THE EXOPLANET CENSUS: A GENERAL METHOD APPLIED TO KEPLER

ANDREW N. YOUDIN

Harvard Smithsonian Center for Astrophysics, 60 Garden Street, MS-16, Cambridge, MA 02138, USA

Received 2011 May 8; accepted 2011 August 19; published 2011 November 3

ABSTRACT

We develop a general method to fit the underlying planetary distribution function (PLDF) to exoplanet survey data. This maximum likelihood method accommodates more than one planet per star and any number of planet or target star properties. We apply the method to announced Kepler planet candidates that transit solar-type stars. The Kepler team’s estimates of the detection efficiency are used and are shown to agree with theoretical predictions for an ideal transit survey. The PLDF is fit to a joint power law in planet radius, down to 0.5 \( R_\oplus \), and orbital period, up to 50 days. The estimated number of planets per star in this sample is \( \sim 0.7-1.4 \), where the range covers systematic uncertainties in the detection efficiency. To analyze trends in the PLDF we consider four planet samples, divided between shorter and longer periods at 7 days and between large and small radii at 3 \( R_\oplus \). The size distribution changes appreciably between these four samples, revealing a relative deficit of \( \sim 3 \) \( R_\oplus \) planets at the shortest periods. This deficit is suggestive of preferential evaporation and sublimation of Neptune- and Saturn-like planets. If the trend and explanation hold, it would be spectacular observational support of the core accretion and migration hypotheses, and would allow refinement of these theories.

Key words: methods: statistical – planetary systems – planets and satellites: detection – planets and satellites: dynamical evolution and stability – planets and satellites: formation

Online-only material: color figures

1. INTRODUCTION

Individual exoplanet discoveries highlight the extraordinary diversity of worlds in the solar neighborhood (Mayor & Queloz 1995; Marcy & Butler 1996; Charbonneau et al. 2009; Lissauer et al. 2011a). For an accurate census of the planet population, the statistical analysis of large samples of exoplanets is required (Cumming et al. 2008; Howard et al. 2010). Trends revealed by the data provide powerful constraints on dynamical theories of planet formation (Goldreich et al. 2004; Kenyon & Bromley 2006; Youdin 2010).

The correlation of giant planets with host star metallicity is perhaps the most interesting trend revealed by radial velocity (RV) surveys (Gonzalez 1997; Fischer & Valenti 2005; Johnson et al. 2010). This trend has been a powerful guide for identifying the mechanisms responsible for planetesimal formation (Youdin & Shu 2002; Youdin & Goodman 2005; Johansen et al. 2009); see Chiang & Youdin (2010) for a review.

The Kepler transit survey is revolutionizing the field of exoplanets from space (Borucki et al. 2010). Borucki et al. (2011, hereafter BKep11) released 1235 partially vetted planet “candidates.” Morton & Johnson (2011) estimate the false-positive rate as \( \lesssim 5\% \) around the brighter stars that are considered here. For brevity and statistical purposes, we mostly refer to the candidates as “planets,” though significant follow-up work remains with only \( \sim 20 \) candidates currently classified as “confirmed.”

Howard et al. (2011, hereafter HKep11) present a detailed statistical analysis of the \( \sim 500 \) BKep11 planets around \( \sim 60,000 \) bright, solar-type stars. Accounting for detection efficiencies, HKep11 report planet occurrence rates of \( \sim 0.2 \) planets per star for radii \( > 2 \) \( R_\oplus \) and orbital periods \( < 50 \) days. The purpose of this paper is to apply different statistical methods to the Kepler data set, making use of the estimates of detection efficiencies—whose importance is on par with the detections themselves—provided by HKep11.

Our method is based on the Tabachnik & Tremaine (2002, hereafter TT02) technique to analyze the mass and period distributions of planets from RV surveys. TT02 emphasized the importance of determining planet mass and period distributions simultaneously. A simultaneous fit is crucial because RV detection thresholds depend on both the mass (times the sine of inclination) and the period. Transit surveys have a similar, though weaker, coupling between planet radius and orbital period in the efficiency of discovering transits.

A primary advantage of the TT02 technique is that it naturally accommodates more than one planet per star. Planet hosting is not treated as a binary proposition. This distinction is not just a technical nicety, because the number of planets per star (NPPS) is of order unity, at least. We show that Kepler data already imply about one planet per solar-type star within \( \sim 0.25 \) AU and with a radius bigger than 0.5 \( R_\oplus \). This number is certain to grow with time as longer periods and smaller sizes are probed. High multiplicity rates in the Kepler sample (Latham et al. 2011) means that the fraction of stars that host a planetary system (at most one, of course) can be significantly smaller than the NPPS. We treat the planet population (including known multiples) as a whole and ignore multiplicity issues. Ragozzine & Holman (2010) discuss the value of multi-transiting systems. Kepler shows that the typical number of planets per planetary system is at least \( \sim 3 \), an estimate that increases with the planets’ mutual inclinations (Lissauer et al. 2011b; Tremaine & Dong 2011).

This paper is organized as follows. Section 2 describes the detection efficiencies relevant to the Kepler transit survey, including an analytic fit to discovery efficiencies reported by HKep11 in Section 2.2. Section 3 describes our method for analyzing exoplanet survey data. Section 4 applies the method by fitting Kepler data to a joint power law in radius and period. Differences between the shorter and longer period planets (in Section 4.2) and also between the smaller and larger planets (in Section 4.3) are analyzed. Section 5 gives non-parametric
estimates of planet occurrence and comments speculatively on the consequences of rising period distributions. We conclude with a summary and discussion in Section 6. The Appendix gives an analytic solution for power-law fits to data from an idealized transit survey.

2. SELECTION EFFECTS FOR TRANSIT SURVEYS

A robust statistical analysis must account for selection effects. We quantify selection effects as detection efficiencies, which give the ratio of detectable events to actual planets. Efficiencies from uncorrelated selection effects multiply to give the net detection efficiency, \( \eta \).

The three main selection effects for transit surveys are: (1) the geometric probability that a randomly oriented planetary orbit transits its host star, with radius \( R_* \), and mean density \( \rho_* \), is \( \eta_{\text{tr}} \approx \sum \eta_{\text{tp}}(1 - 1/r_{\text{p}}) \). Eccentricity tends to increase the detectability of planets at fixed \( a \) (Burke 2008).\(^1\) The duration of \( \text{Kepler} \) transits imply a mean eccentricity \( \lesssim 0.1–0.25 \) (Moorhead et al. 2011). The \( \lesssim 10\% \) effect on \( \eta_{\text{tr}} \) is ignored here.

Detection efficiencies are not constants. They can depend on both planet and stellar properties. We consider how the planet radius, \( R \), and orbital period, \( P \), affect efficiencies. See Section 3.3 for a general discussion of which parameters should be included.

The left panel of Figure 1 shows the \( \text{Kepler} \) discovery efficiency, \( \eta_{\text{disc}} \), calculated as described in Section 2.2. The discovery efficiency is nearly unity over a significant range of planetary radii and periods, thanks to the high photometric precision of \( \text{Kepler} \). There is a sharp drop in discovery efficiency for sufficiently small planets whose transit depths compete with photometric noise. At longer periods, the drop in efficiency shifts to larger radii because fewer transits can occur. The right panel of Figure 1 shows that including the geometric transit probability reduces the probability of finding long period planets.

To connect with the extensive work on RV surveys, we note that RV upper limits can be expressed as a detection efficiency for a given star. Most simply, \( \eta = 1 \) for planet parameters above the detection threshold and \( \eta = 0 \) otherwise. This step function can be smoothed to account for confidence intervals.

