University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

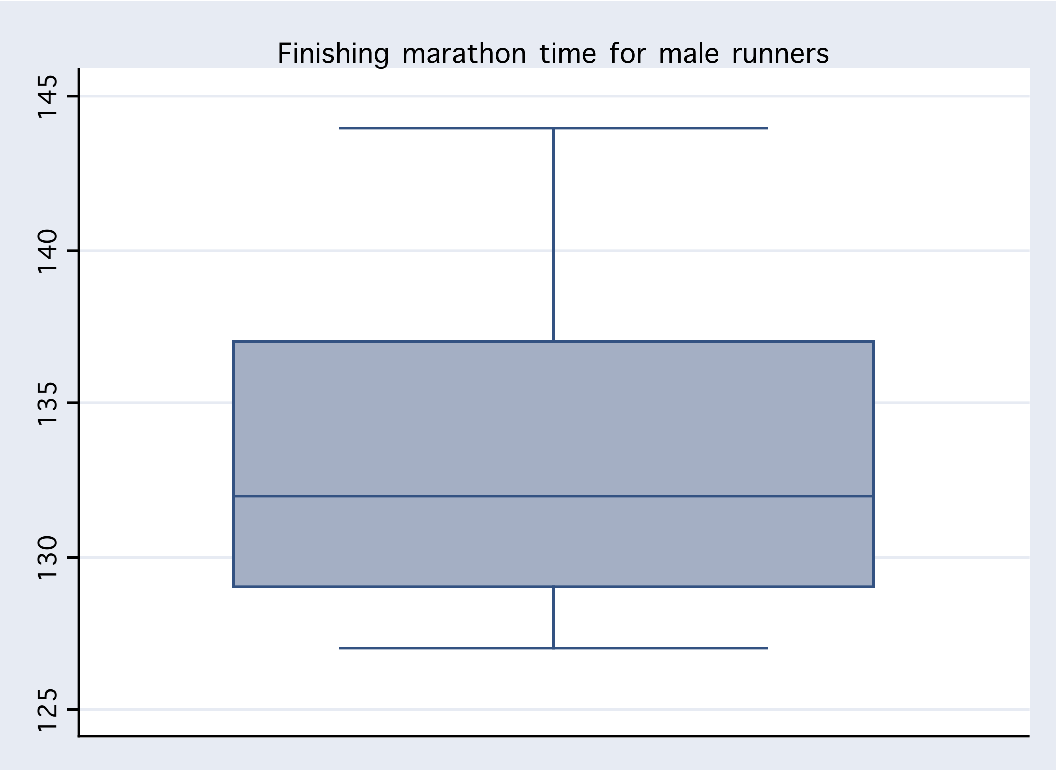
Constructing a boxplot and computing descriptive statistics in Stata and more…

Retrieve the file "textbookex1_38" as follows:

```
. use http://www.stat.ucla.edu/~nchristo/textbookex1_38
```

