NASA ExoPAG Study Analysis Group 13
Exoplanet Occurrence Rates and Distributions

Proposed standards for occurrence rate units and binning
v5, last update 6/2/2016

Belikov, Ruslan (Chair, rulsan.belikov@nasa.gov)
Stark, Christopher (Co-chair)
Batalha, Natalie (Steering Committee)
Burke, Chris (Steering Committee)

Angerhausen, Daniel	Gaudi, Scott	Savransky, Dmitry
Apai, Daniel	Ge, Jian	Serabyn, Gene
Bendek, Eduardo	Gould, Andy	Shabram, Megan
Bennett, David	Hogg, David W	Shao, Mike
Blackwood, Gary	Howard, Andrew	Solmaz, Arif
Boss, Alan	Kane, Stephen	Sparks, William
Brown, Robert	Kasting, Jim	Stahl, Philip
Bryden, Geoff	Kopparapu, Ravi	Stapelfeldt, Karl
Cahoy, Kerri	Macintosh, Bruce	Stark, Christopher
Catanzarite, Joe	Mandell, Avi	Still, Martin
Ciardi, David	Mendez, Abel	Suzuki, Daisuke
Clanton, Christian	Meyer, Michael	Swain, Mark
Cowan, Nick	Morgan, Rhonda	Traub, Wes
Danchi, William	Mulders, Gijs	Turnbull, Margaret
Domagal-Goldman, Shawn	Nielsen, Eric	Unwin, Stephen
Dressing, Courtney	Petigura, Erik	Vanderbei, Bob
Farr, Will	Ragozzine, Darin	Walkowicz, Luzianne
Foreman-Mackey, Daniel	Roberge, Aki	Weiss, Lauren M.
Fressin, Francois	Rogers, Leslie	Wolfgang, Angie
		Youdin, Andrew

Charter:
Over 5000 exoplanets and exoplanet candidates have been discovered to date. Many studies have been published and are on-going to determine exoplanet occurrence rates and distributions, particularly for potentially habitable worlds. These studies employ different statistical and debiasing methods, different definitions of terms such as eta_Earth and habitable zone, different degrees of extrapolation, and present distributions in different units from each other. The primary goal of this SAG is to evaluate what we currently know about planet occurrence rates, and especially eta_Earth, by consolidating, comparing, and reconciling discrepancies between different studies. A secondary goal is to establish a standard set of occurrence rates accepted by as much of our community as possible to be used for mission yield estimates for missions to be considered by the decadal survey.

Key objectives and questions:

1. Propose standard nominal conventions, definitions, and units for occurrence rates/distributions to facilitate comparisons between different studies.

2. Do occurrence estimates from different teams/methods agree with each other to within statistical uncertainty? If not, why?
3. For occurrence rates where extrapolation is still necessary, what values should the community adopt as standard conventions for mission yield estimates?

1. PURPOSE OF THIS DOCUMENT

This document proposes a standard definition for exoplanet occurrence rate units and bin sizes. The goals are: (a) to facilitate comparisons between different studies and estimates by different teams (see SAG13 key objective 1 above) in order to catalyze convergence towards consistent estimates and understanding of discrepancies; (b) help provide a standard interface between scientists who produce occurrence rate distributions and mission designers who use them.

This is the first step in a broader SAG13 effort and for now we are emphasizing simplicity and focusing on Kepler data. However, the intent is to establish the definitions in a way that has a potential path towards generalizations to other data sets (e.g. RV measurements) and conversions to values that would be useful to mission planning, especially direct imaging missions. For example, we use period instead of semi-major axis and planet size instead of msin(i) in this document, but a future SAG13 activity will address how to convert between these variables or define a second set of occurrence rates with respect to those other variables. This is outside the scope of the present document, however.

The definition below has emerged as a result of a SAG13 process (see Appendix) to build a consensus and is endorsed by all the SAG members on the first page.

2. DEFINITION OF STANDARD OCCURRENCE RATE BINS

We propose a 3-dimensional set of planet occurrence rates $\{\eta_{i,j,k}\}$ diagrammed in Figure 1. Each $\eta_{i,j,k}$ is defined as the integral of the number density of planets per star across the *i, j, k'*th bin, where *i*,*j*,*k* are indexes in planet radius, period, and stellar temperature as shown in Figure 1 and described in more detail below. These values are assumed to be debiased to account for all known measurement biases and pipeline completeness.