2.1. Transit Probability

The probability that a planet on a randomly oriented circular orbit of semimajor axis \( a \) transits its host star, with radius \( R_* \) and mean density \( \rho_* \), is

\[
\eta_{\text{tr}} = \frac{R_*}{a} = \left( \frac{3\pi}{G\rho_* P^2} \right)^{1/3} = 0.051 \left( \frac{10 \text{ days}}{P} \right)^{2/3} \left( \frac{\rho_\odot}{\rho_*} \right)^{1/3}
\]

for \( R \ll R_* \). Eccentricity tends to increase the detectability of planets at fixed \( a \) (Burke 2008).\(^1\) The duration of \( \text{Kepler} \) transits imply a mean eccentricity \( \lesssim 0.1–0.25 \) (Moorhead et al. 2011). The \( \lesssim 10\% \) effect on \( \eta_{\text{tr}} \) is ignored here.

For simplicity, we approximate the mean stellar density (or more specifically the mean \( \rho_*^{-1/3} \)) as solar. The densities of planet hosts can be estimated from precise light curve parameters, though corrections for the eccentricity apply (Tingley et al. 2011).

\(^1\) For an instructive explanation see http://oklo.org/2011/03/23/the-eccentricity-distribution/.

Figure 1. Left: the \( \text{Kepler} \) mission’s efficiency of discovering transiting planets around bright solar-type stars vs. planetary radius and orbital period. The discovery efficiency is unity for large planets, but drops sharply below a planet size that increases gradually with orbital period. See the text for details. Right: the net detection efficiency, \( \eta_{\text{disc}} \), combines the discovery efficiency (at left) with the geometric transit probability, which exerts a bias against detecting long period planets of all sizes. (A color version of this figure is available in the online journal.)
The efficiency at the break is quite high, \( \eta_{\text{disc}} = 0.98 \) (and remarkably constant, probably reflecting the high S/N threshold). Note that \( \eta_{\text{disc}} \) drops to 50% \( (r_{\text{disc}} = 1 \text{ in Figure 2}) \) at \( R = 0.54 R_b \), rising roughly from 1 to 2 \( R_{\odot} \) over the periods considered. When discovery efficiencies are small,
\( \eta_{\text{disc}} \approx r_{\text{disc}} \ll 1 \), and the power law in Equation (4) applies directly to \( \eta_{\text{disc}} \).

While the scalings for \( \eta_{\text{disc}} > 0.98 \) are of little practical concern for our study, we report the fit above the break for completeness:

\[
r_{\text{disc}}(R > R_\oplus) = 5.46 \left( \frac{R}{R_\oplus} \right)^{2.74} \left( \frac{P}{\text{day}} \right)^{-0.460}.
\]  

(6)

This fit describes a small fraction of noisy stars, but does not affect our results because \( \eta_{\text{disc}} \approx 1 \) in this regime.

The relevant fit in Equation (4) agrees well with theoretical estimates. For S/N-limited detections, the combined efficiency should scale as

\[
\eta_{\text{disc}}\eta_{\text{tr}} \propto P^{-5/3} R^6; \tag{7}
\]

see, e.g., Equation (7) of Gaudi (2007). The small R limit of Equation (4) gives \( \eta_{\text{disc}}\eta_{\text{tr}} \propto P^{-1.74} R^{6.34} \), rather good agreement.

The efficiencies summarized here only apply to planets detected during Q2 of the Kepler mission (BKep11). Estimates will need to be updated for future data releases. Moreover, future estimates will likely be improved by more detailed studies that include extraction of simulated transits from the actual Kepler data analysis pipeline. Accurate estimates of the Kepler discovery efficiency for all relevant parameters—especially stellar \( T_{\text{eff}} \)—are the most crucial ingredient for determining the frequency of Earth-like planets.

2.3. The Planetary Sample

Tables 1 and 2 of BKep11 list the properties of 1235 planet candidates and their host stars. For the \( \eta_{\text{disc}} \) estimate to be valid, we can only consider the subset of planets that satisfy the conditions set by HKep11, that (a) host stars are within the solar sample set by Equation (2), (b) \( P < 50 \) days, and (c) the transit detection has S/N > 10. The last condition is discussed further below. Applying these filters reduces the number of planet candidates from 1235 to 567, for a detected planet-to-star ratio of just under 1%.

Our “full” planet sample considers

\[
0.5 \text{ days} < P < 50 \text{ days} \tag{8a}
\]

\[
0.5 \; R_\oplus < R < 20 \; R_\oplus. \tag{8b}
\]

Adding these radius and the minimum period cuts only removes five planets (from 567 to 562). Instead of using the BKep11 reported values for \( R \), we compute \( R \) from the reported values for the host star radius, \( R_* \), and \( R/R_* \).

We now discuss relevant differences between our planet sample and that of HKep11. HKep11 conservatively restricted attention to \( R > 2 \; R_\oplus \), because the discovery efficiency is high for these planets (except at longer periods). We extend our analysis down to \( R = 0.5 \; R_\oplus \). This extension includes all small planets in the solar sample, the smallest having \( R = 0.58 \; R_\oplus \). This extension involves (a) trusting the HKep11 efficiencies down to the \( 1 \; R_\oplus \) level to which they were reported and (b) extrapolating the efficiencies down to \( 0.5 \; R_\oplus \). There are 19 planets in the sample with \( 0.5 < R/R_* < 1.0 \). Both the smooth variation of the reported efficiencies and their agreement with theory (shown above) increase our confidence in making the extrapolation. We also report results for samples of larger planets that are unaffected by this potentially risky extrapolation.

HKep11 used different S/N values than the publicly reported BKep11 S/N values on which we rely. BKep11 define S/N based on only one quarter of data (Q2, Kepler’s first full quarter). By contrast, BKep11 report S/N values up to Q5, so that roughly four times more data have been collected. Our goal is to translate the BKep11 S/N values to the Q2 S/N values for which the \( \eta_{\text{disc}} \) estimate is valid. However, this translation is not so simple.

With perfect noise scaling, the BKep11 S/N values should be roughly twice as high as those used by HKep11. Table 2 of HKep11 presents noise scalings for four planets with \( \sim 2.5 \; R_\oplus \) radii and \( \sim 40 \) day periods. The ideal scaling roughly holds for three objects, but the S/N is saturated and only rose modestly for the fourth. It is unclear what the general noise scaling is and how it varies with size and toward shorter periods.

We proceed cautiously by performing our analyses on both an S/N > 10 and an S/N > 20 planet sample (as defined by BKep11). These cases cover the complete range of possibilities between saturated noise and perfect scaling. Fortunately, the adopted S/N threshold has a modest effect on results presented here.

We briefly note a discrepancy between the planet counts reported by BKep11 and HKep11. HKep11 report a sample of 438 planet candidates (of the 1235 released by BKep11) that (a) have hosts that satisfy the stellar parameter cuts of Equation (2), (b) have planet parameters \( P \leq 50 \) days and \( 2 \; R_\oplus \leq R \leq 32 \; R_\oplus \), and (c) satisfy the HKep11 definition of S/N > 10. If we apply cuts (a) and (b) only to the BKep11 tables, but allow all S/N, there are 378 candidates. That number increases to 393 using the reported \( R \) instead of calculating it as \( (R/R_*) \times R_* \). Thus, there are at least 60 (or 45) planets that HKep11 say should appear in the BKep11 tables, but they do not.

Applying any S/N threshold to the BKep11 tables can only make the discrepancy larger. We made no choices that would cause us to miss planets in the BKep11 tables. Cuts were applied inclusively (including any values that fall on a boundary) and all the planets in any system were counted. We conclude that a discrepancy exists between the planet and/or stellar parameters appearing in the current versions of HKep11 and BKep11. Speculatively, parameters may have been refined over time. This mismatch does not affect our results as long as the estimates of the discovery efficiencies in HKep11 are sufficiently accurate.

