Stata 入门：从第一条命令到可复查的统计表

怎样把20份合同原表变成能核对单位、数量和缺失的统计结果?

学习任务 独立交出分组统计、门槛对照和新数据迁移三份作业，解释每项结果的口径。

资料与样本 跟做与改做沿用此前留存的20份中国政府采购网合同公告整理记录，本轮没有重新采集；数据非总体或随机样本。独立作业另用6行自建模拟数据，不对应真实机构。

这堂课要交什么，手里的数据是什么

这堂课的目标是独立完成一张能回查原表的描述统计表。输入《采购合同练习.csv》每行是一份合同，包含合同编号、名称、采购人、采购方式、公告金额万元和详情链接。先用文本编辑器看首行，再浏览前几条：两个名称都叫印刷服务合同的记录有不同编号和金额，不能只看名称就删掉一条。本课明确以合同编号核查这20条课堂样本的唯一性；换数据后必须重新检查。

资料沿用此前留存的20份中国政府采购网合同公告整理结果，没有在本轮重新采集，也没有证明这20条构成随机样本或完整期间的总体。分采购方式的数量与金额仅描述这份课堂表，不据此评价全国采购结构。详情链接供回查来源；网页今天能否访问，不影响已经附包的离线计算。本课不演示绕过验证码、封禁或其他访问限制。

最终交三样东西：一张分采购方式的数量和金额表、一张比较两种金额门槛的核对表、一份在新数据上的明细与分组结果。第一项跟着完整脚本做；第二项改一个规则理解后果；第三项先独立写，再对照答案。网页能自动核对若干数值，但是否保留来源、是否正确说明缺失、是否理解样本边界，仍需按人工标准审阅。

第一次打开Stata：目录、窗口和运行顺序

先把整个实操包解压到自己能写入的文件夹，例如D:/课程/Stata入门；不要在压缩包预览窗口直接双击脚本。启动自己的授权Stata。常用的三块区域是输入一条命令的Command窗口、显示执行信息的Results窗口和显示变量名称的Variables窗口。首次练习前保存自己原有数据；课堂脚本的clear会清除当前内存数据，不会删除磁盘上的CSV。

在Command窗口输入cd "D:/课程/Stata入门"，将示例替换成自己的实际位置，使用英文双引号。再输入pwd查看当前目录，输入dir确认能看到采购合同练习.csv和课堂练习.do。路径中有空格或中文时仍保留引号。目录定位成功后输入doedit "课堂练习.do"打开代码编辑器。先通读分段注释，再输入do "课堂练习.do"运行整份脚本，结果应输出分组统计及几项核验。

也可用编辑器的执行按钮运行代码，具体按钮位置随版本略有变化。第一次先运行整份文件，避免只选中最后几行导致依赖变量不存在。后来逐段练习时，从导入开始按顺序运行。若中途出现红色错误，不要继续跳到下一段；先记下命令和错误内容，再按本课排查表处理。随包的运行课程入口执行的是Python独立复核，不能据它成功就认定Stata客户端已验证。

读入表格后先看类型，不急着做计算

完整导入命令中，using后面是输入文件，varnames(1)说明首行是字段名，encoding("utf-8")说明文本编码。合同编号、合同名称、采购人、采购方式和链接应按字符串读入；代码用stringcols(1/4 6)固定这些列，避免纯数字形式的合同编号被当成数值。公告金额万元是数值列，asdouble用于保留更充分的数值精度。这些列号来自当前CSV字段顺序，若以后换表，必须先核对顺序。

依次执行describe、count和list 合同编号 合同名称 公告金额万元 in 1/3, noobs。describe告诉你变量类型；count显示20条观察；list让你看到真实内容。string类是文字，double等是数值；合同编号即使全由数字组成也不用于加总。数据浏览窗口可用browse打开，只读方式更适合初学检查。不要为了把表格变得好看直接编辑原始输入。

