数据挖掘作业——马的疝病分析

一、 数据摘要及可视化

1. 数据摘要

该部分内容均在 python 下通过 pandas 库完成。

1.1 标称属性频次统计

在该数据集中包含以下标称属性：surgery、Age、Hospital Number、temperature of extremities、peripheral pulse、mucous membranes、capillary refill time、pain、peristalsis、abdominal distension、nasogastric tube、nasogastric reflux、rectal examination – feces、abdomen、abdominocentesis appearance、outcome、surgical lesion、cp_data。

其分别的词频统计为：

（1）Surgery：

```
1.0    214
2.0    152
Name: surgery, dtype: int64
```

（2）Age：

```
1    340
9     28
Name: Age, dtype: int64
```

（3）temperature of extremities：

```
3.0    135
1.0     95
2.0     39
4.0     34
Name: temperature of extremities, dtype: int64
```

(4) peripheral pulse：

```
1.0    151
3.0    116
4.0     12
2.0      6
Name: peripheral pulse, dtype: int64
```

(5) mucous membranes：

```
1.0    98
3.0    81
4.0    50
2.0    38
5.0    28
6.0    25
Name: mucous membranes, dtype: int64
```

(6) capillary refill time：

```
1.0    232
2.0     96
3.0      2
Name: capillary refill time, dtype: int64
```

(7) pain：

```
3.0    82
2.0    77
5.0    50
1.0    49
4.0    47
Name: pain, dtype: int64
```

(8) Peristalsis：

```
3.0    154
4.0     91
1.0     49
2.0     22
Name: peristalsis, dtype: int64
```

(9) abdominal distension：

```
1.0    101
3.0     85
2.0     75
4.0     42
Name: abdominal distension, dtype: int64
```

(10) nasogastric tube：

```
2.0    121
1.0     89
3.0     27
Name: nasogastric tube, dtype: int64
```

(11) nasogastric reflux :

```
1.0    141
3.0     49
2.0     45
Name: nasogastric reflux, dtype: int64
```

(12) rectal examination – feces :

```
4.0    97
1.0    68
3.0    61
2.0    14
Name: rectal examination - feces, dtype: int64
```

(13) abdomen :

```
5.0    96
4.0    55
1.0    31
2.0    24
3.0    19
Name: abdomen, dtype: int64
```

(14) abdominocentesis appearance :

```
2.0    62
3.0    60
1.0    52
Name: abdominocentesis appearance, dtype: int64
```

(15) outcome :

```
1.0    225
2.0     89
3.0     52
Name: outcome, dtype: int64
```

(16) surgical lesion :

```
1    232
2    136
Name: surgical lesion, dtype: int64
```

(17) cp_data :

```
2    244
1    124
Name: cp_ data, dtype: int64
```

1.2 数值型数据统计

在该数据集中包含以下数值属性：rectal temperature、pulse、respiratory rate、nasogastric reflux PH、packed cell volume、total protein、abdomcentesis total protein。

下面分别统计出数值属性的有效总数、平均值、标准差、最小值、二分位数、中位数、四分位数及最大值。

（1）rectal temperature：

```
count    299.000000
mean      38.134448
std        0.711684
min       35.400000
25%       37.800000
50%       38.100000
75%       38.500000
max       40.800000
Name: rectal temperature, dtype: float64
```

（2）pulse：

```
count    342.000000
mean      70.757310
std       28.089867
min       30.000000
25%       48.000000
50%       60.000000
75%       88.000000
max      184.000000
Name: pulse, dtype: float64
```

（3）respiratory rate：

```
count    297.000000
mean      30.521886
std       17.669651
min        8.000000
25%       18.000000
50%       28.000000
75%       36.000000
max       96.000000
Name: respiratory rate, dtype: float64
```

（4）nasogastric reflux PH：

```
count    69.000000
mean      4.962319
std       2.003901
min       1.000000
25%       3.500000
50%       5.400000
75%       6.500000
max       8.500000
Name: nasogastric reflux PH, dtype: float64
```

(5) packed cell volume:

```
count    331.000000
mean      45.656798
std       10.865663
min        4.000000
25%       37.250000
50%       44.000000
75%       52.000000
max       75.000000
Name:  packed cell volume, dtype: float64
```

(6) total protein:

```
count    325.000000
mean      24.771077
std       27.704880
min        3.300000
25%        6.500000
50%        7.500000
75%       58.000000
max       89.000000
Name: total protein, dtype: float64
```

(7) abdomcentesis total protein:

```
count    133.000000
mean       2.948120
std        1.927064
min        0.100000
25%        2.000000
50%        2.100000
75%        3.900000
max       10.100000
Name: abdomcentesis total protein, dtype: float64
```

