第 21 卷 第 4 期　　应 用 数 学 学 报　　Vol. 21 No. 4
1998 年 10 月　　ACTA MATHEMATICA APPLICATAE SINICA　　Oct., 1998

极值分布和威布尔分布异常数据的检验方法*

费鹤良　　陆向薇　　徐晓岭
(上海师范大学数学系，上海 200234)

摘　要　本文对威布尔分布和极值分布异常数据的检验给出了一系列的方法. 首先，导出了极值分布下一般 Dixon 型统计量的精确分布，同时还给出了改进的 G 型统计量，及它们的分位点表. 最后本文提出了一个新的统计量： F 型统计量，并用 Monte-Carlo 模拟的方法给出其分位点表，从而首次给出威布尔分布异常值的直接检验方法. 本文进一步讨论了这些检验方法的功效，且表明 F 型检验是最优的.

关键词　异常值，检验，效，极值分布，威布尔分布

1　一般的 Dixon 型统计量及其精确分布

在异常数据的检验中， Dixon 型统计量是较常用的一个统计量. [1] 将 Dixon 型统计量 $r_{ij}=\frac{x_{(n)}-x_{(n-i)}}{x_{(n)}-x_{(j+1)}}$, $i=1,2$, $j=0,1,2$, 作为剔除特大异常值的统计检验运用到极值分布，导出其在原假设 H_0 下的分布函数，并给出了分位点表. 但这只能用来做 "consecutive" 检验 (即依顺序一个一个地检验), 而不能用在 "block" 检验 (即最大的或最小的几个数据同时得到检验). 本文导出了一般的 Dixon 型统计量： $D^*=\frac{x_{(s)}-x_{(r)}}{x_{(q)}-x_{(p)}}$, $1\le p\le r<s\le q\le n$, $q-p>s-r$. 在假设 H_0 下的精确分布，并可以通过数值积分求出了它的分位点. 这个统计量可用于剔除异常值的 "consecutive" 检验，又可以做 "block" 检验.

设极值分布

$$G(x)=1-\exp\left\{-\exp\left(\frac{x-\mu}{\sigma}\right)\right\},\qquad -\infty<x,\quad \mu<+\infty,\quad \sigma>0. \tag{1}$$

定理 1　设 $x_1,x_2,\cdots,x_n$ 独立同分布于极值分布 $G(x)$, 其中 $-\infty<\mu<+\infty$ 为位置参数， $\sigma>0$ 为刻度参数， $G(x)=1-\exp\{-\exp(\frac{x-\mu}{\sigma})\}$, $-\infty<x<+\infty$, 则 $D^*=\frac{x_{(s)}-x_{(r)}}{x_{(q)}-x_{(p)}}$, $1\le p<r<s\le q\le n$, $q-p>s-r$ 的分布函数为：

$$F(t)=K\int_0^1\frac{1}{(a+bw)^2d}\left[\frac{2a+d+c+2bw}{(a+c+d+bw)^2}-\frac{2a+c+2bw+dw^t}{(a+c+bw+dw^t)^2}\right]\mathrm{d}w, \tag{2}$$

本文 1997 年 9 月 9 日收到. 1998 年 1 月 12 日收到修改稿.
*上海市高等学校科学技术发展基金资助项目.

其中 $a=l+n-q+1,\ b=i+r-p-j,\ c=k+q-s-l,\ d=s+j-r-k,$

$$K=M\sum_{i=0}^{p-1}\sum_{j=0}^{r-p-1}\sum_{k=0}^{s-r-1}\sum_{l=0}^{q-s-1}(-1)^{i+j+k+l+1}\binom{p-1}{i}\binom{r-p-1}{j}\binom{s-r-1}{k}\binom{q-s-1}{l},$$

$$M=\frac{n!}{(p-1)!(r-p-1)!(s-r-1)!(q-s-1)!(n-q)!}.$$

证 令 $y_{(i)}=\frac{x_{(i)}-\mu}{\sigma}$, 则 $y_1,y_2,\cdots,y_n$ 独立同分布于标准极值分布 $G_1(y)$, $G_1(y)$ 是位置参数为 0, 刻度参数为 1 的标准极值分布, 则 $D^*=\frac{x_{(s)}-x_{(r)}}{x_{(q)}-x_{(p)}}=\frac{y_{(s)}-y_{(r)}}{y_{(q)}-y_{(p)}}$, $1\le p<r<s\le q\le n$, $q-p>s-r(y_{(p)},y_{(r)},y_{(s)},y_{(q)})$ 的联合密度为: $(x=y_{(p)},y=y_{(r)},z=y_{(s)},u=y_{(q)})$

$$\begin{aligned}f(x,y,z,u)=&\frac{n!}{(p-1)!(r-p-1)!(s-r-1)!(q-s-1)!(n-q)!}\\&\cdot[G_1(x)]^{p-1}[G_1(u)-G_1(z)]^{q-s-1}[G_1(x)-G_1(y)]^{s-r-1}\\&\cdot[G_1(u)-GA*-1(z)]^{q-s-1}[1-G_1(u)]^{n-q}\\&\cdot g_1(x)g_1(y)g_1(z)g_1(u),\end{aligned}\tag{3}$$

其中 $-\infty<x<y<z<u<+\infty$, $G_1(t)=1-\exp\{-\exp(t)\}$, $g_1(t)=\exp\{-\exp(t)+t\}$, $-\infty<t<+\infty$. 设 $V=Z-y$, $Z=z$, $U=u,|J|=V,R=\frac{z-y}{u-x}$,

$$\begin{aligned}f(R)=&M\int_0^{+\infty}\mathrm{d}v\int_{-\infty}^{+\infty}\mathrm{d}z\int_z^{+\infty}(1-\mathrm{e}^{-\mathrm{e}^{u-v}})^{p-1}(\mathrm{e}^{-\mathrm{e}^{u-v}}-\mathrm{e}^{-\mathrm{e}^{z-RV}})^{r-p-1}\\&\cdot(\mathrm{e}^{-\mathrm{e}^{z-RV}}-\mathrm{e}^{-\mathrm{e}^{z}})^{s-r-1}\cdot(\mathrm{e}^{-\mathrm{e}^{z}}-\mathrm{e}^{-\mathrm{e}^{u}})^{q-s-1}\mathrm{e}^{-(n-q)\mathrm{e}^{u}}\\&\cdot\mathrm{e}^{-\mathrm{e}^{u-v}+u-v}\mathrm{e}^{-\mathrm{e}^{z-RV}}\mathrm{e}^{z-RV}\mathrm{e}^{-\mathrm{e}^{z}+z}\mathrm{e}^{-\mathrm{e}^{u}+u}V\,\mathrm{d}u,\end{aligned}$$

