CrossMark

Boxplots for grouped and clustered data in toxicology

Philip Pallmann[1] · Ludwig A. Hothorn[2]

Abstract The vast majority of toxicological papers summarize experimental data as bar charts of means with error bars. While these graphics are easy to generate, they often obscure essential features of the data, such as outliers or subgroups of individuals reacting differently to a treatment. In particular, raw values are of prime importance in toxicology; therefore, we argue they should not be hidden in messy supplementary tables but rather unveiled in neat graphics in the results section. We propose jittered boxplots as a very compact yet comprehensive and intuitively accessible way of visualizing grouped and clustered data from toxicological studies together with individual raw values and indications of statistical significance. A web application to create these plots is available online.

Keywords Graphics · Statistics · R software · Body weight · Micronucleus assay

Introduction

Preparing a graphical summary is usually the first if not the most important step in a data analysis procedure, and it can be challenging especially with many-faceted datasets

✉ Philip Pallmann
  p.pallmann@lancaster.ac.uk

1 Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK

2 Institute of Biostatistics, Leibniz University Hannover, 30419 Hannover, Germany

as they occur frequently in toxicological studies. However, even in simple experimental setups many researchers have a hard time presenting their results in a suitable manner. Browsing recent volumes of this journal, we have realized that the least favorable ways of displaying toxicological data appear to be the most popular ones (according to the number of publications that use them).

Some researchers refrain from drawing graphs at all and publish their summarized results in a table that typically contains group-specific means, standard deviations (SD), sample sizes, and symbols indicating statistical significance of group comparisons, often for multiple endpoints. An example of such a table from a recent study on long-term intake of the "fat burner" L-carnitine (Empl et al. 2014) is shown in Table 1. The obvious problem with tables is that it can be extremely tough to grasp the big picture.

The dominating type of graphic in toxicological journals to this day is the bar chart. It comprises more or less the same summary measures as most tables (means, SDs, symbols to flag significant effects), as we can see from an example taken from a study on toxicity and bioaccumulation of aluminum nanoparticles (Park et al. 2015) shown in Fig. 1.

A slight variation is line diagrams where the quantities depicted are essentially the same as in bar charts. The only difference is that the means are drawn as points instead of bars and connected across groups. More often than not the connecting lines do not convey any additional information whatsoever or are even misleading in that they suggest linear changes (which may be true or not), as in the example from a study on methanol teratogenicity (Miller-Pinsler et al. 2015) shown in Fig. 2.

Even though tables, bar charts, and line diagrams allow for a compact display of data, they have two major drawbacks: First, the summary statistics involved are only

Table 1 A summary table with means, SDs, sample sizes, and various symbols indicating significant effects [reproduced from Empl et al. (2014)]

	Group 1 (0 g/l)	Group 2 (1 g/l)	Group 3 (2 g/l)	Group 4 (5 g/l)
Number of animals	19	19	19	20
Starting body weight (g)[a]	105.1 ± 7.9	102.4 ± 7.0	101.9 ± 6.9	101.6 ± 6.9
Final body weight (g)[a]	410.5 ± 19.6	408.5 ± 14.6	413.2 ± 14.1	414.5 ± 18.3
Kidney weight (g)[a,*]	1.27 ± 0.08	1.26 ± 0.07	1.28 ± 0.1	1.30 ± 0.09
Liver weight (g)[b]	11.4 ± 1.0	11.4 ± 0.9	11.6 ± 0.8	11.9 ± 0.9
Spleen weight (g)[b]	0.82 ± 0.1	0.83 ± 0.08	0.84 ± 0.06	0.84 ± 0.08
Water uptake (ml/day/animal)[▲,a]	18.1 ± 1.6	18.1 ± 2.1	18.3 ± 1.2	$18.8 ± 1.8^{\alpha,\beta}$

Where applicable, values are shown as mean ± standard deviation

* Weight of one kidney

▲ Calculated on the basis of the mean of the weekly water consumption of two animals/cage

