Memo for internal use
Preprint typeset using LaTeX style emulateapj v. 12/16/11

TOY TRANSIT SURVEYS

LGB, KM, JNW
Memo for internal use

ABSTRACT

What errors does ignoring binarity introduce on the occurrence rates derived from transit surveys? First, we show that in a universe where all stars have fixed properties (except that some are twin binaries), the results of a search for planets of a given radius and orbital period can be expressed analytically.

We then allow the properties of the secondary in binary systems to vary. A distribution of light ratios introduces a detection bias towards equal-mass binaries in magnitude limited samples. We show that dependent upon planet occurrence's functional dependence on host star mass, as well as the specifics of the assumed survey completeness, the occurrence rate correction factor $X_\Gamma \equiv \Gamma_{\rm true}/\Gamma_{\rm obs}$ ranges from $0.6 - 1.6$.

Finally, we perform a synthetic transit survey analogous to *Kepler*. We show that if we assume every star in this *Kepler* survey analog is single, we underestimate occurrence rates for a given planet radius and period by a factor of ≈ 1.4, or a percentage error relative to the true occurrence rate Γ of $\delta\Gamma = -28\%$. This is relevant for measuring accurate values of the rate at which Earth-like planets exist around Sun-like stars.

Subject headings: planets and satellites: detection

1. INTRODUCTION

Ah, transiting planets! We learn so much by studying them. But how much can we actually learn, and how much is messed up by binarity?

Binary companions to single stars introduce a surprising number of biases into transit surveys. For a start, any magnitude limited survey over-represents binaries compared to a volume-limited survey – the familiar Malmquist bias, though in a slightly different context. While transit surveys are ideally signal to noise limited, if the noise properties of the survey are dominated by photon counting noise, a signal to noise limit becomes equivalent to a magnitude limit. Thus stars surveyed in transit surveys must also over-represent binaries compared to those in a volume-limited sample.

A more straight-forward effect of binarity on transit surveys is that any star which is unknowingly a double introduces extra stars about which planets might be detectable. However, this apparent increase in the number of searchable stars is mitigated by dilution: the observed transit depths in such systems are reduced by contaminating flux from the binary companion. This dilution is especially drastic if the planet orbits the secondary, in which case dilution impacts detectability the most.

A related effect to dilution is misinterpretation of transit depths. If the planet to star radius ratio is directly taken as the square-root of the observed transit depth, derived planetary radii in binary systems will systematically too small. Ciardi et al [2015] explored this issue in depth. They showed that for the case of *Kepler*, assuming host stars to always be single will on average underestimate radii by a multiplicative factor of ≈ 1.5.

Finally, binarity may affect transit surveys through actual astrophysics. For the relatively rare case [Duquennoy & Mayor 1991, Raghavan et al 2010] of close binaries, dynamical stability imposes a limited range of periods in which planets may orbit [e.g., Holman & Weigert 1999, Li et al 2017], the effects of which have been observed [Kraus et al 2016]. Also, the formation rates of planets, and thus their intrinsic occurrence rates, may be influenced by properties of circumprimary and/or circumbinary disks. Although the observations are still ongoing, early results seem to indicate that once binary companions are $\gtrsim 100\,$AU apart, binarity does not decrease planet occurrence rates [CITE], and may even increase them for giant planets [Ngo et al 2017].

2. MODEL #1: FIXED STARS, FIXED PLANETS, FIXED LIGHT-RATIO BINARIES

Binarity introduces several convoluted effects on transit surveys. To disentangle them, imagine the following idealized survey:

1. You are going to observe the entire sky for a duration $T_{\rm obs}$, with a detector of area A, and known bandpass. Your detector is photon-noise limited.

2. You are interested only in detecting planets of radius R_p, and orbital period P. For instance, $R_p = R_\oplus$, $P = 1$ year.

3. You are only interested in detecting them around stars of radius R_1, and luminosity L_1. For instance, G2V dwarfs.

4. You can only detect your 'nominal planet' when you observe signals with S/N > $({\rm S/N})_{\rm min}$. For a photon-noise limited survey, this minimum signal to noise ratio is equivalent to a minimum flux, $F_{\rm min}$.

5. However, you opt to observe all the "points" on-sky with apparent magnitude $m < m_{\rm lim}$, or equivalently with energy flux in your bandpass greater than some limit $F > F_{\rm lim}$. (You will not be able to detect planets around the faintest of these, and will need to perform a completeness correction to later

derive occurrence rates.) Generally $F_{\rm lim} < F_{\rm min}$, *e.g.*, whenever the original magnitude cut defines a "source catalog" from which "searchable stars" are selected, as in the KIC [Batalha et al, 2010, Brown et al 2011] and the TIC [Stassun et al 2017].

You get funding, and perform the S/N-limited survey. All planets with $\mathrm{S/N} > (\mathrm{S/N})_{\rm min}$ are detected. The rest are not. You now wish to derive an occurrence rate for planets of radius R_p and orbital period P. Assume your universe is a universe in which:

- The true population of "points" (stellar systems, all unresolved) comprises both single and double star systems. Single star systems have luminosity in the observed bandpass L_1, radii R_1, and effective temperature $T_{\rm eff,1}$. Double star systems have luminosity in the observed bandpass $L_d = (1+\gamma_R)L_1$, for $\gamma_R = L_2/L_1$ the ratio of the luminosity of the secondary to the primary. In this section, γ_R is a constant across the population of star systems. The ratio of the two number densities in a volume-limited sample is the binary fraction[1].

- The true population of planets around these stars is as follows:

  A fraction $\Gamma_{t,s}$ of stars in single star systems have a planet of radius R_p, with orbital period P.

  A fraction $\Gamma_{t,d}$ of each star in a double star system has a planet of radius R_p, with orbital period P. For instance, if $\Gamma_{t,s} = \Gamma_{t,d} = 0.1$, on average each double system contributes 0.2 planets, and each single system 0.1 planets. Any astrophysical difference in planet formation between singles and binaries is captured by these two terms. Note that we are taking an limit by assuming the primary and secondary of a binary system host planets at equal rates. If the secondary is non-identical to the primary, it's clearly wrong.

For simplicity, one might wish to imagine the observer correctly pre-selects all of the "searchable stars", in the style of [Pepper et al, 2003]. In other words, they would begin by knowing $(\mathrm{S/N})_{\rm min}$, and then only observe stars for which they would have adequate flux to achieve it – setting $F_{\rm lim} = F_{\rm min}$. This simplification would naively allow the observer to ignore completeness effects when deriving occurrence rates, since their completeness is 100%.

We do not begin this way because in a stellar population of both singles and binaries, the above approach ignores necessary incompleteness for binary systems. If we set the limiting flux to give 100% completeness for single stars, we will unwittingly select binaries near that flux limit, for which dilution will push transit signals below the detection threshold. This is because a flux limit maps onto three different maximum detection distances for a population of single stars and twin binaries:

1. $d_{{\rm max},s}$: the maximum distance out to which single stars are selected. This can be chosen to coincide with the maximum distance out to which planets are detectable around single stars.

2. $d^\star_{{\rm max},d}$: the maximum distance out to which double stars are selected. This is different from $d^{\rm p}_{{\rm max},d}$, the maximum distance out to which planets are detectable around double stars. For a population with fixed γ_R, since $\mathrm{S/N} \propto \mathcal{D}L^{-1/2}d$,

$$d^{\rm p}_{{\rm max},d} = (1+\gamma_R)^{-1/2} d_{{\rm max},s} = (1+\gamma_R)^{-1} d^\star_{{\rm max},d}. \quad (1)$$

   For $\gamma_R = 1$, this means only 1 in 8 selected binary stars can yield detectable planets.

We develop tools to address this and related completeness effects from the outset.

Consider the following questions:

1. How many single and double star systems, respectively, are in the sample? Correspondingly, how many stars are in the sample?

2. How many planets are in the sample? (Orbiting single stars, and orbiting double stars respectively).

3. What is the true occurrence rate?

4. How many planets are detected?

5. What occurrence rate does astronomer A, who has never heard of binary star systems, or completeness, derive for planets of radius R_p and period P?

6. What occurrence rate does astronomer B, who accounts for the "2 for 1" effect of binarity (*i.e.* that the sample actually has more stars than astronomer A thought) derive? (Astronomer B neglects completeness).

7. What about astronomer C, who accounts for "2 for 1" *and* misclassification due to diluted radii? In other words, astronomer C did a combination of high resolution imaging and RV followup on every candidate, and correctly classifies the planetary radii in every case.

8. What about astronomer D, who additionally notes the importance of completeness?

2.1. *How many stars are in the sample?*

Let N_s be the number of single star systems, and N_d the number of double star systems. Then the total number of stars in the sample is

$$N_{\rm stars} = N_s + 2N_d. \quad (2)$$

In a magnitude-limited sample in which stars are uniformly distributed in volume, the number of stars will be the number density times the volume. If the volume is taken to be a sphere over which the number density is uniform,

$$N_i = n_i \frac{4\pi}{3} d^3_{{\rm max},i}, \quad (3)$$

[1] The binary fraction is equivalent to the multiplicity fraction if there are no triple, quadruple, ... systems.

for $i \in \{\text{single}, \text{double}\} \equiv \{s, d\}$, and

$$\frac{n_d}{n_s} = \text{binary fraction} \equiv \text{BF} \tag{4}$$

by definition. The absolute normalization of the number density is a measured quantity, as is the binary fraction. For G2V dwarfs, the latter is ≈ 0.45 [Duchene & Kraus, 2013]. The former is given by [Bovy 2017].

$d_{\max,i}$ in Eq. 3 is the maximum distance corresponding to the given magnitude limit:

$$d_{\max,i} = \left(\frac{L_i}{4\pi F_{\min}}\right)^{1/2}, \tag{5}$$

where the limiting flux in the bandpass $F_{\min}$ can also be stated in terms of the limiting magnitude $m_{\min}$,

$$m_{\min} = m_0 - \frac{5}{2}\log_{10}\left(\frac{F_{\min}}{F_0}\right), \tag{6}$$

for m_0 a zero-point magnitude and F_0 its corresponding flux (as always, everything is implicitly written in a defined bandpass).

In Eq. 5, again $i \in \{\text{single}, \text{double}\}$, and as a consequence the maximum distance to which binary stars will be selected is greater than that of single stars, simply as a consequence of imposing a magnitude cut. The ratio of double to single systems is

$$\begin{align}\frac{N_d}{N_s} &= \frac{n_d}{n_s}\left(\frac{d_{\max,d}}{d_{\max,s}}\right)^3 \tag{7}\\ &= \text{BF} \times (1+\gamma_R)^{3/2}. \tag{8}\end{align}$$

In the nominal case of twin binaries ($\gamma_R = 1$), with a binary fraction BF $= 0.5$, there are $\sqrt{2}$ more binary systems than single systems in the sample. Correspondingly, there are $2\sqrt{2}$ more stars in binary systems than stars in single systems.

As a comment on Eq. 3, if we wished to write a stellar number density profile that accounted for the vertical structure of the Milky Way, we might choose a profile either $\propto \exp(-z/H)$, or $\propto \text{sech}^2(z/H)$ for z the distance from the galactic midplane and H a scale-height. Both density profiles would lead closed form analytic solutions.

2.2. *How many planets are in the sample?*

The number of planets in the sample is

$$\begin{align}N_{\text{planets}} =& N_{\text{planets in single star systems}} + \tag{9}\\ & N_{\text{planets in double star systems}} \nonumber\\ =& \Gamma_{t,s}N_s + 2\Gamma_{t,d}N_d. \tag{10}\end{align}$$

The factor of 2 accounts for the fact that there are twice as many stars in double star systems.

2.3. *What is the true occurrence rate?*

The "true occurrence rate" is the average number of planets per star. Thus

$$\begin{align}\Gamma_t &= \frac{N_{\text{planets}}}{N_{\text{stars}}} \tag{11}\\ \Gamma_t &= \frac{\Gamma_{t,s}N_s + 2\Gamma_{t,d}N_d}{N_s + 2N_d}. \tag{12}\end{align}$$

2.4. *How many planets are detected?*

The total number of planet detections is the sum of the number of planets detected in single star systems $N_{\text{det},s}$ and the number of planets detected in double star systems $N_{\text{det},d}$. These can be expressed individually. The former is

$$N_{\text{det},s} = N_s\Gamma_{t,s}f_{s,\text{geom}}f_{s,\text{S/N}>(\text{S/N})_{\min}}, \tag{13}$$

where the product $N_s\Gamma_{t,s}$ is the number of planets in the single star systems of the sample, $f_{s,\text{geom}} \equiv f_{s,g}$ is the geometric transit probability, and $f_{s,(\text{S/N}>(\text{S/N})_{\min})} \equiv f_{s,c}$ is the fraction of these transiting planets that are observed with signal to noise greater than the minimum detection threshold (the completeness). Analogously,

$$N_{\text{det},d} = 2N_d\Gamma_{t,d}f_{d,g}f_{d,c}, \tag{14}$$

where now $2N_d\Gamma_{t,d}$ is the number of planets in the double star systems of the sample, the geometric transit probability is the same (if we have twin binaries – not if we consider a distribution of secondaries) and the completeness term must account for any differences in the signal to noise distribution that come from stellar binarity.

