Presenting parasitological data: the good, the bad and the error bar

SOPHIE G. ZALOUMIS[1]*, FREYA J. I. FOWKES[1,2,3], ALYSHA DE LIVERA[1] *and* JULIE A. SIMPSON[1]

[1] *Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Melbourne, Victoria 3010, Australia*
[2] *Macfarlane Burnet Institute of Medical Research, 85 Commercial Road, Melbourne, Victoria 3004, Australia*
[3] *Department of Epidemiology and Preventive Medicine and Department of Infectious Diseases, Monash University, 99 Commercial Road, Melbourne, Victoria 3004, Australia*

(*Received 14 November 2014; revised 21 May 2015; accepted 26 May 2015; first published online 29 June 2015*)

SUMMARY

Visual displays of data in the parasitology literature are often presented in a way which is not very informative regarding the distribution of the data. An example being simple barcharts with half an error bar on top to display the distribution of parasitaemia and biomarkers of host immunity. Such displays obfuscate the shape of the data distribution through displaying too few statistical measures to explain the spread of all the data and selecting statistical measures which are influenced by skewness and outliers. We describe more informative, yet simple, visual representations of the data distribution commonly used in statistics and provide guidance with regards to the display of estimates of population parameters (e.g. population mean) and measures of precision (e.g. 95% confidence interval) for statistical inference. In this article we focus on visual displays for numerical data and demonstrate such displays using an example dataset consisting of total IgG titres in response to three *Plasmodium* blood antigens measured in pregnant women and parasitaemia measurements from the same study. This tutorial aims to highlight the importance of displaying the data distribution appropriately and the role such displays have in selecting statistics to summarize its distribution and perform statistical inference.

Key words: Parasitological data, display, graph, statistical inference, descriptive analysis.

INTRODUCTION

In the parasitology literature, we have noticed as statistical reviewers, that frequently visual displays of the data are presented which are not very informative regarding the distribution of the data, and that better visual displays should have been selected by the authors. Examples of 'poor' quality displays, selected from a search of papers published in 2014 in *Parasitology*, are provided (de-identified) in Fig. 1 and represent common visualizations seen in the biological, veterinarian and clinical literature. The features of these examples that make them 'poor' quality displays are, the raw data are hidden by too few summary statistic(s) (group mean and either top half of error bar or no error bar). In the statistics literature there are many textbooks and articles that advise against using such plots and provide simple alternative displays (Wainer, 1984; Huff, 1993; Campbell, 2009; Freeman *et al*. 2009; Weissgerber *et al*. 2015). The purpose of this article is to summarize the advice on how to select appropriate displays and the simple alternatives proposed in these texts for researchers in parasitology. Visual displays for numerical data (measurements such as parasitaemia and laboratory antibody data) will be the focus of this article, as from our experience, researchers have the most difficulty in selecting appropriate displays for such data due to the variety of plots available.

Statistical concepts and terminology

The purpose of most data analyses is either descriptive or to perform statistical inference. Before we can define these types of analyses and explain how displaying data is crucial to both, some statistical concepts and terminology needs to be introduced. The focus of most studies is to identify covariates (also termed exposures or predictors) that explain the variability in an outcome; for example, does parasitaemia (covariate) explain the variability in haemoglobin levels (outcome) from African children (population of interest). Equally the 'population of interest' may be fish or mammals, or mice or cells in an *in vitro* experiment with relevant covariates and outcomes measured, but for the purpose of this illustrative piece we will use examples relating to human populations. It is rarely feasible to obtain data from the entire population (of humans, fish, mammals, mice or cells) so instead we collect data from a subset of the population, called a *sample*.

* Corresponding author. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Melbourne, Victoria 3010, Australia. E-mail: sophiez@unimelb.edu.au

Parasitology (2015), **142**, 1351–1363. © Cambridge University Press 2015
doi:10.1017/S0031182015000748

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

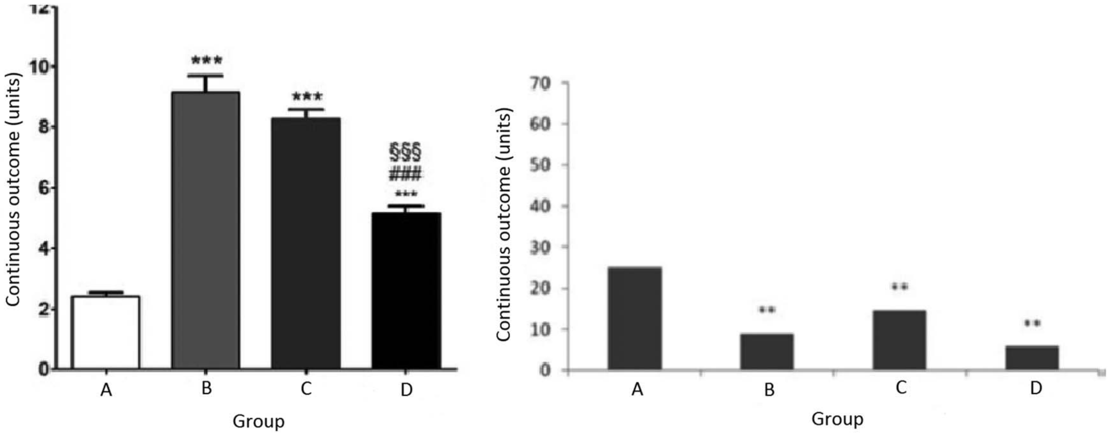

Fig. 1. Examples of poor displays found in a cursory search of this journal. In the left panel the bar indicates the group mean and the error bar is the mean + s.d. ***$P < 0{\cdot}001$ for comparison of means of group B, C and D *vs* group A; ###$P < 0{\cdot}001$ for comparison of means of group D *vs* group B; and §§§$P < 0{\cdot}001$ for comparison of means of group D *vs* group C. In the right panel the bar is the group mean and **$P < 0{\cdot}001$ for comparison of means of group B, C and D *vs* group A. The features of these examples that make them poor quality displays are, the raw data are hidden by too few summary statistic(s). For example in the left panel only the group mean and top half of the error is displayed, while in the right panel only the group means are displayed.

The distribution of a variable measured in a sample is called the *empirical distribution*. *Statistics* are functions of the data values observed in a sample, and they can be viewed in two ways, as: (1) summarizing an empirical distribution of a variable (i.e. that of your sample) or (2) an estimate of a population parameter (more on this when statistical inference is discussed below). The empirical distribution of a variable is commonly summarized using two types of statistics: one that measures location (centre or peak around which the bulk of the data lie) and others that measure spread (how distant observations are from the location or centre). In this tutorial the *sample mean* and *median* will be used to measure the location of an outcome's empirical distribution and the *sample* s.d., *inter-quartile range* and *range* to measure its spread (these statistics are defined in Table 1). *Standard errors* (s.e.) and *95% confidence intervals* (CIs) are also functions of the outcome values observed in a sample, but they *do not* describe the empirical distribution of the outcome, they are used to draw conclusions about the *population* from which the outcome was sampled (these terms are defined in Table 1 and discussed in more detail below).

