Survival Analysis on Duration Data in Intelligent Tutors

Michael Eagle and Tiffany Barnes

North Carolina State University, Department of Computer Science,
890 Oval Drive, Campus Box 8206 Raleigh, NC 27695-8206
{mjeagle,tmbarnes}@ncsu.edu

Abstract. Effects such as student dropout and the non-normal distribution of duration data confound the exploration of tutor efficiency, time-in-tutor vs. tutor performance, in intelligent tutors. We use an accelerated failure time (AFT) model to analyze the effects of using automatically generated hints in Deep Thought, a propositional logic tutor. AFT is a branch of survival analysis, a statistical technique designed for measuring time-to-event data and account for participant attrition. We found that students provided with automatically generated hints were able to complete the tutor in about half the time taken by students who were not provided hints. We compare the results of survival analysis with a standard between-groups mean comparison and show how failing to take student dropout into account could lead to incorrect conclusions. We demonstrate that survival analysis is applicable to duration data collected from intelligent tutors and is particularly useful when a study experiences participant attrition.

Keywords: ITS, EDM, Survival Analysis, Efficiency, Duration Data

1 Introduction

Intelligent tutoring systems have sizable effects on student learning efficiency — spending less time to achieve equal or better performance. In a classic example, students who used the LISP tutor spent 30% less time and performed 43% better on posttests when compared to a self-study condition [2]. While this result is quite famous, few papers have focused on differences between tutor interventions in terms of the total time needed by students to complete the tutor. In many studies of intelligent tutoring systems, time is simply held constant for two groups, and efficiency then boils down to comparing the number of problems each group could solve in the given time and the results of posttest measures. However, it is not clear how to factor students who were not able to complete the tutor into this analysis. In this work, we explore tutor efficiency in terms of time and performance, while taking student *dropout* (ceasing to interact with the tutor before completion) into account.

College students often use computer-based tools to complete homework assignments, but no specific time limits apply. Typical time duration distributions

violate the normality assumptions of many statistical tests and measures of central tendency. Anderson, Corbett, Koedinger, and Pelletier used mean duration data to compare differences between groups of students with and without intelligent feedback in the LISP tutor [1]. The authors state that the mean times (for the control group) are underestimates, as many students in the control (no-feedback group) did not complete all assignments. In other words, if the control group persisted, the time they took to complete tasks would have been longer than the observed durations for the few high-performing students who were able to persist without feedback. This study illustrates how dropout can obscure the true impact of an intervention.

Our exploration of tutor efficiency has three important elements: performance (tutor completion percentage), duration (total time spent interacting with the tutor), and dropout (whether stopped before completion). Dropout can easily confound the results of duration and performance. Different dropout rates between experimental groups can cause attrition bias [10], where groups completing the study are self-selected due to achievement levels; this self-selection causes the sample to become different than the target population and hampers the study's generalizability [8]. When dropout exists, more complex analyses are needed to study learning efficiency; not only are results suspect for generalization purposes, but the data itself contains missing values because of high dropout rates. By modeling tutor data with high dropout rates using survival analysis, we hypothesize that we can build a more detailed understanding of tutor efficiency and explain differences between groups in an educational intervention.

In this study, we investigate data from a prior study of the Deep Thought logic tutor comparing versions with and without hints. Stamper et al. found that the odds of a student in the control group dropping out of the tutor after the first six problems were over 3.6 times higher when compared to the group provided with (data-driven and automatically generated) hints [12]. Students given access to hints also had better tutor performance, as well as higher overall course scores. However, comparison of duration means showed no differences in overall time spent in the Deep Thought logic tutor between the hint and control groups. This is likely because this comparison does not take into account student dropout. In this study, we applied survival analysis to data from Stamper et al.'s study to more fully explore the impact of hints on performance, duration, and dropout. We hypothesize that students given access to hints in the Deep Thought logic tutor, spend less time in tutor while also performing better than students without hints. In other words, the tutor efficiency for Deep Thought with hints is higher than that for they system without hints. We found that students given automatically generated hints take 55% of the time that students in the control needed to complete the tutor.

