Students' strategies of comparing distributions in an exploratory data analysis context

Biehler, Rolf
University of Kassel, Department of Mathematics
Heinrich-Plett-Str. 40
Kassel, 34132, Germany
E-mail: biehler@mathematik.uni-kassel.de

1. Introduction

This paper builds on findings about students' strategies in comparing distributions that we found in several studies at school (Biehler, 2006) and at university level (Biehler, 2007). We studied students' reasoning after an introductory course in statistics, where the emphasis was put on Exploratory Data Analysis. Our students learned to use the software FATHOM as a tool for data exploration. As part of their assessment, students had to do a project and submit a report. The statistical concepts and tools they learned to use consisted in measures of spread and location, histograms, bar graphs, box plots and percentile plots.

As a rule students did one or several group comparisons in their project using various summaries and statistical graphs for this purpose. The university students got detailed guidelines concerning group comparison and report writing. Nevertheless, we got quite a variation of quality in their reports.

In this paper, I try to point out the different strategies and tool uses we observed without going into detail with regard to how often and under which conditions these strategies were found. I will focus on using box plots as a graphical summary of a distribution.

A specific data set is used as an example in this paper: it is a complex data set with 540 cases and about 50 variables that is based on a questionnaire concerning media use and leisure time of 540 11grade high school students: the so-called Muffins data (Biehler, 2003). These data were also used by the students in their projects.

2. The many facets of box plot uses

Box plots seem to be a very nice display for comparing distributions. Tukey (1977) invented them as a graphical summary that shows information about centre, spread, shape and outliers. The measures of location and spread are robust and especially suited for exploratory analyses. The summary values (median and quartiles) are easy to calculate and seemingly easy to understand. Box plots can be used for easy group comparisons simultaneously with several criteria.

Box plots turned out to be much more difficult than expected (Bakker, Biehler, & Konold, 2005; Biehler, 2001). This is partly due to intrinsic complexities of box plots. But the more important question are "What practices in using box plots in statistical practice can we distinguish?", "What practices of using box plots can we distinguish in educational settings", "What practices do we want to foster in educational settings". In asking this question, we adopt a similar perspective to graphs (inscriptions) as Roth et al. (2005) do. We will reconstruct several different uses and interpretations of a box plot that constrain their use in group comparison tasks. We have also observed explicitly wrong uses, for instance when students do not remember the definitions of the elements of the box plot and draw wrong inferences. But this is not the point I like to develop here.

1. Box plot used as location summary. The five values minimum, lower quartile Q1, median, upper quartile Q3, and maximum summarize the location of the data. The Tukey box plot in Fig. 1 has "whiskers" up to the most extreme value inside the "fences".[1] The data are from the Muffins students and show the distribution of the attribute *Time_Sports.* Students had been asked for the amount of hours per week that they devote to actively going in for sports.

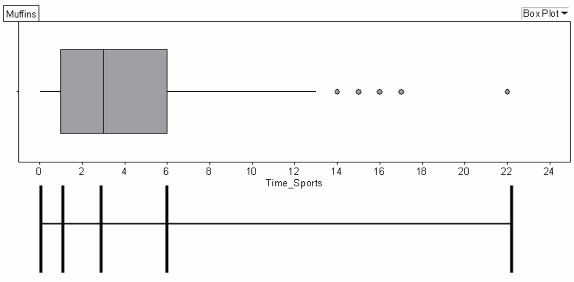


Fig. 1 Box plot as location summary

The median divides the data into two halves, but this is not exactly true, especially if we have ties at the median. The proportion of the data that are lower than the median is at most 50%, and the proportion of the data that are lower or equal to the median is at least 50%.

Many different definitions of quartiles are available (Langford, 2006). We use as FATHOM does the "cdf-definition" that Langford favours, too. The major advantage of this definition is that the following inequalities hold:

$\text{proportion (data} < \text{Q1)} \leq 25\%$ and $\text{proportion (data} \leq \text{Q1)} \geq 25\%$, and consequently

$\text{proportion (Q1} < \text{data} < \text{Q3)} \leq 50\%$ and $\text{proportion (Q1} \leq \text{data} \leq \text{Q3)} \geq 50\%$.