2.4.1. *Analytic completeness*

Since the geometric transit probability is known, the only terms we have yet to compute are the completeness terms, $f_{i,c}$ for $i \in \{\text{single}, \text{double}\}$. We proceed as follows.

The signal S for a box-car train transiting planet is

$$\begin{align}\text{S} &= \delta\mathcal{D} \tag{15}\\ &= \left(\frac{R_p}{R_\star}\right)^2\mathcal{D}, \tag{16}\end{align}$$

for R_p the planet's radius, $R_\star$ that of its host star, and $\mathcal{D}$ the dilution parameter defined as

$$\begin{align}\mathcal{D} &= \begin{cases} L_1/L_d, & \text{if binary and target primary}\\ \gamma_R L_1/L_d, & \text{if binary and target secondary}\\ 1, & \text{if single,}\end{cases}\nonumber\\ &= \begin{cases} (1+\gamma_R)^{-1}, & \text{if binary and target primary}\\ (1+\gamma_R^{-1})^{-1}, & \text{if binary and target secondary}\\ 1, & \text{if single,}\end{cases} \tag{17}\end{align}$$

where L_1, L_d, and γ_R were defined in the opening monograph.

Assuming the only source of noise is Poissonian counting noise, the noise N can be written

$$\text{N} = \frac{1}{\sqrt{N_\gamma}}, \tag{18}$$

for N_γ the number of photons received by the detector. This noise model is a useful simplification – see [Howell 2006, pg 75] for the full CCD equation. The number of received photons can be written

$$N_\gamma = F_\gamma^{\text{N}} A N_{\text{tra}} T_{\text{dur}}, \tag{19}$$

for F_γ^{N} the photon number flux from the system $[\text{ph}\,\text{cm}^{-2}\,\text{s}^{-1}]$, A the detector area, T_{dur} the transit duration, and N_{tra} the number of transits observed, which is multiplied in assuming the transits are "phase-folded".

Thus the signal to noise ratio can be written

$$\mathrm{S/N} = \delta\mathcal{D}\sqrt{F_\gamma^{\mathrm{N}} A N_{\mathrm{tra}} T_{\mathrm{dur}}}. \tag{20}$$

This means we can write the minimum number flux of photons required for a detection at threshold as

$$F_{\mathrm{lim}}^{N} = \left[\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_{\mathrm{min}} \frac{1}{\delta\mathcal{D}}\right]^2 \frac{1}{A N_{\mathrm{tra}} T_{\mathrm{dur}}}. \tag{21}$$

To convert this to F_{lim}, multiply by the average photon energy in the bandpass.

In passing, given the parameters that define a survey and planet type, Eq. 20 would need to be re-expressed with N_{tra} roughly the ratio of the observing baseline to the planet period, and T_{dur} a function of $R_\star, P, a$, and impact parameter b, and then perhaps averaged over b. We leave them as-is for subsequent development.

The interesting term in Eq. 20 that changes between stars of the same binarity class in our idealized sample is the square root of F_γ^{N}. This is the term that leads to a distribution of signal to noises for different stars. The completenesses $f_{i,c}$ can be directly expressed in terms of those probability density functions:

$$f_{i,c} = \int_{(\mathrm{S/N})_{\mathrm{min}}}^{\infty} \mathrm{d}\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_i \ \mathrm{prob}\left(\frac{\mathrm{S}}{\mathrm{N}}\right)_i. \tag{22}$$

We keep the subscript i because the signal to noise distributions are different for the cases of single star systems ($i = s$) and double star systems ($i = d$). Notably:

- The dilution differs (Eq. 17).
- The photon number flux from the system differs.

To simplify notation, we let $x_i \equiv (\mathrm{S/N})_i$, and rewrite Eq. 22 as

$$f_{i,c} = \int_{\mathrm{x_{min}}}^{\infty} \mathrm{d}x_i \ \mathrm{prob}(x_i). \tag{23}$$

2.4.2. *Deriving* prob(x_i)

We want expressions for the probability density function of the observed signal to noise ratio, prob(x_i), for both the single and binary system case.

First, note that a star placed uniformly in the volume of the search space will have a probability density function for its distance r from the origin of

$$\mathrm{prob}(r) = \frac{3r^2}{d_{\mathrm{max}}^3}, \tag{24}$$

where the appropriate maximum distances should be substituted per Eq. 5. Noting the transformation rule for probability density functions, we can evaluate the probability of a star having a observed number flux $F_{i,\gamma}^{\mathrm{N}}$ in the bandpass,

$$\mathrm{prob}(F_{i,\gamma}^{\mathrm{N}}) = \mathrm{prob}(r(F_{i,\gamma}^{\mathrm{N}}))\left|\frac{\mathrm{d}r}{\mathrm{d}F_{i,\gamma}^{\mathrm{N}}}\right| \tag{25}$$

$$= \frac{3}{2d_{\mathrm{max}}^3} c_i^{3/2} (F_{i,\gamma}^{\mathrm{N}})^{-5/2}, \tag{26}$$

where in the latter equality we have written a "number luminosity" c_i (units of inverse time) defined for $i \in \{\mathrm{single}, \mathrm{double}\}$ as

$$c_i = \begin{cases} R_1^2 F_{s1,\gamma}^{\mathrm{N}}, & \text{if single} \\ R_1^2 F_{s1,\gamma}^{\mathrm{N}} + R_2^2 F_{s2,\gamma}^{\mathrm{N}} & \text{if double.} \end{cases} \tag{27}$$

In Eq. 27, $F_{s1,\gamma}^{\mathrm{N}}$ and $F_{s2,\gamma}^{\mathrm{N}}$ are the photon number fluxes at the surfaces of the stars. To derive Eq. 26, we simply scaled these by the distance:

$$F_{i,\gamma}^{\mathrm{N}} = \frac{c_i}{r^2}. \tag{28}$$

The surface photon number fluxes $F_{si,\gamma}^{\mathrm{N}}$ in Eq. 27 are usually evaluated numerically, by convolving the wavelength-specific photon flux density of a star with the dimensionless spectral response function of the instrument. In other words,

$$F_{s,\gamma}^{\mathrm{N}} = \int F_\lambda T_\lambda \,\mathrm{d}\lambda. \tag{29}$$

The wavelength-specific photon flux density F_λ [ph cm^{-2} s^{-1} Å^{-1}] could be from Pickles' library [1998, PASP 110, 863], or could be a blackbody function. The transmission function is, up to factor of order unity, a step function between two wavelengths λ_{min} and λ_{max}. If we assume a blackbody source, Eq. 29 becomes

$$F_{s,\gamma}^{\mathrm{N}} = 8\pi c \left(\frac{kT}{hc}\right)^3 \int_{hc/(\lambda_{\mathrm{max}} kT)}^{hc/(\lambda_{\mathrm{min}} kT)} \frac{u^2}{e^u - 1}\,\mathrm{d}u, \tag{30}$$

which can be evaluated numerically[2].

The importance of the functional form of $F_{s,\gamma}^{\mathrm{N}}$ is that it is to first order only a function of the blackbody temperature and the bandpass wavelength bounds. Thus in the most general case $c_i(R_1, R_2, T_1, T_2, \lambda_{\mathrm{min}}, \lambda_{\mathrm{max}})$ and nothing else. The only random variable involved in the flux being received at the detector is r, so we can indeed write the flux received at the detector as in Eq. 28.

We can finally write out the probability density functions for the signal to noise ratios in the single and double-star cases by using the transformation rule for pdfs, and applying Eq. 20. For single stars,

$$\mathrm{prob}(x_s) = \frac{3}{d_{\mathrm{max,s}}^3} c_s^{3/2} \delta^3 \left(A T_{\mathrm{dur}} N_{\mathrm{tra}}\right)^{3/2} x_s^{-4}. \tag{31}$$

Analogously for double stars,

$$\mathrm{prob}(x_d) = \frac{3}{d_{\mathrm{max,d}}^3} c_d^{3/2} (\mathcal{D}\delta)^3 \left(A T_{\mathrm{dur}} N_{\mathrm{tra}}\right)^{3/2} x_d^{-4}. \tag{32}$$

2.4.3. *Number of detected planets*

Performing the integrals of Eq. 23, we get:

[2] It may help in the numerics to note that infinite series representations of this type of dimensionless integral exist and converge quickly. For instance, one can show that

$$\int_0^a \frac{u^3}{e^u - 1}\,\mathrm{d}u = \sum_{n=1}^{\infty} \frac{6}{n^4} - \frac{e^{-an}}{n^4}(6 + 6an + 3(an)^2 + (an)^3).$$

A similar expression exists for the similar integral in the text. [Michels 1968] explains an analogous derivation.

$$N_{\mathrm{det},s} = N_s \Gamma_{t,s} f_{s,g} f_{s,c} \tag{33}$$

$$= N_s \Gamma_{t,s} \frac{R_\star}{a} \frac{1}{d^3_{\mathrm{max,s}}} c_s^{3/2} \delta^3 \left(A T_{\mathrm{dur}} N_{\mathrm{tra}}\right)^{3/2} x_{\mathrm{min}}^{-3}, \tag{34}$$

and

$$N_{\mathrm{det},d} = 2 N_d \Gamma_{t,d} f_{d,g} f_{d,c} \tag{35}$$

$$= 2 N_d \Gamma_{t,d} \frac{R_\star}{a} \frac{1}{d^3_{\mathrm{max,d}}} c_d^{3/2} (\mathcal{D}\delta)^3 \left(A T_{\mathrm{dur}} N_{\mathrm{tra}}\right)^{3/2} x_{\mathrm{min}}^{-3}. \tag{36}$$

Formally, the $f_{i,c}$ terms should be written as $\min(1, \ldots)$, where $(\ldots)$ is the given expression. This ensures that the fraction of planets above the signal to noise threshold is less than 1. Note that we assumed twin binaries in the geometric transit probabilities. Interestingly, the ratio of the two completeness fractions is

$$\frac{f_{d,c}}{f_{s,c}} = \left(\frac{d_{\mathrm{max,s}}}{d_{\mathrm{max,d}}}\right)^3 \left(\frac{c_d}{c_s}\right)^{3/2} \mathcal{D}^3 = (1+\gamma_R)^{-3}, \tag{37}$$

which, for the $\gamma_R = 1$ case, gives 1/8. (We got the same number from simple scaling laws at the outset, but the above machinery will be useful when generalizing in Sec. 3.)

The number of detected planets N_{det} is the sum of the two appropriate equations above, and can be written

$$N_{\mathrm{det}} = \left(\frac{\delta}{x_{\mathrm{min}}}\right)^3 \left(A T_{\mathrm{dur}} N_{\mathrm{tra}}\right)^{3/2} \frac{R_\star}{a} \times \left[N_s \Gamma_{t,s} \frac{c_s^{3/2}}{d^3_{\mathrm{max,s}}} + 2 N_d \Gamma_{t,d} \frac{c_d^{3/2}}{d^3_{\mathrm{max,d}}} \right]. \tag{38}$$

All terms can be input to a computer, and checked against a Monte Carlo simulation if desired.

2.5. *Astronomer A ignores binarity*

Astronomer A has never heard of binary star systems. Nor has he heard of completeness corrections. He knows about geometric transit probabilities. What occurrence rate does he derive for planets of radius R_p and period P?

The *total* occurrence rate (number of planets divided by number of "stars") for Astronomer A would be $N_{\mathrm{det}}/(N_s + N_d)$, up to the geometric correction. However, even though Astronomer A does not know about binaries, the radii he derives for any planets in binary systems are too small, by a factor $\sqrt{\mathcal{D}}$ (for twin binaries, they are all $R_p/\sqrt{2}$). Astronomer A wants an occurrence rate for planets of radius R_p and period P. The answer is thus

$$\Gamma_{\mathrm{A},R_p} = \frac{N_{\mathrm{det,s}}/f_{s,g}}{N_s + N_d}. \tag{39}$$

This astronomer will also think there is a second population of planets, with radius $R_p\sqrt{\mathcal{D}}$, and will thus rush to Nature claiming to also have derived a second occurrence rate,

$$\Gamma_{\mathrm{A},R_p\sqrt{\mathcal{D}}} = \frac{N_{\mathrm{det,d}}/f_{s,g}}{N_s + N_d}. \tag{40}$$

Note in the second case that this astronomer thinks these are single stars, and generally will miscompute the geometric transit probability.

2.6. *Astronomer B counts host stars correctly*

Astronomer B can somehow account correctly for the "2 for 1" effect of binarity, *i.e.* that the sample actually has more stars than astronomer A thought.

