standard deviation, it can come from the true mean of the test data distribution, or from the mean of the sample randomly drawn. [4]

2.3 Advantages of the t distribution

The t distribution reasonably solves the problem that the sample size is not large enough in the use of the overall distribution. It directly uses the sample standard deviation as the estimation of the overall standard deviation, and in terms of distribution form, because it has one more degree of freedom parameter than the normal distribution, in the sample When the sample size is small, the bell curve is lower in the middle and higher on both sides, and as the sample size increases, the t distribution gradually approaches the standard normal distribution. [5]

The t-distribution embodies the tolerance of small sample outliers, and the t-distribution formula is applicable to all situations where the population is normally distributed, and can simplify the comparison process of the deviation between the true value of the normally distributed variable and the estimated value. [6]

2.4 Inference of the t distribution applied to the estimation of the population mean

Suppose $x_1, x_2, \cdots, x_n$ it is a sample from a normal population $N(\mu, \sigma^2)$, i $x_1, x_2, \cdots, x_n$ ndependent of each other, $\bar{x}$ is the mean of the sample, and is the standard deviation of the sample, then

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{\sigma}{s} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} / \sqrt{\frac{s^2}{\sigma^2}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} / \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)\sigma^2}}$$

set up

$$X = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$Y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{\sigma})^2 = \sum_{i=1}^{n} \left[\frac{(x_i - \mu) - (\bar{x} - \mu)}{\sigma} \right]^2 = \sum_{i=1}^{n} (\frac{x_i - \mu}{\sigma})^2 - n(\frac{\bar{x} - \mu}{\sigma})^2$$

$$= \sum_{i=1}^{n} (\frac{x_i - \mu}{\sigma})^2 - (\frac{\bar{x} - \mu}{\sigma / \sqrt{n}})^2$$

$$\frac{x_i - \mu}{\sigma} \sim N(0,1), \quad \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \Rightarrow Y \sim \chi^2(n-1)$$

Since X and are independent of each other, it can be seen from the definition of the distribution

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{X}{\sqrt{\frac{Y}{n-1}}} \sim t(n-1)$$

Given the confidence level, the use of the distribution t for the population mean interval estimation is required

$$p(\mu_1 < \mu < \mu_2) = 1 - \alpha$$

$$p(-t_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2}}) = 1 - \alpha$$

According to the t-distribution theory

$$\bar{x}\text{-}t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}<\mu<\bar{x}+t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

3 Application Comparison of t Distribution And Normal Distribution

As mentioned earlier, as the degrees of freedom increase, the t-distribution tends to the standard normal distribution. So for large samples, it is useful to compare the difference between means estimation using the normal distribution and the t-distribution.[7]

Question 1: A survey company surveyed the water use of households in a city, and the results of a sample of 36 households showed that the average water consumption per household was 4.9 barrels, the standard deviation was 3.5 barrels, and if the confidence level was 95%, the confidence interval estimated the average monthly water consumption of residents in the city.[8]

Analysis: The overall standard deviation of this question is unknown, calculated as

$$\bar{x}\pm t_{\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}=4.9\pm2.030\cdot\frac{3.5}{\sqrt{36}}\approx4.9\pm1.18=(3.72,\ 6.08).$$

Because n = 36 is a large sample, if calculated as a normal distribution

$$\bar{x}\pm z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}=4.9\pm1.96\cdot\frac{3.5}{\sqrt{36}}\approx4.9\pm1.14=(3.76,\ 6.04).$$

Question 2: A survey company conducted a survey on the water use of households in a city, and the results of a sample of 100 households showed that the average water consumption per household was 4.9 barrels, the standard deviation was 3.5 barrels, and if the confidence level was 95%, the confidence interval estimated the average monthly water consumption of the city's residents.[9]

Analysis: Calculated as a t-distribution

$$\bar{x}\pm t_{\frac{\alpha}{2}}(n-1)\cdot\frac{s}{\sqrt{n}}=4.9\pm1.984\cdot\frac{3.5}{\sqrt{100}}=4.9\pm0.6944=(4.2056,\ 5.5944)$$

Because n=100 is a large sample, if calculated as a normal distribution

$$\bar{x}\pm z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}=4.9\pm1.96\cdot\frac{3.5}{\sqrt{100}}=4.9\pm0.686=(4.214,\ 5.586)$$

Question 3: A survey company conducted a survey on the water use of households in a city, and the results of a sample of 1001 households showed that the average water consumption per household was 4.9 barrels, the standard deviation was 3.5 barrels, and if the confidence level was 95%, the confidence interval estimated the average monthly water consumption of residents in the city.

Analysis: Population variance unknown, calculated as t-distribution

$$\bar{x} \pm t_{\frac{\alpha}{2}}(n-1) \cdot \frac{s}{\sqrt{n}} = 4.9 \pm 1.962 \cdot \frac{3.5}{\sqrt{1001}} \approx 4.9 \pm 0.022 = (4.878,\ 4.922)$$

Because, n=1001 is a large sample if calculated as a normal distribution

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 4.9 \pm 1.96 \cdot \frac{3.5}{\sqrt{1001}} \approx 4.9 \pm 0.022 = (4.878,\ 4.922)$$

It can be seen from the above three examples that other conditions remain unchanged, with the increase of sample size, the calculation results of the population mean interval calculated by normal distribution theory and t-distribution theory are closer and closer, and the calculation results of the two are almost exactly the same when the sample size exceeds 1000.[10]

4 conclusion

When using the sample mean to estimate the overall mean, no matter it is a large sample or a small sample, it can be estimated using the t distribution. The t distribution theory can better reflect the difference in the number of samples, and has the characteristics of unity, accuracy and convenience. Therefore, in a certain sense, the t distribution is an alternative to the normal distribution.

References

1. SONG Lixin. Probability Theory and Mathematical Statistics[M].Beijing:People's Education Press,2003.
2. Liu Hongwei. Principles of Statistics[M].Beijing:Communication University of China Press,2008.
3. LIU Ze. Fundamentals of Statistics[M].Beijing:People's Posts and Telecommunications Press,2017.
4. Wang Xiaoling. Education Statistics[M].Wuhan:Central China Normal University Press,2001.
5. Xiao Zhanfeng. Fundamentals of Statistics[M].Chengdu:Southwestern University of Finance and Economics Press,2009.
   ZHAO Hairong. Fundamentals of Statistics[M].Beijing:Education Science Press,2020.
6. ZHANG Yonglin. Fundamentals of Statistics[M].Qingdao:Ocean University of China Press,2015.
7. Zhou Xuanda. Probability Theory and Mathematical Statistics Learning Guide[M].Beijing:Chinese Minmin University Press,2012.)
8. YUAN Wei. Statistics[M].Beijing:Higher Education Press,2005.
9. Yan Yan. Empirical Research Based on t-Distribution[J], Systems Engineering-Theory & Practice, 2011.
10. ZHANG Qin. Biostatistics[M].Beijing:China Agricultural University Press,2009.