数据挖掘作业——马的疝病分析

# 数据摘要及可视化

## 数据摘要

该部分内容均在python下通过pandas库完成。

### 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：

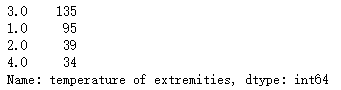
（1）Surgery：



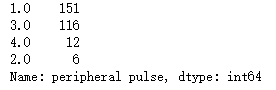
（2）Age：



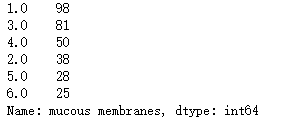
（3）temperature of extremities：



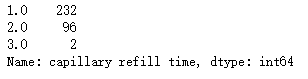
（4）peripheral pulse：



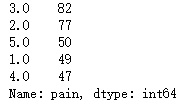
（5）mucous membranes：



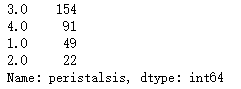
（6）capillary refill time：



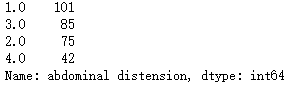
（7）pain：



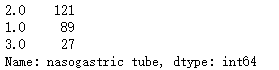
（8）Peristalsis：



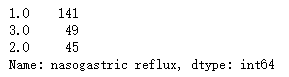
（9）abdominal distension：



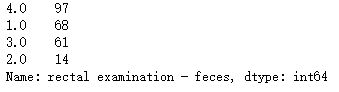
（10）nasogastric tube：



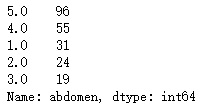
（11）nasogastric reflux：



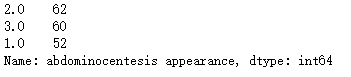
（12）rectal examination – feces：



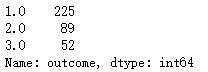
（13）abdomen：



（14）abdominocentesis appearance：



（15）outcome：



（16）surgical lesion：



（17）cp\_data：

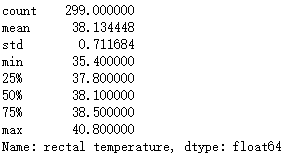


### 数值型数据统计

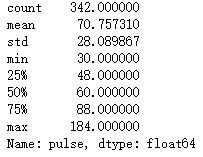
在该数据集中包含以下数值属性：rectal temperature、pulse、respiratory rate、nasogastric reflux PH、packed cell volume、total protein、abdomcentesis total protein。

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。

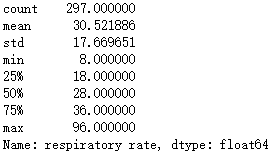
（1）rectal temperature：



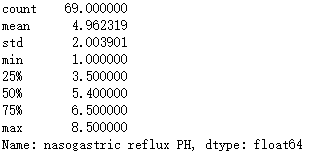
（2）pulse：



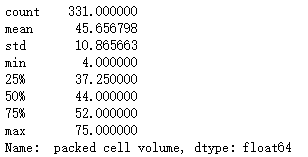
（3）respiratory rate：



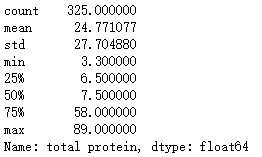
（4）nasogastric reflux PH：



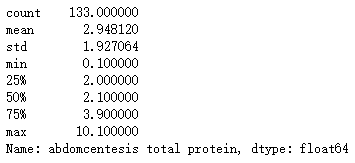
（5）packed cell volume：



（6）total protein：



（7）abdomcentesis total protein：



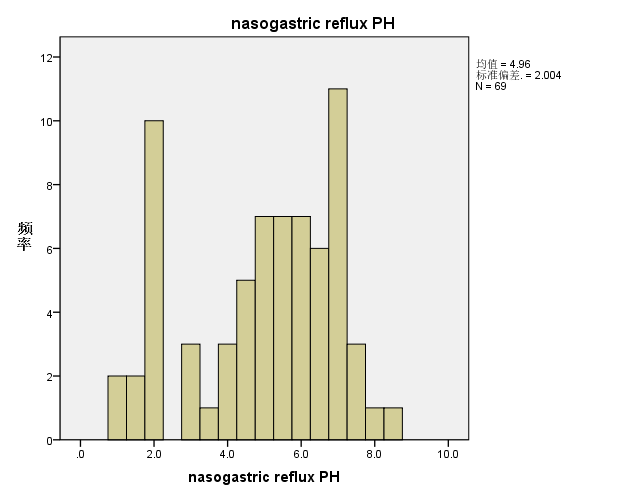
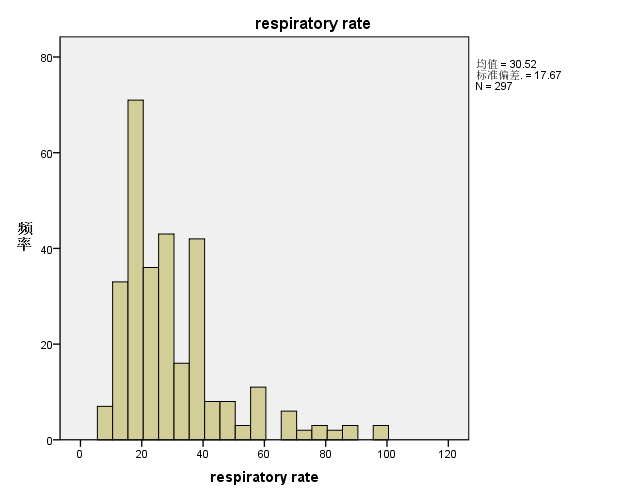
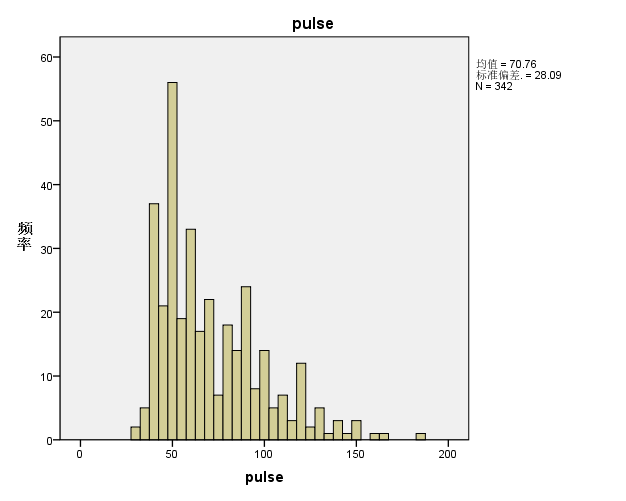
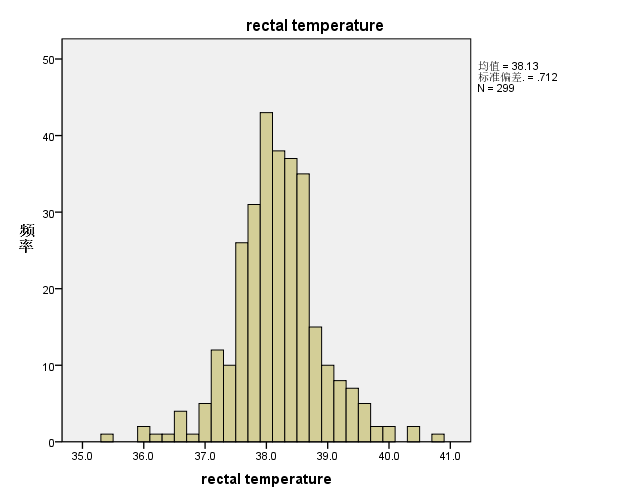
## 2. 数据可视化