其中

$$M=\frac{n!}{(p-1)!(r-p-1)!(s-r-1)!(q-s-1)!(n-q)!}.$$

设 $\mathrm{e}^z=\xi$, 并令 $a=l+n-q+1,\ b=i+r-p-j,\ c=k+q-s-l,\ d=s+j-r-k,$

$$\begin{aligned}f(R)=&M\int_0^1\mathrm{d}\omega\int_0^{+\infty}\mathrm{d}\xi\int_\xi^{+\infty}(1-\mathrm{e}^{-\eta\omega})^{p-1}(\mathrm{e}^{-\eta\omega}-\mathrm{e}^{-\xi\omega^R})^{r-p-1}\\&\cdot(\mathrm{e}^{-\xi\omega^R}-\mathrm{e}^{-\xi})^{s-r-1}(\mathrm{e}^{-\xi}-\mathrm{e}^{-\eta})^{q-s-1}\mathrm{e}^{-(n-q)\eta}\\&\cdot\mathrm{e}^{-\eta\omega}\mathrm{e}^{-\xi\omega^R}\omega^R\mathrm{e}^{-\xi}\xi\mathrm{e}^{-\eta}\eta(-\ln\omega)\,\mathrm{d}\eta\\=&K\int_0^1\mathrm{d}\omega\int_0^{+\infty}\mathrm{d}\xi\int_\xi^{+\infty}\mathrm{e}^{-[(s+j-r-k)\omega^R+(k+q-s-l)]\xi}\xi\\&\cdot\mathrm{e}^{-[(i+r-p-j)\omega+(l+n-q+1)]\eta}\eta\omega^R\ln\omega\,\mathrm{d}\eta,\end{aligned}$$

其中

$$\begin{aligned}K=&M\sum_{i=0}^{p-1}\sum_{j=0}^{r-p-1}\sum_{k=0}^{s-r-1}\sum_{l=0}^{q-s-1}\\=&(-1)^{i+j+k+l+1}\binom{p-1}{i}\binom{r-p-1}{j}\binom{s-r-1}{k}\binom{q-s-1}{l},\end{aligned}$$

$$f(R)=K\int_0^1\left[\frac{2}{(a+b\omega)(a+c+b\omega+d\omega^R)^3}+\frac{1}{(a+b\omega)^2(a+c+b\omega+d\omega^R)^2}\right]\omega^R\ln\omega\mathrm{d}\omega.$$

设 $\omega^R=T$,

$$\begin{aligned}F(t)=&p\{R<t\}\\=&\int_0^t K\,\mathrm{d}R\int_0^1\left[\frac{2}{(a+b\omega)(a+c+b\omega+d\omega^R)^3}\right.\\&\left.+\frac{1}{(a+b\omega)^2(a+c+b\omega+d\omega^R)^2}\right]\omega^R\ln\omega\,\mathrm{d}\omega\\=&K\int_0^1\frac{1}{(a+bw)^2d}\left[\frac{2a+d+c+2bw}{(a+c+d+bw)^2}-\frac{2a+c+2bw+dw^t}{(a+c+bw+dw^t)^2}\right]\mathrm{d}w.\end{aligned}$$

要求置信水平为 $1-\alpha$ 的临界值，只需令 $F(t)=1-\alpha$，求出 t 的数值即可，也可以通过插值的方法得到分位点表. 显然，如果应用到 Weibull 分布上，即 $x_1,x_2,\cdots,x_n$ 独立同分布于两参数 Weibull 分布，则 $D^*=\frac{\ln x_{(s)}-\ln x_{(r)}}{\ln x_{(q)}-\ln x_{(p)}}$，$1\le p<r<s\le q\le n$，$q-p>s-r$.

2　G 型统计量

G 统计量是由陈振民 [2] 提出的，即

$$GH(n)=\frac{x_{(n)}-x_{(n-1)}}{(n-1)x_{(n)}-\sum\limits_{i=1}^{n-1}x_{(i)}},$$

对于截尾样本有

$$GH(n,r)=\frac{x_{(r)}-x_{(r-1)}}{(r-1)x_{(r)}-\sum\limits_{i=1}^{r-1}x_{(i)}},$$

显然，这个检验的统计量比 Dixon 型统计量包涵了更多的样本信息，但是，不能导出其分布函数的明显表达式.

为了讨论的方便，将 $GH(n,r)$ 改成

$$G_1=\frac{x_{(r)}-x_{(r-1)}}{x_{(r)}-\dfrac{1}{r-1}\sum\limits_{i=1}^{r-1}x_{(i)}}.$$

显然，$G_1=(r-1)GH(n,r)$. 当 n 较小时，我们一般通过 10000 次 Monte-Carlo 模拟得 G_r 的分位点.

为了使统计量尽可能地避免遭受屏蔽效应 (关于屏蔽效应另文讨论)，所以本文将 G 型统计量改成

$$G_2=\frac{x_{(r)}-x_{(r-2)}}{x_{(r)}-\dfrac{1}{r-2}\sum\limits_{i=1}^{r-2}x_{(i)}}.$$

我们通过 10000 次随机模拟给出了 $n=4(1)25$，$\alpha=0.99,0.95,0.90,0.10,0.05,0.01$ 时，G_2 的分布的分位点列于表 1.

