Supplementary exercise 1.145 of IPS7e

We start by computing descriptive statistics (including skewness and kurtosis) for the two groups and displaying the five-number summaries graphically as side-by-side boxplots; Minitab commands (shortened) and outputs:

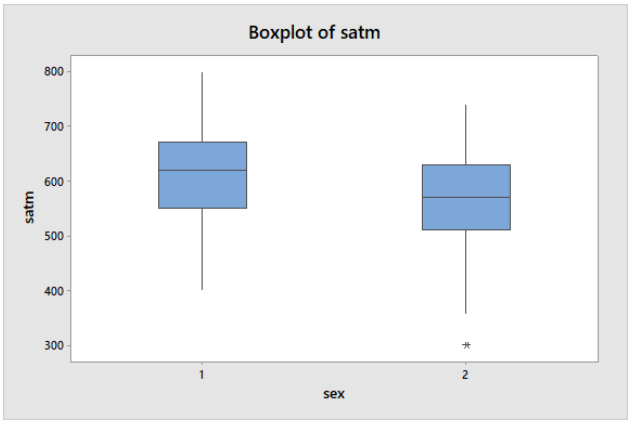
```
MTB > Describe 'satm' 'gpa';
SUBC>   By 'sex';
...
SUBC>   NMissing.
MTB > Boxplot ( 'satm' 'gpa' ) * 'sex';
SUBC>   IQRBox;
SUBC>   Outlier.
```

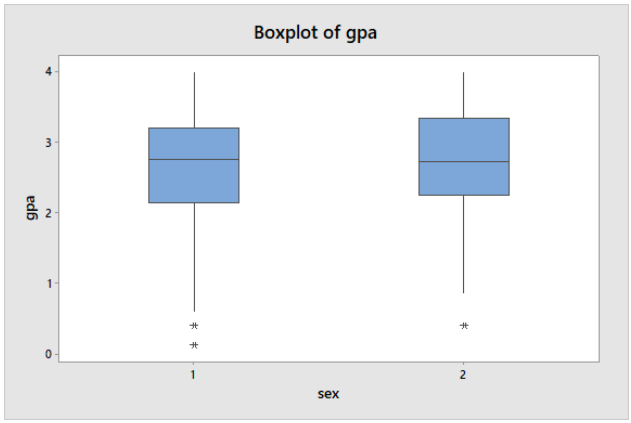
Descriptive Statistics: satm, gpa

Statistics

Variable	sex	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
satm	1	145	0	611.77	6.98	84.02	400.00	550.00	620.00	670.00	800.00
	2	79	0	565.03	9.33	82.93	300.00	510.00	570.00	630.00	740.00
gpa	1	145	0	2.6077	0.0670	0.8068	0.1200	2.1350	2.7500	3.1900	4.0000
	2	79	0	2.6857	0.0820	0.7288	0.3900	2.2500	2.7200	3.3300	4.0000

Variable	sex	Skewness	Kurtosis
satm	1	-0.06	-0.48
	2	-0.47	0.55
gpa	1	-0.71	0.35
	2	-0.59	0.31



