Stata 命令 dtable | 在 Stata 18 中创建描述性统计数据表

在 Stata 17 中，我们介绍了用于创建和自定义表格的新 **collect** 命令集，以及用于创建和导出估计结果表的 **etable** 命令。Stata 18 提供了另一个新命令 **dtable**，它可以轻松地构建和导出描述性统计数据表，在出版物中通常称为 Table 1。现在，为分类变量和连续变量生成描述性统计表比以往任何时候都容易。值得一提的是，**etable** 和 **dtable** 这两个命令都是基于我们在 Stata 17 中引入的 **collect** 框架构建的，因此它们共享许多属性。

在本文中，将演示如何创建和导出描述性统计数据的简单表格和更复杂的数据表，这些数据表按组显示统计数据，测试组间的差异等等。本文还将展示如何使用 **collect** 命令套件来进一步自定义表的外观，以及如何在完整的报告中包括使用 **dtable** 创建的表。

举个简单的实例

在 Stata 18 之前，如果我们想生成一个描述性数据统计表，可以使用 **summary** 来获得连续变量的汇总统计数据，并使用 **tabulate** 来报告分类变量的频率、比例或百分比。我们以 **auto.dta**（1978 年的汽车数据）为例：

```
. sysuse auto, clear
(1978 automobile data)

. summarize price weight mpg

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
      weight |         74    3019.459    777.1936       1760       4840
         mpg |         74     21.2973    5.785503         12         41

. tabulate rep78

     Repair |
record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00
```

这些命令为我们计算了统计数据。然而，手动将所有这些数字输入到一个格式规范的表中是一项繁琐的工作，而且当我们有新数据时，它是不可复制的。

相比之下，使用 **dtable**，我们可以输入

```
. dtable price weight mpg i.rep78

----------------------------------------
                        Summary
----------------------------------------
N                                   74
Price                6,165.257 (2,949.496)
Weight (lbs.)          3,019.459 (777.194)
Mileage (mpg)              21.297 (5.786)
Repair record 1978
  1                             2 (2.9%)
  2                            8 (11.6%)
  3                           30 (43.5%)
  4                           18 (26.1%)
  5                           11 (15.9%)
----------------------------------------
```

就像这样简单的，我们已经建立了一个表，显示了指定连续变量（**price**, **weight**, 和 **mpg**）的数据样本量、平均值和标准差，以及指定分类变量水平的频率和百分比（**rep78**）。

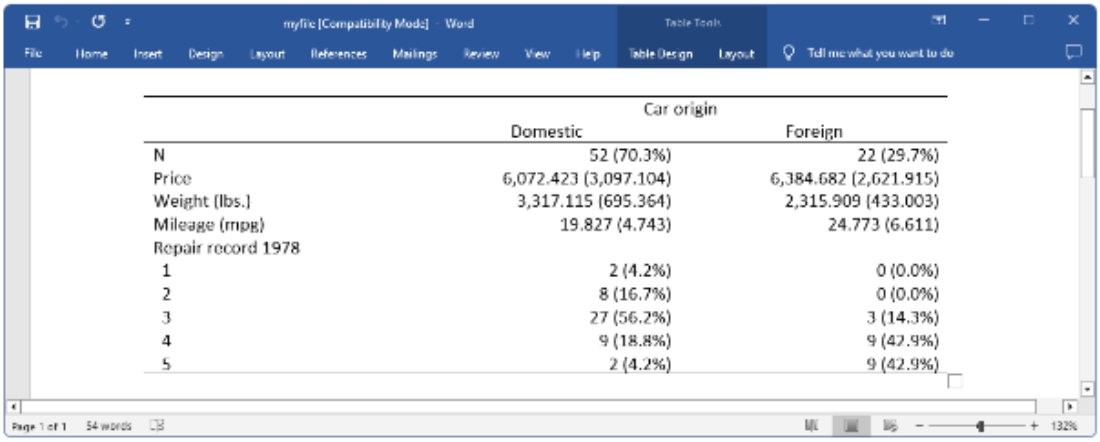
除了完整样本的结果外，我们还可以通过添加 **by ()** 选项，分别请求组变量（比如 **foreign**）的每个类别的上述统计信息：

```
. dtable price weight mpg i.rep78, by(foreign)

--------------------------------------------------------------------------------
                                      Car origin
                    Domestic             Foreign               Total
--------------------------------------------------------------------------------
N                         52 (70.3%)           22 (29.7%)          74 (100.0%)
Price          6,072.423 (3,097.104) 6,384.682 (2,621.915) 6,165.257 (2,949.496)
Weight (lbs.)    3,317.115 (695.364)  2,315.909 (433.003)   3,019.459 (777.194)
Mileage (mpg)          19.827 (4.743)       24.773 (6.611)       21.297 (5.786)
Repair record 1978
  1                        2 (4.2%)             0 (0.0%)             2 (2.9%)
  2                        8 (16.7%)            0 (0.0%)             8 (11.6%)
  3                       27 (56.2%)            3 (14.3%)           30 (43.5%)
  4                        9 (18.8%)            9 (42.9%)           18 (26.1%)
  5                        2 (4.2%)             9 (42.9%)           11 (15.9%)
--------------------------------------------------------------------------------
```