显然，对于总体为两参数 Weibull 分布，可使用

$$WH_1=\frac{\ln x_{(r)}-\ln x_{(r-1)}}{\ln x_{(r)}-\frac{1}{r-1}\sum_{i=1}^{r-1}\ln x_{(i)}},\qquad WH_2=\frac{\ln x_{(r)}-\ln x_{(r-2)}}{\ln x_{(r)}-\frac{1}{r-2}\sum_{i=1}^{r-2}\ln x_{(i)}}$$

作为检验 Weibull 分布异常数据的检验统计量.

3 F型统计量

设两参数 Weibull 分布的分布函数为：

$$F(x)=\begin{cases}1-\exp\left[-(x/\eta)^m\right], & x>0,\\ 0, & x\le 0,\end{cases}$$

其中 $m>0$ 为形状参数，$\eta>0$ 为刻度参数.

当形状参数 $m=1$ 时，$F(x)$ 为指数分布. 在指数分布的异常数据检验中，常用 Fisher 型统计量

$$T=x_{(n)}\Big/\sum_{i=1}^{n}x_{(i)}.$$

我们知道在一定限制条件下，它是最优的[3]. 对于 Weibull 分布可同样提出一个新的检验统计量：

$$T^*=x_{(n)}^{\hat m}\Big/\sum_{i=1}^{n}x_{(i)}^{\hat m},\tag{5}$$

其中 $\hat m=1/\hat\sigma$, $\hat\sigma$ 为与 Weibull 分布相应的极值分布的刻度参数 σ 的不变估计.

定理 2 设 $x_{(1)},\cdots,x_{(n)}$ 为来自 (4) 的样本大小为 n 的随机样本的前 n 个次序统计量，$\hat m=1/\hat\sigma$, 其中 $\hat\sigma$ 为相应的极值分布的刻度参数 σ 的不变估计，则由 (5) 给出的 T^* 是枢轴量.

证 令 $y_{(i)}=(\frac{x_{(i)}}{\eta})^m$, 则 $y_{(1)},\cdots,y_{(n)}$ 是来自标准指数分布样本容量为 n 的次序统计量，所以

$$T^*=\frac{x_{(n)}^{\hat m}}{\sum_{i=1}^{n}x_{(i)}^{\hat m}}=\frac{\left[\left(\frac{x_{(n)}}{\eta}\right)^m\right]^{\hat m/m}}{\sum_{i=1}^{n}\left[\left(\frac{x_{(n)}}{\eta}\right)^m\right]^{\hat m/m}}=\frac{y_{(n)}^{\hat m/m}}{\sum_{i=1}^{n}y_{(i)}^{\hat m/m}}.$$

又 $\frac{\hat m}{m}=\frac{\sigma}{\hat\sigma}$, $\hat\sigma$ 是 σ 的不变估计，则 $\frac{\sigma}{\hat\sigma}$ 是枢轴量，从而可知：$\frac{\hat m}{m}$ 也是枢轴量，则 T^* 的分布函数与参数 m,η 无关，即 T^* 是枢轴量.

定理 3 设 $0<x_{(1)}\le\cdots\le x_{(r)}$ 是来自 (4) 的样本大小为 n 的前 $r(n\le 25)$ 个次序统计量，

$$\hat m=\frac{1}{\sum_{j=1}^{r}C(n,r,j)\ln x_{(j)}}>0,$$

则 T^* 是严格单调增的. 其中 $C(n,r,j)$ 为 BLUE 或 BLIE 系数.

证 首先观察系数 $C(n,r,j)$, $j=1,2,\cdots,r$, 即在 [4] 中的 BLUE 或 BLIE 系数，我们有如下结论：

(1) 对固定的 n 及固定的 r, 有 $\sum_{j=1}^{r} C(n,r,j)=0$.

(2) 对固定的 n,

①如果 $2 \le r \le 5$, 则有: $C(n,r,1) \le C(n,r,2) \le \cdots \le C(n,r,k_0-1) < 0 < C(n,r,k_0) \le \cdots \le C(n,r,r)$, $k_0 \ge 2$.

②如果 $6 \le r \le 25$,

则有: $0 > C(n,r,1) \ge C(n,r,2) \ge \cdots \ge C(n,r,k_0^*) \le C(n,r,k_0^*+1) \le \cdots \le C(n,r,k_0-1) < 0 < C(n,r,k_0) \le \cdots \le C(n,r,r)$, $2 \le k_0^*$, 并且有: 对 $1 \le i_0 \le k_0^*-1$, $-C(n,r,r) - \sum_{j=1}^{i_0} C(n,r,j) - (k_0-i_0)C(n,r,i_0) > 0$. (本结论是通过具体的计算而得到的.)

$$\frac{\partial \widehat{m}}{\partial x_{(r)}} = -\frac{C(n,r,r)}{\Big[\sum_{j=1}^{r} C(n,r,j)\ln x_{(j)}\Big]^2 x_{(r)}}, \qquad \frac{\partial \widehat{x}_{(r)}^{m}}{\partial x_{(r)}} = x_{(r)}^{\widehat{m}}\Big[\frac{\partial \widehat{m}}{\partial x_{(r)}}\ln x_{(r)} + \frac{\widehat{m}}{x_{(r)}}\Big],$$

$$\frac{\partial T^*}{\partial x_{(r)}} = \Big\{\frac{\partial \widehat{x}_{(r)}^{m}}{\partial x_{(r)}}\sum_{i=1}^{r} x_{(i)}^{\widehat{m}} - x_{(r)}^{\widehat{m}}\Big[\frac{\partial \widehat{x}_{(r)}^{m}}{\partial x_{(r)}} + \sum_{i=1}^{r-1} x_{(i)}^{\widehat{m}}\frac{\partial \widehat{m}}{\partial x_{(r)}}\ln x_{(i)}\Big]\Big\} \Big/ \Big\{\Big[\sum_{i=1}^{r} x_{(i)}^{\widehat{m}}\Big]^2\Big\}$$

$$= \frac{x_{(r)}^{\widehat{m}}}{\Big[\sum_{i=1}^{r} x_{(i)}^{\widehat{m}}\Big]^2}\sum_{i=1}^{r-1}\Big\{\frac{\partial \widehat{m}}{\partial x_{(r)}}x_{(i)}^{\widehat{m}}\ln x_{(r)} + \frac{\widehat{m}}{x_{(r)}} - \frac{\partial \widehat{m}}{\partial x_{(r)}}x_{(i)}^{\widehat{m}}\ln x_{(r)}\Big\}$$

