Adjusting for length biased sampling when assessing the impact of interruptions on total time on task in a hospital emergency department setting.

Bruce M. Brown and William T.M. Dunsmuir

January 6, 2010

Abstract

This technical note outlines a method for testing if interruptions to tasks performed by doctors in an emergency department is associated with lengthening or shortening the total time on the task. The method proposed adjusts for length biased sampling under the assumption that interruptions arrive according to a homogeneous Poisson process. This assumption is assessed in relation to the data collected on a work measurement study in a hospital emergency department.

Contents

1 Introduction

In studying the impact of interruptions on the total time required to complete a task, it is important to take account of the phenomenon of length biased sampling. The longer the time required to complete a given task, the higher the chance of it being interrupted, assuming that interrupting tasks arrive at random according to some stochastic process independently of the task being interrupted. Since longer tasks have a higher chance of being interrupted, those that are interrupted will have longer observed task times simply because of this length biased sampling. Thus any study that compares the average, or more generally the whole statistical distribution of the total time required to complete tasks that have been interrupted once, twice or more times, with the total time for tasks that have not been interrupted, may observe longer task times for interrupted tasks. However unless the impact of length biased sampling is adjusted for in this comparison, it is not possible to conclude that the actual act of interrupting a task causes the task time to be lengthened.

In this paper we present a method, based on a simplifying assumption of interruptions arriving according to a Poisson process, for assessing the impact of length biased sampling on total task time.

We find that, in the application of this method to studying total times to complete tasks performed by doctors in a hospital emergency department, average total time on task for tasks which have been interrupted are *shorter* than are expected when length biased sampling is accounted for. This would imply that in this setting tasks which are interrupted are performed more rapidly then they would have been had they not been interrupted.

2 Poisson Interruptions Model

Assume that for tasks that are interrupted k times, the density of the time on task T is $f_T(t;k)$ and that interruptions occur according to a Poisson process with rate λ (per unit of time). Let K denote the number of interruptions. Then the joint distribution of (T,K) is

$$f_{T,K}(t,k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} f_T(t;k).$$

This can be derived as follows:

$$\begin{aligned} f_{T,K}(t,k) &= P(K=k|T=t) f_T(t;k) \\ &= \frac{e^{-\lambda t}(\lambda t)^k}{k!} f_T(t;k) \end{aligned}$$

since the distribution of the number of events in a time interval of length t for a Poisson process with rate λ per unit of time is Poisson with parameter λt.

If interruptions did not cause increased time on task then

$$f_T(t;k) = f_0(t), \text{ for all } k = 0, 1, \ldots$$

This is the null hypothesis assumption (i.e. interruptions do not increase the time on task).

An alternative hypothesis is that interruptions cause additional penalties attributable to the interruptions. So, for a task that is interrupted k times the penalties could be represented by nonnegative random variables S_i for $i = 1, 2, \ldots, k$. In this case the time on task would be

$$T_k = T_0 + S_1 + \cdots + S_k$$

where T_0 has density given by $f_0(t)$ (the density if interruptions do not add penalties). The times S_i represent the time needed to revert from the interruption back to processing the interrupted task. This formulation accomodates constant (i.e. non-random) resumption penalties as well. Under this model the time on task T_k will have density which is moved to the right (higher values) relative to the null hypothesis values (with no interruption penalties). In particular the expected, or mean, value satisfies

$$E(T_k) \geq E(T_0)$$

We next derive the density for the time on task given there are k interruptions.

$$f(t|K=k) = \frac{\frac{e^{-\lambda t}(\lambda t)^k}{k!} f_T(t;k)}{\int_0^\infty \frac{e^{-\lambda s}(\lambda s)^k}{k!} f_T(s;k)ds} = \frac{t^k e^{-\lambda t} f_T(t;k)}{\int_0^\infty s^k e^{-\lambda s} f_T(s;k)ds}$$

Under the null hypothesis that there is no increase in time on task from interruptions we can derive the density for the observed times on task conditional on there being k interruptions by replacing $f_T(t;k)$ by $f_0(t)$ to get

$$f(t|K=k) = \frac{t^k e^{-\lambda t} f_0(t)}{\int_0^\infty s^k e^{-\lambda s} f_0(s)ds}$$

Note that these densities adjust the null hypothesis time on task density $f_0(t)$ for the impacts of length biased sampling. In particular for $k = 0$ we have

$$f(t|K=0) = \frac{e^{-\lambda t} f_0(t)}{\int_0^\infty e^{-\lambda s} f_0(s)ds}$$

and in particular, it can be proved that

$$E_{H_0}(T|K=0) \leq E_{H_0}(T)$$

confirming the impact of length biased sampling on reducing the average time on task for those tasks that are not interrupted.