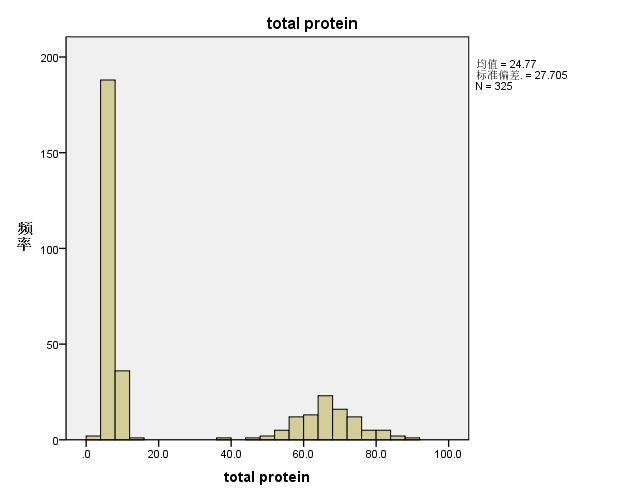
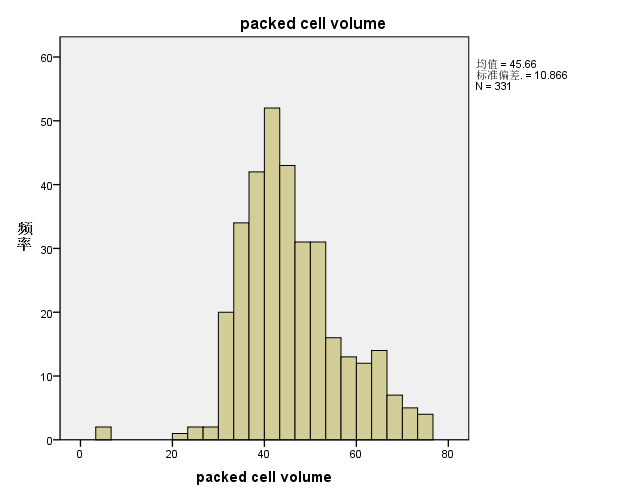
该部分内容通过spss工具完成。

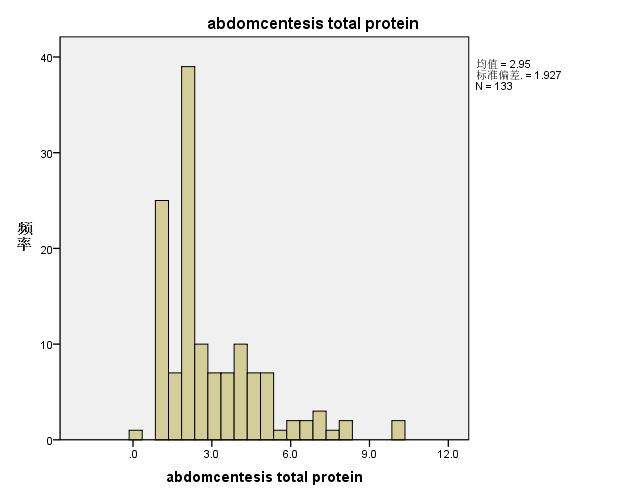
### 2.1 数值属性的直方图及Q\_Q图

#### 2.1.1 直方图

下面分别为数值数据的直方图：

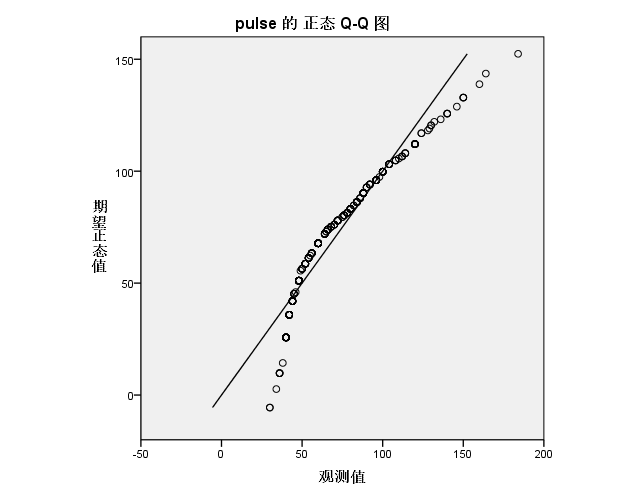
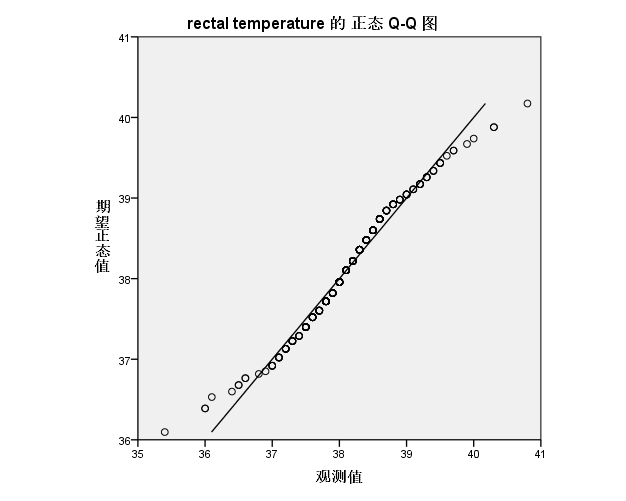


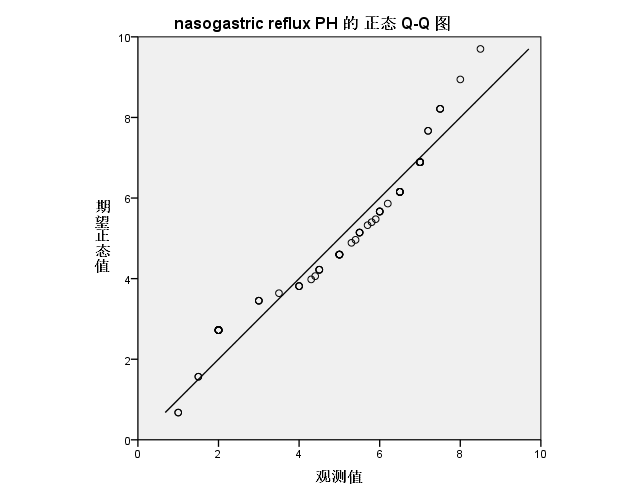
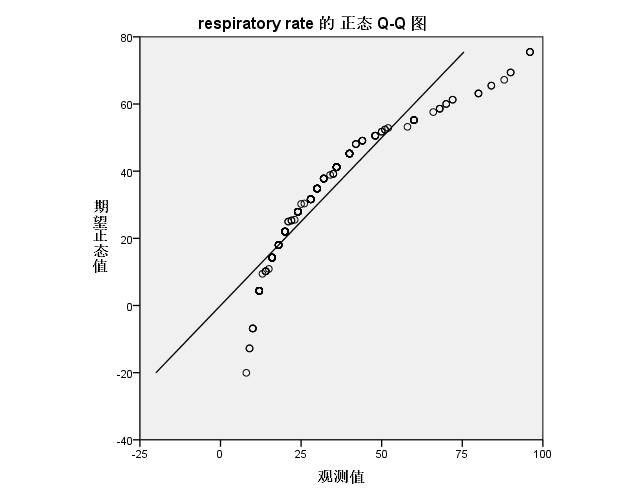