1.1 Methods and Materials

We perform our experiments on the Spring and Fall 2009 Deep Thought propositional logic tutor [6] dataset as analyzed by Stamper, Eagle, and Barnes in 2011[13]. Data was collected from six deductive logic courses, taught by three

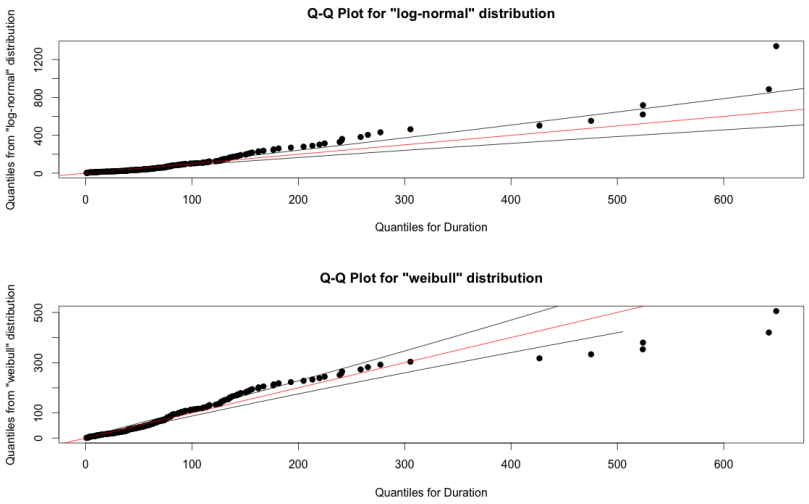

Fig. 1. QQ-Plots for the log-normal and Weibull distributions, the primary difference appears to be that the Log-normal is sensitive to very small durations, while the Weibull distribution is sensitve to very large durations.

professors. Each instructor taught one class using Deep Thought with automatically-generated hints available (hint group) and one without any additional feedback (control). The dataset includes 105 students in the Hint group and 98 students in the Control group. In Deep Thought, students choose the amount of time they spend using the online tutor; however, they were graded on the completion of 13 specific proofs.

The variables we use for this study are:

Group a two level factor (Hint, Control) depicting the student's experimental condition

ProblemDuration the sum of the time taken over all steps in a problem until 1st completed (max 3min per step)

Duration the sum of problem durations over all 13 problems

Performance a number between 0–13 representing the number of proofs solved by the student

Dropout a boolean (True, False) defined as true for students who stop engaging with the tutor without completing the assignment ($Performance \neq 13$)

Duration data often falls into a set of known distributions [3] [9]. Q-Q plots (figure 1) and histogram/density plots (figure 2) allowed us to narrow the possible distributions down to log-normal[5] or Weibull[16]. The primary difference appears to be that the log-normal does not fit well to early dropout (small durations), while Weibull does not fit as well for extremely long durations.

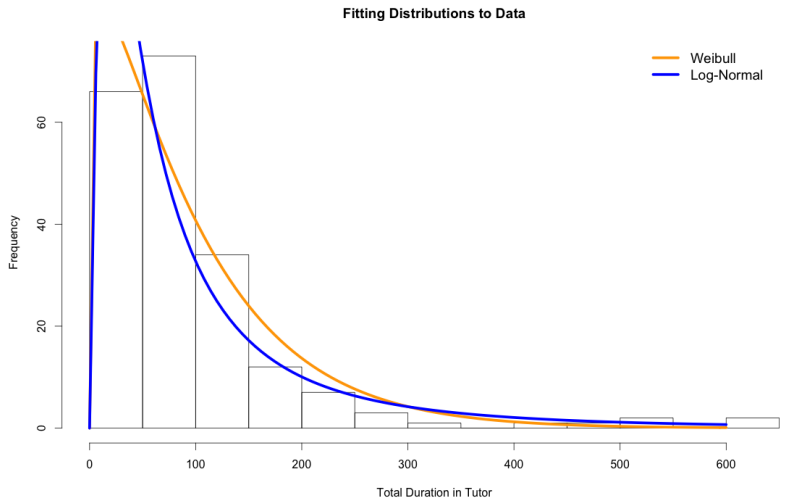

Fig. 2. Histogram with density plots for the Weibull and log-normal distributions. Both seem to fit reasonably well.

