H. Fischer, A. Endmann & M. Krökel (Hrsg.): Mensch und Computer 2015 – Usability Professionals, Berlin: De Gruyter, 2015, S. 401- 407.

Survival-Analyse für UX-Praktiker

Zeitdaten richtig auswerten und verstehen

Bernard Rummel
SAP SE
Dietmar Hopp-Allee 16
69190 Walldorf
bernard.rummel@sap.com

Abstract

Zeiten messen ist nicht schwer, aber was macht man dann damit? Dieses Tutorial gibt eine Einführung in grundlegende Konzepte der Survival-Analyse – statistische Methoden speziell für Zeitdaten, die in der UX-Szene noch wenig bekannt sind. Mit schnell erstellten sog. Probability Plots lassen sich Zeitdaten auf Auffälligkeiten untersuchen, technische Performance und UI-Effizienz getrennt betrachten, Rückschlüsse auf zugrundeliegende Prozesse ziehen und quantitative Modelle bestimmen, mit denen man Zeitdaten effizient beschreiben und vergleichen kann.

Keywords

Methoden, Effizienz, Bearbeitungszeit, Verweildauer, Survival-Analyse

Warum Survival-Analyse?

Die Survival-Analyse (Varianten sind als Zuverlässigkeits – oder Reliability-Analyse bekannt) wurde von Statistikern entwickelt, um herauszufinden, welche Faktoren die Überlebenszeit von Individuen (bzw. Bauteilen) beeinflussen. Ersetzt man Überlebenszeit durch Bearbeitungszeit, Verweildauer auf Webseiten oder Nutzungsdauer von Apps, wird es für UX Professionals sofort interessant. Die Mathematik ist tatsächlich unmittelbar übertragbar, wenn auch komplex. Ziel dieses Tutorials ist, einen verständlichen Einstieg in die Materie zu bieten, mit dem Praktiker einsteigen und Interessierte sich weiter einlesen können.

Für Usability-Praktiker sind in der Regel folgende Fragen relevant:

1. Wie kann ich valide Daten von Ausreißern unterscheiden?

2. Wie kann ich Bearbeitungszeiten bzw. Verweildauer auf Websites modellieren, um Vorhersagen und Vergleiche anzustellen?
3. Wie kann ich auf Ursachenfaktoren schließen, die zu der einen oder anderen Zeitverteilung geführt haben?

Wir beginnen die Diskussion mit der Frage 2 nach der Modellierung. Ausreißer lassen sich dann schnell als diejenigen Werte erkennen, die aus einer ersten Modellschätzung grob herausfallen, während Abweichungen von einem bestimmten Modell – z.B. der Exponentialverteilung – geeignet sind, Frage 3 zu beantworten.

Die Survival-Funktion

Betrachten wir eine Menge von N Individuen. Je nach Fragestellung können das Testteilnehmer, Befragungsteilnehmer, Besucher einer Website etc. sein. Uns interessiert jeweils ein bestimmtes Ereignis: ein Testteilnehmer löst die gestellte Aufgabe, ein Befragungsteilnehmer klickt auf „Abschicken“, ein Besucher verlässt die Webseite. Wir wollen die Zeit analysieren, zu der dieses Ereignis jeweils eintritt.

Abbildung 1 zeigt Daten aus einem Usability-Test. Die Balken geben die Zeit an, die Testteilnehmer zur Bearbeitung der Aufgabe gebraucht haben. Die Balken sind nach dieser Zeit sortiert.

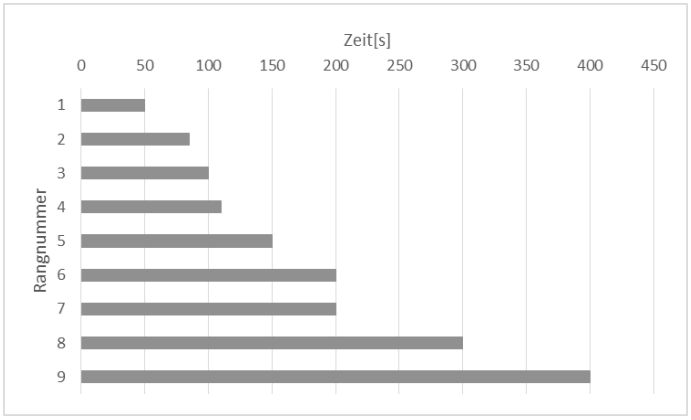

Abbildung 1: Bearbeitungszeiten in einem Usability-Test, aufsteigend sortiert.