We regard as adequate approximation to these relations when students say that "about 50%" of the data are "inside" the box. Or "about 50% of the data lie between Q1=1 and Q3=6 hours", and so on. However, some tasks require a more precise knowledge about percentages in box plots.

We can observe students using box plots without making reference to any quantitative measure of spread such as the interquartile range. This is surprising, because it seems for experts "so natural" to see the iqr "represented" in a box plot.

2. Box plot for classifying data. The box plot divides the data set into 4 natural intervals, but sometimes students' see rather a separation into 3 parts in a box plot. These domains can be called *low, medium, high*, where the "majority" of the data has a "medium" value.

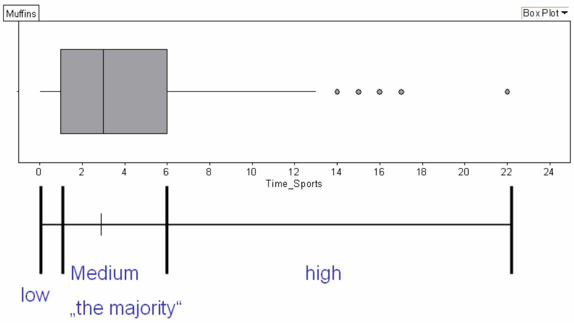


Fig.2 Box plots for classifying data

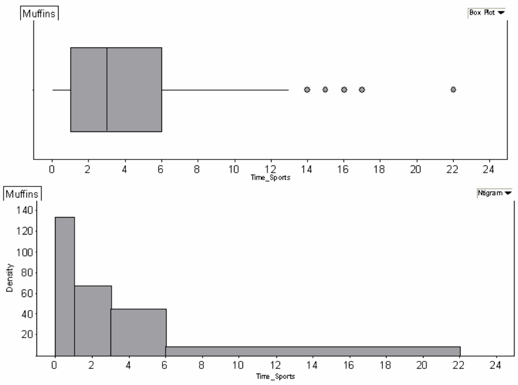


Fig. 3 Assessing "regional density"

3. Regional average density in 4 intervals. Whereas the density histogram with unequal classes shows the density by the height of the column, the density is inversely related to the length of the 4 intervals in the box plot. We call it "regional" because it is neither local nor global and students use box plots for assessing regional density.

4. Regional spread. The lengths of the four quarters, the length of the interval from minimum to

[1] which are defined by $\left[Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)\right]$.

median, from median to maximum, or from Q1 to Q3 are interpreted as what we call “regional” measures of spread. For instance, some students speak of the “spread of the middle half of the data” or of the “spread of the first quarter”.

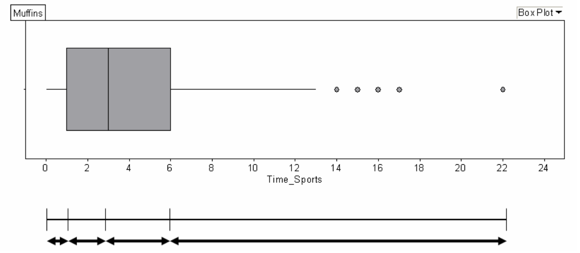


Fig. 4 Assessing “regional spread”

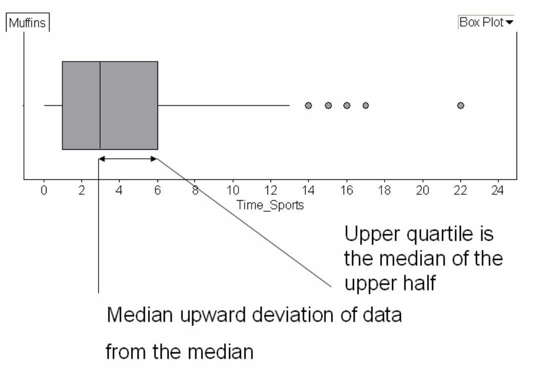


Fig. 5 Box plot interpreted as showing upward and downward spread

This observation forced us to think about possibilities of making students use global measures of spread. We can see the interquartile range also as a kind of global measure of spread similar to the standard deviation, which does not measure the spread of part of the data only.

