PRESENTATION OF STATISTICAL CONCEPTS WITH DYNAMIC GRAPHICS AND SIMULATIONS IN R

Andrej Blejec
National Institute of Biology and University of Ljubljana
Vecna pot 111, SI-1000 Ljubljana, Slovenia
andrej.blejec@nib.si

Understanding statistical concepts is important for proper use of statistics. The idea of using simulations and dynamic graphics to foster understanding of statistical concepts is not new. In recent years, R became the lingua franca *for statistical data analysis. While R graphical devices are not meant for display of animated graphics, my aim is to use base R graphics for display of animated graphical sequences. To enable dynamic graphics in R, I developed a package* `animatoR`*, which supports smooth transitions of graphical elements and simplifies preparation of animated displays. I will show some animations that can be useful for statistics teaching and present basic features of the package. Package* `animatoR` *is freely available at* `https://github.com/ablejec/animatoR`.

INTRODUCTION

Visualizations and graphics are recognized as an effective way to present data and results of statistical analyses. Dynamic and animated graphics are considered as an effective way to display phenomena that are changing over time or space. The importance of dynamic visualisations is recognised in modern statistics education and included in data visualisation courses (Heer, 2018). The importance and effectiveness of animated graphics was demonstrated by Hans Rosling (2010a) in his presentation at ICOTS8 (Lecture video is available at http://videolectures.net/icots2010_rosling_wshdw/).

In statistics teaching, animated graphics can be used for demonstration of stepwise procedures. Animated graphics can augment simulation, randomization and permutation procedures (Budgett and Wild, 2014) and display of information that changes in steps or over time. Several elegant solutions exist for production of web based animated graphics, but they lack the statistical machinery provided in R (R Core Team, 2016). Since R is extensively used in statistics research and teaching, my aim was to extend the R base graphics capabilities with a package `animatoR` (Blejec, 2011, 2016). Other solutions in R are packages `animation`, `tweenr`, or `gganimate`, that also display sequences of frames. Package `animatoR` adds smooth transitions between frames.

METHODS

Since R graphic devices are in a sense static, several approaches towards dynamic graphics are used. The most common way is to plot a series of complete pictures, each one with relocated picture elements. If the pictures are not very complex, R is fast enough (if not too fast) for producing a flicker free dynamic impression. This is the most popular technique, which can provide satisfactory results. To get an impression of smooth movement, the changes in successive pictures should be small and one needs to get many intermediate point or line positions. Package `animatoR` provides such technique and a set of functions that complement base graphics function for production of dynamic graphics. The basic idea is to define the starting and finishing coordinates of moving picture elements (points, lines, segments, ...). Then we plot a series of pictures for successive intermediate positions, which are calculated by homotopy between starting and finishing values. If the start position is x_0 and the end position is x_1 then positions between them can be determined as

$$x(t \mid x_0,\, x_1) = (1-t)\cdot x_0 + t\cdot x_1, \qquad t \in [0,1]$$

for different values of homotopy parameter t. Parameter t can be considered as the time, controlling the display of the animated sequence. Selection of suitable sequence for homotopy parameter t provides an impression of smooth movement along trajectories from starting to finishing positions. In addition, one can control the times of picture elements appearance and disappearance, as well as the times of position changes. Attributes encoded in color, symbol size and line width can also be dynamically changed.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*.
Voorburg, The Netherlands: International Statistical Institute. iase-web.org

EXAMPLES

To produce an animated graphics, one can decompose static graphics into picture elements (points, lines, segments etc) and define when and for how long they are displayed and at what time they change position. In this section we will show few examples of animated graphics, that can be used in statistics teaching to support the explanation of statistical concepts. Animated sequences are embedded into the PDF file with LaTeXpackage `animate`. The animation can be started and controlled by buttons below the graphics if the file is displayed in Adobe® Acrobat Reader with enabled JavaScript (see Preferences).

