# 结构化知识切分

## 切分知识对象

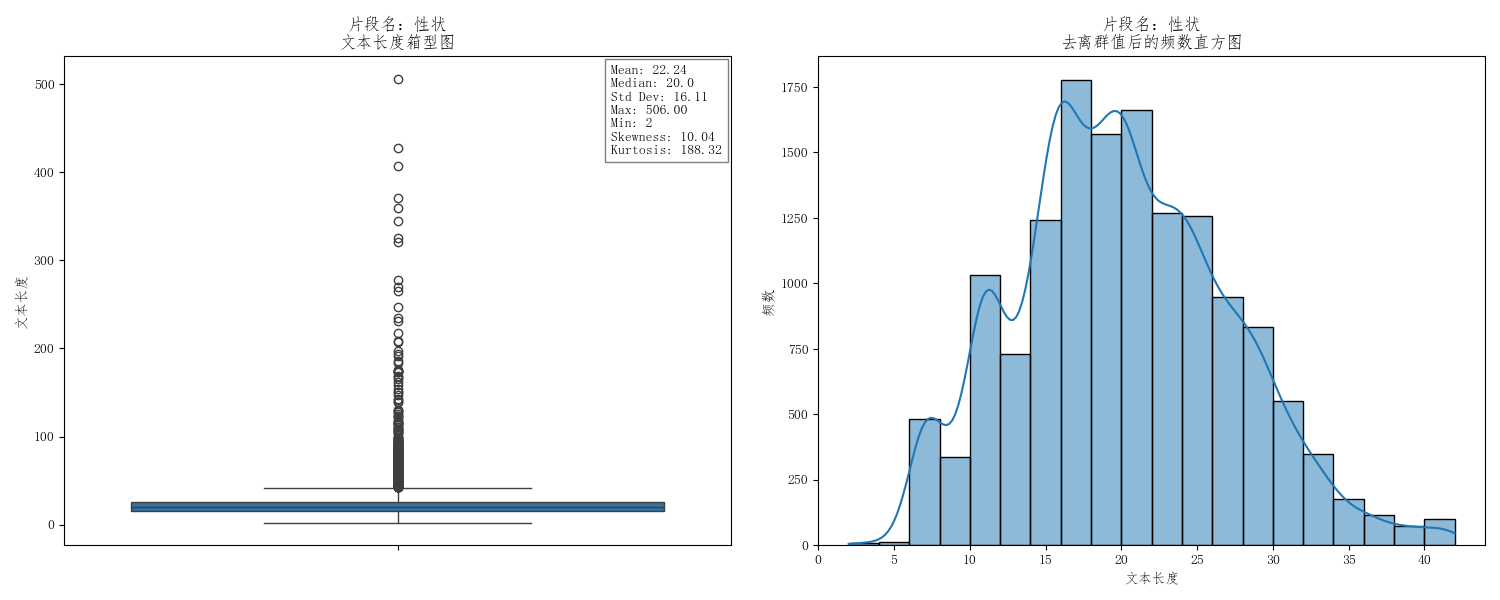
* 1. 疾病知识
     1. 数据量
     2. ..
  2. 药物知识
     1. 总体数据量：107,523条数据，88,805种药物
     2. 各片段数据量

|  |  |  |
| --- | --- | --- |
| **片段名** | **非空值数量** | **去重后的非空值数量** |
| 英文名称 | 14,232 | 11,690 |
| 药理分类 | 38,779 | 1,477 |
| 标准名称 | 34,523 | 10,936 |
| 商品名 | 12,323 | 10,945 |
| 通用名 | 10,859 | 7,605 |
| 规格 | 15,392 | 8,463 |
| 成份 | 14,762 | 12,780 |
| 性状 | 15,185 | 10,507 |
| 适应症 | 61,147 | 50,408 |
| 禁忌证 | 16,890 | 8,104 |
| 用法用量 | 16,058 | 13,254 |
| 药物过量 | 4,870 | 2,760 |
| 不良反应 | 16,253 | 9,957 |
| 老年用药 | 6,084 | 2,494 |
| 儿童用药 | 21,361 | 2,982 |
| 孕妇及哺乳期妇女用药 | 5,471 | 3,119 |
| 药代动力学 | 9,244 | 7,325 |
| 药物相互作用 | 9,596 | 4,798 |
| 药理作用 | 12,796 | 10,725 |
| 注意事项 | 20,338 | 15,319 |

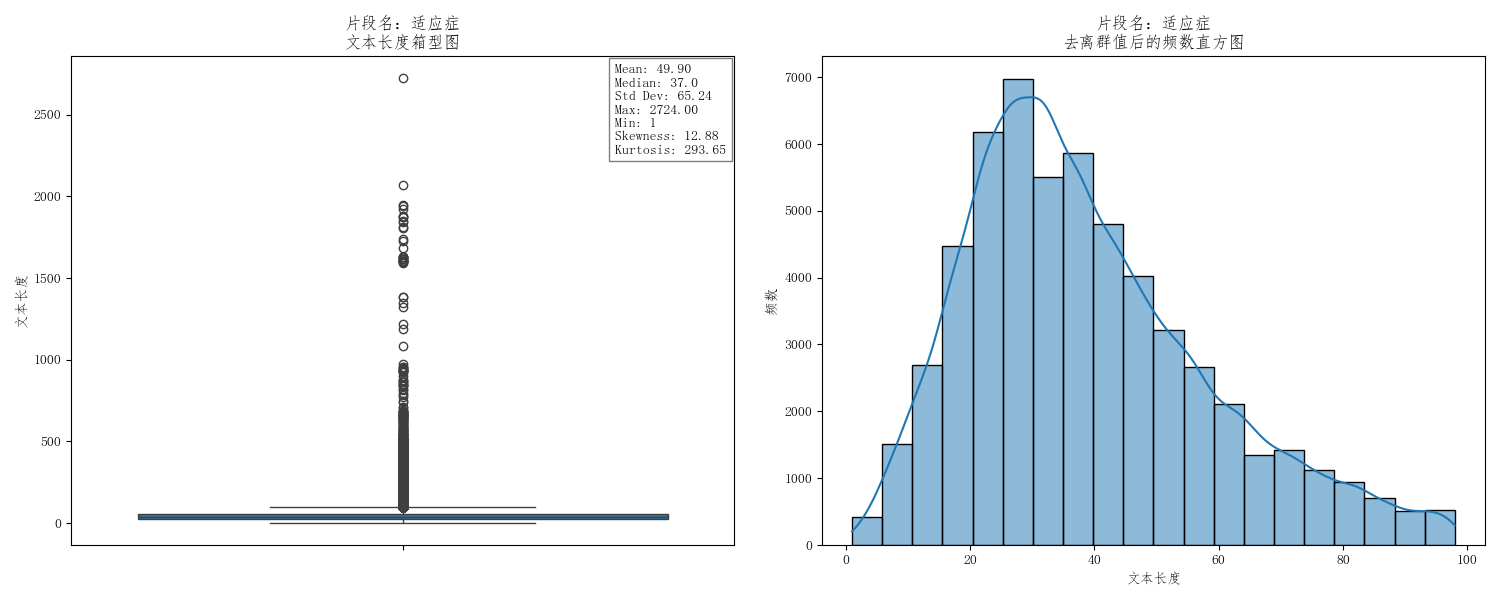
* + 1. 各片段原始文本长度

由于英文名称、药理分类、标准名称、商品名、通用名、规格及成分具有固定格式，大多为短文本，故不统计，后续也不进行切分。

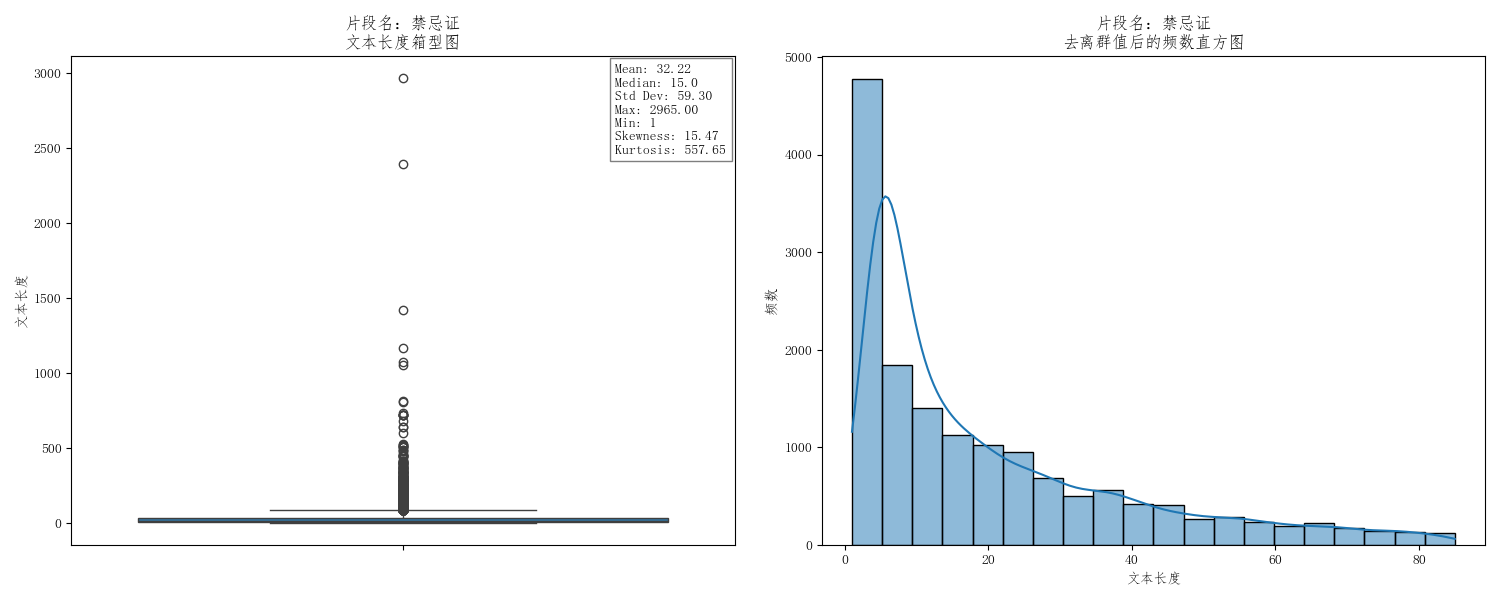
其余片段的原始文本长度分布范围较大，各片段长度的箱型图、去离群值(Q3+1.5IQR)后的频数直方图如下：



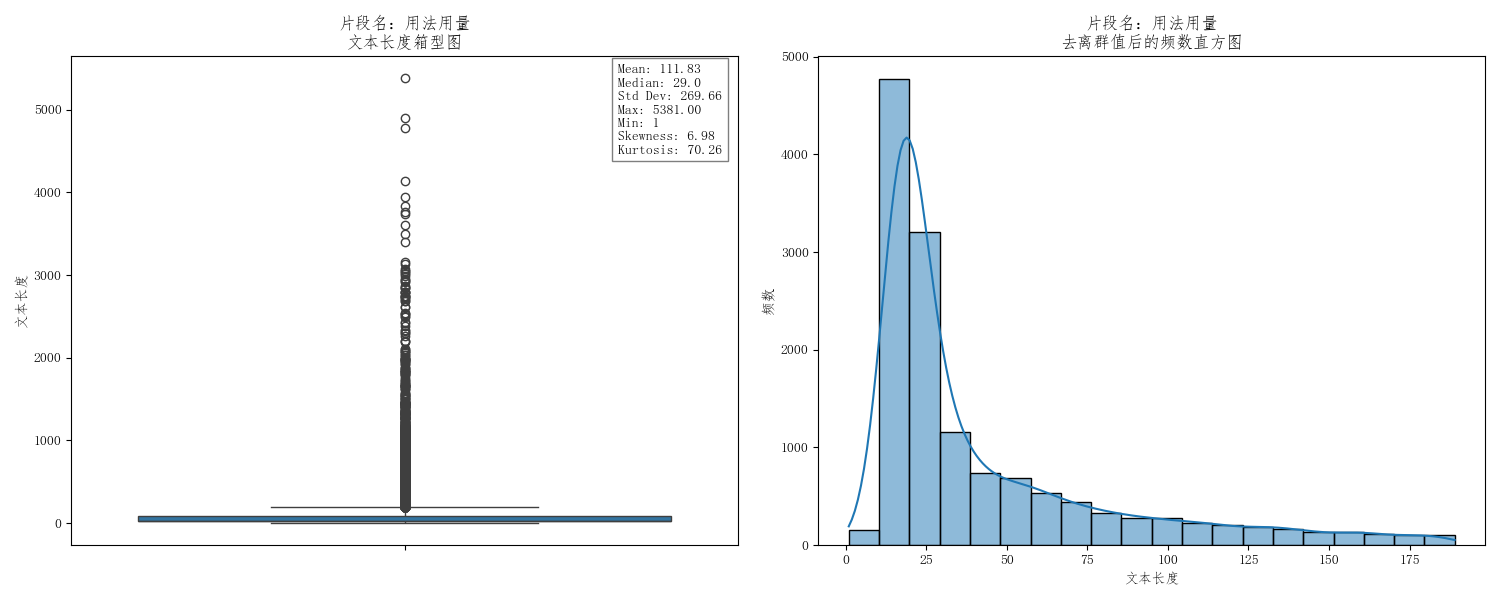
性状描述了药品的外形、颜色、味道、气味、溶解性等物理性质，长度较短，大多集中在2至40字，均值为22字，最长506字。