2.1 Planet radius binning.

The *i*-th bin in planet radius is defined as the interval

$$\mathcal{R}_i = [1.5^{i-2}, 1.5^{i-1})R_\oplus$$

This implies the following bin edges (rounded to 2 significant figures and in units of $R_\oplus$):

$$\{0.67, 1.0, 1.5, 2.3, 3.4, 5.1, 7.6, 11, 17, \dots\}$$

2.2 Planet period binning.

The *j*-th bin in planet period is defined as (in units of days):

$$\mathcal{P}_j = 10\times[2^{j-1}, 2^j)$$

This implies the following bin edges (in units of days):

$$\{10, 20, 40, 80, 160, 320, 640, \dots\}$$

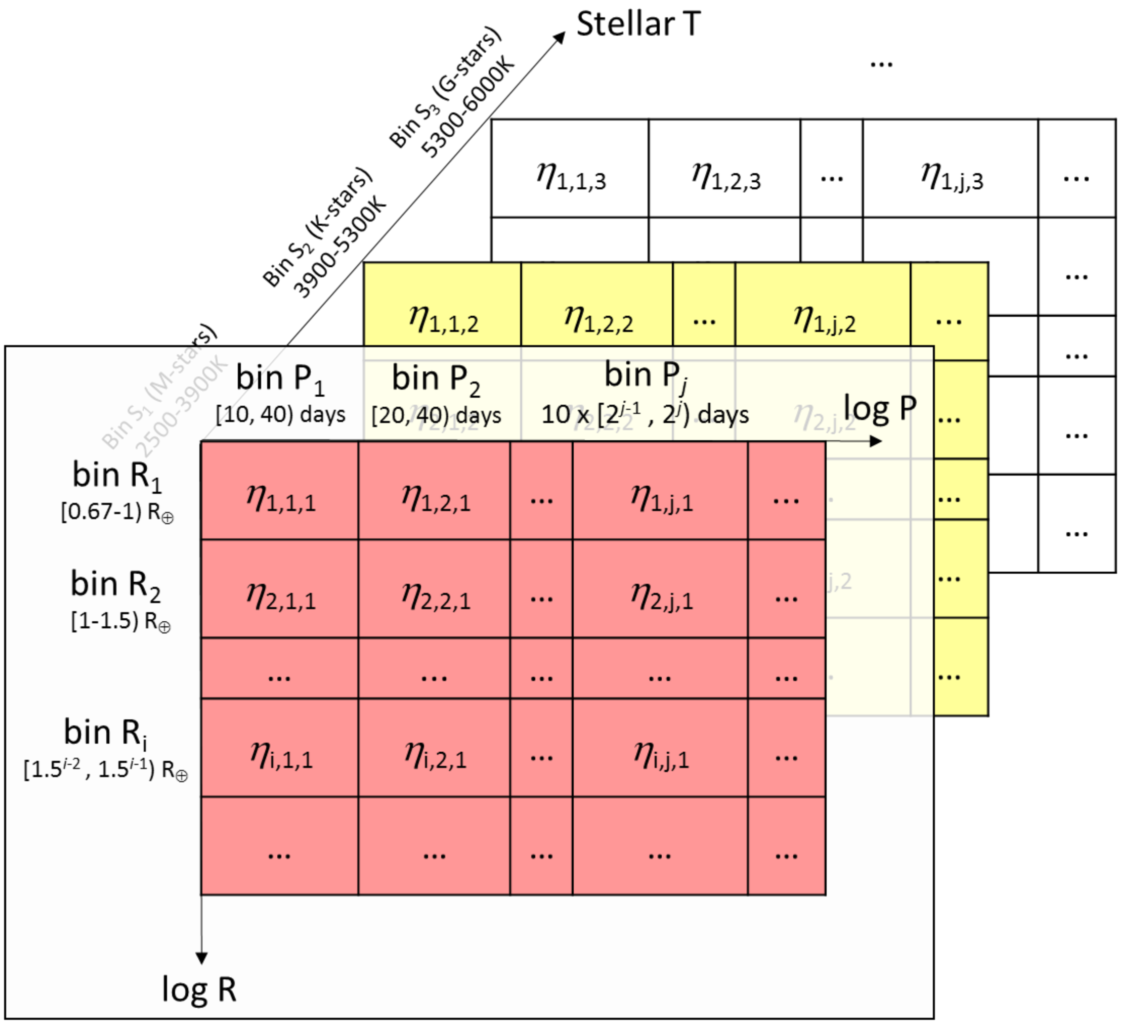

Figure 1. Proposed standardization of occurrence rate bins

2.3 Stellar Temperature binning

The stellar temperature bins are defined as follows:

Bin index *k*	Stellar temperature range	Corresponding spectral type
1	[2400, 3900)	M
2	[3900, 5300)	K
3	[5300, 6000)	G
4	[6000, 7300)	F
5	[7300, 10000)	A

