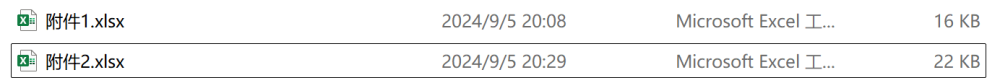
# 编程过程记录

### excel工作簿分离

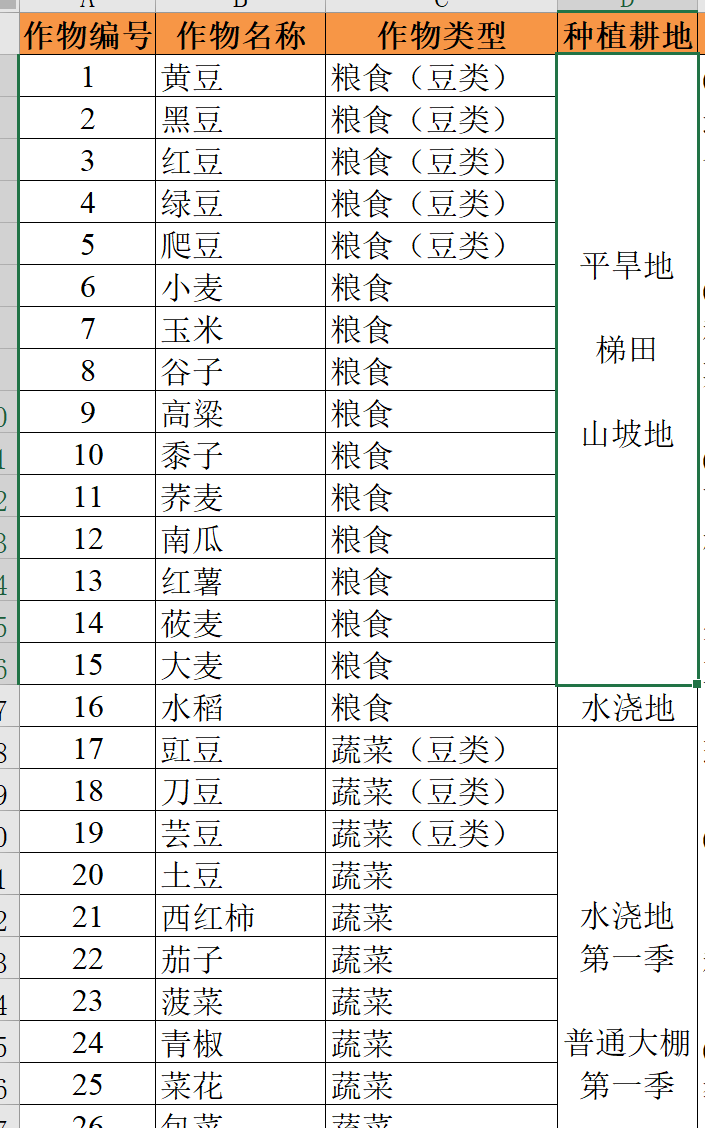


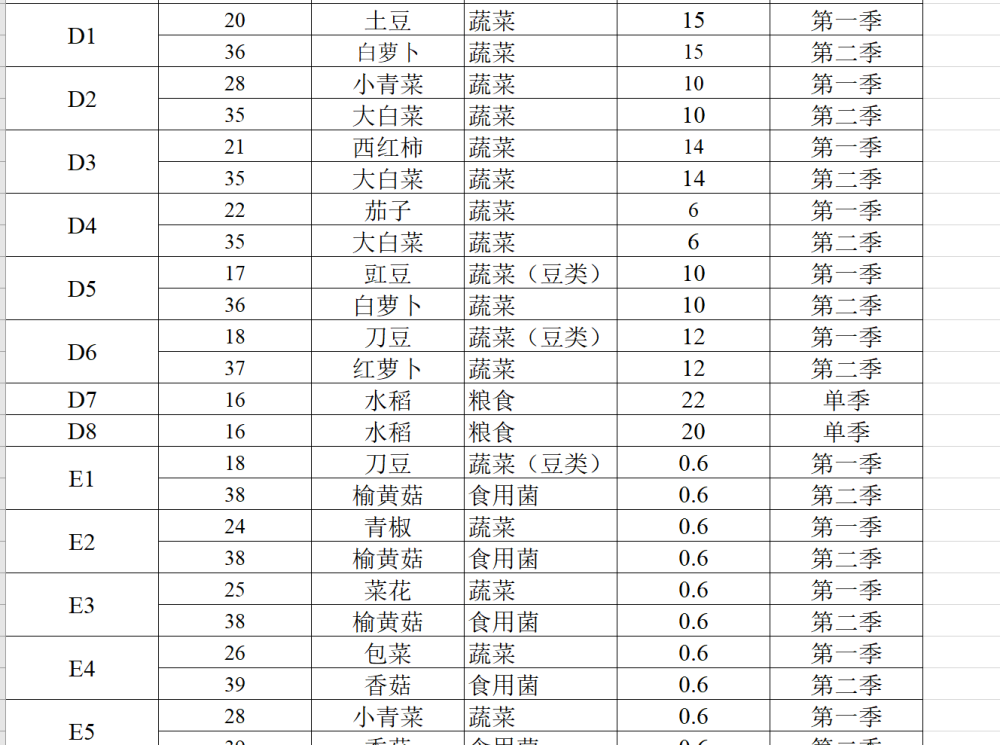
附件1划分为乡村的现有耕地.csv，乡村种植的农作物.csv

附件2划分为2023年的农作物种植情况.csv,2023年统计的相关数据.csv

### 取消合并单元格

为了方便转换成csv文件，需要取消单元格合并





### 合并表

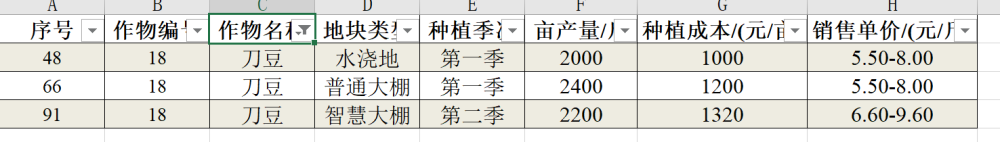