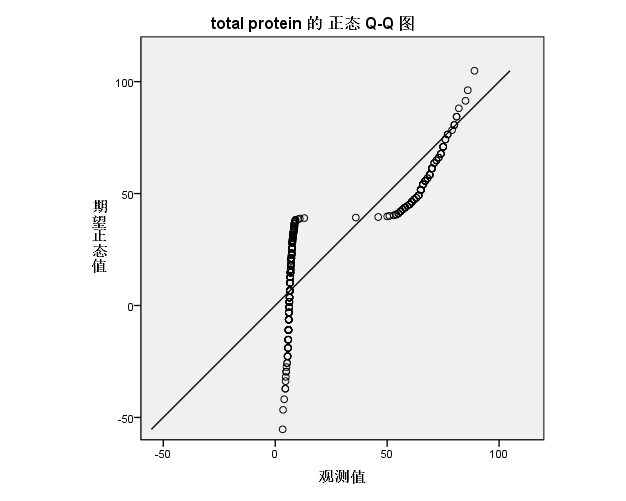
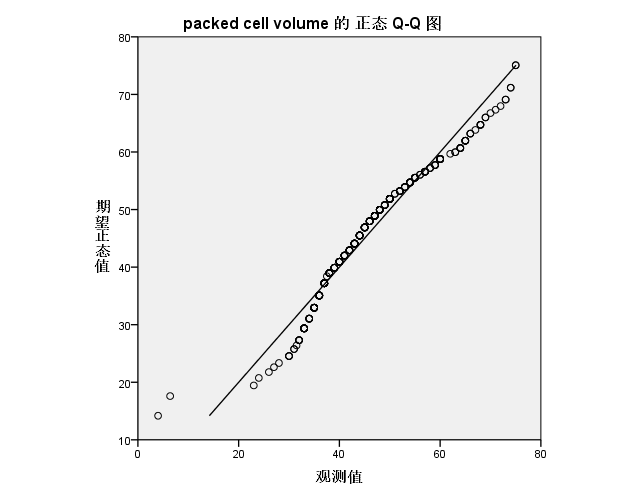


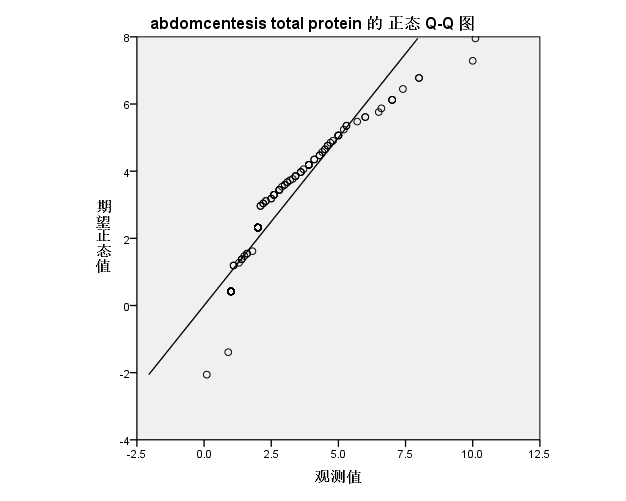
#### 2.1.2 Q\_Q图

下面分别为数值数据的直方图：



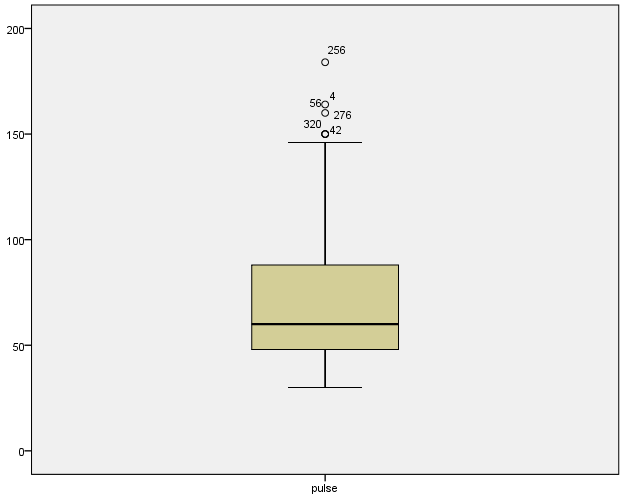
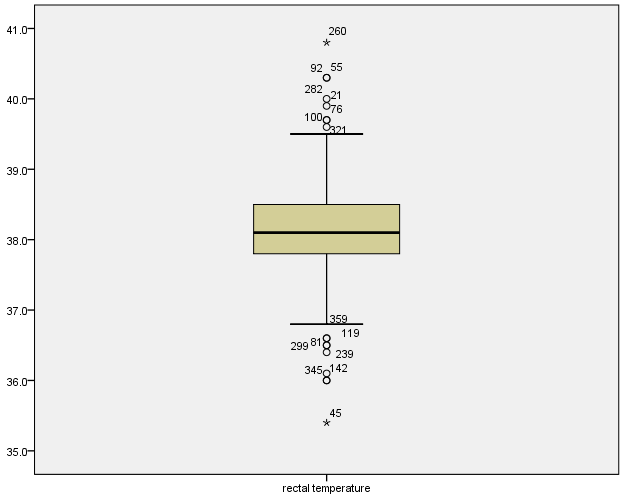


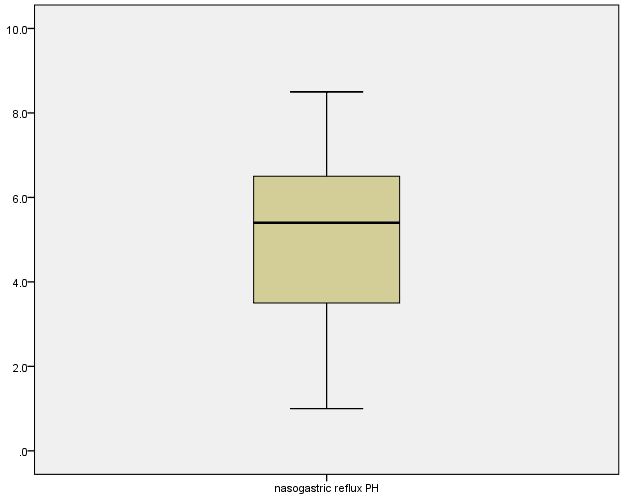
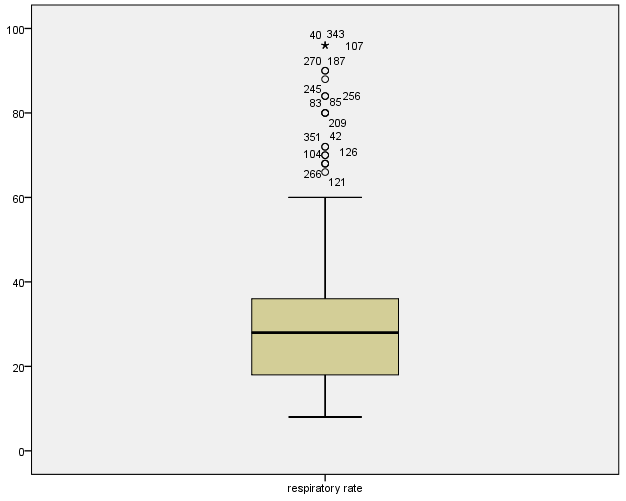


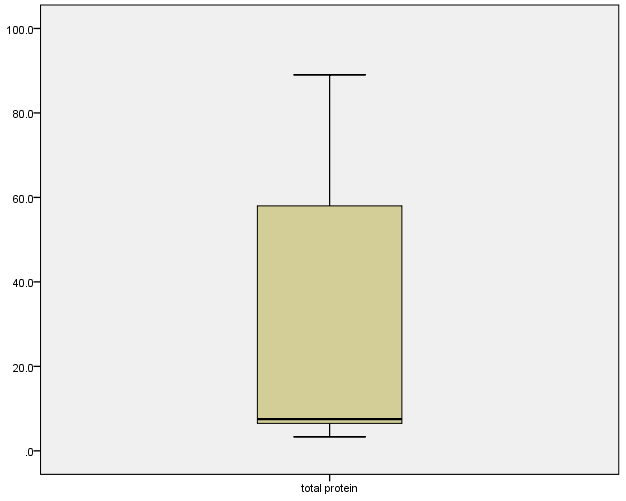
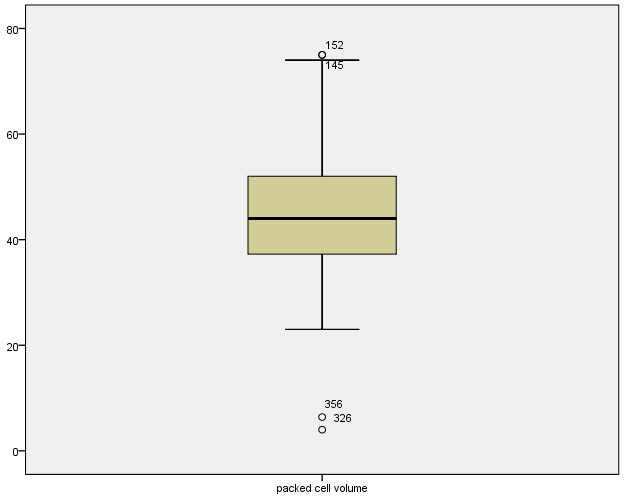


