**Analiza żywotności dysków twardych**

**Autor: Rafał Kaczmarek**

**Streszczenie**

Niniejsza praca porusza tematykę żywotności dysków twardych. W celu analizy tego zjawiska wykorzystano rozkład wykładniczy oraz rozkład Weibulla. Estymacja parametrów została wykonana Metodą Największej Wiarygodności a do weryfikacji hipotez wykorzystano LR test oraz Z test. Z przeprowadzonego badania wynika, że nie można przyjąć oczekiwanej żywotności równej 365 dni, rozkładu Weibulla dla tego zjawiska nie można sprowadzić do rozkładu wykładniczego lub Rayleigha, a parametry dla rozkładu Weibulla wykorzystanego w celu określenia żywotności obrabiarek nie opisują rozkładu Weibulla żywotności dysków twardych.

**Wstęp**

Świat wokół nas dynamicznie się zmienia. Wraz z postępem technologicznym wzrasta ilość danych, które muszą być nie tylko przetworzone, ale także zmagazynowane. Na podstawie artykułu z NodeGraph[[1]](#footnote-1) wynika, że w 2017 roku było 2,7 zetabajtów danych w naszym świecie. Szacuje się, że w 2020 roku będzie to 44 zetabajtów, a w 2025 roku aż 175 zetabajtów (1 zetabajt=1 099 511 627 776 GB). Ukazuje to skalę zapotrzebowania na nośniki do przechowywania danych zarówno teraz, jak i w przyszłości. Jednakże posiadanie dysku do magazynowania danych o odpowiednich parametrach nie jest wystarczające w sytuacji gdy dysk szybko ulegnie awarii i dane zostaną utracone. W związku z tym podjęto się ważnego tematu w kwestii magazynowania danych jakim jest żywotność dysków twardych i dokonano analizy rozkładu czasu działania dysków twardych aż do wystąpienia awarii. Wykorzystano rozkład wykładniczy oraz rozkład Weibulla, a także zweryfikowano hipotezy proste, z których jedna dotyczy rozkładu wykładniczego a dwie rozkładu Weibulla. Hipoteza dotycząca rozkładu wykładniczego zakłada, że wartość oczekiwana badanego zjawiska będzie równa 365 dni, co odpowiada roku bezawaryjnej pracy. Poza tym zweryfikowano hipotezy zakładające, że rozkład Weibulla w przeprowadzanym badaniu można sprowadzić do rozkładu wykładniczego (parametr kształtu *k* = 1) lub rozkładu Rayleigha (parametr *k* = 2).[[2]](#footnote-2) Dodatkowo zweryfikowano hipotezę złożoną zakładającą, że badane zjawisko dla dysków twardych może być opisane rozkładem Weibulla z parametrami *k* = 34.05 oraz λ = 945.25. Parametry te zostały oszacowane dla rozkładu Weibulla wykorzystanym w celu określenia żywotności obrabiarek w artykule autorstwa Jihong Yan oraz Dingguo Hua[[3]](#footnote-3).

**Dane**

Zbiór danych zawiera informacje z okresu styczeń 2016 - grudzień 2017 na temat numerów seryjnych oraz czasu żywotności dysków twardych określonego w dniach. Został pobrany ze strony Kaggle.com[[4]](#footnote-4), a powstał on na podstawie danych udostępnionych przez firmę BlackBaze. Zbiór ten początkowo zawierał 2987 obserwacji, ale usunięto te obserwacje, dla których liczba dni czasu pracy dysku wynosiła 0. Zostało to wykonane, ponieważ czas działania równy zero może sugerować błąd w danych, wadę fabryczną dysku, niepoprawny montaż lub inne nieoczekiwane zdarzenie, a w wykonywanej analizie brane są pod uwagę jedynie działające dyski i nieuwzględnione są niespodziewane zjawiska. Uzyskano zbiór danych o 2964 obserwacjach. Dodatkowe informacje zostały przedstawione w tabeli 1.

Tabela 1. Podstawowe statystyki opisowe danych użytych w analizie (w dniach).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1 Kwartyl** | **Mediana** | **Średnia** | **3 Kwartyl** | **Maksimum** |
| 1 | 123 | 293 | 311 | 482 | 730 |

*Źródło:* Opracowanie własne

**Metoda**

Analiza czasu pracy dysków twardych została wykonana przy użyciu rozkładu wykładniczego oraz rozkładu Weibulla. Rozkłady zostały wybrane na podstawie literatury, w której analizowano podobne zjawiska (m.in. Yan J., Hua D.[[5]](#footnote-5), Schroeder B., Gibson G.A.[[6]](#footnote-6), Li H. et. al.[[7]](#footnote-7)). Poza inspiracją dziełami innych autorów, dokonano także weryfikacji słuszności doboru rozkładów na podstawie informacji o zalecanym wykorzystaniu tychże rozkładów. Z informacji udostępnionych przez StatSoft[[8]](#footnote-8) wynika, że rozkład wykładniczy, podobnie jak rozkład Weibulla, jest często wykorzystywany w celu testowania czasu niezawodności, jednakże różnica tkwi w wskaźniku defektów. Gdy wskaźnik ten jest stały w określonym przedziale czasu to zaleca się rozkład wykładniczy, natomiast gdy jest on zmienny to właściwszym wyborem jest rozkład Weibulla. Przedstawione powyżej wnioski na temat użycia rozkładów potwierdzają słuszność wyboru rozkładów opartego na literaturze.