Note also that the means and variances can be easily derived. We give the formulae only for the null hypothesis case.

$$\mu_k = E(T|K=k) = \frac{\int_0^\infty t^{k+1} e^{-\lambda t} f_0(t)dt}{\int_0^\infty s^k e^{-\lambda s} f_0(s)ds} = \frac{\mu_0(k+1)}{\mu_0(k)}$$

and

$$\sigma_k^2 = \frac{\int_0^\infty t^{k+2} e^{-\lambda t} f_0(t)dt}{\int_0^\infty s^k e^{-\lambda s} f_0(s)ds} - \mu_k^2 = \frac{\mu_0(k+2)}{\mu_0(k)} - \mu_k^2$$

where $\mu_0(k)$ is the kth moment of the density for non-interrupted tasks $f(t|K = 0)$.

These expressions provide a means to test the null hypothesis that interruptions do not increase the average time on task versus the alternative that they increase or decrease them. This is because there is an abundance of data for the case $k = 0$ of no interruptions enabling the values of $\mu_0(k)$ for $k = 1, 2, 3, 4$ and 5 to be estimated quite accurately. Thus, the means and variances μ, σ_k^2 under the null hypothesis can be determined without needing to estimate the underlying Poisson interruptions rate λ.

Comparison of the means as just described can be augmented by a comparison of the distributions that can be expected under length biased sampling from the Poisson process model. Consider the ratios of densities

$$\frac{f(t|K=k)}{f(t|K=0)} = \frac{\frac{t^k e^{-\lambda t} f_0(t)}{\int_0^\infty s^k e^{-\lambda s} f_0(s) ds}}{\frac{e^{-\lambda t} f_0(t)}{\int_0^\infty e^{-\lambda s} f_0(s) ds}} = \frac{t^k}{\mu_0(k)}$$

Hence

$$\mu_0(k) f(t|K=k) = t^k f(t|K=0)$$

thus ensuring the density is properly normalized. If the observed density is skewed further to the right (to larger times on task values) then this would support rejection of the null in favour of the alternative that interrupted tasks take longer to complete because there is an additional workload penalty from interruptions (beyond what is expected under length biased sampling).

Alternatively, and easier to implement, is to base the comparison on cumulative distribution functions. Thus

$$\hat{F}(t|K=k) = \mu_0(k) \int_0^t s^k \hat{f}(s|K=0) ds$$

can be estimated using a density estimate of $f(s|K = 0)$ based on the large ($n > 7000$) sample, multiplying by s^k and then normalizing by dividing by the integral when t is as large as observed in the sample.

3 Application to ED study.

Under H_0 when all 131 sessions are pooled, there are abundant data ($n > 7000$) to estimate the moments $\mu_0(k)$ required to estimate the means and variances of the time on tasks for interrupted tasks. Call these estimates (using the times on tasks for the non-interrupted tasks) $\tilde{\mu}_0(k)$. Then estimates which use the pooled data presented in the paper (ref) are:

k	1	2	3	4
$\tilde{\mu}_0(k)$	86.4501	2.812654×10^4	2.200293×10^7	2.75428×10^{10}

from which estimates of μ_k and σ_k^2 are for $k = 1$

$$\tilde{\mu}_1 = 325.350,\ \tilde{\sigma}_1^2 = 385.57^2$$

and for $k = 2$

$$\tilde{\mu}_2 = 782.2836,\ \tilde{\sigma}_1^2 = 606.0056^2$$

But the sample estimates are $\hat{\mu}_1 = 180$ and $\hat{\mu}_2 = 315$ based on samples of $n_1 = 612$ and $n_2 = 155$.

Therefore the test statistics

$$Z_k = \frac{\hat{\mu}_k - \tilde{\mu}_k}{\tilde{\sigma}_k/\sqrt{n_k}}$$

for $k = 1, 2$ are

$$Z_1 = -9.326,\ Z_2 = -9.600$$

which is a very strong indication to reject H_0 in favour of the alternative that interruptions *reduce* total time on task. These values are also provided as part of the Table A.1 attached.