3. A GENERAL METHOD FOR EXOPLANET STATISTICS

This section describes a general method for estimating the planetary distribution function (PLDF) from the results of a survey with quantified detection efficiencies. The purpose of developing this formalism is twofold.

First, this method can investigate a PLDF that depends on properties of both the planets and the target stars. Furthermore, the PLDF can take an arbitrary functional form.

Second, this formalism allows the inclusion of all detected planets—including those in multi-planet systems. This inclusion is possible because we treat the PLDF as the average NPPS, not as the fraction of stars with planets (FSWP). Following TT02, our technique treats planet occurrence as a Poisson process, i.e., a series of independent random events. This approach naturally gives the NPPS.
By contrast, many statistical analyses treat planet hosting as a binary proposition, i.e., a star either does or does not have a detected planet. The use of binomial statistics naturally gives the FSWP. Including multiple planets per system in this type of analysis is formally incorrect, though discrepancies are small if the multiplicity rate is low. For more discussion of these points, see Cumming et al. (2008) who rigorously analyze the FSWP in RV surveys by only including the most detectable planet around any star.

For transit surveys, the FSWP is more difficult to determine because of multiplicity issues. Should misaligned, non-transiting planets be assigned evenly among all stars or into high-order multi-planet systems? As mentioned in the Introduction, multiplicity issues are being actively studied, but can be ignored when treating the planet population as a whole.

Before developing the general method for fits to a parameterized PLDF in Section 3.2, we consider non-parametric estimates of the NPPS in Section 3.1. Parametric fits also give an estimate of the planet occurrence, but the quality depends on the appropriateness of the assumed functional form. The main reason to consider parameterized fits is to identify and assess trends in the data—and if feeling bold, to extrapolate. Section 3.3 discusses the consequences of fitting some parameters and ignoring others.

3.1. Non-parametric Estimates of Planet Occurrence

For a survey of \( N_e \) stars, we divide planet detections into bins indexed by \( \ell \), with \( N_\ell \) planet detections in a given bin. For generality the meaning of the bin is here kept arbitrary. If the average detection efficiency per bin is \( \eta_\ell \), then the best estimate of the NPPS in each bin is

\[
N_\ell \approx \frac{N_\ell}{\eta_\ell N_e}, \tag{9}
\]

the number of detected planets per star divided by the efficiency.

The total planet occurrence, \( \sum \ell f_\ell \), depends on bin size when \( \eta_\ell \) is not constant. For arbitrarily small bins, there is only one planet per bin, and the efficiency is evaluated for the parameters of each planet and its host. This small bin limit is the only truly assumption-free way to estimate the planet fraction and suggests that binning of data is never required.

A maximum likelihood analysis using Poisson statistics can reproduce the estimates of \( f_\ell \) given by Equation (9) and develop intuition for the general method. The number of expected detections in all bins is

\[
N_{\text{exp}} = N_e \sum \eta_\ell f_\ell. \tag{10}
\]

The likelihood of the detections is a product of the Poisson distributions for each bin,

\[
\mathcal{L} = \left[ \prod \frac{(N_\ell \eta_\ell f_\ell)^{N_\ell}}{N_\ell!} \right] \exp(-N_{\text{exp}}). \tag{11}
\]

We are interested in the values of \( f_\ell \) that maximize the likelihood (and its logarithm). As a consequence, we can ignore any constant multiplicative factors, where constant means independent of the \( f_\ell \) values. This simplification results in a likelihood function,

\[
L = \left[ \prod f_\ell^{N_\ell} \right] \exp(-N_{\text{exp}}). \tag{12}
\]

The efficiencies now appear only in \( N_{\text{exp}} \), a feature that has greater significance for the parameterized fits discussed below.

The maximum likelihood values of \( f_\ell \) are the roots of \( \partial \ln(L)/\partial f_\ell = 0 \), which reproduce the expected result of Equation (9).

3.2. Fitting the Planetary Distribution Function

We now describe a general method to fit a parameterized PLDF to unbinned data. We consider a PLDF that depends on planet properties, \( x \) (an \( N_x \) component vector) and stellar properties \( z \) (with \( N_z \) components). Usually, the elements of these vectors are logarithms of measured quantities such as the orbital period or stellar mass.

The probability that a star with properties \( z \) has a planet with properties \( x \) that lie in a volume \( dx \) of phase space is

\[
df = \frac{\partial f(x, z)}{\partial x^{N_x}} dx. \tag{13}
\]

Defined this way, the integral \( \int df \) represents the NPPS with stellar properties \( z \) over the range of planet properties, \( x \), covered by the integral.

We express the differential distribution,

\[
\frac{\partial f(x, z)}{\partial x^{N_x}} = C g(x, z; \alpha), \tag{14}
\]

as an amplitude, \( C \), times a shape function, \( g \). A functional form for \( g \) must be assumed, and this function can include any number of shape parameters, \( \alpha \). The main result of the maximum likelihood analysis is the best-fit values of \( \alpha \). As a simple example, the values of \( \alpha \) can be power-law exponents.

The total number of planets around \( N_e \) stars (indexed by \( j \)) is

\[
N_{\text{tot}} = C \sum_j \int g(x, z_j; \alpha) dx \tag{15a}
\]

\[
= N_e C \int F_\alpha(z) g(x, z; \alpha) dx dz, \tag{15b}
\]

where this expectation value applies over the range of \( x \) covered by the integral.\(^3\) In Equation (15b), the sum over individual stars is replaced by an integral over the stellar distribution function \( F_\alpha(z) \), which is normalized to one as \( \int F_\alpha(z) dz = 1 \). This replacement approximates individual stellar properties as obeying a continuous distribution. Such an approximation is not required. Indeed, direct summation over all stars is always possible (and preferred when feasible). We proceed with the integral notation to minimize indices.

Consider a survey with a net detection efficiency \( \eta(x, z) \). As described for the case of transit surveys in Section 2, the net efficiency is the product of (a) any number of uncorrelated detection probabilities (all \( \lesssim 1 \)) that arise from individual selection effects and (b) a false-positive efficiency (\( \gtrsim 1 \)) that gives the ratio of all detections (planets and false positives) to actual planets.

\(^3\) For brevity the exponent in the phase space volume is dropped, i.e., \( dx = dx^{N_x} \), here and throughout. We keep this exponent in the denominator of derivatives to be explicit that these are not gradients, as in Equation (13).

\(^4\) The NPPS, \( N_{\text{tot}}/N_e \), would be given by \( C \) if \( g \) was normalized so that the integral in Equation (15b) was one. However, this normalization is unnecessary and would require more care in choosing \( g \), so we do not impose this requirement.
The expected number of planet detections (including false positives when relevant) is then

\[ N_{\text{exp}} = N_s CF, \]  

(16)

where the shape integral,

\[ F \equiv \int \eta(x, z)F_\ast(z)g(x, z; \alpha)dxdz, \]  

(17)

weights the shape function over both the stellar distribution and the efficiency.

### 3.2.1. Maximizing Likelihood

Consider a survey that detects \( N_{\text{pl}} \) planets (some of which could be false positives) around \( N_s \) stars. We define the likelihood of the data as

\[ \mathcal{L} = \prod_{i=1}^{N_{\text{pl}}} \exp(-N_{\text{exp}}). \]  

(18)

Analogous to Equation (12), this likelihood represents the probability of each individual detection, labeled by \( i \), times the exponential probability that no other planets were detected.