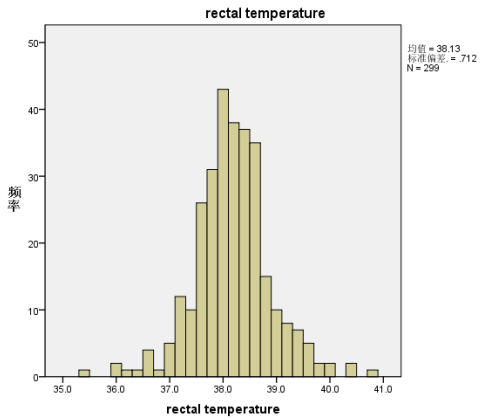
2. 数据可视化

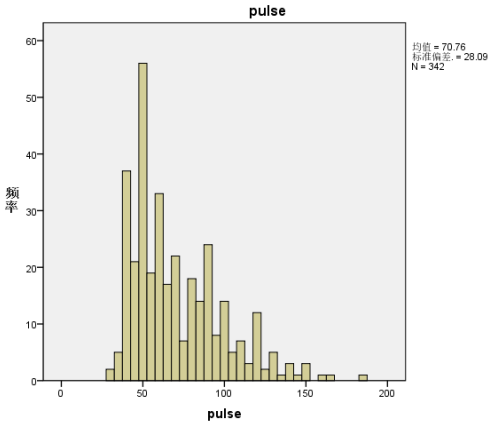
该部分内容通过 spss 工具完成。

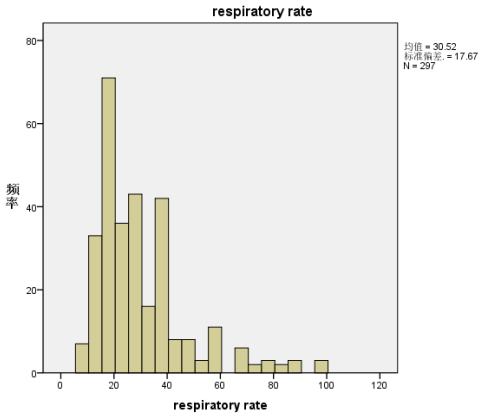
2.1 数值属性的直方图及 Q_Q 图

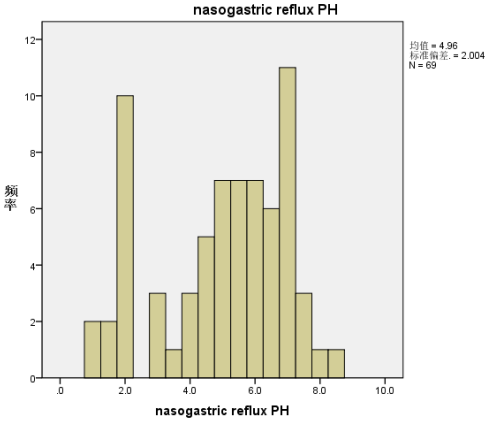
2.1.1 直方图

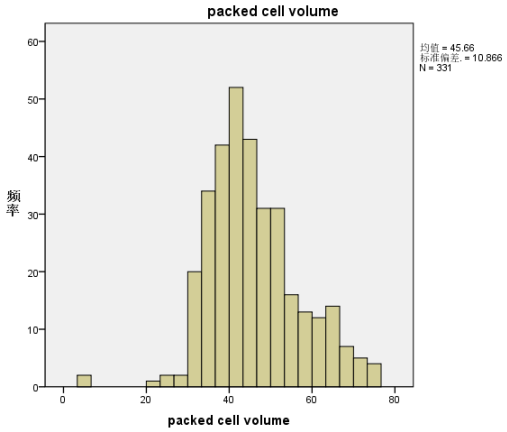
下面分别为数值数据的直方图：

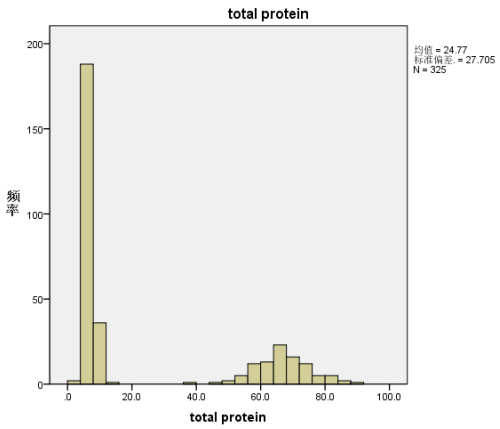


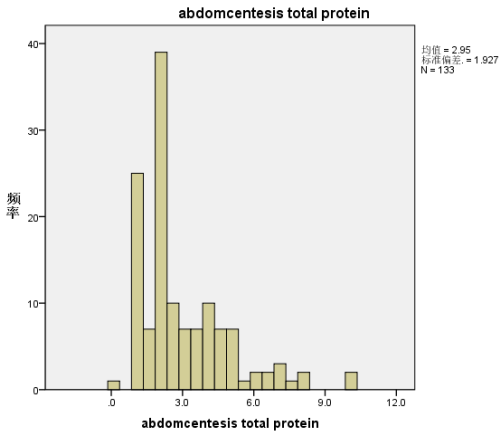






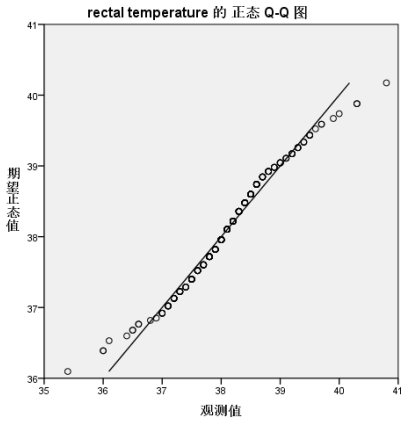


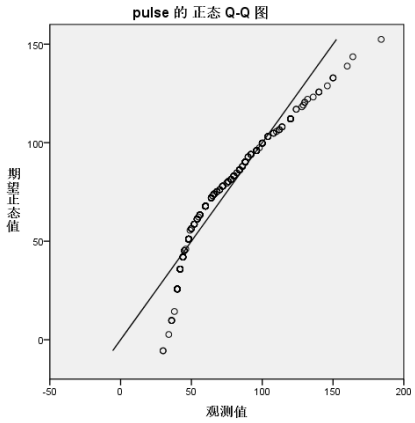


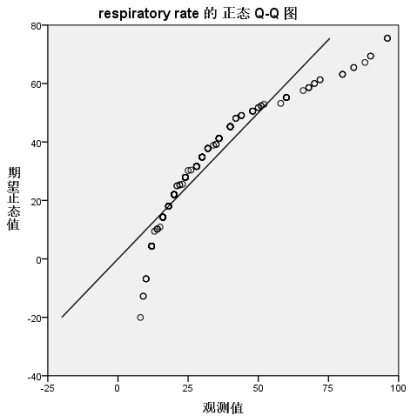


