

离群值检测

离群值、四分位数与盒子图

- 均值的意义与局限性
- 例 姚明给一群小学生讲NBA经历
 - 均值（平均身高）没有意义
 - 中位数(median, 仲数)的定义 $x_1 \leq x_2 \leq \cdots \leq x_n$
 - ✓ n 为奇数 $x_{(n+1)/2}$
 - ✓ n 为偶数 $(x_{n/2-1} + x_{n/2+1})/2$
 - MATLAB求解 `median(x)`
- 离群值 (outliers)

四分位数的定义

- 由中位数将向量 x 分成两部分
- 这两部分各自的中位数称为四分位数(quantile)
 - 第1四分位数 q_1
 - 第2四分位数 q_2 ——中位数 (仲数)
 - 第3四分位数 q_3
- 四分位数向量 $q = [q_1, q_2, q_3]$
- MATLAB求解 $q = \text{quantile}(x, 3)$

四分位距与离群值

- 四分位距(interquartile range , IQR)的定义
 - 第3四分位数与第1四分位数的差 $IQR = q_3 - q_1$
- 离群值的定义
 - 正常值的区间
 $(q_1 - 1.5IQR, q_3 + 1.5IQR)$
 - 正常值区间之外的都是离群值

盒子图的绘制与解释

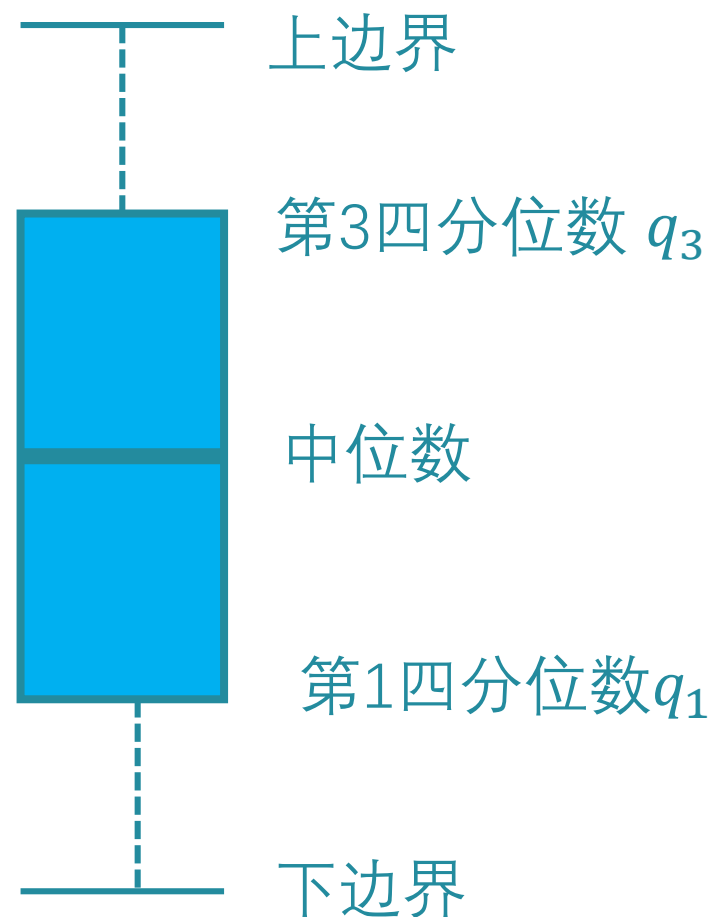
➤ 盒子图的绘制

`boxplot(x)`

➤ 上边界与下边界

$(q_1 - 1.5\text{IQR}, q_3 + 1.5\text{IQR})$

➤ 离群值



例9-26 盒子图的绘制

➤ 一组数据，存于c9dlamp.dat文件

1067	919	1196	785	1126	936	918	1156	920	948	855	1092	1162	1170	929
950	905	972	1035	1045	1157	1195	1195	1340	1122	938	970	1237	956	1102
1022	978	832	1009	1157	1151	1009	765	958	902	923	1333	811	1217	1085
896	958	1311	1037	702	521	933	928	1153	946	858	1071	1069	830	1063
930	807	954	1063	1002	909	1077	1021	1062	1157	999	932	1035	944	1049
940	1122	1115	833	1320	901	1324	818	1250	1203	1078	890	1303	1011	1102
996	780	900	1106	704	621	854	1178	1138	951	1187	1067	1118	1037	958
760	1101	949	992	966	824	653	980	935	878	934	910	1058	730	980
844	814	1103	1000	788	1143	935	1069	1170	1067	1037	1151	863	990	1035
1112	931	970	932	904	1026	1147	883	867	990	1258	1192	922	1150	1091
1039	1083	1040	1289	699	1083	880	1029	658	912	1023	984	856	924	801
1122	1292	1116	880	1173	1134	932	938	1078	1180	1106	1184	954	824	529
998	996	1133	765	775	1105	1081	1171	705	1425	610	916	1001	895	709
610	916	1001	895	709	860	1110	1149	972	1002					

