Effective graphs for data display: recommendations for authors

Kenneth Rice[1] and Thomas Lumley[2]

[1]Department of Biostatistics, University of Washington
[2]Department of Statistics, University of Auckland

September 11, 2015

Abstract

In research papers, graphs are a standard mode of describing data used in analyses. Graphs play crucial roles communicating the raw results of a study, supporting statistical analyses of that data, and helping to corroborate the results of data analyses. However, all too often graphical displays of data in submitted manuscripts are either inappropriate for the task at hand, or poorly-executed, thus requiring revision prior to publication. In this paper we aim to expedite this process by describing good graph construction for forms of data typically seen in *Heart*. Justification for many of these recommendations is drawn from the literature on visual perception; we also provide software instruction and examples, using various popular packages.

1 Introduction

Graphs provide an excellent tool for communicating data to readers. Research in visual perception has shown their superiority, over tables in communicating trends, and differences (1; 2). However, the choice of which graph to use, when communicating different data types and different aspects of a dataset, is often overlooked. In this paper we describe recommended use of graphs for manuscripts submitted to *Heart*. As well as recommending particular types of graph, and giving software examples for producing these graphs, we describe why they are good choices by which we aim to help authors communicate more successfully via the "language" of graphs. In this paper, we consider graphs for single variables, and pairs of variables.

2 Graphs for display of single samples

The simplest form of graph we consider shows observations of a single variable, for example biomarker values for multiple participants in a study. This will efficiently communicate the

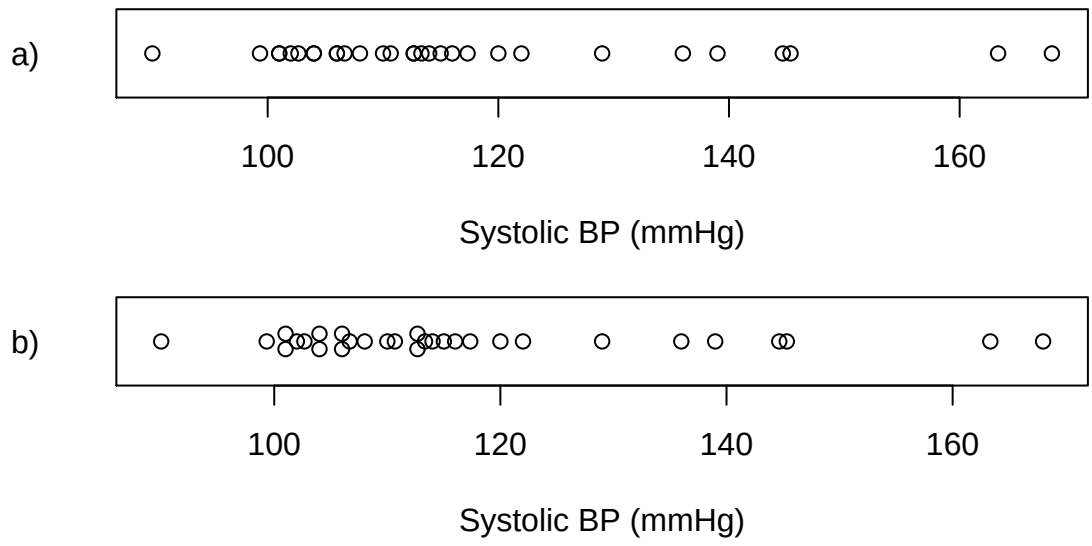

Figure 1: Dotchart (a) and stacked dotchart (b) illustrating systolic blood pressure measurements, for n=30 randomly-selected NHANES participants

range, and distribution of values, while reinforcing other summaries (e.g. mean, median, sample size) that can be communicated precisely in the text.

Different forms of graph are appropriate for quantitative and continuous variables, as described in the subsections below. All the examples are drawn from the publicly-available National Health and Nutrition Examination Survey (NHANES) dataset (3), from which we have taken random subsets of size n=30 (small), n=200 (medium) and n=1000 (large). Code for all the examples is provided at the authors' website `http://faculty.washington.edu/kenrice/heartgraphs/`.

2.1 Quantitative (continuous) variables

To graph the values of a quantitative variable with small samples sizes (i.e. n<30) we recommend use of *dotcharts*, also known as *stripcharts* or *dotplots*. An example is given in Figure 1a. A dotchart plots the observed values on a single axis. When these values are all different, with this sample size using empty circles as the plotting character enables readers to see the raw data underpinning nearby values, that overlap. Multiple filled circles would not permit this. When tied values are present, stacked dotcharts (see Figure 1b) clearly show the multiplicity of the tied values, by stacking them perpendicular to the axis.

For sample sizes above n=50, issues of overlap will typically make dotcharts impractical, either through near-overlapping points becoming obscured in a "cloud" of points (Figure 2a) or so much stacking occurring that plotting point become too small for easy reading (Figure 2b). However, for modest samples sizes (i.e. n between 50 and 200) stacked dotcharts with binned outcomes will often resolve these problems. An example is given in Figure 2c. In it, the raw data are first binned, into bins of width 1mmHg, before the stacked dotchart is constructed. Dotcharts are here preferred over the more well-known boxplot or box-and-

whisker plot (4, pp39-43) as dotcharts shows the actual data values, rather than just a few quantiles plus the most extreme observations, and so enable readers to infer more aspects of the data.

For still larger sample sizes, the problem of points being too small cannot be avoided. However, with such large sample sizes it becomes less likely that individual values affect interpretation to any practically-important extent, and so a direct representation of the range and distribution of the data may be sufficient. In this situation we recommend a *violinplot* (5), as shown in Figure 3a. This is closely related to the more traditional *histogram* (Figure 3b), with two important differences.

First, histograms represent frequency of data in a range of the x-axis by heights of bars, from zero up, which emphasizes the highest bars. Violinplots instead represent frequency by deviations around the x-axis, giving more emphasis to the overall shape, and not just the highest points. Second, while histograms use discrete heights to represent count (or proportion) in each bin, violinplots smooth these heights to produce a continuous curve, the deviation of which around the x-axis gives an estimate of the relative density of observations at any point on the x-axis.

Another advantage of all these plots, over more traditional summaries such as boxplots or histograms, is that they straightforwardly allow authors to superimpose graphical representations of summaries (such as the reference values for the variables, of the sample mean or median) that may be of interest, and related measures of uncertainty such as 95% confidence intervals. Figure 2c shows an example of a superimposed sample mean and corresponding confidence interval; Figure 3b similarly shows a superimposed sample mean and median. Examples of vertical dotcharts, stacked dotcharts and violinplots are given in Figure 6.

2.2 Categorical variables

For categorical variables, such as sex, authors may find that tables suffice for simple and concise recording of data, e.g. numbers male and female, proportions in each group and total sample size. However, if the information in the table is sufficiently important, communicating it graphically may be better choice.

Regardless of sample size, for graphical presentations of binary variables (i.e. categorical variables with only two levels, such as 1/0, or Yes/No, or Alive/Dead) we recommend either a *barchart* or *dotchart of proportion*, as in Figures 4a and b. For graphical presentations of categorical variables with three or more categories, we similarly recommend a *stacked barchart* or dotchart of proportions, as in Figures 4c and d. All of these graphs show the proportions in different groups (Systolic blood pressure above/below 140 mmHg, or the four categories of race-ethnicity) as positions on a common scale, relative to a fixed references point. This is known to be an approach that helps readers assess the relative sizes of these proportions (6).