[a] Kruskal–Wallis test followed by Dunn's post hoc analysis

[b] One-way ANOVA followed by Tukey's post hoc analysis

α Significantly different ($p < 0.01$) when compared to group 1

β Significantly different ($p < 0.001$) when compared to group 2
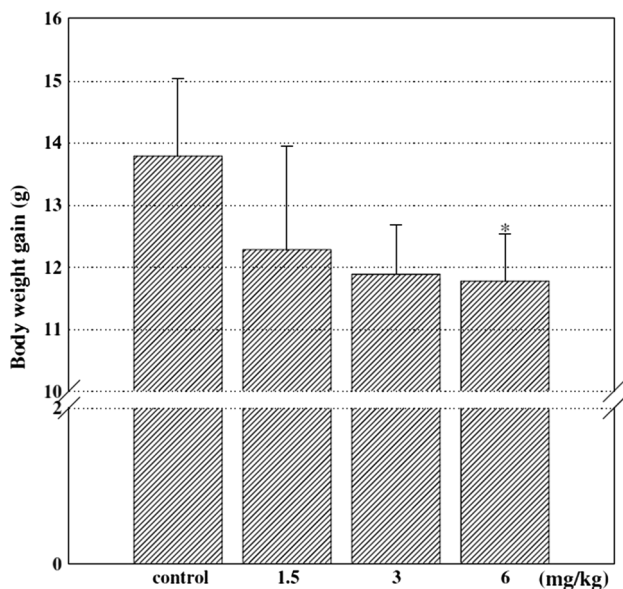

Fig. 1 A mean and SD bar chart with an asterisk indicating a significant effect; note the gap in the *vertical axis* to exaggerate the treatment differences [reproduced from Park et al. (2015)]
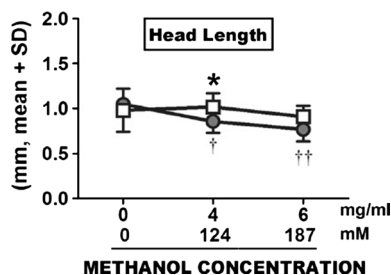

Fig. 2 A *line diagram* with means, SDs, and *symbols* indicating significant effects [reproduced from Miller-Pinsler et al. (2015)]
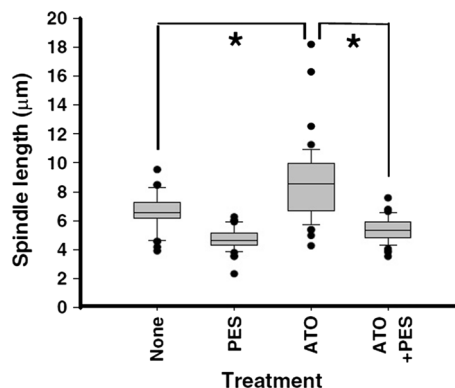

Fig. 3 A boxplot with outlier points and asterisks indicating significant effects [reproduced from Chen et al. (2014)]

meaningful if the data are normally distributed (and we know how often this is violated in toxicological experiments!), and second, they do not provide access to the individual data.

The first issue can be overcome with ordinary boxplots (Tukey 1977). They are surprisingly rarely used in toxicology although being frequently recommended [e.g., by Elmore and Peddada (2009) and Krzywinski and Altman (2014)]. A boxplot in its purest form displays five characteristic measures: median, lower and upper quartiles, minimum and maximum. Possible outliers (based on some definition for boundaries e.g., 1.5 × interquartile range) may be drawn as single points beyond the whiskers. An exemplary boxplot from a study on how the proteins HSP70 and PLK1 affect cells arrested in mitosis (Chen et al. 2014) is shown in Fig. 3. We can see there are a few clear outliers that would just go by the board in a simple mean ± SD chart.

The other issue is individual data. Raw values are of paramount importance in toxicology because sometimes the
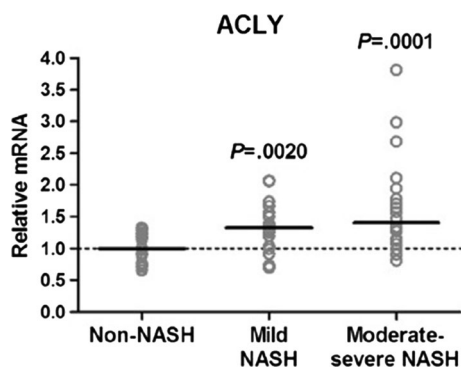
Fig. 4 A *dot plot* of raw values (without horizontal random noise) and their medians [reproduced from Bitter et al. (2014)]
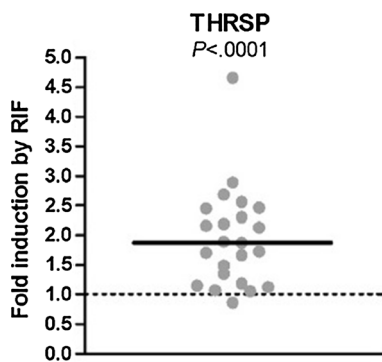