这两张表的地块名称与种植地块的内容相对应，因此可以进行合并



合并结果为



一个作物会对应多个亩产量



因此，各作物的亩产量采取均值记录



### 数据预处理

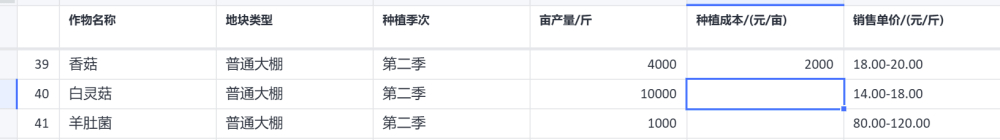
#### 缺失值处理

通过观察附件中2023的种植情况与统计数据发现，文件中不存在缺失值，因此不需要进行缺失值处理，下一步是异常值处理

#### 异常值处理

使用3∂异常识别方法对于附件中的数据进行排查分析发现，2023年统计的相关数组中存在四个异常值，异常的数据如下, 分别是黄瓜，白灵菇，羊肚菌，黄瓜



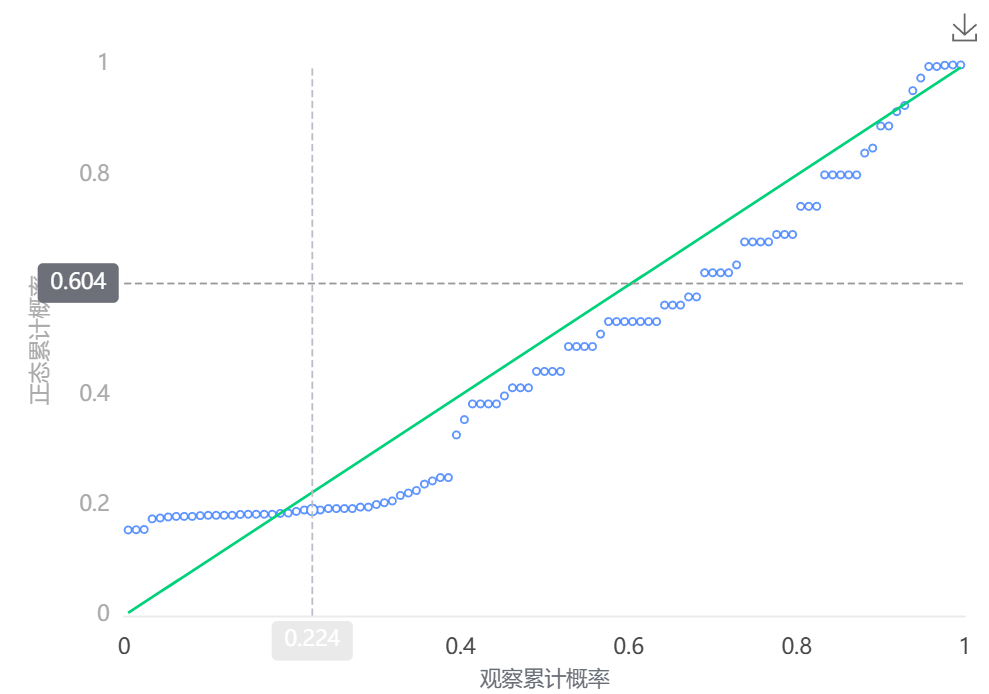




对异常数据进行置空处理

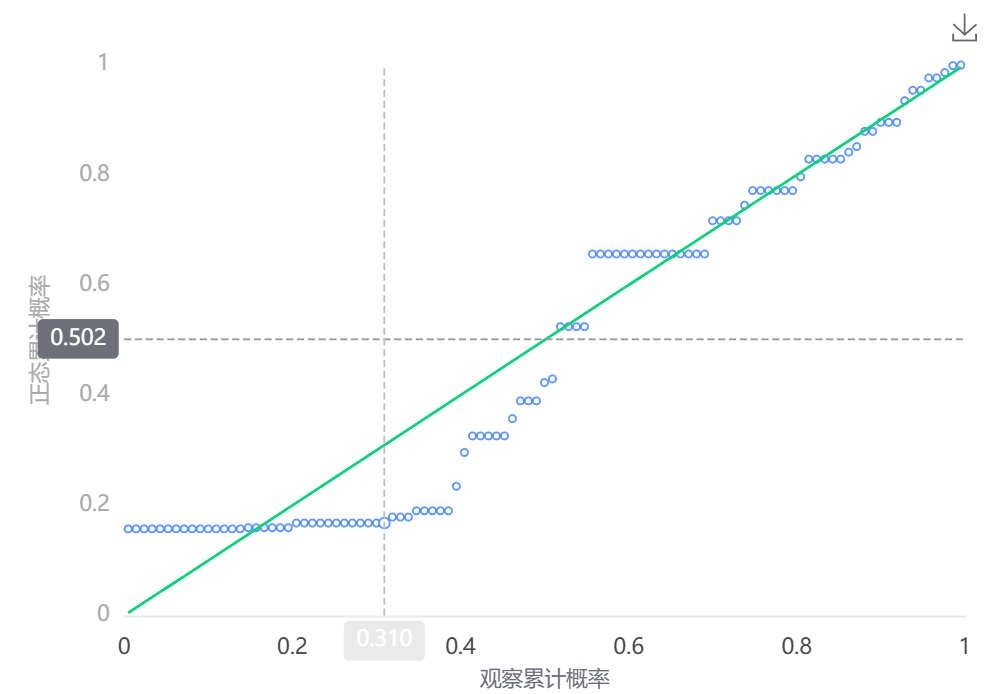
处理完异常值后对数据进行正态性检验，结果如下**输出结果3：正态性检验P-P图**

亩产量/斤



**图表说明：**上图是亩产量/斤计算观测的累计概率（P）与正态累计概率（P）的拟合情况。拟合程度越高越服从正态分布。● 图示最大仅显示1000个样本量信息，若样本量大于1000，则在全样本中进行随机抽样，选取其中1000样本进行绘制。

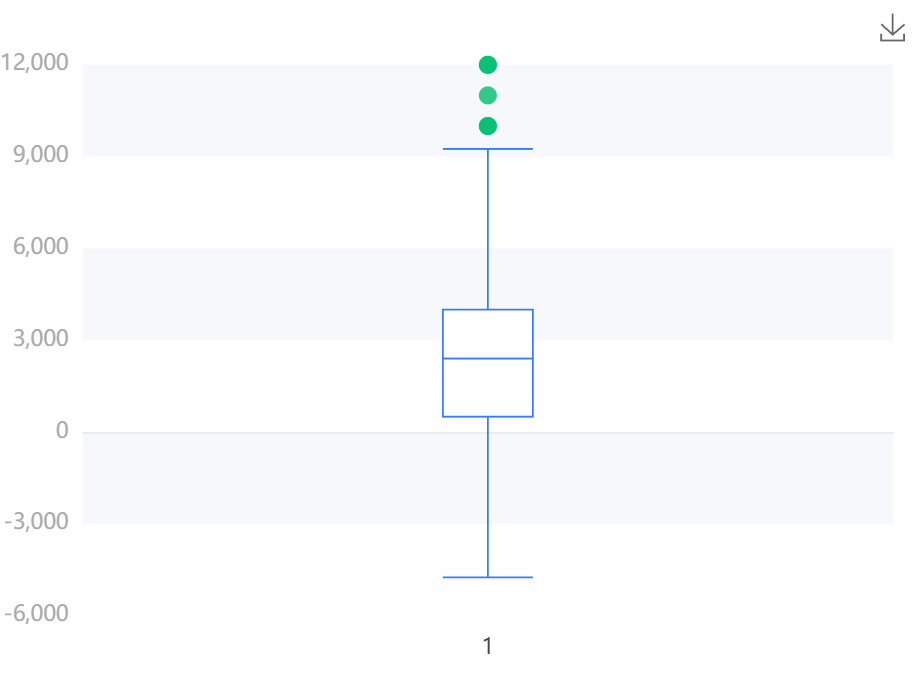
种植成本/(元/亩)

