HERA Memo #123
Received 31 May 2023

LST-Binning Statistics

Steven G. Murray, Josh Dillon, and the HERA Analysis Team

May 31, 2023

Abstract

Averaging nominally redundant visibilities together (redundant either in baseline vector or LST) is an important step in HERA's analysis. While averaging these visibilities together, one can also compute estimates of higher order statistics, in particular the variance. Here we address the question of what the theoretical *distribution* of the measured sample variance should look like under certain conditions, which makes it possible to perform more insightful tests of whether observations exhibit "excess" variance with respect to the prediction.

1 Non-redundantly-averaged LST-binned visibilities

Let us assume a model of visibilities in which any particular baseline, at any particular LST and frequency, has a complex visibility drawn from a complex Gaussian distribution with mean $\mu(b, \mathrm{LST}, \nu)$ and variance $\sigma^2_{\mathrm{C}}(b, \mathrm{LST}, \nu)$. For the rest of this section, we assume baselines, LSTs and frequencies to be independent, and therefore drop their notation. Furthermore, we assume that the real and imaginary components of the visibility are iid, and both drawn from a real-valued Gaussian distribution with variance $\sigma^2 = \sigma^2_{\mathrm{C}}/2$.

In the case that the visibilities are not redundantly-averaged, the unweighted sample mean visibility over nights (in the same LST bin) is an unbiased, minimum-variance estimator of the mean, μ:

$$\bar{V} = \hat{\mu} = \frac{1}{N_d} \sum_{j=1}^{N_d} V_d. \tag{1}$$

Here, N_d is the number of *unflagged* nights/days observed for that LST bin, frequency and baseline. An estimate of σ^2 may be determined in the usual way, via the sample variance:

$$S^2 = \hat{\sigma^2} = \frac{1}{N_d} \sum_{j=1}^{N_d} (V_j - \bar{V})^2, \tag{2}$$

which is itself a random variable. We can easily find the expectation of the sample variance:

$$\langle S^2 \rangle = \frac{1}{N_d} \sum_j \langle V_j^2 \rangle - 2\langle V_j \bar{V} \rangle + \langle \bar{V}^2 \rangle, \tag{3}$$

since $\mathrm{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2$, we have

$$\begin{aligned}
\langle S^2 \rangle &= \frac{1}{N_d} \sum_j \sigma^2 + \mu^2 - \frac{2}{N_d} \left[\langle V_j^2 \rangle + \sum_{k \neq j} \langle V_j V_k \rangle \right] + \mathrm{Var}(\bar{V}) + \mu^2, & (4) \\
&= \frac{1}{N_d} \sum_j \sigma^2 + \mu^2 - \frac{2}{N_d} \left[\sigma^2 + \mu^2 + (N_d - 1)\mu^2 \right] + \frac{\sigma^2}{N_d} + \mu^2, & (5) \\
&= \frac{1}{N_d} \sum_j \sigma^2 - \frac{\sigma^2}{N_d} = \sigma^2 \left(\frac{N_d - 1}{N_d} \right). & (6)
\end{aligned}$$

In practice, the sample variance for a particular baseline/LST/frequency is itself a random variable. We'd like to know the distribution of $\gamma = S^2/\langle S^2 \rangle = S^2 N_d/(\sigma^2(N_d-1))$ – a quantity we'll call the "excess variance", which measures the deviation of measured/sample variance from the expectation. This can help understand whether measurements are conforming with expectation. Note that since the normalization of γ (i.e. the expected variance) is not a random variate, we can write

$$\gamma = \frac{1}{N_d}\sum_{j=1}^{N_d}(X_j - \bar{X}_j)^2, \tag{7}$$

with $X = V\sqrt{\frac{N_d}{\sigma(N_d-1)}}$. This transforms X into a Gaussian random variable with mean zero and variance $\sigma_X^2 = N_d/(N_d-1)$.