Applying Equations (13) and (14), we eliminate unnecessary constants, including the phase space volume. The likelihood function becomes

\[ L = \left[ C^{N_{\text{pl}}} \prod_{i=1}^{N_{\text{pl}}} g(x_i, z_{j(i)}; \alpha) \right] \exp(-N_{\text{exp}}), \]  

(19)

where \( j(i) \) refers to the star that hosts planet detection \( i \). Letting \( g_i \) represent the shape function evaluated at detection \( i \), the log likelihood is

\[ \ln(L) = N_{\text{pl}} \ln(C) + \sum_{i=1}^{N_{\text{pl}}} \ln(g_i) - N_{\text{exp}}. \]  

(20)

The best-fit normalization, \( C \), is found by maximizing the likelihood as the root of \( \partial \ln L / \partial C = 0 \):

\[ C = \frac{N_{\text{pl}}}{N_s F}. \]  

(21)

using Equation (16).

Following TT02, we use this constraint to eliminate \( C \) from the likelihood:

\[ \ln(L) = -N_{\text{pl}} \ln(F) + \sum_{i=1}^{N_{\text{pl}}} \ln(g_i) + \left\{ N_{\text{pl}} \left[ \ln \left( \frac{N_{\text{pl}}}{N_s} \right) - 1 \right] \right\}, \]  

(22)

where the constant term in curly brackets can again be ignored.

To find the best-fit values of \( \alpha \), follow from the roots of the \( \partial \ln(L) / \partial \alpha = 0 \) as

\[ \frac{\partial \ln F}{\partial \alpha} = \frac{1}{N_{\text{pl}}} \sum_{i=1}^{N_{\text{pl}}} \frac{\partial \ln(g_i)}{\partial \alpha}. \]  

(23)

These \( N_\alpha \) (the number of parameters in \( \alpha \)) equations generally couple to each other and require numerical solution, but see the Appendix. In Equation (23), the detection efficiencies affect the left-hand side (via \( F \)), while detections determine the right-hand side.

Next, we summarize the basic steps in obtaining a best-fit solution. First, make a hypothesis for the functional form of the PLDF and express it in the form of Equation (14). Second, solve Equation (23) for the shape parameters. Third, using the best-fit \( \alpha \), calculate the normalization via Equations (21) and (17). The best-fit solution is now complete and can be used (e.g.) to estimate the NPPS as \( N_{\text{tot}}/N_s \) using Equation (15). Fourth, estimate the errors on the fit from the contours of the likelihood function in Equation (20), as explained in Section 3.2.3 below.

### 3.2.2. Interpretation

This formalism has a straightforward interpretation. The normalization \( C \) matches the numbers of detected and expected planets, \( N_{\text{pl}} = N_{\text{exp}} \), as seen by comparing Equations (16) and (21).

The shape parameters similarly match the expected and detected averages of planet and host star observables. Consider the case of power-law distributions with a shape function

\[ g = \exp(\alpha_{\text{pl}} \cdot x + \alpha_s \cdot z). \]  

(24)

Here \( x \) and \( z \) are the logarithms of quantities being fit to a power law, and \( \alpha = [\alpha_{\text{pl}}, \alpha_s] \) distinguishes the power-law exponents for planetary and stellar parameters. With this power-law shape function, Equation (23) gives

\[ \langle x \rangle_F = \langle x \rangle_{\text{obs}}, \quad \langle z \rangle_F = \langle z \rangle_{\text{obs}}, \]  

(25)

where

\[ \langle x \rangle_F \equiv \frac{1}{F} \int x \eta F_\ast g dx dz \]  

(26a)

\[ \langle x \rangle_{\text{obs}} \equiv \frac{1}{N_{\text{pl}}} \sum_i x_i, \]  

(26b)

and similarly for the stellar parameters. Non-power-law shape functions correspond to a more complex averaging of the observables.

We now comment on the validity of treating planet occurrence as a Poisson process, i.e., as random, uncorrelated events. This approach may sound inconsistent with interactions in multi-planet systems, but it is not. In effect, the method investigates whether a random drawing from a PLDF (weighted by selection biases) provides a good fit to the observations. The possibility that planet interactions give rise to a complicated PLDF can be included and does not affect the validity of this procedure. Furthermore, the vast majority of planet pairs (and possible planet pairs as well) involve different stars which should be completely non-interacting. Thus, the ability of planet interactions in individual systems to affect the PLDF of all planets in all systems is quite weak.

This discussion emphasizes one of the main features of the maximum likelihood approach. For better and worse, the functional form of the PLDF must be prescribed. The appropriateness of the assumed family of functions must be evaluated by some external method of hypothesis testing.

### 3.2.3. Errors

To quantify the uncertainty in the shape parameters, we compare contours of the likelihood \( L \) to a Gaussian distribution.
The maximum likelihood, \( L_{\text{max}} \), is given by Equation (22) evaluated at best-fit \( \alpha \). The \( \sigma \) error ellipse covers \( \alpha \) values that satisfy \( \ln(L) \approx \ln(L_{\text{max}}) - 1/2 \). The general \( \sigma \) contours follow \( \ln(\ln(L)) = \ln(L_{\text{max}}) - n^2/2 \). The uncertainty in the normalization \( C \) is only the range of values taken by \( C \) within the error ellipse of the \( \alpha \).

For one-dimensional errors on an individual shape parameter, we hold other parameters fixed at their best-fit values. With strong covariance, and skewed error ellipses, such one-dimensional errors would not be very meaningful. As we will show (in Figure 6), such covariance is mild in our analysis.

We emphasize that a maximum likelihood analysis does not evaluate the quality of a fit. The errors on \( \alpha \) measure the fit’s precision. For instance, a large planet sample will precisely define an average power law even if the actual PLDF deviates strongly from a power law.

Our analysis ignores uncertainties in the planetary and stellar parameters. Here, we describe a simple way to include these errors, mainly to show that this is possible. A more rigorous approach would use Fisher information matrices, but is not developed here.

Uncertainties associated with an individual detection, \( i \), affect the value of the shape function, \( g_i \), that appears in the likelihood analysis. To include these uncertainties, \( g_i \) should be weighted integral over the uncertainties,

\[
g_i = \int \phi_i(x' - x, z' - z_j) g(x', z'; \alpha) dx'dz',
\]

where \( \phi_i \) is an appropriately normalized Gaussian distribution (for instance) centered on the best-fit values of the planet detection and its host star. Ignoring uncertainties is equivalent to setting the error distributions, \( \phi_i \), to \( \delta \)-functions.

Uncertainties in the target star properties affect the number of expected planets via the shape integral \( F \). Errors can be included similarly to Equation (27) as

\[
F = \frac{1}{N_*} \sum_{j} \int \phi_j(z' - z) \eta(x, z') g(x, z'; \alpha) dx'dz' \quad \text{(28a)}
\]

\[
\approx \int \phi(z' - z) C \eta(x, z) g(x, z') dxdzdz'. \quad \text{(28b)}
\]

The two forms above show how errors could be applied to each star \( j \) or (more approximately) how average errors could be assigned to the stellar distribution function. The integration over both \( z \) and \( z' \) in Equation (28b) covers both the distribution of average stellar properties \( z \) (replacing the sum over individual stars in Equation (28a)) and the range of uncertainties \( z' \) allowed by the error distributions \( \phi \). Estimates of the total number of planets \( N_\text{tot} \) of Equation (15) are modified by a similar inclusion of the error distribution \( \phi \).

Systematic uncertainties in the detection efficiencies are more difficult to quantify. Assigning detection efficiencies is the most dangerous step in the statistical analysis of survey data. It is safest to perform multiple analyses that cover a range of (hopefully reasonable and nearly complete) possibilities for detection efficiencies. That approach is the one taken here.