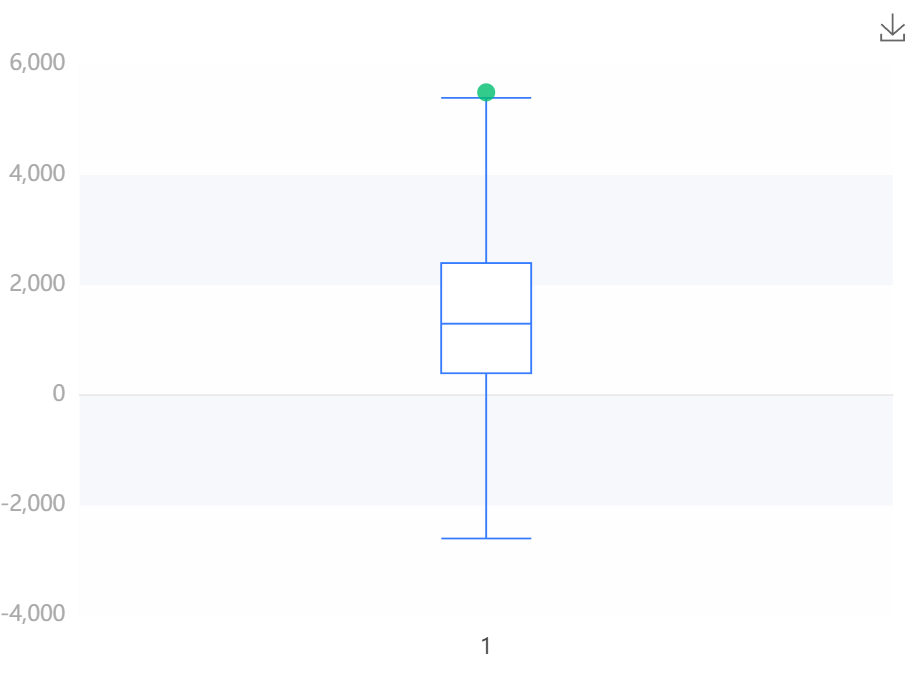


**图表说明：**上图是种植成本/(元/亩)计算观测的累计概率（P）与正态累计概率（P）的拟合情况。拟合程度越高越服从正态分布。● 图示最大仅显示1000个样本量信息，若样本量大于1000，则在全样本中进行随机抽样，选取其中1000样本进行绘制。

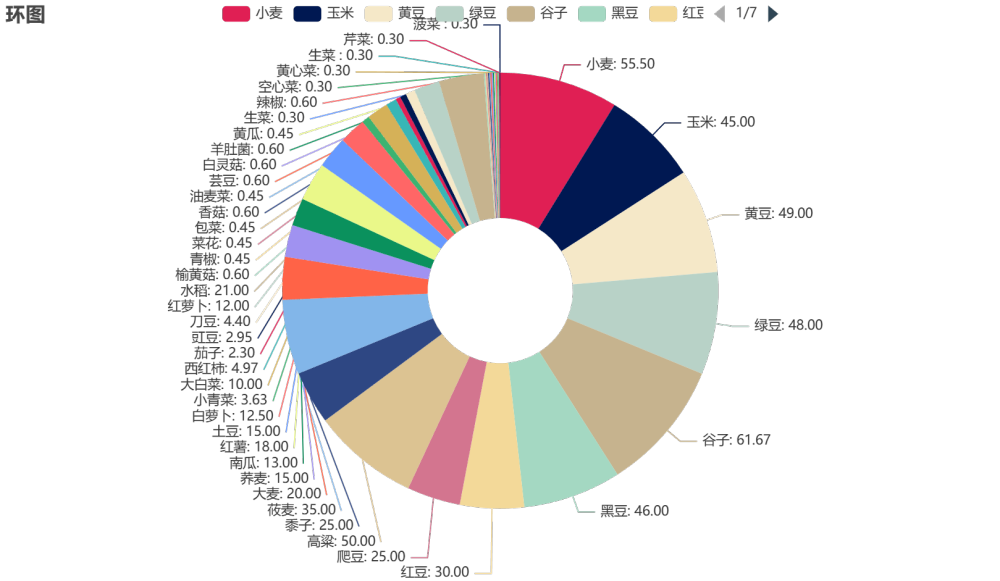
### 数据可视化

亩产量/斤箱型图





**图表说明：**上图以箱线图的形式展示了亩产量/斤、种植成本/(元/亩)频数分析离散趋势分析的结果，离散趋势用极大值、极小值、25%分位数、中位数、75%分位数等统计指标对数据分布进行差异（稳定性）测量。 PS:极大值、极小值并非该数据的最大值、最小值，该值为箱线图的内限，即大于极大值或小于极小值的点视为异常点。

种植面积饼图

作物类型柱状图