1.2 Survival Analysis

Survival analysis is a series of statistical techniques that deal with the modeling of "time to event" data [7]. Survival analysis, also known as reliability analysis or duration analysis in economics, is named for its start as a method to measure survival after applying a medical intervention.

Survival analysis includes techniques for unknown values, non-parametric data, log-normal and Weibull probability distributions, and between-groups testing. We use the survival package for R [15] to perform our analysis of learning efficiency, where the event for survival analysis is tutor completion.

"Censoring" allows for modeling duration with unknown values. Right censoring occurs when participant data is lost before tutor completion, while left censoring would be when completion time is known but start time is not. For our data, the duration for students who drop out, or stop using the tutor is right censored, since we know the start time but do not know how long it would have taken the student to complete the tutor. For example, a student who has completed 5 problems but then quits is considered right censored as we do not know how long it would have taken the student to complete all 13 problems.

The survival function is defined as:

$$S(t) = Pr(E > t) = 1 - F(t) \tag{1}$$

where t is the time in question, E is the time of the event (tutor completion), Pr is probability, F(t) is the duration distribution. This function gives the probability that the time of the tutor completion event, E, is later than t. That is, the probability that the student has not completed the tutor.

The duration distribution function, which is found via the cumulative distribution function $cdf(t)$, is the probability of observing a problem completion time E less than or equal to some time

$$F(t) = Pr(E \leq t) = 1 - S(t) = cdf(t) \tag{2}$$

The derivative of F(t) is the probability density function (pdf) of the duration distribution,

$$f(t)Pr(E = t) = F'(t) = \frac{d}{dt}F(t) = pdf(t), \tag{3}$$

which provides us with the probability of observing a single tutor completion time E at some time t. The hazard function, which tells us the instantaneous completion rate at time t, is:

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq E < t + dt | E \geq t)}{dt} = \frac{f(t)}{S(t)} = \frac{pdf(t)}{1 - cdf(t)}. \tag{4}$$

This is the probability of the event occurring at time t given that the event has not yet occurred.

There are two models we consider for measuring effects of covariates: the accelerated failure time (AFT) model and the Cox proportional hazards model. AFT assumes that the effect results in one group that completes the tutor more quickly, while the Cox proportional hazards model assumes that the tutor completion rate for one group is a constant multiple of that for the other. We have chosen the AFT model, which assumes that the effect of the covariates, θ, is to accelerate the time to tutor completion by some constant factor [14].

$$S(t|\theta) = S(\theta t) \tag{5}$$

The AFT model assumptions fit with our hypothesis that hints shorten the time it takes to finish the tutor. In addition, it is easy to interpret θ as a direct modifier to tutor completion time, and AFT facilitates using data from log-normal and Weibull distributions.

2 Results

To explore the differences between the hint and control groups we submitted the data to an AFT model as both a log-normal and Weibull distributions. The log-likelihood scores were -336.6 and -341.2 respectively. We chose to use the log-normal distribution, however both models fit similarly well and had similar results. Investigation showed the log-normal fit less well for early dropout students, while Weibull fit less well for students with extremely long durations. The probability distribution function (pdf) and cumulative distribution function (cdf) for the log-normal distribution are:

$$pdf(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{[\ln(t)-\mu]^2}{2\sigma^2}}, cdf(t) = \Phi(\frac{\ln t - \mu}{\sqrt{2\sigma^2}}). \tag{6}$$

where Φ(x) is the cumulative distribution function (cdf) of the standard normal distribution. Note that we use $\ln(t)$ when using Φ, we can do this thanks to the assumption that the log of the duration data shows a normal distribution.

The AFT model was statistically significant $\chi^2 = 9.21$ on 1 degree of freedom, $p = 0.0024$, $n = 202$, the coefficients of the model had the intercept (mean) as 5.655, the effect of Hint θ as $-.599$, and the SD (scale) as 0.948. The effect of hints is $e^{-.599} = 0.55$; this means that it takes the Hint group 55% of the time it takes the control group to solve all 13 tutor problems. We have plotted the inverse of the survival curve in figure 4.