Quantiles

In this example (Figure 1) we show determination of quantiles, specifically the popular quartiles. First the data are scattered from some central value as sorted points. Next, the cumulative distribution is shown with cutting the probability axis and knocking down the values that are hit. The quartile values are emphasized below the original sorted points.

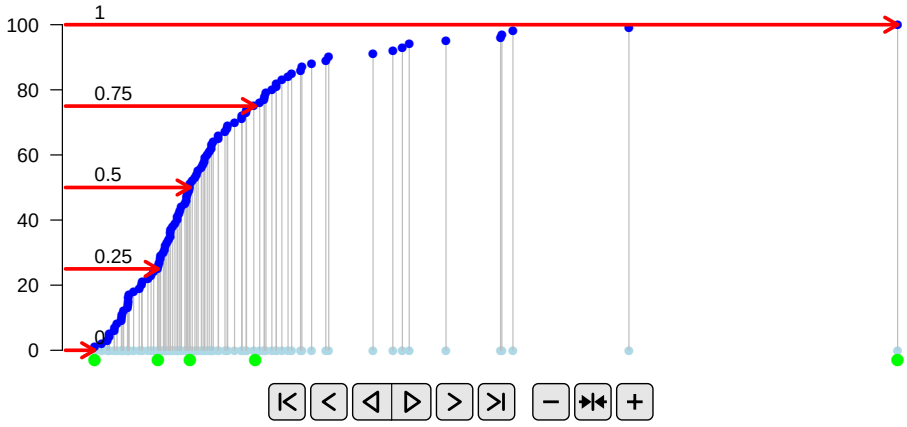


Figure 1: Determine quartiles.

Eigenvectors

What happens to a vector when it is multiplied by a matrix? It changes lengths and rotates. But some vectors do not rotate, just change lengths, and they are called eigenvectors of a matrix. This is displayed in Figure 2. Colored vectors of same lengths are multiplied by a matrix. Pick your favorite color and see if you picked an eigenvector.

```
     [,1] [,2]
[1,]    1    2
[2,]    2    1
```

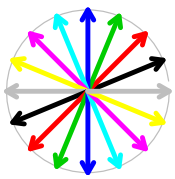


Figure 2: Find the eigenvector.

Goodness of fit

Coefficient of determination (R^2) is not a good measure of the goodness of fit (Stare, 1995). The model which fits data well will have R^2 close to one, the reverse is not necessarily true: in the linear regression, R^2 increases with the increase of the slope. While it is true, that the proportion of explained variance is larger, the fit (the distances of the data from the model) does not change. The animation in the Figure 3 shows the effect of changes in the intercept and slope on the R^2. Side bars show unexplained variation (red) and explained variation (grey and blue, according to the points color).

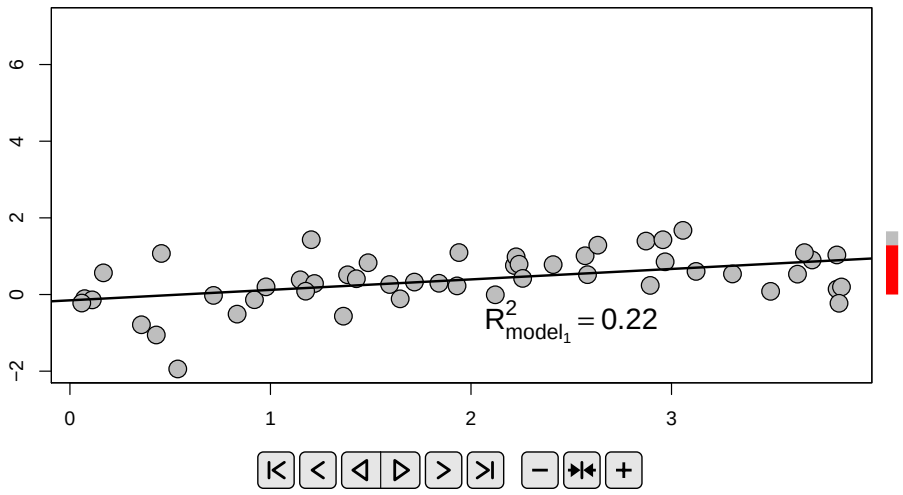


Figure 3: R^2 does not measure the goodness of fit.