Descriptive analyses focus on describing the empirical distribution of a variable, by first plotting the distribution and then calculating suitable statistics to summarize its location and spread. *Statistical inference* is much more complex and uses the empirical distribution of the variable to draw conclusions about the distribution in the population. *First*, a probability distribution is selected for the population distribution of say the outcome variable. Since only data from a sample of individuals (mice, cells or parasite isolates) have been collected, the population distribution is selected to resemble the empirical distribution of the sample. The normal distribution is an example of a widely used probability distribution. *Second*, the *population parameters* of the chosen probability distribution are estimated using statistics. For example, the parameters of the normal distribution are the population mean and population s.d. which are estimated by the sample mean and sample s.d. *Lastly*, measures of how precisely a population parameter is estimated by a statistic, such as s.e. and 95% CIs are derived. Even though these statistics are derived from the single sample of data we have collected, the measures represent characteristics of the sampling distribution. The *sampling distribution* of a statistic is the distribution of the statistic calculated from repeated random samples of a given size, *n*, drawn from the study population.

The main point to note about descriptive analyses and statistical inference is that the first step of *both* should be to display the empirical distribution of the outcome and that the choice of statistics to describe the empirical distribution or estimate population parameters should be informed by the shape of the empirical distribution of the variable. Now that the importance of displaying the empirical distribution of data has been established we can discuss how it should be displayed.

Displaying the empirical distribution of data

A variety of factors should be considered when selecting a plot to display the empirical distribution of data, such as: *research question* (e.g. Are you interested in how an outcome varies with a particular exposure?); *study design* (e.g. Were the data measured at a single time point or several time points?); *measurement/assay factors* (e.g. Is there a lower limit of detection?); *purpose* (Descriptive or statistical inference?); and *measurement scale* of the data collected. The measurement scales of most of the data can be classified as

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

Table 1. Definition of statistics used to measure the location and spread of an outcome's empirical distribution

Statistic	Definition
Measures of location	
Mean	The summed data values divided by the number of observations.
Median	Midway value in a set of observations, 50% of the data values are above and below the median (50th percentile). Can be used to describe the middle or location of data with an empirical distribution that is normal/symmetric or skewed.
Measures of spread	
Standard Deviation (S.D.)	The S.D. is the square root of the variance, and the (sample) variance is the average squared deviation from the mean. The S.D. describes the spread of the data values.
Inter-quartile range (25th–75th percentile)	A range in which the middle 50% of data are contained (i.e. data values between the 25th and 75th percentiles). Used to describe the spread of the middle 50% of data.
Range	The values between which all the observed data values are contained.
Statistical inference only	
Standard Error (S.E.)	Is a measure of the spread of the sampling distribution of an estimate. For example the S.E. of the sample mean is estimated by the sample S.D. divided by the square root of the sample size.
95% confidence interval	A plausible range of values for the population parameter. The confidence level of a confidence interval (CI, i.e. the 95%) is unintuitive to interpret, it tells us if the study were repeated numerous times (say 20) and a 95% CI calculated for each repeated study, that 95% of the CIs (i.e. 19, on average) will contain the true population parameter. When we interpret the 95% CI derived from our single study we assume that our 95% CI is one of the 19 that contain the population parameter and not the one that misses!
P-value	Evaluate/quantifies the strength of evidence against a null hypothesis. A null hypothesis is typically a neutral statement about the population of interest (e.g. there is no difference in population means for two groups). The *P*-value is the probability of obtaining the observed sample statistic (e.g. sample mean difference) or a more extreme result when there is no difference in the population parameter (e.g. population mean) between the groups of interest (e.g. treatment and control).

either numerical or categorical. Numerical data can be divided into continuous (decimal numbers e.g. haemoglobin, temperature, body weight) or discrete (whole numbers or counts, e.g. number of hospital admissions) and categorical data into ordinal (characteristics can be ordered e.g. socio-economic status with three categories; low, medium and high), nominal (characteristics have no ordering e.g. species of malaria infection; *Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale, Plasmodium malariae*, etc.) or binary (characteristic has two categories e.g. sex; males and females).

The plot selected to show the empirical distribution of a data variable depends on the measurement scale. Continuous data are displayed using histograms, boxplots and dotplots, while bar charts are used to display the distribution of discrete and categorical data. Histograms can also be used to display discrete data, if it has a wide range of values. This tutorial will provide advice on how to display continuous and discrete data using such plots. Displaying categorical data using bar charts are not discussed.

Example dataset description

To illustrate how to plot the empirical distribution of continuous data and use it in a descriptive analysis and to inform statistical inference, we use an example dataset of 317 pregnant women (93 malaria infected and 224 parasite-free during pregnancy) attending antenatal clinics of the Shoklo Malaria Research Unit in north-western Thailand which has been published previously (Fowkes *et al.* 2012). The variables used here for illustrative purposes are total IgG titre (units optical density (OD)) in response to three *Plasmodium* blood stage antigens (a *P. falciparum* merozoite (*Pf* merozoite), *P. falciparum* infected erythrocyte (*Pf*-IE) and *P. vivax* merozoite (*Pv* merozoite) antigen). All women in the example dataset had a total IgG titre

Table 2. Summary statistics describing the distributions of IgG titres for three *Plasmodium* blood stage antigens measured from blood samples collected from 317 pregnant women

	Median IgG titer (units OD) {25th–75th percentile} [Range]		
	Pf merozoite antigen	*Pf* infected erythrocyte antigen	*Pv* merozoite antigen
Malaria infected case ($n = 93$)	0·12 {0·07–0·44} [0·002–1·54]	0·34 {0·20–0·80} [0·07–1·16]	0·16 {0·09–0·31} [5×10^{-4}–1·47]
Non-infected control ($n = 224$)	0·04 {0·02–0·08} [$2{\cdot}5 \times 10^{-4}$–0·68]	0·13 {0·07–0·22} [0·01–1·06]	0·07 {0·04–0·14} [0·001–1·16]

Table 3. Glossary of terms used to describe the shape of a variable's empirical distribution

Term	Definition
Normal or Bell-shaped	A *symmetrical* distribution with a *single* peak. The curve either side of the peak resembles the outline of a bell.
Skewed	An *asymmetrical* distribution with a *single* peak. If the tail on the left side of the peak is longer or fatter than the right side the distribution is referred to as *negatively skewed*. If the tail on the right side of the peak is longer or fatter than the left side the distribution is referred to as *positively skewed* (e.g. the empirical distribution of parasitaemia measures and antibody levels determined by immunoassays are typically positively skewed).
Truncated	A distribution whose values are limited to lie above and/or below a given threshold(s) or within a certain range (e.g. parasitaemia determined by microscopy can only detect parasite burdens above 50 parasites μL^{-1} of blood)
Zero-inflated	A distribution characterized by the presence of a large portion of zero values, in addition to continuous or discrete non-zero (i.e. positive) values
Multimodal	A distribution with multiple peaks.

measurement recorded in response to each *Plasmodium* blood stage antigen (Table 2). For the purposes of illustration the dataset has been modified and we assume that IgG titres in response to each antigen have been measured at delivery and are the outcome variable of interest. The research question is: does antibody response to each antigen at delivery differ between women exposed to malaria during pregnancy (parasite-infected cases) and parasite-free controls? To illustrate how to display the empirical distribution of discrete data, parasitaemia measurements (number of asexual parasites μL^{-1} of blood – any species of malaria) from 168 women (a single measurement per woman) taken at one of the follow-up visits (as only three were positive at baseline).

In the following sections, we demonstrate appropriate plots (*The good*) for visualizing the empirical distribution of IgG titres (continuous variable) by a binary variable (malaria infected cases and parasite-free controls), and demonstrate how such plots can be used in a descriptive analysis and to inform statistical inference. We also perform similar demonstrations for discrete parasitaemia measurements, which are much more challenging to visualize and analyse. The 'good' plots for continuous variables only are contrasted with inappropriate displays of such data (*The bad*) that are frequently presented in the literature. All plots were created using the statistical package Stata (StataCorp, 2013); example code for these plots is provided in an online appendix to facilitate practical application.

