1. Introduction

Precision measurements of physical quantities typically require a very large number of individual measurements of the same quantity often taken under varying conditions, such as drifting signal-to-noise or many experimental configurations with different signal sizes. For this reason, as well as for simplification of data analysis and reduction of computational requirements, the data are typically binned together such that measurements in the same bin were taken within a time during which the conditions were similar. In order to check whether the binning is susceptible to the varying conditions, as well as to search for unknown sources of noise, a $\chi^2$ test [1–3] is commonly used. Regardless of whether or not it is an ideal choice of statistic for this case, it is fairly intuitive as a measure of whether the assigned error bars are correctly capturing the statistics of the data. However, some of the simplifying assumptions used to construct the standard $\chi^2$ can give results with a significant bias for large data sets. We discuss why the standard treatment underestimates both the mean and variance of the $\chi^2$ statistic, and then determine the appropriate correction factors.

2. Chi-squared test for binned, Gaussian samples

Consider a quantity $N_x \gg 1$ of measurements $x_i$ without any assigned uncertainties. Say that the measurements are normally distributed with constant, true mean $\mu$ that is not known to the experimenter. We shall not assume that the data has a constant variance. Let us gather these data sequentially into groups $G_j$ with $n$ consecutive points each. Now compute the usual sample mean, standard deviation, and standard error of each group of points:

$$y_j = \frac{1}{n} \sum_{x_i \in G_j} x_i, \quad s_j = \sqrt{\frac{1}{n-1} \sum_{x_i \in G_j} (x_i - y_j)^2}, \quad s_{yj} = \frac{1}{\sqrt{n}} s_j.$$  

(2)

We have now binned our data into a smaller set of $N = N_x/n \gg 1$ mean values $y_j$ and uncertainties $s_{yj}$. As a check to see whether the assigned uncertainties are correctly capturing the statistical fluctuations of the data we can perform a $\chi^2$ test as outlined in many standard texts [1–3]. We will test the hypothesis that the $y_j$ are normally distributed about a constant $\bar{y}$ (though this approach is easily extended to...
models with more degrees of freedom), and that the uncertainties correctly describe the statistical fluctuations of the data about the mean. The reduced-$\chi^2$ value of the data set is

$$\chi^2_{\text{red}} = \frac{1}{N-1} \sum_{j=1}^{N} \left( \frac{y_j - \bar{y}}{\sigma_{yj}} \right)^2 \equiv \frac{1}{N-1} \sum_{j=1}^{N} \chi_j^2,$$

where $\bar{y} = (\sum_j y_j/s_{yj}^2)/(\sum_j 1/s_{yj}^2)$ is the weighted mean of the $y$ data, and $\sigma_{yj}$ is the true (unknown) standard deviation of the points $\{x_i \in G_j\}$, which need not be constant over different values of $j$. If the fluctuations in the data are Gaussian in nature, and correctly accounted for by the uncertainties, then we have the usual result

$$E[\chi^2_{\text{red}}] = 1, \quad \text{Std}[\chi^2_{\text{red}}] = \sqrt{\frac{2}{N-1}}.$$

However, the experimenter does not know the true standard deviation, and therefore actually computes the statistic

$$\tilde{\chi}^2_{\text{red}} = \frac{1}{N-1} \sum_{j=1}^{N} \left( \frac{y_j - \bar{y}}{s_{yj}} \right)^2 \equiv \frac{1}{N-1} \sum_{j=1}^{N} \tilde{\chi}_j^2,$$

using $s_{yj}$ as an estimator for $\sigma_{yj}$. We wish to find the statistical properties of this quantity, which we shall find differ from $\chi^2_{\text{red}}$. Intuitively, the sample standard deviation is computed from a finite number of measurements and therefore has some uncertainty associated with it, and that uncertainty should be propagated through when examining the $\tilde{\chi}^2_{\text{red}}$ statistic. This is a well-known effect when estimating parameters from finite data sets and has been previously explored in a number of contexts, for example Poisson distributions, counting experiments, weighted means, and histogram fitting [4–10].

More specifically, while $\chi_j \sim N(0,1)$ is normally distributed, $\tilde{\chi}_j$ is not:

$$\tilde{\chi}_j \equiv \left( \frac{y_j - \bar{y}}{s_{yj}} \right) \approx \left( \frac{y_j - \mu}{s_{yj}} \right) \sim t(n-1),$$

the $t$-distribution with $n-1$ degrees of freedom, which has larger tails for finite $n$ than a normal distribution. Notice that we are treating $\bar{y} = \mu$ as a constant, which is valid in the limit $N \gg 1$, though for smaller $N$ the statistical properties of the weighted mean cannot be ignored [9, 11–15]. In particular, the weighted mean also has correction factors due to the difference between true and sample standard deviation, and has a non-trivial variance, both of which will impact the $\tilde{\chi}^2_{\text{red}}$ statistic. A good discussion of these complexities can be found in [15].