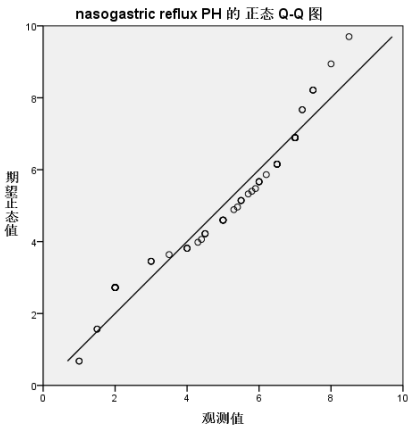
2.1.2 Q_Q 图

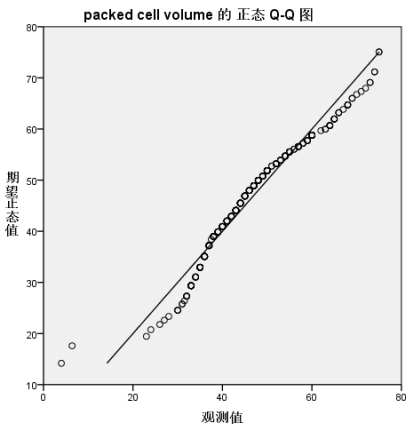
下面分别为数值数据的直方图：

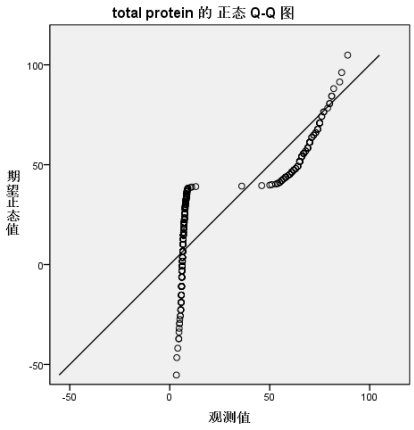


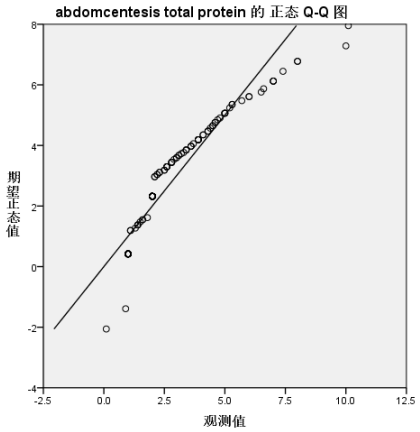






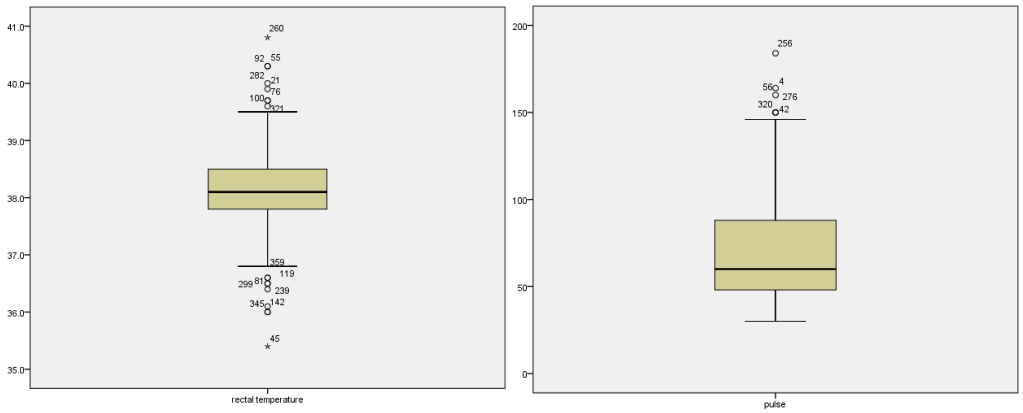


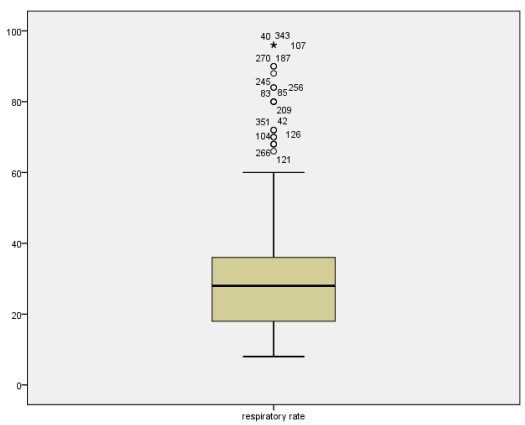


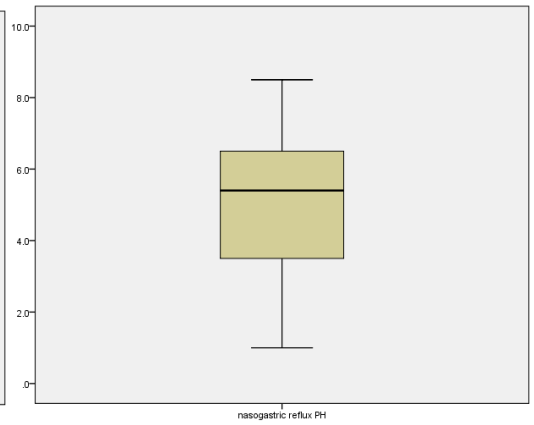


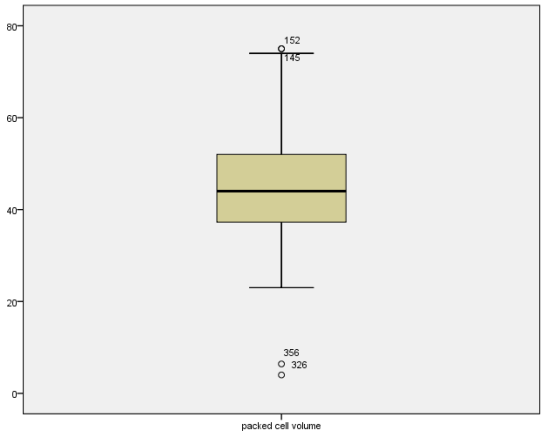
2.2 数值属性的盒图

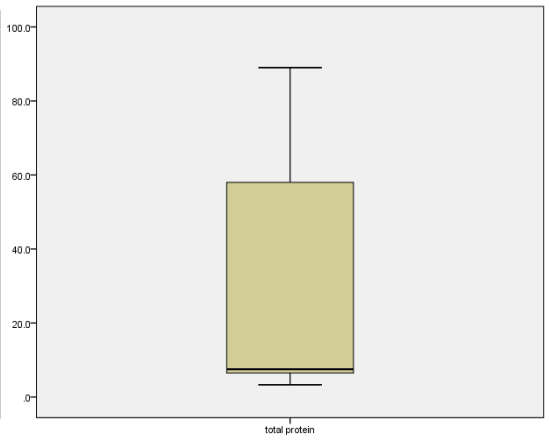
下面分别为数值数据的盒图：