APPROPRIATE DISPLAYS OF THE EMPIRICAL DISTRIBUTION FOR DESCRIPTIVE ANALYSES AND TO INFORM STATISTICAL INFERENCE – The good

Continuous variables

Examining the shape of the empirical distribution of a continuous variable using suitable displays can help you select appropriate statistics to describe its location and spread and to decide on an approach for statistical

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

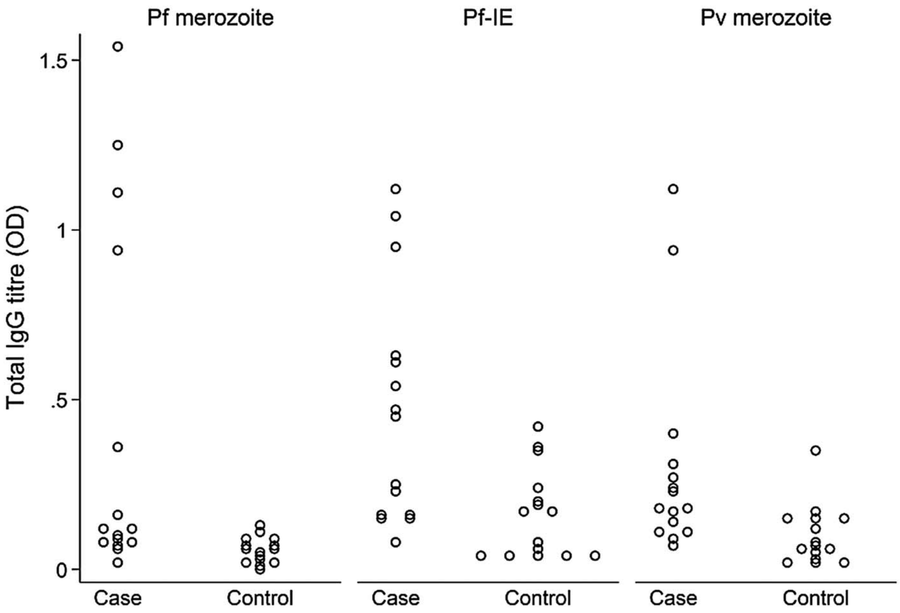

Fig. 2. Dotplot of IgG titres in response to each *Plasmodium* blood stage antigen. Data presented are a random subset of 30 (15 non-infected controls and 15 malaria infected cases) out of the 317 pregnant women included in the example dataset.

inference. The most common shapes that researchers will encounter for continuous variables are normal or skewed (both defined in Table 3). If the display reveals the data's empirical distribution is approximately normal, then the sample mean and S.D. are suitable measures of location and spread, and the normal distribution can be used for statistical inference about the population. This is no longer the case if the display reveals that the data's empirical distribution is highly skewed, where the median should be used as a measure of location and the inter-quartile range (25th–75th percentiles) and/or range as measures of spread. The effect of positive or negative skewness is to pull the mean above or below the median, respectively (i.e. the mean no longer reflects where the bulk of the data lies). Note that if the empirical distribution is approximately normal or symmetric then the mean should be similar to the median. The example in this tutorial deals with only normal and positively skewed data. Advanced statistical techniques are required if the data's empirical distribution has one of the other shapes mentioned in Table 3 or some other shape.

The most accurate display shows all the data points collected in a sample, but this may be difficult to do in an uncluttered way when the sample is large. Accordingly, we recommend the use of different plots depending on sample size. The threshold used to define a small and large sample in the following sections is simply a guide.

Displays for small samples

If the sample is small (⩽30 individuals) then display all the data using a *dotplot*. Figure 2 presents dotplots of IgG titre in response to each *Plasmodium* blood stage antigen. Data are a random subset of 30 (15 parasite-free controls and 15 malaria infected cases) out of the 317 pregnant women. Each dot on the plot represents an observation for an individual, plotted along the *y*-axis is IgG titre in response to each antigen with observations for cases and controls plotted alongside each other for comparison. If women have the same IgG titre in response to a particular antigen the observations are stacked horizontally, so that the frequency of a particular observation is represented, for example, in antigen group *Pf*-IE *merozoite* (Fig. 2) five controls have an IgG titre of 0·04 OD.

How should we interpret Fig. 2 for a descriptive analysis and to inform statistical inference? *Descriptive analysis:* the dotplot clearly shows that IgG titre in response to each antigen tend to be lower with less spread for parasite-free controls compared with malaria infected cases. For both cases and controls, the IgG titres in response to each antigen tend to be bunched towards zero with a tail stretching towards the right, which indicates that the data's empirical distribution for each group is positively skewed (defined in Table 3). Accordingly, the location and spread for each group is best described and compared using the median and interquartile range (and/or range), respectively. Measures of location and spread can be included on dotplots, but make sure you have selected them based on the shape of the distribution. If the dotplot function in the statistical package you are using defaults to adding measures of location and spread to the plot, check what the default settings are and make sure they are suitable. *Statistical inference:* the descriptive analysis

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

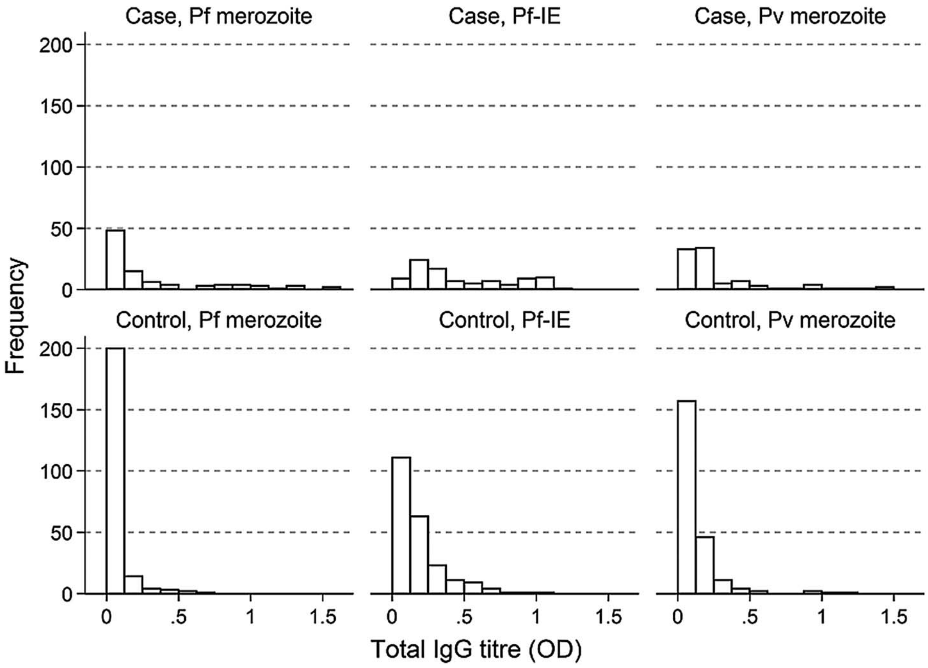

Fig. 3. Histogram of IgG titres in response to each *Plasmodium* blood stage antigen from 317 pregnant women (224 non-infected controls and 93 malaria infected cases) included in the example dataset.

