Musings on Fitting the P(II) Distribution to Single-Event Timing Data

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie
www.brucehardie.com*

July 2020

1 Introduction

The Pareto distribution of the second kind, hereafter P(II), is a robust distribution for modelling duration times / timing data when the event of interest occurs in continuous time.[1] Its pdf is

$$f(t \mid r, \alpha) = \frac{r}{\alpha}\left(\frac{\alpha}{\alpha + t}\right)^{r+1} \tag{1}$$

and its cdf is

$$F(t \mid r, \alpha) = 1 - \left(\frac{\alpha}{\alpha + t}\right)^{r}. \tag{2}$$

One way of "generating" the P(II) is as a gamma-mixture of exponentials. That is, individual duration times are distributed exponential with rate parameter λ and heterogeneity in λ is captured by a gamma distribution with shape parameter r and scale parameter α.

In this note we explore some issues associated with fitting the P(II) to single-event timing data using maximum likelihood estimation.

2 Starting Point

In our teaching, we typically introduce the P(II) distribution as a model for the trial purchasing of a new FMCG product called Kiwi Bubbles. The

[1] It is also used to model other non-negative (continuous) quantities, such as income.

associated data are given in Table 1, which documents the *cumulative* number of households (from a panel of 1499 households) that have made a trial purchase by the end of each week over the first 24 weeks the product is in a test market.[2] (The event of interest is time at which a household makes their first-ever (i.e., trial) purchase of the new product.)

Week	# Households	Week	# Households
1	8	13	68
2	14	14	72
3	16	15	75
4	32	16	81
5	40	17	90
6	47	18	94
7	50	19	96
8	52	20	96
9	57	21	96
10	60	22	97
11	65	23	97
12	67	24	101

Table 1: Cumulative number of households that have made a trial purchase by the end of Weeks 1–24.

What we have are grouped, interval-censored data. Interval-censored, because we do not know exactly when the trial purchase occurred; we only know the interval (i.e., week) within which it occurred. Grouped, because we do not observe each household's time; we simply know how many households made a trial purchase in any given interval.

Let n_t be the number of households that made a trial purchase in week t. Given the grouped, interval-censored nature of the data, we fit the P(II) distribution to the data using the following log-likelihood function:

$$LL(r, \alpha \,|\, \text{data}) = \sum_{t=1}^{24} n_t \ln \left[F(t \,|\, r, \alpha) - F(t-1 \,|\, r, \alpha) \right] + \left\{ 1499 - \sum_{t=1}^{24} n_t \right\} \ln[S(24 \,|\, r, \alpha)] \,. \tag{3}$$

With reference to the worksheet `Grouped data` in the Excel workbook `musings_on_pareto-ii_parameter_estimation.xlsx`, the maximum value of the log-likelihood function is $LL = -681.373$, which occurs at $\hat{r} = 0.050$ and $\hat{\alpha} = 7.973$.

[2] See Fader et al. (2003) and Hardie et al. (1998) for a discussion of models for forecasting new product trial.

3 Individual-level, Interval-censored Data

Table 1 is actually a summary of the household-level data given in the worksheet `Raw Data (I)`. For each of the 1499 households in the panel, we see whether or not they made a trial purchase in the first 24 weeks and, if so, the week in which they made their trial purchase.

In the workbook `Raw Data (Ia)`, we create two new variables:

t_i = the week of purchase if we observe a trial purchase in the 24-week observation period, and 24 (i.e., the time at which right-censoring occurs) if we do not observe a trial purchase in that time period, and

δ_i = 1 if we observe the trial purchase, 0 otherwise (i.e., it is a right-censored observation).

Given data of this form, we fit the P(II) distribution to the data using the following log-likelihood function:

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln \left[F(t_i \,|\, r, \alpha) - F(t_i - 1 \,|\, r, \alpha) \right] + (1 - \delta_i) \ln[S(24 \,|\, r, \alpha)] \,. \tag{4}$$

We code this up in the worksheet `Interval censored (week){week}`. As would be expected, the results are exactly the same as those obtained in `Grouped data`.

4 What If We Know Actual Event Times?

In many situations, the data are not interval-censored; we do know the time at which the event of interest (in this case, the first purchase) occurred for each household.

Suppose we know that the event occurred at t_i. What is the appropriate form of the likelihood function?