The more familiar *pie chart* (Figures 5a and b) instead encodes the proportions as angles, or equivalently "slice" sizes; this is generally less effective for communicating the actual values of the proportions, although it may suffice for conveying relative sizes of different proportions, or the sizes of combinations of two or more of the proportions, i.e. two or more "slices" (7).

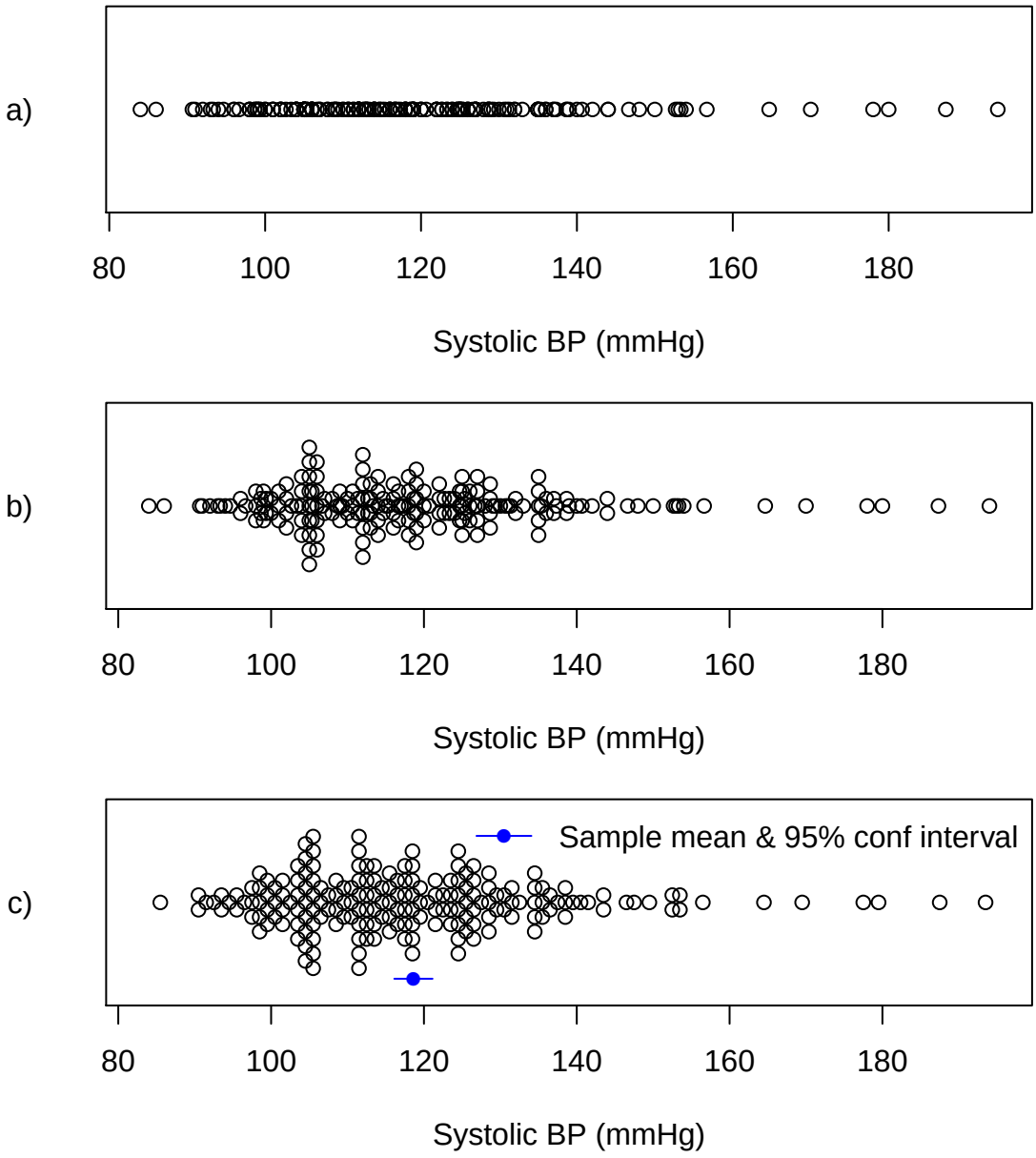

Figure 2: Dotchart (a) stacked dotchart (b) and stacked dotchart with binned outcomes (c) illustrating systolic blood pressure measurements, for n=200 randomly-selected NHANES participants. The sample mean and corresponding 95% confidence interval is superimposed on (c)

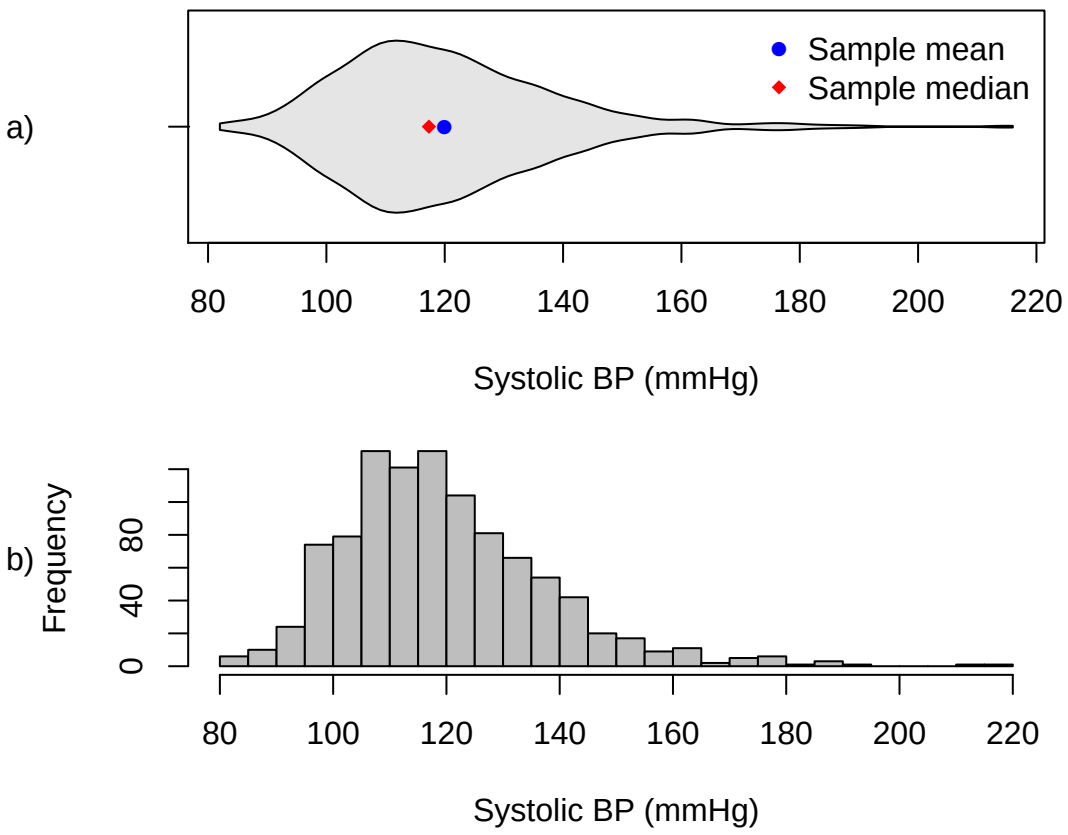

Figure 3: Violinplot (a) and histogram (b) illustrating systolic blood pressure measurements, for n=1000 randomly-selected NHANES participants. The sample mean and sample median are superimposed on (a)