Fig. 5 A *dot plot* of raw values (with horizontal random noise) and their median [reproduced from Bitter et al. (2014)]

relevant information is just in a few extreme values and not necessarily in the group means. There are guidelines that explicitly recommend reporting *both* summary statistics *and* raw data, e.g., for the Ames assay (OECD 1997): "Individual plate counts, the mean number of revertant colonies per plate and the standard deviation should be presented for the test substance and positive and negative controls."

Despite the importance of raw data, graphics that actually show them are incredibly rare in toxicological publications. One positive counterexample can be found in a recent study on the pregnane X receptor's role in hepatic steatosis (Bitter et al. 2014); the authors make excessive use of dot plots, both with and without horizontal random noise ("jitter") to render similar values distinguishable (see Figs. 4, 5).

So we have accumulated evidence that even in fairly simple setups there is much room for improvement of data graphing practices. However, matters are often complicated further because many bioassays have not only a grouped data structure (negative control, several dose or treatment groups, and perhaps a positive control) but in addition some kind of hierarchical substructure, i.e., not all replications can be considered independent. Common examples are:

- technical replicates (e.g., 50 cells per gel and animal in a comet assay),
- sub-units (e.g., multiple pups from the same litter),
- spatial clusters (e.g., several animals caged together),
- temporal clusters (e.g., multiple runs of each animal in a Morris water maze on consecutive days),
- repeated measures (e.g., weekly measured body weights),
- paired organs (e.g., left and right kidney of the same animal),
- multiple donors (e.g., in an in vitro micronucleus assay),
- multi-hierarchical designs (e.g., cells within slides within samples within organs within animals within treatment groups in a comet assay).

In this paper, we spotlight issues critical for visualizing toxicological data that involve one of these or a similar substructure. We elucidate why the widespread bar charts are probably the poorest way of displaying complex grouped and clustered data. Instead we argue that a truly informative graph should incorporate the multi-level structure of the experiment, present raw values, and be based on boxplots.

Since Tukey's original work (1977), various ideas have been put forward how to enhance boxplot graphics. McGill et al. (1978) suggested drawing the boxes' widths proportional to the sample sizes; they also developed a version with the sides of the boxes being notched so that nonoverlapping notches indicate significant differences of medians. Reflections how density estimates could be included have led to "vaseplots" (Benjamini 1988), "violinplots" (Hintze and Nelson 1998), and "beanplots" (Kampstra 2008). These ideas are certainly appealing [as neatly illustrated in Spitzer et al. (2014)], but none of them is suitable for visualizing the hierarchical structure present in many toxicological datasets. To tackle this problem, we propose a composition of boxplots, mean ± SD bars, raw values, and display of other features like sample sizes, covariates, etc.

In Sect. 2 we illustrate with a simple artificial example why boxplots and especially jittered raw values are so much more informative than mean ± SD bar charts. Section 3 is dedicated to a demonstration of our preferred graphic with two real data examples of rats' body weights and a micronucleus assay. We discuss software solutions for drawing jittered boxplots in Sect. 4 and conclude the paper with a few general recommendations in Sect. 5. Executable R code is provided as supplementary material.