3.3. Which Parameters to Study

Any analysis must choose to study some subset of planetary and stellar parameters and to ignore others. The freedom to safely ignore a parameter in the PLDF only exists if detection efficiencies remain constant with respect to the ignored parameter. A detection efficiency that stays near 100% obviously qualifies as constant. As always, these conditions apply only to the region of parameter space being studied.

To understand the inherent danger in ignoring parameters, note that the PLDF is assumed to be averaged over all ignored parameters. However, an accurate prediction of the yield of a survey requires weighting the PLDF by the detection efficiencies before integrating over parameter space, as in Equation (16). Thus, ignoring a parameter in the PLDF makes it impossible to correct for selection biases related to that parameter.

A safer—though admittedly more difficult—way to ignore a parameter (that affects the detection efficiency) is to first include that parameter in the fit and then marginalize over it. With this caveat made explicit, we now show how the general formalism treats the case of a PLDF that depends only on planetary parameters, \( x \), or only on stellar parameters, \( z \), (most famously stellar metallicity).

3.3.1. Planetary Parameters Only

For a shape function, \( g(x; \alpha) \), that is independent of stellar properties, the shape integral simplifies to

\[
F = \int \eta(x) g(x; \alpha) dx,
\]

where the stellar-averaged efficiencies are

\[
\eta_\ast(x) = \int \eta(x, z) F_\ast(z) dz.
\]

The fit procedure is simplified in this case. The stellar averaging is done a single time and not for every choice of \( \alpha \) during optimization.

This case is the one most relevant to this work. Specifically in Section 4 we consider a shape function

\[
g = \exp(\alpha x + \beta y) = \left( \frac{R}{R_o} \right)^\alpha \left( \frac{P}{P_o} \right)^\beta.
\]

This power-law distribution in planetary radius and orbital period identifies the planetary parameters

\[
x = \{x, y\} = \{\ln(R/R_o), \ln(P/P_o)\}
\]

as logarithms in radius and period, relative to reference values \( R_o \) and \( P_o \), respectively. The shape parameters \( \alpha = \{\alpha, \beta\} \) are power laws. The Appendix gives an analytic solution for these best-fit power-law exponents for an ideal transit survey with unity discovery efficiency and no false positives.

3.3.2. Stellar Parameters Only

We now consider fitting a PLDF that depends on stellar parameters only. As cautioned above, this simplification would be a bad idea for transit surveys since detection efficiencies always vary with orbital period (due to the geometric transit probability) and often vary with planetary radius (for smaller planets that give shallow transits). We consider this case for other survey types and for completeness.

The primary conceptual difficulty is that the PLDF was defined in Equation (14) as a differential distribution in one or more planet parameters, \( x \). However, we can remove all references to \( x \) by integrating over the observed volume \( V = \int dx \) of parameter space. The non-differential PLDF,

\[
f(z) = C G(z; \alpha_\ast),
\]

\[
\ln(L) = \ln(L_{\text{max}}) - n^2/2.
\]
gives the NPPS with characteristics \( z \). The integration of the shape function gives \( G \equiv gV \).

The expected planet yield is still \( N_{\exp} = N_s CF \), but the shape integral simplifies to

\[
F = \int \eta(z) F_s(z) G(z, \alpha_s) dz.
\]

The best-fit solutions are still given by Equations (21) and (23), with \( \ln g_r \) replaced by \( \ln G_r \).

This derivation demonstrates how the method could be applied to study the abundance of planets versus stellar metallicity and/or mass alone. We again caution that this approach assumes that the detection efficiencies do not vary with the properties of the detected planets.

4. POWER-LAW FITS TO KEPLER DATA

We now use the planet candidates detected by Kepler (BKep11) to constrain the underlying PLDF as a power law,

\[
\frac{\partial f}{\partial \ln R \partial \ln P} = C \left( \frac{R}{R_s} \right)^\alpha \left( \frac{P}{P_s} \right)^\beta,
\]

in planet radius \( R \), and orbital period \( P \). Equation (35) represents the NPPS per logarithmic interval in \( R \) and \( P \).

The actual PLDF can and does deviate from simple power-law behavior. Nevertheless, power-law distributions are very useful in identifying basic trends in the data. Deviations from power-law behavior help reveal the relevant scales of planet formation and migration, and we show that Kepler data identify these trends quite well.

We study the full planet sample around solar-type stars in Section 4.1. In Section 4.2, we compare the distributions of shorter and longer period planets. Further comparing results for larger and smaller planets in Section 4.3 shows that the sample has a deficit of intermediate-sized planets at the shortest orbital periods. For all these correlations, the \( S/N \) threshold has limited impact (Section 4.4).

4.1. The Full Sample

Our most complete sample of planets covers \( 0.5 < P/\text{day} < 50 \) and \( 0.5 < R/R_s < 20 \) (see Section 2.3 for details). Figure 3 plots planet counts of this full sample. The planet counts are binned by \( R \) (or \( P \)) in the left (right) panel. Raw counts are divided by the number of stars surveyed and the logarithmic bin size.\(^5\) Normalized this way, the data measure the PLDF as weighted by the detection efficiencies, \( \eta \),

\[
\left\{ \eta \frac{df}{d\ln R} \right\} \text{ and } \left\{ \eta \frac{df}{d\ln P} \right\}.
\]

Brackets indicate marginalization over \( R \) and \( P \), respectively. Aside from sampling noise, these values are independent of the survey size or choice of bin width.

The maximum likelihood fits do not use this binned data. Rather, the joint power-law fits use the technique of Section 3.2 (specifically the special case of Section 3.3.1) to analyze the unbinned data. These fits use the detection efficiencies to estimate the true PLDF. For a visual comparison with the planet counts, we project the fits back into observational space in Figure 3. This projection requires weighting the PLDF by the detection efficiencies and marginalizing over \( P \) (or \( R \)) to give the quantities in Equation (36).

We now explain in detail how a power-law PLDF is modified by this projection. Though the best-fit period distribution in the

\(^5\) For example, divided by \( \ln(R_{i+1}/R_i) \) for a bin between \( R_i \) and \( R_{i+1} \) and similarly for period bins.
right panel of Figure 3 has $\beta = 1.3$, the curve in observational space has $\beta_{\text{obs}} \approx 0.3$, smaller by about one. A decrease of precisely 2/3 is due to the transit efficiency. The remainder (here $\approx -1/3$) is due to the period dependence of $\eta_{\text{disc}}$. From Equation (4), this non-geometric correction could be as large as $\approx -1$, if all detections were $S/N > 10$. Thus, $\beta_{\text{obs}}$ could be lower than the underlying $\beta$ by as much as $-5/3$. It is often assumed that the geometric correction of $-2/3$ adequately relates the underlying and observed power laws in $P$, but this approximation is insufficient.

Selection effects not only shift the slope of a power law, they can also introduce curvature. Since Figure 3 is log–log, any deviation of the power-law fits from a straight line is entirely due to selection effects. Though not easily visible, there is modest curvature to the projected period law in the right panel due to lower discovery efficiencies at longer periods.

The curvature in the fit to the size distribution is much more dramatic, as shown in the left panel of Figure 3. The sharp bend coincides remarkably well with the peak in planet counts below $\sim 3 R_\oplus$. The agreement is so good that we must emphasize that the break is controlled by the best estimates of $\eta_{\text{disc}}$, with no adjustable free parameters. Above $\sim 3 R_\oplus$, the Kepler discovery efficiency is unity for all periods considered. Thus, the slope of the projected size distribution at large radii matches the best-fit $\alpha$. The break starts at $R \lesssim 3 R_\oplus$, because $\eta_{\text{disc}}$ drops below 90% for these sizes, as shown in Figure 2. At small $R$, the power-law slope of the projected size distribution is $\alpha + 6.34$, as fixed by the radial dependence of Equation (4).