Figure 3 shows the hazard function for the duration data, in other words, the instantaneous completion rate for each of the groups. It also shows the probability density function for the completion rate. Overall, these plots give us a good overview of the shape of the duration data, showing that the probable total duration for students in the control group, if they were to complete the tutor, would be much longer than that for students in the hint group. One concrete measure of this is illustrated by the median of the survival function, the location where 50% of people have completed the tutor. The median is found by e^{μ}, which is $e^{5.65} = 284.29$ for the control group and $e^{5.65-.599} = 156.18$ for the hint group. Comparing these medians illustrates again the considerable difference in duration, or time to tutor completion, between the groups.

We measure the difference between groups with a Student's t test to explore possible differences in performance between the two groups. We have no reason to believe that the total tutor scores are not normally distributed. We find that the total performance in tutor between the hint group ($M = 9.26$, $SD = 4.26$) and the control group ($M = 6.78$, $SD = 4.62$) was significant, $t(200) = 3.98, p < .001, 95CI = (1.25, 3.71)$, with the Hint group solving between 1.25 and 3.71 more problems than the control group. To illustrate these differences at different points in time, we have added points to the survival curve in figure 4 indicating the mean performance score for students who left the tutor (by completing or dropping out) within the 20%, 40%, 60%, and 80% quantiles of the maximum duration. This lets us compare relative performance in the tutor between the two groups. Both groups have similar scores at about the 30 minute mark, but the hint group experiences a large increase in performance by the 60 minute mark. After this, the rate of growth in score decreases, this is likely because students that take an exceptionally long time are less skilled.

To illustrate the impact of dropout, we compare the results of survival analysis to a more traditional between-groups testing method. To explore differences in overall time in tutor between the two groups, we subjected the total elapsed time on all 13 problems to a 2-tailed Student's t-test. The total time in tutor between the hint group ($M = 86.05$, $SD = 69.80$) and the control group ($M = 122.95$, $SD = 122.94$) was not significant, $t(200) = -1.34, p = 0.183$.

However, since we know the data isn't from a normal distribution, we can improve on this accuracy by using a data transformation. To normalize the data, we use a logarithmic transformation (common log, base 10) to make the data more symmetric and homoscedastic. We subjected the log-transformed data to

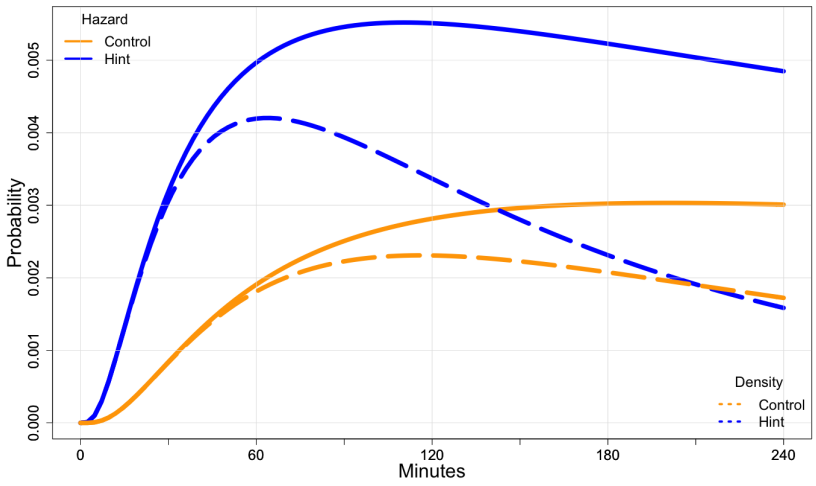

Fig. 3. The probability density functions, represented by the dashed lines, provide the probability of observing tutor completion at a specific time. The hazard functions, the solid lines, are the probability of observing tutor completion at a specific time, given that it has not occurred yet. The probability of completion grows rapidly before becoming stable and eventually decreasing.

a 2-tailed Student t-test. The difference in the logs of duration between the hint group ($M = 8.20$, $SD = 1.02$) and the control group ($M = 8.17$, $SD = 1.21$) was not significant, $t(200) = .168, p = 0.867$. The ratio of the duration between groups is calculated by taking the difference between the means of the groups, since $\lg(x) - \lg(y) = \lg(\frac{x}{y})$. The confidence interval from the log-data estimates the difference between the population means of log transformed data. Therefore, the anti-logarithms of the confidence interval provide the confidence interval for the ratio. The anti-log of the log-transformed means provides us with the geometric mean, the anti-log of the transformed standard deviation is not interpretable. However, we can use the anti-log of the confidence intervals. The most useful statistic we can derive is the difference ratio, and its corresponding confidence intervals. A difference ratio of 0.026 between the means of the logged data equates to $10^{.026} = 1.06$ with a 95% confidence interval of CI (0.52, 2.19).

