

西山矿务局四个矿工作面损失率的 概率分布检验与水平差异分析

杜乃衡 李国荣 左蔚文

〔摘 要〕 本文根据西山矿务局官地、杜儿坪、白家庄和西铭四个矿的工作面损失率数据资料,运用数学方法讨论了其概率分布及检验,并给出相应的参数估计和水平差异分析。以期讨论的结果能为生产管理部门制定合理的损失定额提供理论依据。

〔关键词〕 损失率, 概率分布, 假设检验, 水平效应

0 引 言

煤矿开采中,某一地段(工作面、采区或全矿井)所损失的煤炭储量与该地段全部储量的百分比,称为损失率。科学地制定损失定额对于合理开采煤炭资源、延长矿井寿命与提高经济效益具有重要意义。按照中国统配煤矿总公司关于合理开采煤炭资源的规定,损失率是考核资源利用、开采技术及管理水平等的主要经济技术指标之一。

目前各矿对开采过程中损失率的测算和考查还缺乏深入的定量分析。本文根据西山矿务局提供的官地、杜儿坪、白家庄和西铭(以下顺次简称做G、D、B、X)四矿1983年1月至1989年4月的工作面损失率数据资料(表1)运用数理统计方法讨论了其概率分布,并进

表1 G矿工作面损失率样本数据

19.50	11.40	9.00	8.00	1.40	7.60	7.60	6.00
6.50	8.10	9.20	11.00	14.60	9.60	5.50	7.40
4.30	6.70	12.00	7.90	13.30	7.40	8.30	11.30
7.10	3.80	9.50	9.20	12.60	5.60	8.60	10.80
10.60	3.70	6.40	9.60	13.50	6.20	6.70	7.00
17.80	3.20	4.50	10.50	14.40	8.60	8.70	7.00
10.00	8.50	8.50	11.20	12.60	10.10	8.50	
3.00	7.20	8.50	12.50	14.40	8.60	8.40	
4.00	3.00	11.00	10.40	13.70	9.40	7.30	
4.30	4.30	3.70	10.60	13.10	7.20	6.10	

注:因篇幅限制略去D、B、X矿数据

行了非参数、参数检验,给出了四矿水平差异分析。以期讨论的结果能为有关生产管理部门制定合理的损失定额提供理论依据。

文中前三部分以G矿数据资料为样本总体研究其概率分布和检验,并以三种不同的方法做出参数估计。其余D、B、X矿可按完全类似的方法导出相应结果。所采用的工作面损失率数据由下列公式测算给出:

$$\text{工作面损失率}(\%) = \frac{\text{工作面损失量}}{\text{工作面采出量} + \text{工作面损失量}} \quad (1)$$

1 正态性判定

设G矿损失率数据总体为X,已知样本容量为: $n=76$ 按下列步骤将X分组、计算、列表。

第1步 求得 $\min_{1 \leq i \leq 76} \{x_i\} = 1.4$ $\max_{1 \leq i \leq 76} \{x_i\} = 19.5$

$$R = \max_{1 \leq i \leq 76} \{x_i\} - \min_{1 \leq i \leq 76} \{x_i\} = 18.1$$

第2步 用下列分点将 $(-\infty, +\infty)$ 分成 m 个子区间。 m 的大小由 n 和 R 决定,此处取 $m=11$,即有 $-\infty < 1.5 < 3.5 < 5.5 < 7.5 < 9.5 < 11.5 < 13.5 < 15.5 < 17.5 < 19.5 < +\infty$,从而得到11个半开半闭的区间如表2“组限”列所示。

第3步 计算落入第 i 个子区间的观察值个数即频数 f_i 和频率 f_i/n 。

第4步 列表2,其中序号为 k 的累积频率 $F_n(x_i)$ 按下式计算: $F_n(x_i) = \sum_{i=1}^k f_i/n$ 。

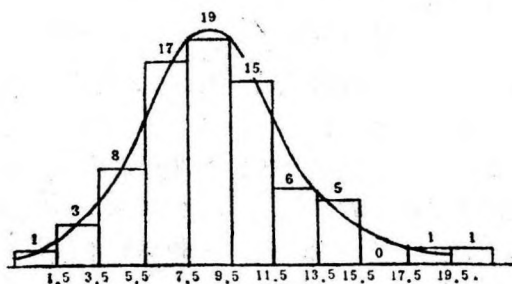


图1 G矿损失率组频直方图

按表2中所列数据,以样本观察值为横坐标,以相应的频数 f_i 为纵坐标在横轴上作出相邻的一排长方形(共 $m=11$ 个),其底宽均取 $\Delta x_i = 2$,而高则为 $f_i \cdot (1/\Delta x_i \cdot n) = f_i/(2 \times 76)$ 单位,见图1。