Thus, the drop in planet counts toward small radii is very well modeled by selection effects. This agreement is strong evidence that the underlying planet counts continue to rise toward small sizes, down to the $\sim 0.5 R_\oplus$ limit of the observations.

The uncertainty of the fits are shown in Figure 3 by the thick red curve covering 1σ deviations from maximum likelihood. Figure 6 shows the corresponding error ellipse in the $\alpha$ and $\beta$ parameters (the small black oval is for the full planet sample). The 1σ errors on $\alpha$ and $\beta$ reported in Table 1 are one-dimensional errors holding the other parameters fixed. As explained in Section 3, the normalization $C$ is not varied during the fit procedure, but follows from the values of $\alpha$, $\beta$, and the planet-to-star ratio. The errors on $C$ indicate the range of values taken inside the 1σ error ellipse of $\alpha$ and $\beta$.

Our best-fit size distribution has $\alpha = -1.73 \pm 0.07$. This value indicates that smaller planets are much more abundant, but that larger planets likely contain most of the mass. At constant planet density, $\alpha < -3$ would give small planets most of the

| Tag     | Planet Sample | $N_{\text{fit}}$ (corrected) | Planets per Star | $\alpha$ (radius law) | $\beta$ (period law) | $C_{\text{fit}}$, yr |
|---------|---------------|-------------------------------|------------------|-----------------------|----------------------|---------------------|
| Full    | 0.5–20 0.5–50 | 1.4 × 10^5                  | 2.38             | -1.73 ± 0.07          | 1.33 ± 0.03          | 23.40 ± 0.50        |
| Full/R2 | 2–20 0.5–50  | 1.2 × 10^4                  | 0.21             | -1.88 ± 0.11          | 1.22 ± 0.04          | 20.30 ± 1.35        |

Notes.

- $R_\oplus$ is planet radii computed form the products of $R/R_e$ and $R_e$ given in Tables 1 and 2 of BKep11.
- $N_{\text{fit}}$: number of Kepler planet candidates with $S/N > 10$ (or 20 where indicated) as reported in BKep11.
- See Section 5 for more accurate non-parametric fits to the planet occurrence.

### Table 1

Joint Power-law Fits to Kepler Planet Candidates in “Solar” Subsample

| Tag     | Radius ($R_\oplus$) | Periods (days) | $N_{\text{fit}}$ | Planets per Star | $\alpha$ (radius law) | $\beta$ (period law) | $C_{\text{fit}}$, yr |
|---------|---------------------|---------------|-----------------|------------------|-----------------------|----------------------|---------------------|
| Full    | 0.5–20 0.5–50       |               | 1.4 × 10^5      | 2.38             | -1.73 ± 0.07          | 1.33 ± 0.03          | 23.40 ± 0.50        |
| Full/R2 | 2–20 0.5–50         |               | 1.2 × 10^4      | 0.21             | -1.88 ± 0.11          | 1.22 ± 0.04          | 20.30 ± 1.35        |

(Short vs. Long Period: “Fast” vs. “Slow”)
mass. However, larger planets tend to have lower density, \( \rho \). Take \( \rho \propto R^{-1} \) as an example.\(^6\) Then \( \alpha \approx -2 \) would give small planets more of the mass (over the range where the density law holds). Though the mass–radius relation remains uncertain, we tentatively conclude that larger planets dominate the mass distribution of \( \text{Kepler} \) planets.

The best-fit period law, \( \beta = 1.33 \pm 0.03 \), indicates that planets are more closely packed further from the star, for logarithmic intervals in \( P \) or semimajor axis \( a \). For \( \beta > 3/2 \), the planet density per linear interval in \( a \) would increase.

However, Figure 3 shows that a single power law is a qualitatively poor fit to the data due to a sharp drop in planet counts below \( P \approx 3 \) days. After a broad maximum, planet counts drop slightly for periods above \( \sim 10 \) days, but (as explained above) the drop is so shallow that the underlying period distribution continues to rise. The obvious break in the period distribution motivates our division of the planet sample below.

### 4.2. Short versus Long Periods

Since the period distribution deviates strongly from a simple power law, we divide the planets into short and long period samples. We use \( P = 7 \) days as the dividing line. The choice is only partly motivated by the proximity to the break in the period distribution, which is gradual and more prominent at shorter periods. We also want each sample to have significant numbers of planets and an adequate range of periods over which to measure a power-law slope. Planet counts for different data cuts are summarized in the \( N_{pl} \) column of Table 1.

Figure 4 shows independent power-law fits to the short and long period data, again projected against the planet counts. Two power laws vastly improve the qualitative fit to the period distribution (right panel). The large exponent at short periods \( \beta_{\text{fast}} \approx 2.2 \) reflects the sharp drop in planet counts to short periods (despite the flattening of the planet counts for \( P > 3 \) days). Though planet counts decline to longer periods, the long period exponent \( \beta_{\text{slow}} \approx 0.6 \) shows that the underlying distribution continues to rise. Longer period planets are needed to determine where the distribution turns over. It is quite exciting that we have not yet found where most of the planets lie.

The size distributions in the left panel shows overlapping histograms for the short and long period samples. The longer period (“slow”) sample has a steeper radius distribution than the shorter period (“fast”) with \( \alpha_{\text{slow}} - \alpha_{\text{fast}} \approx 0.5 \pm 0.2 \). By eye, the difference in the distribution appears driven by the flatness of the fast sample between \( 3 \) and \( 15 \, R_{\oplus} \). To clarify this behavior, we now examine the differences between small and large radius planet samples.

\( \begin{align*}
\alpha_{\text{slow}} &= -1.93 \\
\alpha_{\text{fast}} &= -1.44
\end{align*} \)

\( \beta_{\text{fast}} = 2.23 \quad \beta_{\text{slow}} = 0.63 \)

\( \begin{align*}
\alpha_{\text{big, slow}} &= -1.31 \pm 0.47, \text{ with almost } 3\sigma \text{ significance. Since the discovery efficiency of } \text{Kepler} \text{ is near unity for these}
\end{align*} \)

\( \text{similar to the } \rho \propto R^{-0.94} \text{ suggested by Lissauer et al. (2011b).} \)

\( \text{4.3. Quadrants} \)

We subdivide the planets into four samples by splitting both the shorter and longer period populations (still defined relative to 7 days) into “small” and “big” samples relative to \( 3 \, R_{\oplus} \). The results of the four independent power-law fits are plotted against planet counts in Figure 5. Figure 6 plots error ellipses to compare the slope changes and their statistical significance. The error ellipses of the subsamples are much larger than the full sample due to lower numbers and the reduced range in \( R \) and \( P \) (over which to define a slope).

The period distributions are only modestly affected by this split. At both short and long periods, the best-fit \( \beta \) of the small and big samples agree within the statistical uncertainties. Nevertheless, some qualitatively interesting features appear in the period distributions shown in the right panel of Figure 5. The small planet counts begin their decline below about 7 days (as noted by HKep11). The large planet sample remains flat toward shorter periods—with evidence of a peak at \( P \approx 3 \) days. This peak is responsible for the flat appearance of the overall period distribution down to 3 days, as seen in the right panels of Figures 3 and 4.

The behavior of the size distributions is even more intriguing. For big planets, the best-fit radius law changes by \( \alpha_{\text{big, slow}} - \alpha_{\text{big, fast}} = -1.31 \pm 0.47 \), with almost \( 3\sigma \) significance.
planets the behavior is precisely opposite. The size distribution is quite shallow at large sizes and steepens for the small radius sample.