An artificial example

We can show the benefits of jittered boxplots using a pretty simple example of simulated data (see supplementary material for R code). Imagine we were to compare a

Table 2 Summary statistics of the artificial data example (*n* sample size, *SD* standard deviation, *IQR* interquartile range)

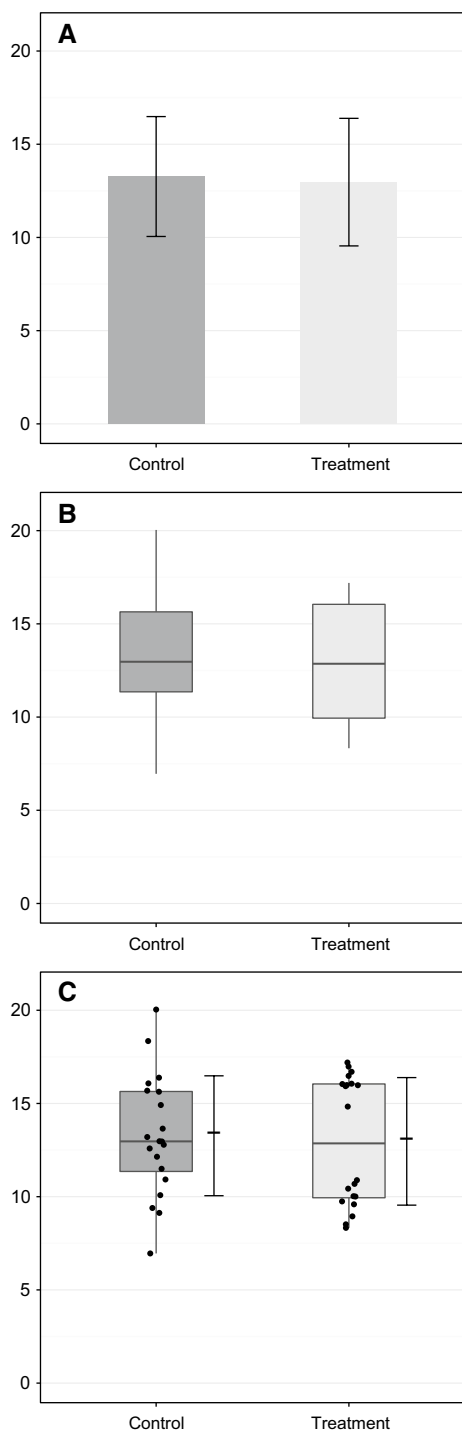|	*n*	Mean	SD	Median	IQR	Range
Control	20	13.27	3.22	12.97	4.29	13.08
Treatment	20	12.97	3.42	12.86	6.10	8.87

Fig. 6 Graphical representations of a set of fake data: **a** *barplot* of means ± SD bars, **b** boxplot, **c** boxplot with jittered raw values and mean ± SD bars

sample of measured values from an active treatment group with a control sample, and they have the summary statistics shown in Table 2.

Figure 6 shows three possible graphical representations of this dataset:

1. Barplots displaying means ± SDs are practically indistinguishable for the two groups.
2. Boxplots displaying medians, interquartile ranges, and total ranges (minimum and maximum) reveal that there is a difference between the two groups: Their quartiles and ranges are clearly dissimilar.
3. Jittered boxplots displaying the raw values (with a bit of horizontal noise added to avoid overplotting) in addition to the boxplot measures bring home the message that really matters: The control sample's distribution is more or less symmetric with most values accumulating near the center and few extremes, whereas the active treatment's values do not aggregate around the center but rather come in two separate clusters (in fact, the treatment sample was generated from a mixture of two normal distributions), and none of them is even close to the overall mean or median.

The biological reason for such an occurrence may be that half of the individuals show a notable reaction to the treatment and the other half do not. Detecting the distinct subgroups in the data is crucial for interpreting the results and also has consequences for the subsequent statistical analysis.

In a nutshell, we have seen that we may fail to spot essential characteristics of the data with simple bar charts. Ordinary boxplots do a better job, but the only way to get the full story is by looking at summary measures *and* raw values.

