Open Journal of Statistics, 2012, 2, 447-451
http://dx.doi.org/10.4236/ojs.2012.24056 Published Online October 2012 (http://www.SciRP.org/journal/ojs)

Study on the Normal and Skewed Distribution of Isometric Grouping

Zhensheng Jia[1], Wenkai Jia[2]
[1]School of Economics and Management, Chongqing Three Gorges University, Chongqing, China
[2]Students Affairs Department, Chongqing Three Gorges University, Chongqing, China
Email: klx716@sina.com

Received July 15, 2012; revised August 20, 2012; accepted August 30, 2012

ABSTRACT

Because of thinking only the number of numbers but not fitting function, it would be adequate to take further a field when calculating group numbers with empirical formula. We have proved the three theorems based on studying the normal distribution, and then reach the conclusion that there is a better method to do the same work. The method is simpler and more practical than empirical method and also works well with any skewed distribution.

Keywords: Mean Value; Variance; Fit Function; Range; One-Quarter Range

1. Introduction

In various books on statistics, when discussing the isometric group, the empirical formula is treated as only a reference or put aside simply. It is attributed that the class interval is only relevant to the number of numbers not the shape of fit function in the empirical formula.

This paper analyzes the isometric group while conforming to the normal distribution of series, and derives simple and practical method to find class interval. Furthermore, the same formula also works well with any skewed distribution.

2. Histogram and the Upper Bound of Class Interval

2.1. Find Out Theorem 1

1) Observe $x_1, x_2, \cdots, x_N$, and find out the minimum value x_1^* and maximum value x_N^*. And select proper c that slightly less than x_1^* and d that slightly greater than x_N^*, and divide (c, d) into $2l+1$ intervals.

Each interval has the same length as $\frac{d-c}{2l+1}$, let $a=\frac{d-c}{2l+1}$, it is easy to see $(2l+1)a > R$.

R is range, normally $l \geq 2$, and then divide into five or more groups.

Denote $d-c$ as $R^* = d-c = (2l+1)a$, we call R^* as closed range [1,2].

The series is bell-shaped distributed in a symmetrical way. Set the axis of symmetry $x-\mu$, its fit density function is $f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$, according to the sample numbers, divide it into $2l+1$ groups, and the scope of each group is

$$\left[\mu-\frac{2l+1}{2}a, \mu-\frac{2l-1}{2}a\right)\cdots\left[\mu-\frac{a}{2}, \mu+\frac{a}{2}\right),$$
$$\left[\mu+\frac{a}{2}, \mu+\frac{3a}{2}\right)\cdots\left[\mu+\frac{2l-1}{2}a, \mu+\frac{2l+1}{2}a\right)$$

and denote it as m_i, which is the number of samples in $\left[\mu+\frac{2i-1}{2}a, \mu+\frac{2i+1}{2}a\right), i=0,1,\cdots,l$,

$$A_i=\left\{x\in\left[\mu+\frac{2i-1}{2}a, \mu+\frac{2i+1}{2}a\right)\right\}$$

$$p(A_i)=\int_{\mu+\frac{2i-1}{2}a}^{\mu+\frac{2i+1}{2}a} f(x)\mathrm{d}x.$$

Drawing a diagram as

$$\left[\mu-\frac{2i+1}{2}a, \mu-\frac{2i-1}{2}a\right),\cdots,\left[\mu-\frac{a}{2}, \mu+\frac{a}{2}\right),$$
$$m_i,\cdots,m_0,\cdots\left[\mu+\frac{2i-1}{2}a, \mu+\frac{2i+1}{2}a\right)\cdots,m$$

This diagram is called frequency distribution of classes of sample numbers. Obviously,

$$m_i = NS_i, m_0+2m_1$$
$$+\cdots+2m_l = N.$$

Set a as base, $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ as height in each interval $\left[\mu+\frac{2i-1}{2}a,\mu+\frac{2i+1}{2}a\right)$, make a rectangle S_i, $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{l^2a^2}{2\sigma^2}}, i=0,1,\cdots,l$, and then the histogram is made out (**Figure 1**).