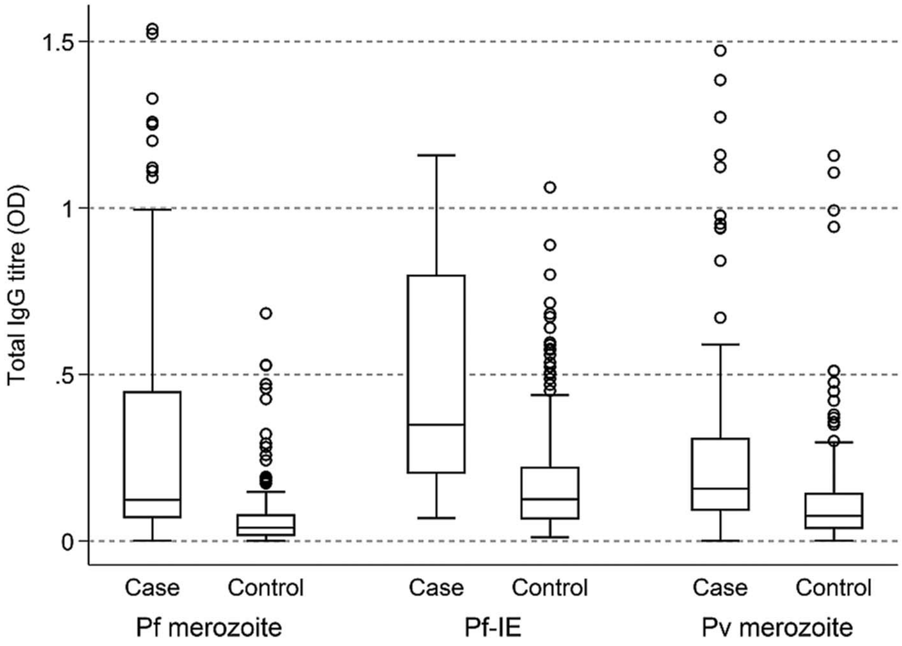

Fig. 4. Box & whisker plots of IgG titres in response to each *Plasmodium* blood stage antigen from 317 pregnant women (224 non-infected controls and 93 malaria infected cases) included in the example dataset. Box represents the inter-quartile range and the horizontal line within the box represents the median IgG titre. The whiskers end at the largest and smallest IgG titre excluding any outliers, and the circles outside the whiskers are outliers.

showed that the data's empirical distribution is positively skewed; therefore, statistical inference based on the normal distribution is not appropriate. The most straightforward approach to statistical inference for skewed data is to transform the data's empirical distribution so that the transformed values are close to a normal distribution. This approach will be adopted in this tutorial and is explained in detail in section '*Appropriate displays for statistical inference*'.

Displays for large samples

If the sample is large (>30 individuals), *histograms* and *box & whisker plots* are appropriate displays. Figures 3 and 4 are histograms and box & whisker plots of IgG

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

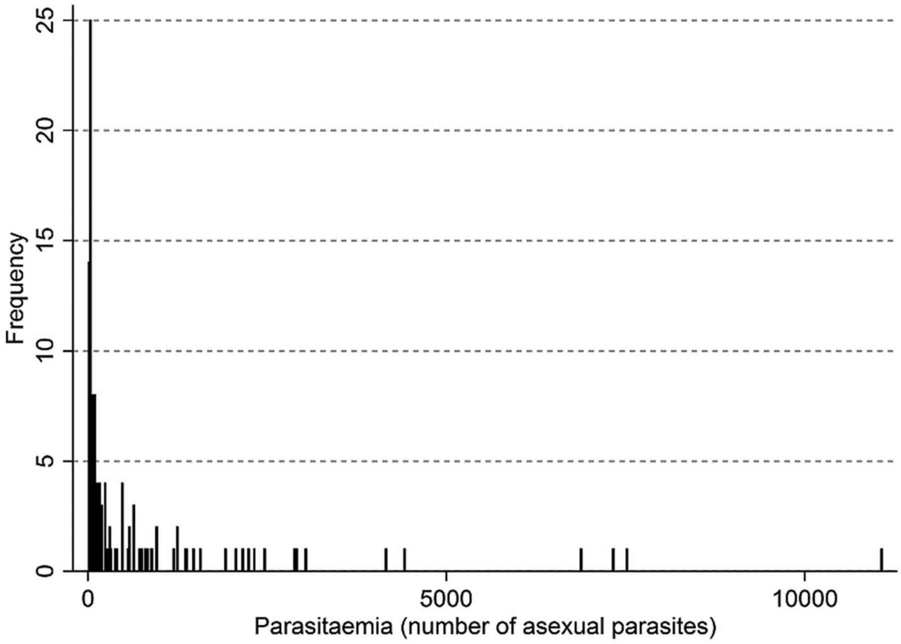

Fig. 5. The distribution of 129 non-zero discrete parasitaemia measurements presented in a bar chart. There were 39 zero measurements out of 168 observations.

titre in response to each antigen for the 93 malaria infected cases and 224 parasite-free controls. A *histogram* is constructed by dividing the data range into several non-overlapping equally sized bins (categories) and the number of observations falling into each bin counted. The bins are displayed on the x-axis and the frequency (or percent or proportion) on the y-axis. A *box & whisker plot* consists of a box which represents the inter-quartile range, that is 25% of the IgG titres lie above the top of the box (i.e. 75th percentile), the box itself contains the middle 50% of the IgG titres and 25% of the IgG titres lie below the bottom of the box (i.e. 25th percentile). The horizontal line within the box represents the median IgG level (i.e. 50th percentile). The whiskers end at the largest and smallest IgG values excluding any outliers. The outliers are defined as those observations greater than 1·5 times the inter-quartile range from the top or bottom of the box, and are represented as points outside the whiskers. Of note, the criterion for determining an outlier may differ between statistical packages, the definition we have given is used by the common statistical packages (e.g. Stata, R, SAS and SPSS). Histograms should be used to identify the shape of the data's empirical distribution. If the histogram reveals that the shape of the data's empirical distribution is normal or skewed, then box & whisker plots are more efficient displays of these shapes and are particularly useful for comparing how the location and spread of normal and skewed distributions vary across several groups (compare Fig. 3 with 4). Both histograms and box & whisker plots are better displays of larger datasets than dotplots, where the latter can look rather messy for a large number of observations. For more details on box & whisker plots, dotplots and other displays see (Freeman *et al*. 2009).

How should we interpret Figs 3 and 4 for a descriptive analysis and to inform statistical inference? *Descriptive analysis:* similar to dotplots (Fig. 2) the histograms and box & whisker plots (Figs 3 and 4) also display the differences in the distribution of IgG titre in response to each antigen between cases and controls, that is the IgG titre for controls are lower and less variable (spread of histogram is narrower and the box & whisker plots have a narrower box and smaller whiskers) than cases. The histograms (Fig. 3) and box & whisker plots (Fig. 4) show a tendency for the observations below the median to be contained in a narrower range than the data above the median, which indicates the IgG titre in response to each antigen for cases and controls are positively skewed. *Statistical inference*: will proceed as outlined in the previous section '*Displays for small samples*'.

Discrete variables

Discrete numerical data (whole numbers such as counts) are also very common in parasitology, in the form of parasitaemia (e.g. number of asexual parasites μL^{-1} of blood), gametocytaemia (e.g. number of gametocytes μL^{-1} of blood). Displaying discrete data are more challenging than displaying continuous data. Bar charts (although not ideal for displaying the empirical distribution of continuous data – see Fig. 1), can be used to display the frequency with which values of a discrete variable occur.