According to MathWorld[1], the distribution of the sample variance for N variates with intrinsic variance σ^2 is a Pearson Type III distribution:

$$f(S^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(S^2)^{\alpha-1}e^{-\beta\gamma}, \tag{8}$$

with $\beta = N/(2\sigma^2)$ and $\alpha = (N-1)/2$. In the case of our γ then, we can substitute σ_X^2 to determine $\alpha = \beta = (N_d-1)/2$, i.e.

$$f(\gamma) = \frac{\beta^\beta}{\Gamma(\beta)}\gamma^{\beta-1}e^{-\beta\gamma}. \tag{9}$$

Note that this is *not* in general a χ^2-distribution, as it has more flexibility in the exponential than the χ^2. It is in fact what you get by drawing samples from a $chi^2_{N_d-1}$ distribution, then dividing each variate by N_d-1, and is a special form of the Gamma-distribution where the shape and rate parameters are equal, and given by β[2]. This distribution has the properties

$$\mu = 1, \tag{10}$$
$$\text{mode} = 1 - 2/(N_d-1) \leq \mu, \tag{11}$$
$$\sigma_\gamma^2 = 2/(N_d-1), \tag{12}$$
$$\text{skew} = \sqrt{8/(N_d-1)}, \tag{13}$$
$$\text{kurt} = 12/(N_d-1). \tag{14}$$

We might also ask what the distribution of an *average* of M γ variates is (e.g. an average over baselines or times or frequencies). According to answers on Math StackExchange[3], the sum of gamma variables with identical scale/rate parameters is another gamma variable with the same rate, and a shape parameter that is the sum of shape parameters of the variates being summed. Here, since we are taking the average, we must modify the scale parameter by a factor of M, and note that the final shape parameter will just be $M\alpha \equiv M(N_d-1)/2$.

2 Redundantly-Averaged Visibilities

Now, let us extend the results of the previous section to the case in which visibilities have been redundantly averaged *before* LST-binning. We thus deal with not a particular baseline, but a baseline *type*. This baseline type may have $N_{\rm blg}$ baselines in it, but not all of them are unflagged for a given LST/frequency on any given night. This means that the theoretical variance for the visibility on any given night can change, depending on the number of baselines unflagged: $\sigma_j^2 = \sigma^2/N_{\rm bl,j}$, $j \in (1, N_{\rm nights})$.

[1] https://mathworld.wolfram.com/SampleVarianceDistribution.html

[2] It turns out that using the gamma distribution within `scipy` is much easier than using the Pearson Type III distribution, which appears to have a different parameterization than the one we use here.

[3] https://math.stackexchange.com/questions/250059/sum-of-independent-gamma-distributions-is-a-gamma-distribution

Now, the minimum-variance unbiased estimator of the mean visibility (assuming perfect LST- and baseline-redundancy) is

$$\bar{V} = \hat{\mu} = \bar{w} \sum_j V_j/\sigma_j^2, \qquad \bar{w} = \left[\sum_j \sigma_j^{-2} \right]^{-1}, \tag{15}$$

which has the theoretical variance $\mathrm{Var}(\bar{V}) = \bar{w}$. We also form a weighted sample variance:

$$S^2 = \bar{w} \sum_j (V_j - \bar{V})^2/\sigma_j^2, \tag{16}$$

for which each term has the expectation

$$\langle S_j^2 \rangle \sigma_j^2 = \langle V_j^2 \rangle - 2\langle V_j \bar{V} \rangle + \langle \bar{V}^2 \rangle, \tag{17}$$

$$= \sigma_j^2 + \mu^2 - 2\bar{w} \left[\langle V_j^2 \rangle / \sigma_j^2 + \sum_{k \neq j} \langle V_j V_k \rangle / \sigma_k^2 \right] + \mathrm{Var}(\bar{V}) + \mu^2, \tag{18}$$

$$= \sigma_j^2 + \mu^2 - 2\bar{w} \left[1 + \mu^2/\sigma_j^2 + \sum_{k \neq j} \mu^2/\sigma_k^2 \right] + \bar{w} + \mu^2, \tag{19}$$

$$= \sigma_j^2 - \bar{w} \tag{20}$$

and so the whole has expectation

$$\langle S^2 \rangle = \bar{w} \left[N_d - \bar{w} \sum_j \sigma_j^{-2} \right] = \bar{w}(N_d - 1). \tag{21}$$

Note that if each night has the same number of samples (i.e. the same variance), we have $\bar{w} = \sigma^2/N_d$, and the expectation of the sample variance reduces to the same form we derived in the previous section.