Two real-world examples

Body weight of rat pups

We illustrate our idea of a well-thought-out graphical representation for toxicological experiments with a set of data where the observations are hierarchically clustered by design. Pinheiro and Bates (2000) present body weights of 322 rat pups from 27 litters obtained in a study of two
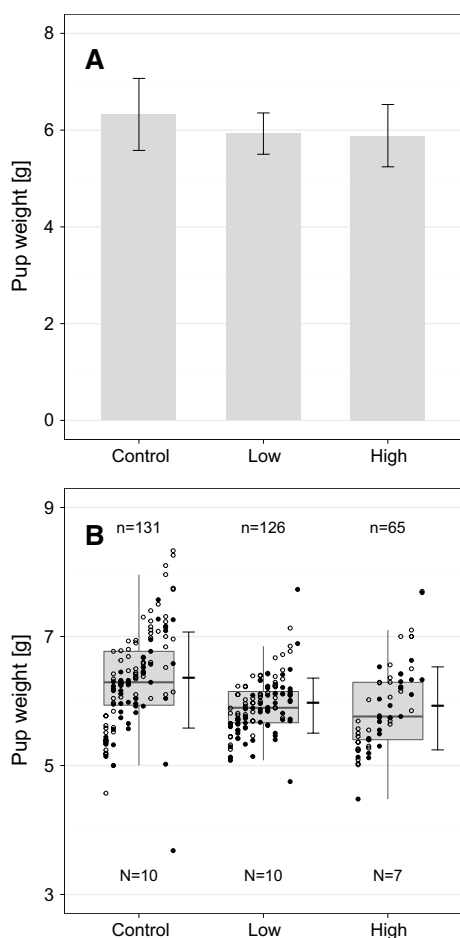
Fig. 7 Graphical representations of the rat pup data: **a** *barplot* of means ± SD bars, **b** boxplot with jittered raw values (strung vertically corresponding to litters, males as open and females as *closed circles*), numbers of pups (*n*) and litters (*N*), and mean ± SD bars

doses (low and high) of an experimental compound and a control; the crucial point with this dataset is that there are not 322 but only 27 independent experimental units, simply because the treatments were randomly assigned to 27 dams and not to their offspring. This clustering gives rise to the assumption that pups from the same litter are more alike (or in statistical terms: correlated) than pups from different litters. Moreover, the dataset is unbalanced in several respects: First, control and low dose were administered to ten dams each but high dose only to seven dams; second, numbers of pups per litter range between two and 18; and third, 171 pups are male and only 151 female. The data are stored as object `RatPupWeight` in the R package `nlme` (Pinheiro et al. 2015).

Panel A of Fig. 7 shows the common but unfavorable bar chart representation. Its informative content is limited to parametric measures of location and scale, i.e., mean and SD. However, there is a lot more behind the data that remains untold with this type of chart. Thus we strongly

advise against confining oneself to mean ± SD plots when faced with complex clustered data.

We strive for a graphical display that conveys as much useful information as possible but is still compact and intuitively understood. With these goals in mind, we propose supplementing standard boxplots with mean ± SD bars, raw values, sample size annotations, and further graphical elements to distinguish clusters and possible covariates. Such a plot is shown for the body weight data in panel B of Fig. 7. It contains:

- nonparametric summary measures of location (median) and scale (interquartile and total range excluding outliers),
- parametric summary measures of location (mean)[1] and scale (SD),
- raw data points (individual body weights) distinguished by a covariate (sex) via point shapes,
- cluster affiliations (which pups belong to the same litter) by points being strung together in vertical direction,
- numbers of randomized units (*N*, here: litters) and subunits (*n*, here: pups) per treatment group.

- Of course further graphical components are conceivable, e.g., we could add information on significant differences between groups (*p* values, asterisks, letters), discriminate cluster affiliations or covariate values using colors, and much more.

A graphical representation like this is highly insightful for many toxicological experiments that involve some kind of clustered structure. What matters is that in addition to the general trend (i.e., an average body weight reduction in comparison with control), our plot reveals a number of aspects that may be of interest:

1. Within-litter variability of body weights is particularly large in the control group.
2. Between-litter variability of body weights is fairly similar in all three treatment groups.
3. Outliers (in both directions) are mostly females.
4. Litter sizes vary considerably and so do the sex ratios within single litters.
5. The average litter size is roughly 13 with control and low dose but only about 9 in the high-dose group.
6. The pup body weight appears to be related to the litter size: The pups from the smallest litters (only two or three animals) are exceptionally heavy on average.