Figure A.1 provides a comparison of the predicted cumulative distribution functions $\hat{F}(t|K = k)$ for $k = 1, 2, 3$ interruptions under length biased sampling alone. The actual empirical cumulative distribution function for the observed total time on task are also shown along with that for the tasks which were not interrupted. For the case $k = 3$ interruptions the actual is calculated by combining data from $k \geq 3$ interruptions which is therefore likely to be an overestimate of the true cumulative distribution for total time on task when $k = 3$. Even so, the predicted distribution under length biased sampling for $k = 3$ lies to the right of the of the observed cumulative distribution based on $k \geq 3$. All three comparisons demonstrate that actual total time on task values are stochastically smaller than those that are predicted under length biased sampling alone. Were there an impact of interruptions which increases total time on task the observed cumulative distributions would be stochastically larger than those predicted under length biased sampling. Also shown as vertical lines on the panels of Figure A.1 are the average values reported in Table A.1 for the last group (corresponding to all sessions).

Table A.1 also reports the average total time on task values predicted under length biased sampling alone and actually observed stratified by three session groups: Sessions 1 to 15, Sessions 16 to 66 and Sessions 67 to 131. The reason for this stratification is the question of whether the underlying Poisson interruption rate λ is homogeneous across the whole study, and whether this influences the result. This question is explained in the next section. Note that the rates of interruptions per hour roughly halved from the first to second groups and then halved again from the second to the third groups. Table A.1 demonstrates that regardless of the rate of interruption occurring the conclusion is that interrupted tasks are shorter on average than the non-interrupted tasks. Cumulative distribution function comparisons similar to Figure A.1 for these separate groups of sessions show similar patterns and lead to similar conclusions. This suggests that by combining observations from different sessions does not lead to a different conclusion than that reached for combinations of sessions in which the rates of interruptions are more homogeneous.

Possible explanations:

1. Doctors work to counteract length biased sampling by refusing interruptions for long or complex tasks.

2. Some types of tasks attract interruptions and these are the shorter tasks and conversely other types of tasks tend to repel interruptions and these are the longer tasks.

4 Assessing the Poisson Process Assumption

The analysis of the complete data set presented in the last section (resulting from 131 separate observations sessions) assumes that the rate of interruptions in the Poisson process is homogeneous across the sessions. Note that the above analysis does not stratify on type of task, type of interrupting task, doctor status or seniority etc.

This section investigates the reasonableness of that assumption. We find that:

1. Within each session the assumption of Poisson sampling process seems reasonable (with inter-interruption times being consistent with the exponential distribution which is expected for these in a Poisson process).

2. The rate of interruptions does vary across the session of observation.

3. Some of the variations in the rate of interruptions is associated with factors characterising the session such as doctor seniority and experience but most of the variation is related to temporal patterns unexplained by other covariates.

4. The observation sessions can be crudely separated into three groups with high, medium and low rates of interruptions corresponding to the first, second and third temporal period of observation.

There will be a limit to the extent to which we can reliably partition the data into sub groups by task type, doctor training etc before the sample sizes for $k = 0, 1, 2$ interruptions start to become too small to be reliable.

4.1 Homogeneiety of interruption rate per hour across sessions.

Figure A.2 shows that the rate of interruptions per hour varies markedly throughout the sample. There is a clear lowering of the rate commencing with session 67 (about half way through the period of data collection). In the first half of the data there is an initial downward trend in rate to around session 15 and then a reasonably level rate from session 16 to session 66. For the purposes of producing Table A.1 we selected these three groups (1 to 15, 16 to 66 and 67 to 131) as the basis for combining session data in order to obtain a reasonable about of data for calculating the cumulative distributions and the mean comparisons of Table A.1. While the interruption rate is not strictly homogeneous with these three groups it is a lot more homogeneous that across the whole 131 sessions combined.