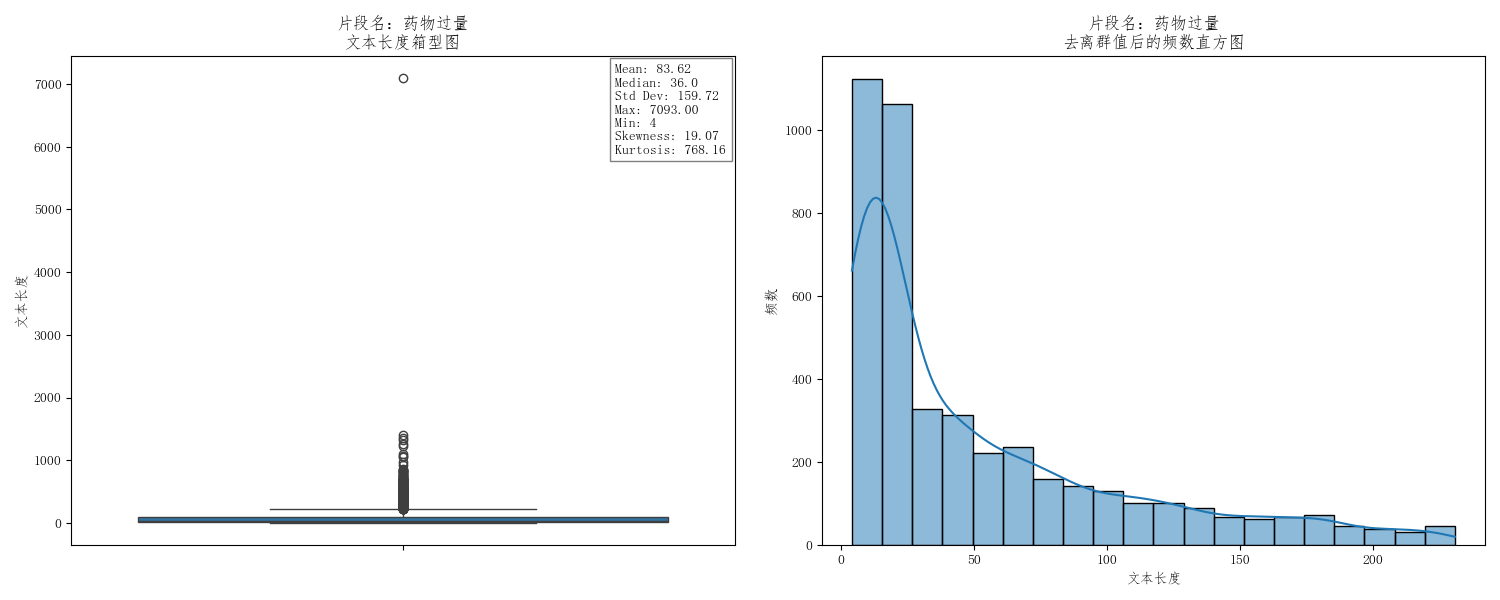
适应症描述了药品的使用时间、对应症状、用药部位，长度适中，大多集中在1至100字，均值为49字，最长为2724字。



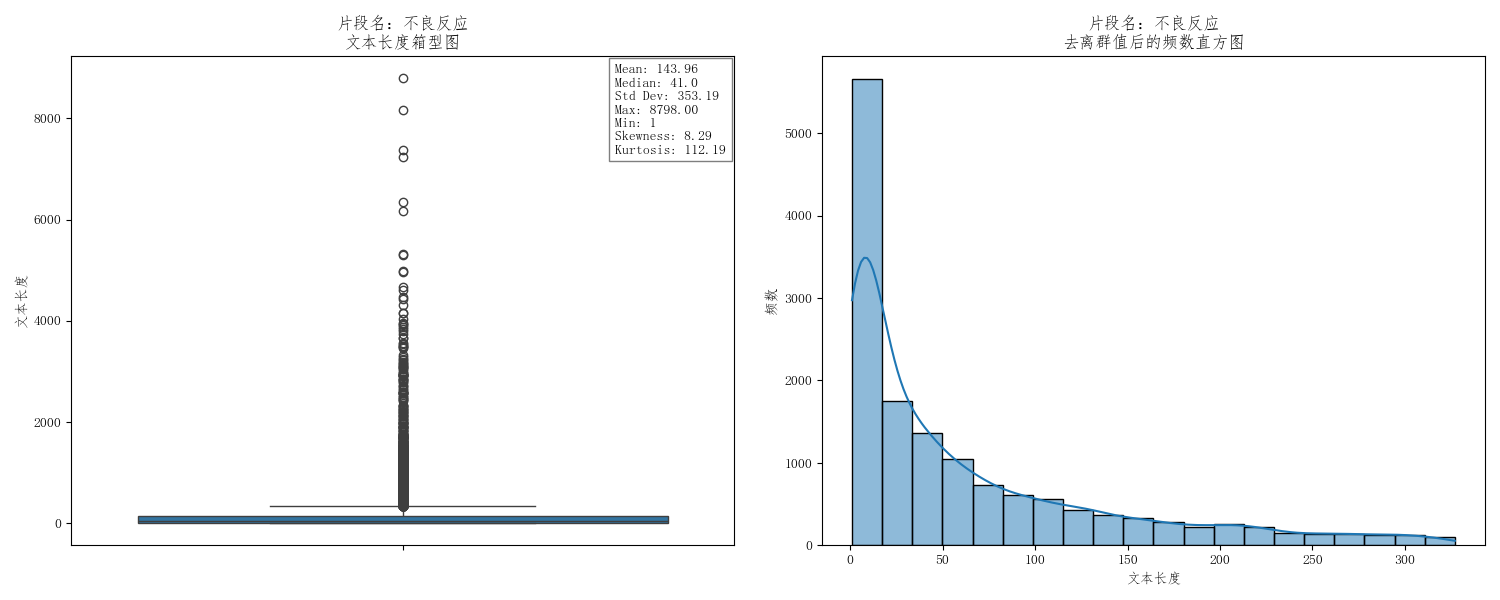
禁忌证规定了禁止使用该药品的人群特征，如患特定疾病等。字段长度大多集中在1至80字。由于多数禁忌证为“尚不明确”，故均值较小，为32字，最长的字段为2965字。



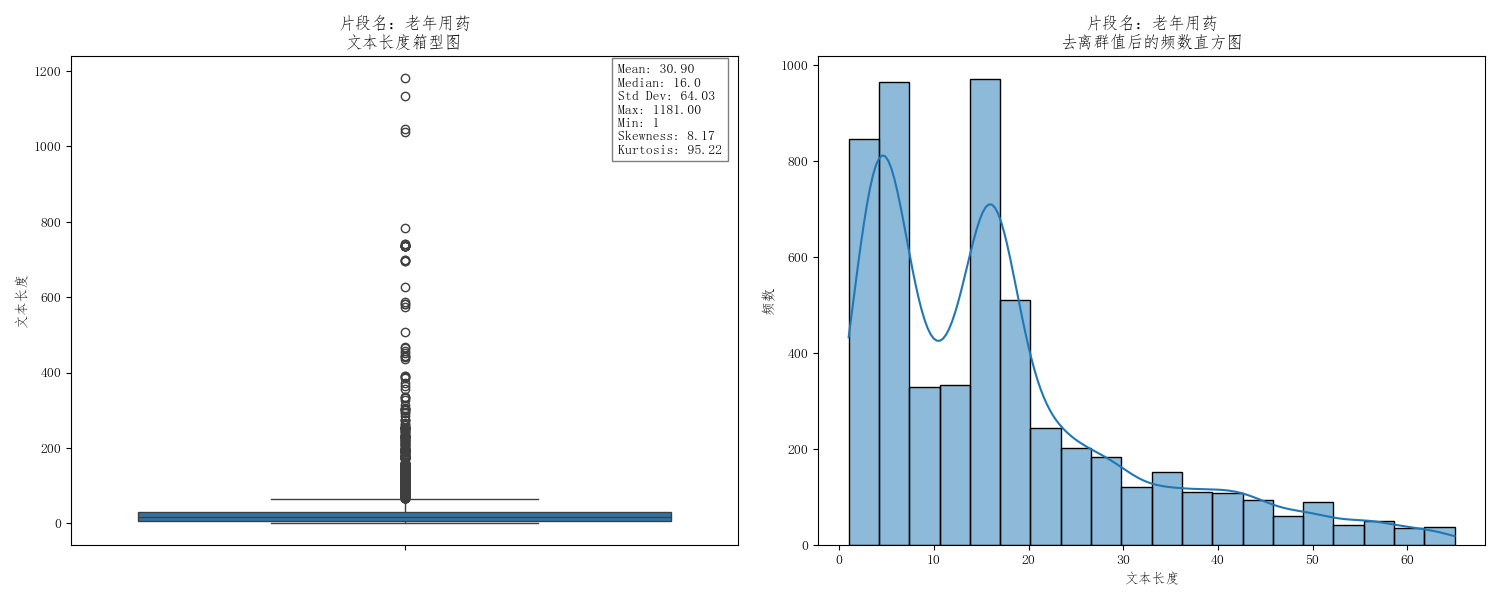
用法用量规定了服药方式、用药规格和用药频率，长度主要集中在1至175字。由于部分药品对用法用量的规定非常详细，拉高了均值，使得均值（111字）与中位数（29字）相差较大。最长的字段为5381字。



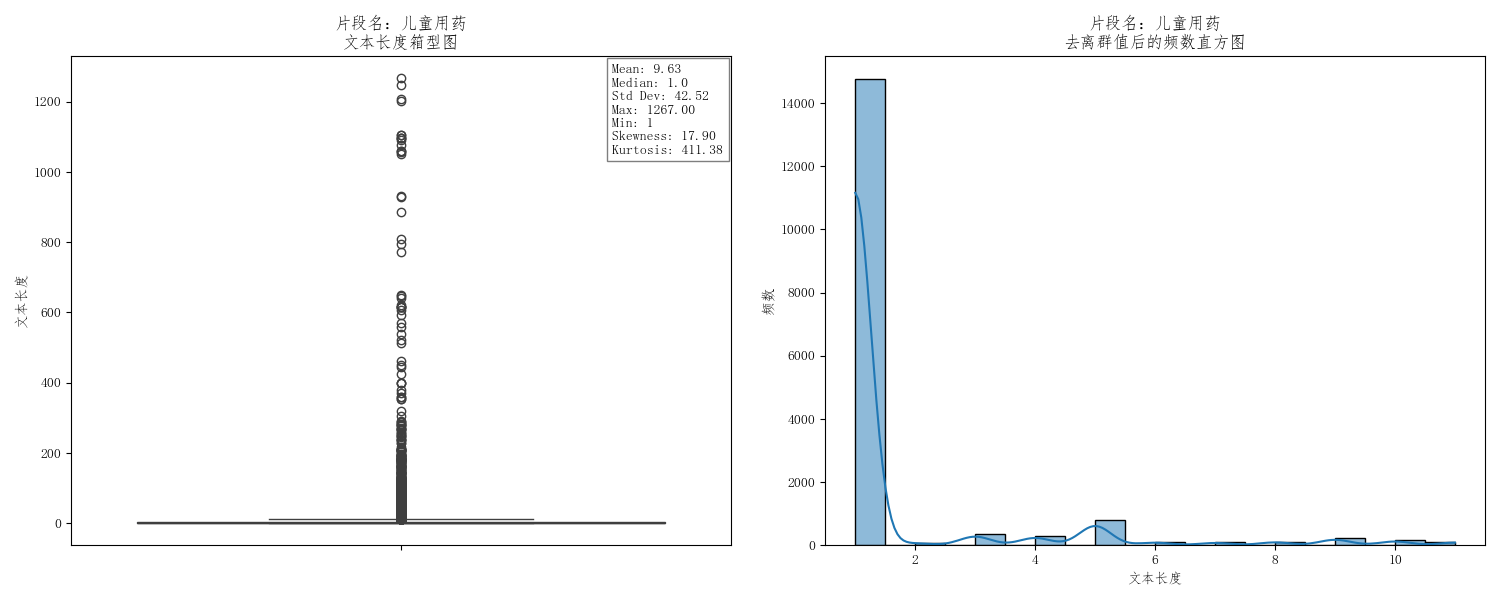
药物过量描述了过量服用药物的症状及解救措施，长度主要集中在4至200字，均值为36字，最长为7093字。



