Draft 0.00
ATLAS Statistics Forum
15 February, 2011

Recommendations for presentation of error bars

1 Introduction

This note summarizes recommendations on how to present error bars on plots. It follows discussions and presentations at the ATLAS Statistics Forum Meeting on 5 October, 2010 [1, 4]. The Statistics Forum regards this as partially a question of statistical practice, but also a question of presentation and thus to some extent a matter of convention and taste. One must also recognize that in different analyses different types of approximate methods for summarizing the a measurement and related uncertainties may be applicable, so that a single recipe will not be appropriate in all cases.

Depending on the needs of the analysis and what one wants to communicate, options for error bars are:

1. Estimate of standard deviation of the measurement, e.g., square root of measured number for Poisson data.
2. No error bars.
3. 68.3% central confidence interval.
4. 68.3% Bayesian credible interval.

For whichever option is chosen, unless it is not relevant to the discussion (rare) or absolutely clear from context, one should state what procedure was used, e.g., in the figure caption. The rationale behind these options and their advantages and disadvantages are discussed below.

Related discussions can be found in the Statistics Forum Note on error analysis for efficiency [2] and in the paper by Casadei [3].

2 Estimate of standard deviation (the "$\sqrt{n}$ method")

Probably the most common recipe for calculating the error bar for a certain number of events, n, is to take the square root of n. For concreteness, suppose we measure an integer value n that is modelled as following a Poisson distribution with some (unknown) mean value ν. One can show that the Maximum likelihood estimator for ν is $\hat{\nu} = \nu$. The estimator is a random quantity characterized by a standard deviation

$$\sigma[\hat{\nu}] = \sqrt{\nu}\,, \tag{1}$$

which itself depends on the true (and *a priori* unknown) ν. We can construct an estimator of the standard deviation of the estimator ν by evaluating Eq. (1) with $\nu = \hat{\nu}$,

$$\hat{\sigma}[\hat{\nu}] = \sqrt{\hat{\nu}} = \sqrt{n} \,, \tag{2}$$

and use this as the error bar. The procedure is simple and it is also easy to communicate what was done, e.g., in a histogram: "error bars are given by the square root of the number of entries in each bin". But it can be misleading in several important cases, such as 0 ± 0 or 1 ± 1. This is not to say that $\hat{\sigma}[\hat{\nu}] = \sqrt{n}$ is "wrong", for these cases, any more than it is wrong when any estimated value does not exactly equal the true value. Rather, the estimated standard deviation in such cases does not meaningfully encapsulate our knowledge about the expected statistical fluctuations implied by the observation. In particular $n = 0$ clearly gives an underestimate of the true standard deviation, assuming that there is nonzero probability to see $n \neq 0$.

A further drawback of the $\sqrt{n}$ recipe is that it can provide a misleading impression of the level of compatibility between a certain hypothesis for the mean value ν and the data n. Suppose one observes $n = 1$ and thus plots 1 ± 1. This appears to be perfectly compatible with $\nu = 0$, but of course this hypothesis is ruled out by observing any nonzero n. The plotted $n = 1$ also appears compatible with, say, the hypothesis $\nu = 10^{-8}$. But the p-value of this hypothesis would be

$$p = P(n \geq 1; \nu) = 1 - P(0; \nu) = 1 - e^{-\nu} \approx \nu = 10^{-8} \,. \tag{3}$$

That is, the hypothesis of a very small mean value is also very incompatible with the data, although it is contained with in the presented interval of 1 ± 1. For these reasons, the $\sqrt{n}$ method is not recommended for cases of small or zero n, unless for example these represent, tail values that are not relevant to the message one wishes to communicate.

In cases where the observed value is not modelled as a Poisson variable, one can follow the same basic procedure using whatever probability model is assumed. For example, when estimating an efficiency, one may find m events passing cuts out of N total, and model m as a binomially distributed variable with an acceptance probability per event ε (the efficiency). The Maximum Likelihood estimator for ε is

$$\hat{\varepsilon} = \frac{m}{N} \,, \tag{4}$$

and the variance of a binomially distributed variable m is $V[m] = N\varepsilon(1-\varepsilon)$. As above if we estimate $\sigma[m] = \sqrt{V[m]}$ by substituting the estimator for ε one finds the estimator of the standard deviation of $\hat{\varepsilon}$,

$$\hat{\sigma}[\hat{\varepsilon}] = \sqrt{\frac{\hat{\varepsilon}(1-\hat{\varepsilon})}{N}} = \frac{1}{N}\sqrt{m\left(1 - \frac{m}{N}\right)} \tag{5}$$

This expression suffers not only from the same pathologies for $m = 0$ or $m = 1$ as seen above for a Poisson distributed value, but has similar behaviour for $m = N$ or $m = N - 1$. For this reason, Eq. (5) is not recommended for error bars on estimated efficiencies or other quantities related to binomially distributed data if the regions $\varepsilon \approx 0$ or $\varepsilon \approx 1$ are relevant to the discussion. Instead, one should consider either frequentist confidence intervals or Bayesian credible intervals as described in Sections 4 and 5, respectively.