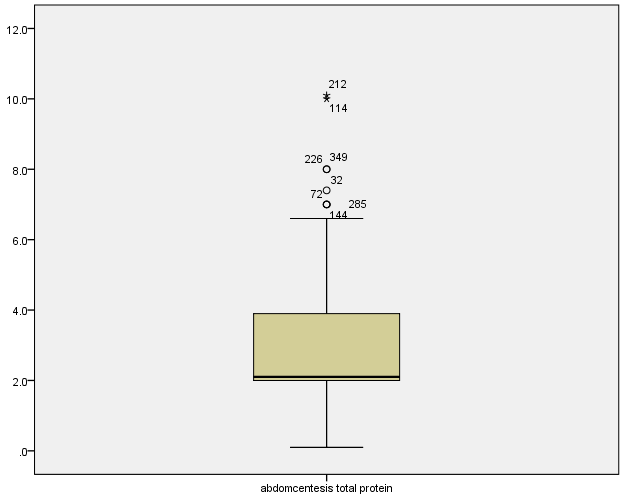
### 2.2 数值属性的盒图

下面分别为数值数据的盒图：









# 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

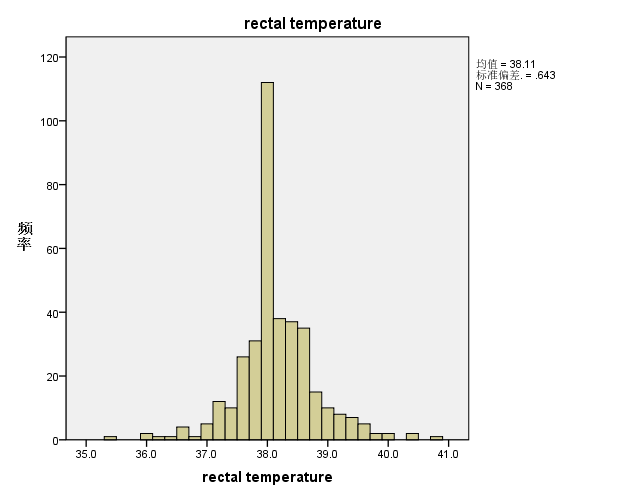
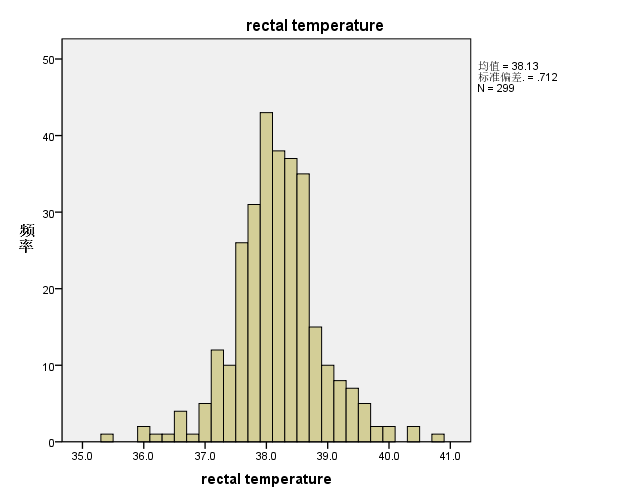
该部分图例已在上一部分展示，不做重复处理。

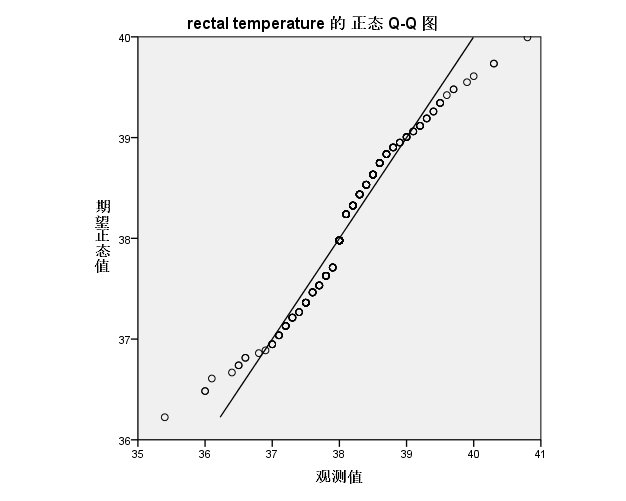
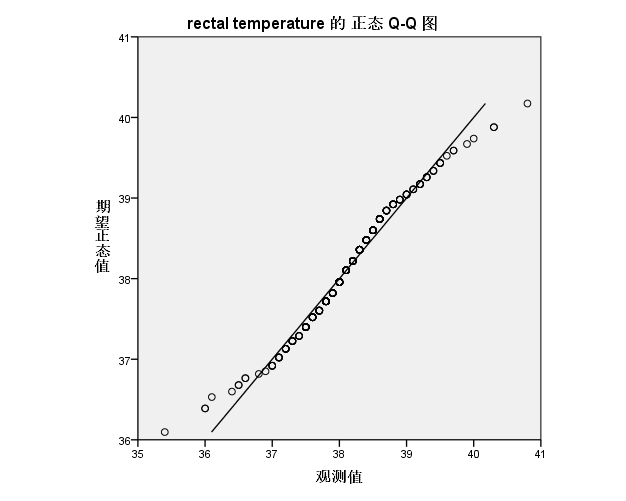
## 用最高频率值来填补缺失值

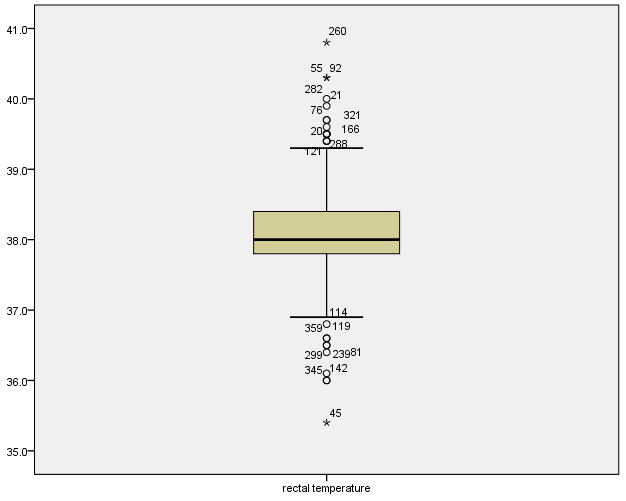
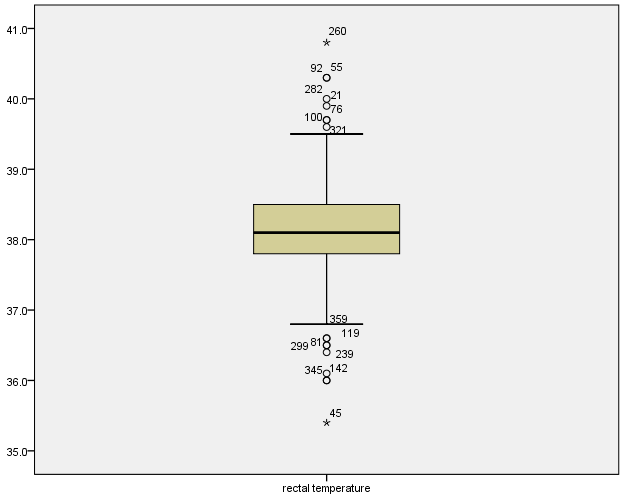
首先通过对每组数值属性进行词频统计，找出该组数值属性的最高频值，并用该值填补缺失值。