3 Discussion

The results of the survival analysis allow us to reveal striking differences between the hint and control groups in terms of the time needed to complete the tutor. Students in the hint group complete the tutor in less than half the time needed for students in the control group. It is interesting that the control group and the hint group do not have *observed* differences in overall tutor time; in other

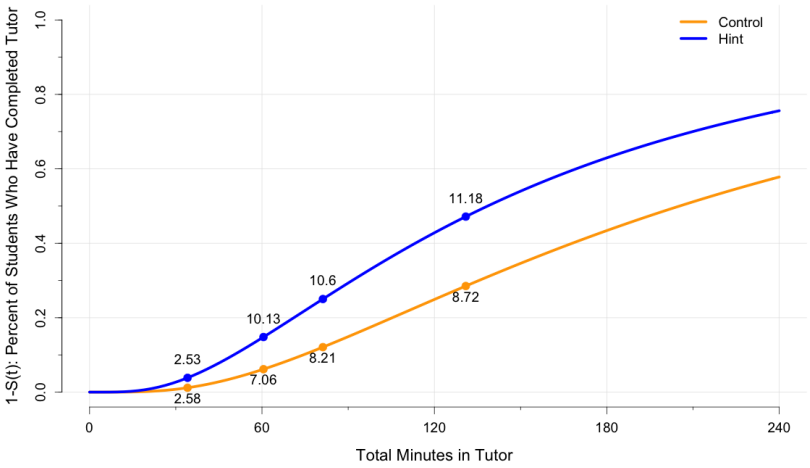

Fig. 4. The percent of students who have completed the tutor over time. We have added points with the mean tutor score (max of 13) to each curve at the 20%, 40%, 60%, and 80% quantiles of total duration.

words - students in the control group don't generally complete the tutor, so we can't observe that they would take twice as long. It is likely that, given the nature of the online access tutor, students are only willing to spend a certain amount of time on this homework assignment. This can explain the observations of differences in tutor progress observed at different times in figure 4.

Using survival analysis, we have estimated that the median duration (tutor completion time) is 284 minutes for the control group and 156 minutes for the hint group. Dividing this by the number of problems in the tutor (13) gives us an estimate of efficiency, since it gives a time per problem needed for solution. The control group is therefore spending about 21 minutes per problem on average, while the hint group is spending an average of 12 minutes per problem. Although this estimate is derived using curves to estimate the (unknown) completion time for the control group, it does in fact fit with the observed data in the first several problems, before significant dropout in the control group occurred. Given these very different rates, we can see that the control group could be discouraged by solving less than 3 problems in an hour, while the hint group could solve 5 in the same amount of time. We were in fact surprised, after realizing these estimates, that students in the control group did not drop out sooner than they did!

This back-of-the-napkin estimate of efficiency is one objective measure that suggests reasons for differences in student behavior (e.g. choice to persist or not). Perception may also play a role in explaining why students in the control group drop out. One possible reason is that the students perceive that the time they are spending is not "worth it." Breen et al. [4] defined the efficiency of a tutor,

for how the student perceives it, as the "belief or judgement that information can be accessed without wasting time or effort." Scanlon and Issroff [11] posit that computer-based instruction can conflict with the student's perceptions of division of labour within learning context. In other words, students using computer-based instruction must be more self-directed and manage their own learning. The feedback provided by the tutor with hints might have helped students in the hint group feel more directed, while also helping them when they were stuck. This could have led to improved student perceptions of efficiency.

Without survival analysis, we would not be able to use observed duration to make any conclusions regarding potential differences between the hint and control groups. Using survival analysis, we can estimate the differences between groups by accounting for unknown values - the total time it would have taken students who dropped out (in both groups) to complete the tutor. Survival analysis has also enabled us to answer questions like "How much time is needed so that 50% of the students can complete the tutor". Using the survival function $S(t) = .5$, we can estimate that the control group needs about 4.76 hours before 50% of students are done, while the hint group needs just 2.61 hours for half the group to complete the tutor. The survival function can be used to decide how much time needs to be allocated in schools for students to use a tutor. We are considering using these estimates to proactively indicate to students when they might need to seek outside help. For example, if a student has taken more than the estimated time for half of students in their group to complete the tutor, we could suggest they speak to a teaching assistant.