我们可以使用 **by（）**中的子选项 **nototal** 来抑制总样本的列。我们可以使用 **export（）**选项将该表导出到 Word 文档 **myfile.docx** 中：

```
. dtable price weight mpg i.rep78, by(foreign, nototal)
> export(myfile.docx, replace)
(output omitted)
```

导出的表如下

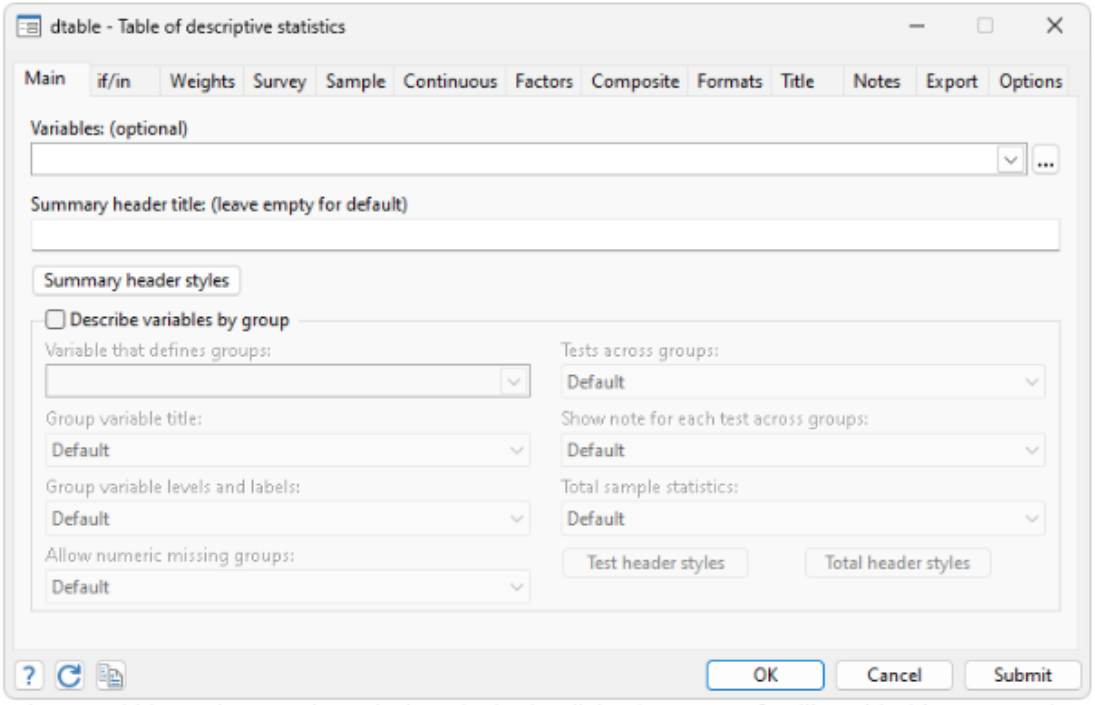


	Car origin	
	Domestic	Foreign
N	52 (70.3%)	22 (29.7%)
Price	6,072.423 (3,097.104)	6,384.682 (2,621.915)
Weight (lbs.)	3,317.115 (695.364)	2,315.909 (433.003)
Mileage (mpg)	19.827 (4.743)	24.773 (6.611)
Repair record 1978		
1	2 (4.2%)	0 (0.0%)
2	8 (16.7%)	0 (0.0%)
3	27 (56.2%)	3 (14.3%)
4	9 (18.8%)	9 (42.9%)
5	2 (4.2%)	9 (42.9%)

请求自定义统计数据和检验

默认情况下，**dtable** 报告数据集的样本量、连续变量的均值和标准差，以及分类变量的频率和百分比。但我们可以要求其他描述性统计数据，如中位数和四分位数范围。我们甚至可以为同一个表中的不同变量指定不同的统计信息。在我们进入更高级的示例之前，您先看看下方 **dtable** 的对话框。