In the example parasitaemia dataset (see section '*Example dataset description*' for a description) there are a large percentage (23% or 39/168) of women who did not have a malaria infection (i.e. parasitaemia measurement recorded as zero) and such data is often

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

referred to as zero-inflated data. For zero-inflated data, we recommend plotting the non-zero measurements (e.g. parasitaemia in those who had a malaria infection) and reporting the percentage of zero measurements. Figure 5 is a bar chart of the parasitaemia data for those women who had a malaria infection (i.e. had non-zero parasitaemia measurements recorded). On the horizontal axis are the numbers of parasites, going from a minimum of 16 parasites to maximum of 11 078 parasites, while on the vertical axis is the frequency with which these measurements occur. The vertical axis could also be rescaled to percentages, which facilitates the comparison of groups. The bar chart shows that the empirical distribution of the parasitaemia measurements is highly positively skewed. However, if the discrete variable has a large number of unique values then plots suitable for continuous data (dot plots for smaller datasets, box plots or histograms for larger datasets) would be appropriate. Since there are 49 unique values for the parasitaemia data plotted in Fig. 5, it would be reasonable to apply plots for continuous data to this discrete variable (see section '*Continuous variables*' for further details). Note the extreme positive skew makes the mean a poor location statistic for these data (mean number of parasites per host is 555, which is much greater than the median of 64 – these sample statistics were calculated including the zero measurements). In the case when the discrete variable consists of a large number of unique categories/counts (and consequently is amenable to analysis with statistical methods for continuous variables), a possible approach to make the distribution approximately normal (or less skewed) would be to take a log transformation (as illustrated in section '*Transformation of IgG titre values*'), but this is often impossible if the discrete variable has a high frequency of zeroes (O'Hara & Kotze, 2010). A brief discussion of how to analyse discrete data and what summary statistics to use is provided under 'Discrete variables' in the next section.

APPROPRIATE DISPLAYS FOR STATISTICAL INFERENCE

Continuous variables

As we described in the introduction, statistical inference is performed to draw conclusions about the distribution of population parameter based on its empirical distribution. The 'good' displays of the IgG titre's empirical distribution showed that its shape in both the small and large sample size examples was positively skewed. In the following sections we will use the complete dataset ($n = 317$) to show how taking the natural log-transformation of positively skewed data can result in an empirical distribution that is approximately normal.

We also recommend statistical inference plots which display estimates (statistics used to estimate a population parameter) and 95% CIs (a plausible range of values for the population parameter of interest) rather than the mean ± one S.E. which corresponds to the 67% CI for the population parameter (i.e. the 67% CI is less likely to contain the true population parameter than the 95% CI).

Transformation of IgG titre values

Data can often be transformed to remove skewness and make the data's empirical distribution resemble a normal distribution. The histograms (Fig. 3) and box & whisker plots (Fig. 4) showed that total IgG titre in responses to all three antigens were positively skewed for cases and controls. When data are positively skewed a log-transformation can often be applied to each data point to make the distribution of the data resemble a normal distribution (note a log-transformation cannot be applied to zero or negative values). As mentioned earlier the parameters that govern the shape of a normal distribution are the population mean and S.D., and the sample mean and sample S.D. are used to estimate the population parameters.

Box & whisker plots are used to display the $\log_e$-transformed IgG titres (Fig. 6). The $\log_e$-transformed total IgG titres appear to be normally distributed, for example each box looks relatively symmetric around the median value and there are considerably fewer outliers. The length of each box also appears similar for cases and controls across antigens, which shows that the spread or variance of the antibody levels (log-transformed) is similar for the malaria infected cases and parasite-free controls. Histograms are also very useful for determining whether data are approximately normally distributed data (if the histogram is symmetric and roughly bell-shaped, and the mean is similar to the median, then the assumption of normality is typically accepted). Quantile–quantile (Q–Q) plots are another useful display for examining normality (see Kohler & Kreuter, 2012 for more details).

Statistical inference plots

Statistical inference plots depict estimates and 95% CIs (defined in Table 1) for the population parameters of interest. We now demonstrate how to make statistical inferences about the population mean $\log_e$-transformed IgG titre to the *Pf merozoite* antigen between cases and controls in the example dataset. In the previous section we established that a normal distribution is an appropriate model for the population distribution of the log-transformed data. The estimates are the sample mean $\log_e$-transformed IgG titre to *Pf merozoite* antigen for cases and controls, and the limits of the 95% CI are calculated from: sample mean ± (1·96 × S.E. of the sample mean) where the S.E. of the sample mean equals the sample S.D. divided by $\sqrt{n}$ (n is 93 for malaria infected cases and 224 for parasite-free controls).

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

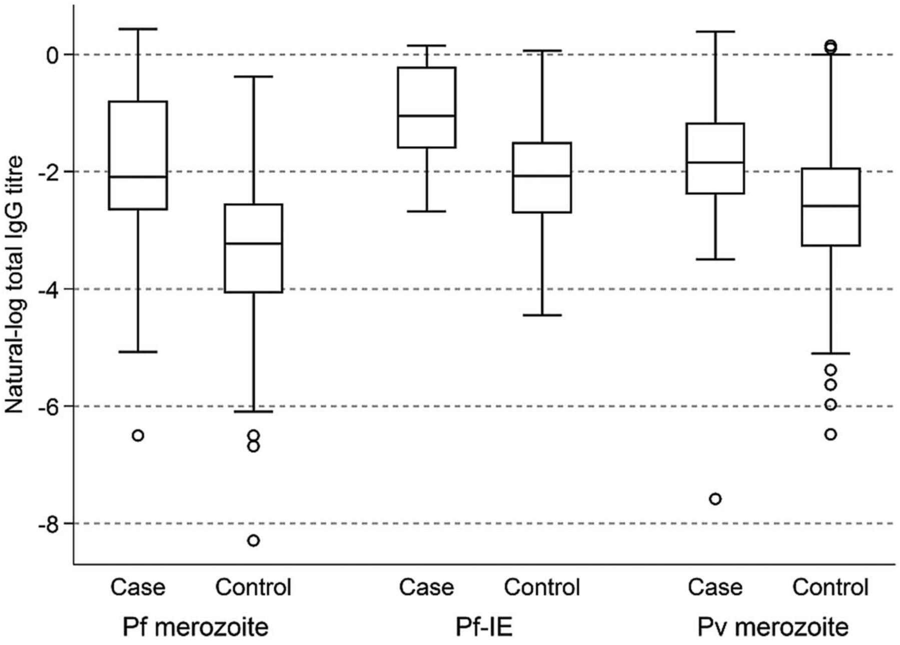

Fig. 6. Box & whisker plots of natural log-transformed IgG titres in response to each *Plasmodium* blood stage antigen from 317 pregnant women (224 non-infected controls and 93 malaria infected cases) included in the example dataset. Box represents the inter-quartile range and the horizontal line within the box represents the median $\log_e$-transformed IgG titre. The whiskers end at the largest and smallest $\log_e$-transformed IgG titre excluding any outliers, and the circles outside the whiskers are outliers.

The value 1·96 in the 95% CI calculation for the population mean is suitable for this example because the sample size is large, however, when the sample size is less than 60 the 1·96 should be replaced with the 97·5th percentile of the t distribution with the degrees of freedom (D.F.) equal to $n-1$. Once the sample size is larger than 60 the 97·5th percentile of the t distribution is close to 1·96, whereas for smaller sample sizes this value increases with decreasing sample size (i.e. more conservative CIs are calculated for smaller sample sizes).