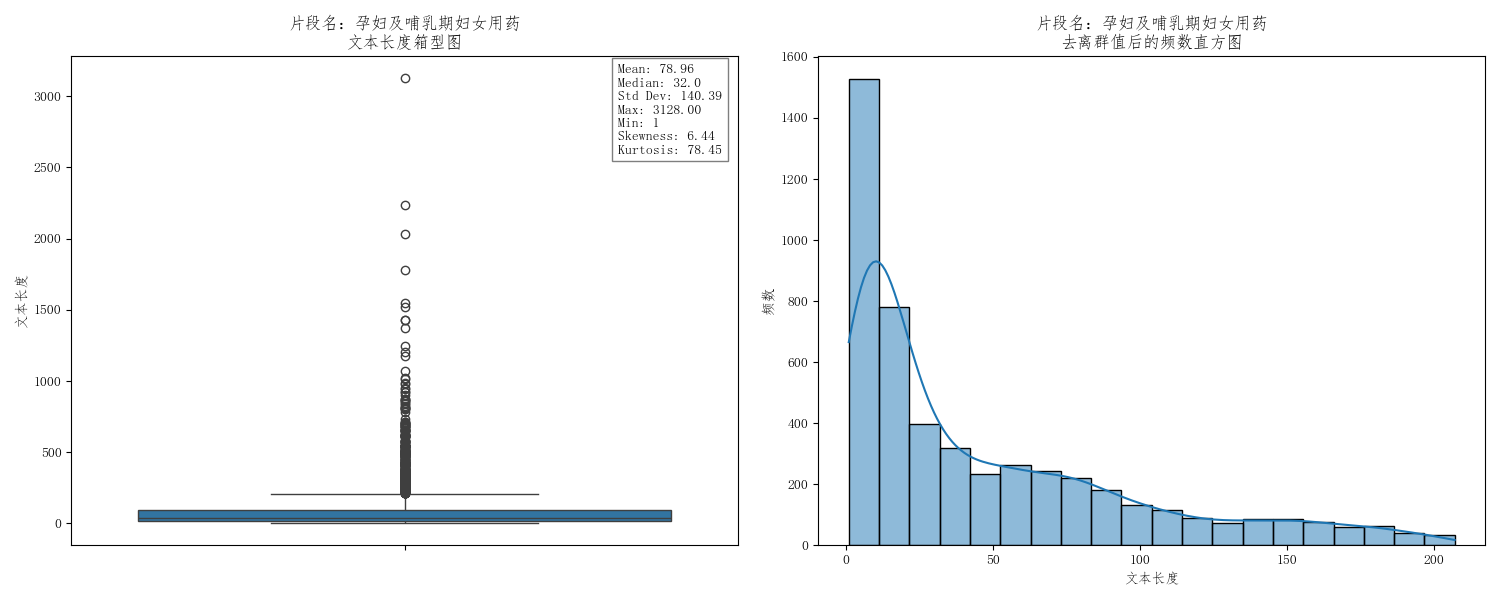
不良反应描述了用药可能导致的副作用及发生率，长度主要集中在1至300字。由于部分药品对不良反应的描述非常详细，拉高了均值，使得均值（143字）与中位数（41字）相差较大。最长的字段为8798字。



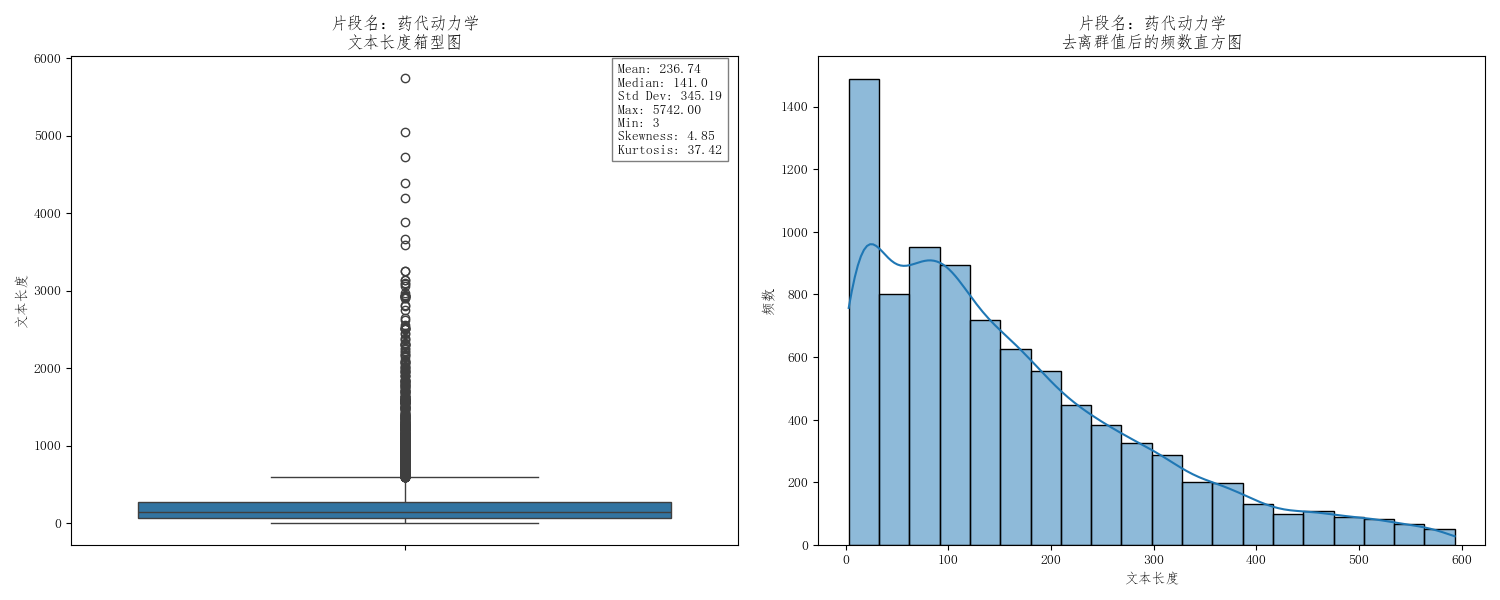
老年用药、儿童用药、孕妇及哺乳期妇女用药分别规定了对三类特殊人群的药物使用方法及禁忌证。老年用药的文本长度集中在1至60字，均值为30字，最长为1181字。



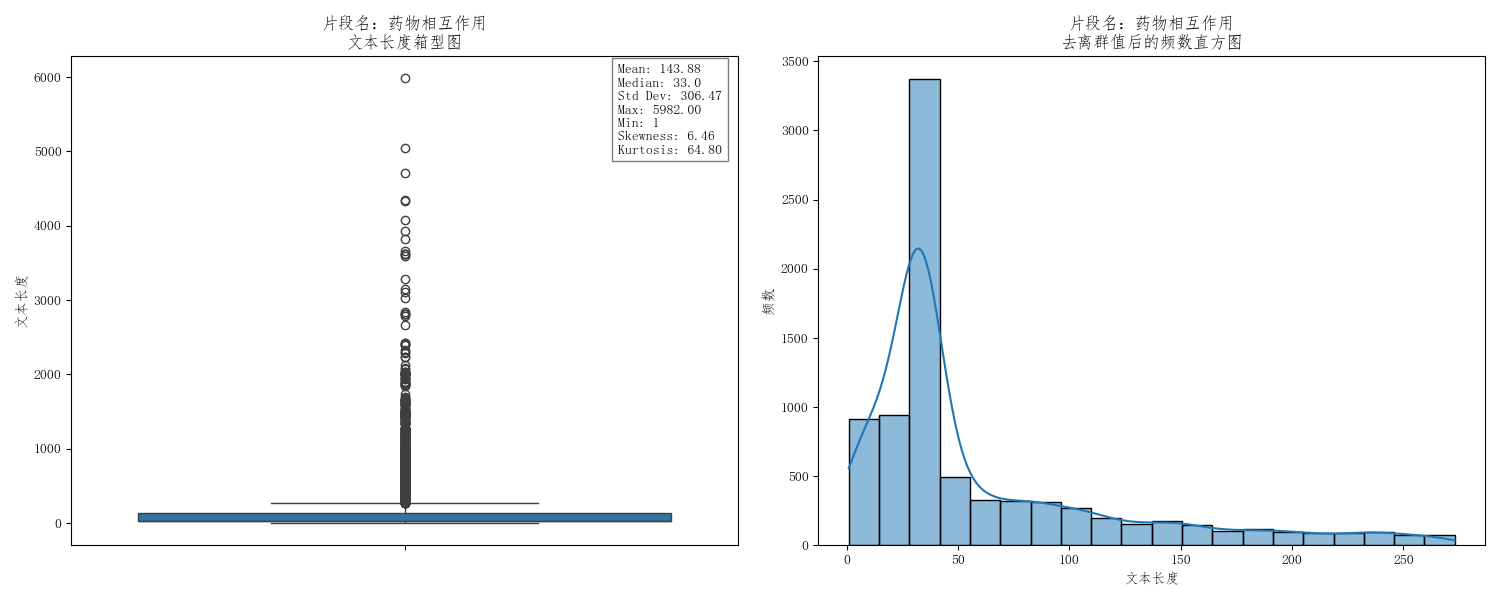
儿童用药的文本长度集中在1至10字，基本为“尚不明确”、“遵医嘱”等短文本，均值为9字，最长为1267字。



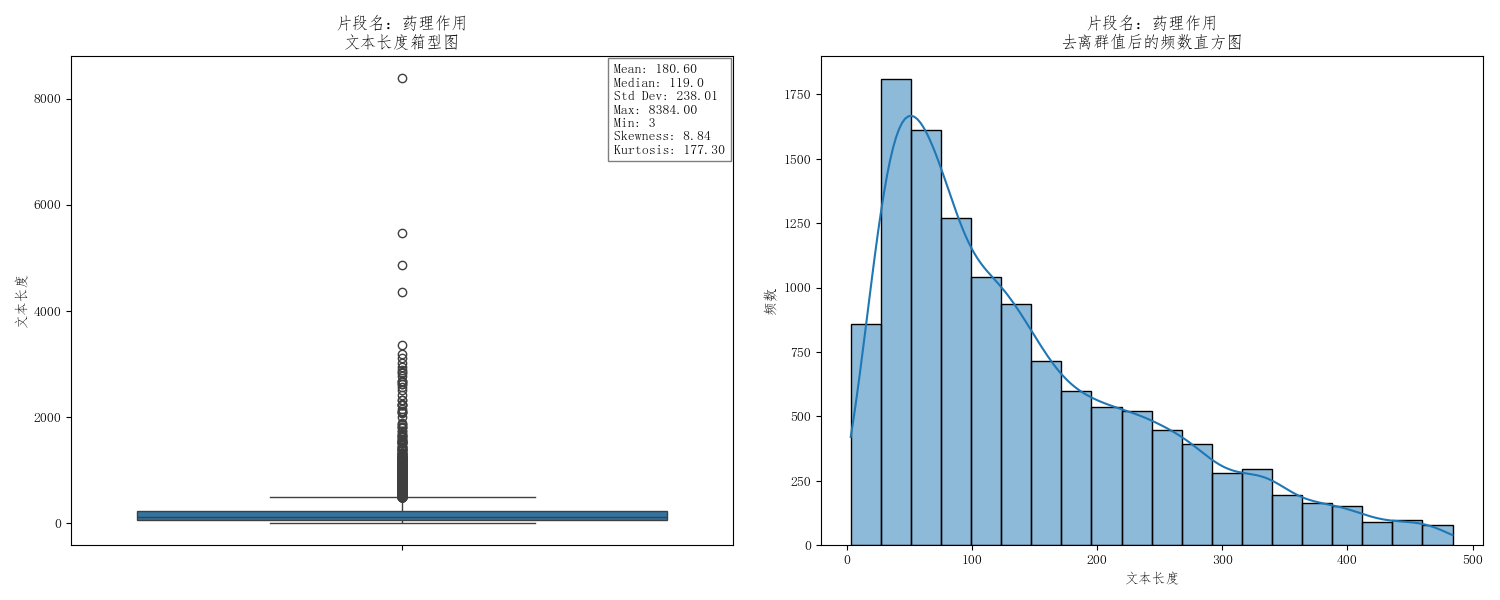
孕妇及哺乳期妇女用药的文本长度集中在1至200字，相比于老年和儿童用药的数据更多、描述更为详细，部分长文本拉高了均值，使得均值（78字）与中位数（32字）相差较大，最长文本为3128字。



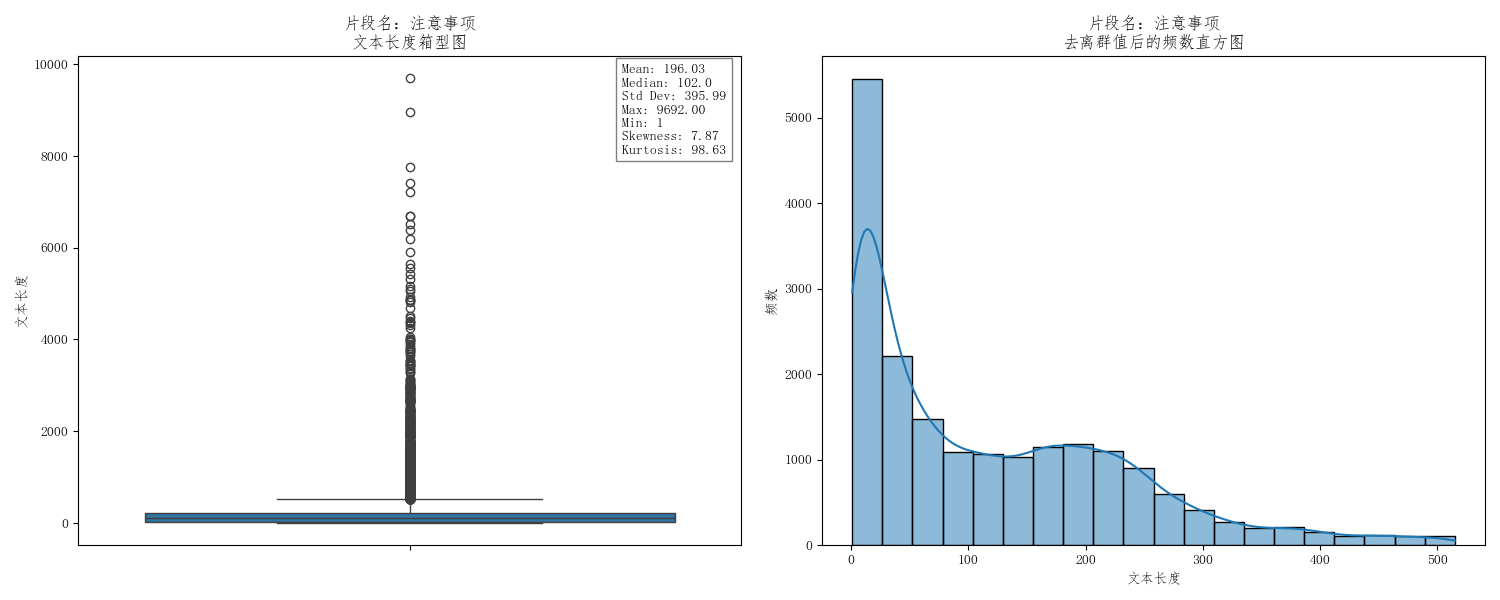
药代动力学描述了体内药物浓度随时间变化的规律，即药物在体内的吸收、分布、代谢、和排泄过程。文本较长，长度主要分布在3至600字，均值为236字，中位数为141字，最大值为5742字。