接着运行isid 合同编号。通过时通常没有醒目的结果表，这是成功检查的一种正常表现；不通过时先找重复或空编号，不要立即删除记录。count只能检查数量，不能证明行没有错配。因此本课还要求核对前几条记录、唯一键和一笔金额。把这些不同检查组合起来，比只看末行出现运行成功更有意义。

用条件定位记录：先分清等号、文字与位置

list 合同编号 合同名称 if 采购人 == "安福县红十字会", noobs按字段内容筛选；这里==表示比较是否相等，双引号表示文字。与之不同，in 1/3按当前排序后的行位置取前三条。排序后前三条可能变化，而按同一采购人筛选的定义保持不变。需要审计某一笔合同时，应保留稳定编号，而不是把“第2行”当成长期标识。

tabulate 采购方式, missing显示类别频数，并显式保留缺失类别供检查。课堂数据有六种采购方式，框架协议有6份。频数告诉你记录多少，不代表金额多少；少量大额合同可能占多数金额。学员应分别写清“合同数占比”和“金额占比”，不要把两个分母混用。本课没有将采购方式转成数值编码，因为当前分组命令可以直接使用字符串类别。

可以自己先写一条筛选命令，查看金额万元等于100的记录，再与完整脚本对应。这里的100万元与后面金额元中的1000000是同一个门槛。若过滤条件中误用一个等号、遗漏文字引号，或把英文字母O当作数字0，应回到命令本身修正。不要通过修改原数据来让错误命令恰好得到想要的结果。

单位换算与手算检查：万元乘一万

先看字段名“公告金额万元”，据此建立新变量金额元。generate double 金额元 = round(公告金额万元 * 10000, 0.01)的含义是保留原列，乘一万得到元，再四舍五入到分。本课20条记录的金额精度与这样的处理相容；如果原资料是不同币种或含税口径，则还需另建转换规则，不能只改字段名。对新规则的解释应写在代码注释与结果说明中。

取合同编号KC-20261103103104-105756做手算：0.999247万元乘10000等于9992.47元。再核对全表合计12405922.20元。一个具体记录检查定位错误，一项总额检查发现整体变化，作用不同。若总额正好相差10000倍，先查单位；若仅差几分钱，再检查舍入和输入精度。不要因为看见很多小数就随意把所有金额先截成整数。

format 金额元 %15.2fc只决定屏幕上保留两位小数、怎样排版，不能代替round改变计算精度。summarize 金额元, detail可看分布；summarize后的Obs是该变量非缺失记录数，不一定等于全表count。本课金额均已观察；后面迁移数据刻意保留一项缺失。quietly summarize..., meanonly之后r(sum)存放本次合计，随后别的命令可能覆盖返回值，因此用后立即检查。

从20行合同到6行分组表，再回到明细

分组统计会改变观察单位：20行合同明细变成每种采购方式一行。先运行preserve保留当前内存状态，再generate byte 一份 = 1给每份合同一个计数标记。collapse (sum) 合同数=一份 合同金额元=金额元, by(采购方式)按采购方式分别求和，并为新字段写出中文名称。这里的合同数是对每行的1求和，不能拿金额非缺失数在所有情形下代替合同总数。

gsort -合同金额元把组按金额从大到小排序，list显示结果，再export delimited写出CSV。导出的表头必须说明金额单位为元。完成后运行restore回到20行合同明细，并再次assert _N == 20；如果你忘了恢复，下一段门槛筛选就会错误地对6个组而非20份合同计算。分组前后数量与金额应分别对得上，这是本课的验收约束。

不要把collapse后的表保存到原始合同CSV名称。输出可以使用replace覆盖本课明确的练习结果，原始输入保持不动。restore恢复的是之前保留的内存状态，不会撤销你已经导出的磁盘文件。若一组的金额全缺失，有些求和结果可能呈现为0，仍须结合有效数与缺失数解释；新数据作业要求同时报告这些数量，避免把未知金额说成已知的零。