5. Box plots interpreted as showing upward and downward spread. If we consider the upper quartile as the median of the upper half of the data then the difference *Q3 – median* is the median upper deviation from the median, i.e. about half of the deviation to higher values are larger than the difference *Q3 – median* and about half of them are less than *Q3 – median*. It is natural that we see two different global measures of spread in the box plot, namely *median – Q1* as the median deviation to lower values and *Q3 – median* as the median deviation of the upper values. The single measure of spread, namely the interquartile range as the sum of upward and downward spread does no longer contain any information about the asymmetry of spreading out. On the other hand, the length of the box a visual cue that some students directly interpret as “the” measure of spread. This latter use and interpretation does emerge spontaneously in students’ practice not very often. This is not really surprising because the amount of reinterpretations and new constructions that are necessary is enormous.

6. Box plots used as centre-plus-spread displays. More traditional centre-plus-spread displays show the mean together with plus/minus the standard deviation. Box plots can be used for displaying a centre (by the median) and a measure of spread (interquartile range, length of the grey box). The asymmetry of the box can be interpreted as shape information (skewness). This use does not necessarily imply using box plots for assessing upward and downward spread as was described above.

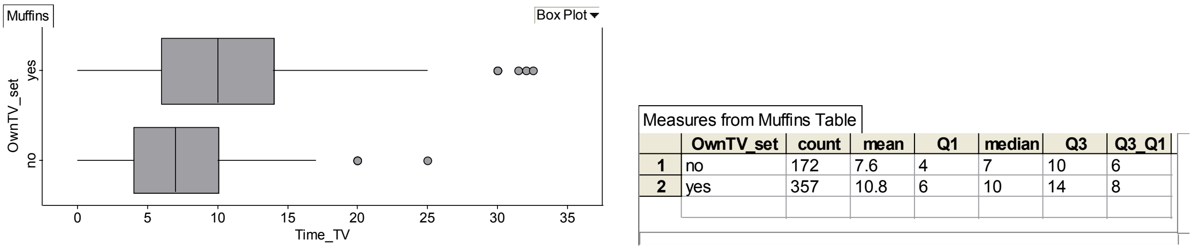
While describing the different uses we have used concepts such as “regional spread” and “median upward deviation from the median”, which are not common in statistics itself. We observed precursors of these concepts in students’ activity. This does not imply that we use all these concepts explicitly in teaching contexts. Some uses such as the “classification use” may be unusual in statistics but may be a reasonable intermediate step that can well take up students intuitive reasoning about modal clumps into account as Konold and colleagues (2002) point out.

3. Different strategies in group comparison tasks

From the work with students we learned that it is even more unclear, what we should regard as a good practices in group comparison tasks in general and with box plots in particular. There is a growing literature on students’ reasoning in group comparison tasks (Hammerman & Rubin, 2006; Pfannkuch, 2006) to which this paper is related.

In the statistical project reports, we got from students in an Elementary Stochastics course (Biehler, 2007), we got two types of poor uses that I called “falling-back to averages” and “distributional overflow". The first means that students do not use much more than averages for comparing groups and neglect all other information. On the other extreme, we find students that collect all the details they can see in each of the box plots and just put them together without much integration. What they pick out from the box plots depends on their use of individual box plots, which may be very different as we have described above.

But we also observed further types of strategies. Let us look at the following example. We have asked the Muffins students for the number of hours that they watch TV and have split the group into two subgroups according to whether the student has a TV set in his living room or not.



Measures from Muffins Table

	OwnTV_set	count	mean	Q1	median	Q3	Q3_Q1
1	no	172	7.6	4	7	10	6
2	yes	357	10.8	6	10	14	8

Fig. 6 TV watching time according to ownership of TV

Centre-plus-spread group comparison. An obvious use of box plots is for comparing the groups with regard to centre and spread. Median and mean are about 3 hours higher in the yes-group; the spread is also higher in the yes-group. We may interpret this as an indication how the ownership of a TV set might influence TV consumption: It is plausible that higher availability can influence the a-mount of consumption and that it may increase variation, because students may react quite differently to the availability of a TV set.

Q-based group comparison. A different use is comparing all the 5 summary values of the box plot in pairs for the two groups, namely that the Q1 (no) is 2 hours less than the Q1 (yes) etc. The *q* stands for *q*uantile-based comparisons.