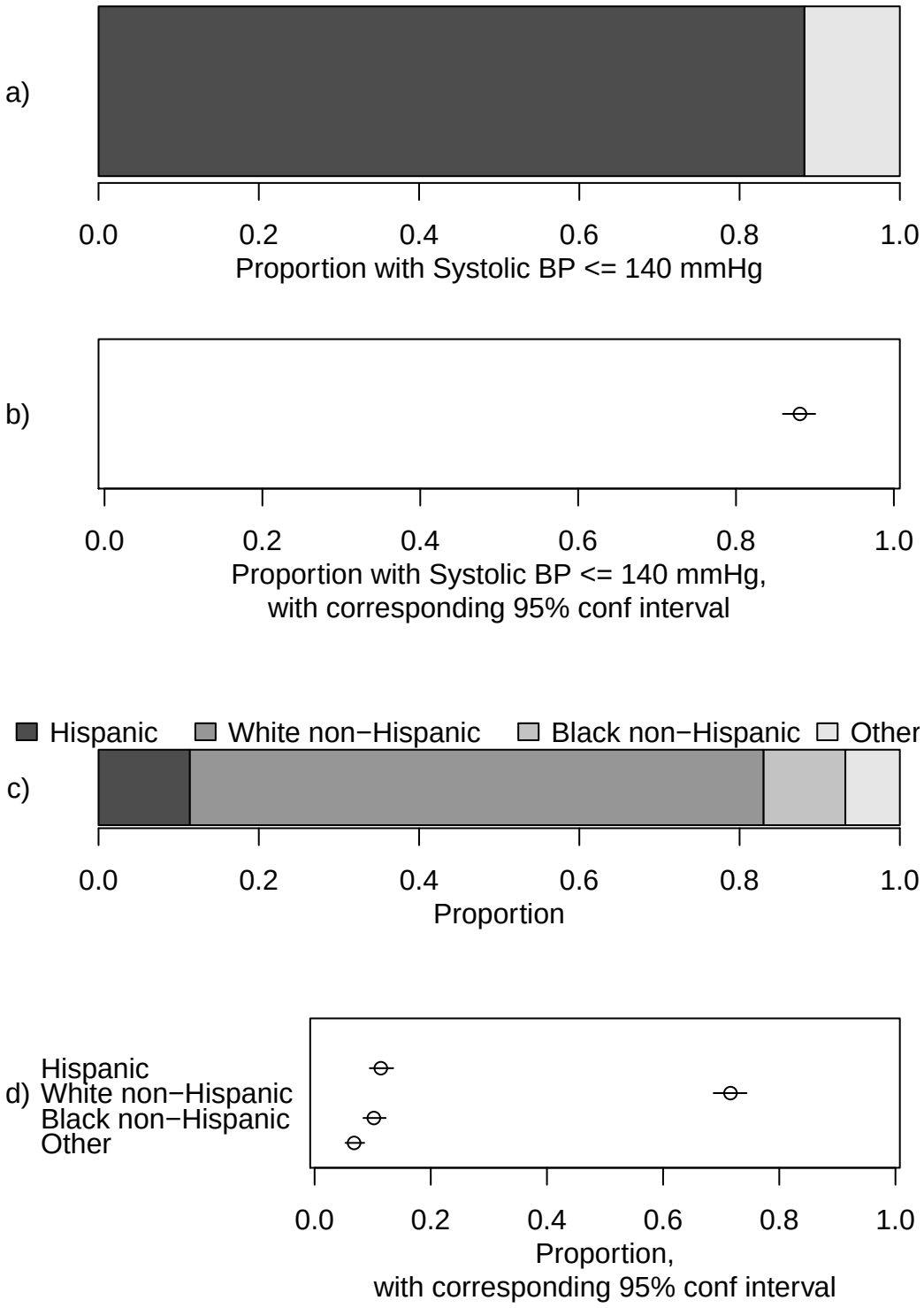

Figure 4: Barchart (a) and dotchart of proportion (b) illustrating proportion with systolic blood pressure measurements no more than 140mmHg, for n=1000 randomly-selected NHANES participants. Stacked barchart (c) and dotchart of proportions (d) for self-reported race-ethnicity, in the same participants. The dotcharts show 95% confidence intervals around each point estimate.

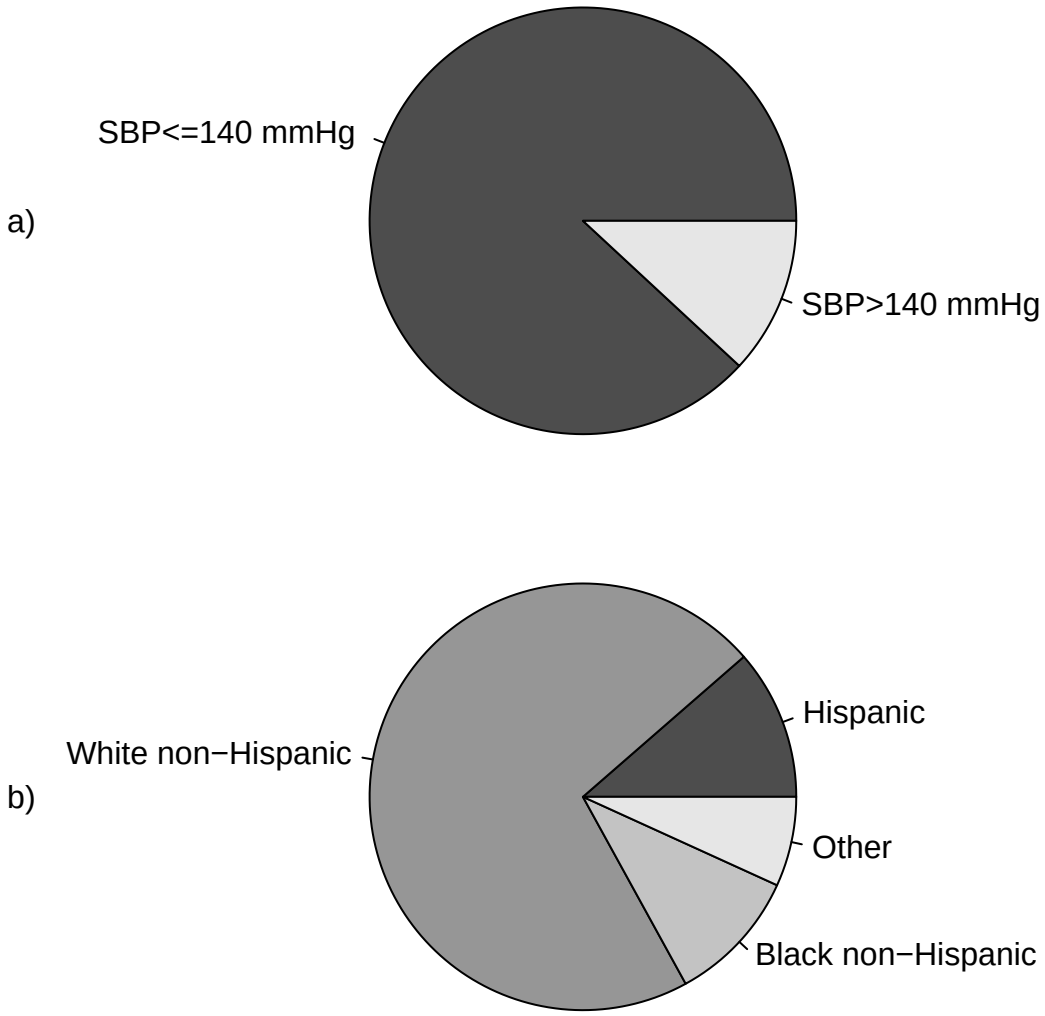

Figure 5: Piecharts of (a) systolic blood pressure measurements $\leq$ 140mmHg and (b) self-reported race-ethnicity, for n=1000 randomly-selected NHANES participants