改做：至少100万元与超过100万元

先用中文写规则：“至少100万元”包含恰好100万元，“超过100万元”不包含。换算到金额元后，分别使用>=1000000和>1000000。本课脚本生成两个0/1变量，便于逐条核对；!missing(金额元)显式排除未观察金额。Stata的数值缺失值在排序比较中高于普通数值，不能省略缺失条件后把未知金额选成大额合同。

对这20份真实课堂记录，至少100万元有4份，合计9717600.63元；超过100万元有3份，合计8717600.63元。差异来自一份恰好100万元的合同，金额差为1000000元。请用list定位该合同编号，解释为什么规则改变会产生这个结果。只修改展示文字而不修改筛选条件，会造成表格口径和实际计算不一致。

这项门槛是课堂筛选练习，不代表政策资格、法律界限或项目质量标准。正式研究中的阈值应来自明确制度或事先研究设计。若试过多个阈值，应说明为什么采用某一个，并保留重要敏感性结果；不能只挑最符合预期的数字。作业要求你提交两条规则的结果和差异解释，而不是替任何一条规则寻找漂亮结论。

独立做：新数据中的前导零、缺失和零

独立作业_模拟合同.csv是另外自建的六行教学数据，采购类别甲乙不对应真实机构。编号001至006需要按字符串保留。原金额有的以元计、有的以万元计，先根据原单位分别赋值，再统一到元。先检查单位只属于已定义的两种；遇到第三种单位应停下核对，不应默认按元处理。按两列共同确定数值含义，是从原课堂迁移到新资料的关键。

编号004的原金额为空，表示尚未观察，保留金额元缺失；编号005明确为0，是一个已观察零值，要计入有效金额数量。六条记录中有效金额5条、缺失1条，已知金额合计22300元。甲类三条合计18800元；乙类有三条记录，其中仅两条金额已知，合计3500元。因此22300元应写为“已知金额合计”，不能说已经掌握六份合同的完整总额。

先在作业起步.do中完成自己的代码，再打开参考答案.do核对。至少交出明细表与分组表：明细保留编号、原金额、原单位、新金额和缺失标记；分组同时列总记录数、有效数、缺失数及已知金额合计。把CSV用文本方式查看一次，确认001仍有三个字符；有些表格软件自动打开时会把它显示为1，需要按文本方式导入后再判断文件是否丢零。

遇到错误怎样缩小问题范围

报file not found时，先pwd和dir，再检查是否完整解压、文件名是否被加成.csv.csv；不要先改几十处路径。报type mismatch时，先describe目标变量，再看是不是把数字读成了字符串，或在文字比较中漏了引号。检查原始字符后再决定转换，destring, force会把无法转换的字符变成缺失，不适合作为不明原因报错的通用修复。

报变量已经定义时，通常是只重跑了generate段而没重新导入。初学阶段最容易核查的办法是从脚本开头重新运行完整流程。报某个变量找不到时，先判断目前是合同明细还是collapse后的分组表，再检查拼写。assert失败表示课堂约束不成立，可能是输入变更、单位变化或上一步状态错误；应保留错误并定位原因，不能删掉assert就宣告问题解决。

本课运行日志保存文本信息，便于比较首次报错与修正后的输出。只在特定可预期清理动作上使用capture，例如关闭可能未开启的日志；不要给整份课堂代码加capture来隐藏失败。提交问题时附出错命令、可复现的输入范围与错误文本，不要上传与练习无关的私人文件。查命令说明可在Stata使用help，随课也列出已核过的官方文档。

从课堂数值走向可复查的交付

自己的成果应包括输入来源说明、do文件、生成的CSV与简短解释。让另一位学习者在新文件夹完整解压后，从导入开始运行，检查观察数、唯一键、单位和分组合计；不要要求对方先手动在数据编辑器里改几格，才能复现你的数字。运行日志能帮助查错，但日志成功打印不等于研究解释正确，结果口径仍须人工阅读。

