Stockholm University

Mathematical Statistics
Stockholm University
Bachelor thesis **2026:16**
https://www.math.su.se

Survival Analysis and Failure Predictors of HDDs from the Backblaze Dataset

Sasha Tomic*

September 2026

Abstract

This thesis applies survival analysis to evaluate the reliability of data-centre hard disk drives (HDDs), using a large, left-truncated dataset of 14TB drives published by Backblaze and covering the period 2021–2025, in order to quantify how SMART error indicators, operating temperature, and power cycling are associated with a drive's remaining lifetime.

Diagnostics based on Schoenfeld residuals showed that the hazard ratios implied by a Cox proportional hazards model vary with drive age, so the proportional hazards assumption does not hold in these data. Because the Weibull distribution is the only accelerated failure time (AFT) family that also assumes proportional hazards, it was excluded on structural grounds; the analysis instead adopts a log-logistic AFT specification, which imposes no such restriction, with Weibull and log-normal fits reported throughout as robustness comparisons.

Reallocated and pending sectors (SMART 5 and 197) were associated with substantially reduced expected lifetimes, with estimated time ratios of approximately 0.49 and 0.15 respectively, and were the most consistent predictors across every distributional family and sensitivity check performed. Mean operating temperature showed a small positive association with lifetime, and drives manufactured by Western Digital survived longer than comparable Seagate and Toshiba units. The estimated log-logistic shape parameter implies a hazard that increases across the full observed age range for drives without SMART faults, but one that is already near its peak for drives reporting pending sectors.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sato7286@student.su.se. Supervisor: Jan-Olov Persson, Martina Favero.