And $S_i=P(A_i)$,

$$S_0=\frac{a}{\sqrt{2\pi}\sigma},2S_1=\frac{2a}{\sqrt{2\pi}\sigma}e^{-\frac{a^2}{2\sigma^2}},2S_2=\frac{2a}{\sqrt{2\pi}\sigma}e^{-\frac{4a^2}{2\sigma^2}},\cdots$$

$$2S_l=\frac{2a}{\sqrt{2\pi}\sigma}e^{-\frac{l^2a^2}{2\sigma^2}}$$

$s_0+2s_1+\cdots+2s_l$ equals

$$S=s_0+2\sum_1^l s_i=\frac{a}{\sqrt{2\pi}\sigma}+\frac{2a}{\sqrt{2\pi}\sigma}\left(e^{-\frac{a^2}{2\sigma^2}}+e^{-\frac{2^2a^2}{2\sigma^2}}+\cdots+e^{-\frac{l^2a^2}{2\sigma^2}}\right)$$

(**Table 1**)

$$H=\lim_{l\to\infty}S=\lim_{l\to\infty}(s_0+2\sum_1^l s_i)$$

$$=\frac{a}{\sqrt{2\pi}\sigma}+\frac{2a\cdot1}{\sqrt{2\pi}\sigma}\lim_{l\to\infty}\left(e^{-\frac{a^2}{2\sigma^2}}+\cdots+e^{-\frac{l^2a^2}{2\sigma^2}}\right)$$

$$=1$$

$$\frac{dH}{da}=\frac{1}{\sqrt{2\pi}\sigma}+\frac{2}{\sqrt{2\pi}\sigma}\lim_{l\to\infty}\left[e^{-\frac{a^2}{2\sigma^2}}\left(1-\frac{a^2}{\sigma^2}\right)+\cdots+e^{-\frac{l^2a^2}{2\sigma^2}}\left(1-\frac{l^2a^2}{\sigma^2}\right)\right](1)$$

1) Set $G=e^{-\frac{a^2}{2\sigma^2}}\left(1-\frac{a^2}{\sigma^2}\right)+\cdots+e^{-\frac{l^2a^2}{2\sigma^2}}\left(1-\frac{l^2a^2}{\sigma^2}\right)$,

when $a=\sigma$, we get the minimum of $G(a)$.

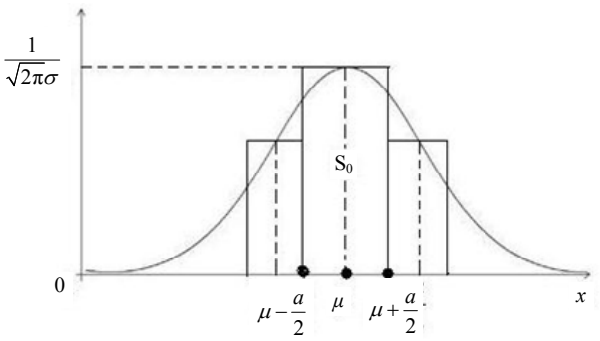

Figure 1. The formation process of histogram.

Table 1. Normal distribution of data (Example 1).