Celem pracy jest analiza rozkładu opisywanego zjawiska, a więc poza doborem rozkładów należy także zająć się oszacowaniem odpowiednich parametrów rozkładu. Aby tego dokonać wykorzystano Metodę Największej Wiarygodności (MNW). Pozwala ona na estymację parametrów rozkładu poprzez maksymalizację funkcji wiarygodności. Aby wykorzystać tę metodę należy spełnić pewne ograniczenia, jednakże są one mniej restrykcyjne niż w Metodzie Najmniejszych Kwadratów. W MNW funkcja wiarygodności jest przedstawiona jako iloczyn funkcji gęstości prawdopodobieństwa dla poszczególnych obserwacji (prób). W tej metodzie najczęściej maksymalizuje się logarytm funkcji wiarygodności, ponieważ przekształcenie logarytmiczne jest przekształceniem monotonicznym, co sprawia że jeśli istnieje maksimum logarytmu funkcji wiarygodności to znajduje się ono w takim samym punkcie co maksimum funkcji wiarygodności. Podczas znajdowania maksimum funkcji wiarygodności należy posłużyć się warunkami pierwszego i drugiego rzędu, co jest związane z wykorzystaniem wektora pierwszych pochodnych (gradientu) oraz macierzy drugich pochodnych (Hessianu). W celu przeprowadzenia maksymalizacji posłużono się metodą Netwona-Raphsona, która jest metodą rekurencyjną i zbiega szybko do poszukiwanego rozwiązania. Do testowania hipotez wykorzystano statystykę Z oraz statystykę ilorazu wiarygodności LR. Statystyka Z, użyta do weryfikacji hipotez prostych, jest opisana wzorem:

gdzie:

- oszacowanie parametru, - wartość hipotezy zerowej, – błąd standardowy oszacowania.   
Wadą statystyki ilorazu wiarygodności jest konieczność uzyskania oszacowań modelu z ograniczeniami oraz modelu bez ograniczeń. Statystykę LR wyliczono z poniższego wzoru:

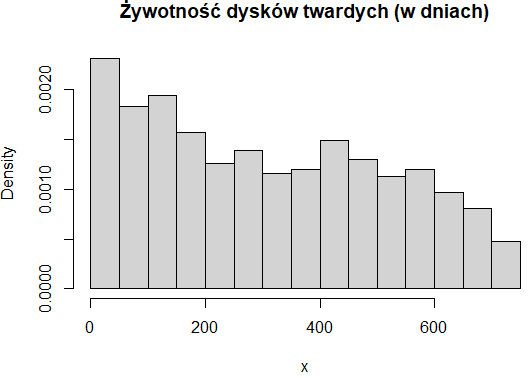
gdzie:  
 - estymator MNW bez ograniczeń, - estymator MNW z ograniczeniami. [[9]](#footnote-9) [[10]](#footnote-10)

**Wyniki**

Wykorzystanie metod opisanych w poprzednim rozdziale pozwoliło na uzyskanie wyników badania. W rozkładzie wykładniczym oszacowany parametr λ był równy 0.0032159, a w rozkładzie Weibulla oszacowania były następujące: *k* = 1.2965, λ = 333.0406. Dla obu rozkładów w celu uzyskania oszacowań posłużono się Metodą Największej Wiarygodności. Powyższe oszacowania wskazują, że używszy rozkładu wykładniczego do analizy badanego zjawiska otrzymano przeciętny czas żywotności dysku twardego równy 310,95 dni (wartość oczekiwana *=* 1/λ), natomiast analizując rozkład Weibulla można zauważyć, że oszacowanie parametru *k* wskazuje, że prawdopodobieństwo uszkodzenia dysku rośnie wraz z jego czasem działania (1 < *k* < 2) a oszacowanie parametru λ oznacza, że szansa, że dysk będzie działał 333.0406 dni wynosi w przybliżeniu 63,21%.[[11]](#footnote-11)

W pracy zweryfikowano wcześniej opisane hipotezy. Wyniki zostały przedstawione w tabeli 2. Z weryfikacji hipotez na poziomie istotności 0,05 wynika, że należy odrzucić wszystkie hipotezy zerowe. Oznacza to, że w rozkładzie wykładniczym wartość oczekiwana żywotności dysku nie jest równa 365 dniom. Analizując rozkład Weibulla stwierdzono, że dla badanego zjawiska nie można sprowadzić rozkładu Weibulla do rozkładu wykładniczego lub rozkładu Rayleigha oraz czas żywotności dysków twardych nie można opisać tymi samymi parametrami rozkładu co czas żywotności obrabiarek. Odrzucenie hipotez może być spowodowane szczególną charakterystyką badanego zjawiska, ponieważ dyski pracują nieprzerwanie oraz mają szerokie zastosowanie w nowych technologiach. Proces produkcji dysków twardych oraz materiały wykorzystywane w tym procesie różnią się od tych wykorzystywanych w produkcji obrabiarek co może mieć przełożenie na ich żywotność.