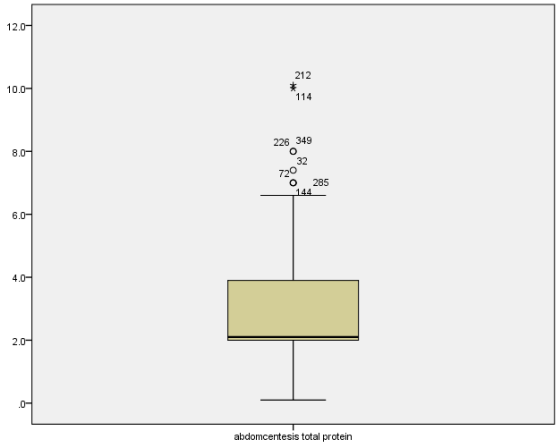












二、 数据缺失的处理

该部分内容先通过 python 对数据进行处理，后通过 spss 制作绘图

1. 将缺失部分剔除

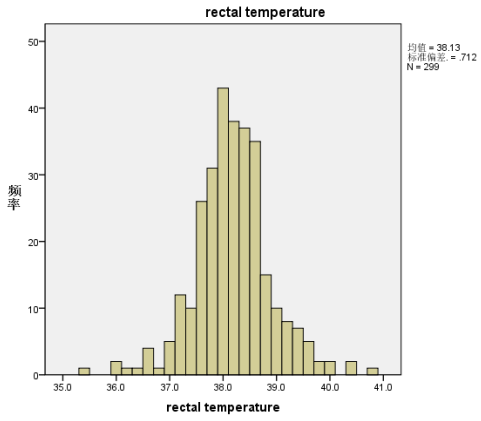
该部分图例已在上一部分展示，不做重复处理。

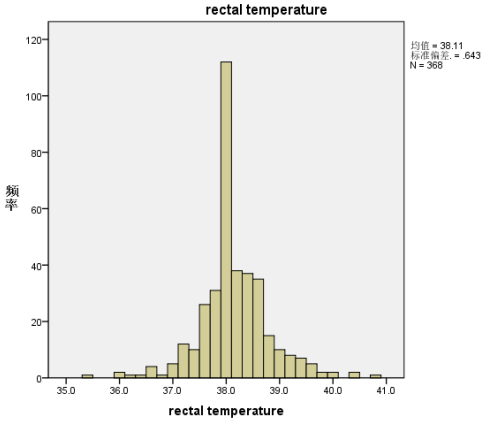
2. 用最高频率值来填补缺失值

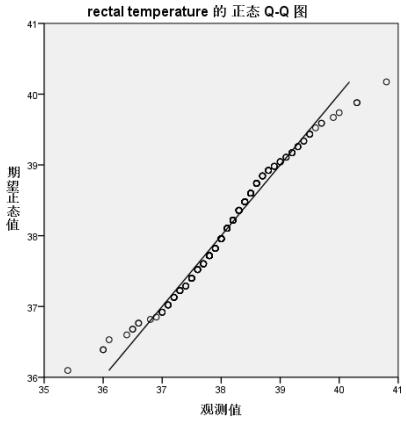
首先通过对每组数值属性进行词频统计，找出该组数值属性的最高频值，并用该值填补缺失值。

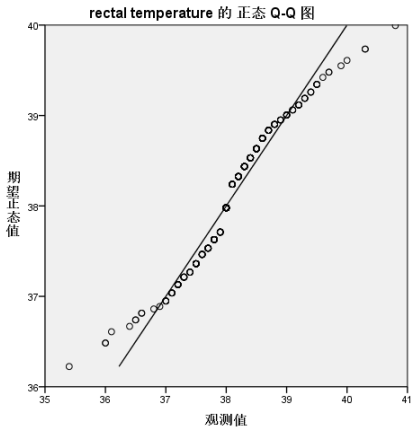
（1）rectal temperature：

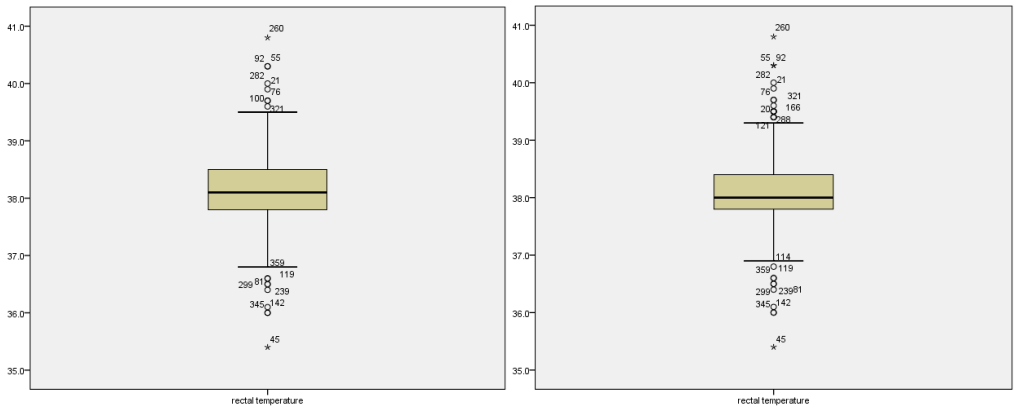
该属性的最高频值为 38.0，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）：