药物相互作用是指在一定时间内先后服用两种及以上药物后产生的复合效应，可能产生药效增强或减弱，也可能会加重或减弱药物的不良反应。文本长度主要分布在1至250字，均值为143字，中位数为33字，最长为5982字。



药理作用阐述药物与临床适应症相关已明确的药理作用，包括药物类别、药理活性、作用机制等，长度主要集中在3至500字，均值为180字，中位数为238字，最长为8384字。



注意事项的文本长度主要集中在1至500字，均值为196字，中位数102字，最长为9692字。

分析去离群值的直方图得知，药代动力学、药理作用、注意事项中长文本较多，其余片段长度均集中在300字以内，属于短文本。

## 切分方法

1. 切分原理

利用BAAI General Embedding（BGE）模型，将关于疾病及药品知识的文本转化为1024维向量。并计算相邻句子的文本相似度，将其作为切分子片段的依据。若相邻句子的语义越相近，文本相似度就越高，表明它们更有可能属于同一子片段；反之，若文本相似度较低，则说明应进行切分。这种切分方法有助于根据语义关系对文本进行划分，提高信息片段的准确性和内容一致性。

1. 切分步骤

首先对原始文本进行判断：若文本属于具有固定格式的短文本，即英文名称、药理分类、标准名称、商品名、通用名、规格及成分，则不进行切分，最终储存的子片段即原始片段；若文本长度小于300字，则同样认为不需要切分；除这两种情况以外，其他文本进入切分流程。

进入切分流程后，将片段切分为单个句子，计算每个句子的向量表示，并进行L2范数归一化。然后计算整个片段中两两句子间的语义相似度，得到矩阵score，score[i,j]即第i个句子与第j+1个句子的语义相似度，只提取score对角线元素即相邻句子的相似度作为切分依据。若相邻句子相似度大于0.7，则认为属于同一子片段，不需切分；若单个句子长度小于30字，则认为该短句子应合并到当前段落，也不需切分；反之，若相邻句子相似度小于等于0.7，或当前段落长度超过最大阈值300字，则进行切分。

切分流程完成后，发现原始字段中有1050个大于350字的片段未作切分，提示“无法计算相似度”。将这些片段提取出来，单独进行二次切分。

## 切分参数

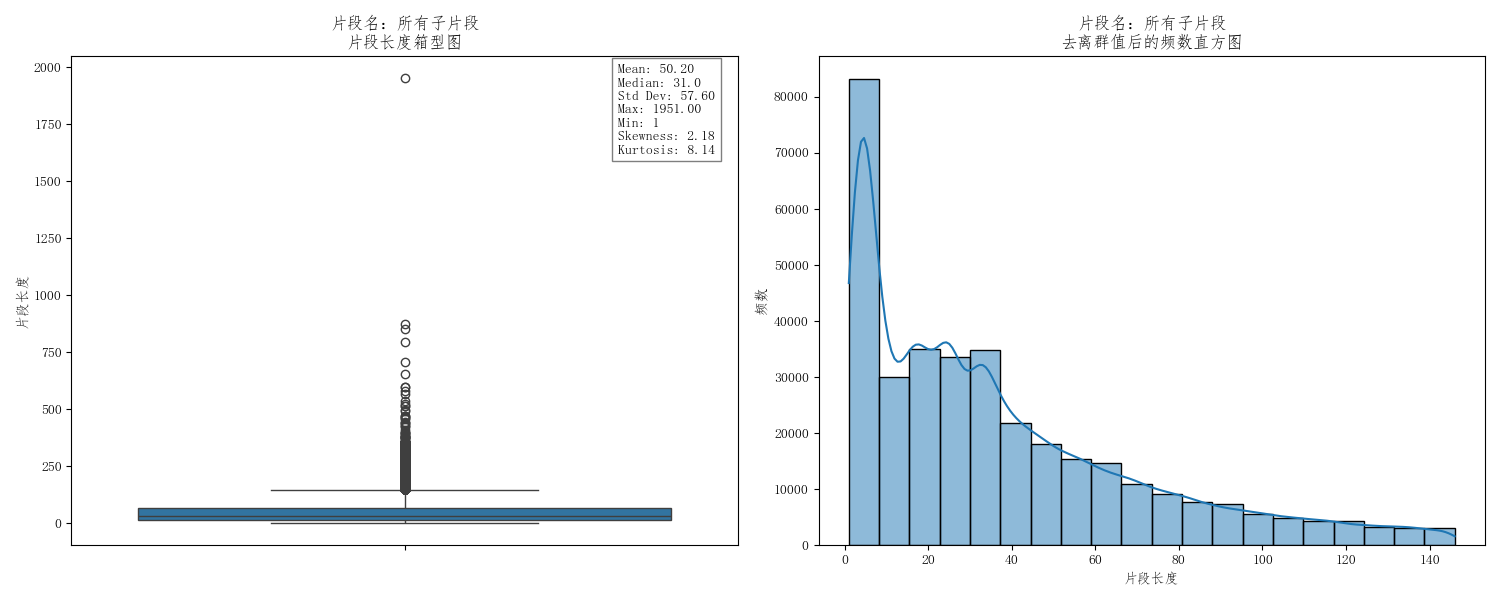
切分过程中需要规定3个参数，分别是子片段字数最大阈值、合并单句的字数最小阈值、句间相似度切分阈值。子片段字数超过最大阈值则截断，单句小于最小阈值则合并，句间相似度超过阈值则合并。最大阈值越大、最小阈值越大、句间相似度阈值越小，则单个片段的字数越多。

在尝试不同的参数范围时，将最大阈值设置在300至700字，最小阈值设置在10至30字，相似度阈值在0.6至0.8之间。目标是既要保证不会出现太多单句切割导致子片段数量过多的情况，也要避免切割次数过少，从而导致子片段内部语义差异较大。最终选择了300字作为最大阈值，30字作为最小阈值，相似度阈值设定为0.7。