Now, we can ask what the distribution of $\gamma = S^2/\langle S^2 \rangle = S^2/(\bar{w}(N_d - 1))$ is (i.e. the distribution of 'excess variance'), as in the previous section. It turns out to be *exactly the same distribution* as in the un-weighted case. This can be shown by a coordinate transformation[4] via an orthoganal matrix. The proof is along the following lines.

Take ξ to be the non-normalized weighted variance, i.e.

$$\xi = \sum_j^{N_d} (V_i - \bar{V})^2/\sigma_i^2. \tag{22}$$

This expands to

$$\xi = \sum_{i=1}^{N_d} Y_i^2 - \bar{V}^2/\bar{w}, \tag{23}$$

where Y_i are independent standard normal variables. Now, the following statement holds, that

$$\sum_{i=1}^{N} Y_i^2 - P_1^2 \sim \chi_{n-1}^2, \tag{24}$$

if Y_i are iid standard normal, and $P = QY$ is a matrix transformation of Y where Q is an orthogonal matrix. This holds because $\sum Y_i^2 \equiv \sum V_i^2$. In our case, $\bar{V}^2/\bar{w}$ can be written as the first row in a matrix $P = QY$, where the first row of Q is $\sqrt{W}\{1/\sigma_1, 1/\sigma_2, \cdots, 1/\sigma_N\}$. Thus, the distribution of ξ is χ_{n-1}^s. To get the distribution of $\gamma = \xi/(n-1)$ merely involves doing a linear transform of variables, which gives back Eq. 9.

This is a little counter-intuitive, because you'd expect that the distribution should have some memory of the relative weights of the input samples, but it depends only on N_d – the number of nights that aren't completely flagged. That is, weights of *zero* mean something, but any other weight, however infinitesimally above zero, makes no difference.

[4] https://math.stackexchange.com/questions/3563361/find-the-distribution-of-xi

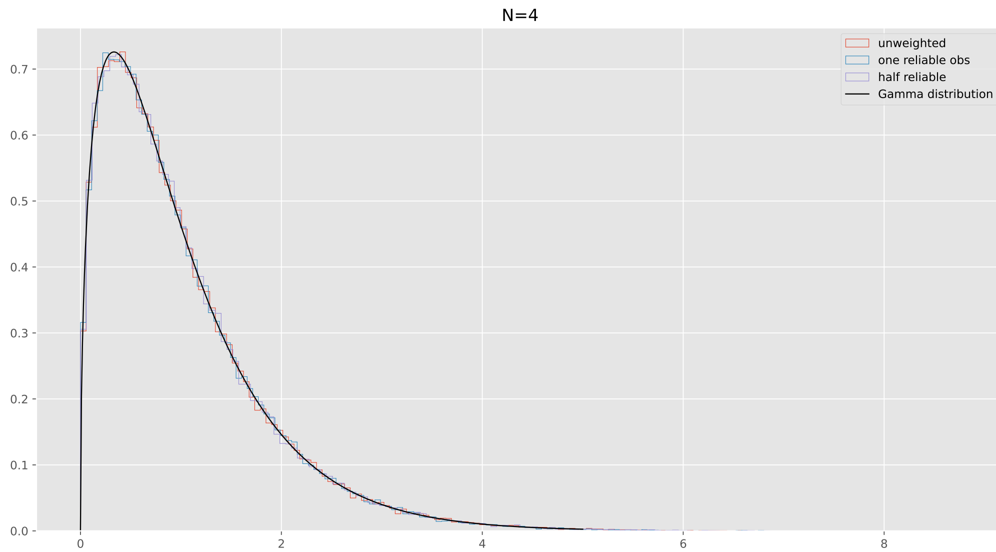

Figure 1: Distribution of γ for $N = 4$.