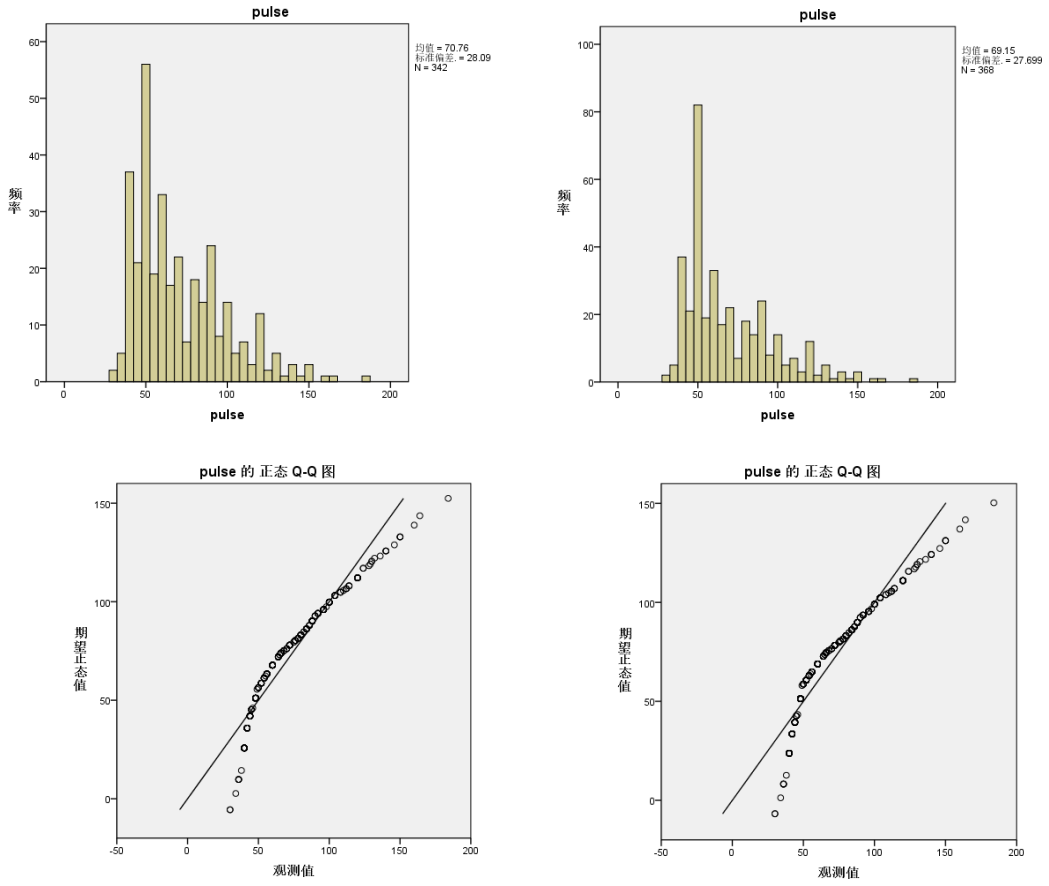


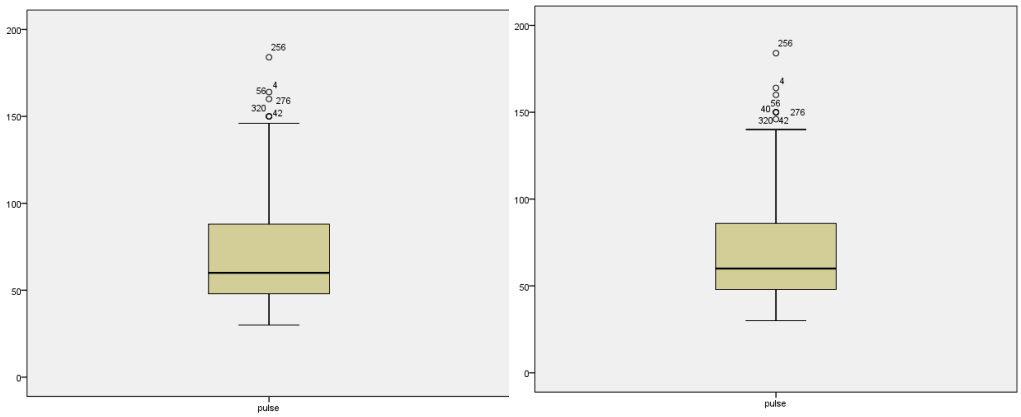




（2）pulse：

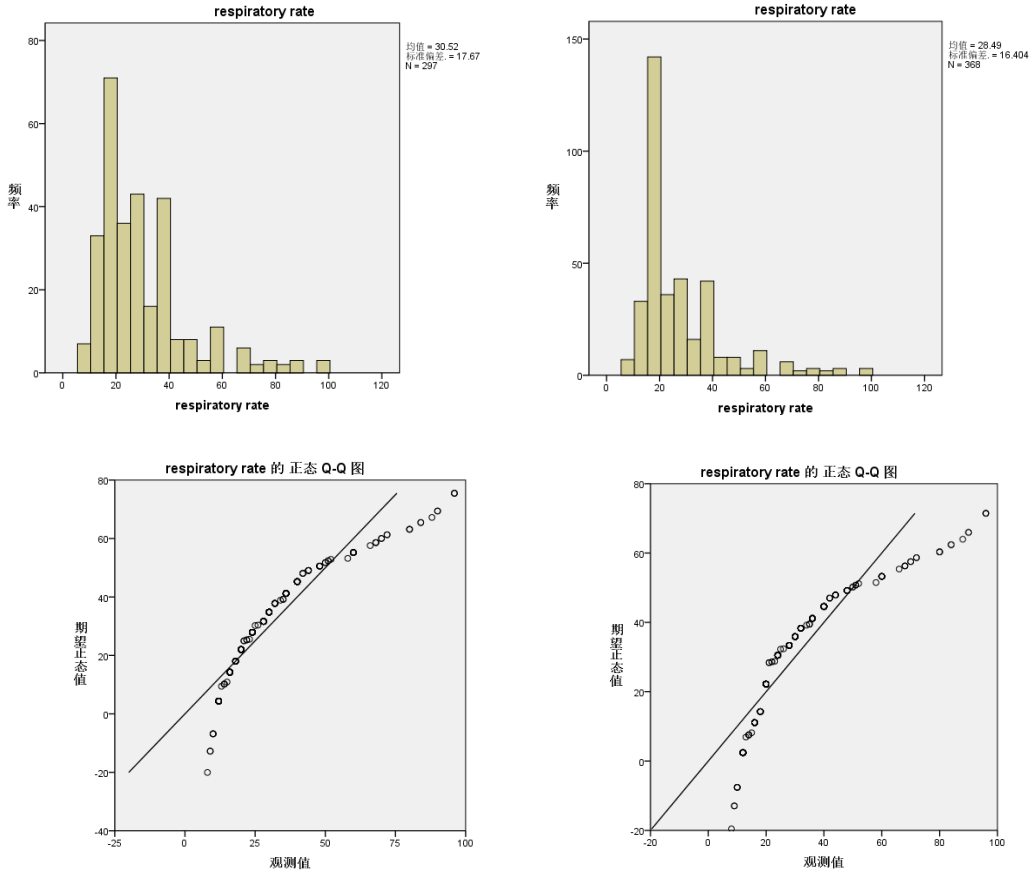
该属性的最高频值为 48，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）：