$$= \frac{x_{(r)}^{\widehat{m}}}{\Big[\sum_{i=1}^{r} x_{(i)}^{\widehat{m}}\Big]^2}\sum_{i=1}^{r-1} x_{(i)}^{\widehat{m}}\Big\{-\frac{C(n,r,r)[\ln x_{(r)} - \ln x_{(i)}]}{\Big[\sum_{j=1}^{r} C(n,r,j)\ln x_{(j)}\Big]^2 x_{(r)}} + \frac{1}{x_{(r)}\sum_{j=1}^{r} C(n,r,j)\ln x_{(j)}}\Big\}$$

$$= \frac{\widehat{m}^2 x_{(r)}^{\widehat{m}-1}}{\Big[\sum_{i=1}^{r} x_{(i)}^{\widehat{m}}\Big]^2}\sum_{i=1}^{r-1} x_{(i)}^{\widehat{m}}\Big[\sum_{j=1}^{r-1} C(n,r,j)\ln x_{(j)} + C(n,r,r)\ln x_{(i)}\Big].$$

由于 $\dfrac{\widehat{m}^2 x_{(r)}^{\widehat{m}-1}}{[\sum_{i=1}^{r} x_{(i)}^{\widehat{m}}]^2} > 0$, 于是要证明 T^* 对 $x_{(r)}$ 是严格单调增, 只要证明:

$$\sum_{i=1}^{r-1} x_{(i)}^{\widehat{m}}\Big[\sum_{j=1}^{r-1} C(n,r,j)\ln x_{(j)} + C(n,r,r)\ln x_{(i)}\Big]$$
$$= \sum_{i=1}^{r-1} x_{(i)}^{\widehat{m}}\sum_{j=1}^{r-1} C(n,r,j)\big[\ln x_{(j)} - \ln x_{(i)}\big] > 0$$

即可.

(I) 对固定的 n 及 r, $r \le 5$.

(1) 固定 $1 \le i_0 \le k_0$, 对 $1 \le j_0 \le r-1$

①不妨设 $i_0 < j_0 < k_0$ 有:

$$x_{(i_0)}^{\widehat{m}} C(n,r,j_0)[\ln x_{(j_0)} - \ln x_{(i_0)}] + x_{j_0)}^{\widehat{m}} C(n,r,i_0)[\ln x_{(i_0)} - \ln x_{(j_0)}]$$
$$= x_{(i_0)}^{\widehat{m}}\Big\{C(n,r,j_0)[\ln x_{(j_0)} - \ln x_{(i_0)}] + \Big(\frac{x_{(j_0)}}{x_{(i_0)}}\Big)^{\widehat{m}}[-C(n,r,i_0)][\ln x_{(j_0)} - \ln x_{(i_0)}]\Big\}$$

$$=x_{(i_0)}^{\widehat{m}}[\ln x_{(j_0)}-\ln x_{(i_0)}]\Big\{C(n,r,j_0)+\Big(\frac{x_{(j_0)}}{x_{(i_0)}}\Big)^{\widehat{m}}[-C(n,r,i_0)]\Big\}$$

$$>x_{(i_0)}^{\widehat{m}}[\ln x_{(j_0)}-\ln x_{(i_0)}]\,[C(n,r,j_0)-C(n,r,i_0)]>0.$$

② $j_0\geq k_0+1$ 有： $x_{(i_0)}^{\widehat{m}}C(n,r,j_0)[\ln x_{(j_0)}-\ln x_{(i_0)}]>0.$

(2) 固定 $i_0\geq k_0+1$, 于是有：

$$x_{(i_0)}^{\widehat{m}}\Big\{\sum_{j=1}^{i_0-1}[-C(n,r,j)]\,[\ln x_{(i_0)}-\ln x_{(j)}]+\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big\}$$

$$=x_{(i_0)}^{\widehat{m}}\Big\{\sum_{j=1}^{k_0-1}[-C(n,r,j)[\ln x_{(i_0)}-\ln x_{(j)}]-\sum_{j=k_0}^{i_0-1}C(n,r,j)\,[\ln x_{(i_0)}-\ln x_{(j)}]$$
$$+\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big\}$$

$$>x_{(i_0)}^{\widehat{m}}\Big\{\sum_{j=1}^{i_0-1}[-C(n,r,j)]\,[\ln x_{(i_0)}-\ln x_{(k_0)}]-\sum_{j=k_0}^{i_0-1}C(n,r,j)\,[\ln x_{(i_0)}-\ln x_{(k_0)}]$$
$$+\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big\}$$

$$=x_{(i_0)}^{\widehat{m}}\Big\{[\ln x_{(i_0)}-\ln x_{(k_0)}]\sum_{j=1}^{i_0-1}[-C(n,r,j)]+\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big\}$$

$$>0.$$

$\Big($ 因为 $\sum_{j=1}^{r-1}[-C(n,r,j)]=C(n,r,r)$, 所以 $\sum_{j=1}^{i_0-1}[-C(n,r,j)]>\sum_{j=i_0}^{r-1}C(n,r,j)>0.\Big)$

(Ⅱ) 对固定的 n 及 $r\geq 6$.