Following Pawitan (2001, pp. 23–24), we recognize that there is limited precision in our measurement of t_i; saying that the event occurred at t_i is really saying that the event occurred in the interval $(t_i - \epsilon/2, t_i + \epsilon/2]$. The associated contribution to the likelihood function is given by

$$F(t_i + \epsilon/2) - F(t_i - \epsilon/2) = \int_{t_i - \epsilon/2}^{t_i + \epsilon/2} f(u) du$$

which, recalling basic concepts of calculus,

$$\approx \epsilon \, f(t_i) \,.$$

For the purpose of finding the values of the model parameters that maximize the value of the likelihood function, ϵ is a constant and can be ignored.

Returning to our dataset on trial purchasing, we actually also know the day on which the trial purchase occurred — see the worksheet `Raw Data (II)`. We create (t_i, δ_i) in the worksheet `Raw Data (IIa)`. Note that a purchase on day 3 of week 5 occurs at $t = 4 + 3/7 = 4.429$ weeks (assuming that week is the underlying unit of time).

If we treat t_i as the actual time at which the trial purchase occurred (i.e., it is not an interval-censored observation), we can replace the $F(t_i \,|\, r, \alpha) - F(t_i - 1 \,|\, r, \alpha)$ term in (4) with $f(t_i \,|\, r, \alpha)$, giving us

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln[f(t_i \,|\, r, \alpha)] + (1 - \delta_i) \ln[S(24 \,|\, r, \alpha)] \,. \qquad (5)$$

This is coded up in the worksheet `"Actual" time {week}` and we find that the maximum value of the log-likelihood function is $LL = -682.441$, which occurs at $\hat{r} = 0.054$ and $\hat{\alpha} = 9.154$.

Comparing these results to those obtained using the interval-censored data, it is not surprising that the parameters are slightly different. What may come as a surprise is the similarity of the values of the log-likelihood functions. Why is this the case?

In order for the two log-likelihoods to be of similar value, it must be the case that $f(t_i) \approx F(\lceil t_i \rceil) - F(\lceil t_i \rceil - 1)$ (where the ceiling function $\lceil x \rceil$ is the smallest integer greater than or equal to x). Would we expect this to be the case?

In Figure 1, we plot $f(t)$ along with the probabilities of a trial purchase occurring in each week — both computed using the parameter estimates from interval-censored dataset — and note that $f(t) \approx F(\lceil t \rceil) - F(\lceil t \rceil - 1)$. This explains the similar log-likelihood function values.

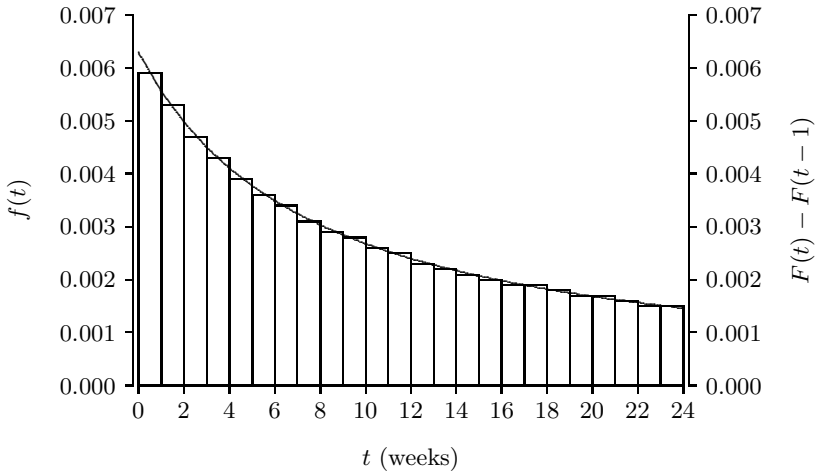

Figure 1: Plot of $f(t)$ $(t \in (0, 24])$ and $F(t) - F(t-1)$ $(t = 1, 2, \ldots, 24)$ for $r = 0.050$ and $\alpha = 7.973$.

5 Is "Actual" Actual?

The reader may be asking whether we can truly treat the day of purchase as the actual time. Shouldn't it be treated as interval-censored data, albeit with a smaller interval of one day (as opposed to one week)? This would see us fitting the model to the data using the following log-likelihood function:

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln \left[F(t_i \,|\, r, \alpha) - F(t_i - 1/7 \,|\, r, \alpha) \right] + (1 - \delta_i) \ln[S(24 \,|\, r, \alpha)] \,. \tag{6}$$

We code this up in the worksheet `Interval censored (day){week}` and find that the maximum value of the log-likelihood function is $LL = -878.497$, which occurs at $\hat{r} = 0.052$ and $\hat{\alpha} = 8.624$. The parameter estimates lie between those obtained using (4) and (5). However, the value of the log-likelihood function is quite different. Why is this the case?