菜单 **Statistics > Summaries, tables, and tests > Table of descriptive statistics**，打开 **dtable** 对话框。



浏览对话框中的选项卡以熟悉此命令。这是探索使用 **dtable** 可以做什么的好方法。我想突出显示三个选项卡，其余的留给您自己探索。

•在 **Main** 选项卡上，我们可以指定感兴趣的连续变量和分类变量（使用 i.因子变量表示法表示分类变量）。我们也可以指定 **by** 变量。我们还可以查看其他结果，比如通过 **by** 组显示检验结果，是否要显示样本统计数据等。
•在 **Continuous** 选项卡上，我们可以指定连续变量（它们可以在 **Main** 选项卡上指定，也可以不指定），并且我们可以请求针对不同变量的自定义统计数据和检验。
•**Factors** 选项卡的工作原理与 **Continuous** 选项卡类似。我们可以指定 **factor** 变量，并为不同的变量选择定制的统计和检验。

例如，我们将加载 Zeng, Mao, and Lin（2016）中提供的修改后的 Modified Bangkok IDU Preparatory Study 数据。我们可以尝试为不同的变量指定自定义的统计信息和检验，而不是生成默认的表。在这里，我使用了对话框（主要是上面提到的三个选项卡）来轻松地构建表，相应的语法显示在下面的输出中。

```
. dtable, by(male, tests testnotes nototal) sample(, statistic(frequency proportion))
> continuous(age, statistics( mean min max) test(kwallis))
> continuous(ltime rtime, statistics(mean skewness kurtosis) test(poisson))
> factor(needle, statistics(fvfrequency fvproportion))
> factor(jail inject, statistics(fvfrequency) test(fisher))
note: using test kwallis across levels of male for age.
note: using test poisson across levels of male for ltime and rtime.
note: using test pearson across levels of male for needle.
note: using test fisher across levels of male for jail and inject.
```

	Male					
	No			Yes		Test
N	76	0.068		1,048	0.932	
Age (in years)	28.776	18.000	46.000	31.656	17.000 52.000	0.002
Last time seronegative for HIV-1	22.129	-0.305	2.017	24.323	-0.353 2.251	<0.001
First time seropositive for HIV-1	11.951	0.951	2.285	14.428	0.749 3.024	0.020
Shared needles						
No	43	0.566		679	0.648	0.149
Yes	33	0.434		369	0.352	
Imprisoned at recruitment						
No	21			351		0.315
Yes	55			697		
Injected drugs before recruitment						
No	47			659		0.902
Yes	29			389		

在该表中，我们要求报告以下描述性统计数据：1） **age** 变量的平均值、最小值和最大值；2） 变量 **ltime** 和 **rtime** 的均值、偏度和峰度；3）**needle** 变量的频率和比例；4）变量 **jail** 和 **inject** 的频率。统计数据分别报告每个级别的组变量 **male**。我们还显示了每组的样本量和比例。

您可能会注意到，我们添加了一列自定义检验来比较组之间的变量。只有在指定了 **by** 变量时，才能包含检验。因为我们在 **testnotes** 上指定了 **by（）**子选项，所以我们为不同变量选择的特定检验在注释中（表前）有明确提及。

连续变量的可用检验类型如下：

regress	main effects test from a linear regression (*t* test)
poisson	main effects test from a Poisson regression
lnormal	main effects test from a log-normal regression

kwallis	Kruskal–Wallis rank test

分类变量的可用的检验类型如下：

pearson	Pearson's chi-squared test
fisher	Fisher's exact test
lrchi2	likelihood-ratio chi-squared test
gamma	Goodman and Kruskal's gamma
kendall	Kendall's τ
cramer	Cramér's V
svylr	survey-adjusted likelihood-ratio test
svywald	survey-adjusted Wald test
svyllwald	survey-adjusted log-linear Wald test
none	suppress the test