(1) 固定 $1\leq i_0\leq k_0^*-1$, 对 $i_0\leq j\leq k_0$

$$x_{(i_0)}^{\widehat{m}}\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]-\sum_{j=k_0+1}^{k_0}x_{(j)}^{\widehat{m}}\,C(n,r,i_0)\,[\ln x_{(j)}-\ln x_{(i_0)}]$$

$$=x_{(i_0)}^{\widehat{m}}\Big[\sum_{j=i_0+1}^{k_0}C(n,r,j)[\ln x_{(j)}-\ln x_{(i_0)}]+\sum_{j=k_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]$$
$$+\sum_{j=i_0+1}^{k_0}\Big(\frac{x_{(j)}}{x_{(i_0)}}\Big)^{\widehat{m}}[-C(n,r,i_0)]\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big]$$

$$>x_{(i_0)}^{\widehat{m}}\Big[\sum_{j=i_0+1}^{k_0}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]+\sum_{j=k_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]$$
$$+\sum_{j=i_0+1}^{k_0}[-C(n,r,i_0)]\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big]$$

$$=x_{(i_0)}^{\widehat{m}}\Big[\sum_{j=i_0+1}^{k_0^*}[C(n,r,j)-C(n,r,i_0)]\,[\ln x_{(j)}-\ln x_{(i_0)}]$$

$$+\sum_{j=k_0^*+1}^{r-1}[C(n,r,j)-C(n,r,i_0)]\,[\ln x_{(j)}-\ln x_{(i_0)}]+\sum_{j=k_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]\Big].$$

易知：存在 $K\geq k_0^*+1$, 有 $C(n,r,K-1)<C(n,r,i_0)$ 但 $C(n,r,K)>C(n,r,i_0)$, 于是

$$x_{(i_0)}^{\widehat{m}}\sum_{j=i_0+1}^{r-1}C(n,r,j)\,[\ln x_{(j)}-\ln x_{(i_0)}]-\sum_{j=i_0+1}^{k_0}x_{(j)}^{\widehat{m}}C(n,r,i_0)\,[\ln x_{(j)}-\ln x_{(i_0)}]$$

$$>x_{(i_0)}^{\widehat{m}}\Big[\sum_{j=i_0+1}^{k_0^*}[C(n,r,j)-C(n,r,i_0)]\,[\ln x_{(j)}-\ln x_{(i_0)}]$$

$$+\sum_{j=k_0^*+1}^{K}[C(n,r,j)-C(n,r,i_0)]\,[\ln x_{(K)}-\ln x_{(i_0)}]$$

$$+\sum_{j=K+1}^{k_0}[C(n,r,j)-C(n,r,i_0)]\,[\ln x_{(K)}-\ln x_{(i_0)}]+\sum_{j=k_0+1}^{r-1}C(n,r,j)\,[\ln x_{(K)}-\ln x_{(i_0)}]\Big]$$

$$=x_{(i_0)}^{\widehat{m}}\,[\ln x_{(K)}-\ln x_{(i_0)}]\Big[\sum_{j=i_0+1}^{k_0^*}[C(n,r,j)-C(n,r,i_0)]+\sum_{j=k_0^*+1}^{K}[C(n,r,j)-C(n,r,i_0)]$$

$$+\sum_{j=K+1}^{k_0}[C(n,r,j)-C(n,r,i_0)]+\sum_{j=k_0+1}^{r-1}C(n,r,j)\Big]$$

$$=x_{(i_0)}^{\widehat{m}}\,[\ln x_{(K)}-\ln x_{(i_0)}]\Big[\sum_{j=i_0+1}^{k_0}C(n,r,j)-(k_0-i_0)C(n,r,i_0)+\sum_{j=k_0+1}^{r-1}C(n,r,j)\Big]$$

$$=x_{(i_0)}^{\widehat{m}}\,[\ln x_{(K)}-\ln x_{(i_0)}]\Big[\sum_{j=i_0+1}^{r-1}C(n,r,j)-(k_0-i_0)C(n,r,i_0)\Big]$$

$$=x_{(i_0)}^{\widehat{m}}\,[\ln x_{(K)}-\ln x_{(i_0)}]\Big[-C(n,r,r)-\sum_{j=1}^{i_0}C(n,r,j)-(k_0-i_0)C(n,r,i_0)\Big]>0.$$

(2) 固定 $k_0^*\leq i_0\leq k_0$, 易知即成为 (I) 中的 (1) 的情形.

(3) 固定 $i_0\geq k_0+1$, 即成为 (I) 中的 (2) 的情形.

综上： T^* 对 $x_{(r)}$ 是严格单调增的.

由定理 2 和定理 3 可知： T^* 的分布函数与 m,η 无关，仅与样本的大小 n 有关，故可以通过 Monte-Carlo 模拟方法，对于给定的 n 和 α 来获得 T^* 的临界值. 由 T^* 对 $x_{(r)}$ 单调增，可用统计量 T^* 来检验 $x_{(r)}$ 是否是特大异常值. 对于 $n\leq 25$ 的情况，用 σ 的 BLUE 和 BLIE 来作为 $\hat{\sigma}$, 显然这两个均是不变估计，对于 $n>25$, 可用 σ 的 GLUE, GLIE 来作为 $\hat{\sigma}$, 这三个估计也均是不变估计.

本文对 $n=4(1)25$, 用 Monte-Carlo 方法模拟 10000 次，给出了 $\hat{\sigma}$ 为 BLUE 和 BLIE 时，对 $\alpha=0.99,0.95,0.90,0.10,0.05,0.01$ 情况下 T^* 的分布的分位点表，列于表 2, 表 3. 计算 $\hat{\sigma}$ 的 BLUE 和 BLIE 的系数均可在 [4] 中查到.

4 检验的功效

极值分布或 Weibull 分布异常数据检验，相当于假设检验：

$H_0: x_1, x_2, \cdots, x_n$ 独立同分布，x_1 服从极值分布 $F(x) = e^{-e^{\frac{x-\mu}{\sigma}}}$，$-\infty < x$，$\mu < +\infty$，$\sigma > 0$.

$H_1: x_1, \cdots, x_n$ 中，$n-1$ 个服从参数为 (μ, σ) 的极值分布，另一个服从参数为 $(\mu+\lambda, \sigma)$ $(\lambda \neq 0, \lambda > 0)$ 的极值分布.

为了比较上述三种类型统计量在异常数据检验中的效果，不妨将备择假设改为：

$H_1: x_1, x_2, \cdots, x_n$ 中，x_1 服从参数为 $(\mu+\lambda, \sigma)$ $(\lambda \neq 0, \lambda > 0)$ 的极值分布，$x_2, \cdots, x_n$ 服从参数为 (μ, σ) 的极值分布.

这种修改不会对比较结果产生不良的影响，以下只对 $\lambda > 0$ 的情况进行讨论，至于 $\lambda < 0$ 可类似地加以讨论.