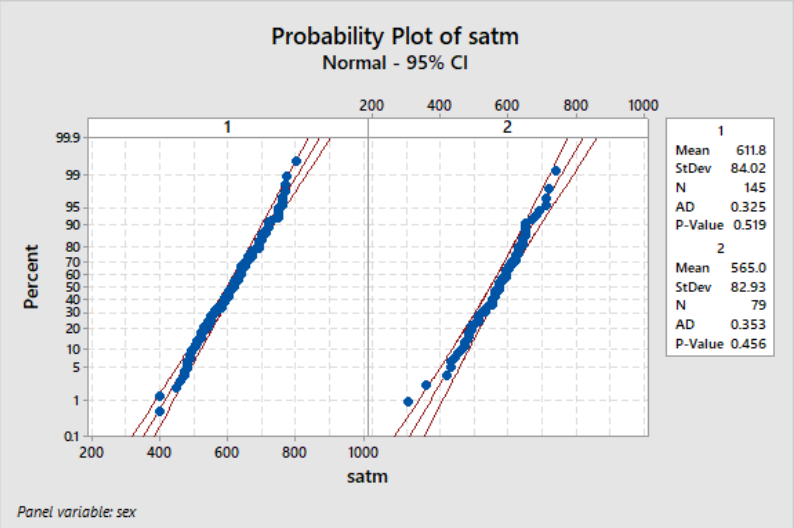


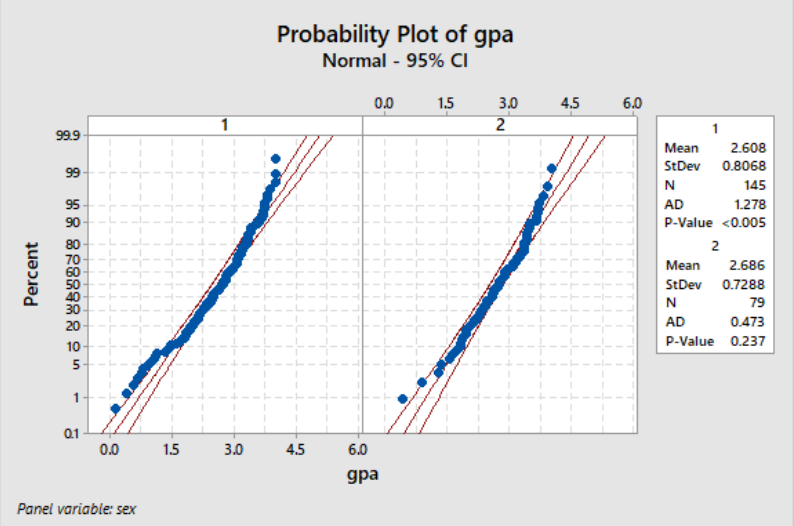
Comments:
The boxplot of SAT math scores show distributions of similar shape, but the distribution for men (sex =1) appears to be centred at a higher value than for women. The difference in means is approximately 50 units. Both distributions appear fairly symmetrical, and their spread is almost the same (when measured by the standard deviation).

In contrast, the two distributions of gpa values appear to be centred at approximately the same value (around 2.7), but both distributions appear a bit left-skewed with longer left than right tails, with a more pronounced left-skewness for the men.

Next, we produce the probability plots. It is probably easier to look at the plots in separate panels for men and women. This is most easily achieved in Minitab by choosing Probability Plot-Single, and adding the panels under the Multiple Graphs-By Variables tab.

```
MTB > PPlot 'satm' 'gpa';
SUBC>   Normal;
...
SUBC>   Panel 'sex'.
```





Comments:
The normal probability plots for satm are nicely straight, and the A-D normality tests are clearly non-significant ($P \gg 0.05$), leading us to the conclusion that there these values can be viewed as approximately normally distributed. The boxplot for women shows one low suspected outlier, and this point is also outside the normal probability limits, but overall the distribution (with this observation included) seems well approximated by a normal distribution.

The normal distribution plots for gpa look less straight because of some upwards curvature at the lower end, reflecting the left skewness. The A-D normality test is clearly significant ($P < 0.005$) for

the men and non-significant for the women. The distributions do not appear that different, and the difference in conclusion from the normality test also reflects the different sample sizes (almost twice as many men as women). With a larger sample size, any deviations from a normal distribution are less likely to appear as the result of random fluctuations. When comparing the points to the overlaid CI lines, more points are outside these bounds at both ends of the distribution for the men; this is reflected in the stronger skewness and would seem to cause the much stronger and significant value for the normality test. In conclusion, we should be cautious in assuming normal distributions for both the gpa distributions.

In response to the question asked in the exercise, ignoring outliers does not seem to substantially alter our impression of the distributions. Generally speaking it is not recommended to "ignore outliers" for the sole purpose of achieving distributions that seem closer to normal. Later in the course we will discuss how to deal with non-normal distributions in the context of statistical inference.