By the same token as above,

$$\Gamma_{\mathrm{B},R_p} = \frac{N_{\mathrm{det,s}}/f_{s,g}}{N_s + 2N_d}, \tag{41}$$

and

$$\Gamma_{\mathrm{B},R_p\sqrt{\mathcal{D}}} = \frac{N_{\mathrm{det,d}}/f_{d,g}}{N_s + 2N_d}. \tag{42}$$

2.7. *Astronomer C counts host stars correctly and figures out diluted radii*

Astronomer C did high resolution imaging followup on every candidate, and correctly classifies the planetary radii. Thus, she also knows which planets are in binary systems, and which are in single star systems.

She knows that the purported population of planets with radii $R_p\sqrt{\mathcal{D}}$ does not exist. All detected planets from this survey have radii R_p. She computes an occurrence rate

$$\Gamma_{\mathrm{C},R_p} = \frac{N_{\mathrm{det,s}}/f_{s,g} + N_{\mathrm{det,d}}/f_{d,g}}{N_s + 2N_d} \tag{43}$$

the closest yet to the true rate (Sec. 2.3).

2.8. *Astronomer D counts host stars correctly, figures out diluted radii, and accounts for completeness*

Astronomer D, knows which detections were around binaries and the associated radius correction. They do injection recovery, and derive correct estimates for their completeness functions about single stars $f_{s,c}$, and about double star systems, $f_{d,c}$. With this knowledge in hand, they compute the respective single and binary occurrence rates

$$\Gamma_{t,s} = \frac{N_{\mathrm{det,s}}}{N_s f_{s,g} f_{s,c}}, \tag{44}$$

$$\Gamma_{t,d} = \frac{N_{\mathrm{det,d}}}{2N_d f_{d,g} f_{d,c}}. \tag{45}$$

With these in hand, they derive the overall occurrence rate

$$\Gamma_{\mathrm{D},R_p} = \frac{N_{\mathrm{det,s}}/(f_{s,g}f_{s,c}) + N_{\mathrm{det,d}}/(f_{d,g}f_{d,c})}{N_s + 2N_d} \tag{46}$$

$$= \frac{\Gamma_{t,s}N_s + 2\Gamma_{t,d}N_d}{N_s + 2N_d}. \tag{47}$$

Although we admit it's been a bit of a slog, it happens that Astronomer D's occurrence rate is the true occurrence rate (cf. Eq. 12).

All it takes is ≈ a full semester at Keck, a system-by-system analysis, and perfect understanding of the completeness of the detection efficiency for single and double star systems.

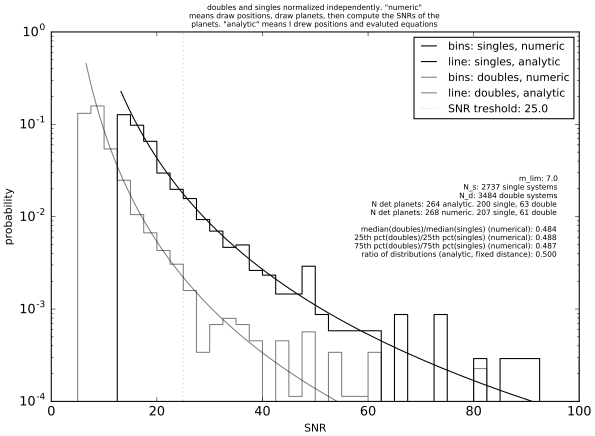

FIG. 1.— Comparison of analytic and numeric probability density functions of the SNR in an idealized transit survey. The analytic lines are Eqs. 31 and 32 for the planet populations orbiting single and binary stars. The underlying stepped histogram is output from Monte Carlo simulations. Poisson noise leads to a small deviation at the faint and bright limits, but the numerics and analytics otherwise agree.

2.9. *Numerical verification*

To check the preceding analytic development, we implemented a Monte Carlo simulation of this idealized transit survey. To run the survey, we defined the instrument specifications (detector area and transmission function), the stellar population (binary fraction, total number density of a given stellar class, fixed stellar properties), the planet population (fixed planet radius, period, and occurrence rate about single and binary stars), and finally the survey parameters (observing baseline, minimum SNR for "detection"). We then randomly drew star positions, randomly assigned planets to stars in single and binary systems, and computed the resulting signal to noise (Eq. 20) with which the transits would be observed. As in the preceding analytics, we assumed "twin" binaries (same stellar radii, same effective temperature, and dilution does not depend on which stellar binary is the "target").

The results are shown in Fig. 1, and indicate that the analytic probability distribution functions Eqs. 31, 32, and the number of detections (Eq. 38) are correct.

A point evident in Fig. 1 is that, for fixed planet parameters, and fixed stellar parameters ($R_\star, L_\star$, and distance r) the SNR distribution for planets in binaries is poorer than that of planets in single star systems. We can see analytically that this simply due to dilution:

$$\frac{\mathrm{prob}(x_d)}{\mathrm{prob}(x_s)} = (1+\gamma_R)^{-1} = \mathcal{D}. \tag{48}$$

Deriving this simple form requires noting that the ratios of the bandpass-specific number luminosities (Eq. 27) is equal to the ratio of the bandpass-specific energy luminosities (otherwise a term with c_s/c_d must be included).

2.10. *Representative numbers for a few cases*

2.10.1. *Twin binaries: if we ignore binarity, for what fraction of detections do we misclassify the radii?*

Ignoring binarity, we will detect $N_{\rm det,s}$ planets around single stars, and $N_{\rm det,d}$ planets around double stars. The latter set will be assumed to have radii $R_p\sqrt{\mathcal{D}}$. The fraction of detections with misclassified radii can then be written

$$\frac{N_{\rm det,d}}{N_{\rm det,s}+N_{\rm det,d}} = \frac{1}{1+\alpha}, \tag{49}$$

for

$$\alpha \equiv \frac{1}{2(\mathrm{BF})}(1+\gamma_R)^{3/2}\frac{\Gamma_{t,s}}{\Gamma_{t,d}}. \tag{50}$$

For the nominal G2V dwarf case of BF = 0.45, twin binaries with equal occurrence rates this produces a misclassification rate of 24%, in agreement with Fig. 1.

2.10.2. *Twin binaries: if we ignore binarity, how wrong is our occurrence rate for planets of radius R_p?*

This is almost simply asking "what is the relative difference between the occurrence rates derived by Astronomers D and A for planets of radius R_p?" However, in the more realistic case, Astronomer A also has derived a completeness, which we assume is the same as for Astronomer D in the single star case. So Astronomer A now misclassifies planetary radii, and miscounts the total number of stars, but knows his completeness for single stars. Astronomer D corrects all these errors. Then

$$\Gamma_{\mathrm{A},R_p} = \frac{N_{\rm det,s}/(f_{s,g}f_{s,c})}{N_s+N_d}. \tag{51}$$

To express "how wrong" our occurrence rate is, we define a correction factor X_Γ such that the true occurrence rate is the product of the correction factor with the measured occurrence rate: $\Gamma_D \equiv \Gamma_A X_\Gamma$.

The correction factor is then

$$\begin{aligned} X_\Gamma &= \frac{\Gamma_{\mathrm{D},R_p}}{\Gamma_{\mathrm{A},R_p}} && (52)\\ &= \frac{\Gamma_{t,s}N_s + 2\Gamma_{t,d}N_d}{N_s+2N_d}\cdot\frac{N_s+N_d}{\Gamma_{t,s}N_s} && (53)\\ &= \frac{(1+\beta)(1+2\beta\Gamma_{t,d}/\Gamma_{t,s})}{(1+2\beta)}, && (54)\end{aligned}$$

for

$$\beta \equiv N_d/N_s = \mathrm{BF}\times(1+\gamma_R)^{3/2}. \tag{55}$$

For the nominal G2V dwarf case of BF = 0.45 with twin binaries ($\gamma_R = 1$) and $\Gamma_{t,d} = \Gamma_{t,s}$ this gives $X_\Gamma = 2.27$. For instance, in the numerical simulation corresponding to Fig. 1, Astronomer A finds $\Gamma_{\mathrm{A},R_p} = 0.22$, while Astronomer D derives the true (input) occurrence rate of $\Gamma_{\mathrm{D},R_p} = 0.5$. This gives a relative percentage error of $\delta\Gamma_{\mathrm{D},R_p} \equiv (\Gamma_{\rm A,R_p} - \Gamma_{\rm D,R_p})/\Gamma_{\rm D,R_p} = -56\%$.

3. MODEL #2: FIXED STARS, FIXED PLANETS, VARYING LIGHT-RATIO BINARIES

Same story as Model #1, except now γ_R varies across the population of star systems. It does so because the underlying mass ratio varies. Since we are interested in solar type binaries, we take the distribution of the mass ratio $q = M_2/M_1$ by approximating [Rhagavan et al 2010, Fig 16]:

$$\mathrm{prob}(q) = \begin{cases} c_q & 0.1 < q \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{56}$$

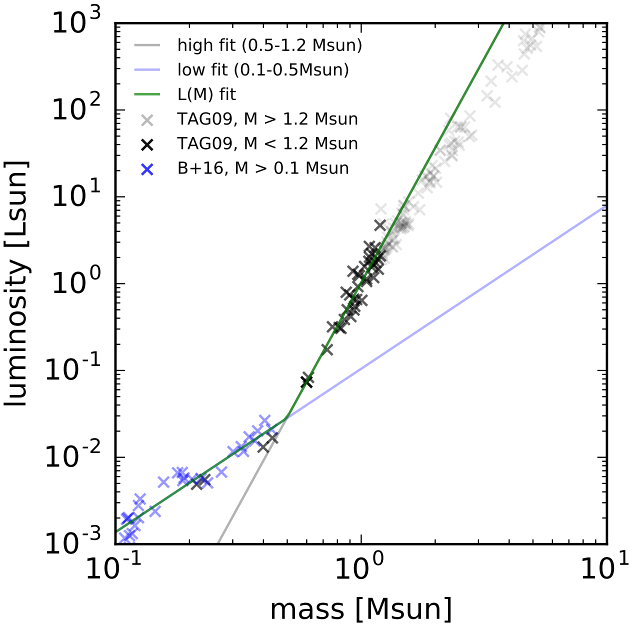

FIG. 2.— Empirical fit to main sequence dwarf mass luminosity data compiled from Torres et al. [2009] and Benedict et al. [2016]. The "low fit" is a least squares fit to data from $0.1 - 0.5M_\odot$, and the "high fit" is to data above that, and below the Kraft break ($1.2M_\odot$). The $L(M)$ relation taken for subsequent numerics is the maximum of the two fits.

for $c_q = 1/9$ to normalize the distribution to 1. The mass-luminosity relation can, for analytic convenience, be approximated as $L = M^\alpha$, with the lore-value of α being 3.5. While we use this in subsequent analytic development, for numerics we fit a lines to mass-luminosity data collected by Torres et al. [2009] for dwarfs above M, and Benedict et al. [2016] for dwarfs below M. To convert Benedict et al. [2016]'s reported M_V values to absolute luminosities, we interpolated over E. Mamajek's table[3]. In log-log space, we let the intersection point of the lines float, and make various cuts on the data as indicated in Fig 2. The fit parameters are available in a footnote[4].

The other changes between Model #2 and Model #1 are:

1. We allow a third occurrence rate, splitting $\Gamma_{t,d} \rightarrow (\Gamma_{t,d1}, \Gamma_{t,d2})$ where the latter terms represent the occurrence rate around the primary of a double star system, and the occurrence rate around the secondary of a double star system. For the case of *e.g.*, $q = 0.1$, this seems more relevant. We can then take the limit that says "any planet around the secondary is as likely as around the primary", or go to the opposite extreme of "there are only planets around the primaries, not the secondaries".

2. Our binary systems have secondaries with varying masses. Thus they have varying radii. Taking [Demircan & Kahraman 1991]'s empirical fit to eclipsing binary data,
$$R = 1.06M^{0.945} \tag{57}$$
for all the stars in our desired mass range ($M < 1.66M_\odot$, D&K 1991's stated bounds), and for each in solar units. This only affects the transit probabilities about the secondaries of binary star systems.

We then ask the same questions as for Model #1.

3.1. *How many stars are in the sample?*

Let N_s be the number of single star systems, and N_d the number of double star systems. N_s is the same as in Sec. 2.