3 Graphs comparing two variables

Many of the principles for plotting single variables, such as use of position on a common scale, and basic tools such as dotplots, violinplots and stacked barplots, can be adapted to communicate the relationship between two variables. The graphs typically show how the distribution of an "outcome" or "dependent" variable, depends on a "predictor" or "independent" variable or "covariate" (8, Ch.9). Below, we describe recommended approaches for pairs of variables, with separate approaches depending on whether the variables are continuous or categorical.

3.1 Continuous versus categorical

For plotting a continuous outcome versus a categorical covariate, we recommend *multiple* dotcharts, stacked dotcharts and violinplots. Examples are given in figures 6a, b and c. These illustrate distributional information for the outcome for each group defined by levels of the outcome. Comparison across groups is enabled by plotting their data on the same common scale, i.e. using the same y-axis. The sample size of each group is reflected naturally in the width of each stacked dotchart or violinplot.

As for single groups, sample means, medians and other information can be superimposed on the results from each graph. These have a direct connection to the statistical analyses that are typically used to compare the groups; the t-test compares the means in two groups, and Analysis of Variance compares the mean across multiple groups both procedures can be implemented directly given sample means for each group with corresponding 95% confidence intervals, and knowledge of the sample size in each group. We do not recommend that "stars" indicating levels of statistical significance (e.g. * for p <0.05, ** for p <0.01, etc) are superimposed. These distract from direct comparison of the group means which will usually be of more relevance to the underlying science than whether a somewhat arbitrary significance threshold is achieved. The use of stars is particularly problematic when many groups are being compared.

3.2 Continuous versus continuous

For plotting a continuous outcome against another continuous covariate, and for small or modest samples sizes, we recommend a *scatterplot*. Examples are given in Figures 7a and b. As with the other recommended graphs, these code the information in the two variables as positions on common scales the x-axis and y-axis. As with dotplots, we recommend that points are plotted using empty circle symbols, rather than filled circles, as the empty circles better enable readers to distinguish individuals points when there are multiple points nearby. When observations are not unique (i.e. when observations have identical X and Y values) then usually adding a small amount of "jitter" (9) to the points will be sufficient to make individual points distinguishable, while retaining the overall pattern of the original data. Where available, overplotting using transparent colors may also be effective (10, Ch.5).

As with earlier plots, relevant statistical summaries can be interpreted in terms of the graph, and superimposed on it. For example, linear regression of outcome on covariate (as in Figure 7a can be interpreted as finding the straight line through the graph that minimizes

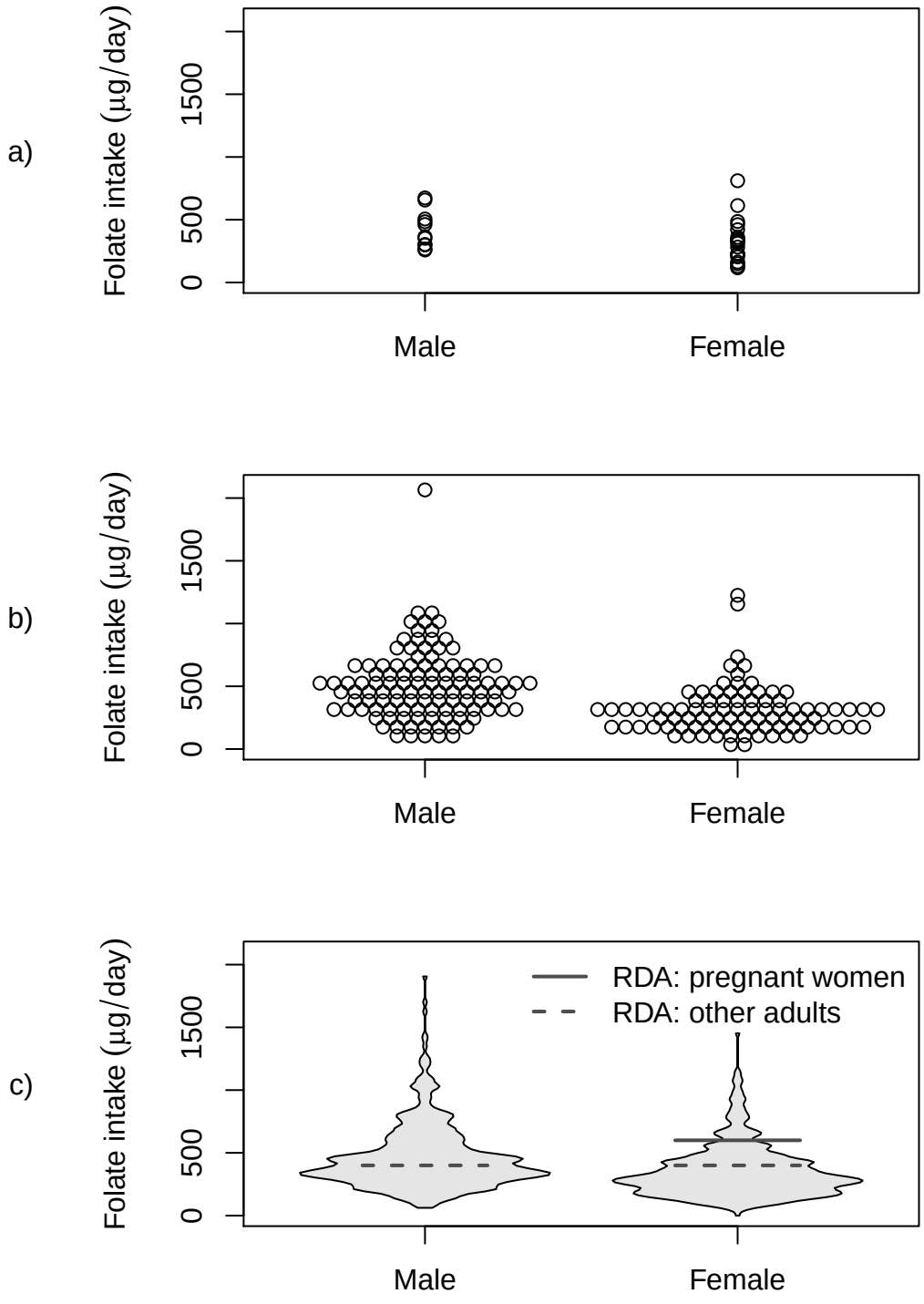

Figure 6: Multiple dotchart (a) multiple binned stacked dotchart (b) and multiple violinplot (c) illustrating folate intake, respectively for n=30, 200 and 1000 randomly-selected NHANES participants. Recommended Daily Allowances (RDAs) for pregnant women and other adults are superimposed on the violinplot