有了这些选项，**dtable** 可以非常方便地执行跨组比较变量的许多检验，并一步到位将 p 值放入表中。

自定义格式和样式

从上表可以看出，我们可以对其外观进行改进。例如，我们想在列标题中而不是在第一行中显示子组样本大小和比例。我们可能还想增加或减少某些统计数据报告的小数位数。我们可能希望将 **min** 值和 **max** 值的显示格式更改为“**min**-**max**”，并将其放入括号中，我们也可能希望将比例放入括号中。所有这些更改都可以通过 **dtable** 选项完成，而无需额外编码。以下是修改后的 **dtable** 语法和输出。

```
. dtable, by(male, tests testnotes nototal)
> sample(, statistic(frequency proportion)
> place(seplabels) ) continuous(age, statistics(mean minmax) test(kwallis))
> continuous(ltime rtime, statistics(mean skewness kurtosis) test(poisson))
> factor(needle, statistics(fvfrequency fvproportion))
> factor(jail inject, statistics(fvfrequency) test(fisher))
> define(minmax = min max, delimiter(-)) nformat(%9.1f mean minmax)
> sformat("(%s)" fvproportion minmax proportion)
> nformat(%9.2f proportion fvproportion) export(myfile.docx, replace)
note: using test kwallis across levels of male for age.
note: using test poisson across levels of male for ltime and rtime.
note: using test pearson across levels of male for needle.
note: using test fisher across levels of male for jail and inject.

------------------------------------------------------------------------------
                                                Male
                                      No              Yes         Test
                                   76 (0.07)       1,048 (0.93)
------------------------------------------------------------------------------
Age (in years)                  28.8 (18.0-46.0)  31.7 (17.0-52.0)  0.002
Last time seronegative for HIV-1  22.1 -0.305 2.017 24.3 -0.353 2.251 <0.001
First time seropositive for HIV-1  12.0 0.951 2.285  14.4 0.749 3.024  0.020
Shared needles
  No                                  43 (0.57)        679 (0.65)  0.149
  Yes                                 33 (0.43)        369 (0.35)
Imprisoned at recruitment
  No                                        21               351  0.315
  Yes                                       55               697
Injected drugs before recruitment
  No                                        47               659  0.902
  Yes                                       29               389
------------------------------------------------------------------------------
(collection DTable exported to file myfile.docx)
```

在上面的语法中，我使用选项 **define（）**来定义一个新的复合统计数据 **minmax**，使用现有的统计数据 **min** 和 **max**（分隔符“-”用于组合它们）。我们还使用选项 **nformat（）**和 **sfomat（）**分别更改一些统计数据的数字显示格式和字符串显示格式。请注意，“**%s**”是我们正在编辑字符串格式的统计数据的占位符。

如上面例子所示，如果喜欢现在的表，我们可以使用 **export（）**选项将表导出到文档中。下面列出了所有支持的导出表的文件类型:

Suffix	File format	Output format
docx	**as(docx)**	Microsoft Word
html	**as(html)**	HTML 5 with CSS

pdf	**as(pdf)**	PDF
xlsx	**as(xlsx)**	Microsoft Excel 2007/2010 or newer
xls	**as(xls)**	Microsoft Excel 1997/2003
tex	**as(latex)**	LaTeX
smcl	**as(smcl)**	SMCL
txt	**as(txt)**	Plain text
markdown	**as(markdown)**	Markdown
md	**as(markdown)**	Markdown