本地制作环境已用固定Python对相同输入和三层练习独立复算。当前没有在授权Stata客户端实际执行这些do文件，也没有完成macOS实机验证，因此课内保留该说明。脚本采用Stata 16语法约束，但version命令不会安装软件或验证许可证。你在自己的客户端运行后，可逐项记录实际版本、输入文件、结果与错误；不要把他人的运行记录复制为自己的验证。

继续学习时，可从Stata基础组进入数据导入与保存、merge和append、reshape以及面板课程；准备政策评估时再学习从零做DID。进入下一课前，至少能解释本课为什么有20行明细和6行分组、为什么两种门槛相差一条、为什么一个缺失金额不等于零。软件操作为研究判断服务，这三个解释也是判断自己是否真正完成本课的标准。

分层作业：跟做：导入并交出第一张汇总表

使用采购合同练习.csv，完成导入、唯一键检查、万元换元和分采购方式汇总。先核对一份合同，再核对总额。

提交内容：跟做结果表、完整do文件及原样本范围说明。

提示：从pwd、dir和import delimited开始；合同编号应读作字符串。

提示：每行生成一份=1，preserve后collapse求和，导出后restore。

自查：保留原合同编号与来源文件，未按合同名称误删记录。

自查：表头写清数量与元金额；分组数量和金额能分别回到20条原表。

自查：说明这是有限课堂样本，不概括为全国结构。

参考解释：20份，公告金额合计12405922.20元；六种采购方式，其中框架协议6份。示例首份0.999247万元=9992.47元。

数值核对：合同记录数 = 20 份；检查是否只导入部分记录，或按合同名称误去重。

数值核对：公告金额合计 = 12405922 元；检查万元乘10000及保留到分的口径。

数值核对：框架协议合同数 = 6 份；在原表按采购方式计数，不按分组表行数作答。

分层作业：改做：改变门槛是否包含边界

仍使用20份真实课堂合同，分别筛选至少100万元和超过100万元，报告合同数量、合计及两规则的金额差。

提交内容：两规则对照表、筛选代码及差异来自哪条合同的说明。

提示：先把门槛换成1000000元，检查>=与>的区别。

提示：在原表定位金额元==1000000；条件中显式排除缺失。

自查：规则文字与实际比较运算符一致。

自查：解释差异来自恰好100万元的一份合同，而非数据被删除。

自查：不把课堂门槛解释为法定采购标准。

参考解释：至少100万元4份，合计9717600.63元；超过100万元3份，合计8717600.63元；两规则金额差1000000元。

数值核对：至少100万元合同数 = 4 份；“至少”包含恰好100万元的边界记录。

数值核对：超过100万元合同数 = 3 份；“超过”使用严格大于，排除边界。

数值核对：两规则金额差 = 1000000 元；逐笔定位边界合同，并核对金额单位。

分层作业：独立做：混合单位与一笔未知金额

使用独立作业_模拟合同.csv六行自建资料。保留编号文本，按原单位统一到元，区分空金额与零，导出明细和分组结果。先完成作业起步.do，再看参考答案.do。

提交内容：明细、分组表、自写do文件及缺失金额的文字说明。

提示：第一列用stringcols(1)保留001；金额元先设缺失再按单位replace。

提示：同时报告总记录数、有效数与缺失数，已知金额合计不能冒充完整总额。

自查：001至006保留前导零，原金额与原单位仍可回查。

自查：004为空保留缺失，005为0保留为有效观察。

自查：甲类已知合计18800元，乙类3500元且注明一条缺失；数据明确标为模拟。

参考解释：有效金额5条，缺失1条，已知金额合计22300元；甲类3条/3条有效/18800元，乙类3条/2条有效/3500元。

数值核对：有效金额记录数 = 5 条；真实零值005应计入有效观察，004空金额不计入。

数值核对：缺失金额记录数 = 1 条；只对金额是否观察计数，不能把零也当缺失。

数值核对：已知金额合计 = 22300 元；先按元或万元换算；空金额不补零，也不把已知合计称为全部总额。

常见问题与处理

file not found，找不到CSV或do文件。原因：当前目录错误、未完整解压或文件名不一致。操作：先pwd，再dir；将cd后的路径改为实际解压目录，保留英文引号。复查：dir能看到采购合同练习.csv；重新导入后count为20。

type mismatch，金额不能乘10000。原因：金额列包含文字或被读成字符串。操作：describe查看类型；list查看原字符，回查异常值；不要直接destring, force。复查：原金额列类型为数值，单笔0.999247万元变成9992.47元。