由科尔莫戈罗夫强大数定理当 n, m 充分大时这图形将近似于某个概率密度函数 $\varphi(x)$ 的图形。现从该图的变化可描出G矿损失率的概率密度曲线,它是近似正态分布的概率密度曲线。但此图仅能对G矿损失率总体X作出粗略反映。由于该图的制作往往因子区间的划分不同而出现较大误差,甚至得出错误的判断,故工程实际应用中对上述估计应进一步以曲线拟合法判断。具体做法如下:

按表2所列数据以 $(x_i, F_n(x_i))$ 为坐标在正态概率纸上描点,当这些点的位置几乎

表2 G矿损失率数据频数分布表

组序	组限	频数 f_i	频率 f_i/n	累积频率 $F_n(x_i)$
1	$(-\infty, 1.5)$	1	0.01	0.01
2	$[1.5, 3.5)$	3	0.04	0.05
3	$[3.5, 5.5)$	8	0.11	0.16
4	$[5.5, 7.5)$	17	0.22	0.38
5	$[7.5, 9.5)$	19	0.25	0.63
6	$[9.5, 11.5)$	15	0.20	0.83
7	$[11.5, 13.5)$	6	0.08	0.91
8	$[13.5, 15.5)$	5	0.07	0.98
9	$[15.5, 17.5)$	0	0	0.98
10	$[17.5, 19.5)$	1	0.01	0.99
11	$[19.5, +\infty)$	1	0.01	1.00

处于同一直线 l 上时,即可认为该总体是服从正态分布的。图 2 给出此例的描点结果。由图 2 可见,关于 G 矿损失率数据总体正态性的判断是较为明显的。至此可以作出关于 G 矿工作面损失率数据总体 X 分布的判断,即其概率密度函数为

$$\varphi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (2)$$

2 参数估计

经过 1 的判断后,再需对式(2)中的 μ 、 σ 作出参数估计而后加以检验。下面给出三种不同方法

对参数 μ 、 σ 估计以便说明其一致性(数理统计方法在工程问题应用中,常对参数予以不同方法进行估计、加以比较、互相验证以说明其一致性)。

2.1 图估计

图估计是工程中常用的一种简便方法,但其结果较为粗略,方法如下:

利用图 2 中正态坐标纸上描点所确定的直线 l ,由 $p=0.50$ 与 l 的交点所对应的横坐标值是 X 总体期望 $E(X)$ 的估计值 $\hat{\mu}$;而 $p=0.159$ 与 l 的交点所对应的横坐标值即是 $\hat{\mu} - \hat{\sigma}$ (其中 $\hat{\sigma}$ 是总体均方差 $\sqrt{D(X)}$ 的估计值)。此处由图 2 按上法可求得: $\hat{\mu}=8.6$, $\hat{\mu} - \hat{\sigma}=5.4$ 。于是有

$$\begin{cases} \hat{\mu}=8.6 \\ \hat{\sigma}=3.2 \end{cases}.$$

2.2 极大似然估计

Fisher 引进的极大似然估计,从理论观点看至今仍是参数估计中最重要的方法,但其利用数学工具较多,这里 μ 、 σ 的似然函数为:

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^n \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \right] \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2\right) \end{aligned}$$

似然方程为

$$\begin{cases} \frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases} \quad (3)$$

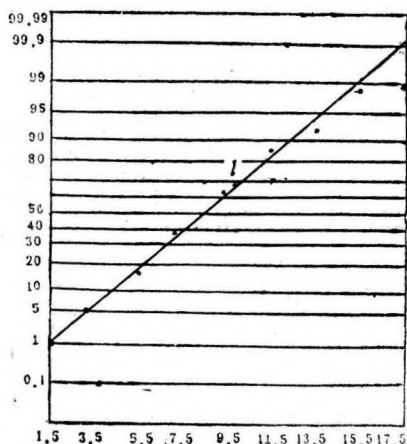


图 2 正态坐标纸拟合

由(3)式解得:

$$\begin{cases} \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases} \quad (4)$$

代入样本数据值及 $n=76$ 得

$$\begin{cases} \hat{\mu} = 8.685 \\ \hat{\sigma} = 3.42 \quad \hat{\sigma}^2 = 11.71 \end{cases}$$

2.3 最优无偏估计

所谓参数的最优无偏估计是指具有最小方差的无偏估计量,其详细理论依据这里不赘述,仅给出如下结论和估值。

这里 $E(X)$ 的估值同 2.2 的极大似然估计结果一致, 即有 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$.