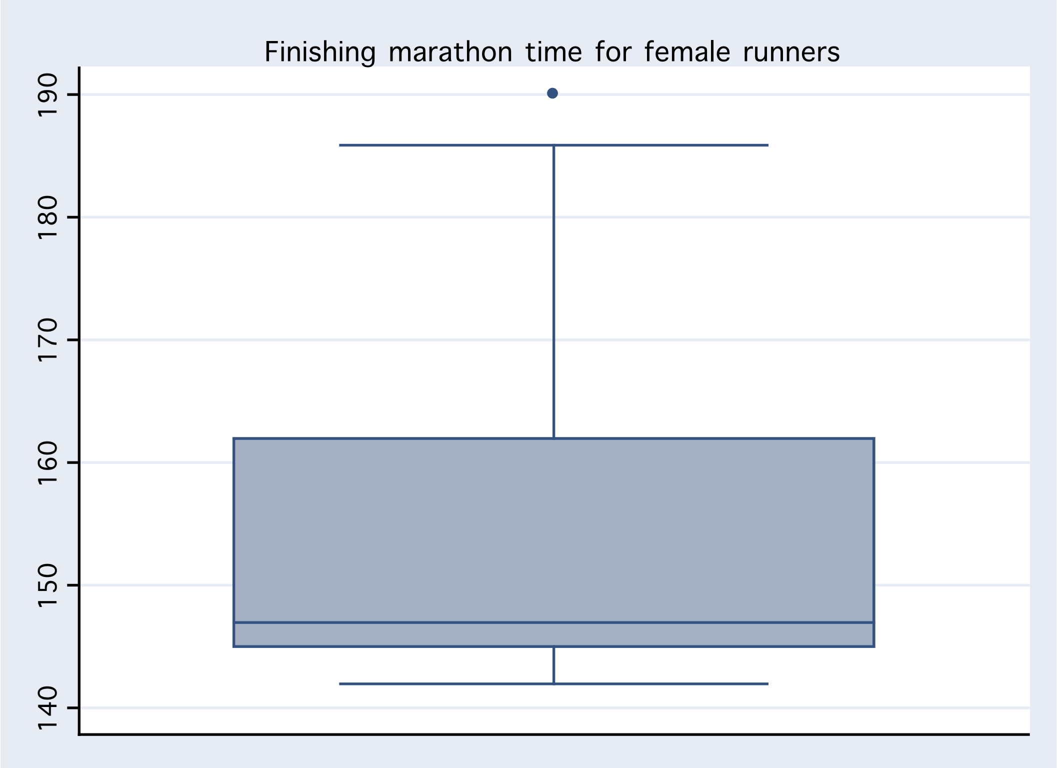
In this file there are four variables.. The first (year1) is the year in which the Boston marathon was completed by men. The second (male_tim) is the completion time rounded to the nearest minute for the years 1959-97. The third (year2) is the year in which women were allowed to compete in the Boston Marathon. The fourth (female_tim) is the completion time also rounded to the nearest minute. You can see the data by typing `. edit` ***or*** `. list`
Let's construct a box plot for the finishing time of men and women.
First the box plot of the finishing time for men:

```
. graph box male_tim, t1title(Finishing  marathon time for male runners)
```



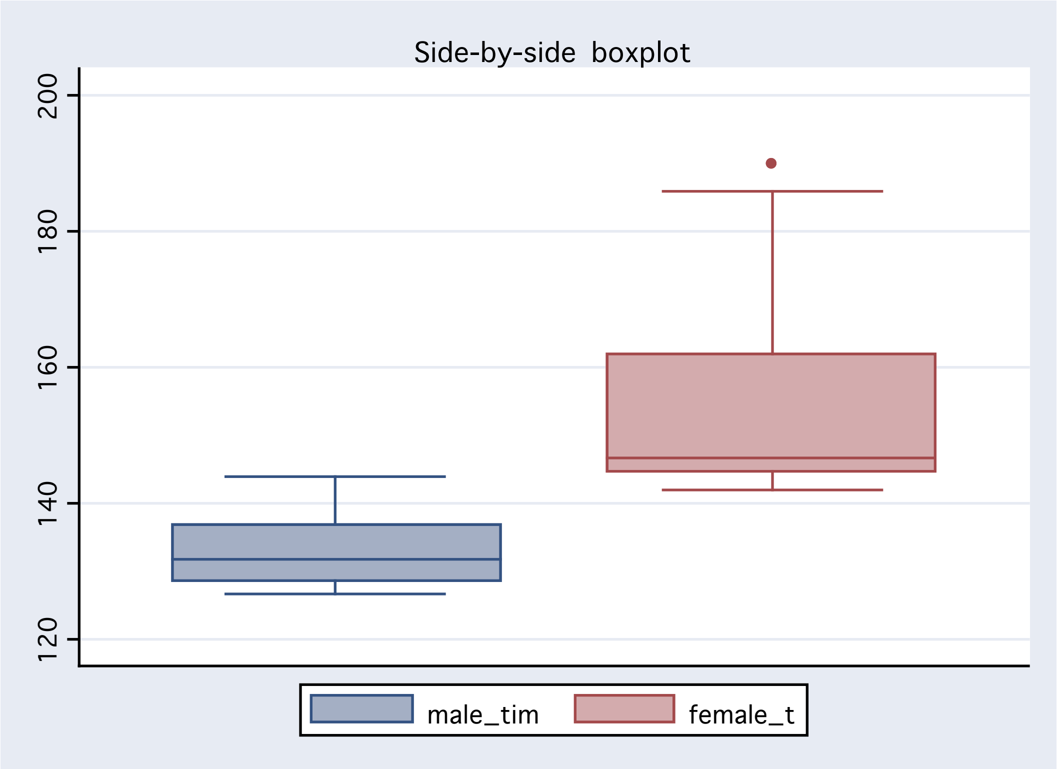
Then the boxplot for the finishing time for women:

```
. graph box female_t, t1title(Finishing marathon time for female runners)
```