## 切分结果分析

1. 疾病知识
2. 药物知识
3. 所有子片段

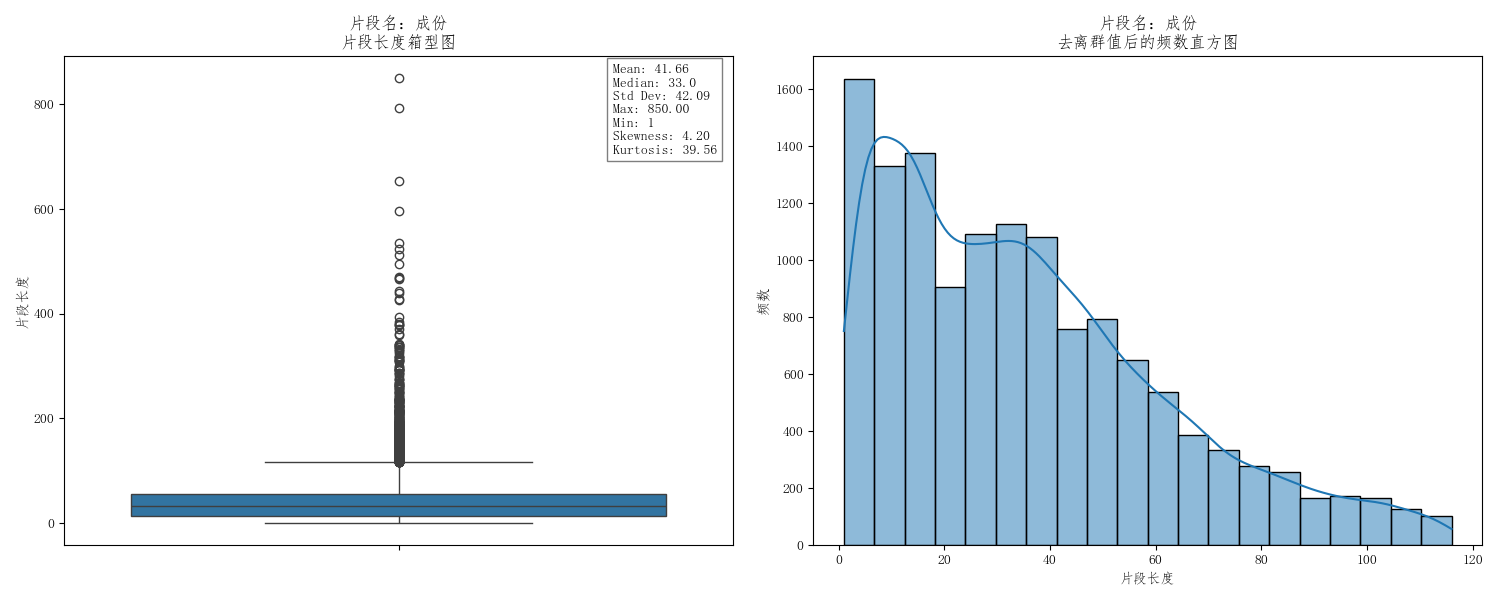
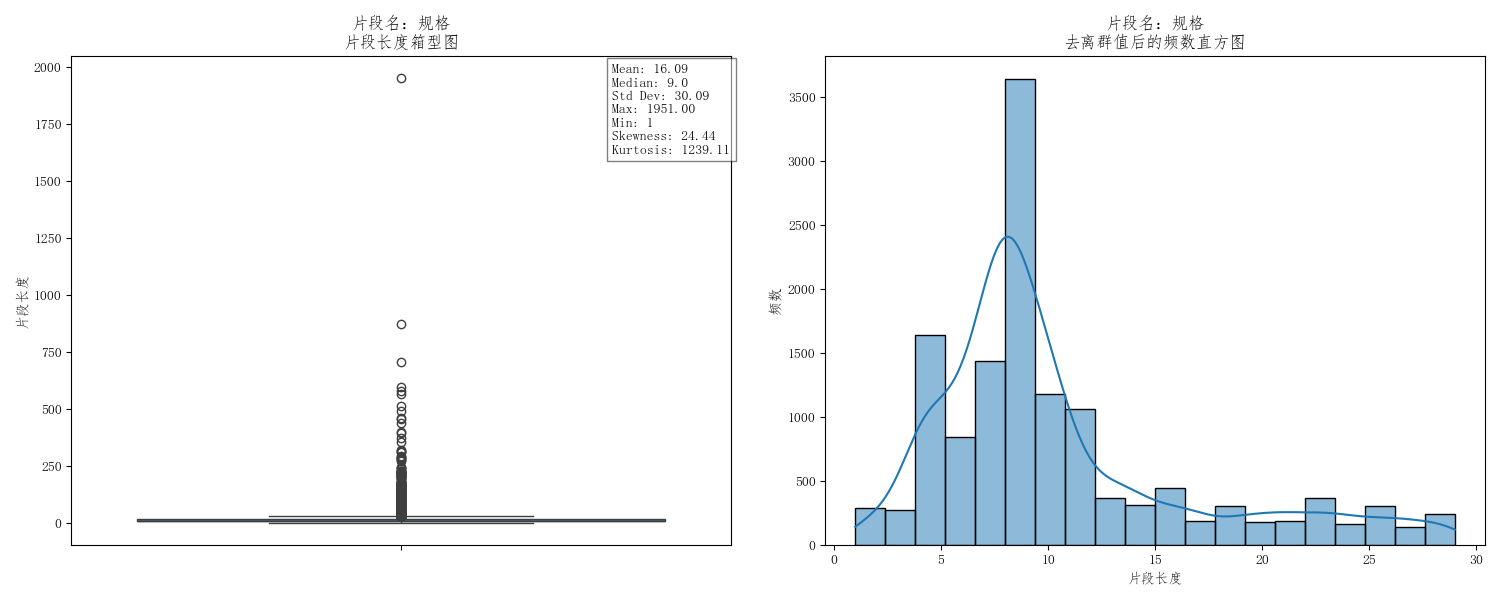
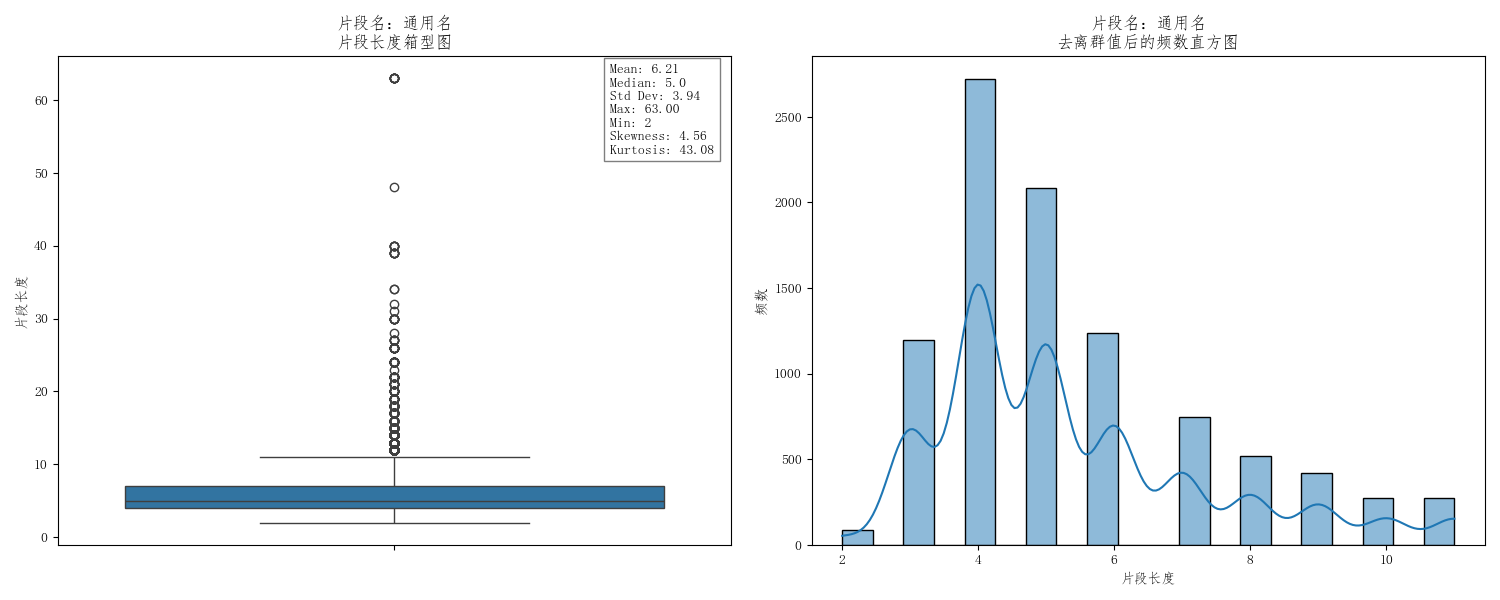
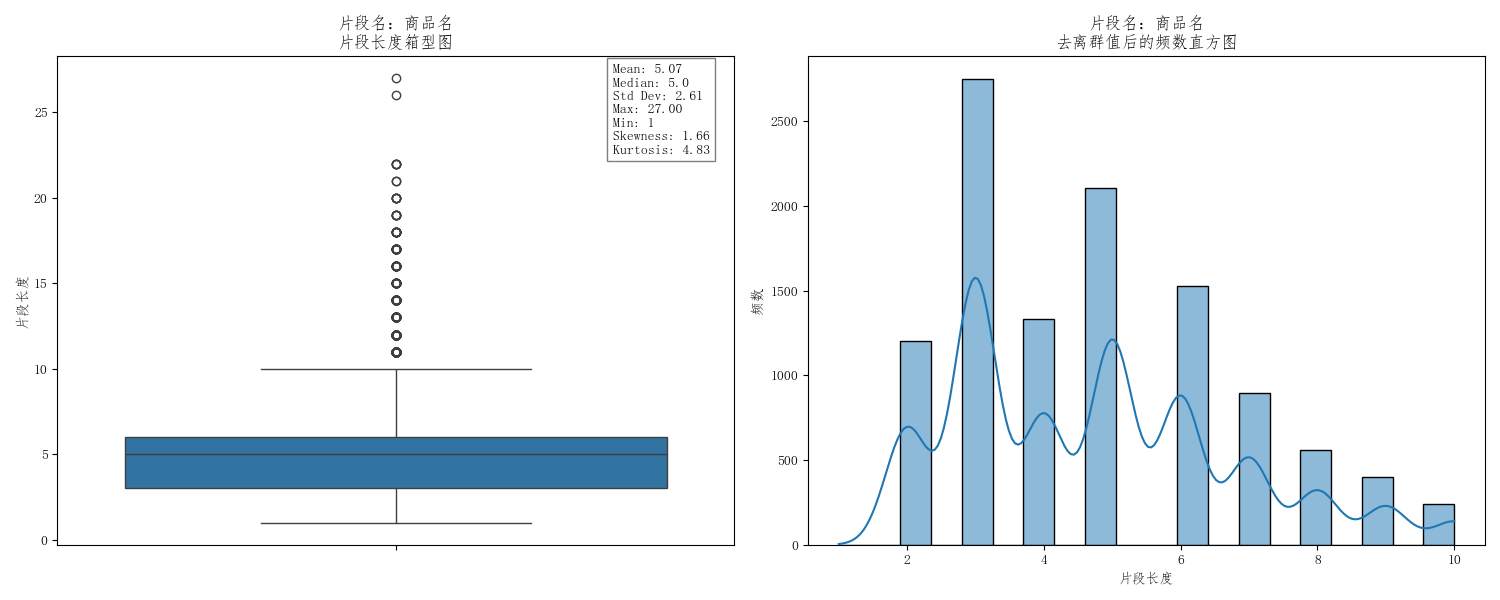
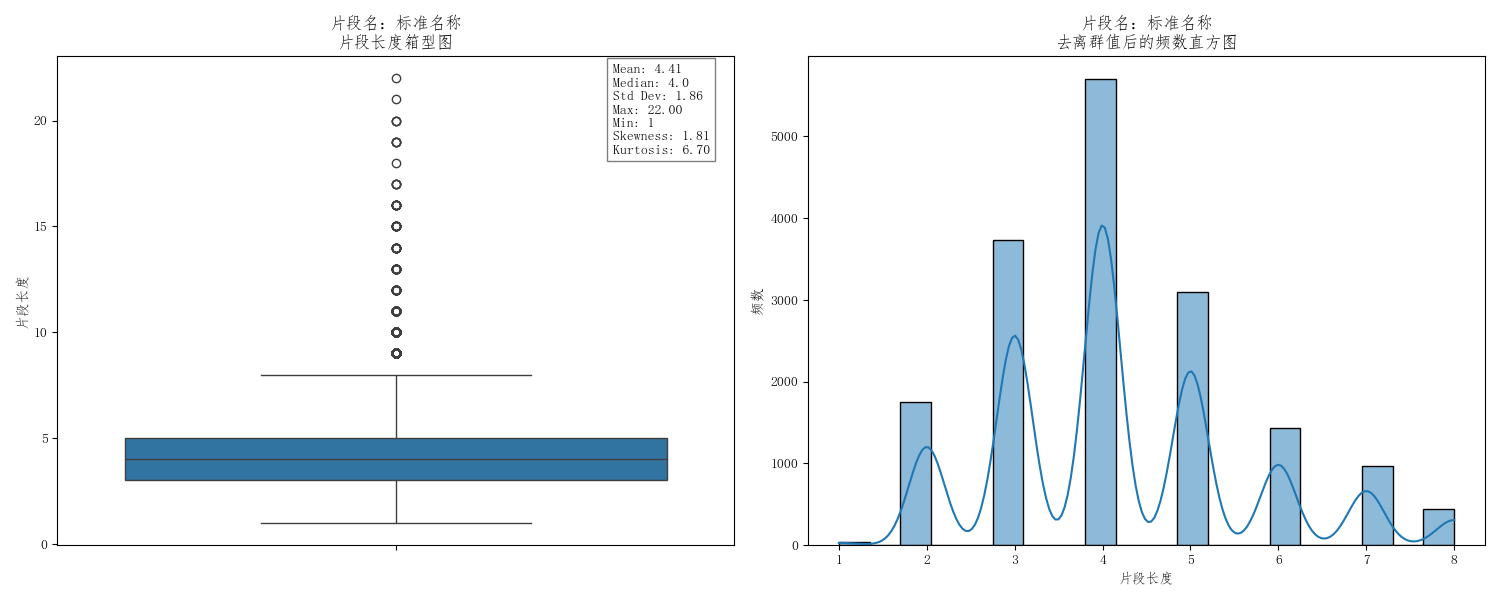
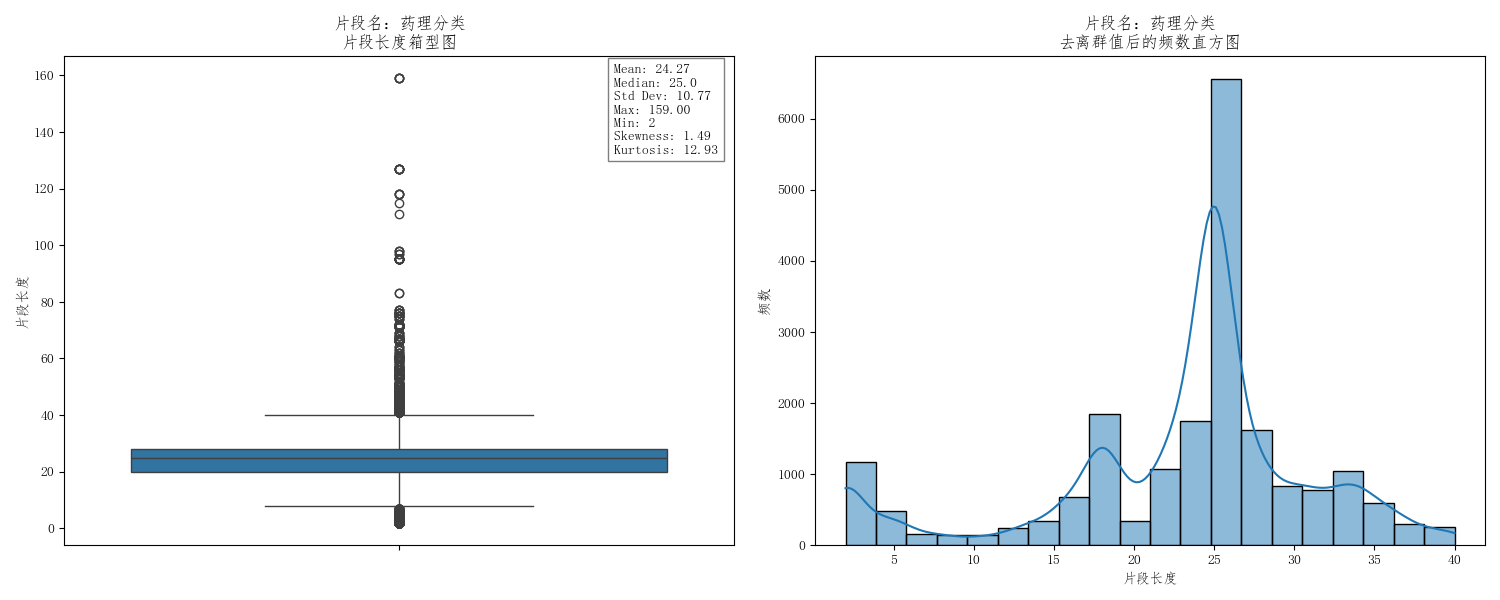
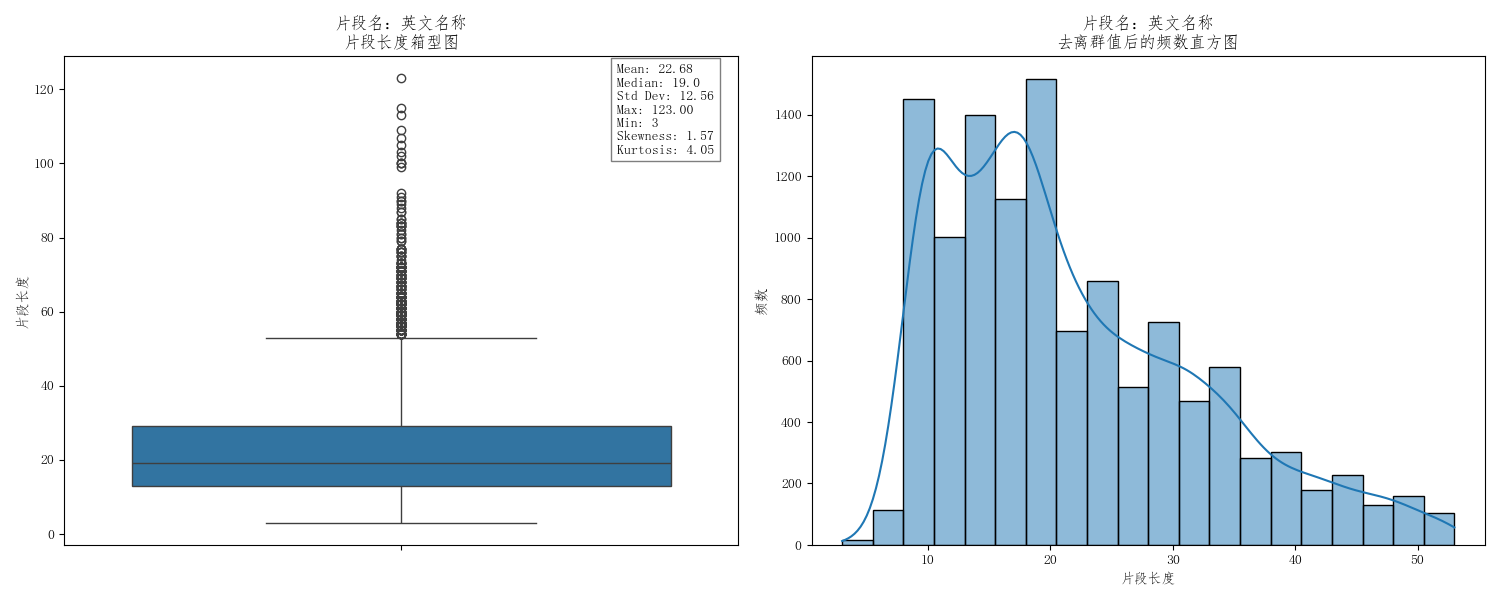
切分后生成子片段378032个，去重后剩余377829个。