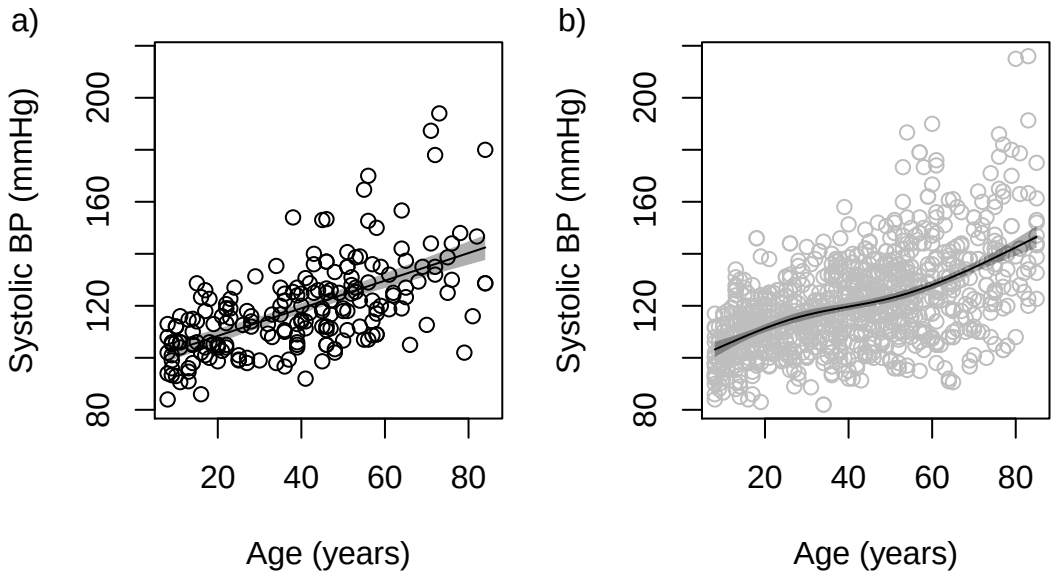

Figure 7: Scatterplot of systolic blood pressure versus age, for (a) n=200 and (b) n=1000 randomly-selected NHANES participants. For n=200, the best-fitting linear slope is superimposed, surrounded at each point by a 95% confidence interval. For n=1000 a flexible spline representation is instead used, allowing a more subtle signal to be observed.

the mean squared vertical distance from that line to the observed data points, i.e. finding the "best-fitting line". This line can be superimposed on the graph of the corresponding data, together with a region depicting 95% confidence intervals for its fitted value, i.e. the fitted mean outcome, at each value on the x-axis. For small-to-moderate sample sizes, this analysis provides a pragmatic summary of increasing or decreasing "trend" relationships, although this straight-line summary may be unhelpful if the true relationship is e.g. strongly U-shaped (11). For large sample sizes, the pragmatic choice to fit just a straight line is less well-justified; a more flexible spline representation of the covariate (12) can be fitted instead, that can capture non-linear relationships between the mean outcome and the covariate. Pointwise 95% confidence intervals around this fitted curve can be added, as in Figure 7b. A alternative to the use of splines "smooths" the y-axis values (13). This is in spirit close to regression with splines (14, Ch.2), but calculation of corresponding confidence intervals is more involved. (The confidence intervals around the fitted regression lines here use standard methods, that assume there is identical variance of observations around the mean at each point on the x-axis (8, Ch.9).)

3.3 Categorical versus categorical

As with single categorical variables, numeric tables (i.e. contingency tables) may suffice for describing the relationship between two categorical variables. Where graphics are needed, for plotting a categorical outcome against a categorical covariate for any sample size, we recommend a *mosaicplot*, also known as a *multiple stacked barplot*, or alternatively a multiple dotchart of proportions. Examples are shown in Figures 8a and b. In these plots, the

primary comparison is of relative proportions of outcome categories, within levels of covariate category these proportions are indicated by the heights of the elements in each stacked bar. However, as the widths of each stacked bar is proportional to the number of observations in its corresponding covariate category, the areas of the elements also indicate absolute counts, across covariate categories.

In a mosaicplot the covariate and outcome play different roles, determining the widths of the bars and the heights within each bar respectively. For situations where the graph is to illustrate the agreement or disagreement between two measures, this distinction is not appropriate, and we instead recommend a *fluctuation diagram*, as shown in Figure 9. Here, the area of the square corresponding to each combination of covariate values is proportional to the count of that observation. Patterns of asymmetry around the 45° degree line indicate disagreement between the covariates. For further discussion see Hoffman (15, Ch.2) and Unwin et al (16).

Unfortunately neither the stripchart nor fluctuation diagram can be modified easily to add confidence intervals, point estimates, or other summaries of the raw data. We recommend that these are reported in text, or through separate graphs illustrating just the summaries and not the raw data.

3.4 Categorical versus continuous

For plotting a categorical outcome against a continuous covariate, we again consider different options when the outcome has two levels (i.e. is binary), versus having more levels. For a binary outcome, a standard scatterplot of outcome on covariate will show the data, although stacking or other adjustments may be needed to reduce overplotting of nearby or tied points. This may be a minor concern, however, when interest lies in the proportion of outcomes in either category, in different regions of the x-axis. This proportion can be shown on the scatterplot by superimposing the fitted mean from a logistic regression, or similar analysis. An example is given in Figure 10a. With larger datasets, using a spline representation of the covariate in regression analyses again adds flexibility, as shown in Figure 10b, and this approach is again similar to adding a smoother to the scatterplot. As with the fitted summaries in Figure 7, regions indicating pointwise 95% confidence intervals can be added around these lines.

With more than two categories, representing the outcome data on a single axis is not possible, so there is no straightforward way to plot all the data points individually. Dichotomizing the outcome into "one category versus all the others", one can plot multiple regression lines (or smoothers) indicating proportion in each outcome group, at different covariate levels; an example is shown in Figure 11a. With careful annotation, this plot can be effective with up to 5 or 6 colored lines, though we recommend no more than 4 if the graph is produced in black and white. Distinguishing between lines may also be difficult if, due to the data, the lines are very close over large ranges of the x-axis. With ordered categories, this problem can be reduced by instead splitting the outcome into "upper categories versus lower", for multiple cutpoints; an example is shown in Figure 11b. To aid interpretation, it may help to note that this plot is equivalent to splitting the data up by ranges of the covariate, giving a stacked barplot of outcomes for each range, and smoothing the resulting series of barplots. Omitting the stacks and showing just the smoother is visually cleaner,