Now
$$N_d(\gamma_R) = n_d \frac{4\pi}{3} d^3_{\rm max,d}(L_1^N, \gamma_R, F^N_{\rm lim}) \tag{58}$$
for
$$d_{\rm max,d}(L_1^N, \gamma_R, F^N_{\rm lim}) = d_{\rm max,s}(L_1^N, F^N_{\rm lim}) \times (1+\gamma_R)^{1/2}. \tag{59}$$
Since q is a random variable, γ_R is a random variable, and $d_{\rm max,d}$ is a random variable. Since N_d is a function of $d_{\rm max,d}$, the number of double star systems becomes a random variable. The expected (mean) number of double star systems in the sample is
$$\langle N_d \rangle = \int_0^\infty N_d \, \mathrm{prob}(N_d) \, \mathrm{d}N_d. \tag{60}$$
By applying the chain rule for probability density functions, the distribution $\mathrm{prob}(N_d)$ can be written
$$\mathrm{prob}(N_d) = \mathrm{prob}(q(\gamma_R)) \left| \frac{\mathrm{d}q}{\mathrm{d}\gamma_R} \right| \left| \frac{\mathrm{d}\gamma_R}{\mathrm{d}d_{\rm max,d}} \right| \left| \frac{\mathrm{d}d_{\rm max,d}}{\mathrm{d}N_d} \right|. \tag{61}$$
Doing some algebra, and assuming $\gamma_R = q^3$, this can be shown to be
$$\mathrm{prob}(N_d) = \frac{2}{81} N_d^{-1/3} \left(\frac{3}{4\pi n_d} \right)^{2/3} d_{\rm max,s}^{-2/3} \times \left[\left(\frac{3N_d}{4\pi n_d} \right)^{2/3} - d^2_{\rm max,s} \right]^{-2/3} \tag{62}$$
over the interval
$$\text{lower bound} = \frac{4\pi n_d}{3} (\sqrt{0.1^3 + 1} d_{\rm max,s})^3 \tag{63}$$
$$\text{upper bound} = \frac{4\pi n_d}{3} (\sqrt{2} d_{\rm max,s})^3, \tag{64}$$
and outside the stated interval $\mathrm{prob}(N_d) = 0$.

While a closed analytic expression for $\langle N_d \rangle$ does exist, it is messy, and it does not yield much intuition. Instead, summarizing the important points and supporting them numerically:

- In a volume-limited sample of binary star systems in which the primary mass is fixed, and the mass ratio is drawn from a bounded uniform distribution, the distribution of γ_R will be biased towards low values ($\gamma_R \approx 0.1$). This is shown in Fig. 3.

[3] pas.rochester.edu/~emamajek/EEM_dwarf_UBVIJHK_colors_Teff.txt, downloaded 2017.08.02

[4] m lo: 1.8818719873988132 – c lo: -0.9799647314108376 – m hi: 5.1540712426599882 – c hi: 0.0127626185389781 – M at merge: 0.4972991257826812 – L at merge: 0.0281260412126928.

- In a magnitude-limited sample of binary star systems in which the primary mass is fixed, and the mass ratio of the *population* is drawn from a bounded uniform distribution, the observed distribution of mass ratios will be biased towards high values. You will see more twins, because they are detectable out to a greater distance[5]. This is shown in Fig. 4, and explained analytically below.
- In a magnitude-limited sample of binary star systems in which the primary mass is fixed, the distribution of γ_R will be biased towards low values (≈ 0.1), but less so than in a volume-limited sample. This is shown in Fig. 5.

In passing, the bias in prob(γ_R) towards low luminosity ratios can be seen analytically. If we assume $\gamma_R = q^\alpha$, then

$$\mathrm{prob}(\gamma_R) = \frac{1}{9\alpha}\,\gamma_R^{\frac{1-\alpha}{\alpha}} \quad \text{for } (0.1)^\alpha < \gamma_R < 1, \tag{65}$$

and otherwise zero. For instance if $\alpha = 3$, $\mathrm{prob}(\gamma_R) \propto \gamma_R^{-2/3}$ and the domain extends from 1 to 10^{-3}, where the probability distribution peaks.

Similarly, the probability distribution function for the mass ratio of binaries in a magnitude limited sample can be found by marginalizing over the joint distribution for a binary star's position and mass ratio. Explicitly, using 'ml' as a subscript for 'magnitude-limited', and assuming $L \propto M^3$,

$$\mathrm{prob}_{\mathrm{ml}}(q) = \int_0^{d_{\max}(q)} \mathrm{prob}_{\mathrm{ml}}(r,q)\,\mathrm{d}r \tag{66}$$

$$= \mathscr{Z}\cdot(1+q^3)^{3/2} \quad \text{if } 0.1 < q \leq 1, \tag{67}$$

and otherwise 0, for $\mathscr{Z}$ the appropriate normalization constant. Eq. 67 is the "analytic" form overplotted in Fig. 4. Its independence from both the primary star mass as well as from the magnitude cutoff will be used in Sec. 4. The form given in Eq. 67 is found by writing $\mathrm{prob}_{\mathrm{ml}}(r,q) = \mathrm{prob}(q|r)\mathrm{prob}_{\mathrm{ml}}(r)$, using Eq. 56 for $\mathrm{prob}(q|r)$, and marginalizing over the mass ratio

$$\mathrm{prob}_{\mathrm{ml}}(r) = \int_0^1 \mathrm{prob}_{\mathrm{ml}}(r|q)\,\mathrm{prob}(q)\,\mathrm{d}q \tag{68}$$

$$= \int_{0.1}^1 \frac{1}{3}\frac{r^2}{d_{\max}^3(q)}\,\mathrm{d}q \quad \text{if } r < d_{\max}(q), \tag{69}$$

and otherwise zero. The resulting joint probability distribution is plotted in Fig. 6.

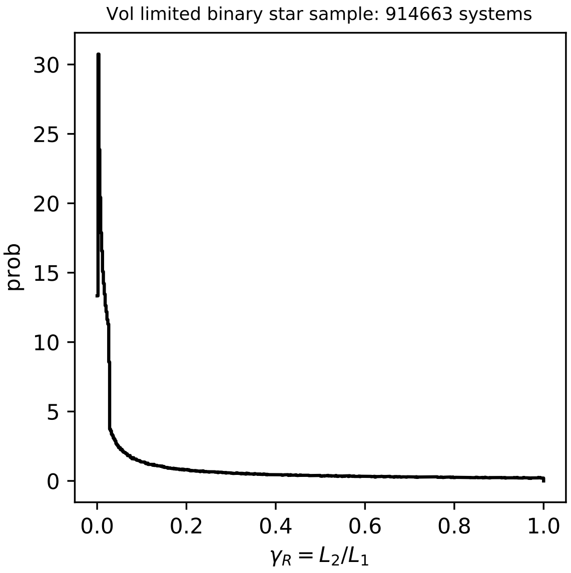

FIG. 3.— The distribution of the luminosity ratio for a volume limited sample of binary stars. BF = 0.45 [Duchene and Kraus 2013]; total number density from Bovy 2017; $M(L)$ relation from Fig. 2; mass ratios are drawn from Eq. 56.

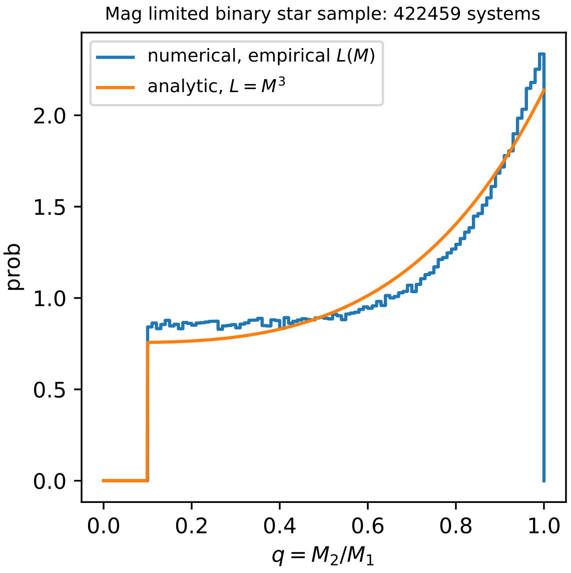

FIG. 4.— The distribution of the mass ratio for a magnitude limited sample of binary stars. For numerics, BF = 0.45 [Duchene and Kraus 2013], total number density from Bovy 2017, empirical $M(L)$ from Fig. 2, and mass ratios are drawn from Eq. 56 – a *uniform* distribution in a volume-limited sample! The analytic line (Eq. 67) is shown to indicate that the entire effect, albeit with a simpler $M(L)$, can be understood without resorting to numerics.

How many stars are in the sample?— We return to the original question: how many stars?

$$\langle N_{\mathrm{stars}}\rangle = N_s + 2\langle N_d\rangle, \tag{70}$$

as before. Since we are not giving analytic expressions for either of the $\langle\ldots\rangle$ terms, we just note that in the case of a population with a single γ_R value, $N_d/N_s = \mathrm{BF}\times(1+\gamma_R)^{3/2}$. A lower bound for this ratio will always be the binary fraction. For a single run of a Monte Carlo code, where $\mathrm{mean}(\gamma_R) = 0.28$, $\mathrm{median}(\gamma_R) = 0.13$, and the distribution was that given in Fig. 5, the observed ratio was $N_d/N_s = 0.59$. This corresponds to a single-valued population with $\gamma_R \sim 0.20$, a number between the mean and median of the true distribution. It means ~ 1.2 stars in binary star systems for every star in a single star system from this sample.

We also note in passing that $N_d/N_s = 0.59$ is a smaller ratio than for the case of a fixed γ_R population with

[5] This is a scary systematic w.r.t. the claimed intrinsic excess of twin binaries.

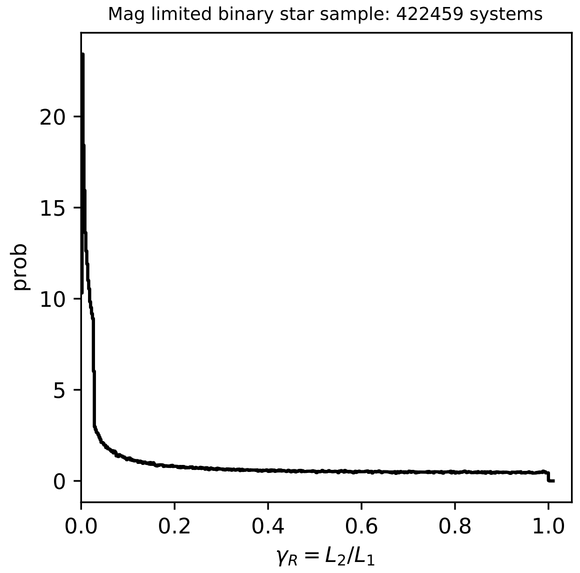

FIG. 5.— The distribution of the luminosity ratio for a magnitude limited sample of binary stars. BF = 0.45 [Duchene and Kraus 2013]; total number density from Bovy 2017; $M(L)$ relation from Fig. 2; mass ratios are drawn from Eq. 56.

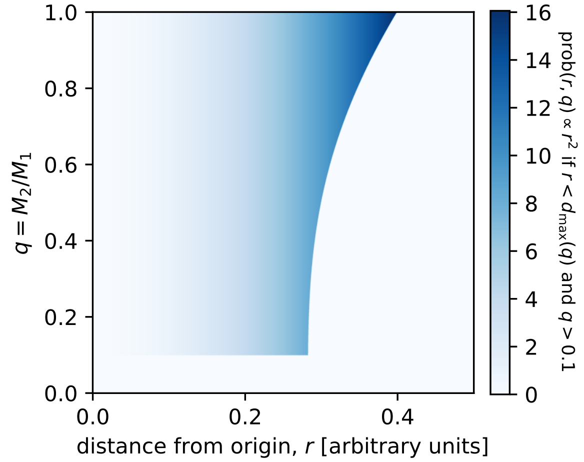

FIG. 6.— The joint probability distribution of a binary star's position and mass ratio in a magnitude-limited sample. This plot assumes a different $M(L)$ relation than previous plots: $L = M^3$. Note that the marginalized mass ratio distribution is uniform at any given distance.

$\gamma_R = 1$, which gave $N_d/N_s = 1.27$. The difference is that in the latter population, the volume of searchable stars is greater. Based on the fixed-γ_R population scaling law $N_d/N_s \propto (1+\gamma_R)^{3/2}$, we would expect the ratio of single stars to be $\sim (2/1.1)^{3/2} = 2.4$ times smaller, which is roughly (though not exactly) what we observe.

3.2. *How many planets are in the sample?*

The mean number of planets in the sample is

$$\begin{aligned}\langle N_{\rm planets}\rangle =& N_{\text{planets in single star systems}} + \\ & N_{\text{planets in double star systems}} \quad (71)\\ =& \Gamma_{t,s} N_s + (\Gamma_{t,d1} + \Gamma_{t,d2})\langle N_d\rangle. \quad (72)\end{aligned}$$

What was formerly a factor of 2 has now been split into the true occurrence rates about the primaries and secondaries of double star systems.