Um abzuschätzen, welcher Anteil der Benutzer zu einer bestimmten Zeit noch an einer Aufgabe arbeitet bzw. auf der Webseite verbleibt, können wir für jeden Zeitpunkt den Anteil der „Überlebenden“ durch die Zahl der Teilnehmer dividieren: anfangs „leben“ 100%, später weniger, am Ende 0%. Das ist die Survival-Funktion S(t), in der Zuverlässigkeitsanalyse auch „Reliability Function“ R(t) genannt. Das Komplement 1-S wird auch als

Ausfallwahrscheinlichkeit (probability of failure) F bezeichnet – in unserem Fall bedeutet F die Wahrscheinlichkeit, dass ein Testteilnehmer die Aufgabe löst bzw. die Webseite verlässt.

Abbildung 1 zeigt eine *empirische* Survival-Funktion, die wir tatsächlich beobachtet haben. Natürlich wollen wir aus dieser Beobachtung auf die Gesamtheit aller Benutzer schließen, und das Ganze in einer übersichtlichen Formel darstellen.

Schritt 1: von der Rangnummer zur Survival-Statistik

Die Rangnummer der sortierten Zeiten in Abbildung 1 gibt als Zählindex nur ungenau an, welchem Prozentsatz in der Gesamtstichprobe sie entspricht. Dazu gibt es eine einfache Korrekturformel, bekannt als Bénard- oder Median Rank - Näherung (NIST 2012):

$$(1)\quad S_i = 1 - [(i-0.3)/(N+0.4)]$$

wobei i die Rangnummer des Individuums und N die Anzahl Individuen in der Stichprobe bezeichnet; S_i ist dann der Prozentsatz, den wir der entsprechenden Zeit zuweisen können. Die Bénard-Näherung ist immer dann verwendbar, wenn alle Individuen das betreffende Ereignis auch tatsächlich „erlebt“ haben, d.h. alle Testteilnehmer die Aufgabe lösten bzw. Besucher einer Website diese auch zu einer bekannten Zeit wieder verlassen haben.

Das ist nun nicht immer der Fall. In der Survival-Analyse spricht man von „censoring“, wenn nicht alle interessierenden Daten zur Verfügung stehen, z.B. weil Individuen bis zum Studienende nicht verstorben sind (bzw. die Aufgabe lösten), oder nur zu fixen Zeitpunkten Daten erhoben wurden. Für verschiedene Fälle von censoring gibt es spezifische Schätzmethoden, eine davon (modified Kaplan-Meier Product Limit) ist in Rummel (2014) für Usability-Testdaten ausführlich beschrieben. Hier soll der Hinweis genügen, dass man mit dieser Methode auch Daten aus Tests analysieren kann, bei denen nicht alle Teilnehmer die gestellte Aufgabe lösten. Ebenso lassen sich Verweilzeiten auf Webseiten auswerten, auch wenn der Zeitpunkt des Verlassens nicht exakt bekannt ist (Rummel, 2015).

Schritt 2: Schätzung der Survival-Funktion

Nun wollen wir den Zusammenhang zwischen den geschätzten Prozentwerten und den beobachteten Zeiten modellieren. Offensichtlich hängt dieser von der statistischen Verteilung der Zeiten ab hat. Wir können jedoch mit einer einfachen Annahme schon weitreichende Aussagen machen: nehmen wir an, die beobachteten Zeiten seien ausschließlich vom Zufall abhängig.

Reine Zufallsprozesse führen mathematisch gesetzmäßig zu exponentialverteilten Zeiten; ein Beispiel ist radioaktiver Zerfall. Charakteristisches Merkmal dieser Prozesse ist, dass in gleichen Zeitintervallen ein konstanter Anteil der betrachteten „Teile“ „zerfällt“. Das vereinfacht die Schätzung der Survival-Funktion erheblich: wenn wir die S-Achse logarithmisch auftragen, müssten exponentialverteilte Zeiten als Gerade erscheinen!