610	690	870	910	930	1020	1050	1110	1120
1130	1150	1220	1240	1260	1270	1280	1320	1340
1360	1400	1410	1410	1420	1440	1450	1450	1490
1500	1540	1540	1550	1570	1580	1580	1590	1630
1650	1670	1710	1720	1730	1750	1770	1840	1860
1870	1880	1940	1970	2060	2080	2120	2300	2380

2) When $\frac{dH}{da}=0$, we have the turning point $a=\sigma$, Now

$$\lim_{l\to\infty}\left[e^{-\frac{4\sigma^2}{2\sigma^2}}\left(1-\frac{4\sigma^2}{\sigma^2}\right)+\cdots+e^{-\frac{l^2\sigma^2}{2\sigma^2}}\left(1-\frac{l^2\sigma^2}{\sigma^2}\right)\right]$$

$$=-3e^{-2}-8e^{-\frac{9}{2}}-15e^{-8}+\cdots=-\frac{1}{2}$$

when $a=\sigma$, the maximum of H is 1 and S is the partial sum of H. According to the Reference [3], $4\sigma\le(2l+1)a\le6\sigma$, so σ is the upper bound of class interval a [4].

2.2. Major Theorem

Theorem 1. If a group of numbers shows a normal distribution, its fit density is $\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ [3].

When we discuss the isometric group, σ is the upper bound of a (a is the class interval).

1) Definitions

Definition 1

We divide $\left[\mu,\mu+\left(l+\frac{1}{2}\right)a\right)$ into four equal parts, Their points are Q_0,Q_1,Q_2,Q_3,Q_4, $Q_4-Q_0=\frac{R^*}{2}$ is the mid-range. $Q_3-Q_1=Q_2-Q_0=\frac{R^*}{4}$ is the quartile. $Q_1-Q_0=\frac{R^*}{8}$ quantile of order eight.

Definition 2

The midpoint of interval $\left[\mu,\frac{R^*}{2}\right]$ (or the symmetric interval) is called one-quarter range, and denoted as $\frac{R^*}{4}$.

Definition 3

If N numbers are distributed in R^*, then $\frac{|R^*|}{N}$ indicated the average distance each number shares.

Definition 4

If there is an interval which makes $\frac{a}{m_i} \approx \frac{R^*}{N}$, then the interval is called similar interval.

2) Major theorem

Lemma 1. Mean value theorem integrals.

If $f(x)$ in $[-a,a]$ is consecutive and symmetrical accidentally function there is a point ξ makes $\int_{-a}^{a} f(x)\mathrm{d}x = 2f(\xi)a$.

Theorem 2. There are at least two similar intervals in a normally distributed series [3].

Proof. By Lemma 2, there is $\frac{a}{m_i} \le \frac{R^*}{N} \le \frac{a}{m_{i+1}}$, find $\min\left\{\left|\frac{R^*}{N}-\frac{a}{m_i}\right|,\left|\frac{R^*}{N}-\frac{a}{m_{i+1}}\right|\right\}$.

Let

$$\min\{\ \} = \left|\frac{R^*}{N}-\frac{a}{m_i}\right|,$$

then the intervals in m_i are similar intervals. By Lemma 2, the similar intervals are at least 2.

Theorem 3. A normally distributed series has a similar interval in $\left(\frac{R^*}{8}+\frac{a}{2},\frac{3}{8}R^*-\frac{a}{2}\right)$, where a is the class interval [3].

3. Work out the Way to Find Class Interval by Theorem 3

1) Arrange the data in ascending order, and calculate the average μ and variance σ of the numbers.

2) Find the closed range R^*, and calculate $\overline{f} = \frac{R^*}{N}$; to Expanding from $\frac{1}{4}R^*$ ($\pm\frac{1}{4}R^*$, to the two points).

3) Find out the minimum similar intervals $s_1,\cdots,s_n$; and their numbers are $|s_1|,\cdots,|s_n|$ respectively. Let it satisfy $\min\left\{\left||s_1|-\frac{R^*}{N}\right|,\cdots,\left||s_n|-\frac{R}{N}\right|\right\}$.

Let $\left||s_1|-\frac{R^*}{N}\right|$ be the minimum value, thus let $|s_i| = s$.

4) Fix on the class interval a: $s\cdot\frac{R^*}{N} \approx a \le \frac{3\sigma^2}{R^*}$.