（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）：

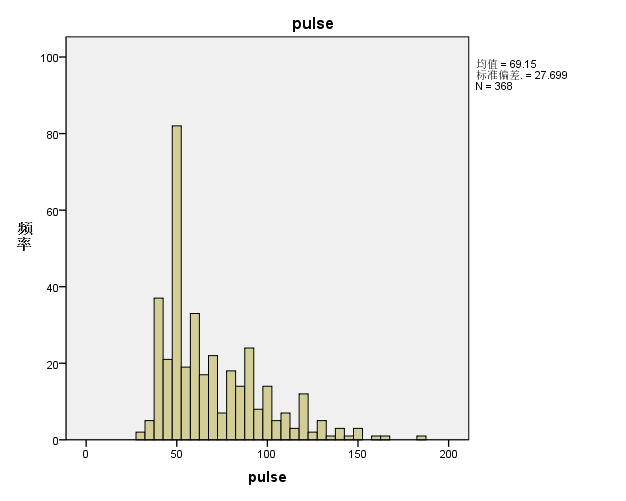
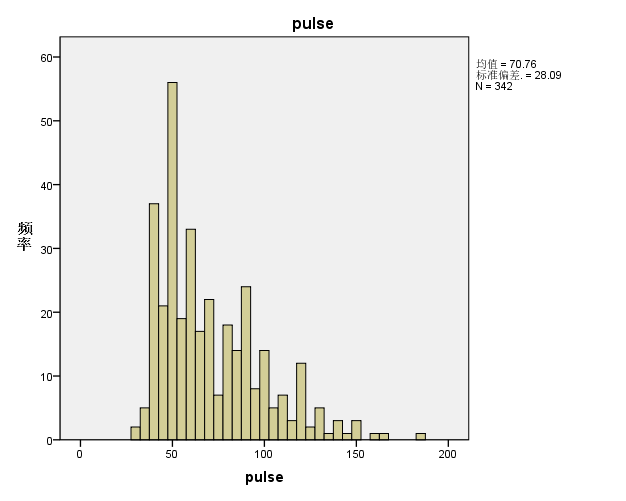


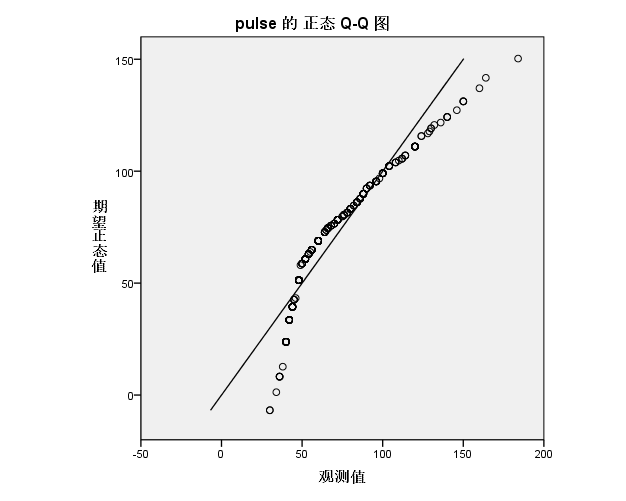
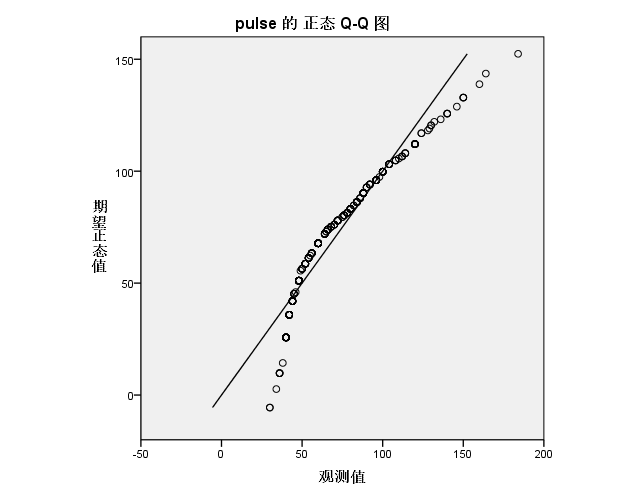


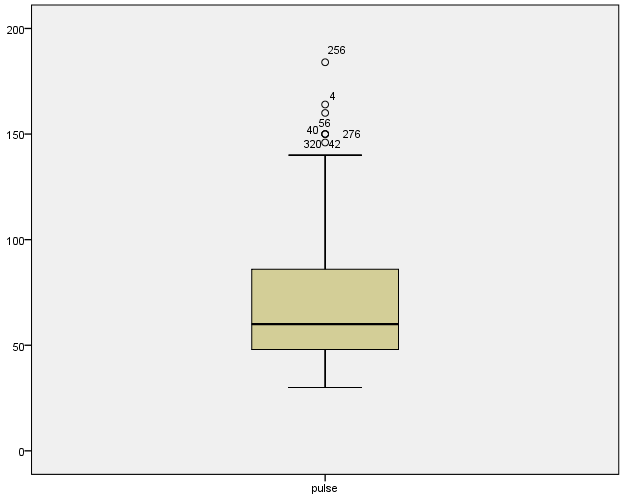
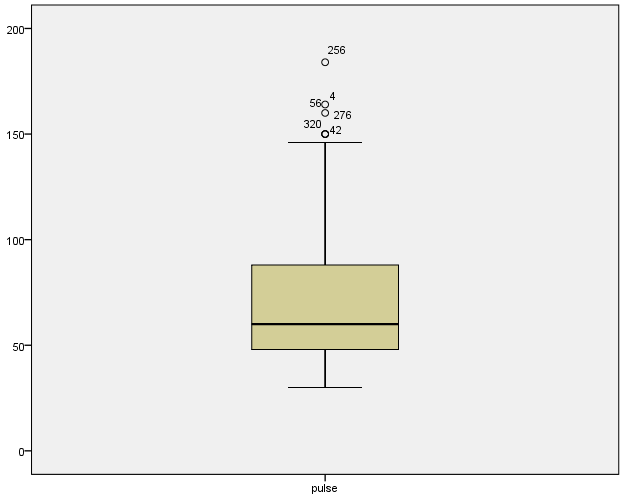


（2）pulse：

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）：

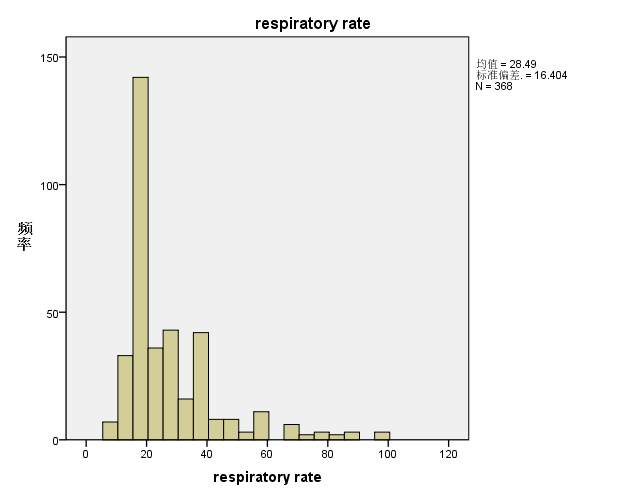
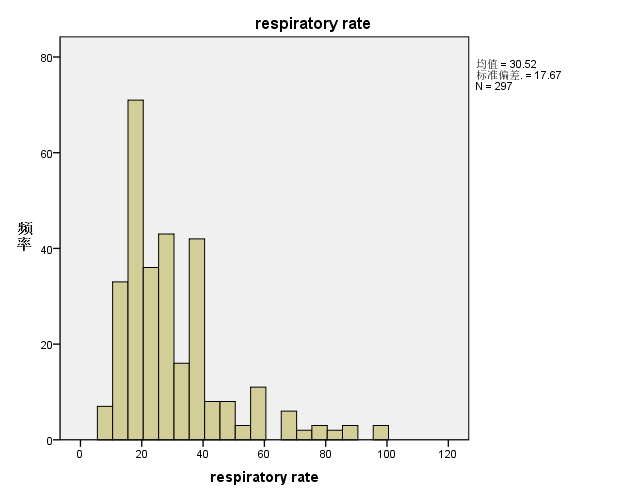


