# 马的疝病数据集分析报告

1. 问题描述

疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。

2. 数据说明

共368个样本，27个特征。其中数值属性有7个：rectal temperature，pulse，respiratory rate，nasogastric reflux PH，packed cell volume，total protein，abdomcentesis total protein；

3. 数据分析过程

3.1 数据摘要

* 对标称属性，给出每个可能取值的频数：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **surgerynan** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 214 | 58.2 | 58.2 | 58.2 |
| 2 | 152 | 41.3 | 41.3 | 99.5 |
| nan | 2 | .5 | .5 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |
| **Age** | | | | | |
|  | | 频率 | 百分比 | 有效百分比 | 累积百分比 |
| 有效 | 1 | 340 | 92.4 | 92.4 | 92.4 |
| 9 | 28 | 7.6 | 7.6 | 100.0 |
| 合计 | 368 | 100.0 | 100.0 |  |

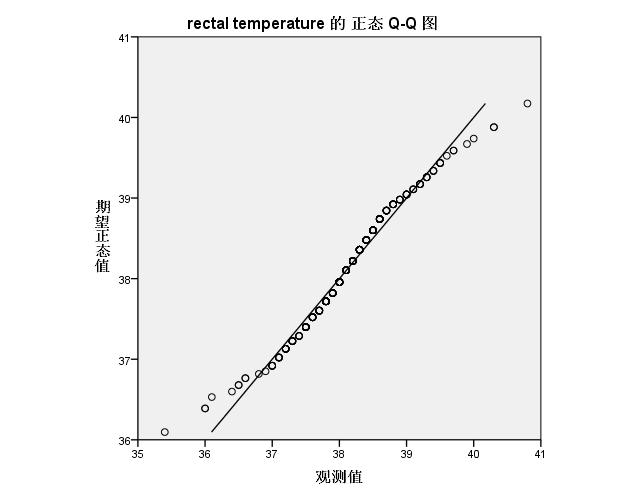
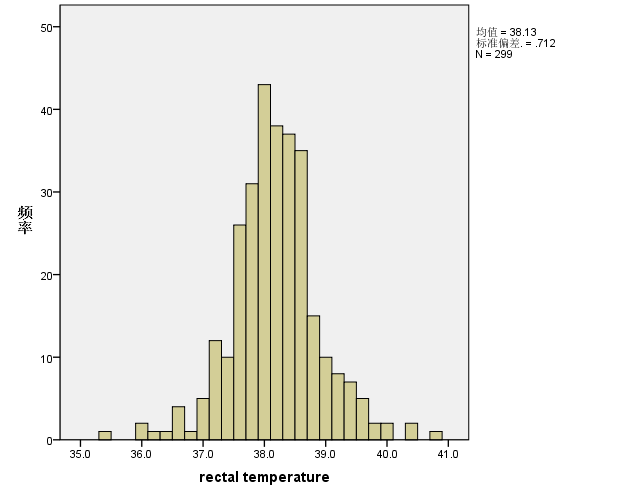
（注：以下省略，详见上传的数据完整版）

* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | rectal temperature | pulse | respiratory rate | nasogastric reflux PH |
| N | 有效 | 299 | 342 | 297 | 69 |
| 缺失 | 69 | 26 | 71 | 299 |
| 均值 | | 38.134 | 70.76 | 30.52 | 4.962 |
| 中位数 | | 38.100 | 60.00 | 28.00 | 5.400 |
| 最小值 | | 35.4 | 30 | 8 | 1.0 |
| 最大值 | | 40.8 | 184 | 96 | 8.5 |
| 百分位数 | 25 | 37.800 | 48.00 | 18.00 | 3.250 |
| 50 | 38.100 | 60.00 | 28.00 | 5.400 |
| 75 | 38.500 | 88.00 | 36.00 | 6.500 |