The square of $\tilde{\chi}_j$ is therefore distributed as $\tilde{\chi}^2_j \sim F(1, n-1)$, the $F$-distribution with $(1, n-1)$ degrees of freedom, which has

$$E[F(1, n-1)] = \frac{n-1}{n-3}, \quad \text{Var}[F(1, n-1)] = 2 \left( \frac{n-1}{n-3} \right)^2 \frac{n-2}{n-5}.$$

This is as opposed to the $\chi^2$ statistic, which has (appropriately) a $\chi^2$ distribution. $\tilde{\chi}^2_{\text{red}}$ is therefore distributed as a sum of $F$-distributions, which is complicated [16]. However, the expectation value and variance are straightforward to calculate,

$$E[\tilde{\chi}^2_{\text{red}}] = \frac{N}{N-1} E \left[ \tilde{\chi}^2_j \right] = \frac{n-1}{n-3} + O(N^{-1}),$$

$$\text{Var}[\tilde{\chi}^2_{\text{red}}] = \frac{2}{N-1} \left( \frac{n-1}{n-3} \right)^2 \frac{n-2}{n-5} + O(N^{-2}). \quad (9)$$

This implies that the mean and standard deviation of the $\tilde{\chi}^2_{\text{red}}$ statistic are larger than those of the $\chi^2_{\text{red}}$ statistic by

$$\frac{E[\tilde{\chi}^2_{\text{red}}]}{E[\chi^2_{\text{red}}]} = \frac{n-1}{n-3}, \quad \frac{\text{Std}[\tilde{\chi}^2_{\text{red}}]}{\text{Std}[\chi^2_{\text{red}}]} = \frac{n-1}{n-3} \sqrt{\frac{n-2}{n-5}}. \quad (10)$$

up to further corrections of order $O(N^{-1})$. A plot of these correction factors is shown in figure 1. In the limit $n \to \infty$ we recover the usual result, but for finite $n$ we will always expect larger values for both mean and standard deviation. We can also see that choosing $n \leq 5$ is not advisable, since the statistic will have a non-convergent variance.

3. Conclusion

In summary, we find that the standard $\chi^2$ statistic computed from binning finite data sets underestimates the mean and variance for binned Gaussian samples, and derive simple, closed expressions for the biases. For very large data sets with finite bin sizes, such as those commonly found in precision physics measurements, these corrections can be significant and should not be neglected.
Acknowledgments

I would like to acknowledge helpful discussions with David Watson, and many helpful discussions with the ACME Collaboration, in particular David DeMille, John M Doyle, and Brendon O’Leary.

Appendix. A simple example

We can see how the ‘usual’ chi-squared statistic gives an incorrect result by performing a simple numerical test on some simulated data. Generate 1,000,000 points $x_i \sim \mathcal{N}(0, 1)$, bin into groups of $n = 10$, and then compute means $y_j$, standard errors $\sigma_y$, and the reduced chi-squared statistic $\tilde{\chi}_2$ (as described in the main text) for the resulting 100,000 binned points.

```matlab
Nx = 1000000
nbin = 10
for j = 1:(Nx/nbin)
    x = randn(1,nbin)
    y(j) = mean(x)
    sy(j) = std(x)/sqrt(nbin)
end
ybar = sum(y./sy.^2)/sum(1./sy.^2)
chi = (y-ybar)./sy
chi2 = sum(chi.^2)
dof = length(y)-1
redchi2 = chi2/dof
redchi2sigma = sqrt(2/dof)
```

If we run this piece of code, we will find $\text{redchi2} = 1.2868$ and $\text{redchi2sigma} = 0.0045$ (though of course the former will be different each time due to the random nature of the calculation.) This value differs considerably from the naïve expectation of $1 \pm 0.0045$ based on the usual treatment that ignores the difference between sample and true standard deviations, but is quite close to the expected value of $1.2857 \pm 0.0073$ from equations (8) and (9).

ORCID iDs

Nicholas R Hutzler © https://orcid.org/0000-0002-5203-3635

References

[1] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 2007 Numer Recipes 3rd edn (Cambridge: Cambridge University Press)
[2] Bevington P R and Robinson D K 2003 Data Reduction and Error Analysis for the Physical Sciences 3rd edn (Boston, MA: McGraw-Hill)
[3] Taylor J R 1996 An Introduction to Error Analysis 2nd edn (Mill Valley, CA: University Science Books)
[4] Baker S and Cousins R D 1984 Nucl. Instrum. Methods Phys. Res. 221 437–42
[5] Jading Y and Riisager K 1996 Nucl. Instrum. Methods Phys. Res. A 372 289–92
[6] Hammersley A and Antoniadis A 1997 Nucl. Instrum. Methods Phys. Res. A 394 219–24
[7] Mighell K J 1999 Astrophys. J. 518 380–93
[8] Hauschild T and Jentschel M 2001 Nucl. Instrum. Methods Phys. Res. A 457 384–401
[9] Zhang N F 2006 Metrologia 43 195–204
[10] Gagunashvili N 2010 Nucl. Instrum. Methods Phys. Res. A 614 287–96
[11] Hutzler N R 2014 A new limit on the electron electric dipole moment PhD Thesis Harvard University
[12] Cochran W G 1937 Suppl. J. R. Stat. Soc. 4 102–18
[13] Graybill F A and Deal R B 1959 Biometrics 15 543–50
[14] Böckenhoff A and Hartung J 1998 Biometrical J. 40 937–47
[15] Hartung J, Knapp G and Sinha B K 2008 Statistical Meta-Analysis with Applications (New York: Wiley)
[16] Morrison D F 1971 J. Am. Stat. Assoc. 66 383