Plotting Two Empirical CDFs on the Same Graph

(to Illustrate What the Kolmogorov-Smirnov Test is Comparing)

Aim: To plot the Empirical cumulative distribution functions for two (or more) samples on the same graph.

Result: Set out below are descriptions of how to do this in R and JMP.

The data are from p. 130 of S. Siegel, "Non-Parametric Statistics", and show the percentages of errors made by 10 seventh-graders and 10 eleventh-graders in remembering the first half of a series.

Students	Percentage in error									
Seventh-grade	39.1	41.2	45.2	46.2	48.4	48.7	55.0	40.6	52.1	47.2
Eleventh-grade	35.2	39.2	40.9	38.1	34.4	29.1	41.8	24.3	32.4	32.6

In **R**, the following commands produce the graph that appears in Figure 1. It is suggested that it would be more effective to show the two empirical CDFs in different colours, but I obtained the error *parameter "color" couldn't be set in high-level plot() function* when I tried to do this. If you delete the option *verticals = T* in the *plot* or *lines* commands, the dotted uprights do not appear. The *xlim* and *ylim* options ensure that both CDFs will fit on the same graph.

```
x7<-c(39.1,41.2,45.2,46.2,48.4,48.7,55.0,40.6,52.1,47.2)
x11<-c(35.2,39.2,40.9,38.1,34.4,29.1,41.8,24.3,32.4,32.6)
plot(ecdf(x7),xlim=c(20,60),ylim=c(0,1),xlab="No. of errors",main="ECDFS for
    seventh- and eleventh-graders",verticals=T)
lines(ecdf(x11),lty=3,verticals=T)
legend(20,0.9,c("7th grade","11th grade"),lty=c(1,3))
```

Most of the hard work for the *R* program was done by Damian Collins.

If you follow Pam's instructions below, you will obtain the two empirical CDFs in JMP, as shown in Figure 2:

JMP can also do multiple cdf's, but a few tricks are needed. Here is one method which works as long as the data are non-negative. First use Stack in the Tables menu to put all of the data in a single column, with a second column indicating group membership. (Or possibly the data are already structured this way.) Next set up a new numeric column with every entry equal to 0. Go to Survival and Reliability in the Analyze menu, and select Survival/Reliable in the pop-up menu. Specify the data column as Y, the group column as Grouping, the column of zeros as Censor, and check 'Plot Failure instead of Survival'. Voila!

The labels on the graph will need to be edited, especially the 'Failing', but that is easy to achieve by mouse clicking in the right place.

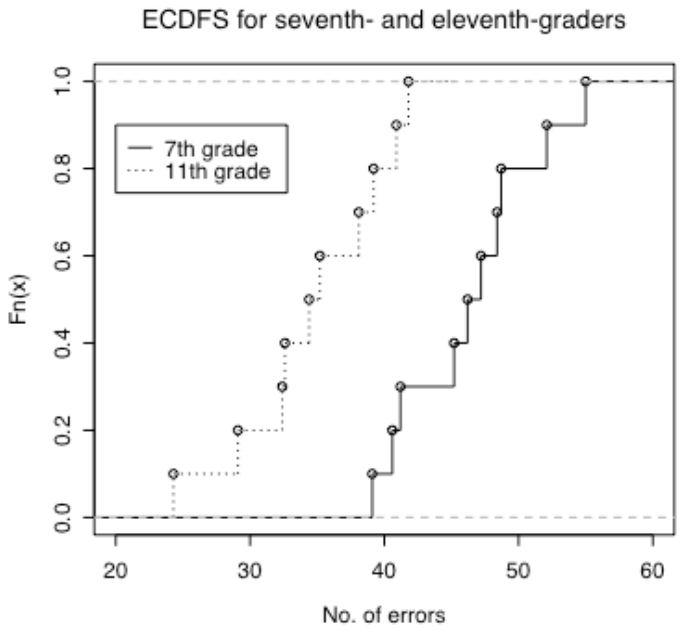
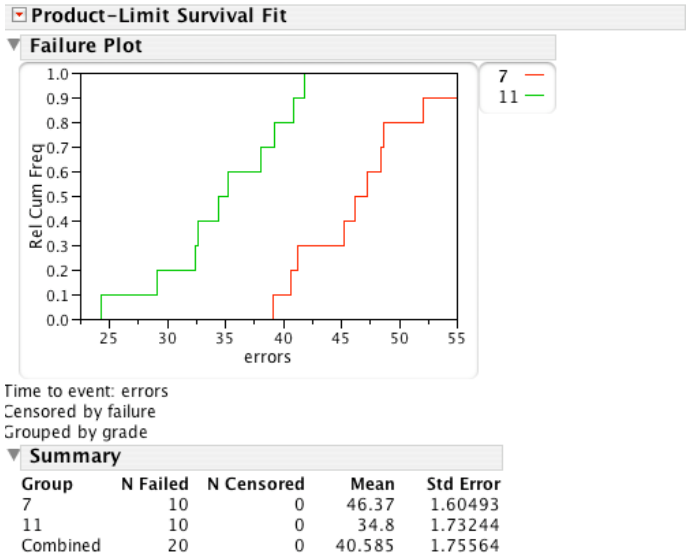


Figure 1: *The output from R.*



Group	N Failed	N Censored	Mean	Std Error
7	10	0	46.37	1.60493
11	10	0	34.8	1.73244
Combined	20	0	40.585	1.75564

Figure 2: *The output from JMP.*

Author: Ken Russell