---

[1] The summary measures (e.g., mean and median) are unweighted with respect to litter sizes, which may be slightly distortive due to the data's substantial imbalance.
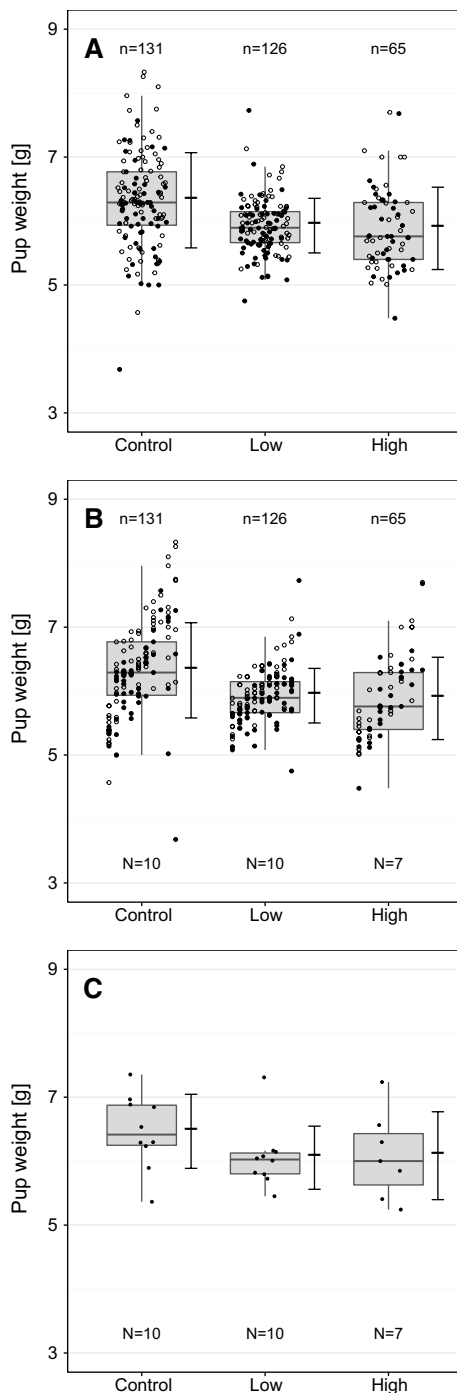
Fig. 8 Jittered boxplots illustrating three models for the rat pup data: **a** per-fetus analysis, **b** per-litter analysis, **c** per-mean analysis

All these details cannot be determined from a bar chart and neither from standard boxplots.

On top of that, our jittered boxplots prove very useful for visualizing and distinguishing between different models that may be fitted to the data. In principle, the pups' weights can be analyzed based on either of three statistical approaches:

1. Per-fetus analysis, i.e., the single pup is (incorrectly) considered as an independent experimental unit,
2. per-litter analysis, i.e., the single pup is treated as a sub-unit within the randomized unit litter,
3. per-mean analysis, i.e., using each litter's average pup weight.

The jittered boxplots in Fig. 8 illustrate the differences between the approaches. The per-fetus analysis (A) uses unduly large sample sizes because all observations are lumped together; as a consequence, tests for treatment differences will not keep the desired type I error level (Edler 2002). Averaging over the single pups and using the litter means (C) ignores that the litter sizes differ considerably and should thus be weighted relative to their contribution; moreover, it disregards the covariate sex. In fact, the per-litter analysis (B) is the only appropriate way to go (Hothorn 1991; ICH 1993), and the clustered structure of pups within litters—which is nicely visualized by the vertical strings of beads—is best reflected in a mixed-effects model with treatment as fixed and litter as random factor.

Micronucleus assay

Assays without a negative control group are unthinkable in toxicology, and statistical inference of treatment means versus control is typically attained through a many-to-one comparison procedure [e.g., Dunnett's test (1955) for normally distributed endpoints]. Including positive controls is less common but can be used either for demonstrating assay sensitivity, or to underpin the relevance of a change (that is significantly different from the negative control) by testing for noninferiority (Laster and Johnson 2003). Indications of significance obtained from such tests can be conveniently included in our jittered boxplots.

We consider data of a micronucleus assay involving a vehicle control, four doses (30, 50, 75, and 100 mg/kg) of hydroquinone, and a positive control (25 mg/kg cyclophosphamide). The original experiment was published by Adler and Kliesch (1990); the subset used here (only male mice) is available as dataset `Mutagenicity` in the R package `mratios` (Djira et al. 2012).