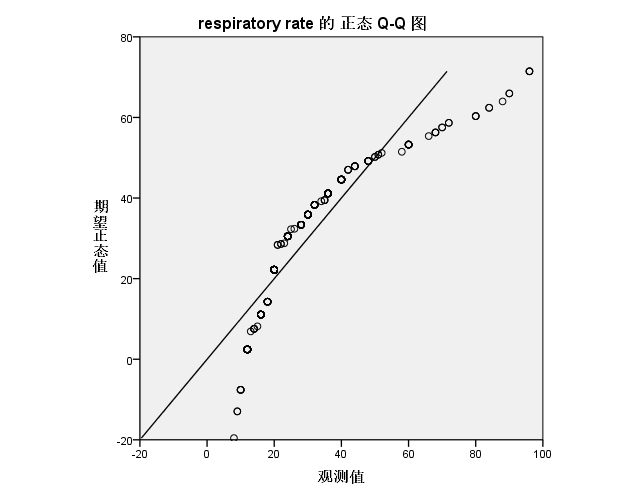
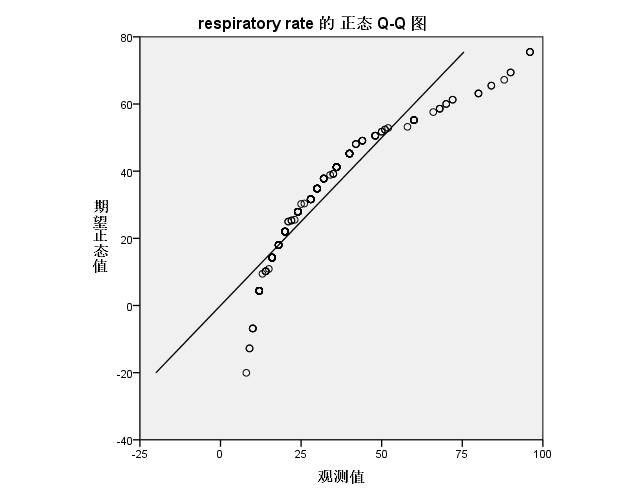


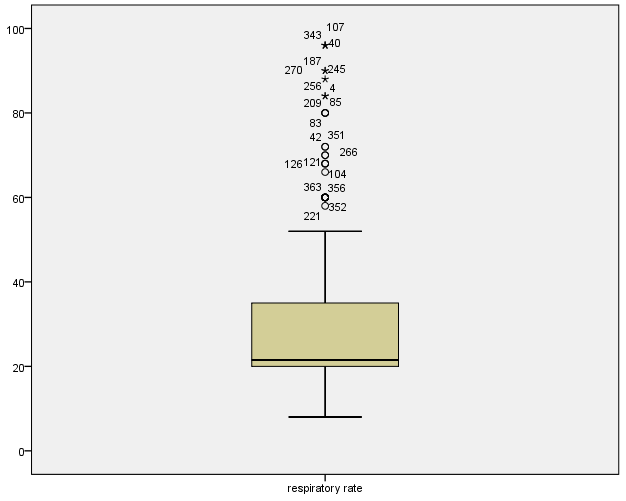
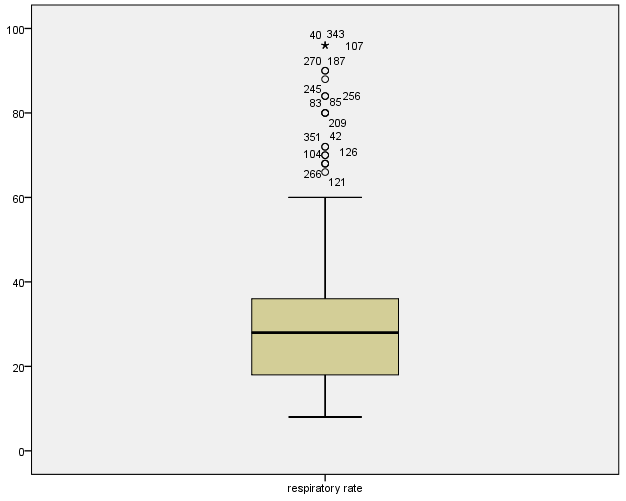


（3）respiratory rate：

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）：

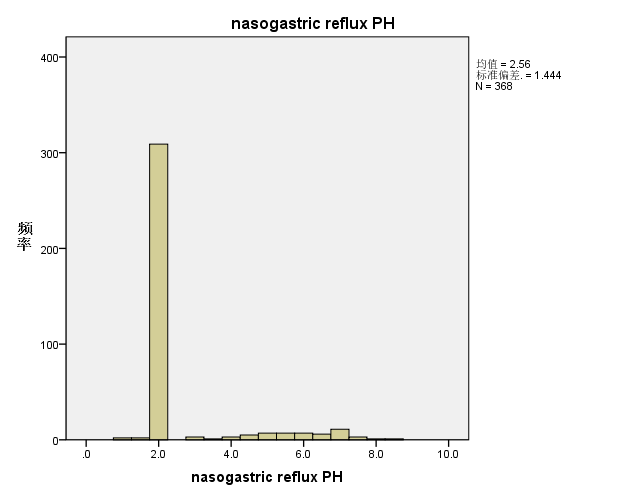
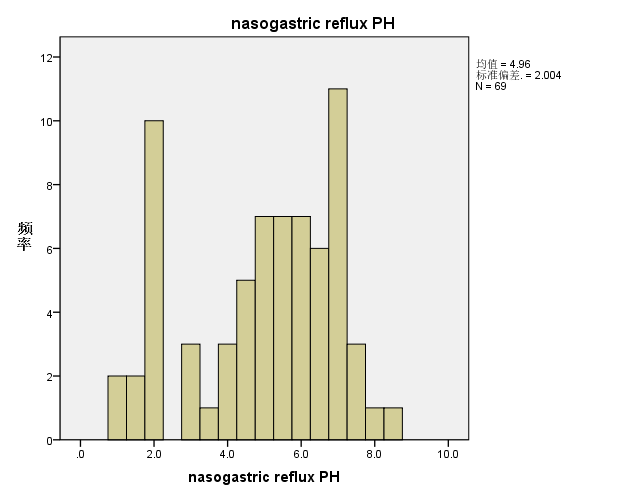