5) Grouping a group of numbers, link the groups up carefully. If it is done well, it can reflect the overall trend. The chain of numbers is $\cdots b_0 \le b_1 \le \cdots \le b_n \le b_{n+1}\cdots$, where $b_1,b_2,\cdots,b_n$ are included in one group. How to insert this class into the chain of data? We set a rule that this group should be included in (a,c), where

$$\begin{cases} a = \dfrac{b_1+b_0}{2}; \\ c = \dfrac{b_n+b_{n+1}}{2} \end{cases}$$

6) If it is a skewed distribution, the above method is also available, but need to do twice, referring to the flowing example.

3.1. Example 1

By a sample survey of living conditions of urban households, we get the following numbers of per capita monthly household income (already arranged).

The minimum number is 610, and the maximum number is 2380. It is a marmal distribution progression.

Mean:

$$\mu = \frac{610+690+\cdots+2380}{54} = 1495$$

$$\sigma^2 = \frac{(610-1495)^2+\cdots+(2380-1495)^2}{54} = 377.6^2$$

$c = 600$, $d = 2400$, $R^* = 2400-600 = 1800$.

2) Because it is not completely symmetrical, we just consider the data from 1500 to 2400. From 600 to 2400 there are 27 numbers, and the average distance between them: $F = \frac{900}{27} = 33.3$.

$\frac{R^*}{4} = 450$ its coordinate is 1500 + 450 = 1950;

$\frac{R^*}{8} = 225$ its coordinate is 225 + 1500 = 1725;

$\frac{R^*}{8} = 675$ its coordinate is 1500 + 675 = 2175.

3) There are 12 points included in $\left(\frac{R^*}{8},\frac{3R^*}{8}\right)$, that is (1725, 2175), the length of this interval is 450. The 12, points are 1730, 1750, 1770, 1840, 1860, 1870, 1880, 1940, 1970, 2060, 2080 and 2120, When calculating the average distance of the 12 points, the interval we should consider is $\frac{450}{12} = 37.5 > 33.3$ (**Table 2**).

It is suggested that the similar interval is closer to $\frac{1}{8}R^*$ than to $\frac{3}{8}R^*$.

It is easy to see that there are six numbers between 1800 and 2000, which are 1840, 1860, 1870, 1880, 1940, and 1970, and $F_1' = \frac{200}{6} = 33.3$, which is similar to

$F_2 = 33.3$.

Let a' be the class interval, then 6 × 33.3 = 200 is the valuation of a'. We could get the following distribution series after further arrangement.

3.2. Example 2

By a sample survey of living conditions of urban households, we get the following numbers of per capita monthly household income (already arranged).

1) **Table 3**: Skewed distribution of data for Example 2 [1].

The minimum number is 810, and the maximum number is 2380. And the average is

$$\mu = \frac{810+840+\cdots+2300+2380}{54} = 1497.2$$

$$\sigma^2 = \frac{(810-1497.2)^2+\cdots+(2380-1497.2)^2}{54} = 365^2$$

where $c = 800$, $d = 2400$, the closed range is $R^* = 2400-800 = 1600$ (**Table 4**).

2) Because it is not completely symmetrical, we divide into two steps to finish it. From 800 to 1497.2, there are 27 numbers, and the average distance between them is

Table 2. Normal distribution of data (Example 1).

810	840	870	920	990	1050	1070	1080	1100
1120	1120	1160	1200	1240	1280	1300	1310	1330
1350	1350	1360	1390	1400	1420	1460	1460	1490
1500	1540	1540	1550	1570	1580	1580	1590	1630
1650	1670	1710	1720	1730	1750	1770	1840	1860
1870	1880	1940	1970	2060	2080	2120	2300	2380

Table 3. Here choose C = 600, D = 2400 for Example 1.