Measurement scale determines the selection of procedures

Students usually consider measurement scales as a boring part of the course. But measurement scale consideration might be the most important part of analysis preparation. Can we compute the mean (or the sum) and what would be the meaning of such a value? The answer to that question leads to decision of parametric or nonparametric statistics or selection of procedure that combines numerical and categorical measurement scales. A nice example is the analysis of dependence (not necessarily in the causal sense) of two phenomena giving four combinations of variables measured either numerically or in categories.

We can start with the difficulty of getting the joint distribution from two marginal distributions. Without some relationship, any combination of x and y is possible. Figure 4 shows several possible joint distributions for the same marginals.

Scatterplot on Figures 4 shows the relationship if both variables are measured on the numerical scale. One can try to use correlation and regression, possibly getting better predictions than given by the average of the dependent variable alone (animation in Figure 5a). The other panels in Figure 5 show the situations, where one or both variables are converted to the categorical scale. Conversion of the predictor to a categorical variable is shown in the animation in Figure 5b. If predictor is measured on the categorical scale, mean value of categories is meaningless and analysis of variance can be appropriate. Figure 5c shows the case where the predicted variable is cut into two categories, which leads to use of logistic regression. Finally, if both variables are measured in categories, contingency analysis comparing the group and marginal counts is a possible choice (Figure 5d). The shown animations are all based on the same original data and transformed in a dynamic way. Controls allow inspection of the first and final frame as well as the manual stepwise inspection. Similar animations can be used for presentation of statistical concepts based on simulated data with known statistical properties (Blejec, 2002).

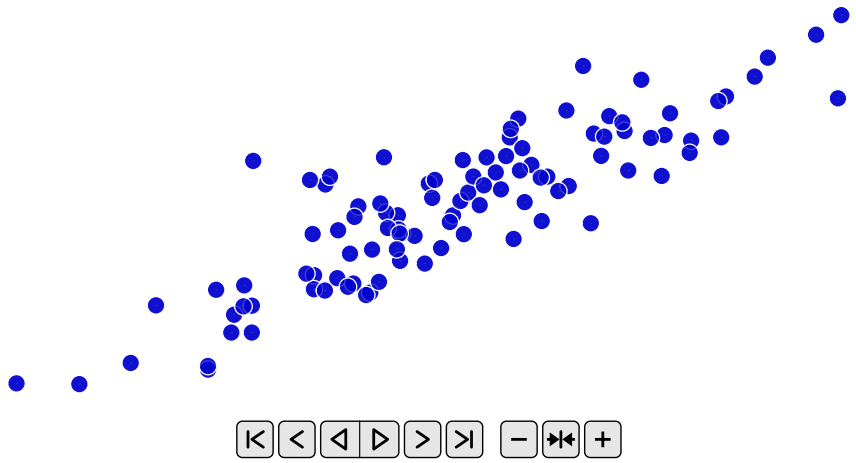
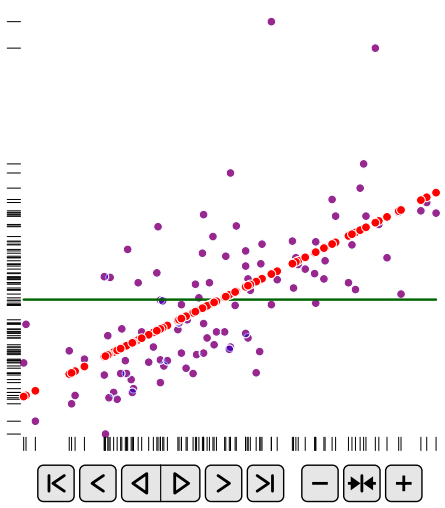
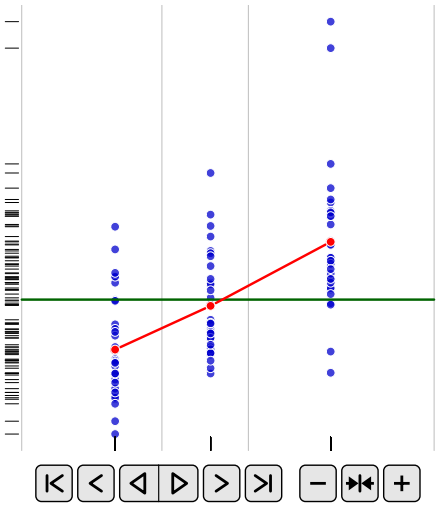