The implications of this behavior are intriguing. The processes that form and/or migrate planets inside \( \sim 7 \) days clearly disfavor planets of a few to several \( R_\oplus \). Atmospheric evaporation can at least qualitatively explain these trends. At short periods, it seems that only the most massive giants—true Jovians with \( R \gtrsim 11 R_\oplus \)—can efficiently retain their atmospheres. Lesser giants are more readily stripped of their atmospheres, including ices that sublimate from the surface. The demotion of ice and gas giants to small rocky cores can explain the flattening of the large planet size distribution at short periods.

While the data are strongly suggestive, the evaporation hypothesis does have complications. First, even at periods below 7 days, lower gravity ice giants might retain their atmospheres (Ehrenreich & Désert 2011). Also, evaporation can only partly explain the sharp rise in small planets at short periods. Evaporation could provide some small cores, but not nearly enough. An additional mechanism for the preferential parking of smaller versus larger planets appears to be required. Nevertheless, it appears likely that atmospheric retention plays a key role in the observed size distributions, providing another pillar of support for the core-accretion hypothesis.

### 4.4. Systematic Uncertainty

To investigate how systematic uncertainties in the discovery efficiency might affect our results, we experimented with different S/N thresholds. As discussed in Section 2.3, S/N thresholds of 10 and 20 should provide a reasonable bound on the uncertainties.

The results so far have described the more inclusive S/N > 10 sample. Figure 3 includes dotted histograms that correspond to the higher S/N > 20 sample. Planet counts are noticeably lower below 3 \( R_\oplus \) and across all periods.

Table 1 includes fits with both S/N thresholds. All of the qualitative trends discussed in this section persist in the higher S/N sample, with only mostly reduced significance. In particular, the shifts to the size distributions of even the small planet

---

**Figure 5.** Planet sample is divided into four quadrants with separations in period at 7 days and in radius at 3 \( R_\oplus \). Each quadrant is separately fit to a power-law PLDF (\( \propto R^\alpha P^\beta \)). The result is plotted in observational space, and the values of the best-fit exponents are shown. The differences between the subsample distributions and their significance are discussed in the text and shown in Figures 6 and 7.

(A color version of this figure is available in the online journal.)

**Figure 6.** Error ellipses (1\(\sigma \), 2\(\sigma \), and 3\(\sigma \)) for the power-law fits to the PLDFs (\( \propto R^\alpha P^\beta \)) of different planet samples. The black contours show the fit to the full planet sample. Fits to the four quadrants of the planet sample shown in Figure 5 are also given. The difference in the period distribution of short and long period planets is highly significant. The behavior of the size distributions is more complex and has \( \sim 2\sigma \)–3\(\sigma \) significance. For shorter periods, the size distribution steepens (more negative \( \alpha \)) between the large and small planet samples. Longer periods show the opposite trend—the size distribution flattens, going from bigger to smaller planets.

(A color version of this figure is available in the online journal.)

---

large planets, no obvious systematic uncertainties should be at work. For small planets, the change in the size distributions, \( \alpha_{\text{small,slow}} - \alpha_{\text{small,fast}} = 0.70 \pm 0.47 \). While of more modest significance, this change is of the opposite sign.

Figure 7 plots the best-fit size distributions for each quadrant, with the full planet sample shown for comparison. At longer periods the size distribution flattens toward smaller radii, i.e., \( \alpha \) is less negative for the small planets. For the shortest period
The Astrophysical Journal, 742:38 (13pp), 2011 November 20
Youdin

Figure 7. Size distributions, marginalized over period. Fits are to a power-law PLDF, for both our full planet sample and the four quadrants shown in Figures 5 and 6. Planets around \(\sim 3 R_{\oplus}\) appear preferentially depleted at orbital periods below \(\sim 7\) days. This deficit likely arise due to the loss of volatiles from (and/or inability to attract gas onto) lower mass giant planets that approach their host stars too closely.

(A color version of this figure is available in the online journal.)

The probable existence of more than one planet per solar-type star, especially within 50 days, is remarkable. However, it does not imply that the Sun is exceptional for lacking a planet so close. For \(R > 1.5 R_{\oplus}\), there are \(\geq 2–3\) planets per planetary system (Lissauer et al. 2011b). Multiplicity rates will increase further when planets down to 0.5 \(R_{\oplus}\) are considered.

It is inherently speculative to extrapolate into unobserved regions of parameter space. Nevertheless, all the power-law fits of the previous section have distributions that rise with period. To emphasize the implications of this rise, we calculate the extrapolated density of Earth-sized planets at 1 AU. The final column of Table 1 expresses the normalization to the fits as

\[
C_{\oplus, yr} \equiv \frac{d^2 f}{d \ln R d \ln P} (R_{\oplus}, \text{year}),
\]

(37)

the NPPS in a logarithmic interval centered on an Earth radius and a year.

The most relevant fit is to the small planet, long period sample, which gives \(C_{\oplus, yr} = 2.75 \pm 0.33\). Integrating this power law gives on average \(f_{1.365} \approx 3\) Earth-like planets with periods below a year. If the distribution function of planets does not turn over shortly beyond periods of 50 days, the implications for habitable Earths around Sun-like stars is truly staggering.

6. CONCLUSIONS

We develop a general technique for the analysis of exoplanet survey data. The merits of the approach, a generalization of the TT02 maximum likelihood analysis, include the following.

1. Data are analyzed without binning in order to preserve the statistical significance of each detection.
2. Describing the NPPS allows multi-planet systems to be included.
3. Selection effects are used to calculate the number of expected detections. They are not used to enhance (or diminish) the number of actual detections, which affect error estimates.
4. The overall normalization is analytically removed from the likelihood function. With one less parameter in the numerical optimization, the fit is simpler and possibly more robust.
5. Different surveys can be jointly analyzed, as shown by TT02. Even different survey methods that probe different regions of parameter space can be combined if selection effects are well characterized.

We apply the technique to Kepler planet candidates released by BKepp11. We focus on planets orbiting solar-type stars for which HKepp11 has quantified the Kepler discovery efficiency. Fits to these efficiencies (see Figure 2) reveal a remarkably smooth variation with planet radius and orbital period that agrees with analytic predictions for transit surveys. This regular behavior gives us confidence to analyze detections down to 0.5 \(R_{\oplus}\). Nevertheless, results for the smallest planets are tentative and certain to be improved on future data releases.

From this analysis, the best estimate of the NPPS larger than 0.5 \(R_{\oplus}\) and with periods below 50 days is \(f \approx 0.7–1.4\). Systematic uncertainty in the discovery efficiencies dominate these estimates. Since planet occurrence rises toward both small sizes and longer periods, the more relevant question is when (not if) we can claim with certainty that there is more than one planet per star.

The shape of the radius and period distributions is even more informative than absolute number counts. The most notable trend is a difference between the size distributions of shorter and longer period planets (below and above 7 days), plotted in Figure 7. While the shorter period planets are less abundant overall, their relative deficit is most pronounced around \(\sim 3 R_{\oplus}\). This trend could arise from the preferential evaporation of ice and lower mass gas giants that migrated too close to their host star. Alternately, the in situ formation of short period planets (Hansen & Murray 2011) might explain these trends based on the difficulty of atmospheric capture.

This discussion assumes that most Kepler planets are likely formed by the core accretion mechanism (Pollack et al. 1996), a hypothesis that is not seriously challenged at masses below \(M_{\text{Jup}}\) and possibly much higher (Kratter et al. 2010). By probing...
the transition from terrestrial to giant planets, *Kepler* offers unprecedented opportunities to improve formation models.