The outcome of the assay is a rate (number of micronuclei counted per 2000 cells after 24 h) and therefore a priori not normally distributed (not to mention that the variance evidently increases with the mean). In fact, the data are appropriately evaluated by fitting a Poisson generalized linear model (GLM) with logarithmic link function (McCullagh and Nelder 1989). Multiple tests of GLM parameters are conveniently performed with R packages such as `multcomp` (Hothorn et al. 2008) or `mcprofile` (Gerhard 2014) in the presence of small sample sizes (see supplementary material for R code). We are particularly interested
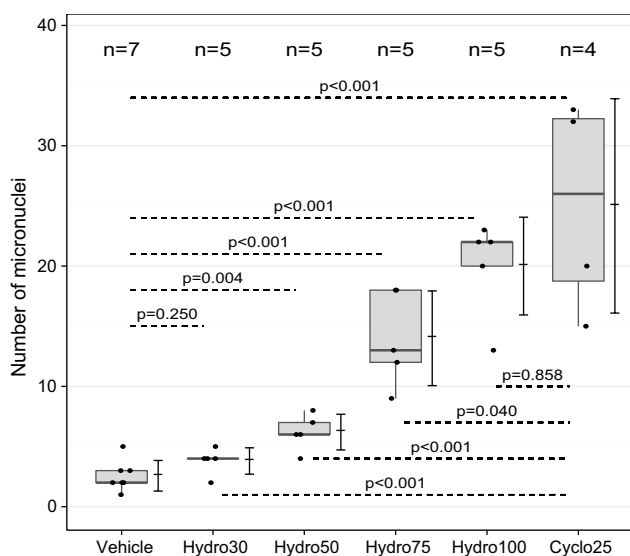
Fig. 9 Jittered boxplot of the micronucleus assay data: multiplicity-adjusted *p* values refer to comparisons against vehicle control (*above boxes*), against positive control (*below boxes*), and positive versus vehicle control (*top*)

in the following three comparisons, each of which is carried out at a type I error level of 5 %:

1. a one-sided two-sample test of positive versus negative control (test for assay sensitivity);
2. one-sided Dunnett-type tests of the hydroquinone doses versus negative control (test for superiority);
3. one-sided Dunnett-type tests of the hydroquinone doses versus positive control (test for noninferiority in relation to cyclophosphamide, with a noninferiority margin of 80 %).

Figure 9 shows the jittered boxplot with multiplicity-adjusted *p* values for all relevant comparisons. We see that all doses but the lowest (30 mg/kg) induce significantly more micronuclei than the vehicle control, whereas only the highest dose (100 mg/kg) is noninferior to the positive control at a margin of 80 %. The tiny *p* value for the comparison between positive and negative control indicates that the assay is adequately sensitive.

Implementation in R

Drawing jittered boxplots with additional elements is straightforward using the `ggplot2` graphics system (Wickham 2009) inside R (R Core Team 2015). Assuming that `ggplot2`'s high flexibility may overwhelm R novices, we provide a web application to facilitate getting started. It is available online at https://lancs.shinyapps.io/

ToxBox, and we showcase its use with a short tutorial in the supplementary material.

There is no doubt that similar graphs can be realized with different pieces of statistical software as well. However, the major advantages of R are that (a) it is open source and free to anybody, (b) it makes writing and extending one's own functions much easier than many commercial software packages, and (c) it allows to save graphics in various different formats and include them smoothly in multipanel figures.

Conclusion

We recommend jittered boxplots as an informative tool not only for exploratory data inspection but also for display of clustered and grouped datasets in toxicological publications. Software that creates such plots is readily available in R and can be easily modified to meet any data-specific requirements.