Abbildung 2 zeigt die Daten aus Abbildung 1 in einem solchen sogenannten Probability Plot (NIST, 2012). Die Zeitachse ist linear, die vertikale (S-)Achse logarithmisch skaliert. Tatsächlich erscheinen die Datenpunkte entlang einer Geraden, die hier als

Regressionsgerade mit entsprechender Regressionsgleichung eingetragen ist. Der Parameter R^2 gibt an, dass dieses lineare Modell 98% der beobachteten Varianz erklärt – die Modellanpassung ist also sehr gut, wir haben eine Exponentialverteilung vor uns.

Mit diesem Modell können wir nun für beliebige Zeiten angeben, welchen Prozentsatz an Benutzern wir erwarten können, der die Aufgabe innerhalb dieser Zeit löst. Umgekehrt können wir die Zeiten bestimmen, zu denen wir eine beliebig gegebene Erfolgsquote erwarten.

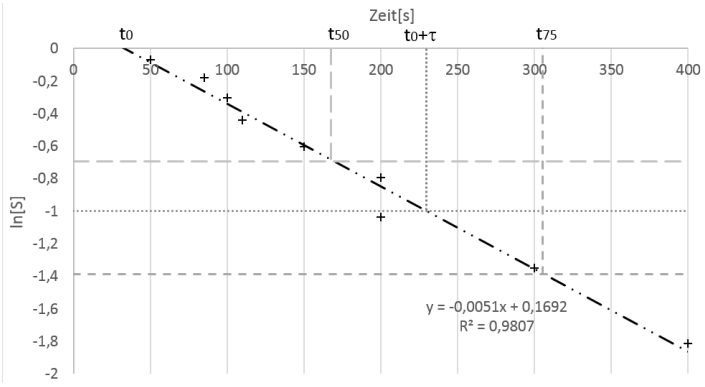

Abbildung 2: Probability Plot der Daten aus Abbildung 1. Die drei horizontalen Linien entsprechen S=0,5 bzw. 0,37 und 0,25 (50% bzw. 63% und 75% der Teilnehmer haben die Aufgabe gelöst). Weitere Erläuterungen im Text

„Normal"-Verteilung für Zeiten: die Exponentialverteilung

Abbildung 2 zeigt einige, für usability professionals unmittelbar bedeutsame Merkmale einer solchen, typischen Verteilung von Bearbeitungszeiten. Die Regressionsgerade schneidet die Zeitachse nicht im Ursprung, sondern bei einer Zeit t_0: erst ab dieser Zeit (hier 33s) können wir erwarten, dass überhaupt jemand die Aufgabe löst. Der eigentliche Zufallsprozess ist dieser konstanten Zeit überlagert.

Das ist plausibel: in einem Usability-Test hat man typischerweise eine mehr oder weniger konstante Systemreaktionszeit. Weiterhin brauchen Benutzer eine gewisse Zeit allein zum Durchklicken des Lösungspfades – diese Zeit variiert zwischen Testteilnehmern, aber nur um wenige Sekunden. t_0 beschreibt den konstanten Anteil an der Verteilung der Zeiten, ist also vor allem durch diese „mechanischen" Anteile an der Gesamtzeit bestimmt.

Die restliche Zeit verbrachten die Testteilnehmer mit der Suche nach Funktionen, Fehlern und Korrigieren derselben etc. – es ist schon bemerkenswert, dass ein Zufallsmodell diesen Zeitanteil so gut abbildet. Dieser für die Gebrauchstauglichkeit entscheidende Zeitanteil wird vollständig durch die Steigung der Regressionsgeraden beschrieben! Der Kehrwert dieser Steigung hat die Dimension einer Zeit, nennen wir ihn τ (hier 196s). Zur Zeit $t_0+\tau$ lösten ca.

63% die Aufgabe; von den verbleibenden brauchten wiederum 63% zusätzlich die Zeit τ, usw.

Wo investieren?

Ein Plot wie Abbildung 2 gibt Hinweise darauf, wie Verbesserungen der Systemeffizienz zu bewerten sind. Investiert ein Kunde in schnellere Hardware, wird das t_0 beeinflussen, aber nicht τ - die Gerade in Abbildung 2 würde im Wesentlichen ein wenig nach links verschoben. Eine Reduktion von t_0 um die Hälfte würde gerade einmal 16,5s bringen.