3 No error bars

The observed data are "exact". That is, if one finds $n = 3$ events, then this observation is definitely 3, not 2 or 4. Thus it is not necessary or even desirable to plot the observation together with an error bar. Consider, for example, the histogram in Fig. 1.

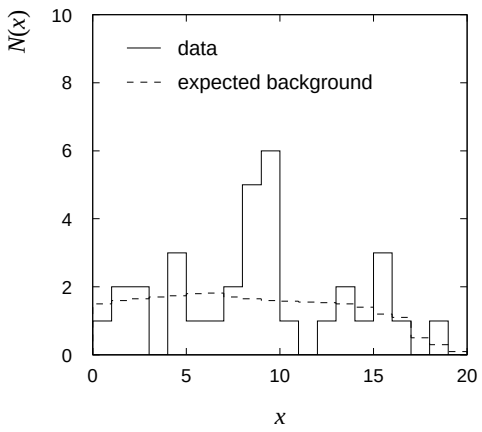

Figure 1: Observed and expected histograms of a variable x (see text).

Suppose the number of events observed in each bin is modelled as a Poisson variable with a mean given by the dashed curve. One could judge whether the peak is significant by computing a p-value that gives the probability, assuming the mean values of the dashed curve, to find a peak as significant or more so relative to the one seen. Error bars play no role in this analysis and could in fact be misleading, for the reasons discussed in Sec. 2.

What the reader can do "by eye" in viewing a plot such as Fig. 1 is to mentally compute the square root of the value of the dashed line to obtain the standard deviation of the Poisson value for that bin, and see how far the observed data is from the hypothesis as a multiple of this standard deviation. This is not as informative as the p-value but still gives some meaningful measure of the level of compatibility between the data and hypothesis. One might therefore consider plotting the hypothesized mean together with plus-or-minus one square root of the hypothesized value, but this becomes awkward, especially if one wants to display more than one hypothesis on the same plot. It is also not common practice, so this latter procedure is therefore not recommended.

4 Confidence interval

Frequentist confidence intervals provide a means to display error bars that do not suffer from the pathologies seen with the $\sqrt{n}$ method but which are (usually) easy to compute and which coincide with the $\sqrt{n}$ recipe when n is sufficiently large.

If n follows a Poisson distribution with mean ν, the lower confidence limit at confidence level $1 - \alpha$ and upper limit at level $1 - \beta$ are given by (see [5] Eq. (33.59)),

$$\nu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) \; , \tag{6}$$

$$\nu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) \; , \tag{7}$$

where $F^{-1}_{\chi^2}$ is the quantile of the chi-square distribution for the indicated number of degrees of freedom. These can be computed using the routine `TMath::ChisquareQuantile` in the ROOT library. For error bars we recommend 68.3% *central* confidence intervals. Here central means that the upper and lower tail probabilities, α and β, are equal and $1-\alpha-\beta = 0.683$, i.e., $\alpha = \beta = 0.159$ (more precisely, 0.158655). The displayed measurement and error bars are thus $\hat{\nu}^{+(\nu_{\rm up}-\hat{\nu})}_{-(\hat{\nu}-\nu_{\rm lo})}$.

For an efficiency ε estimated using binomially distributed data, the lower and upper confidence limits are given (see [5] Eq. (33.60)),

$$\varepsilon_{\rm lo} = \frac{mF_F^{-1}[\alpha; 2m, 2(N-m+1)]}{N-m+1 \,+\, mF_F^{-1}[\alpha; 2m, 2(N-m+1)]}\,, \tag{8}$$

$$\varepsilon_{\rm up} = \frac{(m+1)F_F^{-1}[1-\beta; 2(m+1), 2(N-m)]}{(N-m) \,+\, (m+1)F_F^{-1}[1-\beta; 2(m+1), 2(N-m)]}\,. \tag{9}$$

where F_F^{-1} is the quantile of the F distribution (also called the Fisher–Snedecor distribution; see [6]). The function F_F^{-1} can be obtained using `ROOT::Math::fdistribution_quantile` from the ROOT's MathCore library [7].

. As before, one would generally choose the 68.3% central confidence intervals, i.e., $\alpha = \beta = 0.159$ and display the point as $\hat{\varepsilon}^{+(\varepsilon_{\rm up}-\hat{\varepsilon})}_{-(\hat{\varepsilon}-\varepsilon_{\rm lo})}$.