3.3. *What is the true occurrence rate?*

The "true occurrence rate" is the average number of planets per star. Thus

$$\Gamma_t = \frac{\langle N_{\rm planets}\rangle}{\langle N_{\rm stars}\rangle} \quad (73)$$

$$\Gamma_t = \frac{\Gamma_{t,s} N_s + (\Gamma_{t,d1} + \Gamma_{t,d2})\langle N_d\rangle}{N_s + 2\langle N_d\rangle}. \quad (74)$$

3.4. *How many planets are detected?*

The total number of planet detections is the sum of the number of planets detected in single star systems $N_{{\rm det},s}$ and the number of planets detected in double star systems $N_{{\rm det},d}$. The latter of these is a random variable, and can be further split into the primary and secondary contributions $N_{{\rm det},d1}$, $N_{{\rm det},d2}$.

The number of planets detected in single star systems is

$$N_{{\rm det},s} = N_s \Gamma_{t,s} f_{s,g} f_{s,c}, \quad (75)$$

where the product $N_s\Gamma_{t,s}$ is the number of planets in the single star systems of the sample, $f_{s,g}$ is the geometric probability of the planets transiting, and $f_{s,c}$ is the fraction of transiting planets around single star systems that are detected (the completeness).

The mean number of planets detected in double star systems is

$$\langle N_{{\rm det},d}\rangle = \langle N_d\rangle(\Gamma_{t,d1} f_{d1,g} f_{d1,c} + \Gamma_{t,d2} f_{d2,g} f_{d2,c}). \quad (76)$$

Now $\langle N_d\rangle\Gamma_{t,d1}$ is the mean number of planets orbiting the primaries of double star systems, ditto the corresponding expression for the secondaries. For this population, $f_{d1,g} = f_{s,g}$, but $f_{d1,g} \neq f_{d2,g}$. The completeness fractions differ between the primary and secondary because the dilution as a function of the light ratio, $\mathcal{D}(\gamma_R)$, has different behavior for the two cases. We show this in Fig. 7. Specifically, a planet orbiting the secondary will have worse dilution than a planet orbiting the primary, and consequently lower completeness.

We could probably derive analytic expressions for all the terms between brackets in Eq. 76, completeness included. I'm not yet convinced we need to – we should get cancellations for most of the questions we want to ask.

3.4.1. *Analytic completeness*

`todo, if necessary`

3.4.2. *Deriving* prob(x_i)

`todo, if necessary`

3.4.3. *Number of detected planets*

`todo, if necessary`

3.5. *Astronomer A ignores binarity*

Just as in Sec. 2, Astronomer A

- has never heard about binary star systems.

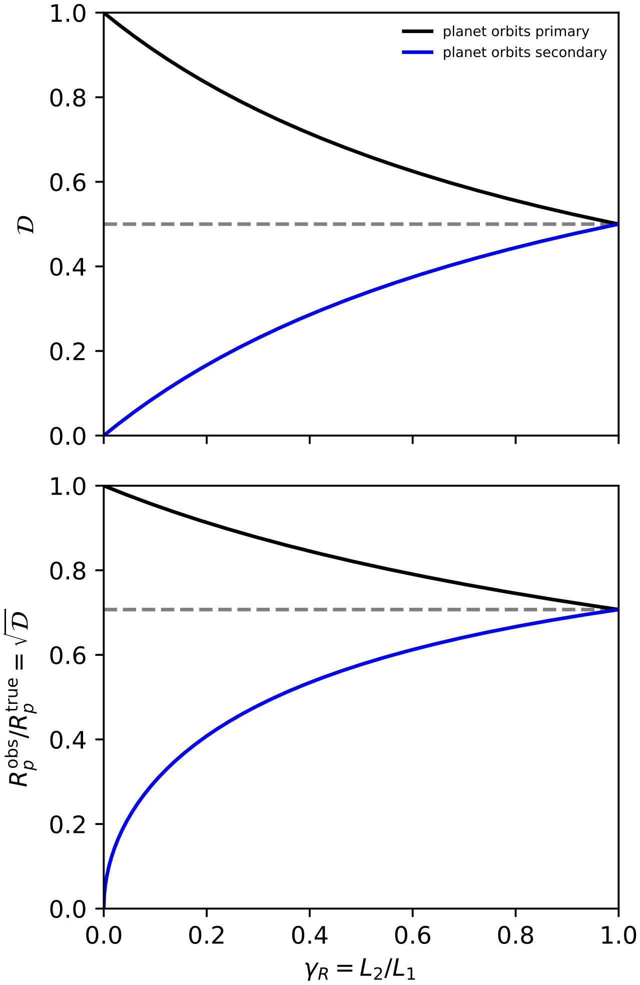

FIG. 7.— Dilution $\mathcal{D}$ and observed to true radius ratio $\sqrt{\mathcal{D}}$ plotted against binary light ratio γ_R. This illustrates Eq. 17. The lower panel's horizontal line is at $1/\sqrt{2}$. The implication is that if a detected planet orbits the secondary and you do not know it, you cannot measure R_p to better than $\approx 29\%$ accuracy. Conversely, if it orbits the primary and you do not know it, you will get the radius wrong by at most $\approx 29\%$.

- has never heard about completeness corrections.
- knows about geometric transit probabilities.

What occurrence rate does he derive for planets of radius R_p and period P?

The answer is the same as in Sec. 2,

$$\Gamma_{{\rm A},R_p} = \frac{N_{\rm det,s}/f_{s,g}}{N_s + N_d}. \tag{77}$$

However, now rather than planets detected in binary systems being perceived as a second population of planets with fixed radius $R_p\sqrt{\mathcal{D}}$, there will be an apparent spectrum of diluted radii, since $\mathcal{D}$ varies by system. Writing the number of detections in double systems as a function of observed radius, $N_{\rm det,d}(R_p^{\rm obs})$, Astronomer A would derive an occurrence rate for this separate population of

$$\Gamma_{\rm A}(R_p^{\rm obs}) = \frac{N_{\rm det,d}(R_p^{\rm obs})/f_{s,g}}{N_s + N_d}. \tag{78}$$

Note that this astronomer thinks these planets orbit single stars, and so miscomputes the geometric transit probability for the secondaries.

A more realistic question is "what fraction of multiples have γ_R smaller enough so that you'd still observe a radius close to the true one?". We discuss this in Sec. 3.10.3.

3.6. *Astronomer B counts host stars correctly*

Same as astronomer A above, except the denominator becomes $N_s + 2N_d$.

3.7. *Astronomer C counts host stars correctly and figures out diluted radii*

In Sec. 2, Astronomer C only had to do high resolution imaging to measure γ_R for each system. Since for $\gamma_R = 1$, the dilution is single-valued (Fig. 7), this immediately told her the true planet radius (since the primary/secondary distinction was meaningless).

In Model #2's universe, the dilution is double-valued for a given γ_R. To correctly work out the diluted radii, Astronomer C now needs more than high resolution imaging – she needs to resolve the stars during transit to discover which star the planet orbits. She also needs to derive masses for the primaries and secondaries (to make the geometric transit probability correction).

Only after doing this does she find that all detected planets from this survey have radii R_p. She computes an occurrence rate

$$\Gamma_{{\rm C},R_p} = \frac{N_{\rm det,s}/f_{s,g} + N_{\rm det,d1}/f_{d1,g} + N_{\rm det,d2}/f_{d2,g}}{N_s + 2N_d} \tag{79}$$

the closest yet to the true rate (Sec. 2.3).

3.8. *Astronomer D counts host stars correctly, figures out diluted radii, and accounts for completeness*

Same story as for Model #1. Now, they do injection recovery and derive correct estimates for their completeness functions about single stars $f_{s,c}$, the primaries of double star systems, $f_{d1,c}$, and the secondaries of double star systems $f_{d2,c}$. They compute the respective single and binary occurrence rates

$$\Gamma_{t,s} = \frac{N_{\rm det,s}}{N_s f_{s,g} f_{s,c}}, \tag{80}$$

$$\Gamma_{t,d1} = \frac{N_{\rm det,d1}}{N_d f_{d1,g} f_{d1,c}}, \tag{81}$$

$$\Gamma_{t,d2} = \frac{N_{\rm det,d2}}{N_d f_{d2,g} f_{d2,c}}. \tag{82}$$

With these in hand, they derive the overall occurrence rate

$$\Gamma_{{\rm D},R_p} = \frac{N_{\rm det,s}/f_s + N_{\rm det,d1}/f_{d1} + N_{\rm det,d2}/f_{d2}}{N_s + 2N_d} \tag{83}$$

$$= \frac{\Gamma_{t,s} N_s + (\Gamma_{t,d1} + \Gamma_{t,d2})N_d}{N_s + 2N_d}, \tag{84}$$

where $f_s \equiv f_{s,g} f_{s,c}$, $f_{d1} \equiv f_{d1,g} f_{d1,c}$, $f_{d2} \equiv f_{d2,g} f_{d2,c}$ for brevity. Astronomer D's occurrence rate is the true occurrence rate (cf. Eq. 74).

3.9. *Numerical verification*

```
todo, if needed
```

3.10. *Representative numbers for a few cases*

We have not given analytic expressions for the number of double star systems N_d because even with simplified $M(L)$ relations, they are unwieldy. However, we can ask similar numerical questions as from Sec. 2, and compare the answers.

3.10.1. *If we ignore binarity, for what fraction of detections do we misclassify the radii?*

Ignoring binarity, we will detect $N_{\rm det,s}$ planets around single stars, and $N_{\rm det,d}$ planets around double stars. The latter set will be assumed to have radii $R_p\sqrt{\mathcal{D}}$, where $\mathcal{D}$ is the vector of dilutions appropriate for each binary star system. The fraction of detections with misclassified radii can then be written

$$\frac{N_{\rm det,d}}{N_{\rm det,s}+N_{\rm det,d}} = \frac{1}{1+\alpha}, \tag{85}$$

for

$$\alpha \equiv \frac{N_d(\Gamma_{t,d1}f_{d1}+\Gamma_{t,d2}f_{d2})}{N_s\Gamma_{t,s}f_s}, \tag{86}$$

where $f_s \equiv f_{s,g}f_{s,c}$, $f_{d1} \equiv f_{d1,g}f_{d1,c}$, $f_{d2} \equiv f_{d2,g}f_{d2,c}$ for brevity.

```
The different secondary completeness matters
here if proceeding analytically.  Would be
easier to do numerically.  Regardless, this is
not the most interesting question right now.
```

3.10.2. *If we ignore binarity, how wrong is our occurrence rate for planets of radius R_p?*

Write $\Gamma_{\text{A, planets of }R_p} = \Gamma_{{\rm A},R_p}$, and similarly for D. Just as in Model #1, the answer to "what is the relative difference between the occurrence rates derived by Astronomers D and A for planets of radius R_p?" is simpler when we assume that Astronomer A has also derived a completeness, which we assume is the same as for Astronomer D in the single star case. So Astronomer A now misclassifies planetary radii, and miscounts the total number of stars, but somehow knows his completeness for single stars. This is possible, because Astronomer A can correctly derive a completeness estimate and geometric transit probability for single stars only. For Astronomer A' below, including the planets in binary star systems in the numerator introduces mistakes in the estimated completeness, and so the picture is more complicated. Then

$$\Gamma_{{\rm A},R_p} = \frac{N_{\rm det,s}/(f_{s,g}f_{s,c})}{N_s+N_d}. \tag{87}$$

The correction factor X_Γ is

$$\begin{aligned} X_\Gamma &= \frac{\Gamma_{{\rm D},R_p}}{\Gamma_{{\rm A},R_p}} && (88)\\ &= \frac{\Gamma_{t,s}N_s+(\Gamma_{t,d1}+\Gamma_{t,d2})N_d}{N_s+2N_d}\cdot\frac{N_s+N_d}{\Gamma_{t,s}N_s} && (89)\\ &= \frac{(1+\beta(\Gamma_{t,d1}+\Gamma_{t,d2})/\Gamma_{t,s})(1+\beta)}{(1+2\beta)} && (90)\\ &= (1+\chi\beta)\cdot\frac{1+\beta}{1+2\beta} && (91)\end{aligned}$$

for

$$\beta \equiv N_d/N_s, \quad \chi \equiv (\Gamma_{t,d1}+\Gamma_{t,d2})/\Gamma_{t,s}. \tag{92}$$

In the limit where the R_p planet has the same occurrence rate about every host type regardless of its mass, $\chi = 2$, and this simplifies to $X_\Gamma = 1+\beta$. This is same expression as what we found for Model #1, but now the value of β will be smaller in this model because of the smaller volume of searchable binary stars. From numerics, we find $\beta = N_d/N_s = 0.59$, so the occurrence rate correction factor is $X_\Gamma(\chi=2) = 1.59$.

If we consider the opposite limit of the R_p planet *only* existing around single stars and the primaries of double star systems with equal occurrence, $\chi = 1$, and $X_\Gamma = (1+\beta)^2/(1+2\beta)$. For $\beta = 0.59$, this gives $X_\Gamma(\chi=1) = 1.16$. The true correction factor is somewhere between these two limits.