Anders eine Investition in verbessertes Interaktionsdesign: schafft man es, τ um die Hälfte zu reduzieren, gewinnt man nicht nur deutlich mehr Zeit insgesamt, sondern reduziert vor allem den Anteil an Individuen, die besonders lange im Prozess verbleiben – in den meisten Geschäftsanwendungen ist dies der entscheidende Faktor. Gelingt es, τ substanziell zu verbessern, darf dies sogar auf Kosten von Systemreaktionszeit und click count gehen – der Nettogewinn ließe sich mit unserem Modell sogar quantifizieren.

Alles nur Zufall?

Probability Plots nutzen die mathematischen Eigenschaften statistischer Verteilungen, indem die Achsen des Plots genau so eingestellt werden, dass die Datenpunkte als gerade Linie erscheinen, sofern die entsprechende Verteilung vorliegt. Man kann anhand entsprechender Plots daher sofort sehen, welche Verteilungsmodelle infrage kommen (Rummel, 2014); detailliertere Analysen können daran anschließen.

Man kann jedoch auch ohne komplizierte Verteilungsanalyse pragmatische Aussagen machen. Wenn reine Zufallsprozesse exponentialverteilte Zeiten erzeugen, deutet jede Abweichung vom Exponentialverteilungsmodell auf einen nicht-zufälligen Einfluss hin – zumindest in Form einer Hypothese, die der Überprüfung wert ist.

Abbildung 3 zeigt vier Formen der Abweichung von der Exponentialverteilungs-„Ideallinie“:

1. Unsystematische Schwankungen um die Regressionsgerade können ignoriert werden.
2. Auffällig kurze Zeiten in der linken oberen Ecke des Plots deuten darauf hin, dass einige Individuen an dem betrachteten Prozess gar nicht teilgenommen haben – Test- oder Befragungsteilnehmer könnten geschummelt haben, irrtümliche Besucher einer Website könnten ihren Irrtum sofort gesehen und zurücknavigiert haben – oder gar keine Besucher sein, sondern bots.
3. Systematische Abweichungen von der Geraden nach links zeigen, dass Individuen kürzere Zeit in dem Prozess verbracht haben als erwartet. In einem Test können Lerneffekte aufgetreten sein, bzw. der Inhalt einer Webseite war so langweilig, dass er Leser eher abgeschreckt hat, oder eine App erwies sich schnell als nutzlos.

4. Systematische Abweichungen von der Geraden nach rechts zeigen, dass Individuen längere Zeit in dem Prozess verbracht haben als erwartet – die Webseite enthielt „fesselnden“ Content, Testteilnehmer waren von Usability-Problemen so zermürbt, dass ihre Leistungsfähigkeit nachließ, etc.

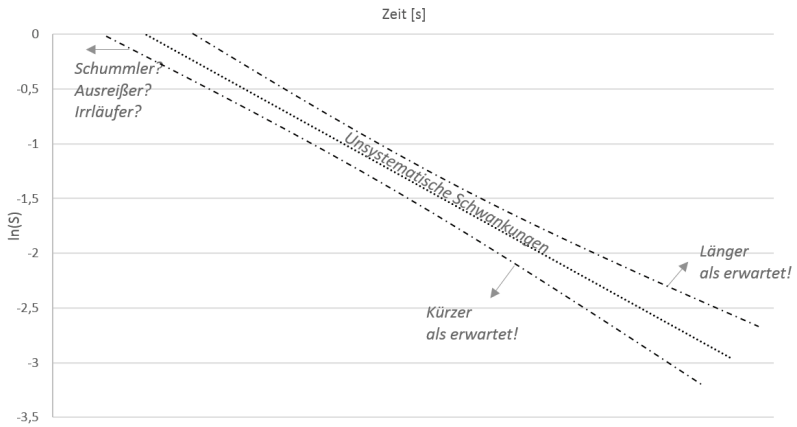

Abbildung 3: Interpretation von Abweichungen vom Exponentialverteilungsmodell

Selbstverständlich kann es auch vorkommen, dass die beobachteten Daten einer anderen als der Exponentialverteilung folgen. Eine dahingehende Analyse lohnt sich immer dann, wenn die Datenpunkte im Plot einer deutlich gekrümmten Linie folgen. Rummel (2014) beschreibt Analysemöglichkeiten für derartige Fälle.