使用 collect 进一步自定义表格

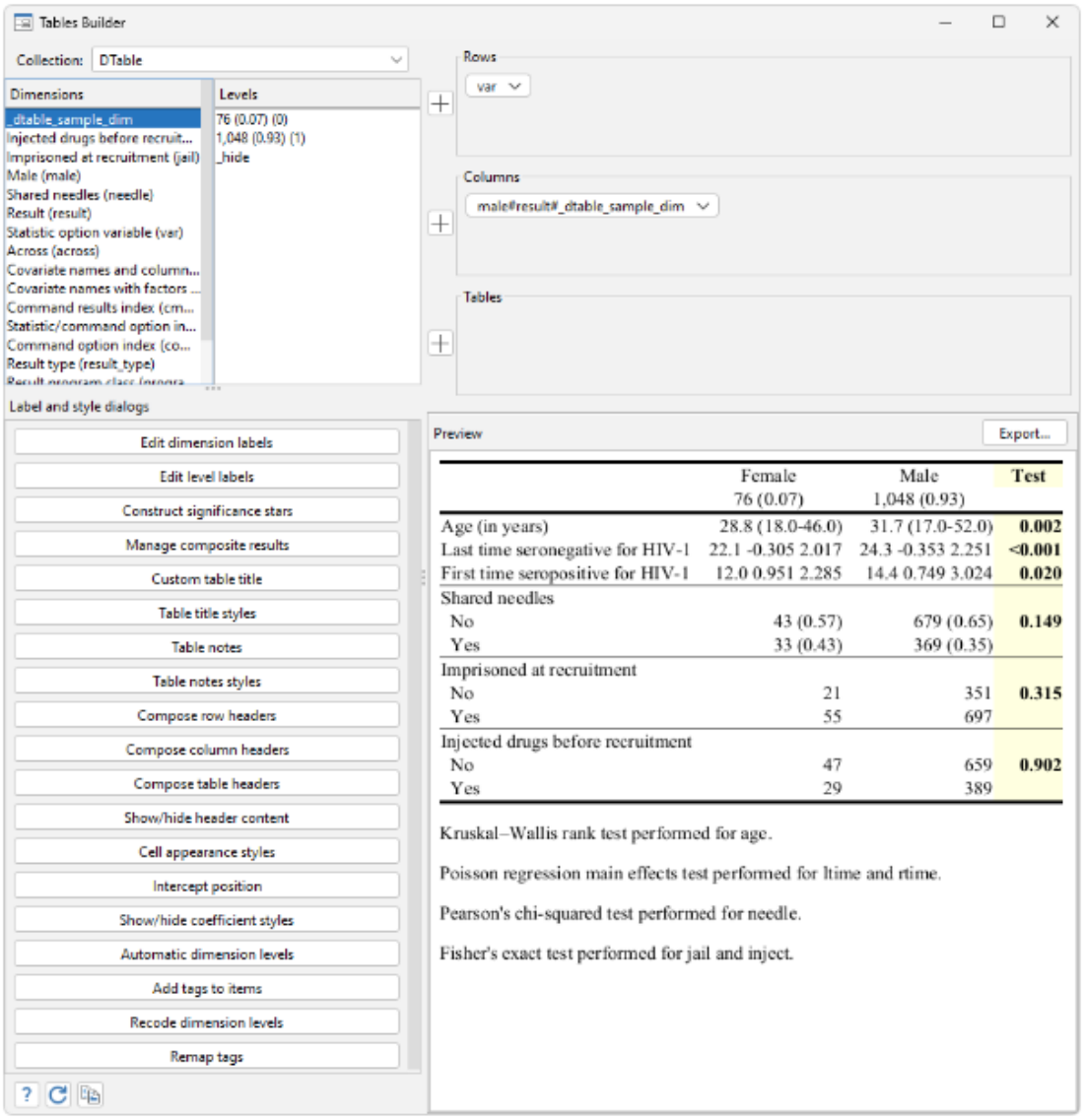
上面的表格看起来不错。我还将演示如何进行一些 **dtable** 无法直接使用的其他更改。由于 **dtable** 是使用 **collect** 实现的，因此我们可以使用 **collect** 命令集来进一步管理使用 **dtable** 创建的表，并以各种方式对其进行编辑。顺便说一句，**collect** 命令一开始需要花点功夫来熟悉所有工具，但我相信您会掌握这些技能，并在稍加练习后喜欢使用这组命令来创建您需要的任何表。

进一步更改表信息，我想 1）隐藏表格标题中的变量名称 **male**，并将组标签 **No** 和 **Yes** 分别更改为 **Female** 和 **male**，2）在连续变量和分类变量之间以及不同分类变量之间添加水平线，3）将检验的 p 值加粗，并用浅黄色阴影突出显示检验列，4）在表中添加自定义注释，显示不同变量的检验类型。让我们使用下面的 **collect** 命令来进行这些更改:

```
. collect style header male, title(hide)

. collect label levels male 0 "Female", modify

. collect label levels male 1 "Male", modify

. collect style cell var[rtime 1.needle 1.jail], border( bottom, width(1))

. collect style cell male[_dtable_test], shading( background(lightyellow)) font(, bold)

. collect notes "Kruskal-Wallis rank test performed for age."

. collect notes "Poisson regression main effects test performed for ltime and rtime."

. collect notes "Pearson's chi-squared test performed for needle."

. collect notes "Fisher's exact test performed for jail and inject."

. collect layout
```

请注意，Stata Results 窗口可以显示其中的一些更改，但它不能显示诸如阴影颜色之类的修改。我们可以打开 Tables 生成器，并在那里确认我们拥有所需的确切表格样式。我们可以从菜单中打开表格生成器，打开菜单 **Statistics > Summaries, tables, and tests > Tables and collections > Build and style table**。