3.10.3. *What if we ignore binarity, but count derived planet radii that are "close enough"?*

Let x be the acceptable margin of error on the observed radius, so that we are interested in counting the number of planets detected with $R_p < R_p^{\rm obs} < (1\pm x)R_p$. For Model #2, since all planets have radius R_p, we only care about the smaller radius case, *i.e.*, we want to count $R_p > R_p^{\rm obs} > (1-x)R_p$, for instance with $x = 0.1$.

The diluted detections only occur in binary systems. The number of planets detected in double star systems with observed radius $R_p^{\rm obs}$ is

$$\begin{aligned} N_{\rm det,d}(R_p^{\rm obs}) &= N_{\rm det,d1}(R_p^{\rm obs}) + N_{\rm det,d2}(R_p^{\rm obs}) && (93)\\ &= \Gamma_{t,d1}f_{d1}(R_p^{\rm obs})N_{d1}(R_p^{\rm obs}) + \\ &\quad \Gamma_{t,d2}f_{d2}(R_p^{\rm obs})N_{d2}(R_p^{\rm obs}). && (94)\end{aligned}$$

As always, the fractional terms f_{di} for $i \in 1,2$ encompass geometric and completeness corrections. We're writing both them and the number of stars for which a given observed radius would be seen as functions of $R_p^{\rm obs}$.

To save on notational horribleness, write $R_p' \equiv \{(1-x)R_p < R_p^{\rm obs} < R_p\}$, in other words let R_p' denote the set of planets that are observed in the desired radius range.

Then writing $N_{d1}(R_p')$, the number of double systems for which a (R_p,P) planet would be observed in R_p' if it orbiting the primary, is relatively straight-forward:

$$N_{d1}(R_p') = \int_0^{\min(1,\gamma_{R,u})} N_d(\gamma_R)\,{\rm prob}(\gamma_R)\,{\rm d}(\gamma_R), \tag{95}$$

where the upper limit is set by equating the limiting radius $(1-x)R_p$ to the observed diluted radius $(1+\gamma_R)^{-1/2}R_p$, giving

$$\gamma_{R,u} \equiv \frac{x(2-x)}{(1-x)^2}. \tag{96}$$

If $x > (1-2^{-1/2})$, bad things happen (cf. Fig. 7), hence the need for the minimum. Note that Eq. 95 is not an expression for the number of detected planets about primaries – that quantity is expressed as $\Gamma_{t,d1}f_{d1}(R_p^{\rm obs})N_{d1}(R_p^{\rm obs})$.

The analogous equation to Eq. 95 for $N_{d2}(R_p')$, the number of double systems for which a (R_p,P) planet would be observed in R_p' if it orbited the secondary, is

$$N_{d2}(R_p') = \begin{cases} \int_{\gamma_{R,l}}^{1} N_d(\gamma_R)\,\mathrm{prob}(\gamma_R)\,\mathrm{d}(\gamma_R), & \\ & x > 1 - 2^{-1/2} \\ 0, & x \leq 1 - 2^{-1/2} \end{cases} \quad (97)$$

where

$$\gamma_{R,l} \equiv \frac{(1-x)^2}{x(2-x)}. \quad (98)$$

Eq. 97 accounts for the fact that you cannot measure R_p to better than $\approx 29\%$ if the planet orbits the secondary of an unidentified binary. This was originally noted in Fig. 7.

If we were forward-modelling, *i.e.* if we actually wanted to compute $N_{\mathrm{det,d}}(R_p')$, we would also need to evaluate the total detection efficiency term $f_{d1}(R_p^{\mathrm{obs}})$, which is a product of the geometric transit probability and the fraction of signals of a given depth that are detectable. We would need to integrate this product over the domain of allowable radii R_p' to find the number of detected planets.

However, the procedure of estimating the geometric transit probability, as well as the completeness, is complicated by the errors that are introduced by including binary star systems that are thought to be single in the numerator. These stars will have incorrectly estimated masses and radii, which influences both $f_c(R_p')$ and $f_g(R_p')$.

`If we ask the right question, we will not need to actually compute either term.`

Astronomer A misclassifies planetary radii, miscounts the total number of stars, and (incorrectly) estimates his completeness and geometric correction for "single" stars as a function of radius $f(R_p')$ (omitting a "s" subscript because the stars for which this incorrect occurrence rate are derived are not all single). His occurrence rate for planets of radius *near* R_p, *i.e.* the set of planets with $\{(1-x)R_p < R_p^{\mathrm{obs}} < R_p\}$, is

$$\Gamma_{\mathrm{A},R_p'} = \frac{(N_{\mathrm{det,s}} + N_{\mathrm{det,d}}(R_p'))/f(R_p')}{N_s + N_d}. \quad (99)$$

For him,

$$f(R_p') = f_g(R_p') f_c(R_p'). \quad (100)$$

Astronomer D, who had everything right, changes nothing.

The new correction factor is

$$X_\Gamma = \frac{\Gamma_{\mathrm{D},R_p}}{\Gamma_{\mathrm{A},R_p'}} \quad (101)$$

$$= \frac{1+\chi\beta}{1+\xi} \cdot \frac{1+\beta}{1+2\beta}, \quad (102)$$

for

$$\beta \equiv N_d/N_s, \quad \chi \equiv (\Gamma_{t,d1} + \Gamma_{t,d2})/\Gamma_{t,s}, \quad (103)$$

and

$$\xi \equiv \frac{N_{\mathrm{det,d}}(R_p')}{f(R_p')} \cdot \frac{1}{N_s \Gamma_{t,s}}. \quad (104)$$

If we let $x < 1 - 2^{-1/2}$, for instance $x = 0.1$, then $N_{d2}(R_p') = 0$. The expression for ξ in this case becomes

$$\xi = \frac{N_{d1}(R_p')}{N_s} \cdot \frac{\Gamma_{t,d1}}{\Gamma_{t,s}} \cdot \frac{f_{d1}(R_p')}{f(R_p')} \quad (105)$$

$$= \frac{N_{d1}(R_p')}{N_s} \cdot \frac{\Gamma_{t,d1}}{\Gamma_{t,s}} \cdot \frac{f_{g,d1}(R_p')}{f_g(R_p')} \cdot \frac{f_{c,d1}(R_p')}{f_c(R_p')}, \quad (106)$$

provided that $x < 1 - 2^{-1/2}$.

We have now parametrized our ignorance. The values that the latter two ratios of Eq. 106 take depend on what we assume Astronomer A' does to derive the stellar parameters for the systems that he thinks are singles but are in fact doubles. For instance, assume he somehow observed the bolometric flux from every point on the sky, and used it with a distance to obtain single-star luminosities, and thus radii and masses. His binary systems (thought to be singles) would be bluer than either component star is in reality. On the main sequence, $\rho_\star$ decreases with increasing stellar radius, so $\rho_\star$ would be underestimated for all binaries. Since $f_g \propto \rho_\star^{1/3}$, this means that $f_{g,d1}(R_p')/f_g(R_p') > 1$. However, by the same token we can place an *upper* bound on this ratio. In the worst case scenario a given star's luminosity is twice the individual star's luminosity. If $L \propto M^3 \propto R^3$, then the radius is over-estimated by at most a factor of $2^{1/3}$. Thus the density is underestimated by at most a factor of $2^{1/2}$, and $f_g(R_p')$ is underestimated by at most a factor of $2^{1/6} = 1.12$.

This begins to take us somewhere. The left-most term of Eq. 106 is bounded above by β. This is because $N_{d1}(R_p') \leq N_{d1} = N_d$. If we assume $\Gamma_{t,d1} = \Gamma_{t,s}$, as we were doing anyway, then the product of the three left-most fractions must be less than 1.12β. The only thing standing in our way of an analytic upper bound to ξ is the completeness fraction $f_{c,d1}(R_p')/f_c(R_p')$. Note that over a fixed radius interval dilution will always yield a lower completeness fraction for double star systems than singles, so $f_{c,d1}(R_p')/f_{c,s}(R_p') \leq 1$.

For the completeness term, an argument that I don't wholly believe is that

$$\frac{f_{c,d1}(R_p')}{f_c(R_p')} = \frac{f_{c,d1}(R_p')}{f_{c,s}(R_p')} \cdot \frac{f_{c,s}(R_p')}{f_c(R_p')} \lesssim 1 \cdot (1 + \langle\gamma_R\rangle)^3, \quad (107)$$

where the first part of the inequality is obvious, but the second is by assuming it's $\lesssim f_{s,c}/f_{d,c}$, which might be wrong because of the radius dependence.

A way of making the estimate without deriving the actual completeness fractions, or even running numerics, is just to plot the error as a function of ξ. We do this in Fig. 8, which shows that the correction factor varies from 0.6 to 1.6, depending on what fraction of secondaries you assume host planets.

I could run numerics for this to make the result more interesting.

4. MODEL # 3: A SYNTHETIC *KEPLER* ANALOG

If we assume every KIC star to be single, what order of magnitude of error do we make in occurrence rates estimated for planets of different sizes (and periods)?

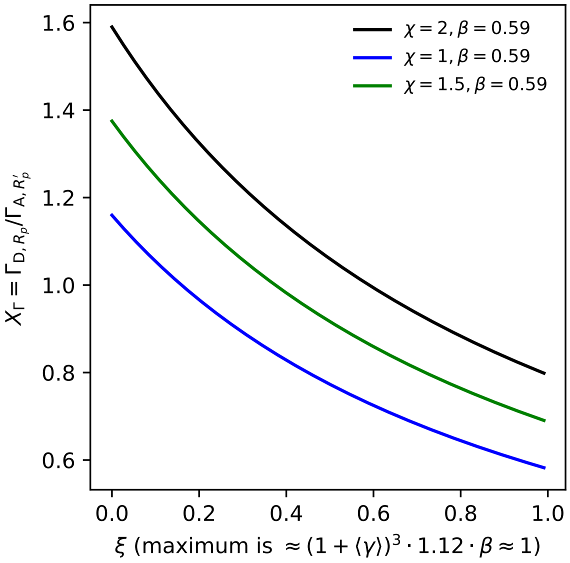

FIG. 8.— *Y axis*: correction factor for the astronomer who ignores binarity but counts planetary radii that are "close enough" (A') and the astronomer who knows everything (D). *X axis*: χ as defined in Eq. 106.

We can attempt an answer with Monte Carlo simulations of something like the *Kepler* field. I say "something like" because the binarity properties of the *Kepler* field (or more specifically the stars selected in the KIC) are not known.

but it might be better to do something directly on the KIC's selected target stars, like what Ciardi did! The following is a good example, but at most it can "suggest" that a similar level of error has been made from the *Kepler* survey.

The plan is as follows:

- Verify Galaxia can produce something resembling the KIC.
- Introduce binaries (Galaxia does not, by default, have any).
- Use Galaxia [Sharma et al 2010] to construct an analog of the KIC.
- Introduce a planet population.
- Numerically assess what occurrence rates would be derived, and compare to previous sections.

4.1. *Verifying Galaxia can produce a KIC analog*

Galaxia [Sharma et al 2010] is a stellar population synthesis code for creating surveys of the Milky Way. It samples an assumed analytic distribution function. In other words, it assumes a number density of stars as a function of position, velocity, age, metallicty, and mass, from which it then samples to construct a "realistic" stellar population. The assumed model includes:

- A metallicity distribution (log-normal),
- A velocity distribution (triaxial Gaussian),

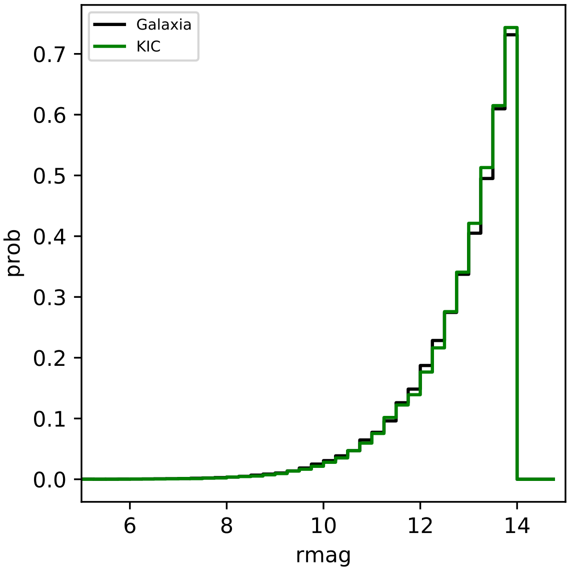

FIG. 9.— Probably density distributions of r magnitude for $r < 14$ stars within 7.5° of the center of the *Kepler* field, for Galaxia and the KIC.

- The stellar density and galactic components specified by Robin et al [2003] (the Besançon model) – see Sharma et al [2010] Table 1. This includes the IMF for a young and old thin disk, a thick disk, a spheroidal component, a bulge component, as well as an ISM and dark halo. It also includes the star-formation rate.