3 Simulations

Setup the following Python code to create a simulation:

```python
def simulate(variance: np.ndarray, size=100000):
    nd = len(variance)
    w = 1/np.sum(1/variance)

    Z = np.random.normal(size=(size, nd))
    first_term = np.sum(Z**2, axis=1)
    second_term = np.sum(Z / np.sqrt(variance), axis=1)**2 * w
    return (first_term - second_term) / (nd - 1)
```

Then we can create plots of the distribution for different variance vectors, (i) including all ones, (ii) a single very low variance, and (iii) half-low half-high variances.

4 Conclusion

Somewhat counter-intuitively, we find that the distribution of "excess variance" (that is, the ratio of measured sample variance to expected variance based on e.g. auto-correlations) follows a special Gamma distribution (equivalently, a re-scaled χ^2-distribution), Eq. 9, that depends *only* on the number of unflagged samples (and not the individual variances of the samples). While counter-intuitive, this result is borne out by simulation as well as theoretical derivation.

The following points are potentially useful caveats/extensions of this work:

1. Here we have assumed we *know* the 'true' variance of the samples when we perform the averaging and sample variance calculation. If the weights applied to the averaging are biased, or represent *estimates* of the variance (both of which are invariably true in practice), the distributions derived here will be invalid (in proportion to the precision and accuracy of the estimate). In particular, we *suspect* (but have not proven) that inflating the prior on σ_j^2 will significantly affect the distribution for small N_d.

2. In practice, two forms of weighting can be used. The first is to weight directly by $N_{\rm samples,j} \equiv N_{\rm red,j}$. For this to be consistent with the derivation in this memo, one must assume both that the intrinsic

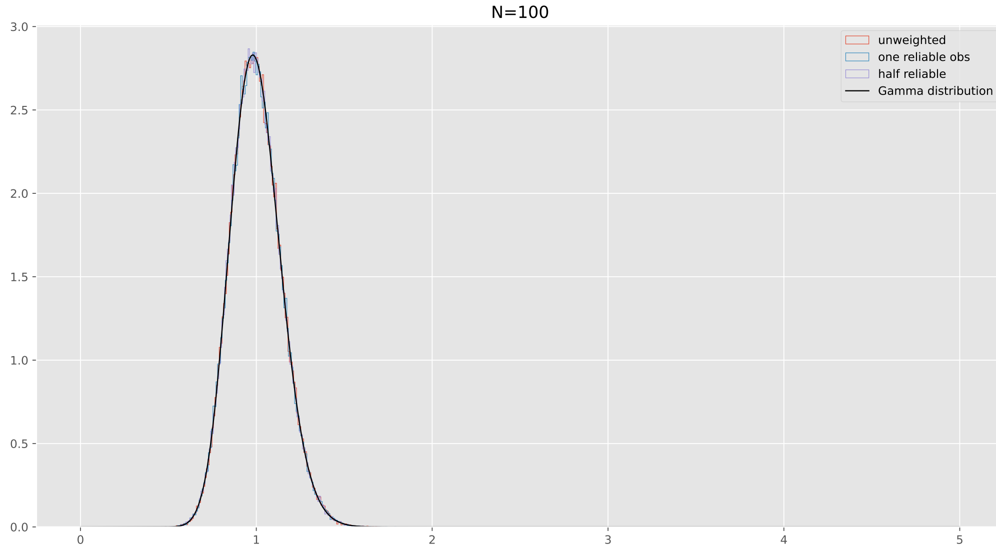

Figure 2: Distribution of γ for $N = 100$.

variance of a redundantly averaged visibility on a particular night j is given by $\sigma^2_{j,\rm bl} = \sigma^2_{\rm bl}/N_{{\rm red},j}$, *and* that one can get a very precise and accurate estimate of $\sigma^2_{\rm bl}$ (e.g. via the averaged autocorrelations of the constituent antennas). The second is to weight by $\sigma^2_{\rm j,bl}$ directly, which accounts for potential night-to-night deviations of the variance, but at the expense of potentially introducing spectral structure via systematics in the autocorrelations. It is our recommendation to use the first approach (weighting by $N_{\rm samples}$), and treating observed deviations in the auto-correlations night-to-night as indications of *bad things happening* rather than trying to inverse-variance-weight your way out of them.