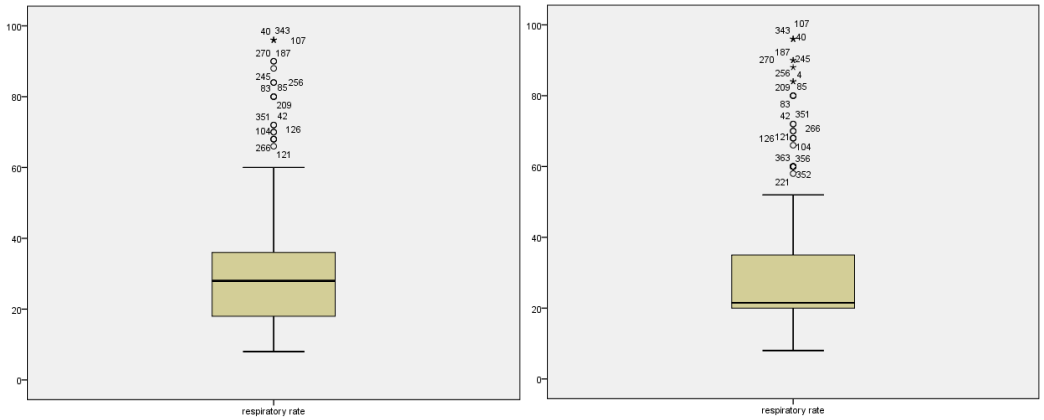




（3）respiratory rate：

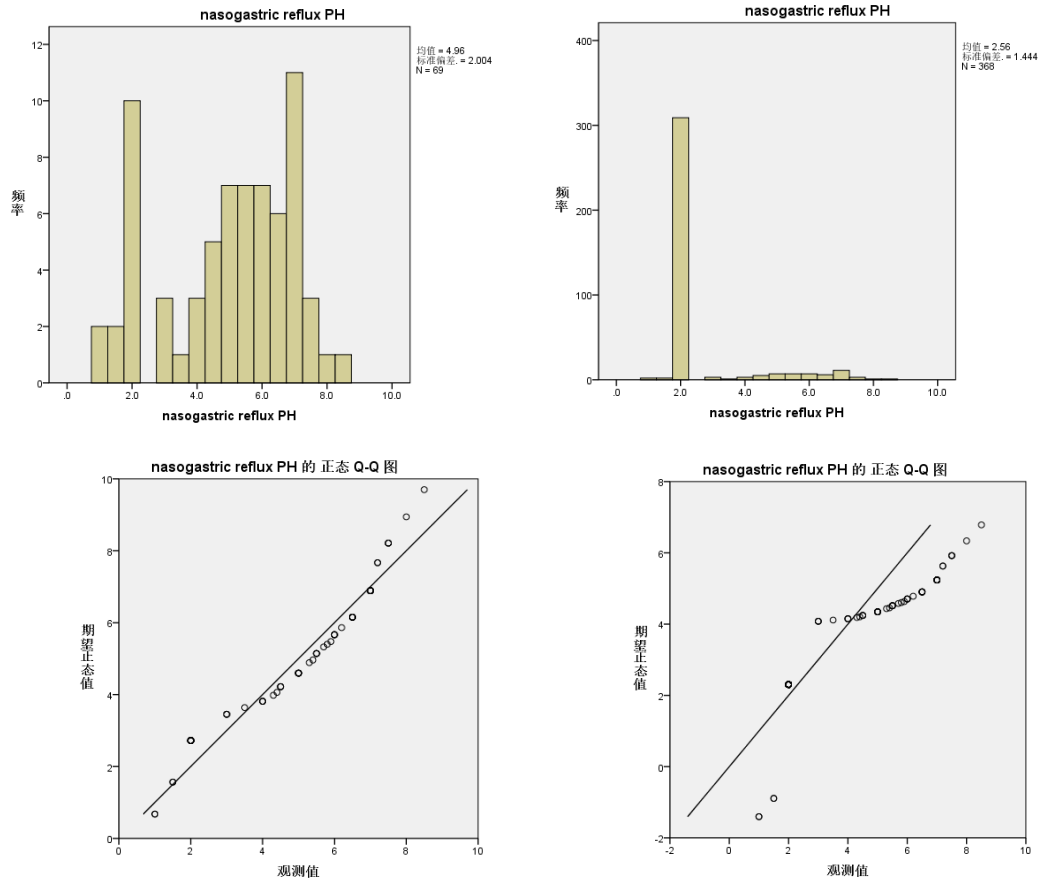
该属性的最高频值为 20.0，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）:

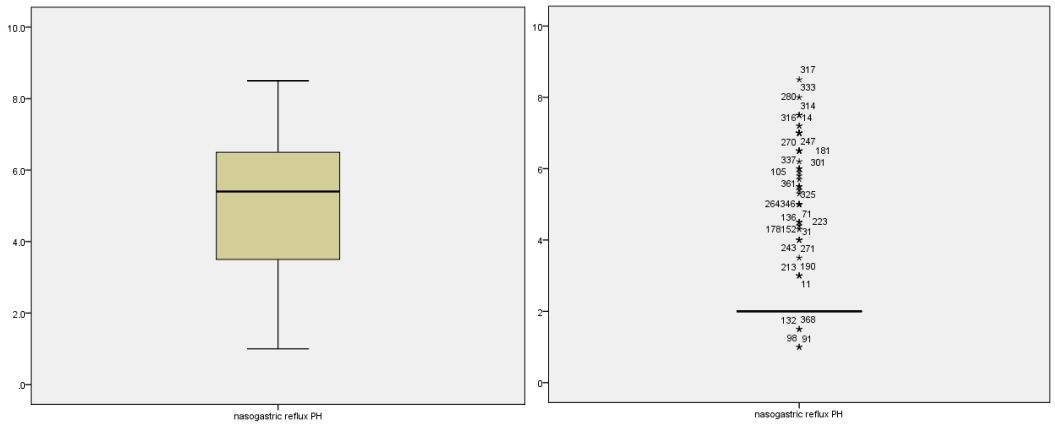




（4）nasogastric reflux PH：

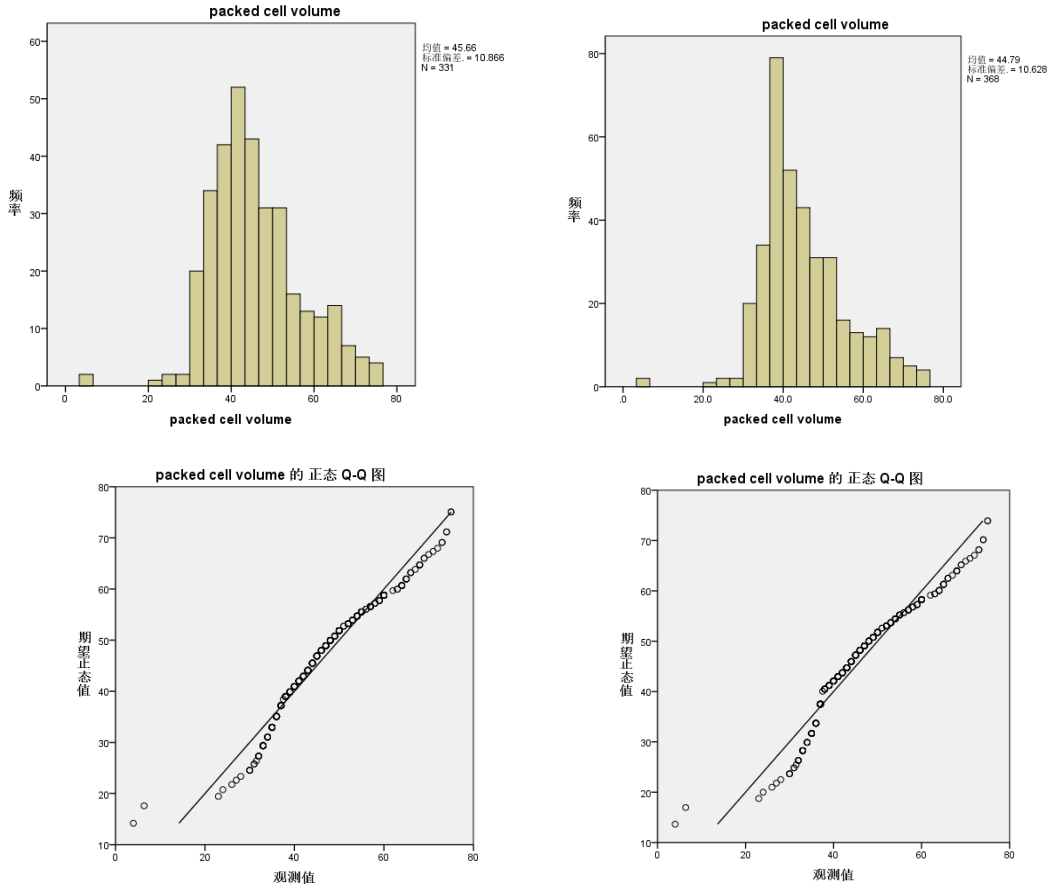
该属性的最高频值为 2.0，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）：