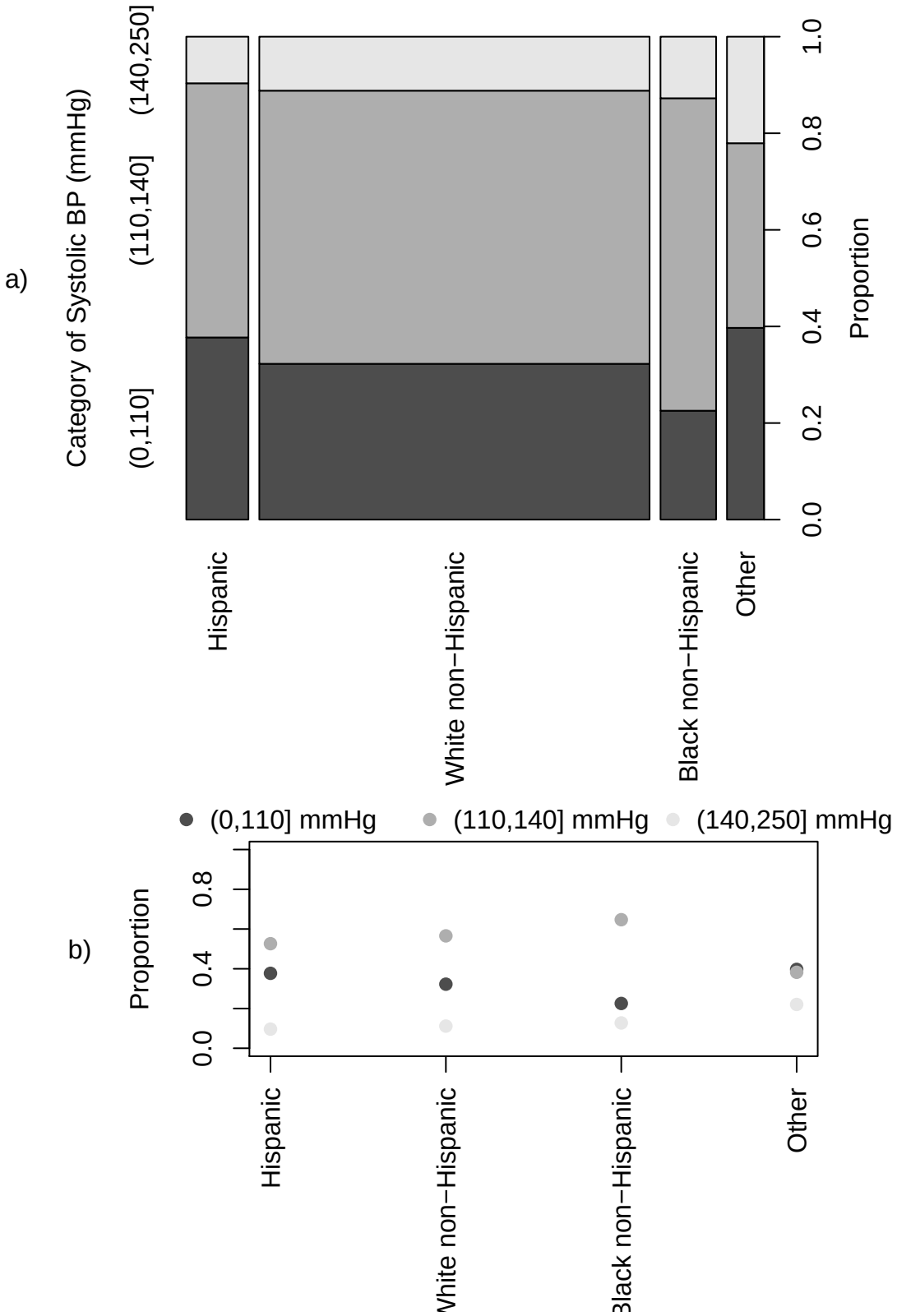

Figure 8: Mosaicplot (a) and multiple dotchart of proportions (b) of categorized systolic blood pressure versus self-reported race-ethnicity, for n=1000 randomly-selected NHANES participants.

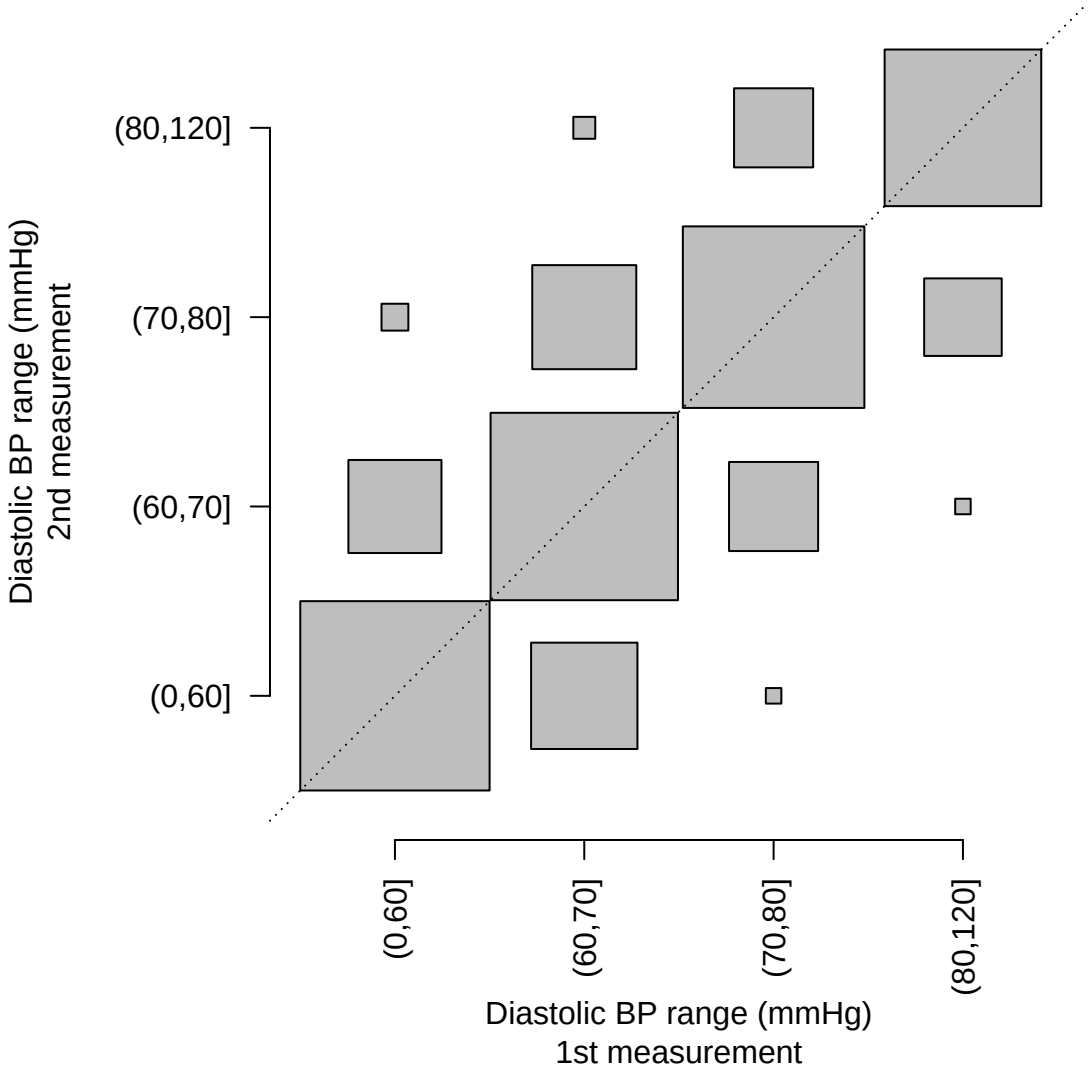

Figure 9: Fluctuation diagram, comparing categorized first and second diastolic blood pressure measurements. Symmetry around the 45◦ line indicates no strong evidence of systematic differences between measurements, based on this data.