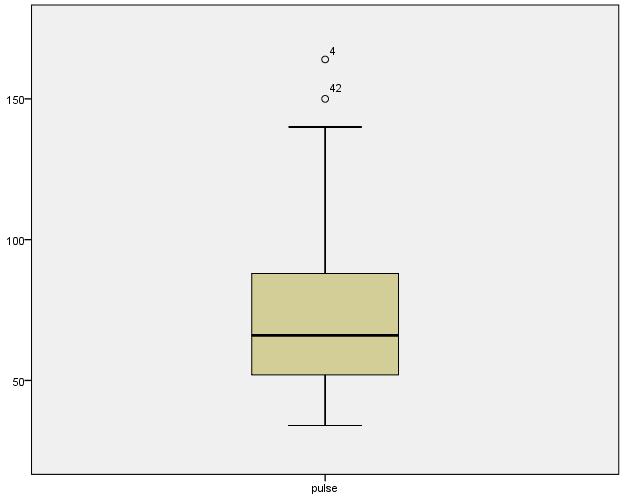
|  |  |  |
| --- | --- | --- |
| packed cell volume | total protein | abdomcentesis total protein |
| 331 | 325 | 133 |
| 37 | 43 | 235 |
| 45.66 | 24.771 | 2.948 |
| 44.00 | 7.500 | 2.100 |
| 4 | 3.3 | .1 |
| 75 | 89.0 | 10.1 |
| 37.00 | 6.500 | 1.900 |
| 44.00 | 7.500 | 2.100 |
| 52.00 | 58.000 | 3.900 |

3.2数据的可视化

针对数值属性：绘制直方图，用qq图检验其分布是否为正态分布（以“rectal temperature”属性为例）：

（由于数据点分布非常贴合正态分布，可认为其服从正态分布）

* 绘制盒图，对离群值进行识别

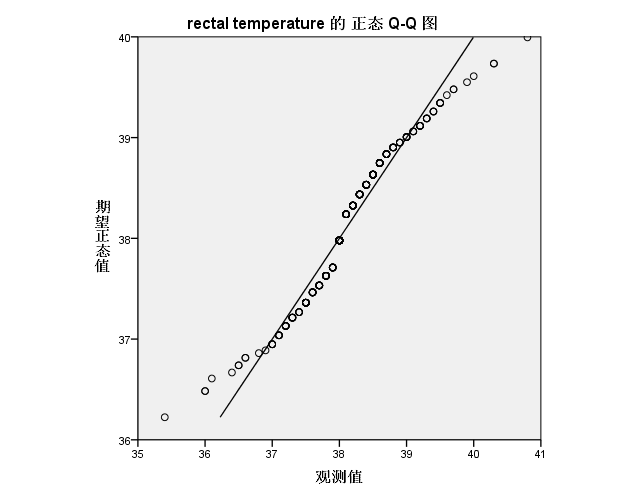
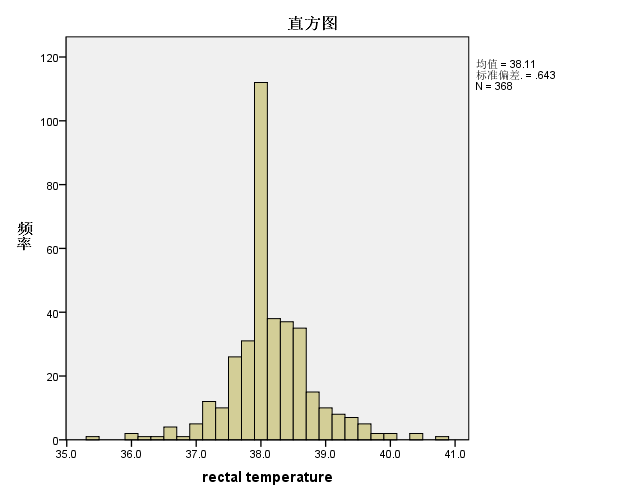