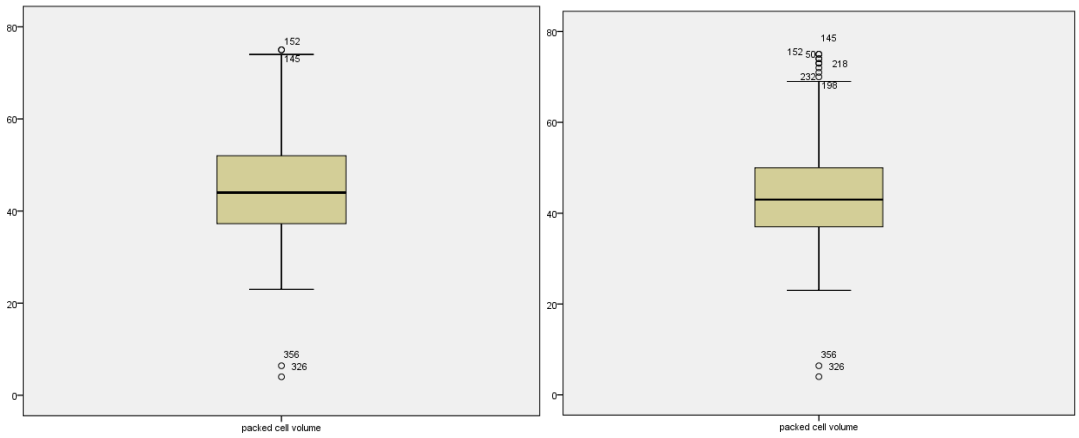




（5）packed cell volume：

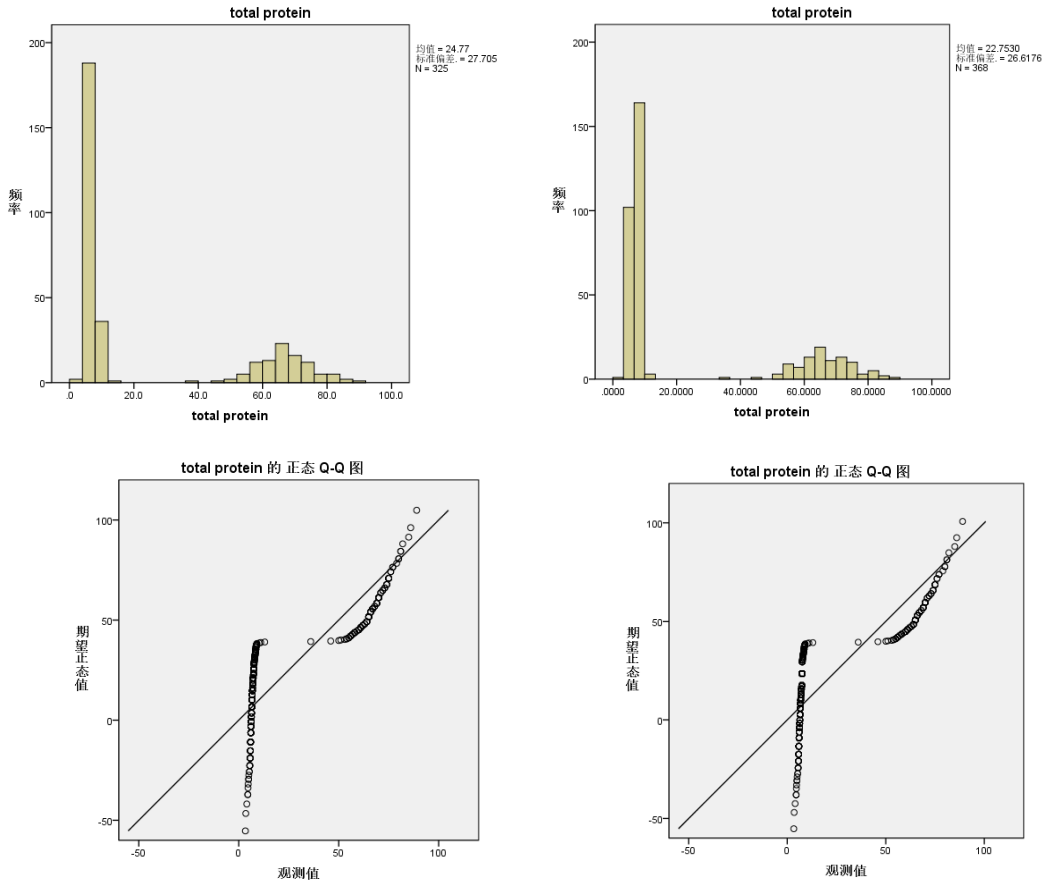
该属性的最高频值为 37.0，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）:

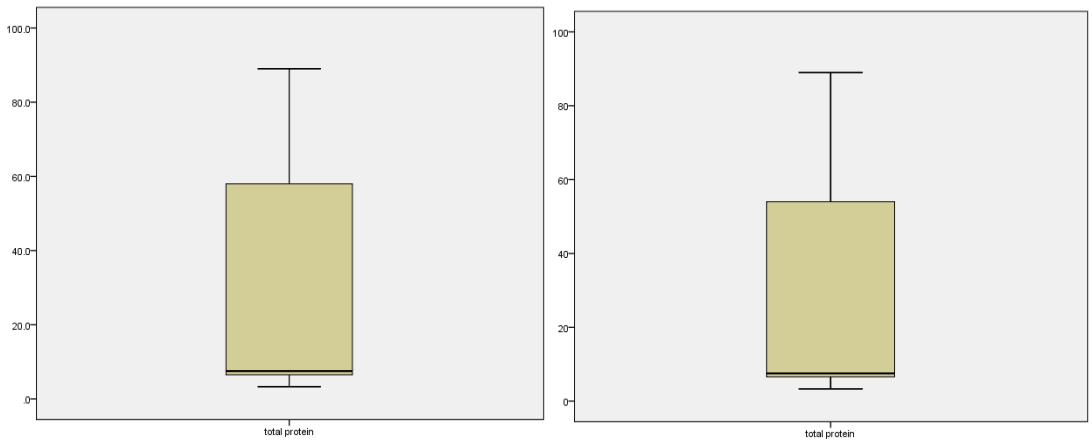




（6）total protein：

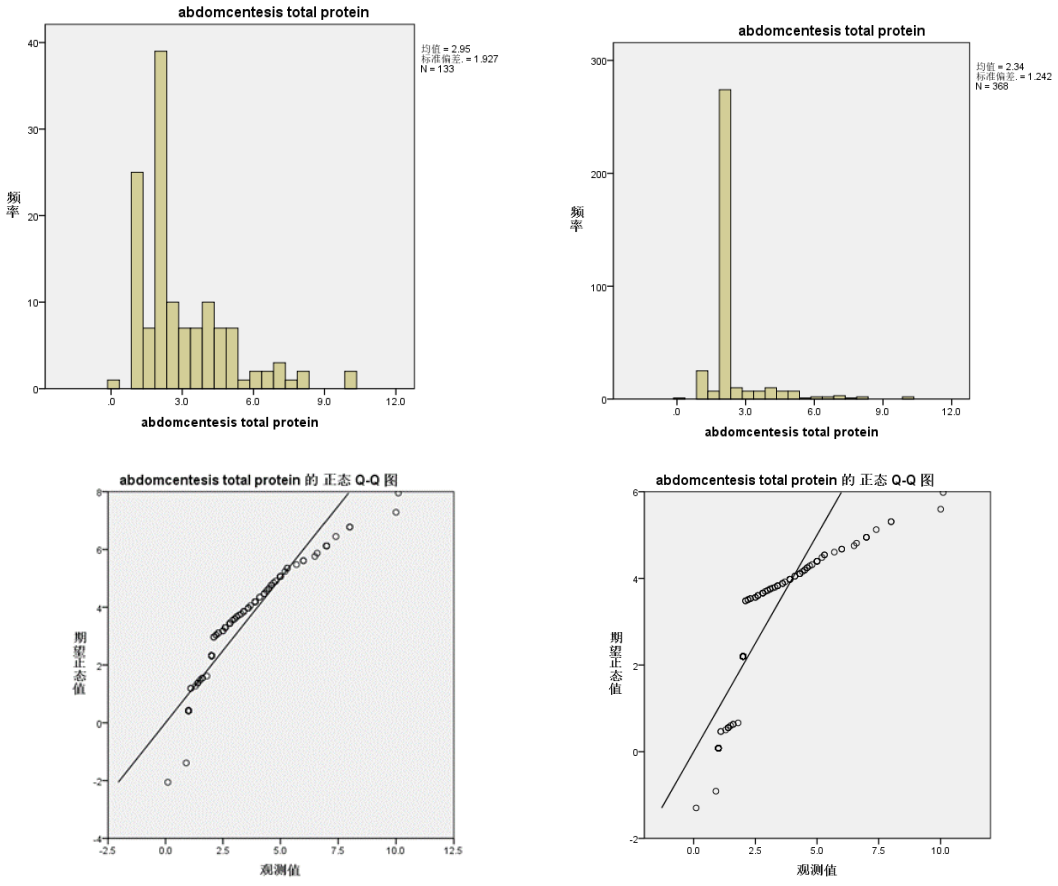
该属性的最高频值为 7.5，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）：

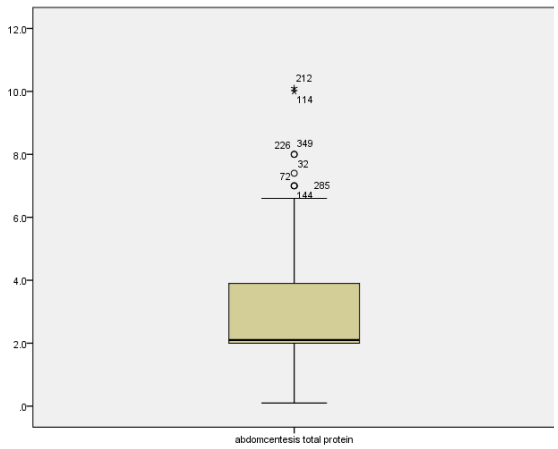


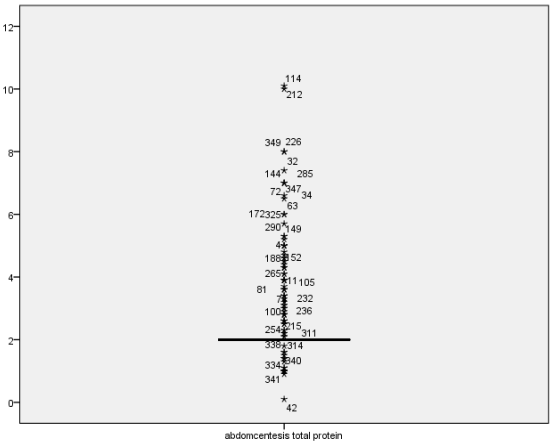


（7）abdomcentesis total protein：

该属性的最高频值为 2.0，用该值填充缺失值后，与省略缺失值图的对比如下（左侧为省略缺失图）：




12.0
10.0
8.0
6.0
4.0
2.0
.0
212
114
226
349
32
72
285
144
abdomcentesis total protein


12
10
8
6
4
2
0
114
212
349
226
32
144
285
72
347
34
63
172
325
290
149
4
188
52
265
111
105
81
232
100
236
254
215
311
338
314
340
334
341
42
abdomcentesis total protein