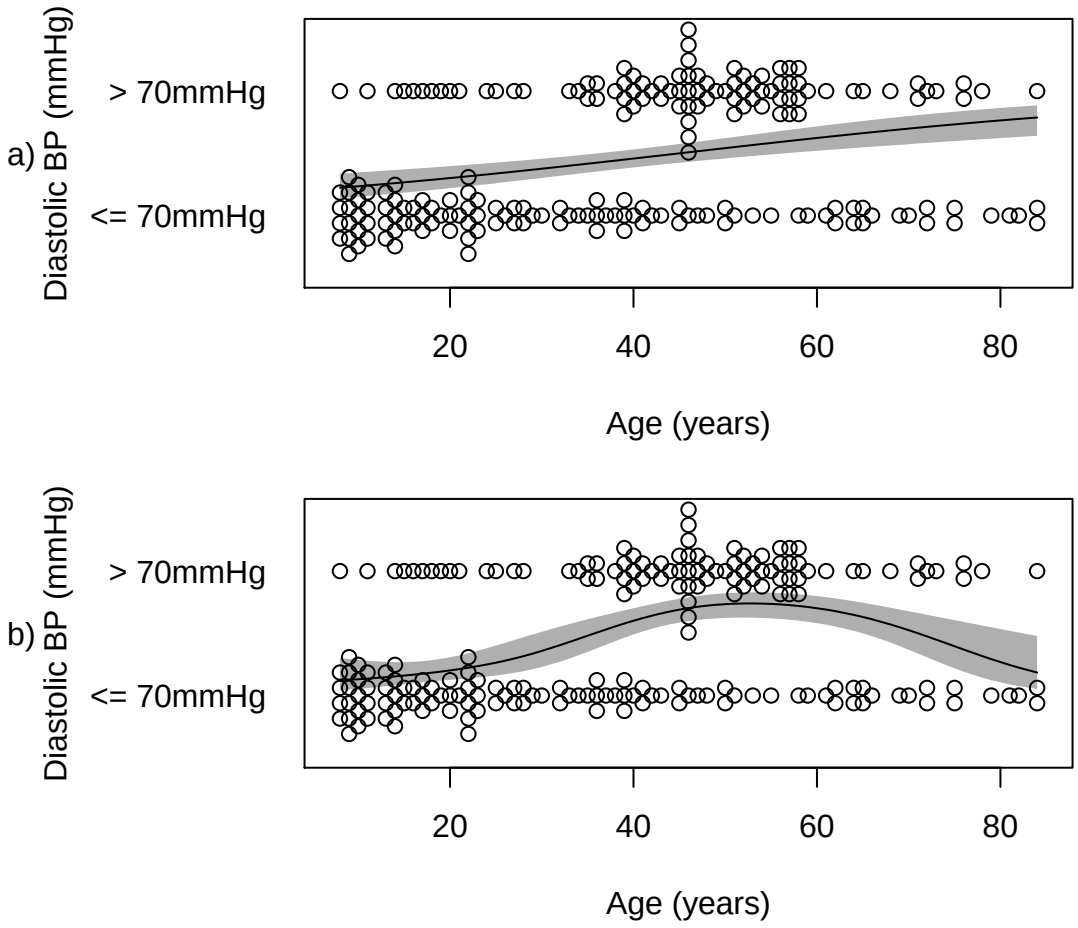

Figure 10: Simple logistic regression (a) and spline-based logistic regression (b) fit, showing the relationship between dichotomized diastolic blood pressure and age, for n=200 randomly-selected NHANES participants. The shaded regions indicate 95% confidence interval around the fitted values, for each point on the x-axis.

and can facilitate adding measures of uncertainty.

4 Graphs illustrating more than two variables

To compare relationships between outcome and covariate at different levels of a third, stratifying, variable that takes only a few values, points and lines can be color-coded by strata. To avoid printing in more than one color, different plotting symbols (circles, squares, etc) and line types (solid, dashed, dotted) may also be used. Effective color schemes for points can be found using the ColorBrewer software (17), for both ordered and unordered stratifying variables. Guidance on which plotting symbols best distinguish multiple groups is given by Krzywinski and Wong (18). For different line types, few choices are typically available, with solid and dashed lines used predominantly. However, even with judicious choices of colors and symbols/line types, distinctions between more than 4–5 strata are likely to be difficult for the reader, and so the number of strata should be considered carefully by the authors, and kept to a minimum.

When showing how the relationship between continuous outcome and covariate depends on a third continuous variable, plotting in three dimensions becomes appealing, but this is naïve. As journal articles are necessarily printed on 2-dimensional paper, without animation or stereoscopic techniques, the "depth" of a 3-dimensional plot cannot be communicated accurately. Techniques for providing visual clues to depth are available (19) but cannot cope with comparisons between multiple points, needed in a display of data. Instead, it may be effective to stratify the data according to the third variable, and plot "small multiples" (20, pg 67) of a scatterplot, with one subplot per strata. An example is given in Figure 12. The idea of stratification can also be used for violinplots, barcharts and other graphs, enabling an outcome/covariate relationship to be shown conditioned on values of one covariate, or two covariates if a grid of small multiples is provided.

5 Software and fine-tuning

The graphs in this paper have all been produced in R (21), a free and widely-used statistical computing environment. Data and code to reproduce them is available on the authors' website, `http://faculty.washington.edu/kenrice/heartgraphs/`. Several examples of the use of Excel to produce graphs of the same style, using the same data, are also available there. Regardless of which software is used in their production, the final production-quality version of a graph will likely require that several details are considered.

Axes labels should be provided, with text that is legible at the size at which the graph is to be produced. Similarly, legends should be provided denoting plotting points. Lines may be labeled directly, or described via a legend. We do not recommend that legend information be provided in the caption below the figure; switching between the figure and the caption text distracts the reader. In *Heart*'s style, graphs are not supplied with titles; all text should appear in the caption. For some variables it is common to use the data only after transformation (e.g. log-transformation of gene expression). If the transformed units are familiar to the audience, these should be used throughout; if not then a transformed axis