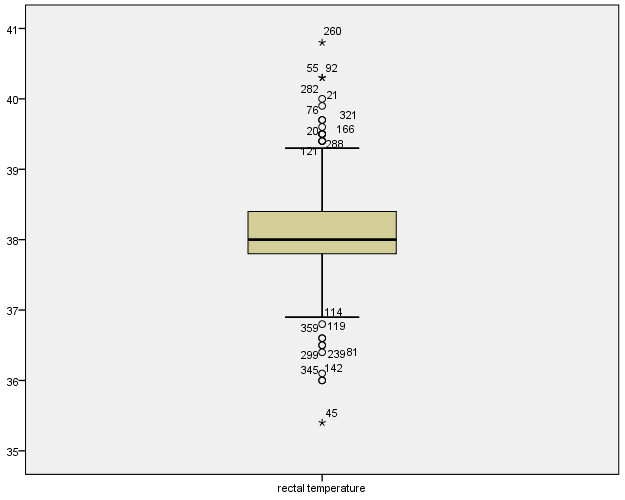
（其中可看出离群点有两个，对应数据的标号为4和42）

3.3 数据缺失的处理

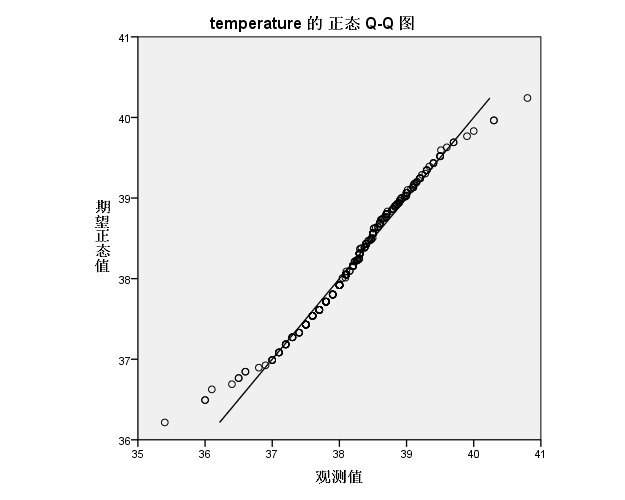
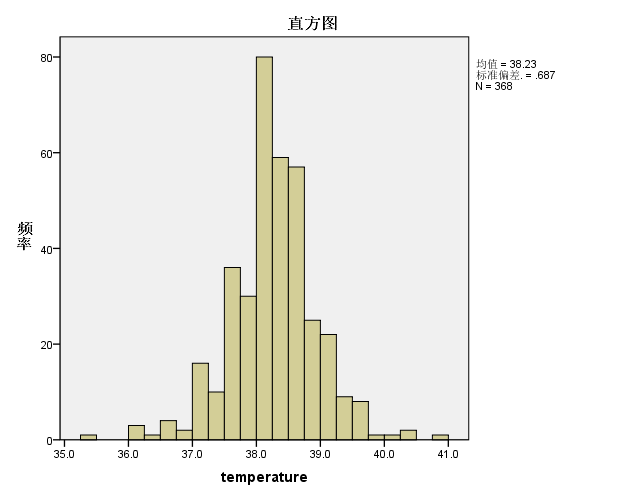
数据集中有30%的值是缺失的，分别使用下列四种策略对缺失值进行处理:

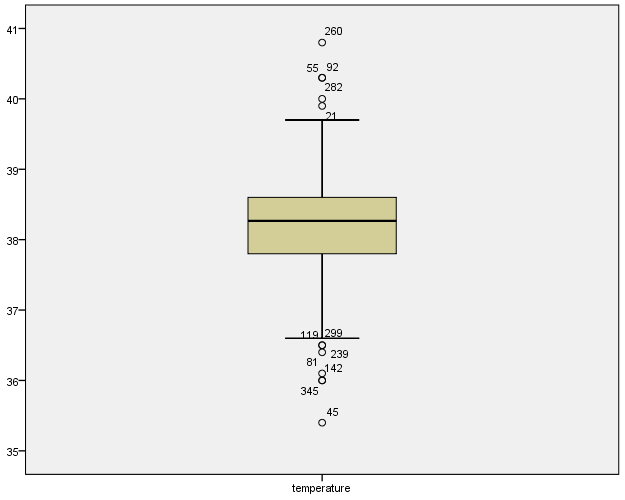
* 将缺失部分剔除（SPSS处理的时候默认忽略空值，即为上方所示图）
* 用最高频率值来填补缺失值（仍然以“rectal temperature”属性为例）





* 通过属性的相关关系来填补缺失值





* 通过数据对象之间的相似性来填补缺失值