We can also do a side-by-side boxplot to compare the 2 variables.

```
. graph box male_tim female_t, t1title(Side-by-side boxplot)
```



Now let’s use Stata to compute descriptive statistics for the completion time of men and women. Here is the command:

```
. summarize male_tim female_t
```

And here is what Stata gives you:

```
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
male_tim |      39     133.641   5.106689        127        144
female_t |      26    153.6538   13.07499        142        190
```

If you want more detailed output you should type:

```
. summarize male_tim female_t, detail
```

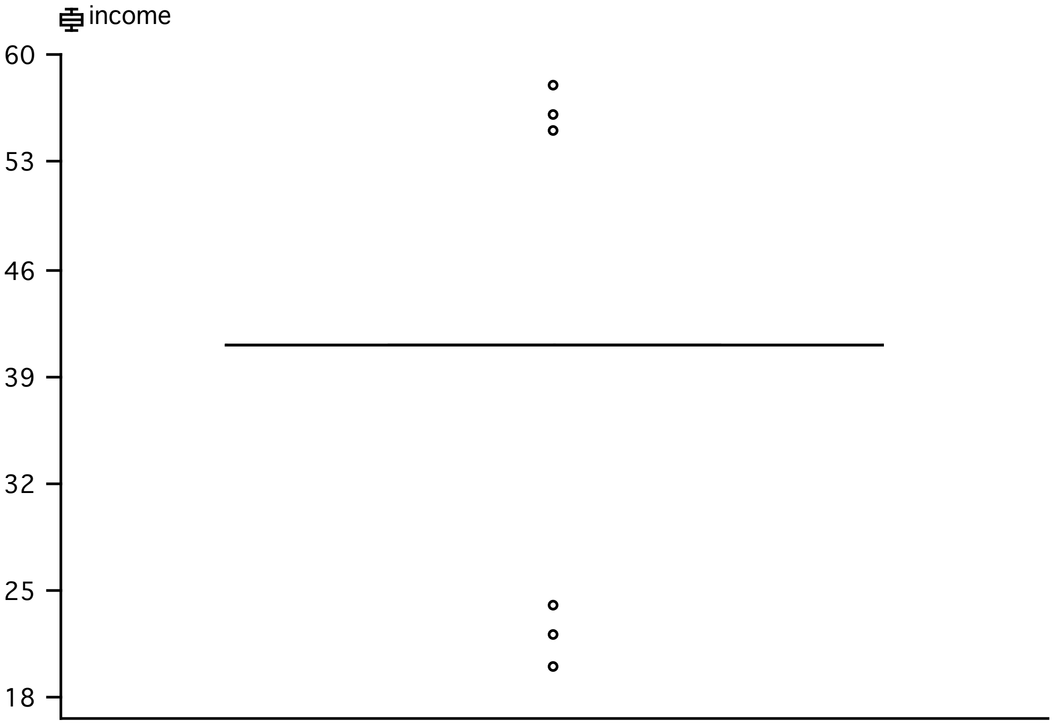
And you will receive this:

```
                          male_tim
-------------------------------------------------------------
      Percentiles      Smallest
 1%          127            127
 5%          128            128
10%          128            128       Obs                  39
25%          129            128       Sum of Wgt.          39

50%          132                      Mean            133.641
                        Largest       Std. Dev.      5.106689
75%          137            142
90%          142            143       Variance       26.07827
95%          144            144       Skewness       .6219865
99%          144            144       Kurtosis       2.113501

                          female_t
-------------------------------------------------------------
      Percentiles      Smallest
 1%          142            142
 5%          143            143
10%          144            144       Obs                  26
25%          145            144       Sum of Wgt.          26

50%          147                      Mean           153.6538
                        Largest       Std. Dev.      13.07499
75%          162            167
90%          168            168       Variance       170.9554
95%          186            186       Skewness       1.482637
99%          190            190       Kurtosis       4.361222
```