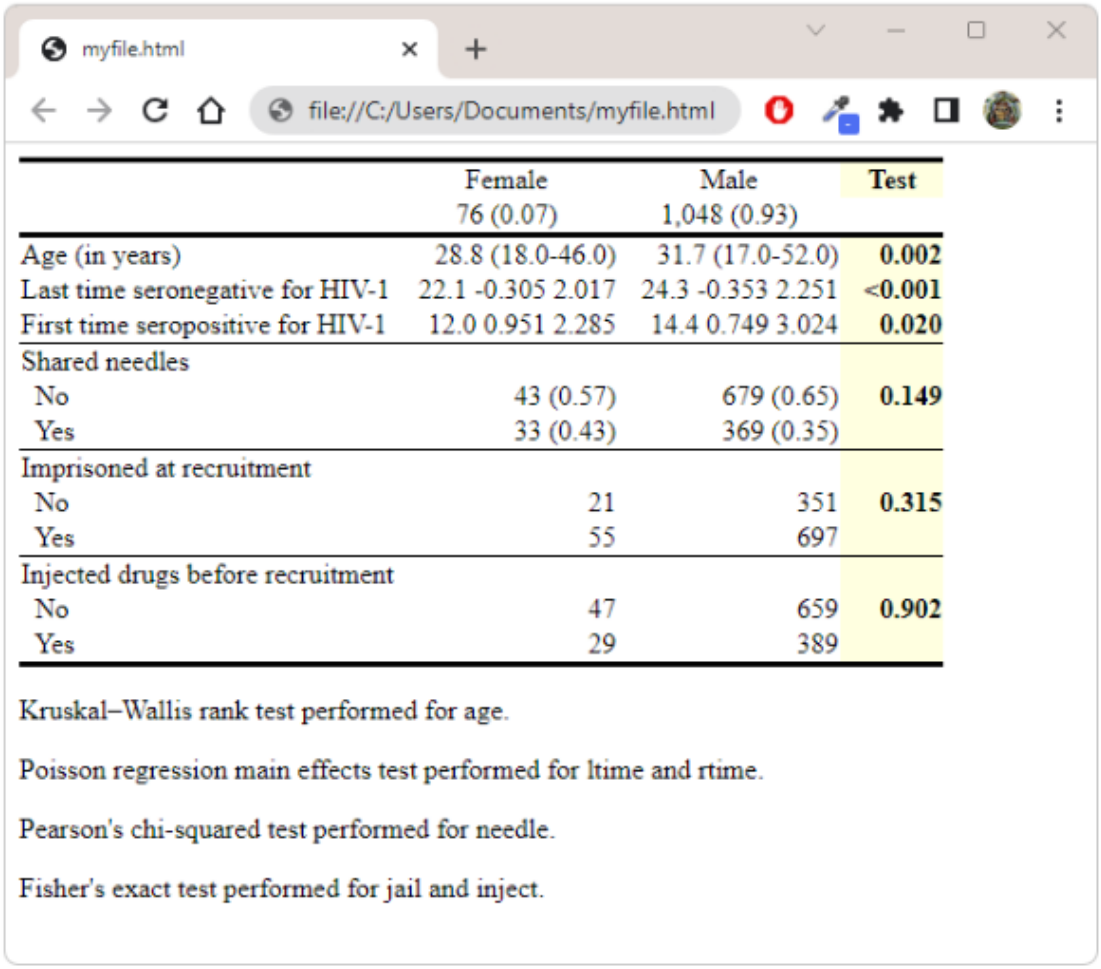
我们可以在 Tables 生成器的预览窗口中看到表格的外观。



当我们将该表导出到其他文档时，导出的表将与这里显示的表相同。我们将该表导出为**.html**文件中。

```
. collect export myfile.html, replace
```

以下是我们的最终文档：



	Female 76 (0.07)	Male 1,048 (0.93)	Test
Age (in years)	28.8 (18.0-46.0)	31.7 (17.0-52.0)	**0.002**
Last time seronegative for HIV-1	22.1 -0.305 2.017	24.3 -0.353 2.251	**<0.001**
First time seropositive for HIV-1	12.0 0.951 2.285	14.4 0.749 3.024	**0.020**
Shared needles			
No	43 (0.57)	679 (0.65)	**0.149**
Yes	33 (0.43)	369 (0.35)	
Imprisoned at recruitment			
No	21	351	**0.315**
Yes	55	697	
Injected drugs before recruitment			
No	47	659	**0.902**
Yes	29	389	

Kruskal–Wallis rank test performed for age.

Poisson regression main effects test performed for ltime and rtime.

Pearson's chi-squared test performed for needle.

Fisher's exact test performed for jail and inject.