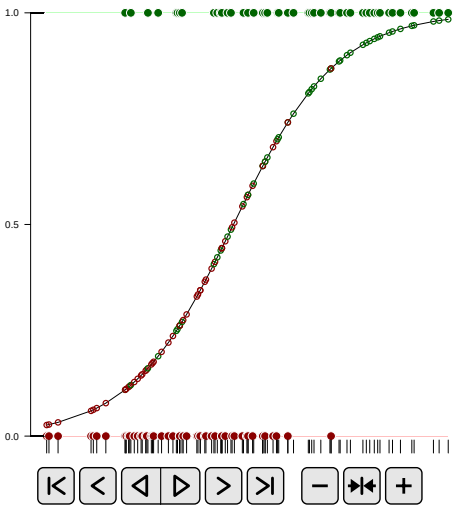
Figure 4: From joint to marginal distributions and back.



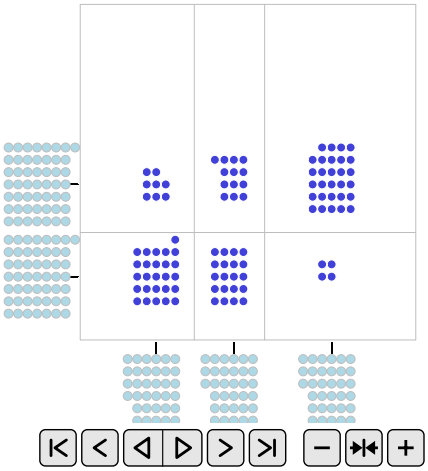
(a) N ~ N: regression



(b) N ~ C: analysis of variance



(c) C ~ N: logistic regression



(d) C ~ C: contingency analysis

Figure 5: Methods to analyse dependence (C: categorical and N: numerical scale).

Dynamic scattergram

The last example (Figure 6) shows a dynamic scattergram (bubbleplot) showing changes of several phenomena in time. In addition to the coordinates for the main two variables, population size and geographic origin are encoded as the size and color of the bubble. Rosling (2010b) demonstrated on several occasions, that such displays can be very effective, especially in combination with a passionate explainer.