Cutoff-point based comparisons. In our work with students we often observed the following type of comparison. Students mentally draw a line at the median of the yes group and summarize the data as: Whereas it is only about 50% of the TV-owners who watch less than 10 hours is about 75% of the no-group that watch less than 10 hours. Therefore the owners watch longer TV. We call this cut-off point based group comparisons. One of the quantiles is taken as a cutoff-point, and the frequencies up to these cut-points are compared.

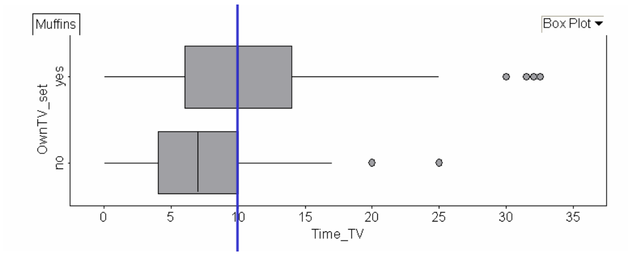
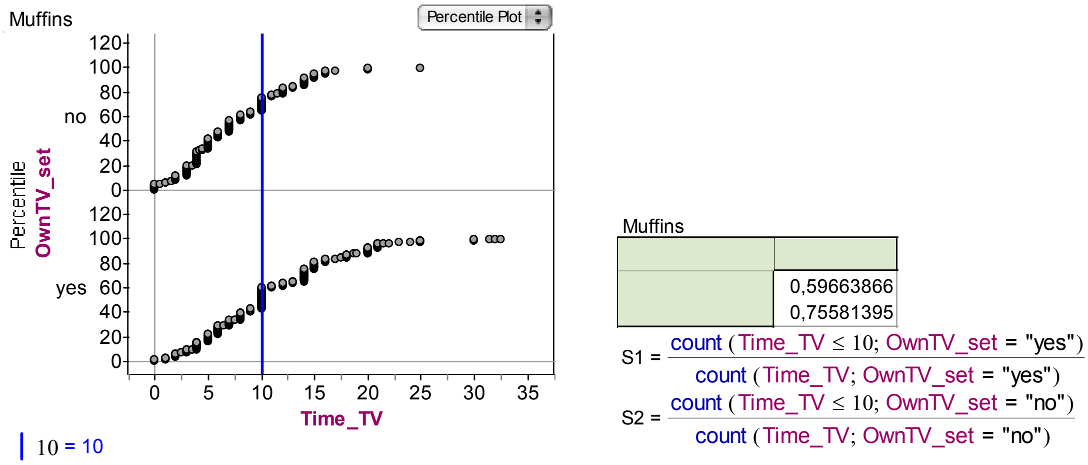


Fig. 7 Cutoff-point based comparisons

This comparison is a special case, because the median of the yes-group coincides with the upper quartile of the no-group. If this is not the case and a cut-off point is somewhere in-between two quartiles, students tend to estimate the frequency below the cut-off point by rough interpolations. The use of box plots for cut-off point based comparisons is certainly something that box plots were not made for, originally. However, cut-off point based comparison strategies are quite common in the media in addition to comparing groups by averages. We also observed the "invention" of this strategy by young students, when they were working with FATHOM, where they could calculate the cut-off point (cumulative) frequencies directly. As Fig. 8 shows working directly with cut-off points is superior to box plots, because the frequency estimations in box plots are only approximate. The example shows that, because of ties, there are about 60 % of the data below or equal the median in the yes-group, whereas the estimation of 75 % from the box plot of the no group is quite accurate.



Muffins

	0,59663866 0,75581395

$$S1 = \frac{\text{count}\,(\text{Time\_TV} \le 10;\ \text{OwnTV\_set} = \text{"yes"})}{\text{count}\,(\text{Time\_TV};\ \text{OwnTV\_set} = \text{"yes"})}$$

$$S2 = \frac{\text{count}\,(\text{Time\_TV} \le 10;\ \text{OwnTV\_set} = \text{"no"})}{\text{count}\,(\text{Time\_TV};\ \text{OwnTV\_set} = \text{"no"})}$$

Fig. 8 Cutoff-point based comparisons with percentile diagrams

For cut-off point comparisons, it would be a better strategy to either use percentile diagrams or calculate cumulative frequencies directly.