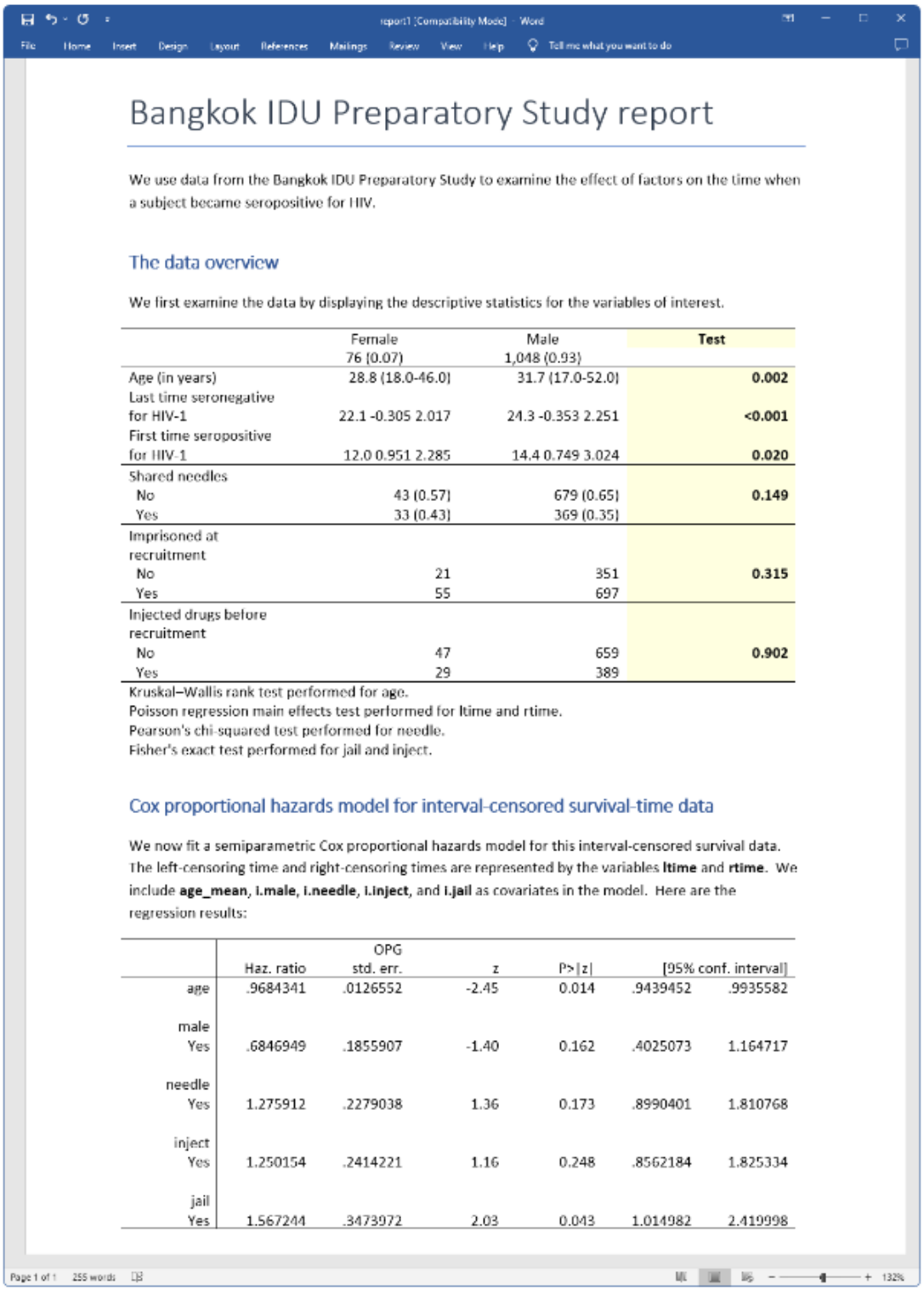
生成包含表格的完整报告

由于 **dtable** 创建描述性统计数据表，而这种类型的表通常作为 Table 1 包含在技术文档中，因此您可能希望将使用 **dtable** 获得的表插入到更大的文档中，而不是单独将表导出为文档。如果是这种情况，如果您分别使用 **putdocx**、**putpdf** 或 **putexcel** 创建文档，则可以使用 **putdocx collect**、**putpdf collect** 或 **putexcel ul_cell=collect** 导出表格。通过这种方式，该表可以与其他内容一起放在文档中的任何位置。以下是使用 **putdocx** 创建包含上述表的文档的示例：

```
webuse idu, clear

putdocx clear

putdocx begin

// Add a title

putdocx paragraph, style(Title)

putdocx text ("Bangkok IDU Preparatory Study report")

putdocx textblock begin

We use data from the Bangkok IDU Preparatory Study to examine
the effect of factors on the time when a subject became
seropositive for HIV.

putdocx textblock end

// Add a heading

putdocx paragraph, style(Heading1)

putdocx text ("The data overview")

putdocx textblock begin

We first examine the data by displaying the descriptive
statistics for the variables of interest.

putdocx textblock end

dtable, by(male, tests testnotes nototal) ///
 sample(, statistic(frequency proportion) ///
 place(seplabels) ) continuous(age, statistics(mean minmax) test(kwallis)) ///
 continuous(ltime rtime, statistics(mean skewness kurtosis) test(poisson)) ///
 factor(needle, statistics(fvfrequency fvproportion)) ///
 factor(jail inject, statistics(fvfrequency) test(fisher)) ///
 define(minmax = min max,  delimiter(-)) nformat(%9.1f  mean minmax) ///
 sformat("(%s)" fvproportion minmax proportion) ///
 nformat(%9.2f proportion fvproportion)
```

```
collect style header male, title(hide)

collect label levels male 0 "Female", modify

collect label levels male 1 "Male", modify

collect style cell var[rtime 1.needle 1.jail], border( bottom, width(1))

collect style cell male[_dtable_test], shading( background(lightyellow)) ///
 font(, bold)

collect notes "Kruskal-Wallis rank test performed for age."

collect notes "Poisson regression main effects test performed for ltime and rtime."

collect notes "Pearson's chi-squared test performed for needle."

collect notes "Fisher's exact test performed for jail and inject."

putdocx collect

putdocx paragraph, style(Heading1)

putdocx text ("Cox proportional hazards model for interval-censored survival-time data")

putdocx textblock begin

We now fit a semiparametric Cox proportional hazards model for this
interval-censored survival data. The left-censoring time and
right-censoring times are represented by the variables
<<dd_docx_display bold: "ltime">> and
<<dd_docx_display bold: "rtime">>.  We include