[Note: this is based on table 5 from Pecaut & Mamajeck(2013), ApJ, 208, 1, http://iopscience.iop.org/article/10.1088/0067-0049/208/1/9/meta;jsessionid=698F3A9F5272B070DC62876C1764BFDB.c1#apjs480616s3]

3. DATAFILE FORMAT

The SAG also recommends the following data format conventions for machine-readable data tables to facilitate automatic processing and comparing of occurrence rate data:

- ASCII file
- tab (or any white space) delimited, floating-point values in scientific notation
- use "NaN" when you cannot or do not wish to provide a particular $\eta_{i,j,k}$ (e.g. due to incompleteness)
- each line spans index *j* (log period bins) for a fixed *i.* and *k*
- lines span index *i* (log period planet radius)
- double carriage return separating blocks of data corresponding to different *k* (stellar T) bins.
- Any supplementary information can be included in a header with "%" in front of each line

The resulting data file should look like this:

% [header]\n

[$\eta_{1,1,1}$]\t[$\eta_{1,2,1}$]\t[…]\n
[$\eta_{2,1,1}$]\t[$\eta_{2,2,1}$]\t[…]\n
[…]
\n
[$\eta_{1,1,2}$]\t[$\eta_{1,2,2}$]\t[…]\n
[$\eta_{2,1,2}$]\t[$\eta_{2,2,2}$]\t[…]\n
[…]
\n
[…]

Whenever providing a datafile of occurrence values, if possible please also provide 2 files corresponding to 1-sigma positive and negative uncertainties for each $\eta_{i,j,k}$, following the format above. (Just the sigma values, not "eta+sigma".) Filenames should start with "eta", "sigma_p", "sigma_n".

4. README FILE

If possible, please include a standard readme file with each your datasets, with the following fields:

Planet Catalog: [e.g. Kepler Q1 – Q17, SOC 9.2 DR24]
Filters: [e.g. remove MES < 15, logg < 4]
Completeness Model: […]
Vetting efficiency: […]
Reliability: […]
Methodology: […]
i0, j0, k0: [indexes of the first element in the datafiles $\eta_{i0,j0,k0}$. If this field is omitted, they are assumed to all be 1]
Any other information that you think is important.

For an example of such a file, please see https://drive.google.com/drive/folders/0B_xHopbVhE_-MmI5UkYxRkJOVk0

5. OTHER PROPOSED CONVENTIONS

We do not constrain any parameter, estimation or debiasing method, or other choice that is not explicitly specified above, but the general rule is that any unusual choice or circumstance that may affect results should be noted in the header whenever supplying values for $\{\eta_{i,j,k}\}$. There are some specific cases worth noting and we propose conventions for them below:

5.1 Choosing the maximum value of period, radius, and star temperature (i.e. farthest bin), and extrapolation.

The SAG standard above does not specify the maximum period, radius, and star temperature that should be considered (i.e. maximal values of i, j, k). That choice should be constrained by the data, but we encourage estimating all the values (and properly debiasing them) for which there is some data, even if the statistical error bars are high (this is ok as long as uncertainty files are produced). Actual extrapolation is not encouraged, but can be done *as long as all extrapolated values are flagged as such*. (This can be done by a verbal description in the header or a file of flags (0s and 1s) in the same format as in section 3.)

5.2 Minimum values of period, radius, and star temperature.

By default, the minimal values are implicit because the indexes i,j,k all start at 1. Starting an index at a value other than 1 (e.g. planet radii < 0.67) is possible, but in that case the starting value of the index needs to be specified in the readme file (see section 4). Conventions for extrapolating in this case are same as in subsection 5.1.

5.3 Other parameters and dimensions

For simplicity we are ignoring several potentially important dimensions, such as stellar multiplicity, metallicity, galactic latitude, planet eccentricity, and many others. It is understood that there may be structure in the occurrence rates across these parameters, and that if the dataset is biased in any of these dimensions, this may bias occurrence rates. The default choice is marginalizing over the entire range of all those dimensions in the dataset, and any other choice should be noted in the readme file.

6. APPENDIX

6.1 Background information

This standard emerged as a result of 4 1-hour teleconference meetings of the SAG (complete meeting notes available upon request), several online and offline discussions, and an initial poll which captured different alternatives and suggestions that were considered and helped reach a constrained consensus. The results of the poll can be seen here:

https://docs.google.com/forms/d/14dBTg7hHmqxvfwoNXfsTENQ4afIUtBxfw4WBt3WtR78/viewanalytics