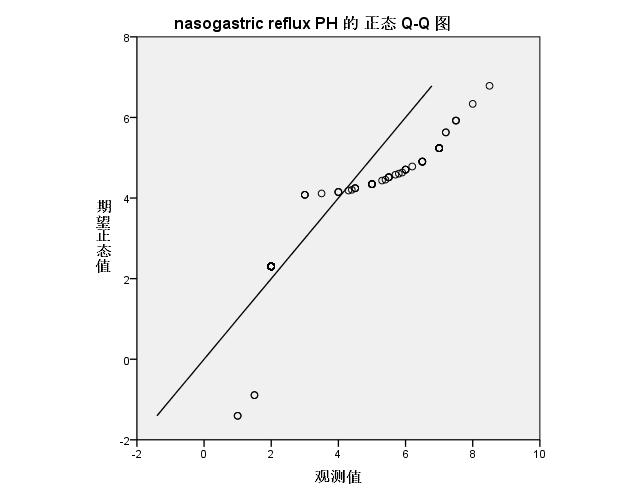
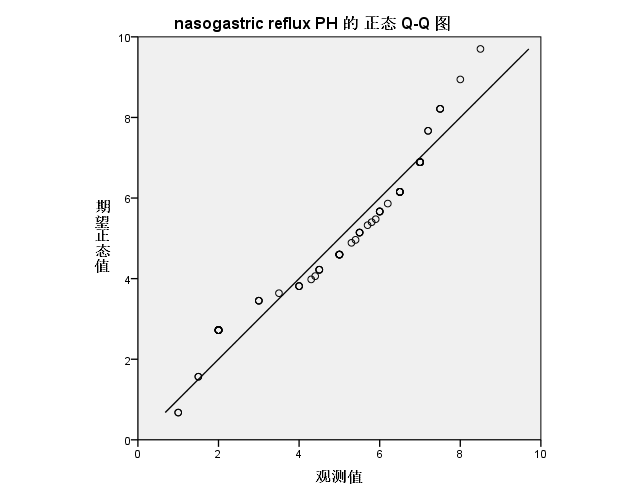


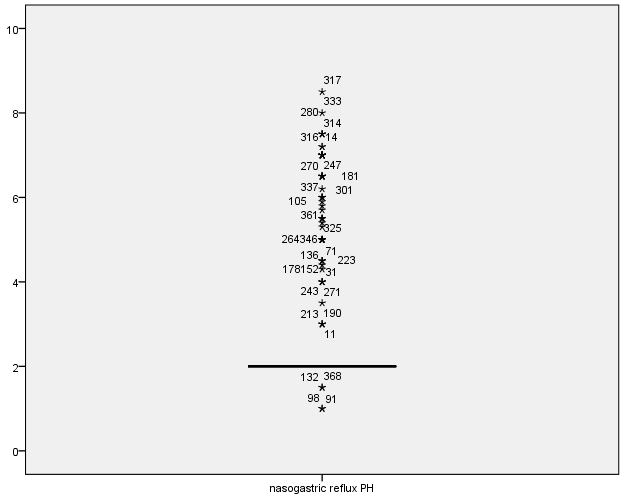
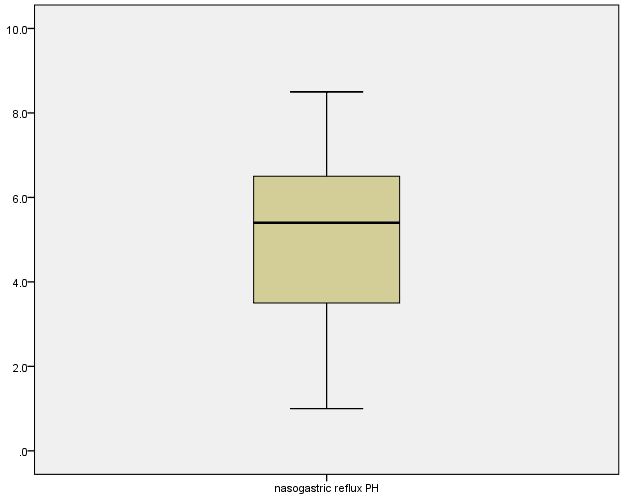


（4）nasogastric reflux PH：

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）：

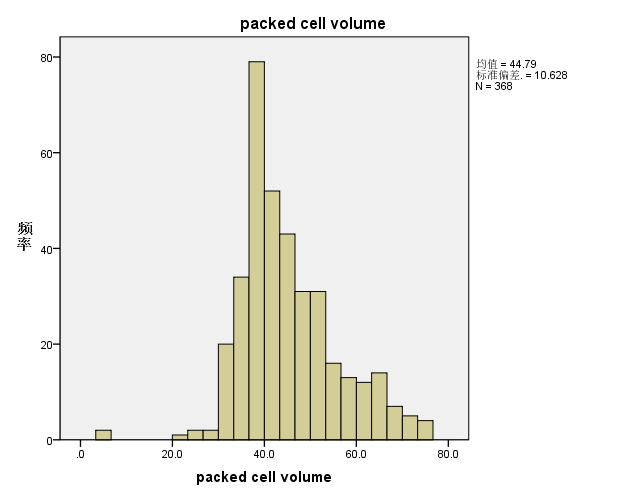
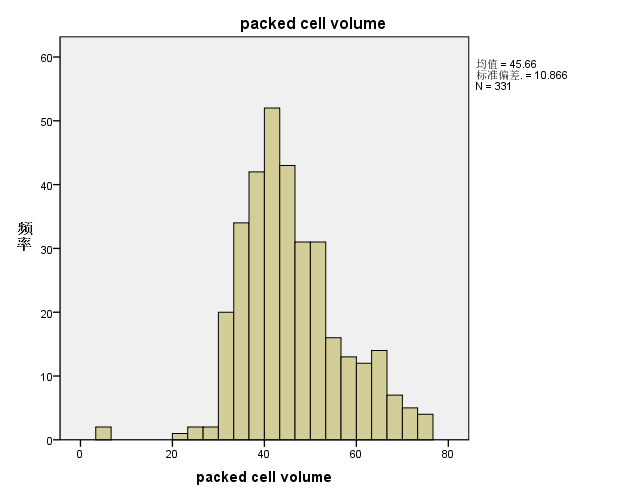


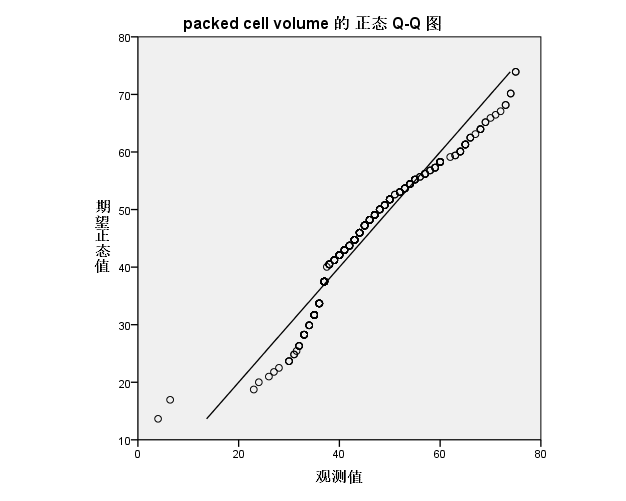
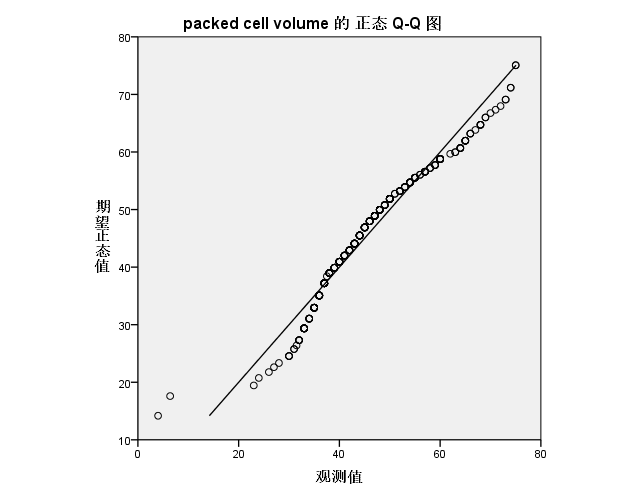