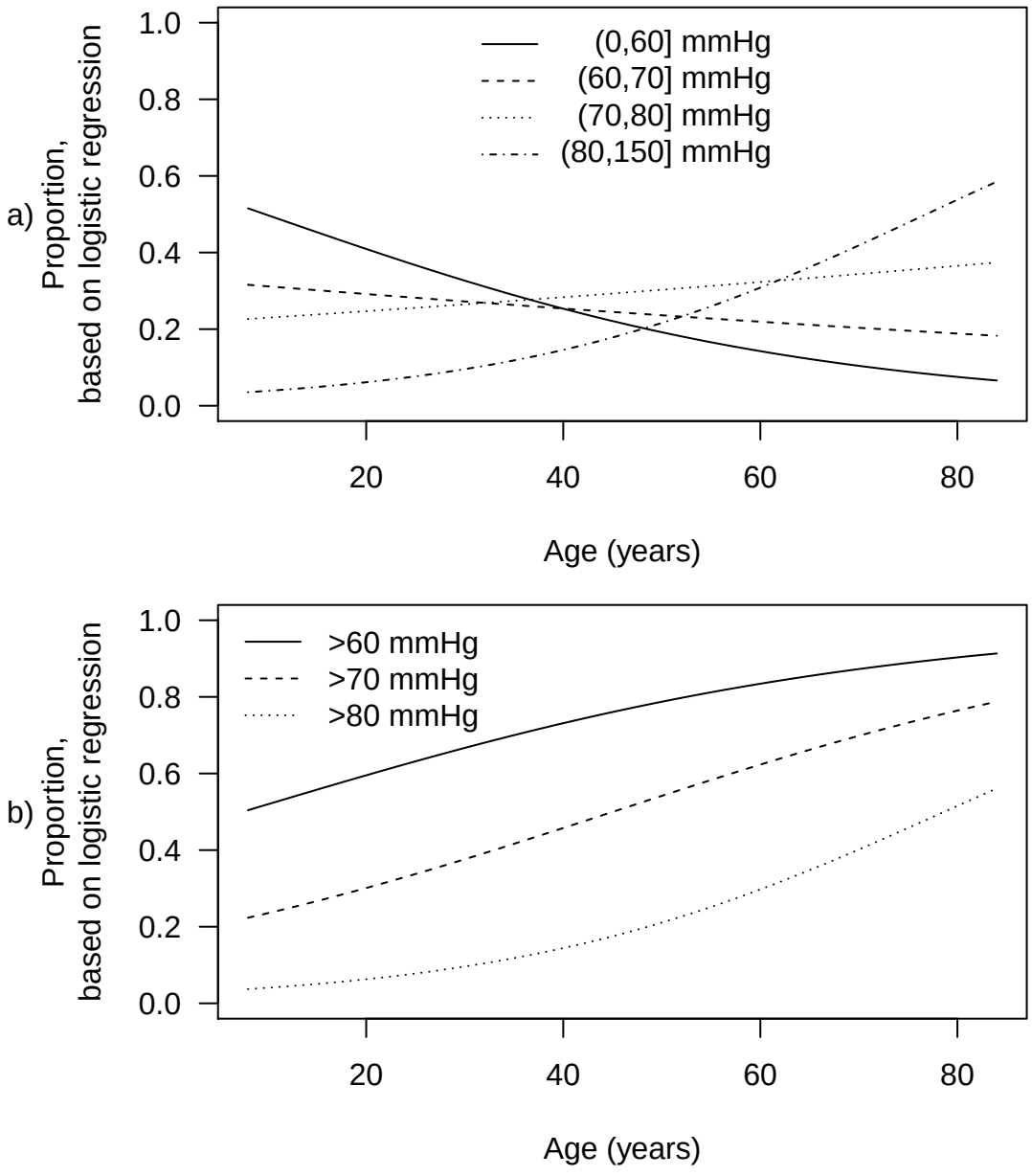

Figure 11: Plots of fitted proportions, by age, of observations (a) in each unique category of diastolic blood pressure and (b) exceeding each specified threshold. In all cases the fitted proportions are from logistic regression of binary outcome on age, using data from n=1000 randomly-selected NHANES participants.

Figure 12: Small multiples of systolic blood pressure versus BMI, with sex-specific smoothers, conditioned on three age ranges. The plot uses data from n=1000 randomly-selected NHANES participants; in this data the difference in blood pressure-BMI relationship by sex appears different in older participants.

should be used (e.g. a logarithmic axis, for gene expression levels). If neither transformed nor untransformed are ubiquitous, both may be indicated, by using both left and right vertical axes, or upper and lower horizontal axes.

A key consideration in fine-tuning a graph is considering the size at which it will be reproduced; much statistical software will automatically re-size axis labels and legend text depending on whether the graph will appear, for example, in a single column within the text, or as a whole page. However, to take advantage of this, authors must specify their figure's size. In the printed journal, a full page-width figure is 7.1 inches wide. On a 2-column page, a full column-width figure is 3.4 inches wide, while on a 3-column page it is 2.25 inches wide. In supplementary material typical graphs will be full page-width.

6 Summary

Simple visualization techniques can be used to enhance communication of scientific data; the choice of how to plot data is not just an issue of taste, house style, or simply accepting the literature default. For effective communication of the distribution of a variable, or showing relationships between pairs of variables, we have recommended plots that are straightforward, and can be produced in standard software. In manuscripts, poor graphs, as well as poor writing, hinder communication of findings, yet compared to running studies the time and resources required to improve them is often minimal. We therefore hope that the recommendations in this paper provide authors with a useful way to expedite publication of their scientific findings.

References

[1] Meyer J, Shamo MK, Gopher D. Information structure and the relative efficacy of tables and graphs. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1999;41(4):570–587.

[2] Gelman A, Pasarica C, Dodhia R. Let's practice what we preach: turning tables into graphs. The American Statistician. 2002;56(2):121–130.

[3] Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey Data; 2015. U.S. Department of Health and Human Services. Available from: `http://www.cdc.gov/nchs/nhanes.htm`.

[4] Tukey J. Exploratory Data Analysis. Reading: Addison-Wesley; 1977.

[5] Hintze JL, Nelson RD. Violin plots: a box plot-density trace synergism. The American Statistician. 1998;52(2):181–184.

[6] Cleveland WS, McGill R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association. 1984;79(387):531–554.

[7] Spence I. No humble pie: The origins and usage of a statistical chart. Journal of Educational and Behavioral Statistics. 2005;30(4):353–368.

[8] Van Belle G, Fisher LD, Heagerty PJ, Lumley T. Biostatistics: a methodology for the health sciences. vol. 519. John Wiley & Sons; 2004.

[9] Chambers JM, Cleveland WS, Kleiner B, Tukey PA. Graphical methods for data analysis. Wadsworth, Belmont, CA: Wadsworth International Group; 1984.

[10] Few S. Now You See it: Simple Visualization Techniques for Quantitative Analysis. Burlingame, CA: Analytics Press; 2009.

[11] Freedman DA. On the so-called "Huber sandwich estimator" and "robust standard errors". The American Statistician. 2006;60(4).

[12] Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. 1995;p. 356–365.

[13] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 1979;74(368):829–836.

[14] Hastie TJ, Tibshirani RJ. Generalized additive models. vol. 43. Boca Raton, FL: CRC Press; 1990.

[15] Hofmann H. Graphical Tools for the Exploration of Multivariate Categorical Data. Norderstedt: BoD–Books on Demand; 2001.

[16] Unwin A, Hofmann H, Cook D. Let Graphics Tell the Story – Datasets in R. R Journal. 2013;5(1).

[17] Brewer C, Harrower M, Sheesley B, Woodruff A, Heyman D. ColorBrewer 2.0; 2015. Accessed July. Available from: `http://www.colorbrewer2.org/`.

[18] Krzywinski M, Wong B. Points of view: Plotting symbols. Nature methods. 2013;10(6):451–451.

[19] Ellis SR. Pictorial communication: pictures and the synthetic universe. Leonardo. 1990;p. 81–86.

[20] Tufte ER. Envisioning Information. No. v. 914 in Envisioning Information. Cheshire, CT: Graphics Press; 1990.

[21] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org/`.