Per Capita Monthly Income (RMB Yuan)	Households	Frequency (%)
600 - 800	2	3.70
800 - 1000	3	5.56
1000 - 1200	6	11.11
1200 - 1400	8	14.81
1400 - 1600	16	29.63
1600 - 1800	8	14.81
1800 - 2000	6	11.11
2000 - 2200	3	5.56
2200 - 2400	2	3.70
Total	54	100.00

Table 4. Here choose C = 800, D = 2400 for Example 3, [1].

Per Capita Monthly Income (RMB Yuan)	Households	Frequency (%)
800 - 1000	5	9.26
1000 - 1200	7	12.96
1200 - 1400	10	18.52
1400 - 1600	13	24.07
1600 - 1800	8	14.82
1800 - 2000	6	11.11
2000 - 2200	3	5.56
2200 - 2400	2	3.70
Total	54	100.00

$$F_1 = \frac{1497.2-800}{27} = 25.8.$$

The half range is $\frac{R^*}{2} = 697.2$; $\frac{R^*}{4} = 348.6$, its coordinate is 800 + 348.6 = 1148.6; $\frac{R^*}{8} = 174.3$, its coordinate is 800 + 174.3 = 974.3.

3) $\frac{3R^*}{8} = 800 + \frac{3R^*}{8} = 800 + 3\times 174.3 = 1322.9$. There are 13 points included in $\left(\frac{R^*}{8}, \frac{3R^*}{8}\right)$, that is (974, 1323), the length of this interval is 349. The 13 points are 990, 1050, 1070, 1080, 1100, 1120, 1120, 1160, 1200, 1240, 1280, 1300, and 1310.

When calculating the average distance of the 13 points, the interval we should consider is

$$\left(990-\frac{990-920}{2}, 1310+10\right) = (955, 1320),$$

and the average distance is $\frac{1320-955}{13} = \frac{365}{13} = 28.08$, which is greater than $F_2 = 25.8$. It suggests that when selecting similar intervals, it is skewed to the right.

Thus we should delete 990 in the following discussion.

We have the result as the following.

$$\xrightarrow{30} 1050 \xrightarrow{20} 1070 \xrightarrow{10} 1080 \xrightarrow{20}$$

$$1100 \xrightarrow{20} 1120 \xrightarrow{20} 1140 \xrightarrow{20}$$

$$1160 \xrightarrow{40} 1200 \xrightarrow{40} 1240 \xrightarrow{40} 1280$$

We add a horizontal line between numbers.

We find that $\frac{R^*}{4} = 1148.6$ is a point on the right side, the distance is 40, the average distance of points on the left is 20;

We could further ascertain that there are two points on the right side of 1148.6 at most, which are 1160 and 1200. It is easy to find that there are 8 numbers in interval (1020, 1220) (see the fifth points in the third part), which are 1050, 1070, 1080, 1100, 1120, 1120, 1160, and 1200.

4) Since $F_1' = \frac{200}{8} = 25 \approx 25.8 = F_1$, then the interval (1020, 1220) is a similar interval. And $S = 8$, $206.4 = 8 \times 25.8 \approx a'$, $a = 200$.

The Length of similar interval is a'.

$a' = 6 \times 33.3 = 199.8$ is the luation of a.

By the conclusion of 1) and 2), a is 200, we could get the following distribution series after further arrangement.

Mathematics diagram Frequency distribution of the per capita monthly income available for living expenses of urban households in a certain city.

REFERENCES

[1] Q. N. Xie and Z. Z. Han, "Principle of Statistics," 6th Edition, Jinan University Press, Jinan, 1991, pp. 53-64.

[2] C. S. Wu, "Probability and Statistics," Higher Education Press, Beijing, 2004, pp. 128-144.

[3] B. H. Qian and L. W. Huang, "Statistics," Sichuan People's Publishing House, Chengdu, 2001, p. 107.

[4] Z. S. Jia, All the Proof Are Published in the *Mathematics Practices and Theory*, Vol. 40, No. 20, 2011, pp. 238-244.