4. Different contextual framings of group comparisons

4.1 The "*X* is larger in group A than in group B" – hypothesis

We have introduced some terminology for distinguishing different uses of box plots and group comparisons in the preceding section. Group comparison in the spirit of EDA (Exploratory Data Analysis) means to compare distributions in many different aspects.

It was surprising for us that students often frame their group comparisons in more limited terms. In our example of TV watching an obvious question could be "Do students watch more TV when they own a TV set (as compared to those who own no TV set)?" It is obvious that posing such questions influences the way students do group comparisons. We call such type of hypothesis the "is larger hypothesis". Looking at our box plots we can just say: yes, this hypothesis is confirmed by the data. This student strategy may be called the "hypothesis testing strategy". Ending an analysis with just stating the confirmation is its simplest version, where the data are not much exploited. An elaboration would be stating *how much more* one group watches TV as compared to the other. Konold & Pollatsek (2002) report that taking the difference of means to answer to the how-much-question is not natural for young students. Our elder students find this more natural, but do not always use it in this context.

What are the different ways students deal with an "is larger hypothesis"? For most students the questions that is asked seems to be clear, although from an expert point of you it is far from being clear, what it could mean that *X* is larger in group *A* than in *B* given that there is variation in each group. The students use one or more criteria to check whether the statement is true, they choose mean, median, and the quartiles. Some students argue that the more of these values are higher in group *A* than in *B* the higher is the "evidence" in favour of the hypothesis. If some of the pairwise differences of quartiles or measures of centre are negative and some are positive, this strategy comes to a limit, as the "evidence" seems to point in different directions. An intuitive strategy that some students use in this context is just to count the number of measures that are larger in group *A* than in group *B* and if this is the case for the majority of measures, the students consider the hypothesis as confirmed.

Some students use cut-off point-based comparisons to even better support the "is larger hypothesis". For instance, let us look at a cut-off point c = 10 in our above example, we have:

$$rel.frequency\left(X \le 10 \,\middle|\, own\_TV\right) \approx 60\% < 75\% \approx rel.frequency\left(X \le 10 \,\middle|\, no\,TV\right).$$

It is intuitively clear for most students that this is a criterion for that TV consumption is higher in the owner group (less are below the cut-off point of 10, respectively more are above the cut-off point 10). The general criterion is: If $rel.frequency\left(X \le c \middle| A\right) < rel.frequency\left(X \le c \middle| B\right)$, this is an indication that X is larger in group B then in group A.

We observed these strategies especially in contexts, where students had not been very much taught about group comparisons in any explicit way. We have experimented with some remedies. First of all, we have to expound the problems of phrasing "The owners watch more television than the non-owners". In careful media reporting, we find formulations such as "The owners watch more television than the non-owners, *on average*". Some students use this phrasing but then tend to just compare means or medians. On the other hand, this usage is an indication that they may be aware that there is variation within the groups and that not all TV owners watch more than the non-owners.

As we intend to support a group comparison beyond averages, we suggest rephrasing such hypotheses as "The owners *tend to* watch more television than the non-owners". The next important step is using the pairwise comparison of measures of location and cut-off point comparisons not as accumulating evidence for the truth of the hypothesis but as answer to the descriptive question "In which respect and by which amount do the students watch more TV in the yes-group?

Generally, it could well be that there is no unique answer to the question which group tends to be larger with regard to a certain attribute.

4.2 Group comparison in a decision context

Let us look at another example of group comparison for clarifying the issue. In the context of a research and development project where Paul Cobb, Koeno Gravemeijer and Arthur Bakker were involved the "battery problem" was used as an example for group comparison (Bakker & Gravemeijer, 2004). The students get a sample of the lifetimes from two brands of batteries and have to decide, which brand of batteries is "better".