切分后的所有子片段长度分布如上图所示。可以看到，子片段长度主要集中在1至140字，均值为50字，中位数为31字；最大值为1951字，字段名为“规格”，属于不切分字段，且该片段为英文，因此片段异常长。

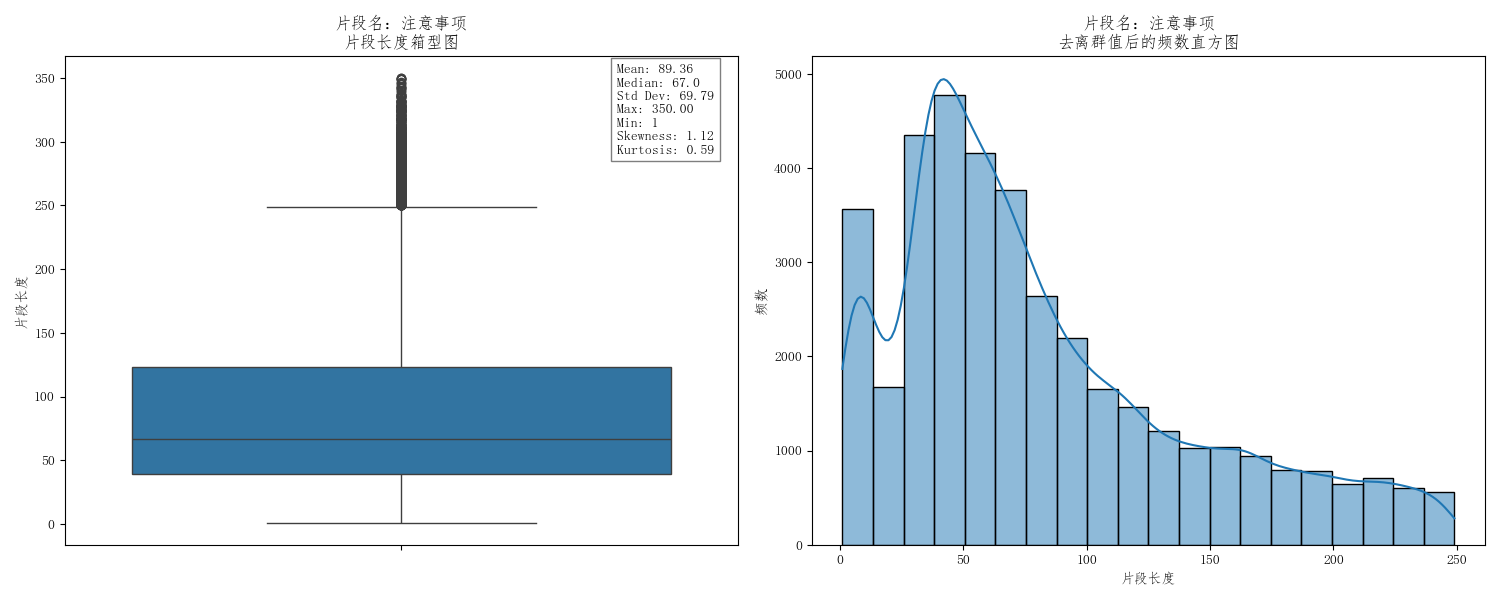
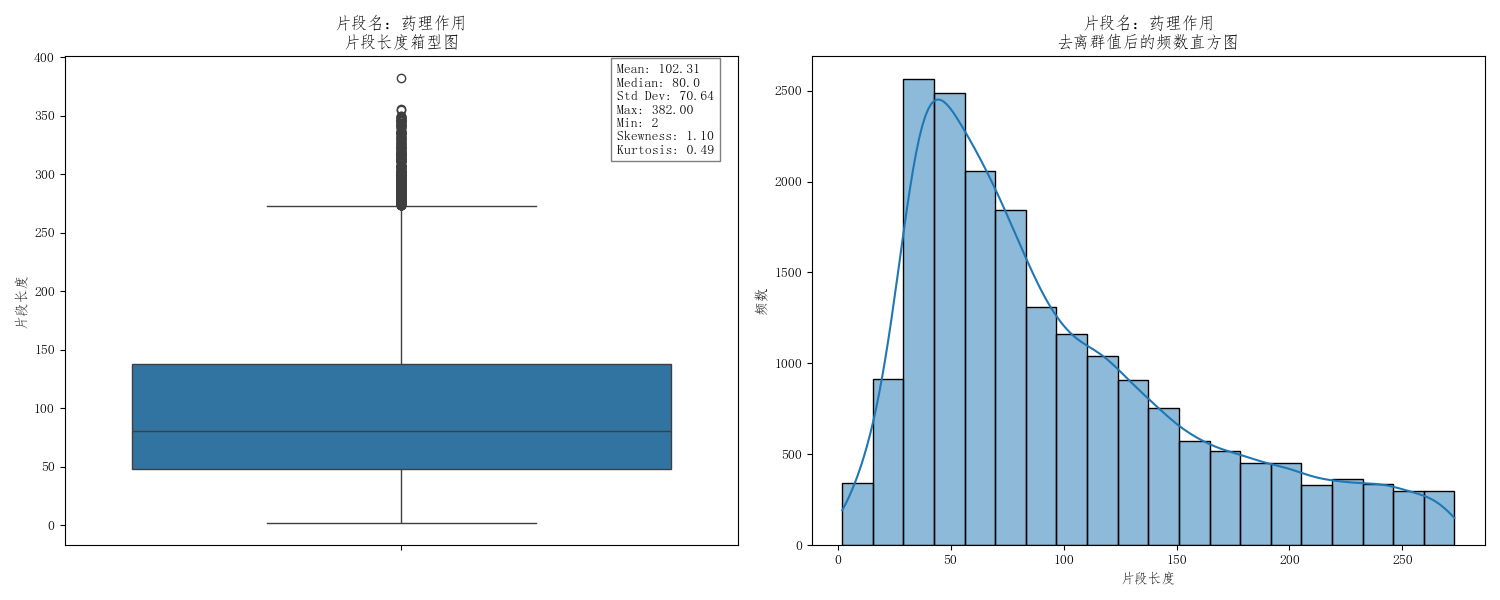
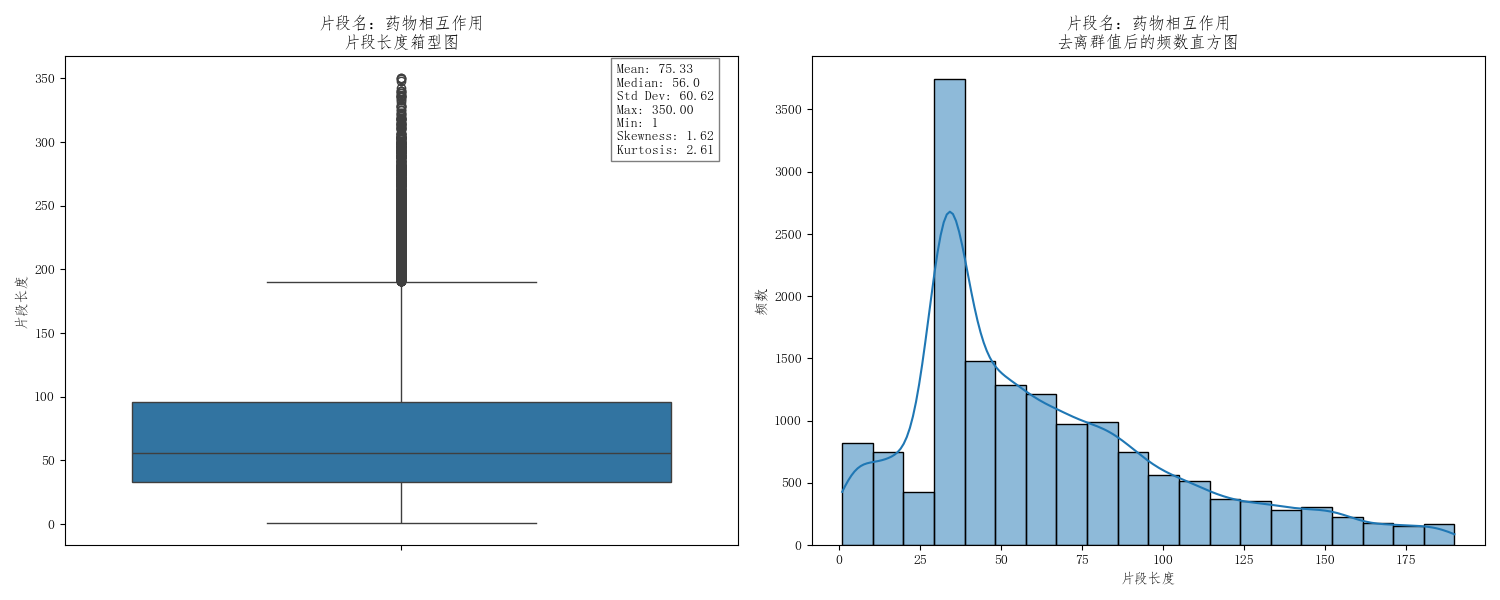
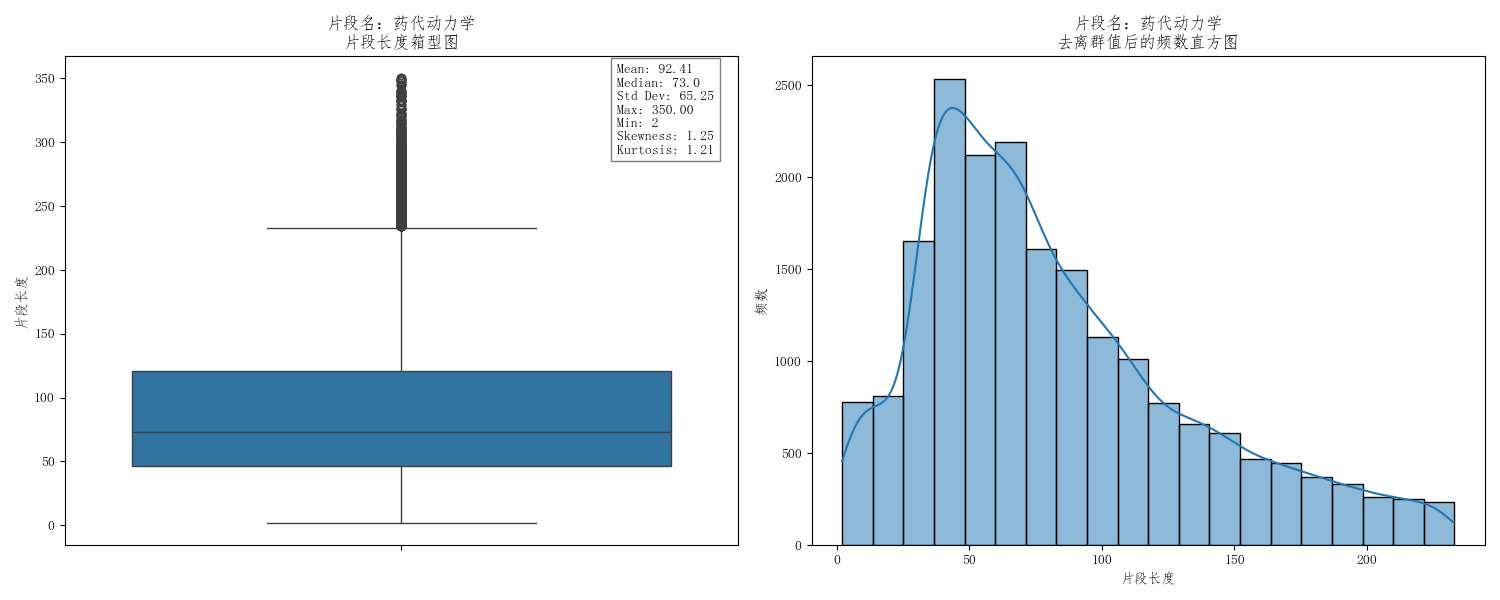
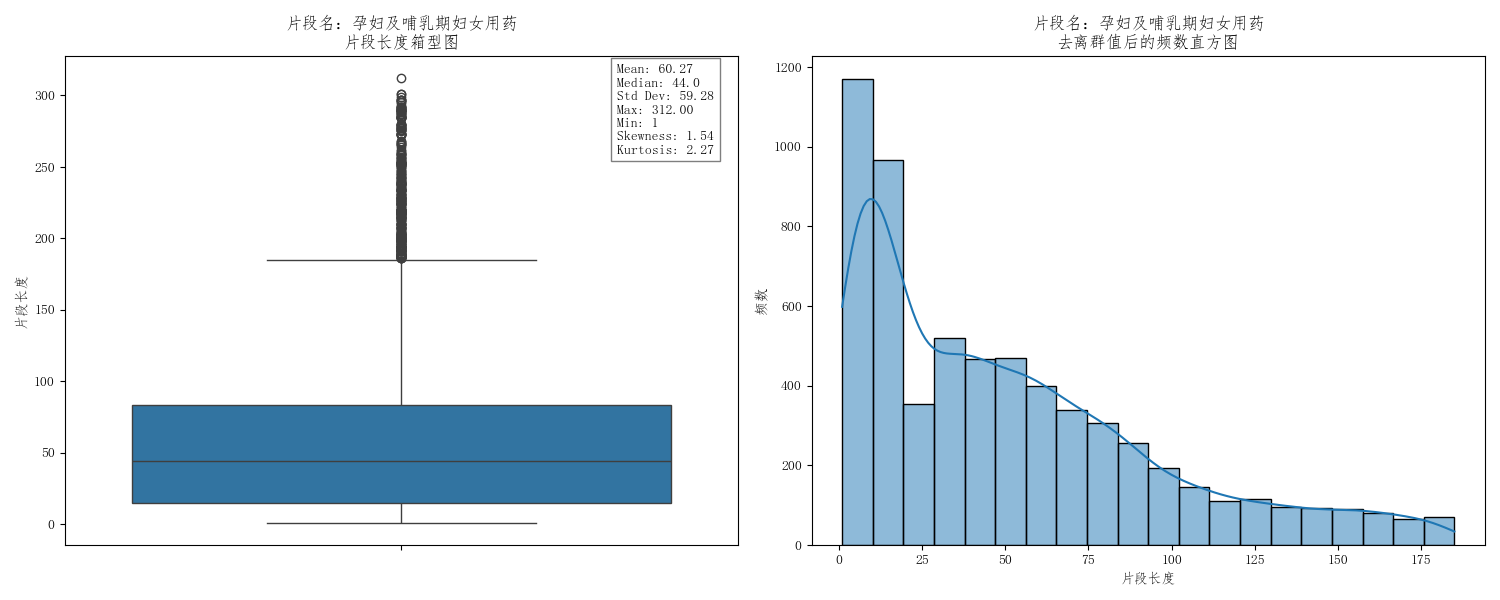