We can interpret the results on the original scale (as log-transformed variables do not retain the original units of measurement) by exponentiation (typically e^x on a calculator or the exp()function in most statistical packages) of the estimate and the limits of the CI calculated on the $\log_e$-scale. The exponentiated estimate is the geometric mean and the exponentiated 95% CI is for the population geometric mean, which have the same units as the outcome (in our example IgG titre measured in OD units). The geometric mean should closely approximate the median of the untransformed data if the $\log_e$-transformed data are normally distributed.

In Fig. 7 the x-axis represents malaria exposure group (malaria infected or parasite-free), and the y-axis is the mean $\log_e$ IgG titre in the left panel and the back transformed results (i.e. the exponentiated mean $\log_e$ IgG response and corresponding 95% CI for population mean $\log_e$ IgG titre) on the original OD units in the right panel, with the dots representing the sample mean (arithmetic left panel and geometric right panel) for controls, the squares representing the sample mean for cases and the error bars portraying the 95% confidence limits for the population mean. The statistical inference plots in Fig. 7 suggest the population geometric mean IgG level (or, alternatively, the population mean $\log_e$ IgG level) in response to each antigen is higher for cases than controls.

Some researchers like to add P-values examining the strength of evidence against the null hypothesis (which, in this example, is that the *population* mean IgG level for cases is the same as controls) to statistical inference plots. If it is the researcher's preference to include P-values on such plots, we recommend stating the exact P-value and not presenting stars or symbols to indicate whether the P-value is below a particular threshold, e.g. <0·05 (see Fig. 1). When the P-value is very small, then stating, for example, the P-value is <0·001 is acceptable.

In situations where continuous data (original data values) are not normally distributed and a transformation of the data cannot alleviate this problem, then statistical inference plots of medians and 95% CIs for the population median (using statistical methods not covered in this tutorial, such as those presented in (Campbell & Gardner, 1988), an approach based on the beta distribution and the bootstrap method) or box & whisker plots along with P-values from an applicable non-parametric test (e.g. Mann–Whitney U (Hart, 2001)) are recommended.

Statistical inference plots are not restricted to displaying statistical inferences regarding population

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

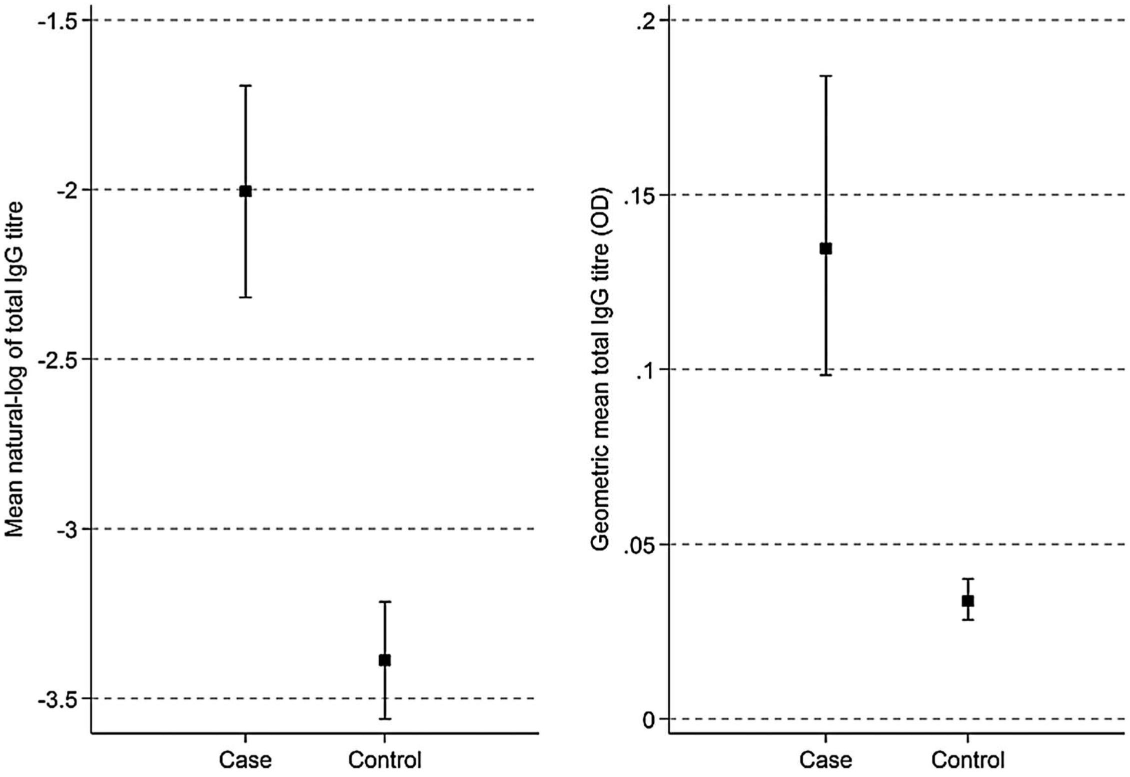

Fig. 7. Left panel: sample mean $\log_e$ IgG titre in response to antigen *Pf* merozoite between cases and controls, and the 95% confidence interval (CI, error bars) for the population mean. Right panel: geometric mean $\log_e$ IgG titre (OD) into antigen *Pf* merozoite between cases and controls, and the 95% CI (error bars) for the population geometric mean.

means and medians, but statistical inferences concerning any population parameter of interest can be displayed (e.g. proportions, odds, rates etc.) assuming suitable estimates and 95% CIs can be derived (see online Supplementary Figure 1 as an example of a statistical inference plot for the population proportion).

Discrete variables

Statistical inference plots are also difficult for discrete data, due to the combination of skew and zeroes mentioned above. Generally the mean or expected count (discrete value) is estimated by modelling the counts with an appropriate probability distribution for discrete data such as Poisson or negative binomial, or zero-inflated versions of these distributions that can account for the high frequency of zeros (Bolker *et al*. 2009; O'Hara & Kotze, 2010). Such modelling is beyond the scope of an accessible article on data display.

INAPPROPRIATE DISPLAYS FOR DESCRIPTIVE ANALYSES AND STATISTICAL INFERENCE – The bad and the error bar

In parasitological research bar charts (bar starts at zero and extends to the group mean) with the upper half of the error bar on top (comically referred to as detonator plots) are often used to describe the distribution of a continuous variable or display statistical inferences about the population mean (Fig. 1). There are two categories of error bars commonly displayed on detonator plots: descriptive and inferential. In the following sections detonator plots with descriptive error bars (descriptive detonator plots) and detonator plots with inferential error bars (inferential detonator plots) will be compared with box & whisker plots and statistical inference plots, respectively. For further details on detonator plots and the error bars typically displayed see (Cumming *et al*. 2007; Vaux, 2008).

Descriptive detonator plots vs *box & whisker plots*

Descriptive detonator plots can obscure the shape of the data distribution by the use of summary statistics that are not robust (i.e. interpretation of the statistic changes depending on the shape of the data distribution) to skewness/outliers (e.g. mean and S.D.) or through the use of too few robust summary statistics (displaying only the mean and maximum). For example, in Fig. 8 detonator plots comparing the distribution of the raw (i.e. untransformed) IgG titres with each antigen between cases and controls are presented. The descriptive error bars are the sample mean plus one S.D. in panel A and the maximum in panel B. No information is presented for the distribution of the data below the top of the bar (i.e. below the sample mean). Much of the discussion relating to Fig. 8, panel A, is applicable to Fig. 1.