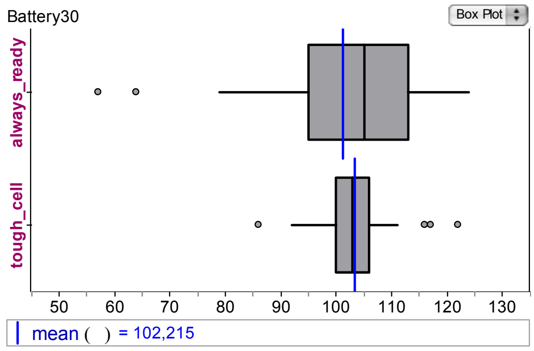


Fig. 9 Box plots "battery problem"

The box plots with added means in Fig. 9 show a contradictory picture: mean and median convey different information with regard to which type lasts longer (on average). The spread is different too and the comparison of the quartiles, minimum and maximum gives different directions. The answer whether one type lasts longer than the other cannot be answered unambiguously. The authors do not use a descriptive context but set up a situation for decision making to achieve a unique solution. For instance, if we intend to take 10 batteries and we are interested to maximize the total sum of life span, then the mean would be a criterion for our choice: "tough cell" is "better" in this respect. If we need batteries that have to last at least 110 hours for some reason (cutoff point strategy) you will find them more likely in the brand "always ready". If you wish to avoid - as far as possible - batteries that last less than 95 hours (another cut-off point), you will have to choose "tough_cell". In these varying decision contexts, developing different criteria about which group is better seems to be quite natural. The descriptive ambiguity with regard to the question which brand tends to have a "higher lifespan" will lead to different unique decisions depending on the respective criterion. By this "trick" the descriptive ambiguity is made visible first but then eliminated in a second step. A major disadvantage of the battery problem is of course that the decision is based on a very small sample and it is implicitly assumed that future samples will be very similar. It is therefore questionable whether this example is a good elementarization of statistical practice if the sampling problem and related uncertainties are never made explicit. The descriptive contexts we use avoid this problem, however, we also observe students that make comparisons and inferences from their comparisons that only been done if the data were a random or representative sample of something. A description task as such may result into ambiguities. We think that it is reasonable to keep these descriptive ambiguities, however, in some examples we may wish to show that the plurality of criteria that can lead to a unique decision if the context provides further constraints.

4.3 Some remedies: the shift model

In the context of analysing a single distribution, we observed strategies that I call *distributional itemizers* versus *distributional integrators* (Biehler, 2007). The first class of students diligently collects a list of all kinds of features whereas the second group tries to integrate information into an overall picture of a distribution. The situation is even worse in group comparison tasks, where integration of results is more difficult. It is certainly naïve assuming that students spontaneously can provide acceptable group comparisons.

Konold & Pollatsek (2002) argue for putting the mean as a signal more into the foreground as compared to the distribution (around the mean) which is considered as noise. In many real statistical studies, hypotheses concerning different means are tested, and students have to learn that the mean can be seen as such a signal for a process, and that a changed mean is often an indication that the process has changed. How can we combine this insight with a distributional perspective that takes the whole distribution as an indication of an underlying process? Many (parametric) tests concerning means such as the two sample *t*-test are essentially not only concerned with means but the tests are concerned with two distributions (often normal ones) that only differ by the mean and that are identical in all other aspects. In nonparametric statistics, we have two different generalizations of this model. The first generalization is the "additive shift model". The question to be tested is whether a whole distribution is shifted by a certain number *a*. This implies that the mean and the median and all quantiles are also shifted by the same number *a.* A further generalization is the notion that a random variable *X* is *stochastically larger* than a random variable *Y* if $\mathrm{Prob}\left(X \le c\right) \le \mathrm{Prob}\left(Y \le c\right)$ for all *c*. If the random variables have a continuous distribution this is equivalent to the following statement about quantiles: $Q_p(X) \ge Q_p(Y)$ for all $p \in \left(0,1\right)$. If the additive shift model holds, we have X has the same distribution as $Y + a$, and if $a > 0$ this implies that *X* is stochastically larger than *Y* and that the properties hold $Q_p(X) = Q_p(Y) + a$ for all $p \in \left(0,1\right)$. If $a < 0$, *Y* is stochastically larger than *X.* Thus the shift model is a special case of being "stochastically larger". Another special case is the multiplicative shift model: X has the same distribution as $a \cdot Y$, $a > 0$. If $a > 1$, X is stochastically larger than *Y* and all the quartiles multiply by the same factor *a,* which is also the factor by which any measure of spread increases.