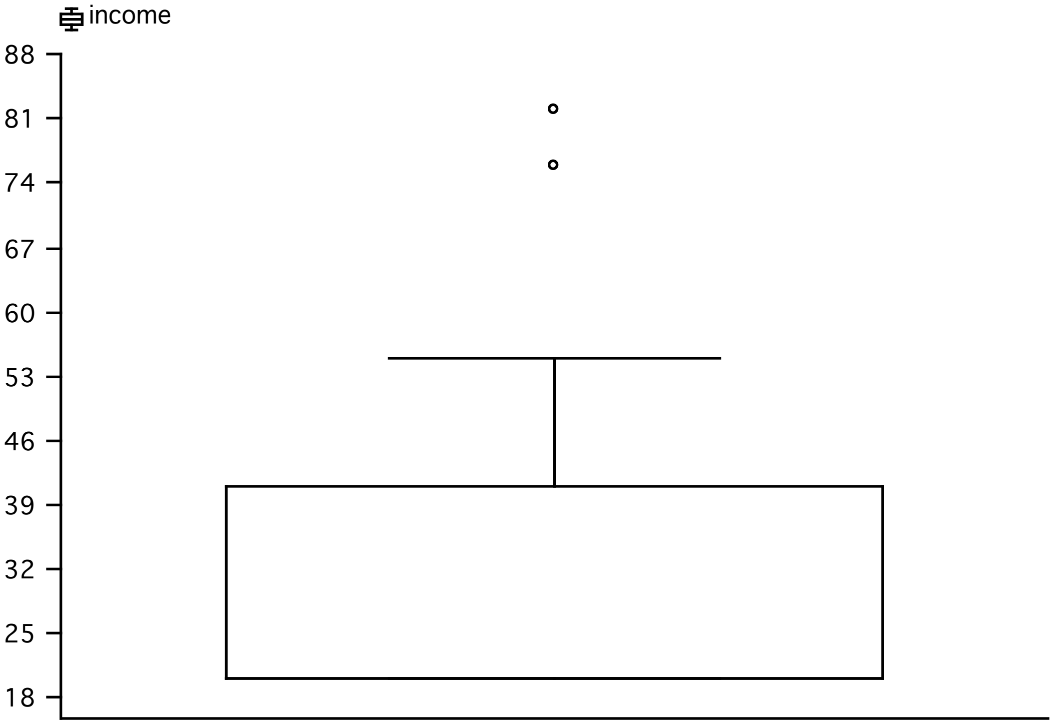
- Question:
Find the median, the first and third quartiles, and compute the interquartile range of the completion time for men and women. Now go back to the boxplots and locate these numbers. Which dataset has larger variation? Try to find a reason for that.

Think about this…

Another data set gives the boxplot below.
What happened here? Generate data that give you approximately the following boxplot.



A different data set has the following boxplot.
Why? Generate data that give you approximately the following boxplot.


income