While this model of course lacks some detail, Sharma et al [2010] show that agreement with the volume-limited Hipparcos sample is probably good enough for our purpose of estimating the number of "solar-like" (by definition hereafter, $0.7M_\odot < M_\star < 1.3M_\odot$) stars in something like the *Kepler* field.

We then run the Galaxia in its "circular survey" mode, towards the direction of the *Kepler* field[6]. We apply the Schlegel, Finkbeiner et al [1999] extinction map when converting absolute to apparent magnitudes.

To verify the results are in reasonable agreement with the stars that actually exist in that direction, we compare the output with the KIC[7] [Brown et al 2011]. Following Sharma et al [2016, ApJ 822:15], for each catalog we selected all stars within 7.5° of the center of the *Kepler* field, and with Sloan $r < 14$. The KIC is complete to magnitudes fainter than $r = 14$, so the comparison should not be affected by its completeness.

A few distributions are shown in Figs. 9- 11. From Fig. 9, we can see that the stellar counts as a function of Sloan r magnitude are in close agreement with the KIC. From Fig. 10, the KIC reports relatively more dwarf stars, and fewer giants than Galaxia does. This is likely related to KIC's inability to distinguish between dwarfs and giants with only limited photometry – although Brown et al [2011]'s idea of using a specific "D51"

[6] galactic longitude 76.532562, galactic latitude 13.289502.

[7] From http://archive.stsci.edu/kepler/kic.html, we downloaded the 13.1 million row ”—”-delimited gzipped ASCII file containing the complete Kepler Input Catalog (version 10).

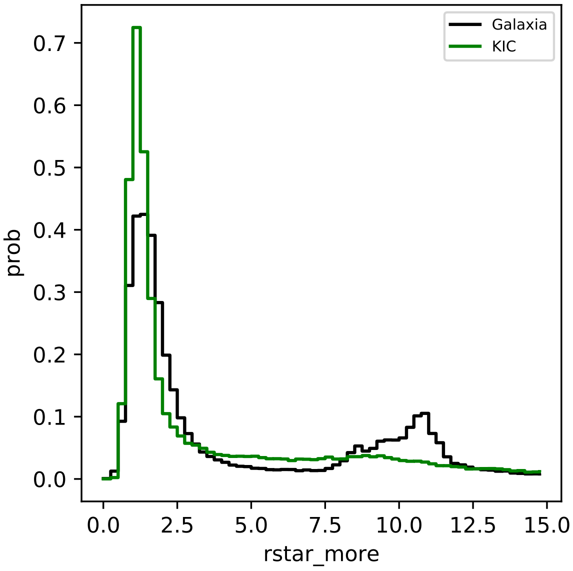

FIG. 10.— Same as Fig. 9, for $R_\star$.

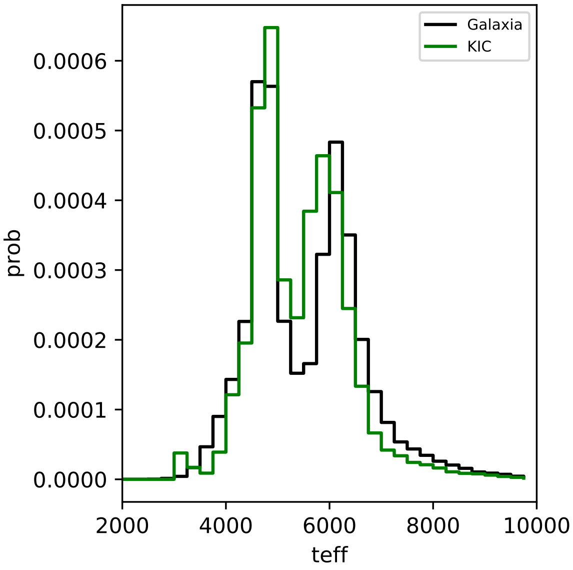

FIG. 11.— Same as Fig. 9, for $T_{\rm eff}$.

photometric band as a $\log g$ diagnostic was reported to have helped with this problem. In Fig. 11, we see something resembling the K dwarf desert again (5000K ≈ K3V dwarf). Although there are a few slightly worrying discrepancies (notably in Fig. 10), we get the idea that Galaxia at least roughly resembles "reality" (inasmuch as the reported KIC parameters represent reality!).

4.2. *Introduce binaries*

By default, Galaxia does not include binaries. However, from Fig. 9, we've seen that its number counts as a function of apparent magnitude are good. So we assign binarity to the Galaxia stars as follows.

- Assume a binary fraction BF = 0.45 [Raghavan et al 2010].
- Ignore higher order multiples.
- If a star is drawn to be a binary:

  Its reported magnitude becomes a system magnitude.

  Draw the binary's mass ratio q from the probability distribution function for a magnitude limited survey, i.e. from Fig. 4. This is sensible so long as the pdf is independent of the primary's mass. We assume this to be the case over $0.7 - 1.3 M_\odot$[8].

  For $q \geq 5/7$, it turns out the above is sufficient information to analytically specify the individual stellar masses, and thus luminosities with the relation shown in Fig. 2.

  For $q < 5/7$, the individual stellar masses are not uniquely specified[9]. Thus for this latter fraction of the population, we draw the primary mass from the pdf of single star primary masses.

  With the individual masses known, use Eq. 57 to assign a radius to each star in the binary system.

Benefits of the above procedure include that it does not change Galaxia's star counts as a function of magnitude. It produces the correct distribution of the mass ratio for the binaries in a magnitude limited sample, and thus a reasonable light ratio distribution.

The drawbacks are that the individual component masses this procedure produces are not drawn from the same distribution throughout. This might bias any subsequent synthetic transit survey. From Fig. 14, it's likely not excessively important (it's more important to get the mass ratio and light ratio distributions correct).

In addition, the procedure produces a different mass-radius relation for stars in binary systems, vs. stars in single star systems. This might also bias any subsequent synthetic transit survey. The mass-radius relation for single stars comes from the Padova isochrones [Marigo et al 2008, Marigo & Girardi 2007, Girardi et al. 2000, Bertelli et al 1994] used in Galaxia. Although Galaxia does not directly report radii, it reports luminosities and effective temperatures, from which we compute the radii. The alternative choice would be to use the reported Galaxia masses with our Eq. 57 to produce radius estimates. This would produce consistency in the mass-radius relation, but would break $L = 4\pi R^2 \sigma T_{\rm eff}$ for single star systems. The different mass-radius relations are shown in Fig. 16. Generally, it seems that the at a given mass, the stars in binary systems are biased to have slightly greater radii. (Ignoring the evolved stars). However this is a small effect – perhaps of order 10% in

[8] It was not immediately self-evident to me that this distribution was the correct choice, since for fixed mass ratio as the primary mass changes so does the detectable volume. However, as long as prob(q) is independent of the primary's mass, the average distribution over all the possible detectable volumes will remain the same.

[9] See the derivation of 2017/08/25. This is because of the mass luminosity relation we specified, and the fact that we have q, and L_d, but want individual masses, and individual luminosities.

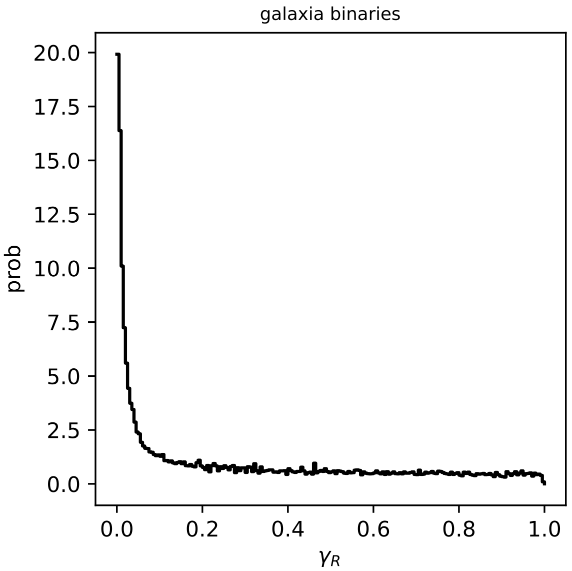

FIG. 12.— Light ratio distribution for the $r < 17$ Galaxia sample discussed in Sec. 4.2.

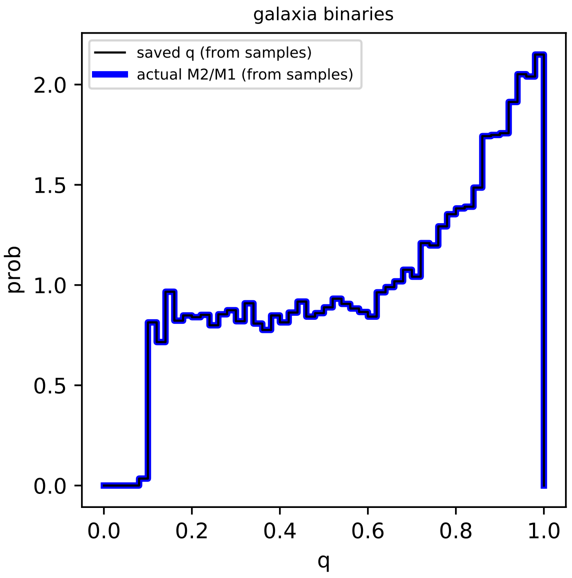

FIG. 13.— Mass ratio distribution for the $r < 17$ Galaxia sample discussed in Sec. 4.2.

the radius. This will affect the transit depths in a synthetic transit survey, making it more difficult to detect planets around binaries.

4.3. *Using Galaxia to construct a KIC analog*

With glowing confidence, we then proceed to construct a KIC analog as follows. We take all Galaxia stars within 7.5 degrees of the center of the *Kepler* field, and take stars with $r < 17$.

a different magnitude cut, e.g. $r < 16$, might matter?

This is now beyond where the KIC is complete, so direct comparisons of distributions with the KIC will be confused. We then apply an analog of the prioritization

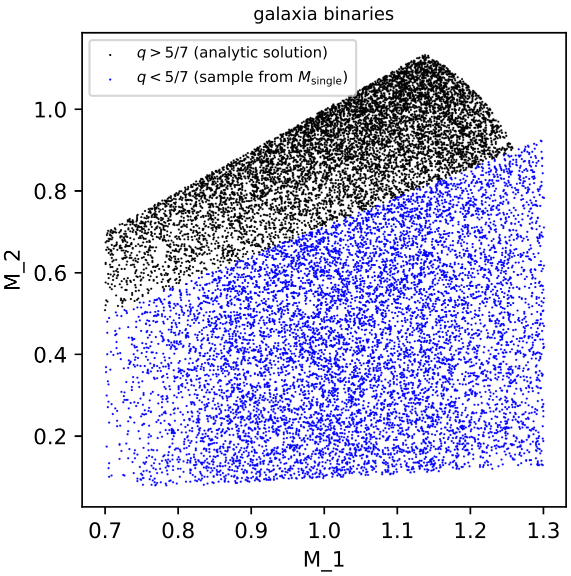

FIG. 14.— Scatter plot of primary and secondary mass (solar units).

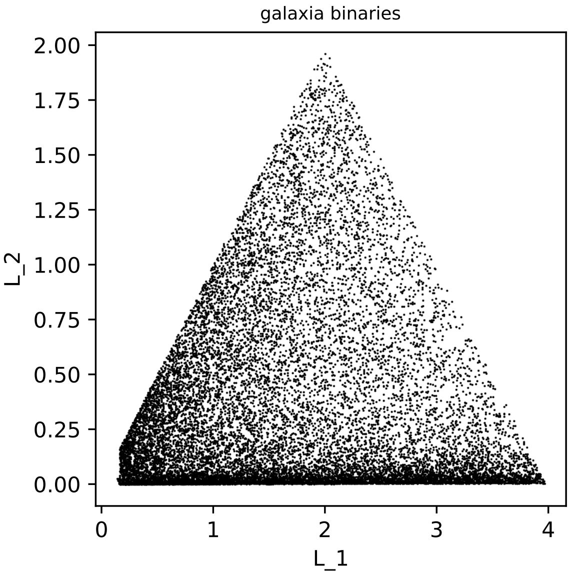

FIG. 15.— Scatter plot of primary and secondary luminosity (solar units).

scheme used by the *Kepler*mission to select target stars, described by Batalha et al [2010]'s Table 1.