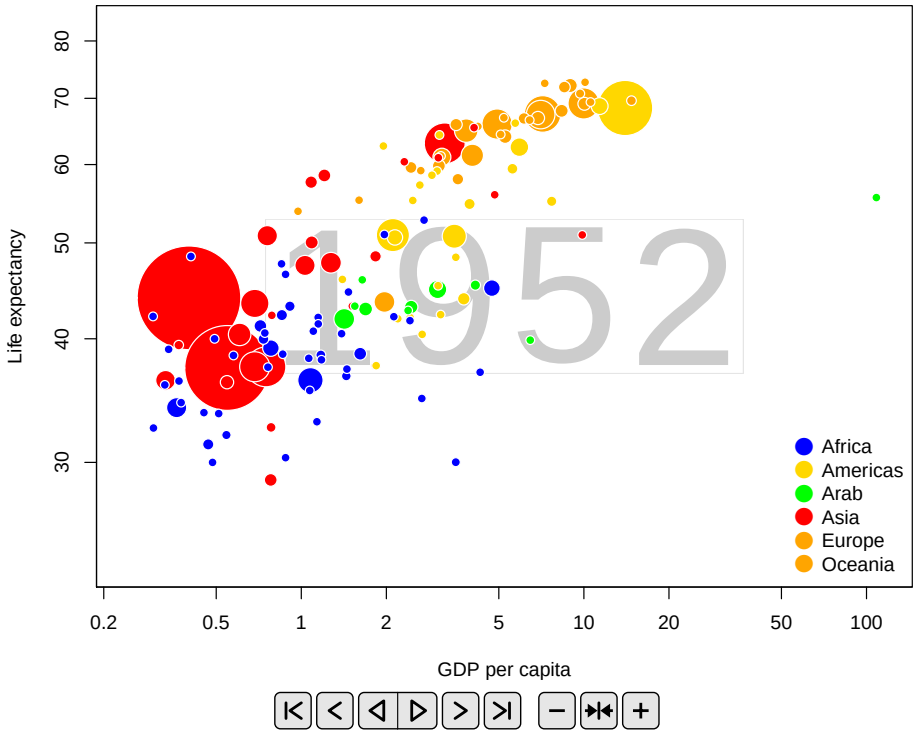


Figure 6: Dynamic scattergram.

CONCLUSION

Visualization and graphical displays are important parts of statistical practice. Good visualization can reveal the structure of presented data. Dynamic visualization can uncover the changes over time and position. Dynamic visualization can be used also to display changes in statistical procedures and their properties. They can be particulary handy if the procedure develops in several stages. Procedures related to resampling (e.g. confidence intervals), permutation and randomization (e.g. permutation tests) or combine/distribute steps (e.g. various non-parametric tests) can be effectively illustrated with dynamic graphics.

R is a number one software for doing and teaching statistics. Our aim was to provide support for use of dynamic graphics in R environment. In the modern reproducibility of research setup it is particularly interesting to include dynamic graphics in the final documents and reports. This paper is an example of such reproducible report, composed in combination of R, `knitr` and LaTeX. For animations we used package `animatoR` which is freely available on GitHub at `https://github.com/ablejec/animatoR`.

For illustration of properties of statistical methods and procedures, statistical simulation and resampling are often used. With the support of animated graphics they can be shown in a dynamic way and hopefully enhance student's understanding. To some extent, simulations and dynamic graphics can be an answer to the question asked by Moore (1996): *"If an audience is not convinced by the proof, why do proof?"*

REFERENCES

Blejec, A. (2002). Teaching statistical concepts with simulated data. In Phillips, B., editor, *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS6, July, 2002), Cape Town, South Africa*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/ iase/publications.php.

Blejec, A. (2011). animatoR: dynamic graphics in R. In *The R User Conference, useR! 2011*, Coventry, UK. University of Warwick. http://www.warwick.ac.uk/statsdept/user-2011/abstract_booklet.pdf.

Blejec, A. (2016). *animatoR: Support for Animated Graphics in Base R Graphics.* [Software] Available from https://github.com/ablejec/animatoR.

Budgett, S. and Wild, C. (2014). Student's visual reasoning and the randomization test. In Makar, K., de Sousa, B., and Gould, R., editors, *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/ iase/publications.php.

Heer, J. (2018). CSE442 Data visualization https://courses.cs.washington.edu/courses/cse442/17sp/.

Moore, D. S. (1996). New pedagogy and new content: The case of statistics. In Phillips, B., editor, *Papers on Statistical Education presented at ICME-8 (International Congress on Mathematics Education-8) Seville, Spain, July 14-21, 1996*.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rosling, H. (2010a). What showbiz has to do with it. In Reading, C., editor, *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/ iase/publications.php.

Rosling, H. (2010b). What showbiz has to do with it. [Video] Available at http://videolectures.net/icots2010_rosling_wshdw/.

Stare, J. (1995). Some properties of R^2 in ordinary least squares regression. *Metodoloski zvezki*, 10:133–145.