变量金额元已经定义。原因：在已有生成变量的内存状态重复执行generate。操作：保存自己的代码后从脚本开头重新导入并执行；不要为了跳过报错随意覆盖变量。复查：完整脚本执行一次后生成目标变量和结果，编号唯一、记录仍为20。

采购人变量找不到，内存只剩采购方式和合计。原因：执行了collapse后没有restore，观察单位已变为采购方式。操作：从preserve之前开始完整重跑该段，确保collapse后执行restore。复查：restore后count为20，describe能再次看到采购人。

isid或assert失败。原因：重复编号、缺失键、输入改变或规则计算不符。操作：定位触发的检查与相关记录，核对输入版本、单位和前一步状态；保留错误而不删检查。复查：说明并修复具体问题后，对同一约定再次运行检查通过。

缺失金额被选为大额合同。原因：只写金额元>=门槛，没有排除Stata数值缺失。操作：条件加入!missing(金额元)，另行统计缺失数。复查：迁移表有效金额5条、缺失1条，零金额005仍有效。

CSV用表格软件打开时001显示成1。原因：打开程序自动推断编号为数字，不一定是CSV原文丢零。操作：先用文本编辑器检查文件；在表格软件按文本导入编号，在Stata使用stringcols(1)。复查：编号001原文及Stata字符串长度均为3。

计算结果与口径

真实课堂合同数 20

公告金额合计元 12405922.20

模拟迁移缺失金额数 1

采购方式	合同数	合同金额元
公开招标	4	8707600.63
竞争性磋商	4	2907100.00
询价	1	429609.00
框架协议	6	197272.57
电子卖场	4	132640.00
单一来源	1	31700.00

前三列是既有20份合同留存资料的描述汇总，不代表总体采购结构。迁移数据另表标为模拟；22300元是5条已知金额合计，并非6条完整总额。数值由Python复核，Stata客户端未实测。

练习与参考答案

1. 使用采购合同练习.csv，完成导入、唯一键检查、万元换元和分采购方式汇总。先核对一份合同，再核对总额。

参考答案 20份，公告金额合计12405922.20元；六种采购方式，其中框架协议6份。示例首份0.999247万元=9992.47元。

2. 仍使用20份真实课堂合同，分别筛选至少100万元和超过100万元，报告合同数量、合计及两规则的金额差。

参考答案 至少100万元4份，合计9717600.63元；超过100万元3份，合计8717600.63元；两规则金额差1000000元。

3. 使用独立作业_模拟合同.csv六行自建资料。保留编号文本，按原单位统一到元，区分空金额与零，导出明细和分组结果。先完成作业起步.do，再看参考答案.do。

参考答案 有效金额5条，缺失1条，已知金额合计22300元；甲类3条/3条有效/18800元，乙类3条/2条有效/3500元。

阅读材料与运行说明

中国政府采购网合同公告：课堂表保留逐条详情链接

Stata官方：Do-file Editor操作

Stata官方：文本表格导入与导出

Stata官方：唯一标识检查isid

Stata官方：generate与replace

Stata官方：summarize统计及返回值

Stata官方：collapse分组汇总

Stata官方：preserve与restore

Stata官方：表达式、比较与缺失值

完整注释代码、输入资料和成果文件见随课实操包。Stata命令采用version 16.0语法约束。运行课程入口执行Python复核，不会启动Stata。