Numerous factors were recorded to reflect the load on the ward and the doctor's characteristics for each session of observation. Measures of load on the ward included numbers of patients admitted on the day of the observation session, average age of patients in and immediately prior to the observation session, numbers of admissions from the emergency department to the hospital, presentations to the ward just prior and during the observation session, numbers of medical staff rostered to the ward just prior and during the session, numbers of registrars present and numbers of junior medical officers present. None of these factors was significant in explaining the observed fluctuations in interruption rates across sessions when incorporated in a generalized linear model for counts using a Poisson response distribution, a negative binomial response distribution or using a quasi-likelihood assuming overdispersion in the variance relative to the mean for Poisson responses. Measures on the doctor being observed were their experience, their training and their position. Of these, doctors' experience and seniority were statistically significant when a negative binomial response distribution was assumed for the counts of interruptions in each session - see Table A.2 for results. Additionally we investigated if morning or afternoon or day of the week impacted the interruption rates in sessions and these did not. These two doctor related variables were not significant when the quasi likelihood procedure was used in which overdispersion in the variance relative to the mean relationship was allowed for. These mixed results, together with the smaller contribution relative to the temporal means for the three periods, suggest that interruption rates cannot be reliably predicted by the factors measured. Other factors are required to capture the complex manner in which interruptions arise in the emergency department setting.

Of all the variables considered the significant ones (period of data collection, experience and seniority of doctors being observed) are listed in Table A.2 for the negative binomial fit and Table A.3 for the quasi -likelihood fit. All modelling was done using the Splus statistical package or the R statistical package.

4.2 Times Between Interruptions

Given the lack of homogeneity across sessions discussed in the last subsection it is not possible to use the times between interruptions aggregated over all sessions in a test for the exponential inter interruptions distribution that would be expected for a homogeneous Poisson process. The empirical distribution function of the observed times between interruptions was compared with the exponential distribution for each session using the Kolmogorov-Smirnov goodness of fit statistic. Over all sessions for which there was sufficient data to compute this statistic three sessions had P-values for rejecting the assumption of exponential times between interruptions slightly less than 0.05, which is as to be expected if the null hypothesis of exponential times is true. Some sessions have very few interruptions so the power to detect departures from the exponential distribution is low. However there are 22 sessions with at least 20 interruptions, 8 with at least 30 and 4 with at least 40 interruptions. In none of these sessions was there evidence to suggest that the exponential distribution should be rejected.

Further work could be carried out to test the exponential assumption and thereby the reasonableness of the Poisson process assumption used here.

However since we have observed no marked deviation from the assumption of Poisson process for interruptions and since the conclusions presented above concerning the impact of interruptions on shortening total time on task are robust to the levels of interruptions in three time periods with progressively halving of interruption rates it is unlikely that a more complex model for the interruptions process than the Poisson will alter the conclusion substantively.

5 Analysis Excluding Tasks Which Are Not Resumed

The analysis presented above is based on all tasks. However some tasks that were interrupted were not resumed. These tasks are clearly incomplete and the total time on task will likely be less than those recorded for tasks which are completed. Additionally some tasks, interrupted or not, were recorded as ending precisely at the end of the observation session and there is a possibility that these tasks were truncated prematurely. In view of these comments the question arises whether or not the above results on the impact of length biased sampling would continue to hold when tasks that were not resumed and tasks that may have been prematurely ended are removed from the analysis. This section briefly addresses this question.

Of the 8369 new tasks considered in the above analysis 881 were interrupted one or more times. Of these interrupted tasks 163 (18.5%, approximate 95% CI [15.9%, 21.1%]) were not resumed by the end of the observation session in which they occurred. Of course, for tasks that were not interrupted the question of resumption is irrelevant and so there is 0% of non-interrupted tasks that are completed in this sense. Hence interruption of a task results in nearly 1 in 5 tasks being left incomplete. The impact of excluding these nonresumed interrupted tasks on the analysis of length biased sampling presented in Table A.1 is shown in Table A.4 for data set (B). These results can be compared with the results from Table A.1 presented also as data set (A) in Table A.4. As might be expected the observed means for total time on tasks are higher for data set (B) (which removes incomplete tasks) than for data set (A). Since the predicted means are based on non-interupted tasks these do not change between data sets (A) and (B). The result of removing non-resumed and therefore incomplete tasks on our assessment of the impact of length biased sampling is clearly minor and the overall conclusion we have reached above do not materially change. Figure A.3 provides the graphical results for data set (B) and can be compared with Figure A.1 for data set (A).