4 Conclusions and Future Work

As more learning systems become used outside of traditional classrooms it is imperative that educational data mining researchers leverage methods such as survival analysis that can handle non-normal data with high dropout rates. In this paper, we have used survival analysis to re-analyze the data from six 2009 logic courses using the Deep Thought logic tutor both with and without hints. The original paper showed that students without hints were over 3.6 times more likely to drop after the first six problems when compared to students offered hints. However, standard analyses were insufficient to show the impact of hints on the time needed to complete the tutor between the two groups. Using survival analysis, we have been able to estimate the total duration for both hint and control groups while taking into account dropout data, showing that students in the hint group take 55% of the time to complete the tutor than students in the control group. Using these estimates, we were able to explain approximate time per problem in the tutor for each group. This analysis sheds light on the probable reasons for dropout in the control group. Without these analyses, we might have concluded that students in the control group gave up sooner or were not persistent. However, in reality we see that these students are in fact persistent and spend a considerable amount of time in the tutor - equal to the amount of time spent in the tutor by the hint group. The difference is tutor efficiency:

students in the hint group performed more efficiently, and were therefore able to complete the tutor, while the control group spent a similar amount of time but was less likely to be able to finish. This is a much richer understanding of the differences in effects between the two groups than traditional methods provide. The survival function also allows us to make predictions on how much time is needed for tutor completion, both for teacher planning and student feedback. These results suggest that survival analysis is a powerful toolbox for investigating the impact of interventions on learning efficiency while accounting for performance, duration, and dropout.

References

1. J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):pp. 167–207, 1995.
2. J. R. Anderson and B. J. Reiser. The lisp tutor. *Byte*, 10(4):159–175, 1985.
3. W. R. Blischke and D. P. Murthy. *Reliability: modeling, prediction, and optimization*, volume 767. Wiley, 2011.
4. R. Breen, R. Lindsay, A. Jenkins, and P. Smith. The role of information and communication technologies in a university learning environment. *Studies in Higher Education*, 26(1):95 – 114, 2001.
5. E. L. Crow and K. Shimizu. *Lognormal distributions: Theory and applications*, volume 88. CRC PressI Llc, 1988.
6. M. J. Croy. Graphic interface design and deductive proof construction. *J. Comput. Math. Sci. Teach.*, 18:371–385, December 1999.
7. D. W. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley-Interscience, New York, NY, USA, 2nd edition, 2008.
8. K. A. McGuigan, P. L. Ellickson, R. D. Hays, and R. M. Bell. Adjusting for attrition in school-based samples bias, precision, and cost trade-offs of three methods. *Evaluation Review*, 21(5):554–567, 1997.
9. W. Q. Meeker and L. A. Escobar. *Statistical methods for reliability data*, volume 314. Wiley. com, 1998.
10. R. B. Miller and C. S. Hollist. Attrition bias. 2007.
11. E. Scanlon and K. Issroff. Activity theory and higher education: evaluating learning technologies. *Journal of Computer Assisted Learning*, 21(6):430–439, 2005.
12. J. Stamper, M. Eagle, T. Barnes, and M. Croy. Experimental evaluation of automatic hint generation for a logic tutor. *International Journal of Artificial Intelligence in Education (IJAIED)*, 22(1):3–18, 2012.
13. J. C. Stamper, M. Eagle, T. Barnes, and M. Croy. Experimental evaluation of automatic hint generation for a logic tutor. In *Proceedings of the 15th international conference on Artificial intelligence in education*, AIED'11, pages 345–352, Berlin, Heidelberg, 2011. Springer-Verlag.
14. Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.
15. T. M. Therneau. *A Package for Survival Analysis in S*, 2014. R package version 2.37-7.
16. W. Weibull et al. A statistical distribution function of wide applicability. *Journal of applied mechanics*, 18(3):293–297, 1951.