➤ 四分位数求值  >> A=load('c9dlamp.dat');

q=quantile(A,3)

➤ 盒子图绘制



>> boxplot(A)



>> sort(A')

离群值的检测

➤ 单变量数据

- Rutgers大学 Niccolo Battistini编写的outliers() 函数

$[v_1, v_2] = \text{outliers}(v, \text{opts}, \alpha)$

- 选项 'grubbs', 'quartile'

- 显著性水平 α

➤ 多变量数据

- Antonio Trujillo-Ortiz编写的函数 moutlier1()

$\text{moutlier1}(X, \alpha)$

例9-27 单变量离群值

➤ 数据表格，存于c9dlamp.dat文件

1067	919	1196	785	1126	936	918	1156	920	948	855	1092	1162	1170	929
950	905	972	1035	1045	1157	1195	1195	1340	1122	938	970	1237	956	1102
1022	978	832	1009	1157	1151	1009	765	958	902	923	1333	811	1217	1085
896	958	1311	1037	702	521	933	928	1153	946	858	1071	1069	830	1063
930	807	954	1063	1002	909	1077	1021	1062	1157	999	932	1035	944	1049
940	1122	1115	833	1320	901	1324	818	1250	1203	1078	890	1303	1011	1102
996	780	900	1106	704	621	854	1178	1138	951	1187	1067	1118	1037	958
760	1101	949	992	966	824	653	980	935	878	934	910	1058	730	980
844	814	1103	1000	788	1143	935	1069	1170	1067	1037	1151	863	990	1035
1112	931	970	932	904	1026	1147	883	867	990	1258	1192	922	1150	1091
1039	1083	1040	1289	699	1083	880	1029	658	912	1023	984	856	924	801
1122	1292	1116	880	1173	1134	932	938	1078	1180	1106	1184	954	824	529
998	996	1133	765	775	1105	1081	1171	705	1425	610	916	1001	895	709
610	916	1001	895	709	860	1110	1149	972	1002					

➤ 离群值的直接检测



```
>> A=load('c9dlamp.dat');  
[v1 v2]=outliers(A,'grubbs',0.05)
```


例9-28 多变量离群值检测

➤NBA球队的数据


单位：百万美元

球队序号	球队价值	场馆价值	收入	球队序号	球队价值	场馆价值	收入
1	447	149	22.8	2	401	160	13.5
3	356	119	49	4	338	117	-17.7
5	328	109	2	6	290	97	25.6
7	284	102	23.5	8	283	105	18.5
9	282	109	21.5	10	280	94	10.1
11	278	82	15.2	12	275	102	-16.8
13	274	98	28.5	14	272	97	-85.1
15	258	72	3.8	16	249	96	10.6
17	244	94	-1.6	18	239	85	13.8
19	236	91	7.9	20	230	85	6.9
21	227	63	-19.7	22	218	75	7.9
23	216	80	21.9	24	208	72	15.9
25	202	78	-8.4	26	199	80	13.1
27	196	70	2.4	28	188	70	7.8
29	174	70	-15.1				

离群值检测

➤ MATLAB求解语句

➤ 数据读入



```
>> X=[447,149,22.8; 401,160,13.5; 356,119,49; 338,117,-17.7; 328,109,2;  
290,97,25.6; 284,102,23.5; 283,105,18.5; 282,109,21.5; 280,94,10.1;  
278,82,15.2; 275,102,-16.8; 274,98,28.5; 272,97,-85.1; 258,72,3.8;  
249,96,10.6; 244,94,-1.6; 239,85,13.8; 236,91,7.9; 230,85,6.9;  
227,63,-19.7; 218,75,7.9; 216,80,21.9; 208,72,15.9; 202,78,-8.4;  
199,80,13.1; 196,70,2.4; 188,70,7.8; 174,70,-15.1];
```

➤ 离群值检测



```
>> moutlier1(X,0.05)
```

检测结果的图形表示

- 由检测结果可知，第14队数据是离群值
- 三维散点图显示



```
>> plot3(X(:,1),X(:,2),X(:,3),'o')
```

- x - z 平面图



```
>> plot(X(:,1),X(:,3),'o')
```

- y - z 平面图



```
>> plot(X(:,2),X(:,3),'o')
```