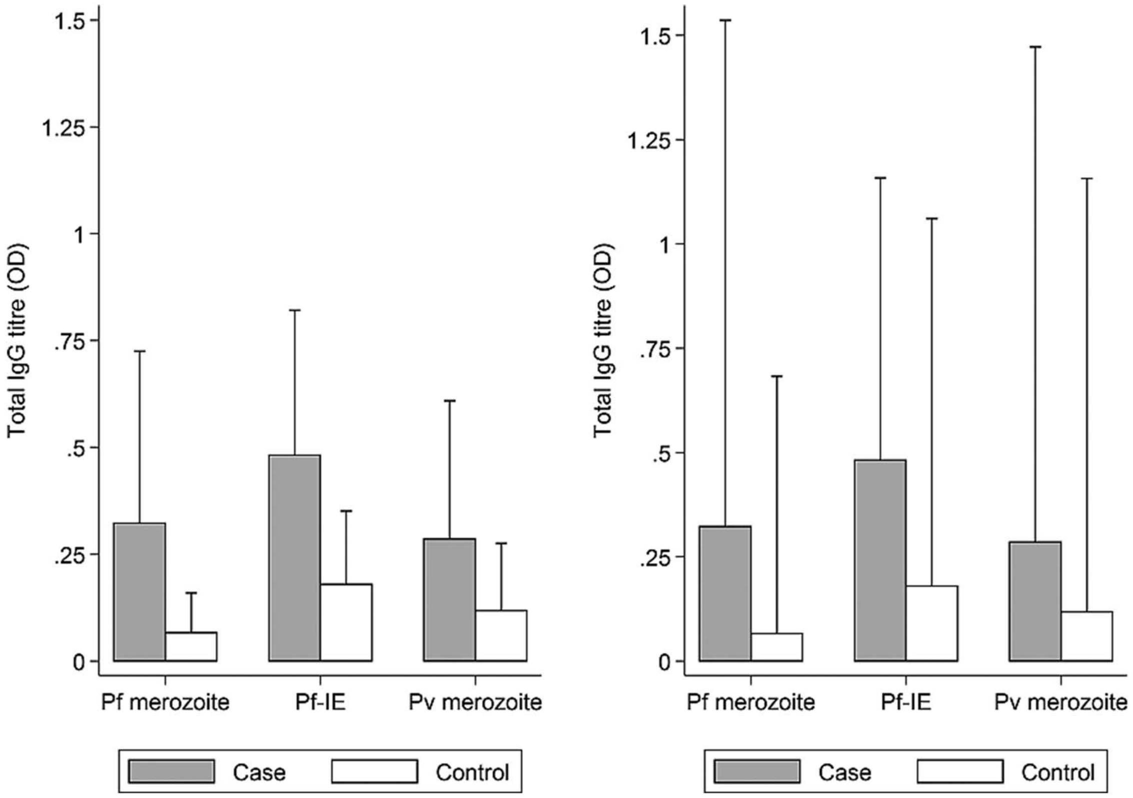

Fig. 8. Detonator plot of the distribution of IgG titres in response to each *Plasmodium* blood stage antigen from 317 pregnant women (224 non-infected controls and 93 Malaria infected cases) included in the example dataset. In both panels the bar indicates the mean. Left panel: mean plus one S.D. is the error bar. Right panel: maximum is the error.

Our examination of appropriate displays for the data's empirical distribution revealed that the distribution of IgG titres to each antigen for cases and controls was positively skewed. As mentioned earlier the effect of positive or negative skewness is to pull the mean above or below the median, respectively (i.e. the mean no longer reflects where the bulk of the data lies). For example in Fig. 8, the mean of the untransformed IgG tires is plotted, and without our previous investigation using the 'appropriate' displays the reader would not know that the mean is not the most appropriate measure of location.

In addition, the reader would also conclude from Fig. 8 that the difference in IgG titres for cases and controls is much larger than it appears due to the effect of the positive skewness (e.g. difference in means between cases and controls is 0·26, 0·30 and 0·17 for *Pf* merozoite, *Pf*-IE and *Pv* merozoite groups). The median (a measure of central tendency that is robust to skewness), however, indicates that the difference in IgG titres to each antigen for cases and controls is not as large as that displayed in Fig. 8 (e.g. difference in medians between cases and controls is only 0·08, 0·22 and 0·08 for *Pf* merozoite, *Pf*-IE and *Pv* merozoite groups – the medians for the groups are also displayed in Fig. 4 and are provided in Table 2).

We know from the boxplots in Fig. 4 that there is less variability in measurements from controls compared with cases and the IgG titres to each antigen for cases and controls are positively skewed, but such useful information about the spread/distribution of the IgG titres cannot be easily ascertained from either (or both) of the detonator plots in Fig. 8. The detonator plot with error bars equal to one S.D. suggests that the distribution of IgG titres is normally distributed, and can be summarized using the mean and S.D. If the data were normally distributed then 67% of the observations should fall between the mean ± one S.D., which is not the case for the IgG titres to *Pf merozoite* antigen from cases (the mean minus one S.D. is −0·1 but the 16·5th percentile is 0·04 OD units) and raises questions that the assumption of normality may not be well supported by these data. Drawing this conclusion from Panel A of Fig. 8 requires the viewer to roughly calculate the mean minus the S.D. (or the lower limit of the error bar). Skewness can only be diagnosed using this plot if the data are positive valued or the error bars contain implausible values. In all other situations it is impossible to diagnose skewness or deviations from normality using a detonator plot with error bar equal to mean + one S.D. The detonator plot displaying the mean and maximum is also not very informative and only shows you how far the maximum is from the mean. Since the mean (not the median) is being displayed we cannot be certain whether 50% of the observations lie between the