Based on this analysis, we suggest several descriptive *models* for group comparisons to our students. The function of the model is to help them with approximately summarizing the differences and relations between the groups. Students have the choice between the following

- *X* is (statistically larger) in group *A* than in group *B*: we speak of a shift to higher values
- The distribution of *X* in group *B* can be described as an *uniform additive shift* of the distribution of *X* in group *A*
- The distribution of *X* in group *B* can be described as an *multiplicative shift* of the distribution of *X* in group *A*
- The difference in distribution between both groups is more complex

The quartiles and the median in the box plot can be taken as pragmatic indicators for assigning a descriptive model to the data. An extension would be to check the whole cumulative distribution function or to use QQ-plots for analyzing the relation between two distributions. Cleveland (1993) discusses how to use QQ-plots for diagnosing additive and multiplicative shifts between two data sets. We do not recommend this extension for teaching in introductory statistics education, but this is of course important background knowledge. If we look back to our example of TV watching, we can model the comparison by a multiplicative shift model now: Those with a TV tend to watch about 35% more than those without a TV. This factor of 1,35 does not only apply approximately to the mean and median but also accounts for the increase in spread from an interquartile range of 6 in the no-group to 8 in the yes-group. This descriptive modelling also allows the students to express not only the comparative statement that there is a shift, but they can also quantify the shift. Most important, the shift relates to the distribution as a whole and applies not only to the mean. We think that it is important that students can associate different mental pictures when hearing about mean differences: Various shift models can hold but distributions may also differ in relevant other aspects.

5. Conclusions

We tried to make aware of a large variety of students' strategies of comparing distributions, especially when using box plots. We have mentioned some remedies for overcoming limited strategies and replacing them by more advanced ones. Further research and development work has to show how to organize teaching in order that more adequate practices for group comparisons may develop.

SOFTWARE

FATHOM http://www.keypress.com/fathom/ or http://www.mathematik.uni-kassel.de/~fathom

REFERENCES

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 28 June-3 July 2004. [www.stat.auckland.ac.nz/~iase/publications.php]* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147-168). Dordrecht: Kluwer.

Biehler, R. (2001). Statistische Kompetenz von Schülerinnen und Schülern - Konzepte und Ergebnisse empirischer Studien am Beispiel des Vergleichens empirischer Verteilungen. In M. Borovcnik, J. Engel & D. Wickmann (Eds.), *Anregungen zum Stochastikunterricht* (pp. 97 – 114). Hildesheim: Franzbecker.

Biehler, R. (2003). Interrelated learning and working environments for supporting the use of computer tools in introductory courses. . In *Proceedings of the IASE Satellite Conference on Teaching Statistics and the Internet[http://www.stat.auckland.ac.nz/~iase/publications/6/Biehler.pdf]*.

Biehler, R. (2006). Working styles and obstacles: Computer-supported collaborative learning in statistiscs. In *Proceedings of ICoTS 7, Salvador de Bahia, Brazil [www.stat.auckland.ac.nz/~iase/publications/17/2D2_BIEH.pdf]*.

Biehler, R. (2007). Assessing students' statistical competence by means of written reports and project work. In B. Chance & B. Philipps (Eds.), *Proceedings of the IASE Satellite Conference on Assessing Student Learning in Statistics, Guimaraes, Portugal, August 2007*.

Cleveland, W. (1993). *Visualizing data*. Murray Hill, NJ: AT & T Bell Laboratories.

Hammerman, J. K., & Rubin, A. (2006). Saying the same (or different) thing: How shape affects ideas about distribution in a software exploration environment. In *Proceedings of ICoTS 7 [http://www.stat.auckland.ac.nz/~iase/publications/17/6E3_HAMM.pdf]*.

Konold, C., & Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. *Journal for Research in Mathematics Education, 33*, 259-289.

Konold, C., & Robinson, A., et al. (2002). Students' use of modal clumps to summarize data. In *Proceedings of ICOTS 6 [www.stat.auckland.ac.nz/~iase/publications/1/8b2_kono.pdf]*.

Langford, E. (2006). Quartiles in Elementary Statistics. *Journal of Statistics Education, 14*(3).

Pfannkuch, M. (2006). Informal inferential reasoning. In *Proceedings of ICoTS 7 [http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]*.

Roth, W.-M., Pozzer-Ardenghi, L., & Han, J. Y. (2005). *Critical Graphicacy*. Dordrecht: Springer.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.