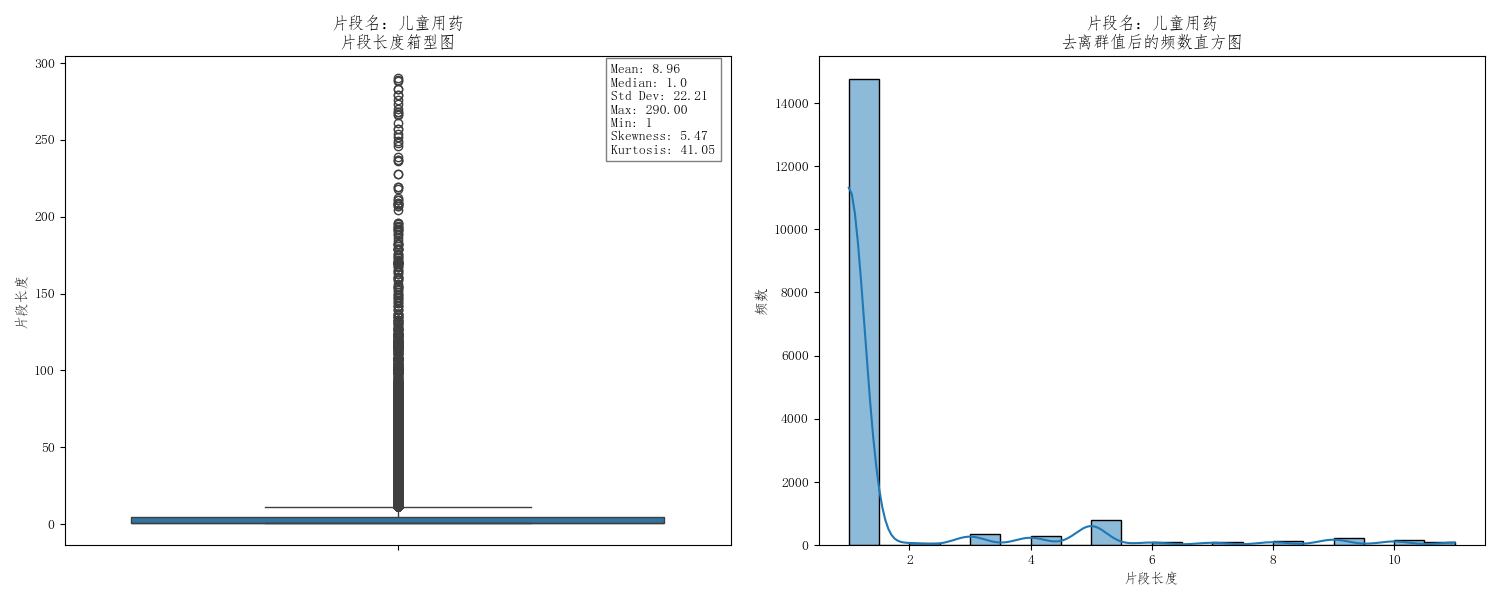
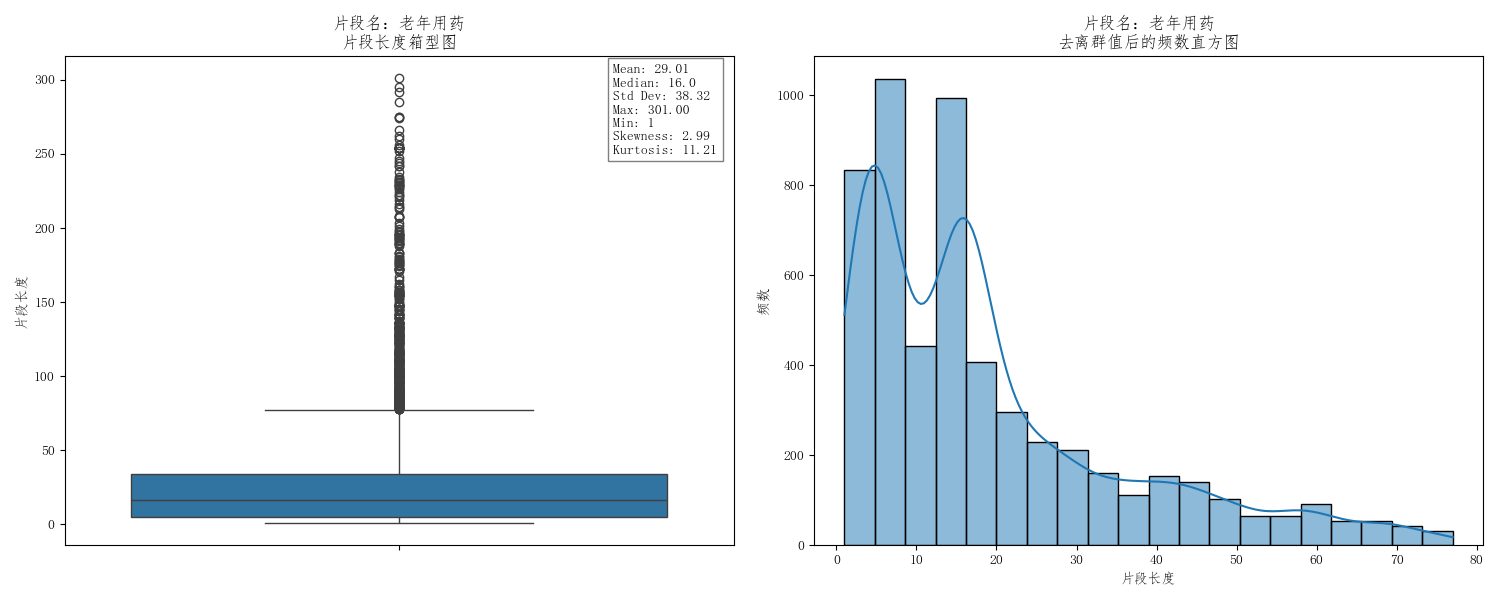
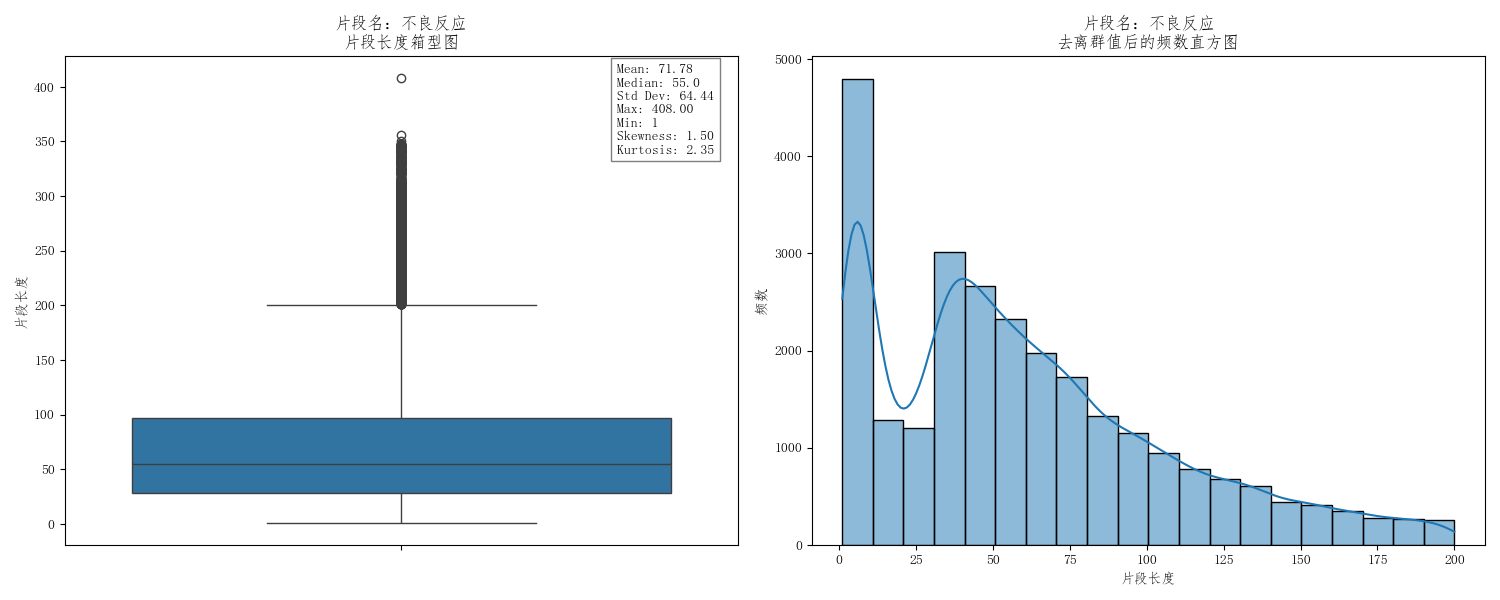
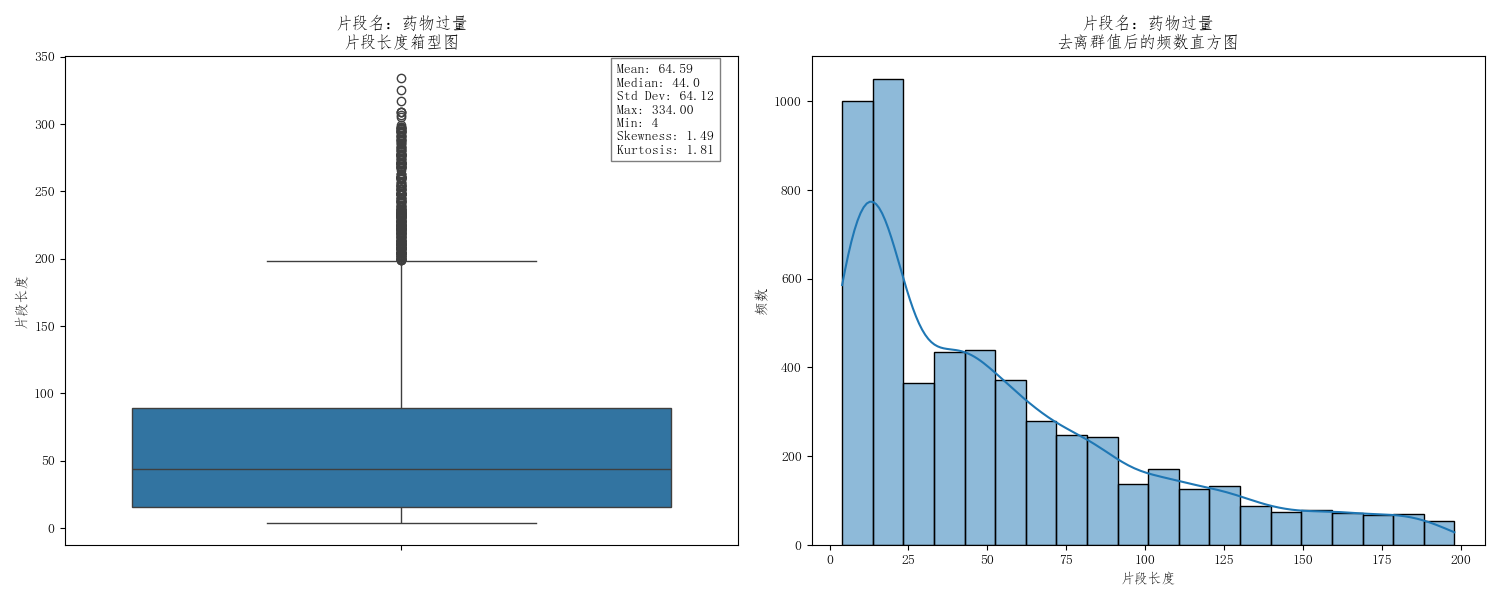
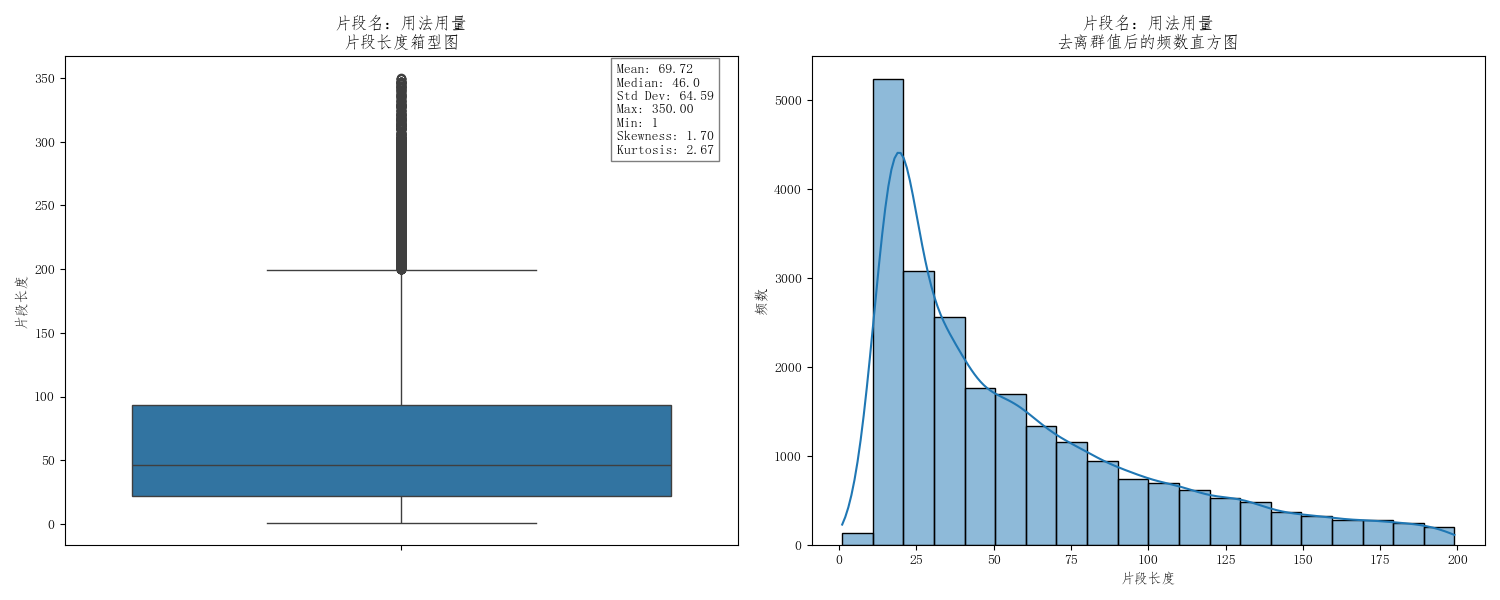
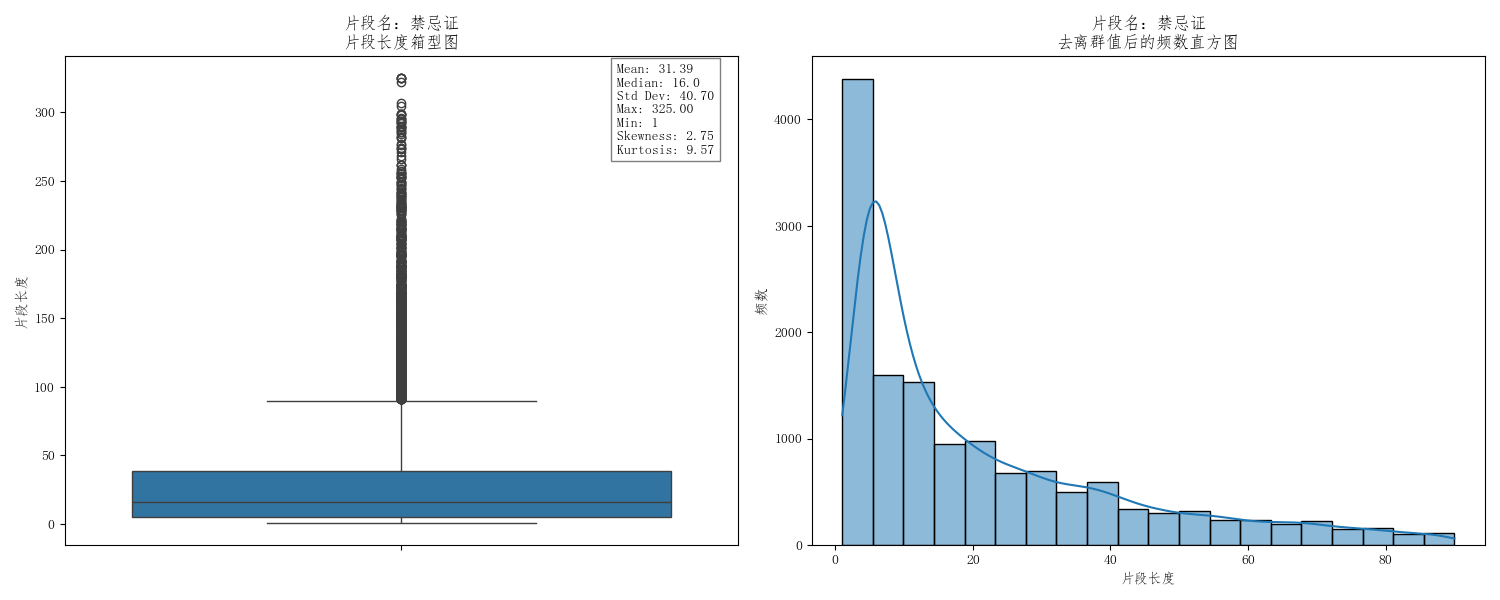
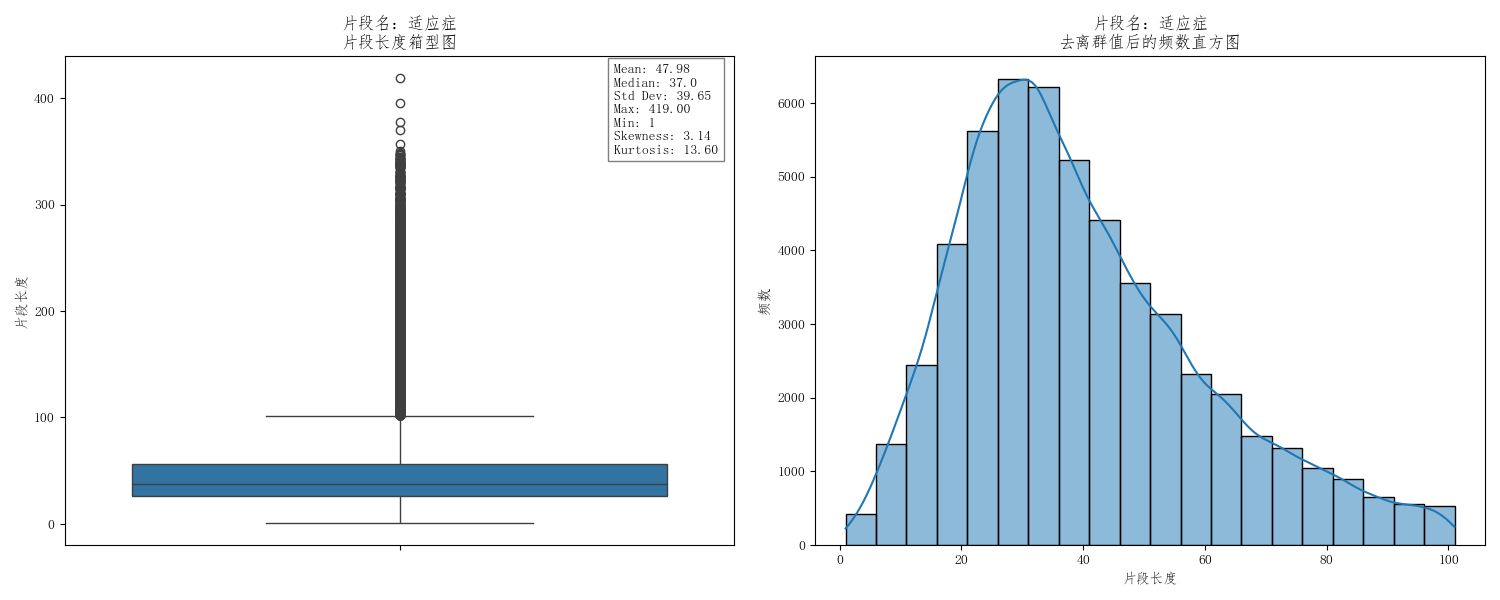
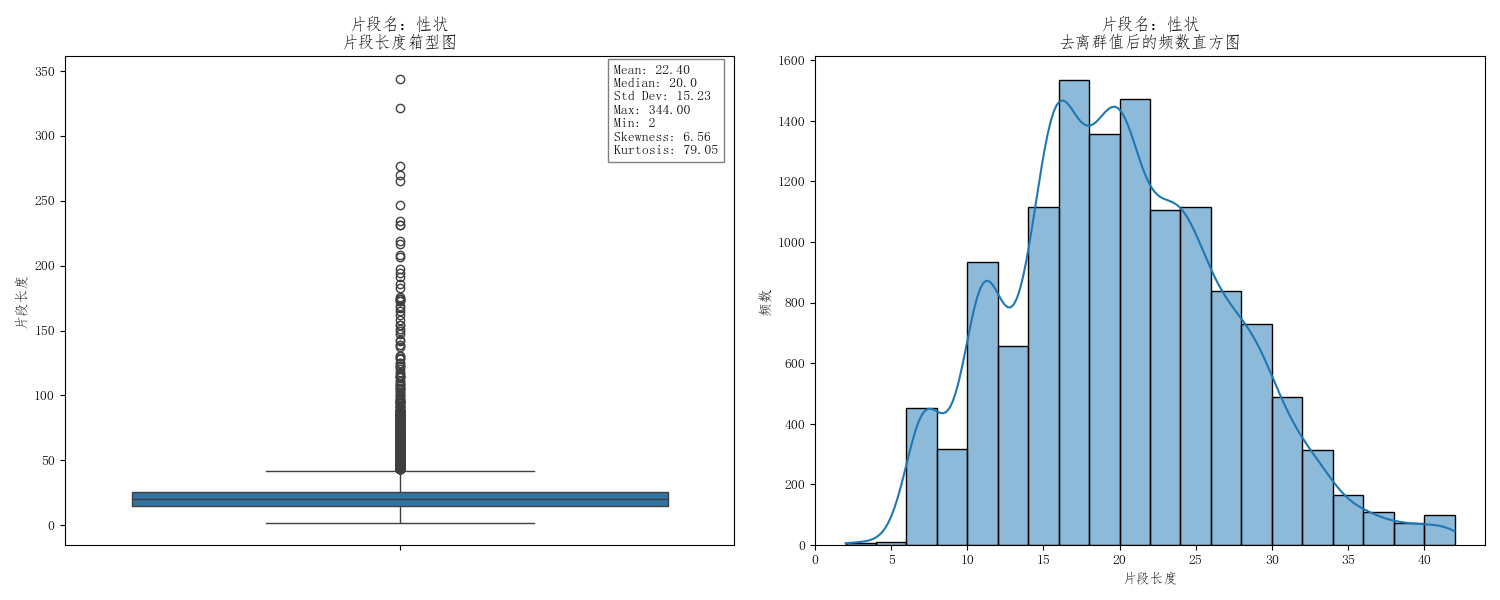
1. 不切分字段的原始片段