Confidence intervals have the advantage that they only extend over the physically defined range of the parameter and show meaningful intervals also for $n = 0$ (Poisson) and also $m = 0$ and $m = N$ (binomial). They are in general asymmetric about the measured value, but they approach the $\pm\sigma$ standard error bars if in limiting cases (e.g., Poisson for large n; binomial for large m and $N-m$).

5 Bayesian credible interval

Instead of frequentist confidence intervals one can also base the error bars on Bayesian credible intervals. For the case of binomial data these are discussed at length in [3] and also in [2].

In Bayesian statistics one computes a posterior probability density for the parameter of interest. For example, for Poisson distributed n with mean ν, one uses Bayes' theorem to find the posterior pdf $p(\nu|n)$, which characterizes that ν takes on different values in the light of the observed n,

$$p(\nu|n) = \frac{L(n|\nu)\pi(\nu)}{\int L(n|\nu)\pi(\nu)\,d\nu}\,. \tag{10}$$

Here $L(n|\nu)$ is the Poisson likelihood function and $\pi(\nu)$ is the prior pdf, which characterizes the uncertainty in the parameter before making the measurement. Selecting this prior is one of the nontrivial aspects of a Bayesian analysis, and for present purposes we will assume this has been considered and a choice made. Often a uniform prior for $\nu \geq 0$ is used (advantages and drawbacks of the uniform prior are discussed, e.g., in [5]).

In the case of, say, an efficiency based on binomially distributed data, one would have a corresponding posterior distribution

$$p(\varepsilon|n) = \frac{L(m|\varepsilon, N)\pi(\varepsilon)}{\int L(n|\varepsilon)\pi(\varepsilon)\, d\varepsilon} \; . \tag{11}$$

The prior $\pi(\varepsilon)$ could, for example, be be taken as uniform in $[0, 1]$; the Jeffreys prior $\pi(\varepsilon) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$ is also often used. A detailed discussion is given in [3].

Once one has obtained the posterior distribution $p(\theta)$ for a parameter θ(e.g., Poisson ν or binomial ε), the upper and lower bounds can be found by requiring

$$\alpha = \int_{-\infty}^{\theta_{\rm lo}} p(\theta)\, d\theta \tag{12}$$

$$\beta = \int_{\theta_{\rm up}}^{\infty} p(\theta)\, d\theta \; . \tag{13}$$

As with the case of frequentist confidence intervals, one would usually take $\alpha = \beta = 0.159$, so that the resulting interval $[\theta_{\rm lo}, \theta_{\rm up}]$ has a probability content of 68.3%.

6 Discussion

Finally, it is interesting to see the practice is for other experiments. The CDF Statistics Committee has also addressed the question of error bars for Poisson data, and the following is extracted from their (public) report [8]:

> The Statistics Committee had been asked by the Spokespersons to make a recommendation about the magnitude of error bars to be shown on histograms in CDF publications. This produced very animated discussions in the Statistics Committee ...
>
> ... it was decided that it is simplest to keep to the traditional practice of using sqrt(n) for the error bars, where n is the observed number of events.

Thus for CDF the overriding argument was the ease with which one could communicate what the error bar meant, even if the message in certain cases (e.g., 0 ± 0, 1 ± 1) is not very informative or even misleading.

References

[1] Glen Cowan, *Burning Issues for the Statistics Forum*, presentation at ATLAS Statistics Forum, 5 October, 2010, `indico.cern.ch/conferenceDisplay.py?confId=109265`.

[2] G. Cowan, *Error analysis for efficiency*, `https://twiki.cern.ch/twiki/pub/AtlasProtected/ATLASStat` see also the related Statistics Forum presentation from 28 July, 2008 at `indico.cern.ch/conferenceDisplay.py?confId=38556`.

[3] Diego Casadei, *Efficiency measurement: a Bayesian approach*, arXiv:0908.0130.

[4] Diego Casadei, *Estimating the trigger efficiency — a tutorial*, presentation at ATLAS Statistics Forum, 5 October, 2010, `indico.cern.ch/conferenceDisplay.py?confId=109265`.

[5] K. Nakamura et al. (Particle Data Group), J. Phys. G 37, 075021 (2010); `pdg.lbl.gov`.

[6] F.E. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, Singapore, 2007.

[7] The ROOT MathCore library, `project-mathlibs.web.cern.ch/project-mathlibs/sw/html/MathCore.html`.

[8] CDF Statistics Committee, *Error Bars for Poisson Data*, `www-cdf.fnal.gov/physics/statistics/notes/pois_eb.txt` (2005).