而 $D(X)$ 的估计则为 $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$. (5)

于是有:

$$\begin{cases} \hat{\mu} = 8.685 \\ \hat{\sigma} = 3.44 \quad \hat{\sigma}^2 = 11.86 \end{cases}$$

现将上面三种不同估计方法所得结果归纳为如下的表 3 .

由表3可见三种估计结果的一致性。下面就 1 的正态性判断和 2.3 的最优无偏估计所得 $\hat{\mu}$ 和 $\hat{\sigma}$ 进行假设检验。

表 3 分布(2)参数估值比较

估计法 参 数	图估计	极大似然估计	最优无偏估计
损 失 率 均 值 μ	8.6	8.685	8.685
损失率波 动均方差 σ	3.2	3.42	3.44

3 Pearson 标准检验

由 Pearson 给出的统计量

$$\eta = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} \quad (6)$$

其极限分布为 $\chi^2(k-r-1)$, 其中 r 为待估参数。不论总体 X 属于什么类型分布,均可以此统计量作出相应的假设检验。检验过程可按如下步骤进行:

第 1 步 选定显著性水平 $\alpha=0.05$, 给出假设

$$H_0: F(x) = F_0(x) \quad H_1: F(x) \neq F_0(x)$$

其中 $F_0(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$, μ, σ 为待估计参数。下面计算中将按 2.3

的估计结果选取 $\hat{\mu} = 8.685$, $\hat{\sigma} = 3.44$.

第2步 按 G 矿损失率数据总体 X 取值情况 (可结合表1) 将 $(-\infty, +\infty)$ 划分为 k 个子区间, 此处暂取 $k=11$.

第3步 计算 f_i ——实际观察频数;

$$p_i \text{——} p_i = F_0(x_{i+1}) - F_0(x_i);$$

np_i ——理论频数。

并做相继运算后列成如下的表4。

表4 Pearson 检验计算表

序号	组 限	f_i	np_i	$f_i - np_i$	$(f_i - np_i)^2$	$(f_i - np_i)^2 / np_i$
1	$(-\infty, 1.5)$	1	0.9			
2	$[1.5, 3.5)$	3	3.6	-0.9	0.81	0.063
3	$[3.5, 5.5)$	8	8.4			
4	$[5.5, 7.5)$	17	14.4	2.6	6.76	0.47
5	$[7.5, 9.5)$	19	17.4	1.6	2.56	0.147
6	$[9.5, 11.5)$	15	15.1	0.1	0.01	0.0006
7	$[11.5, 13.5)$	6	9.5	3.5	12.25	1.29
8	$[13.5, 15.5)$	5	4.3			
9	$[15.5, 17.5)$	0	1.4			
10	$[17.5, 19.5)$	1	0.4	0.9	0.81	0.133
11	$[19.5, +\infty)$	1	0			
Σ						2.1

注意: 计算中对一些理论频数小于5的组, 应予适当合并。这里经过并组后取 $\bar{k} = 6$ 。

第4步 确定自由度。因 $r=2$, 故 $\bar{k} - r - 1 = 3$, 根据显著性水平 $\alpha=0.05$, 查表得

$$\chi^2_{0.05}(3) = 7.815 > 2.1$$

故接受假设 H_0 。即认为 G 矿工作面损失率总体 X 服从正态分布 $N(8.685, 11.86)$ 。

对于 D 、 B 、 X 矿可采用与上述相同的方法估计和检验得

D 矿工作面损失率 $\sim N(14.5, 21.935)$;

B 矿工作面损失率 $\sim N(15.7, 21.28)$;

X 矿工作面损失率 $\sim N(11.14, 18.77)$ 。

4 水平差异分析

将四个矿看做四个水平, 设其总体分别为: $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), N(\mu_3, \sigma_3^2), N(\mu_4, \sigma_4^2)$ 。我们检验四个水平下的工作面损失率在方差齐性的前提下有无显著差异; 即判断四个矿之间工作面损失率的差异, 除因随机波动引起的误差外, 是否有因各水平间效应引起的实质性差异。

检验步骤如下:

第1步 检验四个总体的方差齐性。

取显著性水平 $\alpha=0.01$, 假设 $H_0: \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ 具有方差齐性。

分别对四个总体计算

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (7)$$

式中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $n=76$.