One might argue that tasks that end precisely at the end of the observation session to which they belong could be truncated or censored by the end of session. There are 116 such tasks obtained over the 131 observation sessions. Of the 7488 tasks that were not interrupted 113 (1.5%) ended at session end. Of the 718 (881 less 163) tasks that were interrupted and resumed 3 terminated at the end of session. If these additional 116 tasks are removed from the analyis we arrive at data set (C) which excluded non-resumed tasks and tasks that might be considered to be prematurely truncated by session end. The results for this case are also included in Table A.4. The main difference between the results for data sets (B) and (C) are in the predicted length under length biased sampling and this is as expected because the majority of extra excluded tasks are from the 7488 that were not interrupted. However the results are scarely different between these two data sets. If tasks which were interrupted and not resumed are combined with those that terminate at end of observation session to arrive at a group of tasks that were 'incomplete' in either sense then the proportion of interrupted tasks that are incomplete is 18.8% (166 tasks out of 881 interrupted tasks) and proportion of non-interrupted tasks that are incomplete is 1.5%. Thus in this sense interruptions lead to significantly higher incomplete tasks (Chi square 734.88, df =1, $P < 0.0001$).

For the noninterrupted cases those that ended at the end of the observation session had significantly larger mean total time on task than those that ended before the end of session using a two sample t-test on the logarithms of the total time on task values (results: $t = 6.02$, $P < 0.0001$). This result is not consistent with the hypothesis that tasks which are recorded as ending with the observation session are truncated prematurely or, in other words, have right censored total time on task values. Approximate 95% confidence intervals for the average total time on task are $[2.10, 3.89]$ using $n = 113$ observations which ended at session end and $[1.36, 1.47]$ using $n = 7375$ observations which ended before session end.

Tables

Table A.1

Caption: Comparison of observed average total time on task (minutes) with that predicted under length biased sampling and no additional penalty due to interruptions. Note the predicted average total time on task for 3 or more interruptions is actually calculated using the formula for 3 interruptions.

Sessions	Average Interruption Rate per hour	Number of interruptions	Sample size (n)	Average predicted under Length Biased Sampling	Observed Average	Percentage Observe Mean to Predicted	Z (see note)
1 to 15	15.4	1	145	4.50	2.02	45%	-5.0
		2	40	12.60	2.47	20%	-6.4
		3 or more	42	18.63	4.98	26%	-11.2
16 to 66	7.7	1	263	5.00	3.33	67%	-4.8
		2	92	11.17	5.45	49%	-7.0
		3 or more	59	16.75	6.80	41%	-9.8
67 to 131	3.3	1	204	5.88	3.28	56%	-5.3
		2	33	14.23	7.08	73%	-3.6
		3 or more	13	23.07	6.27	27%	-4.6
ALL	6.2	1	612	5.41	3.00	55%	-9.3
		2	155	13.03	5.25	40%	-9.6
		3 or more	114	20.87	6.05	29%	-13.4

Note: All test statistics reject the null hypothesis of no impact due to interruptions with P-values no larger than 0.0005 in all cases.

Table A.2

Caption: Results from fitting of a generalized linear model with negative binomial response distribution to the interruptions in 131 sessions of observation.

Model Term	Estimate	Standard Error	P-value (significance)
Intercept (sessions 1 to 15)	3.77	0.47	<0.0001
Sessions 16 to 66	-0.68	0.17	<0.0001
Sessions 67 to 131	-1.56	0.17	<0.0001
Doctor's Position	-0.33	0.13	0.0113
Doctor's Experience	-0.059	0.031	0.0575

Table A.3

Caption: Results from quasi likelihood fitting of a generalized linear model to the interruptions in 131 sessions of observation. Overdispersion is allowed for in this fit.

Model Term	Estimate	Standard Error	P-value (significance)
Intercept (sessions 1 to 15)	3.47	0.49	<0.0001
Sessions 16 to 66	-0.71	0.13	<0.0001
Sessions 67 to 131	-1.56	0.15	<0.0001
Doctor's Position	-0.25	0.14	0.074
Doctor's Experience	-0.032	0.033	0.330

Table A.4

Caption: Comparison of observed average total time on task (minutes) with that predicted under length biased sampling and no additional penalty due to interruptions. Three selections of tasks: (A) All tasks including those that were not resumed after interruption; (B) Excluding tasks which were interrupted and not resumed; (C) Excluding tasks which were interrupted and not resumed and tasks which were shown as ending exactly at the end of the observation session.