Scott Kenyon’s advice and encouragement made this paper possible. I thank Kevin Schlaufman for making Tables 1 and 2 of Borucki et al. (2011) available in ascii format at http://www.ucolick.org/~kcs/research/planets/kepler_table.dat. I am very grateful to Elisabeth Adams, Jean-Michel Desert, Andrew Howard, Matt Holman, Jonathan Irwin, and Darin Ragozzine for insightful discussions and to all responsible for maintaining the venue of these conversations: the daily SSP coffee at the CfA. This project was supported by the NASA Astrophysics Theory Program and Origins of Solar Systems Program through grant NNX10AF35G and by endowment funds of the Smithsonian Institute.

**APPENDIX**

**ANALYTIC SOLUTIONS FOR POWER-LAW FITS**

This appendix derives an analytic solution to the best-fit power laws of the planetary radius and orbital period. The transit probability is included, but all other detection efficiencies are set to unity, as appropriate for a perfect transit survey. Since numerical fits are more flexible, this derivation is intended to provide insight and to provide a check on numerical solutions (when only the transit probability is included as a selection effect).

Consider the detection of $N_{\text{pl}}$ planets in a transit survey. We fit the PLDF to the power law of Equation (31) (equivalent to Equation (35) used in our main analysis of *Kepler* data). The logarithms of radius and period are defined as the parameters $x$ and $y$ as in Equation (32). We define the reference planet radius, $R_{\text{r}}$, and orbital period $P_{\text{o}}$ so that the (here rectangular) domain of the survey is centered on $x = y = 0$ and all detections (indexed by $i$) fall within

$$-x_m \leq x_i \leq x_m; -y_m \leq y_i \leq y_m. \quad (A1)$$

The transit efficiency of Equation (1) can be written as

$$\eta_{tr,i} = \eta_{tr,o} \exp\left(-\frac{2y_i}{3}\right); \quad \eta_{tr,o} \equiv \left(\frac{3\pi}{G\rho_{\ast}R_{\ast}^2}\right)^{1/3}, \quad (A2)$$

and we ignore other selection effects.

The number of expected transits for such a perfect survey is $N_{\text{exp}} = N_{\ast}C_{\text{perf}}$ with

$$C_{\text{perf}} = 4\eta_{tr,o} \left\{ \frac{\sinh(\alpha x_m) \sinh(\beta - 2/3)y_m)}{\alpha} \right\}, \quad (A3)$$
given by Equation (29) (which averages the general Equation (16) over stellar parameters).

The best-fitting $\alpha$ and $\beta$ follow from Equation (23) as

$$\coth(\alpha x_m) - \frac{1}{\alpha x_m} = \frac{\langle x \rangle_{\text{obs}}}{x_m}; \quad \coth(\beta - 2/3)y_m - \frac{1}{\beta - 2/3} = \frac{\langle y \rangle_{\text{obs}}}{y_m}, \quad (A4)$$

where the observational mean $\langle x \rangle_{\text{obs}} = N_{\ast}^{-1} \sum x_i$ and similarly for $\langle y \rangle_{\text{obs}}$. The relevant transcendental function, $T(y) = \coth(y) - 1/y$, is monotonic in $y$ with $T \rightarrow \pm 1$ as $y \rightarrow \pm \infty$ and $T(y) \approx y/3$ for $|y| \ll 1$. The closed form solutions of Equation (A4) have several features that relate to more complete analyses.

1. The power laws increase monotonically with the average value of the (log of the) observables, independent of how the values are distributed around that mean. Higher-order shape functions can describe more detailed features in the data.

2. The period power-law $\beta$ in the PLDF is $2/3$ larger than the period law describing the detections. The transit efficiency, $\eta_{tr}$, causes this increase. In general, any power-law selection effect adjusts from the observed to the underlying distribution this way.

3. The overall magnitude of the detection efficiency does not affect power laws or shape parameters in general. For instance, uncertainty in $\eta_{tr,o}$ due to uncertainty in stellar densities does not affect the shape function. The overall planet occurrence of course does depend on the magnitude of detection efficiencies via the normalization given by Equation (21).

4. Solutions for the best-fit power laws decouple. In general, the solution for two-shape parameters decouple if the shape parameters (and the physical parameters they describe) are separable in the shape function $g$ and the relevant physical parameters are separable in the detection efficiencies. This separability condition is not met for our full treatment of the *Kepler* data. While the efficiency ratio $r_{\text{disc}}$ of Equation (4) is separable, the efficiency itself, $\eta_{\text{disc}} = r_{\text{disc}}/(1 + r_{\text{disc}})$, is not. This leads to the modest covariance between $\alpha$ and $\beta$ seen in Figure 6.

**REFERENCES**

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Borucki, W. J., Koch, D. G., Basri, G., et al. 2011, arXiv:1102.0541 (BKep11)

Burke, C. J. 2008, ApJ, 679, 1566

Catalanize, J., & Shao, M. 2011, arXiv:1103.1443

Charbonneau, D., Berta, Z. K., Irwin, J., et al. 2009, Nature, 462, 891

Chiang, E., & Youdin, A. N. 2010, Annu. Rev. Earth Planet. Sci., 38, 493

Cumming, A., Butler, R. P., Marcy, G. W., et al. 2008, PASP, 120, 531

Ehrenreich, D., & Desért, J.-M. 2011, A&A, 529, A136

Fischer, D. A., & Valenti, J. 2005, ApJ, 622, 1102

Gaudi, S. 2007, in ASP Conf. Ser. 366, Transiting Exoplanetar planets Workshop, ed. C. Afonso, D. Wolsdrake, & T. Henning (San Francisco, CA: ASP), 273

Goldreich, P., Lithwick, Y., & Sari, R. 2004, ARA&A, 42, 549

Gonzalez, G. 1997, MNRAS, 285, 403

Hansen, B. M. S., & Murray, N. 2011, arXiv:1105.2050

Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2011, arXiv:1103.2541 (HKep11)

Howard, A. W., Marcy, G. W., Johnson, J. A., et al. 2010, Science, 330, 653

Johansen, A., Youdin, A., & Mac Low, M. 2009, ApJ, 704, L75

Johnson, A., Aller, K. M., Howard, A. W., & Crepp, J. R. 2010, PASP, 122, 905

Kenyon, S. J., & Bromley, B. C. 2006, AJ, 131, 1837

Krautter, K. M., Murray-Clay, R. A., & Youdin, A. N. 2010, ApJ, 710, 1375

Latham, D. W., Rowe, J. F., Quinn, S. N., et al. 2011, arXiv:1103.3896

Lissauer, J. J., Fabrycky, D. C., Ford, E. B., et al. 2011a, Nature, 470, 53

Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., et al. 2011b, arXiv:1102.0543

Marcy, G. W., & Butler, R. P. 1996, ApJ, 464, L147

Mayor, M., & Queloz, D. 1995, Nature, 378, 355

Moorhead, A. V., Ford, E. B., Morehead, R. C., et al. 2011, arXiv:1101.5043

Morton, T. D., & Johnson, J. A. 2011, ApJ, 738, 170

Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Icarus, 124, 62

Ragozzine, D., & Holman, M. J. 2010, arXiv:1006.3727

Tabachnik, S., & Tremaine, S. 2002, MNRAS, 335, 151 (TT02)

Tingley, B., Bonomo, A. S., & Deeg, H. J. 2011, ApJ, 726, 112

Tremaine, S., & Dong, S. 2011, arXiv:1106.5403

Youdin, A. N. 2010, in EAS Publications Series, Vol. 41, ed. T. Montmerle, D. Ehrenreich, & A.-M. Lagrange (Les Ulis, France: EDP Sciences), 187

Youdin, A. N., & Goodman, J. 2005, ApJ, 620, 459

Youdin, A. N., & Shu, F. H. 2002, ApJ, 580, 494