作为比较统计量在异常数据检验中效果的衡量标准，最易想到的是用势函数，但是势函数并不能完全反映检验出异常值 x_1 的效率. [5] 针对正态分布的情况，在与前面类似的假设和备择假设下，提出对于 $V = \max V_j$ $(1 \le j \le n)$ 形式的统计量，可以用下列五个概率作为衡量统计量在异常数据检验中效果的标准：

1. $P_1 = Pr\{V > V_\alpha \mid H_1\}$.
2. $p_2 = Pr\{V_1 > V_\alpha \mid H_1\}$.
3. $p_3 = Pr\{V_1 > V_\alpha, x_1 > x_2, \cdots, x_n \mid H_1\}$.
4. $p_4 = Pr\{V_1 > V_\alpha, V_2, \cdots, V_n < V_\alpha \mid H_1\}$.
5. $p_5 = Pr\{V_1 > V_\alpha \mid x_1 > x_2, \cdots, x_n, H_1\}$

这里 V_α 是统计量 V 的 α 上侧分位点.

这五个概率比较全面地反映了在控制犯第一类错误概率的条件下，统计量能检验出异常数据的效果. 因此，只要所选的统计量是 $\max V_j$ $(1 \le j \le n)$ 的形式，即可用上述五个概率作为衡量其检测异常数据的效果. 统计量 $D_1, D_2, G_1, G_2, T_1, T_2$ 皆可以写成这种形式. 记

$$D_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}} = \max_{1 \le j \le n} D_1^j, \qquad D_1^j = \frac{x_{(j)} - x_{(n-1)}}{x_{(n)} - x_{(1)}},$$

$$D_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} = \max_{1 \le j \le n} D_2^j, \qquad D_2^j = \frac{x_{(j)} - x_{(n-2)}}{x_{(n)} - x_{(1)}},$$

$$D_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}} = \max_{1 \le j \le n} D_2^j, \qquad D_2^j = \frac{x_{(j)} - x_{(n-2)}}{x_{(n)} - x_{(1)}},$$

$$G_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - \frac{1}{n-1}\sum_{i=1}^{n-1} x_{(i)}} = \max_{1 \le j \le n} G_1^j, \qquad G_1^j = \frac{x_{(j)} - x_{(n-1)}}{x_{(n)} - \frac{1}{n-1}\sum_{i=1}^{n-1} x_{(i)}},$$

$$G_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - \frac{1}{n-2}\sum_{i=1}^{n-2} x_{(i)}} = \max_{1 \le j \le n} G_2^j, \qquad G_2^j = \frac{x_{(j)} - x_{(n-2)}}{x_{(n)} - \frac{1}{n-2}\sum_{i=1}^{n-2} x_{(i)}},$$

$$T_1 = \frac{\widehat{y}_{(n)}^m}{\sum_{i=1}^{n} \widetilde{y}_{(i)}^m} = \max_{1 \le j \le n} T_1^j, \qquad T_1^j = \frac{\widetilde{y}_{(j)}^m}{\sum_{i=1}^{n} \widehat{y}_{(i)}^m},$$

$$T_2 = \frac{\widetilde{y}_{(n)}^m}{\sum_{i=1}^{n} \widetilde{y}_{(i)}^n} = \max_{1 \le j \le n} T_2^j, \qquad T_2^j = \frac{\widetilde{y}_{(j)}^m}{\sum_{i=1}^{n} \widetilde{y}_{(i)}^m},$$

其中 $x_1, x_2, \cdots, x_n$ 为来自于极值分布的样本容量为 n 的前 n 个次序统计量，$y_1, y_2, \cdots, y_n$ 为来自于两参数 Weibull 分布的样本容量为 n 的前 n 个次序统计量.

在极值分布和 Weibull 分布場合，对 $n = 4(1)25$, $\lambda = 1, 2, 5$, $\alpha = 0.10, 0.05, 0.01$ 用 Monte-Carlo 方法作了模拟，对每组 n, λ, α 重复 10000 次，从而计算出六个检验统计量的 p_1 至 p_5 的数值.

现将计算的部分结果列于表 4- 表 7, 从随机模拟的结果看出：

(1) 对于 p_1, p_2, p_3, p_4, p_5, 不论 n 的大小， F 型检验一致地好.

(2) 对于 p_1, p_2, p_3, p_4, p_5, 不论 n 的大小， Dixon 型和 G 型检验效果都还可以，有时 Dixon 型检验效果好些，有时 G 型检验效果好些.

(3) 随着异常参数 λ 的增长，所有检验统计量的剔除效率都在增长.

(4) 在检验仅限于只有一个异常值的情况下， D_1 和 G_1 明显地比 D_2 和 G_2 的效果好.

(5) F 型检验中， T_1 和 T_2 的差异较小.

5 结束语

通过对于 Weibull 和极值分布异常数据检验方法及其有效性的研究，可以看到本文提出的一个新的方法：在文献 [5] 提出的检验异常数据效果的标准下，通过数值模拟表明，在极值分布和 Weibull 分布异常数据检验中， F 型检验的效果比 Dixon 型和 G 型检验一致地好. 理论研究表明，这个检验还可以较好地避免屏蔽效应 (另文讨论). 本文给出几种检验的部分分位点表，便于实际应用.