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

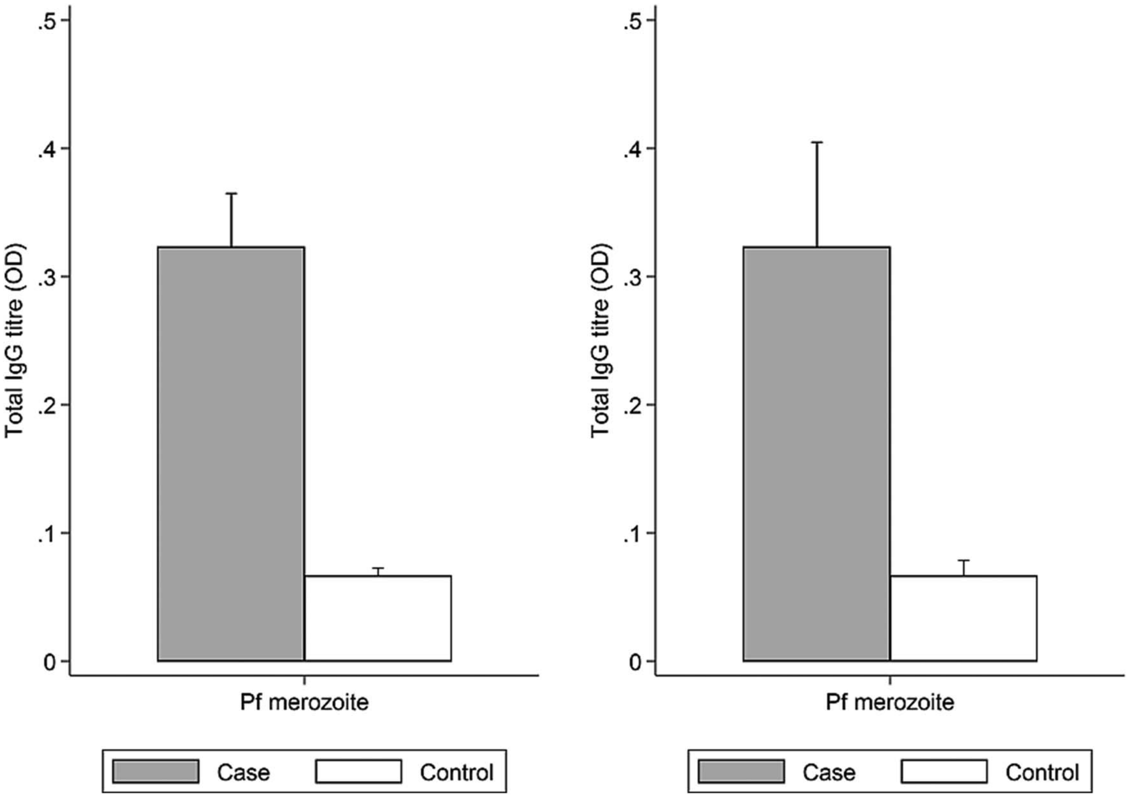

Fig. 9. Inferential detonator plots of the sample mean IgG titre in response to antigen *Pf* merozoite between cases and controls (bars). Left panel: mean plus S.E. is the error bar. Right panel: upper limit of the 95% confidence interval for the population mean is the error bar.

mean and the maximum (without assuming the data are normal or symmetric) and cannot unambiguously determine whether the data are skewed (even if the minimum were included it may be difficult to diagnose skewness or deviations from normality because these single minimum and maximum values may be outliers). The scale of detonator plots beginning at zero is also misleading as zero values may be biologically implausible or not observed in the study sample.

The need to make strong assumptions (e.g. normality) or perform additional calculations in order to draw conclusions from descriptive detonator plots about the shape of the data illustrates that descriptive detonator plots cannot faithfully display the data and that box & whisker plots are a superior alternative.

Inferential detonator plots vs *statistical inference plots*

Inferential detonator plots are essentially statistical inference plots with the lower half of the inferential error bar removed and the mean displayed using a bar extending from zero rather than a point. Omitting the lower half of the error bar impedes the viewer's ability to visualize the lower range of plausible values for the population parameter. As mentioned previously, the sample mean of the original data values may not represent the true centre of the empirical distribution, as is the case for our data example (see Fig. 9).

DISCUSSION

Misrepresentation of the shape of the data distribution can lead to incorrect statistical inferences about the population being made (e.g. making inferences about the population mean when data are positively skewed, as illustrated in sections '*Descriptive detonator plots vs box & whisker plots*' and 'Inferential detonator plots *vs* statistical inference plots'). Potential differences in conclusions can also arise when examining data using different plots.

Dotplots, histograms and box & whisker plots can be produced in all standard statistical software packages and are far superior to descriptive detonator plots commonly used to display the distribution of continuous variables in the parasitology literature. The 'good' displays discussed in this article either show all the data, the frequency of observations in small groupings of the continuous variable, or summary statistics that are robust to skewness (e.g. median, inter-quartile range, minimum, maximum and outliers), respectively. Another advantage of these 'good' displays for descriptive analyses over descriptive detonator plots is that the same features are presented in the display, unlike detonator plots

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press

where the error bars need to be selected and may be inconsistent from plot to plot.

Inferential detonator plots and statistical inference plots for displaying statistical inferences concerning a population parameter (estimates and 95% CIs) are similar, but inferential detonator plots display too little information wastefully (e.g. display the estimate (a single point) using a bar extending from zero to the estimate and only the top half of the error bar), whereas statistical inference plots display the estimate as a point and the upper and lower halves of the error bar.

The example used in this tutorial is very simple (with the outcome variable measured at a single time point from independent individuals). More advanced topics, such as appropriate data displays to help inform linear regression analyses (used to examine the association between an outcome and multiple continuous and categorical covariates), are provided in (Zuur *et al*. 2010). Note that while the 'good' displays illustrated in this tutorial can be applied to dependent data (e.g. data collected in longitudinal studies), they do not accurately represent trends over time or clustering of repeated measurements collected from the same individual (for simple approaches to visualizing and analysing repeated measurement data see (Matthews *et al*. 1990)).

In order to promote the wider adoption of the 'good' displays discussed in this article, it might be helpful for major parasitology journals to provide guidance to authors on data visualization or link to such guidance (which is what the BMJ, a major medical journal, does (Group, 1997)), and inform reviewers of the importance of transparent data display and the potential impact on downstream inference. This combined with better dissemination of the wealth of statistical resources available (e.g. textbooks, webpages etc.) to senior researchers and PhD students, either through access to statistics courses or online resources, would also improve the quality of displays used in the parasitology literature.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0031182015000748

ACKNOWLEDGEMENTS

We thank James Beeson, Rose McGready and Francois Nosten for providing us with the example dataset.

FINANCIAL SUPPORT

Sophie Zaloumis and Alysha M De Livera were supported under an Australian National Health and Medical Research Council (NHMRC) Centre of Research Excellence grant, ID# 1035261, awarded to the Victorian Centre for Biostatistics (ViCBiostat). Freya JI Fowkes is supported by a Future Fellowship, Australian Research Council.

REFERENCES

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. and White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**, 127–135.

Campbell, M. J. (2009). *Statistics at Square One. [electronic resource]*, 11th Edn. John Wiley & Sons, Ltd., Chichester.

Campbell, M. J. and Gardner, M. J. (1988). Calculating confidence intervals for some non-parametric analyses. *British Medical Journal (Clinical Research ed.)* **296**, 1454–1456.

Cumming, G., Fidler, F. and Vaux, D. L. (2007). Error bars in experimental biology. *Journal of Cell Biology* **177**, 7–11.

Fowkes, F. J., McGready, R., Cross, N. J., Hommel, M., Simpson, J. A., Elliott, S. R., Richards, J. S., Lackovic, K., Viladpai-Nguen, J., Narum, D., Tsuboi, T., Anders, R. F., Nosten, F. and Beeson, J. G. (2012). New insights into acquisition, boosting, and longevity of immunity to malaria in pregnant women. *Journal of Infectious Diseases* **206**, 1612–1621.

Freeman, J. V., Walters, S. J. and Campbell, M. J. (2009). *How to Display Data*. Wiley, Hoboken.

Group, B. P. (1997). Statistics at Square One.

Hart, A. (2001). Mann–Whitney test is not just a test of medians: differences in spread can be important. *BMJ (Clinical Research ed.)* **323**, 391–393.

Huff, D. (1993). *How to Lie with Statistics*. Norton, New York.

Kohler, U. and Kreuter, F. (2012). *Data Analysis using Stata/Ulrich Kohler, Frauke Kreuter*, 3rd Edn. Stata Press, College Station, Tex.

Matthews, J. N., Altman, D. G., Campbell, M. J. and Royston, P. (1990). Analysis of serial measurements in medical research. *BMJ (Clinical Research ed.)* **300**, 230–235.

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution* **1**, 118–122.

StataCorp (2013). *Stata Statistical Software: Release 13*. StataCorp LP, College Station, TX.

Vaux, D. L. (2008). Ten rules of thumb for the presentation and interpretation of data in scientific publications. *Australian Biochemist* **39**, 37–39.

Wainer, H. (1984). How to display data badly. *American Statistician* **38**, 137–147.

Weissgerber, T. L., Milic, N. M., Winham, S. J. and Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biology* **13**, e1002128.

Zuur, A. F., Ieno, E. N. and Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1**, 3–14.

https://doi.org/10.1017/S0031182015000748 Published online by Cambridge University Press