Data Set based on all 131 sessions	Number of interruptions	Sample size (n)	Average predicted under Length Biased Sampling	Observed Mean	Percentage Observed Mean to Predicted Mean	Z (see note)
(A) ALL Tasks	1	612	5.41	3.00	55%	-9.3
	2	155	13.03	5.25	40%	-9.6
	3 or more	114	20.87	6.05	29%	-13.4
(B) Tasks not resumed	1	480	5.41	3.45	64%	-6.7
	2	143	13.03	5.46	42%	-9.0
	3 or more	95	20.87	6.67	32%	-11.7
(C) Tasks not resumed and ended before session end	1	477	5.25	3.47	66%	-6.3
	2	143	12.70	5.46	43%	-8.5
	3 or more	95	20.78	6.67	32%	-13.3

Note: All test statistics reject the null hypothesis of no impact due to interruptions with P-values no larger than 0.0005 in all cases.

Figures

Figure A.1

Caption: Comparison of cumulative distribution functions of total time on task predicted under length biased sampling with that observed aggregated over all 131 observation sessions. Vertical lines mark the mean values listed in Table A.1.

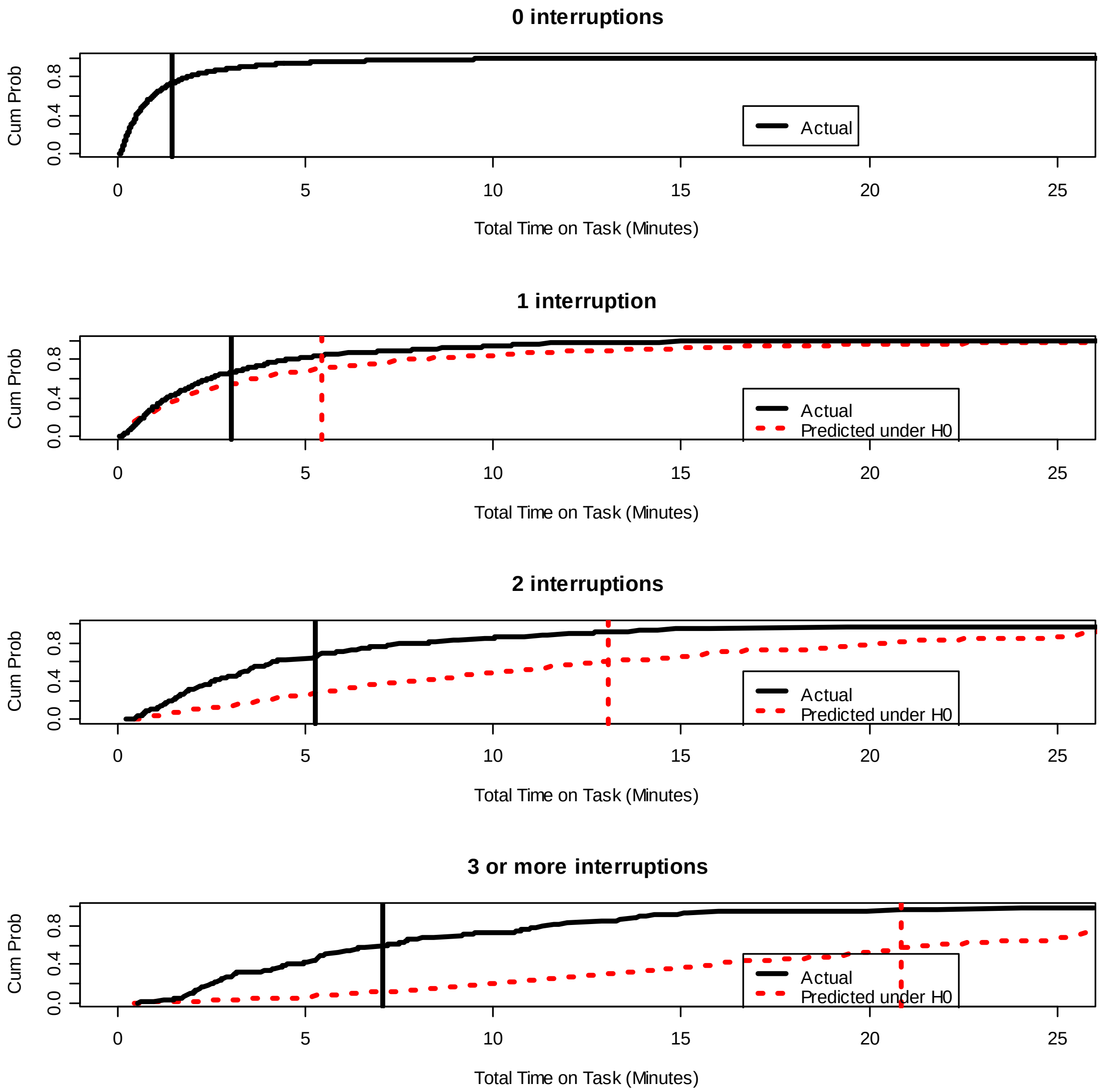

Figure A.2

Caption: Observed Interruption Rates per Session over 131 observation sessions with average interruption rates for three periods (Sessions 1 to 15, 16 to 66 and 67 to 131).

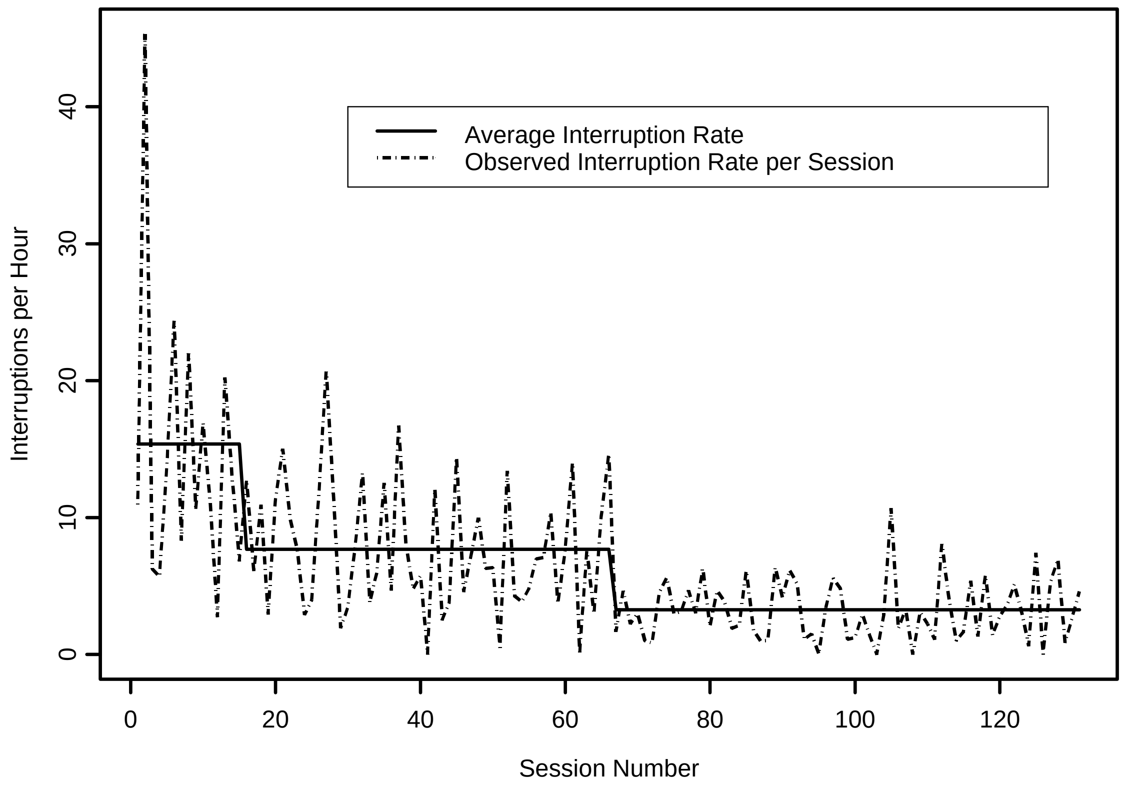

Figure A.3

Caption: Comparison of cumulative distribution functions of total time on task predicted under length biased sampling with that observed aggregated over all 131 observation sessions excluding tasks that were interrupted and not resumed before session end – data set (B) from Table A.4. Vertical lines mark the mean values listed in Table A.1.