In Figure 2, we plot $f(t)$ along with the probabilities of a trial purchase occurring on each day, both computed using the parameter estimates obtained from (6). (For visual clarity, we just consider the first three weeks.) It is clear that the pdf and the differences in the cdf are no longer nearly equivalent, which is why the maximized values of (5) and (6) are so different. But why do we see this gap?

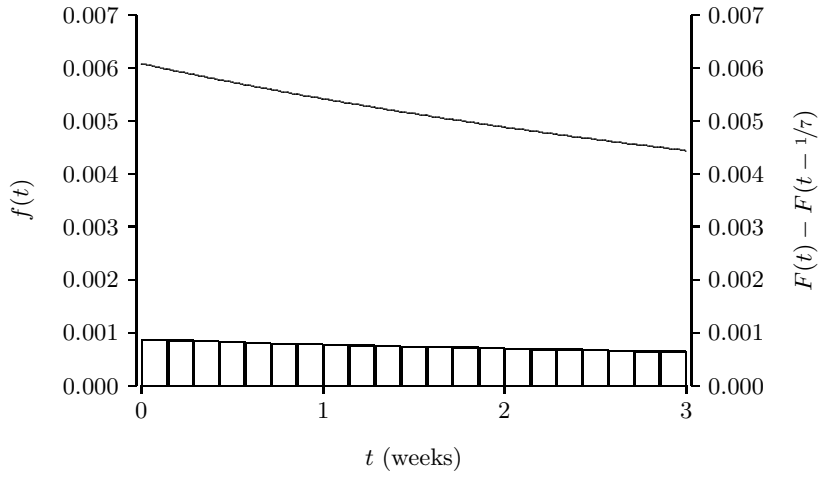

Figure 2: Plot of $f(t)$ $(t \in (0, 3])$ and $F(t) - F(t - 1/7)$ $(t = 1/7, 2/7, \ldots, 2\,6/7, 3)$ for $r = 0.052$ and $\alpha = 8.624$.

Recall that

$$F(t) - F(t - \Delta t) = \int_{t-\Delta t}^{t} f(u)du \approx f(t)\Delta t \,.$$

In Figure 1, $\Delta t = 1$, and we observe that the approximation is reasonably good.[3] However Figure 2 reflects a setting where $\Delta t = 1/7$. If we multiply $f(t)$ by $1/7$ — see Figure 3 — we note that the approximation is very good.

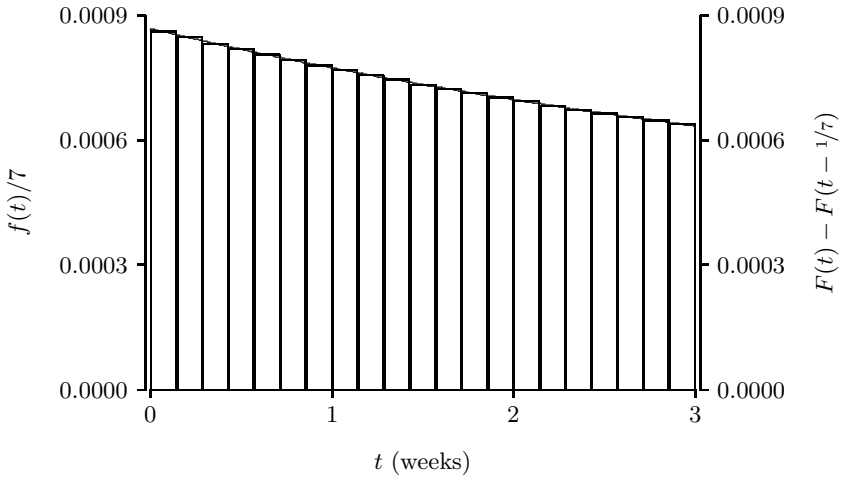

Figure 3: Plot of $f(t)/7$ $(t \in (0,3])$ and $F(t) - F(t - 1/7)$ $(t = 1/7, 2/7, \ldots, 2\,6/7, 3)$ for $r = 0.052$ and $\alpha = 8.624$.