<<dd_docx_display bold: "age_mean">>, <<dd_docx_display bold: "i.male">>,
<<dd_docx_display bold: "i.needle">>, <<dd_docx_display bold: "i.inject">>,
and <<dd_docx_display bold: "i.jail">> as covariates in the model.
Here are the regression results:

putdocx textblock end

stintcox age i.male i.needle i.inject i.jail, interval(ltime rtime)

putdocx table results = etable

putdocx save report1, replace
```

使用上面的代码，我们创建了文件 **report1.docx**，如下所示



report1 [Compatibility Mode] - Word

File Home Insert Design Layout References Mailings Review View Help Tell me what you want to do

Bangkok IDU Preparatory Study report

We use data from the Bangkok IDU Preparatory Study to examine the effect of factors on the time when a subject became seropositive for HIV.

The data overview

We first examine the data by displaying the descriptive statistics for the variables of interest.

	Female 76 (0.07)	Male 1,048 (0.93)	**Test**
Age (in years)	28.8 (18.0-46.0)	31.7 (17.0-52.0)	**0.002**
Last time seronegative for HIV-1	22.1 -0.305 2.017	24.3 -0.353 2.251	**<0.001**
First time seropositive for HIV-1	12.0 0.951 2.285	14.4 0.749 3.024	**0.020**
Shared needles			
No	43 (0.57)	679 (0.65)	**0.149**
Yes	33 (0.43)	369 (0.35)	
Imprisoned at recruitment			
No	21	351	**0.315**
Yes	55	697	
Injected drugs before recruitment			
No	47	659	**0.902**
Yes	29	389	

Kruskal–Wallis rank test performed for age.
Poisson regression main effects test performed for ltime and rtime.
Pearson's chi-squared test performed for needle.
Fisher's exact test performed for jail and inject.

Cox proportional hazards model for interval-censored survival-time data

We now fit a semiparametric Cox proportional hazards model for this interval-censored survival data. The left-censoring time and right-censoring times are represented by the variables **ltime** and **rtime**. We include **age_mean**, **i.male**, **i.needle**, **i.inject**, and **i.jail** as covariates in the model. Here are the regression results:

	Haz. ratio	OPG std. err.	z	P>|z|	[95% conf.	interval]
age	.9684341	.0126552	-2.45	0.014	.9439452	.9935582
male						
Yes	.6846949	.1855907	-1.40	0.162	.4025073	1.164717
needle						
Yes	1.275912	.2279038	1.36	0.173	.8990401	1.810768
inject						
Yes	1.250154	.2414221	1.16	0.248	.8562184	1.825334
jail						
Yes	1.567244	.3473972	2.03	0.043	1.014982	2.419998

Page 1 of 1 255 words 132%

本报告也可复制。随时重新运行命令并重新创建报告。

结语

在这篇博文中，向您展示了 Stata 18 中使用 **dtable** 可以实现的一些功能和有趣的操作。它有很多功能，我无法在一篇文章中全部展示。您可以申请 Stata 18 试用来体验一下它给您带来的便利。