Rys 1. Histogram przedstawiający dane dotyczące czasu działania dysku do momentu wystąpienia awarii



*Źródło:* Opracowanie własne

Tabela 2. Wyniki weryfikacji hipotez na poziomie istotności równym 5%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rozkład** | **Hipoteza zerowa** | **Użyty test** | **Wartość statystyki** | **p-value** |
| Wykładniczy | λ = 1/365 | Z - test | 8,0615 | 0 |
| Weibulla | *k* = 1 | Z - test | 14,7063 | 0 |
| Weibulla | *k* = 2 | Z - test | -34,8909 | 0 |
| Weibulla | *k* = 34,05; λ = 945,25 | LR test | 282621,9547 | 0 |

*Źródło:* Opracowanie własne

**Wnioski**

Przeprowadzona analiza ukazała, że zarówno przy użyciu rozkładu wykładniczego jak i rozkładu Weibulla możliwe jest osiągnięcie przydatnych informacji na temat badanego zjawiska, aczkolwiek rozkład Weibulla w dokładniejszy sposób ukazuje jego charakterystykę. Odrzucenie hipotezy zakładającej *k* = 1 sugeruje, że w tym przypadku rozkład Weibulla nie może być sprowadzony do rozkładu wykładniczego, co informuje o prawdopodobieństwie wystąpienia awarii, które to nie jest stałe. Dodatkowo badanie pokazuje, że nie można wyciągać takich samych wniosków na podstawie analizy tego zjawiska w kontekście innych urządzeń. Badacze zainteresowani tym tematem powinni wziąć pod uwagę rozwój technologii, dyski SSD oraz wykorzystać bardziej zaawansowane metody.

**Bibliografia**

Li H., et. al., *Reliability Evaluation of NC Machine Tools considering Working Conditions*, Mathematical Problems in Engineering, wydanie 2016, luty 2016

Mycielski J., *Ekonometria*, Uniwersytet Warszawski. Wydział Nauk Ekonomicznych, Warszawa, 2010,s.245-260

NodeGraph, *How much data is on the internet? The Big Data Facts Update* 2020, 26.03.2020, https://www.nodegraph.se/how-much-data-is-on-the-internet/ Dostęp 29.12.2020

Rinne H., *The Weibull Distribution: A Handbook*, Chapman and Hall/CRC, 2008, edycja 1

Schroeder B., Gibson G. A., *Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?,* Parallel Data Laboratory, Carnegie Mellon University, wrzesień 2006

StatSoft, *Dopasowanie rozkładu,* https://www.statsoft.pl/textbook/stathome.html, Dostęp 30.12.2020

Woźniak R., Wykłady z Programowania Narzędzi Analitycznych, 2020

Yan J., Hua D., *Energy Consumption Modeling for Machine Tools after Preventive Maintenance*, IEEE, 2010

1. NodeGraph, *How much data is on the internet? The Big Data Facts Update* 2020, 26.03.2020, Dostęp 29.12.2020, https://www.nodegraph.se/how-much-data-is-on-the-internet/ [↑](#footnote-ref-1)
2. Rinne H., *The Weibull Distribution: A Handbook*, Chapman and Hall/CRC, 2008, edycja 1, s.47 [↑](#footnote-ref-2)
3. Yan J., Hua D., *Energy Consumption Modeling for Machine Tools after Preventive Maintenance*, IEEE, 2010 [↑](#footnote-ref-3)
4. https://www.kaggle.com/ezeydan/hard-drive-failure-data, Dostęp 27.12.2020 [↑](#footnote-ref-4)
5. Yan J., Hua D., op.cit. [↑](#footnote-ref-5)
6. Schroeder B., Gibson G. A., *Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?,* Parallel Data Laboratory, Carnegie Mellon University, wrzesień 2006 [↑](#footnote-ref-6)
7. Li H., et. al., *Reliability Evaluation of NC Machine Tools considering Working Conditions*, Mathematical Problems in Engineering, wydanie 2016, luty 2016 [↑](#footnote-ref-7)
8. StatSoft, *Dopasowanie rozkładu,* Dostęp 30.12.2020 https://www.statsoft.pl/textbook/stathome.html [↑](#footnote-ref-8)
9. Mycielski J., *Ekonometria*,Uniwersytet Warszawski, Wydział Nauk Ekonomicznych,Warszawa, 2010,s.245-260 [↑](#footnote-ref-9)
10. Woźniak R., Wykłady z Programowania Narzędzi Analitycznych, 2020 [↑](#footnote-ref-10)
11. Rinne H., op.cit., s.30-50 [↑](#footnote-ref-11)