For each star, we compute the minimum detectable planet radius at three semimajor axes: (1) the inner radius of the so-called "habitable" zone,

$$a_{\rm HZ} = 0.95\,{\rm AU}\left(\frac{L}{L_\odot}\right)^{1/2}, \qquad (108)$$

(2) $0.5a_{\rm HZ}$, and (3) $5R_\star$. We make the same assumptions as Batalha et al [2010] did about the transit duration. We assume a noise model of pure shot noise, rather than using any CDPP estimates. We then apply the Batalha et al [2010] Table 1 prioritization, except that we use SDSS

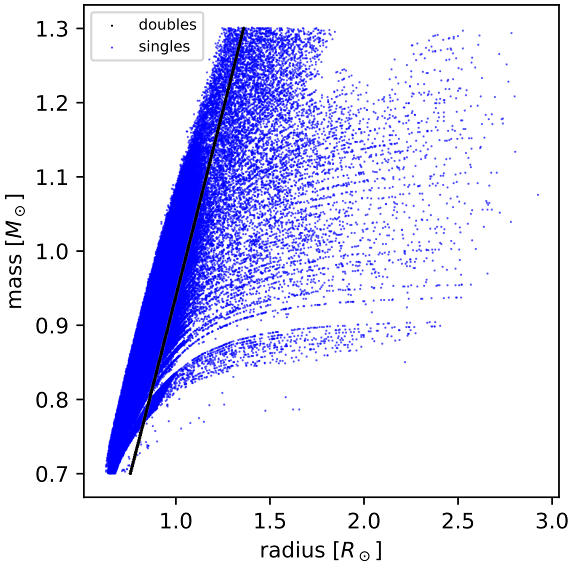

FIG. 16.— Scatter plot of mass vs radius for singles and binaries. For binaries, different points are shown for the primary and secondary of a given system. Only points with $0.7 < M_\star/M_\odot < 1.3$ are selected.

Criteria
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 1R_e$; $a =$ HZ; $Kp < 13$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a =$ HZ; $Kp < 13$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 1R_e$; $a =$ HZ; $Kp < 14$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a =$ HZ; $Kp < 14$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 1R_e$; $a =$ HZ; $Kp < 15$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 1R_e$; $a =$ HZ; $Kp < 16$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 1R_e$; $a = \frac{1}{2}$HZ; $Kp < 14$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a = \frac{1}{2}$HZ; $Kp < 14$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a = 5R_*$; $Kp < 14$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a =$ HZ; $Kp < 15$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a =$ HZ; $Kp < 16$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a = 5R_*$; $Kp < 15$
$N_{\rm tr} \geqslant 3$; $R_{p,\min} \leqslant 2R_e$; $a = 5R_*$; $Kp < 16$

FIG. 17.— Prioritization applied by Batalha et al [2010], from whom I copy-pasted this Table. I did the same, but using r magnitudes instead of Kp.

r magnitudes instead of Kepler Kp magnitudes. The rankings are shown in Fig. 17. We get similar (within a factor of two) counts as they did towards the end of the table, though at the beginning we seem to have fewer stars[10].

Note that, since we assigned binarity to Galaxia stars, applying the Batalha et al [2010] prioritization procedure requires estimating *incorrect* single star parameters for what are really binary star systems! We do this by keeping the total system luminosity, inverting Fig. 2 to find an incorrect mass, and then applying Eq. 57 to get an incorrect radius. We can then calculate the minimum detectable planet radius as

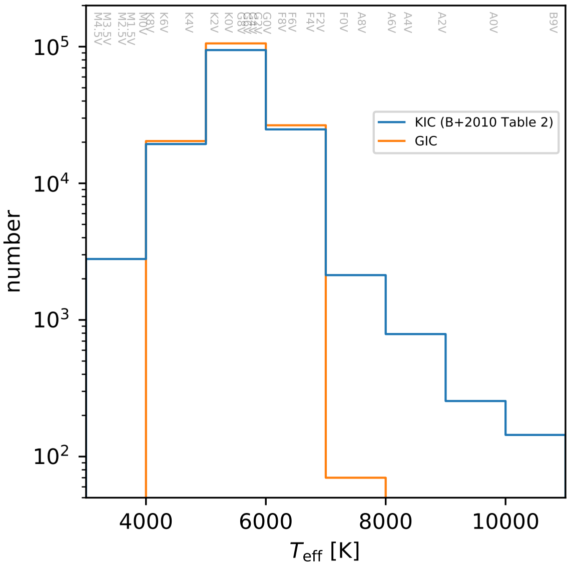

FIG. 18.— Effective temperature distribution of the stars in our synthetic Kepler input catalog analog – the Galaxia input catalog. Note that the effective temperatures of binary star systems in the GIC are intentionally wrongly estimated, as if they were single star systems. The listed $T_{\rm eff}$ to spectral type correspondence is from Mamajek's tables.

$$R_{p,\min} = \left(\frac{7.1\sigma_{\rm tot}}{r}\right)^{1/2} R_\star, \tag{109}$$

for $r = 1$ the dilution that we ignore, and

$$\sigma_{\rm tot} \equiv {\rm N} = (F_\gamma^{\rm N} A N_{\rm tra} T_{\rm dur})^{-1/2} \tag{110}$$

In applying the prioritization, we omit the latter two priority classes of Fig. 17, to obtain a catalog of 152682 stellar systems, which we hereafter refer to as the *Galaxia Input Catalog (GIC)*. We compare the effective temperature distribution of the GIC with that of the KIC, reported by Batalha et al [2010] Table 2, in Fig. 18. Evidently, the GIC is dominated by sun-like stars, in a very similar manner to the exoplanet target stars of the KIC.

4.4. *Introduce a planet population*

As in Sec. 3, we allow for three different occurrence rates: $\Gamma_{t,s}$, the fraction of stars in single systems with a planet of radius R_p and orbital period P, and also $\Gamma_{t,d1}$ and $\Gamma_{t,d2}$ for the fraction per sun-like primary and secondary (respectively) of double star systems with a planet of (R_p, P).

We then randomly select which stars get a planet, and using the true stellar masses and radii compute the corresponding impact parameters for the planets as

$$b = \frac{a}{R_\star}\cos i, \quad \cos i \sim \mathcal{U}(0,1). \tag{111}$$

Any planet with $|b| < 1$ is transiting, and has its transit duration evaluated as

$$T_{\rm dur} = 13\,{\rm hr}\left(\frac{P}{\rm yr}\right)^{1/3}\left(\frac{\rho_\star}{\rho_\odot}\right)^{-1/3}\sqrt{1-b^2}, \tag{112}$$

[10] I suspect this is because of the slightly different noise calculation – and perhaps Batalha using $N_{\rm sample}$ rather than just counting photons matters

which assumes circular orbits.

4.5. *Run the survey*

We run the following survey: $A = 0.708\,\mathrm{m}^2$ (*Kepler* effective area), $T_{\rm obs} = 20\,$years (to detect more Earth-like planets than *Kepler* did), and $x_{\min} = 7.1$, the same signal to noise threshold as *Kepler*.

For simplicity, we assign a r band zero point of $10^6\,\mathrm{ph/s/cm}^2$ to correspond to an $r = 0$. We use this to compute the observed photon number fluxes for every system, regardless of whether it has a planet. We then compute the signal to noise distribution of transit events, and the number of detections, according to Eq. 20.

4.6. *Numerically assess what occurrence rates would be derived, and compare to previous models*

4.6.1. *If we ignore binarity, how wrong is our occurrence rate for planets of radius R_p?*

This is similar to what we did in Sec. 3.10.2, with the exception that we need to specify the calculation more precisely.

Specifically, our "Astronomer A" in this case will estimate an occurrence rate for planets of radius R_p and orbital period P as

$$\Gamma_{\mathrm{A},R_p} = \frac{N_{\rm det,s}}{Z}, \tag{113}$$

for

$$Z \approx (N_s + N_d)\frac{1}{J}\sum_{j=1}^{J} Q_j, \tag{114}$$

for Q_j the total detection efficiency, indexed over single and double systems, given as the product of the geometric transit probability and the system-specific completeness:

$$Q_j \equiv f_g^{(j)} f_c^{(j)}. \tag{115}$$

Keep in mind that for Astronomer A, we need to compute the *incorrect* transit probability for double star systems (since Astronomer A is assuming that they are single stars).

To evaluate the system-level completeness, we assume that Astronomer A performs something equivalent to perfect injection-recovery. For a stellar system with known properties (stellar and planetary), and a perfectly known noise model, the probability of detection is a binary function: either $\mathrm{S/N} > (\mathrm{S/N})_{\min}$, or it is not[11]. So we compute $f_c^{(j)}$ as Astronomer A would: an array (over star systems) of zeros wherever the (R_p, P) planet could not be detected, and ones where it could. We then evaluate the total detection efficiency Q_j for each system, and compute the occurrence rate as in Eq. 113. The resulting occurrence rates are only slightly smaller than those in Sec. 4.6.2 discussed below and shown in Fig. 19.

4.6.2. *What if we ignore binarity, but count planets whose derived radii are "close enough" to R_p?*

As in Sec. 3, we now count planets with observed radii from $(1-x)R_\oplus < R_p^{\rm obs} < R_\oplus$. We'll use $x = 0.1$, but it can be whatever we want. The occurrence rate in this case will be

$$\Gamma_{A,R_p'} = \frac{N_{\rm det,s} + N_{\rm det,d}(R_p')}{Z'}, \tag{116}$$

for

$$Z' \approx (N_s + N_d)\frac{1}{J}\sum_{j=1}^{J} Q_j', \tag{117}$$

where the geometric transit probability is the same as before, but the system-specific completeness is now a function of the desired radius interval:

$$Q_j' = f_g^{(j)} f_c^{(j)}(R_p'). \tag{118}$$

There are three possible cases for the completeness, specified by the minimum detectable planet radius $R_{p,\min}$ (computed according to Eq. 109):

$$f_c^{(j)}(R_p') = \begin{cases} 1 & \text{if } R_{p,\min} > R_p, \\ 0, & \text{if } R_{p,\min} < (1-x)R_p, \\ \frac{R_p - R_{p,\min}}{R_p - (1-x)R_p}, & \text{otherwise.} \end{cases} \tag{119}$$

The latter term above is the case in which the minimum detectable planet radius happens to be between $(1-x)R_p$ and R_p. In that case the completeness becomes the fraction of planets in that interval that would be detected. Assuming a uniform prior for the planet radius distribution $R_p \sim \mathcal{U}((1-x)R_p, R_p)$, this becomes the expression given in Eq. 119.

These occurrence rates can then be compared to the "true occurrence rate" derived by the all-knowing Astronomer D:

$$\Gamma_{\mathrm{D},R_p} = \frac{\Gamma_{t,s}N_s + (\Gamma_{t,d1} + \Gamma_{t,d2})N_d}{N_s + 2N_d}. \tag{120}$$

A note in passing: the more information-rich parameters for Astronomer D to report would be $(\Gamma_{t,s}, \Gamma_{t,d1}, \Gamma_{t,d2})$, rather than Γ_{D,R_p}. But I digress.

Similar to Fig. 8, Fig. 19 shows that the magnitude of the correction needed by the $\Gamma_{\mathrm{A},R_p'}$ estimate depends strongly on the fraction of secondaries that are assumed to host the (R_p, P) planets. The simplest assumption is that the occurrence rate is the same for all solar type stars: any "Sun-like" $(0.7 < M_\star/M_\odot < 1.3)$ star hosts planets at the "true rate" Γ_t, and any other mass of star does not. Given how we have defined the GIC, this means that all single stars and primaries of double star systems are sun-like. The fraction of Sun-like secondaries in this case is 37.4% of all secondaries. Reading the appropriate value off Fig. 19, this corresponds to $\Gamma_{\rm D} = 1.39\Gamma_{\rm A'}$. This gives a relative percentage error of $\delta\Gamma_{\mathrm{D},R_p} \equiv (\Gamma_{\rm A,R_p'} - \Gamma_{\rm D,R_p})/\Gamma_{\rm D,R_p} = -28\%$.

[11] There are probabilities involved if we allow for any uncertainty in the parameters, but this is too complicated for our current goal.

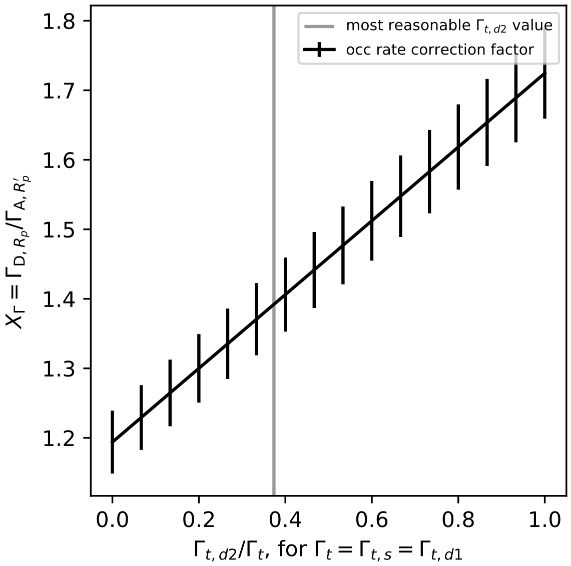

FIG. 19.— The "most reasonable" $\Gamma_{t,d2}$ value is obtained by assuming that the occurrence rate of (R_p, P) planets is the same for all solar type stars, and zero otherwise. This applies most immediately to the question of Earth-like planets around Sun-like stars.