表 1　G_2 分布的分位点

n \ α	0.99	0.95	0.90	0.10	0.05	0.01
4	0.996	0.976	0.953	0.388	0.289	0.131
5	0.952	0.892	0.841	0.283	0.210	0.095
6	0.891	0.807	0.752	0.227	0.164	0.076
7	0.840	0.750	0.690	0.199	0.144	0.064
8	0.801	0.697	0.641	0.177	0.125	0.055
9	0.764	0.659	0.601	0.158	0.113	0.052
10	0.730	0.622	0.567	0.150	0.108	0.051
11	0.692	0.598	0.547	0.140	0.099	0.041
12	0.666	0.576	0.522	0.135	0.093	0.038
13	0.646	0.558	0.503	0.124	0.084	0.036
14	0.634	0.538	0.486	0.118	0.082	0.036
15	0.609	0.525	0.470	0.114	0.080	0.036
16	0.595	0.508	0.459	0.107	0.075	0.034
17	0.585	0.501	0.449	0.107	0.074	0.033
18	0.578	0.490	0.439	0.102	0.070	0.033
19	0.557	0.476	0.429	0.100	0.070	0.030
20	0.555	0.466	0.421	0.097	0.066	0.028
21	0.540	0.454	0.412	0.094	0.064	0.028
22	0.526	0.447	0.408	0.092	0.064	0.028
23	0.522	0.443	0.400	0.090	0.062	0.026
24	0.516	0.436	0.392	0.087	0.061	0.025
25	0.504	0.428	0.385	0.086	0.060	0.025

表 2 T^* 分布的分位点 ($\hat{\sigma}$ 是 σ 的BLUE)

n \ α	0.99	0.95	0.90	0.10	0.05	0.01
4	0.6567	0.6400	0.6251	0.4399	0.4166	0.3751
5	0.6118	0.5811	0.5581	0.3757	0.3557	0.3228
6	0.5731	0.5283	0.5040	0.3306	0.3119	0.2820
7	0.5304	0.4837	0.4569	0.2942	0.2802	0.2527
8	0.4995	0.4542	0.4244	0.2693	0.2557	0.2330
9	0.4735	0.4222	0.3950	0.2471	0.2346	0.2128
10	0.4459	0.3943	0.3671	0.2300	0.2179	0.1982
11	0.4259	0.3714	0.3416	0.2149	0.2040	0.1852
12	0.4064	0.3509	0.3263	0.2011	0.1912	0.1752
13	0.3832	0.3349	0.3100	0.1901	0.1810	0.1660
14	0.3667	0.3186	0.2958	0.1807	0.1714	0.1585
15	0.3466	0.3023	0.2813	0.1719	0.1628	0.1508
16	0.3396	0.2888	0.2663	0.1633	0.1549	0.1410
17	0.3262	0.2777	0.2548	0.1564	0.1487	0.1353
18	0.3168	0.2679	0.2469	0.1504	0.1419	0.1297
19	0.3060	0.2599	0.2373	0.1439	0.1364	0.1245
20	0.2987	0.2500	0.2299	0.1387	0.1321	0.1204
21	0.2847	0.2428	0.2216	0.1335	0.1267	0.1158
22	0.2739	0.2329	0.2134	0.1296	0.1228	0.1125
23	0.2710	0.2273	0.2068	0.1251	0.1190	0.1087
24	0.2596	0.2183	0.2000	0.1217	0.1152	0.1051
25	0.2518	0.2133	0.1946	0.1176	0.1116	0.1022

表 3 T^* 分布的分位点 ($\hat{\sigma}$ 是 σ 的BLUE)

n \ α	0.99	0.95	0.90	0.10	0.05	0.01
4	0.7389	0.7199	0.7022	0.4646	0.4376	0.3873
5	0.6806	0.6453	0.6128	0.3957	0.3720	0.3341
6	0.6311	0.5803	0.5518	0.3462	0.3248	0.2915
7	0.5792	0.5263	0.4956	0.3072	0.2909	0.2608
8	0.5416	0.4908	0.4569	0.2801	0.2648	0.2398
9	0.5104	0.4532	0.4228	0.2562	0.2422	0.2187
10	0.4782	0.4212	0.3910	0.2381	0.2250	0.2032
11	0.4547	0.3945	0.3668	0.2219	0.2099	0.1897
12	0.4324	0.3717	0.3447	0.2075	0.1966	0.1796
13	0.4063	0.3533	0.3262	0.1956	0.1858	0.1699
14	0.3875	0.3353	0.3106	0.1856	0.1759	0.1629
15	0.3650	0.3170	0.2942	0.1763	0.1668	0.1539
16	0.3568	0.3021	0.2780	0.1674	0.1586	0.1439
17	0.3420	0.2900	0.2654	0.1601	0.1519	0.1378
18	0.3315	0.2790	0.2566	0.1538	0.1449	0.1323
19	0.3196	0.2704	0.2456	0.1471	0.1392	0.1267
20	0.3116	0.2599	0.2384	0.1417	0.1347	0.1225
21	0.2965	0.2521	0.2295	0.1363	0.1291	0.1176
22	0.2848	0.2413	0.2207	0.1322	0.1251	0.1143
23	0.2815	0.2353	0.2135	0.1276	0.1212	0.1104
24	0.2693	0.2257	0.2063	0.1240	0.1172	0.1067
25	0.2609	0.2202	0.2004	0.1198	0.1134	0.1037

表 4

n=8		P_1	P_2	P_3	P_4	P_5
	D_1	0.1509	0.1025	0.1025	0.1025	0.241517
	D_2	0.1394	0.0916	0.0888	0.0844	0.209236
α=0.1	G_1	0.1463	0.0985	0.0985	0.0985	0.232092
λ=1	G_2	0.1373	0.0864	0.0851	0.0826	0.200518
	T_1	0.8724	0.3965	0.3890	0.3853	0.916588
	T_2	0.8508	0.3858	0.3824	0.3817	0.901037
	D_1	0.0856	0.0614	0.0614	0.0614	0.144675
	D_2	0.0757	0.0506	0.0498	0.0477	0.117342
α=0.005	G_1	0.0819	0.0577	0.0577	0.0577	0.135957
λ=1	G_2	0.0729	0.0497	0.0494	0.0485	0.116400
	T_1	0.8186	0.3751	0.3720	0.3717	0.876532
	T_2	0.7849	0.3631	0.3615	0.3631	0.851791
	D_1	0.0179	0.0133	0.0133	0.0133	0.031338
	D_2	0.0156	0.0106	0.0106	0.0102	0.024976
α=0.01	G_1	0.0199	0.0150	0.0150	0.0150	0.035344
λ=1	G_2	0.0151	0.0108	0.0108	0.0106	0.025448
	T_1	0.7220	0.3367	0.3351	0.3367	0.789585
	T_2	0.6848	0.3230	0.3215	0.3230	0.757540