得 $S_1^2=11.86$, $S_2^2=21.94$, $S_3^2=21.28$, $S_4^2=18.77$.

及 $S_2^2/S_1^2=1.85$, $S_2^2/S_3^2=1.03$, $S_3^2/S_4^2=1.134$.

查表得 $F_{0.05/2}(n_1-1, n_2-1)=F_{0.005}(75, 75)=1.93$.

由 $1.85 < 1.93$, $1.03 < 1.93$, $1.134 < 1.93$ 判断接受假设 H_0 , 即认为四个总体具有方差齐性.

第2步 取显著性水平 $\alpha=0.01$, 给出假设:

H_0 : 四总体均值无显著差异;

H_1 : 四总体均值有显著差异.

$$\text{第3步 计算总变差 } S_T = \sum_{j=1}^s \sum_{i=1}^n (x_{ij} - \bar{x})^2 \quad (8)$$

式中 $\bar{x} = \frac{1}{n} \sum_{j=1}^s \sum_{i=1}^n x_{ij}$;

s ——水平个数;

n ——每一水平下样本数据个数.

按照Fisher对总变差 S_T 的分解

$$S_T = S_E + S_A \quad (9)$$

其中 $S_E = \sum_{j=1}^s \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})^2$;

$S_A = \sum_{j=1}^s \sum_{i=1}^n (\bar{x}_{\cdot j} - \bar{x})^2$.

这里, S_E 反映样本组内的随机波动, 称为样本组内平方和; S_A 则反映样本组间的差异, 即由不同水平效应引起的变差, 称为样本组间平方和. 经计算得

$$\bar{x} = 12.5, S_E = 5511.61, S_A = 2328.8.$$

第4步 求 F 比.

$$[(n-s)/(s-1)] \cdot (S_A/S_E) = 42.25.$$

查表得: $F_{0.01}(s-1, sn-1) = F_{0.01}(3, 300) = 3.90 < 42.25$

故在水平 0.01 下拒绝假设 H_0 , 而接受假设 H_1 , 即认为各水平间的差异是显著的.

我们进一步还可以同样的方法验知, 即使除去有较明显差异的 G 矿外, 其余 D, B, X 三矿间工作面损失率的差异仍是显著的.

上述分析结果表明: G, D, B 和 X 四矿之间工作面损失率的差异绝不仅仅 是由于各种作用微小的随机因素所引起; 而是存在某些作用较明显的特殊因素引起的. 比如管理或技术操作水平以及地质因素等. 按统计语言讲, 此即所谓水平效应. 这将有待实际生产管理部门进一步考察, 以便对各矿做出合理的损失率定额.

此文中所采用的四个矿工作面损失率数据资料由西山矿务局常玉跃, 刘树森等收集提供,

在此致谢。

参 考 文 献

1. 王梓坤. 概率论基础及其应用. 北京: 科学出版社, 1976.7: 213—244
2. 中山大学数力系. 概率论及数理统计. 北京: 人民教育出版社, 1980.3: 下册
3. 朱晓岚等. 矿体几何学. 中国矿业学院出版社, 1987.8: 156—166
4. B.B 诺里莫夫. 数学统计学在材料分析中的应用. 莫斯科. 1960: 98—121

The Testing for Probability Distribution and Analysis of Horizontal Difference of the Observational Data of the Loss Rate in the Coal Faces of Four Mines in Xishan Coal Bureau

Du Naiheng, Li Guorong, Zuo weiwen

Abstract

Using mathematical methods, the authors of the paper discusses the probability distribution of the observational data of the loss rate in the coal faces of four mines, i.e. Du'erping, Baijiazhuang Guandi and Ximing, all of them belong to Xishan Coal Bureau. The paper also discusses the methods of testing for the probability distribution and gives an estimation of certain parameters and an analysis of horizontal difference. The authors hope that the results of the analysis can be used as a theoretical basis for the management of production in setting reasonable norms of loss rates.

Key words: loss rate, probability distribution, hypothesis testing, horizontal effects