如上所述，20个字段中有7个字段为不进行切分的字段，即英文名称、药理分类、标准名称、商品名、通用名、规格、成分，这些字段具有固定格式，且大多为短文本，故不进行切分。这里展示着7个字段的原始片段长度，可以看出，除个别异常值，绝大多数片段均小于200字。个别异常值已整理到Excel.



1. 经切分的子片段

经切分，13个经切分的子片段（性状、适应症、禁忌证、用法用量、药物过量、不良反应、老年用药、儿童用药、孕妇及哺乳期妇女用药、药代动力学、药物相互作用、药理作用、注意事项）绝大部分长度均在350字以内，切分合理。只有适应症、不良反应、药物作用的最大长度超过350字，存在10个长度在350至419字之间的异常值，已单独整理到Excel.



## 遇到的问题及解决方法

1. 多进程不能在交互式的jupyter里运行。
2. Windows使用多进程要放到if \_\_name\_\_ == "\_\_main\_\_":下面。
3. 如果子进程的输出是赋值，而不直接写入文件，需要在进程循环结束后，先写入文件，再使用pool.close()和pool.join()关闭进程池。
4. 计算相似度矩阵并取对角线元素后，得到的是两两句子间的相似度列表score，长度比句子数少1，如4个句子得到3个相似度。因此应将第一个句子直接添加到当前段落，后续根据score中的第i个元素判断第i+1个句子是否合并到当前段落，即第i个和第i+1个句子间是否需要切分。
5. “药理分类”字段中存在占位符号，如“中药>>中成药>>开窍剂>>凉开”，对实际内容的长度有影响。