表 5

n=9		P_1	P_2	P_3	P_4	P_5
	D_1	0.1550	0.1013	0.1013	0.1013	0.252555
	D_2	0.1431	0.0917	0.0899	0.0863	0.224134
α=0.1	G_1	0.1511	0.0992	0.0922	0.0922	0.247320
λ=1	G_2	0.1438	0.0930	0.0917	0.0899	0.228621
	T_1	0.8951	0.3833	0.3734	0.3642	0.930940
	T_2	0.8783	0.3748	0.3693	0.3648	0.920718
	D_1	0.0837	0.0589	0.0589	0.0589	0.146846
	D_2	0.0745	0.0510	0.0498	0.0485	0.124158
α=0.05	G_1	0.0833	0.0584	0.0584	0.0584	0.145600
λ=1	T_1	0.0795	0.0540	0.0533	0.0524	0.132884
	T_2	0.8514	0.3644	0.3598	0.3558	0.897033
	T_2	0.8280	0.3569	0.3549	0.3535	0.884817
	D_1	0.0215	0.0163	0.0163	0.0163	0.040638
	D_2	0.0153	0.0107	0.0106	0.0104	0.040638
α=0.01	G_1	0.0244	0.0185	0.0185	0.0185	0.046123
λ=1	G_2	0.0173	0.0131	0.0131	0.0130	0.032660
	T_1	0.7496	0.3289	0.3280	0.3284	0.817751
	T_2	0.7153	0.3163	0.3157	0.3163	0.787086

表 6

$n=8$		P_1	P_2	P_3	P_4	P_5
$\alpha=0.1$ $\lambda=5$	D_1	0.9258	0.9247	0.9247	0.9247	0.942321
	D_2	0.8814	0.8798	0.8783	0.8783	0.896326
	G_1	0.9208	0.9201	0.9201	0.9201	0.937643
	G_2	0.8833	0.8819	0.8818	0.8816	0.898604
	T_1	0.9928	0.9794	0.9791	0.9788	0.997758
	T_2	0.9921	0.9789	0.9788	0.9786	0.997452
$\alpha=0.05$ $\lambda=5$	D_1	0.8775	0.8771	0.8771	0.8771	0.893814
	D_2	0.7777	0.7769	0.7768	0.7768	0.791603
	G_1	0.8696	0.8693	0.8693	0.8693	0.885866
	G_2	0.7891	0.7884	0.7883	0.7883	0.803322
	T_1	0.9889	0.9772	0.9771	0.9770	0.995720
	T_2	0.9878	0.9769	0.9769	0.9769	0.995516
$\alpha=0.01$ $\lambda=5$	D_1	0.6302	0.6302	0.6302	0.6302	0.642209
	D_2	0.4346	0.4343	0.4343	0.4343	0.442576
	G_1	0.6410	0.6409	0.6409	0.6409	0.653113
	G_2	0.4201	0.4198	0.4198	0.4198	0.427800
	T_1	0.9813	0.9720	0.9720	0.9720	0.990523
	T_2	0.9793	0.9704	0.9704	0.9704	0.98889

表 7

$n=9$		P_1	P_2	P_3	P_4	P_5
$\alpha=0.1$ $\lambda=5$	D_1	0.9312	0.9299	0.9299	0.9299	0.947041
	D_2	0.9030	0.9015	0.9013	0.9005	0.917914
	G_1	0.9290	0.9281	0.9281	0.9281	0.945208
	G_2	0.9093	0.9081	0.9080	0.9077	0.924738
	T_1	0.9944	0.9806	0.9799	0.9787	0.997963
	T_2	0.9933	0.9796	0.9794	0.9786	0.997454
$\alpha=0.05$ $\lambda=5$	D_1	0.8922	0.8919	0.8919	0.8919	0.908341
	D_2	0.8257	0.8251	0.8251	0.8249	0.840310
	G_1	0.8869	0.8865	0.8865	0.8865	0.902841
	G_2	0.8376	0.8372	0.8372	0.8372	0.852633
	T_1	0.9919	0.9788	0.9787	0.9781	0.996741
	T_2	0.9908	0.9782	0.9781	0.9778	0.996130
$\alpha=0.01$ $\lambda=5$	D_1	0.7147	0.7146	0.7146	0.7146	0.727773
	D_2	0.5016	0.5015	0.5015	0.5015	0.510744
	G_1	0.7186	0.7185	0.7185	0.7185	0.731744
	G_2	0.5270	0.5269	0.5269	0.5269	0.536613
	T_1	0.9853	0.9745	0.9745	0.9744	0.992464
	T_2	0.9834	0.9730	0.9730	0.9730	0.990936

致谢 感谢审稿人提出的宝贵意见！

参 考 文 献

[1] 马逢时，许其洲. 极值分布异常数据检验. 数理统计与应用概率, 1986, 1 (1): 81–91.
[2] 陈振民. $G(\frac{x-\mu}{\sigma})$ 型分布样本中异常值的统计检验. 上海师范大学学报 (自然科学版), 1987, (3): 13–18.
[3] 费鹤良，徐锦龙，陈振民. 指数分布样本中异常数据检验的有效性. 应用概率统计, 1989, 5 (4): 289–294.
[4] 中国电子技术标准化研究所编著. 可靠性试验用表 (增订本). 北京：国防工业出版社， 1987.
[5] H.A. David. Order Statistics. New York: John Wiley & Sons, Inc., 1981.

THE PROCEDURES OF TESTING OUTLYING OBSERVATIONS WEIBULL OR EXTREME-VALUE DISTRIBUTION

FEI HELIANG LU XIANGWEI XU XIAOLING

(*Department of Mathematics, Shanghai Normal University, Shanghai* 200234)

Abstract In this paper, we discuss the outlier procedures on test for Weibull or Extreme-value distribution. The exact distribution of the Dixon statistic is derived. We present a new F-type statistic,and give its percentile values in Weibull or Extreme-value distribution case. We also compare some methods for testing outliers in Weibull or Extreme-value sampling by Monte-Carlo method, its is shown that the F-type procedure is the best.

Key words Outlier, test, effectiveness, extreme-value distribution, Weibull distribution