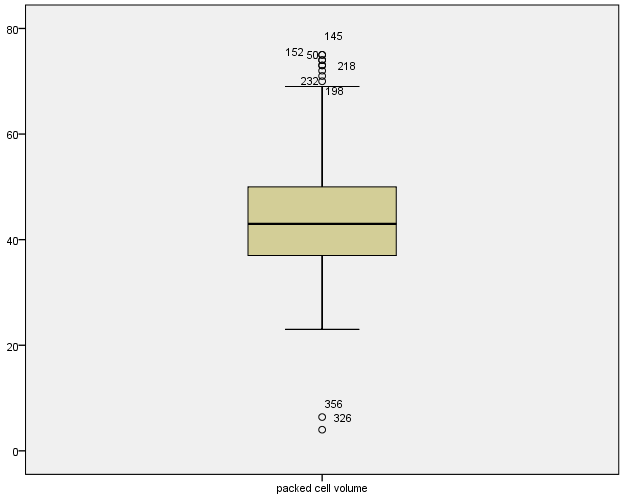
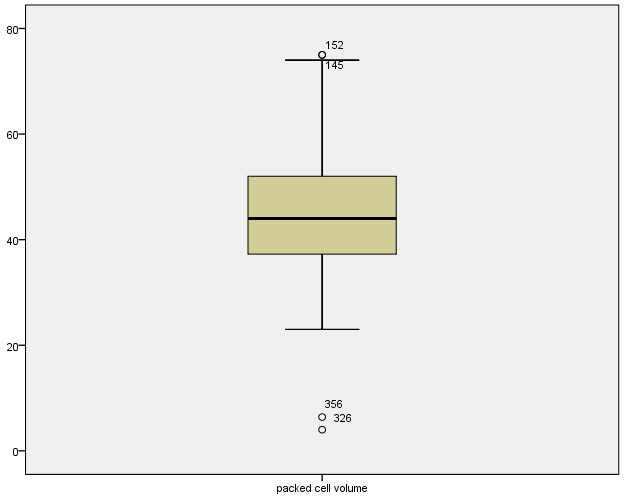


（5）packed cell volume：

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）：

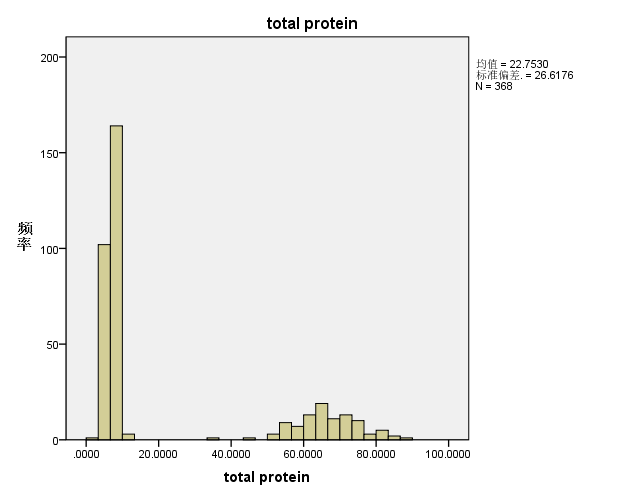
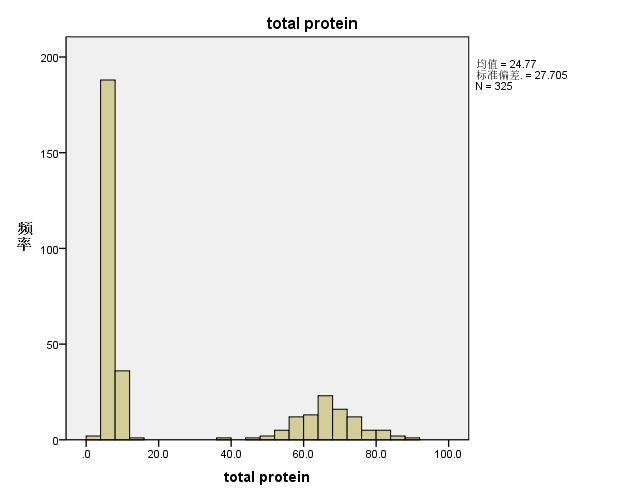


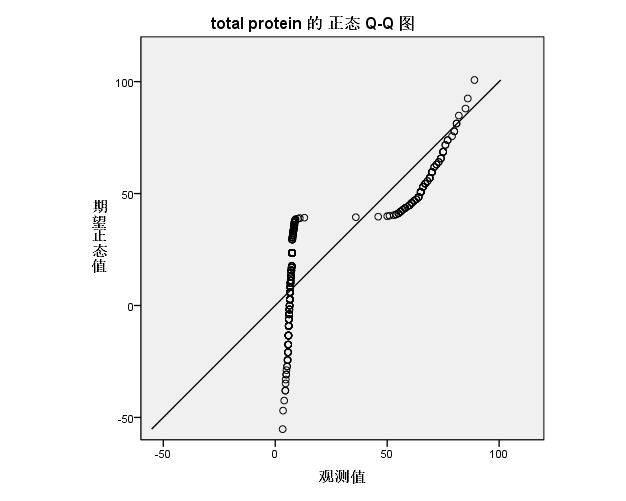
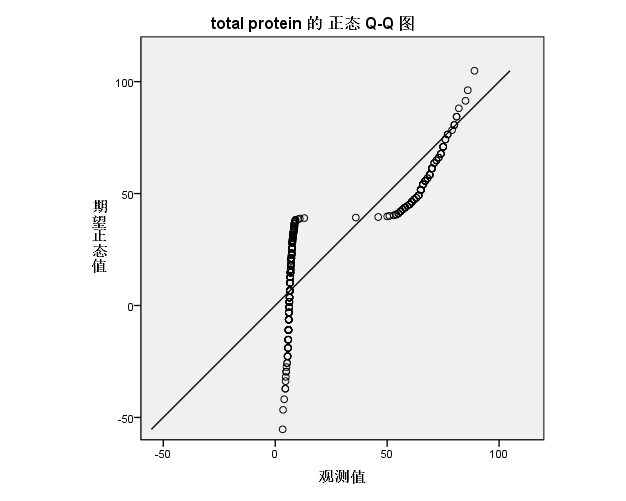


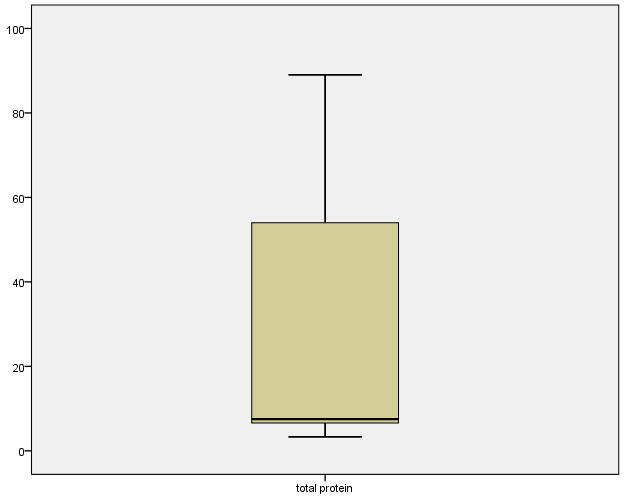
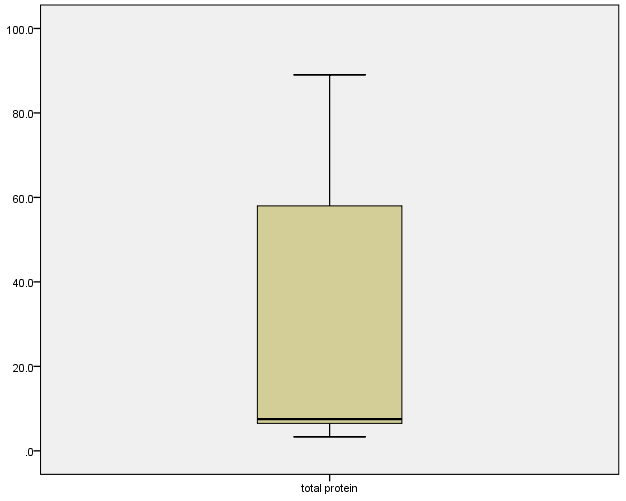


（6）total protein：

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）：