References

Adler ID, Kliesch U (1990) Comparison of single and multiple treatment regimens in the mouse bone marrow micronucleus assay for hydroquinone (HQ) and cyclophosphamide (CP). Mutat Res 234(3–4):115–123. doi:10.1016/0165-1161(90)90002-6

Benjamini Y (1988) Opening the box of a boxplot. Am Stat 42(4):257–262. doi:10.2307/2685133

Bitter A, Rümmele P, Klein K, Ba K, Rieger JK, Nüssler AK, Zanger UM, Trauner M, Schwab M, Burk O (2014) Pregnane X receptor activation and silencing promote steatosis of human hepatic cells by distinct lipogenic mechanisms. Arch Toxicol. doi:10.1007/s00204-014-1348-x

Chen YJ, Lai KC, Kuo HH, Chow LP, Yih LH, Lee TC (2014) HSP70 colocalizes with PLK1 at the centrosome and disturbs spindle dynamics in cells arrested in mitosis by arsenic trioxide. Arch Toxicol 88(9):1711–1723. doi:10.1007/s00204-014-1222-x

Djira GD, Hasler M, Gerhard D, Schaarschmidt F (2012) mratios: inferences for ratios of coefficients in the general linear model. R package version 1.3.17. http://cran.r-project.org/package=mratios

Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50(272):1096–1121. doi:10.2307/2281208

Edler L (2002) Statistical methods for toxicity detection and testing. In: Chyczewski L, Niklinski J, Pluygers E (eds) Endocrine disrupters and carcinogenic risk assessment. IOS Press, Amsterdam, pp 290–306

Elmore SA, Peddada SD (2009) Points to consider on the statistical analysis of rodent cancer bioassay data when incorporating historical control data. Toxicol Pathol 37(5):672–676. doi:10.1177/0192623309339606

Empl MT, Kammeyer P, Ulrich R, Joseph JF, Parr MK, Willenberg I, Schebb NH, Baumgärtner W, Röhrdanz E, Steffen C, Steinberg P (2014) The influence of chronic l-carnitine supplementation on the formation of preneoplastic and atherosclerotic lesions in the colon and aorta of male F344 rats. Arch Toxicol. doi:10.1007/s00204-014-1341-4

Gerhard D (2014) Simultaneous small sample inference for linear combinations of generalized linear model parameters. Commun Stat Simul Comput. doi:10.1080/03610918.2014.895836

Hintze JL, Nelson RD (1998) Violin plots: a box plot-density trace synergism. Am Stat 52(2):181–184. doi:10.1080/00031305.1998.10480559

Hothorn LA (ed) (1991) Statistical methods in toxicology. Springer, Berlin

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50(3):346–363. doi:10.1002/bimj.200810425

ICH (1993) Guideline S5, Part I: detection of toxicity to reproduction for medicinal products. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Safety/S5_R2/Step4/S5_R2__Guideline.pdf

Kampstra P (2008) beanplot: A boxplot alternative for visual comparison of distributions. J Stat Softw 28(code snippet 1):1–9. http://www.jstatsoft.org/v28/c01/paper

Krzywinski M, Altman N (2014) Visualizing samples with box plots. Nat Methods 11(2):119–120. doi:10.1038/nmeth.2813

Laster LL, Johnson MF (2003) Non-inferiority trials: the 'at least as good as' criterion. Stat Med 22(2):187–200. doi:10.1002/sim.1137

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton

McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. Am Stat 32(1):12–16. doi:10.2307/2683468

Miller-Pinsler L, Sharma A, Wells PG (2015) Enhanced NADPH oxidases and reactive oxygen species in the mechanism of methanol-initiated protein oxidation and embryopathies in vivo and in embryo culture. Arch Toxicol. doi:10.1007/s00204-015-1482-0

OECD (1997) Guideline for testing of chemicals, Test No. 471: bacterial reverse mutation test. doi:10.1787/9789264071247-en

Park EJ, Sim J, Kim Y, Han BS, Yoon C, Lee S, Cho MH, Lee BS, Kim JH (2015) A 13-week repeated-dose oral toxicity and bioaccumulation of aluminum oxide nanoparticles in mice. Arch Toxicol 89(3):371–379. doi:10.1007/s00204-014-1256-0

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2015) nlme: Linear and nonlinear mixed effects models. R package version 3.1-120. http://cran.r-project.org/package=nlme

Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS. Springer, New York

R Core Team (2015) R: a language and environment for statistical computing. http://www.r-project.org

Spitzer M, Wildenhain J, Rappsilber J, Tyers M (2014) BoxPlotR: a web tool for generation of box plots. Nat Methods 11(2):121–122. doi:10.1038/nmeth.2811

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading

Wickham H (2009) ggplot2: Elegant graphics for data analysis. Springer, New York. http://docs.ggplot2.org/current/