So, the reason why the value of the log-likelihood function given in (5) is not the same as that given in (6) is because $f(t_i) \not\approx F(t_i) - F(t_i - 1/7)$. The $f(t_i)$ need to be rescaled by $1/7$ for the approximation to hold. Noting that there are 101 uncensored observations in the dataset, this implies an adjustment to the log-likelihood function of $101 \times \ln(1/7) = -196.537$. The actual difference is -196.055.

6 Further Explorations

Suppose we change the underlying unit of time from week to day. The three individual-level-data log-likelihood functions above become

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln \left[F(7 \times t_i \,|\, r, \alpha) - F(7 \times t_i - 7 \,|\, r, \alpha) \right] + (1 - \delta_i) \ln[S(24 \times 7 \,|\, r, \alpha)] \,, \tag{7}$$

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln[f(t_i \,|\, r, \alpha)] + (1 - \delta_i) \ln[S(24 \times 7 \,|\, r, \alpha)] \,, \tag{8}$$

[3] Strictly speaking, Figure 1 compares $f(t) \times 1$ with $\displaystyle\int_{\lceil t \rceil - 1}^{\lceil t \rceil} f(u)du = F(\lceil t \rceil) - F(\lceil t \rceil - 1)$.

$$LL(r, \alpha \,|\, \text{data}) = \sum_{i=1}^{1499} \delta_i \ln \left[F(7 \times t_i \,|\, r, \alpha) - F(7 \times t_i - 1 \,|\, r, \alpha) \right] + (1 - \delta_i) \ln[S(24 \times 7 \,|\, r, \alpha)] \,. \quad (9)$$

We code-up these in the worksheets `Interval censored (week){day}`, `"Actual" time {day}`, and `Interval censored (day){day}`, and summarize the results (along with those given above) in Table 2.

	Time unit: Week			Time unit: Day		
	r	α	LL	r	α	LL
Int. cens. (week)	0.050	7.973	−681.373	0.050	55.808	−681.373
"Actual"	0.054	9.154	−682.441	0.054	64.076	−878.978
Int. cens. (day)	0.052	8.624	−878.497	0.052	60.371	−878.497

Table 2: Estimation results for the three different likelihood function specifications for two different time scales.

Sure enough, when we change the unit of time to day, the maximum values of the log-likelihood functions given in (8) and (9) are very close. And, as expected, the difference in the values of the log-likelihood functions associated with "actual" times when we change the underlying unit of time from week to day is $101 \times \ln(1/7) = -196.537$.

Comparing the parameter estimates across the two time scales, we see that the values of r do not change, while the estimates of α change by a factor of 7 as we go from week to day. This should not be surprising since α is called a scale parameter.

7 Conclusion

The key lesson is that, when comparing the value of an interval-censored log-likelihood function with that of one formulated assuming we know the exact times at which the event of interest occurred, the two log-likelihoods will be very similar in their values when the underlying unit of time is the same as the width of the censoring interval. The closeness will depend on how well $f(t_i)$ approximates $F(\lceil t_i \rceil) - F(\lceil t_i \rceil - 1)$.

The final point to consider is at what level of coarseness in the data should we treat time as being interval-censored rather than being a precise measurement of the actual time at which the event of interest occurred? A purely pragmatic answer would say that it all depends on how well $f(t)$ approximates

$$\int_{\lceil t \rceil - 1}^{\lceil t \rceil} f(u) du = F(\lceil t \rceil) - F(\lceil t \rceil - 1) \,,$$

assuming the underlying unit of time is the same as the width of the censoring interval. Another answer would say that it depends on whether you feel it is appropriate given the time scale associated with the problem behind the model-building exercise.[4] Finally, it worth noting that students without a strong statistical background tend to find it much easier to make sense of an interval-censored likelihood function than one that directly includes the pdf. As such, the choice could be influenced by ease of exposition.

References

Fader, Peter S., Bruce G. S. Hardie, and Robert Zeithammer (2003), "Forecasting New Product Trial in a Controlled Test Market Environment," *Journal of Forecasting*, **22** (August), 391–410.

Hardie, Bruce G. S., Peter S. Fader, and Michael Wisniewski (1998), "An Empirical Comparison of New Product Trial Forecasting Models," *Journal of Forecasting*, **17** (June–July), 209–229.

Pawitan, Yudi (2001), *In All Likelihood*, Oxford: Clarendon Press.

[4] In this example, we would not be comfortable treating week-level measurements as "actual" times but are happy to treat day-level measurements as "actual" times. You may feel differently.