Verweildauer auf Websites und Nutzungsdauer von Apps sind solche Fälle, die typischerweise nicht einer Exponentialverteilung folgen, sondern einer Weibull-Verteilung. Das Weibull-Verteilungsmodell ist eine Verallgemeinerung des Exponentialverteilungsmodells, das systematische Veränderungen in der „Halbwertszeit“ beschreibt. Nimmt diese zu, ist das z.B. für eine Blogging-Website durchaus erwünscht: man will ja gerade, dass Besucher sich „festlesen“ und mit zunehmender Besuchsdauer umso länger bleiben. Inwieweit dies systematisch geschieht, kann man mittels einer Weibull-Analyse (Liu et al., 2010) quantifizieren; eine kurze Einführung in das Verfahren gibt Rummel (2015).

Mittelwert Ade?

Wir haben bereits gesehen, dass sich mit den Parametern t_0 und τ Aussagen treffen lassen, die deutlich über den Informationsgehalt eines Mittelwerts hinausgehen. Ebenso informative Parameter gibt es für andere gängige Zeitverteilungen, wie die Weibull- und Lognormalverteilung (Liu et al., 2010; Rummel, 2014, 2015).

Haben andere Parameter nun ausgedient? Das *Common Industry Format* (CIF) nach ISO 25062 (2006) fordert Mittelwert, Range und Standardabweichung, die *Summative Test*

Method nach ISO 20282 (2013) geometrisches Mittel bzw. Median und schließt sich damit einer Forderung von Sauro (2011) an. Damit sind alle gängigen Parameter im Rennen – welche sollte man nun verwenden, und was kann man daraus lesen?

In einer reinen, nicht verschobenen Exponentialverteilung wären Standardabweichung und Mittelwert mathematisch gleich τ. Da die Verschiebung nur den Mittelwert, nicht aber die Standardabweichung betrifft, wäre die Differenz von beobachtetem Mittelwert und Standardabweichung die „lösungsfreie Zeit" t_0.

Leider ist t_0 statistisch nicht leicht zu bestimmen; schon geringe systematische Abweichungen von der Exponentialverteilung können den numerischen Wert stark verzerren. Eine direkte Messung von t_0 als Systemreaktions- und „Durchklickzeit" ist in der Praxis dagegen meist problemlos durchführbar; ansonsten gibt der Minimalwert der beobachteten Zeiten in der Regel einen guten ersten Anhaltswert.

Der Range der beobachteten Zeiten, also Minimum und Maximum, erlaubt eine weitere Abschätzung von τ. Ist die Teilnehmerzahl bekannt, lassen sich aus der Bénard-Formel (1) die ln(S)-Werte des schnellsten und langsamsten Testteilnehmers ermitteln. Der Range der Zeiten geteilt durch diese Differenz ist ein Schätzwert für τ.

Im Fall von nicht exponentialverteilten Zeiten (ca. 15-20% in Usability Tests) müssen unter Umständen komplexere Verteilungsmodelle geschätzt werden. Im Fall von Lognormalverteilungen – ebenfalls eine sehr häufig beobachtete Verteilungsform in Usability-Testdaten – sind Mittelwert und Standardabweichung der logarithmierten Zeiten aussagekräftige Verteilungsmaße; ersterer entspricht dem geometrischen Mittel.

Literatur

ISO/IEC 25062 (2006). *Software engineering–Software product quality requirements and evaluation (SQuaRE)–Common Industry Format (CIF) for usability test reports*. Berlin: Beuth

ISO/TS 20282-2 (2013). *Usability of consumer products and products for public use — Part 2: Summative test method*. Berlin: Beuth

Liu, C., White, R. W., & Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 379-386. New York: ACM.

NIST/SEMATECH (2012): Probability Plotting. In: *E-handbook of statistical methods*. National Institute of Standards and Technology. Zugriff April 2015 von http://www.itl.nist.gov/div898/handbook/apr/section2/apr221.htm

Rummel, B. (2014): Probability Plotting: A Tool for Analyzing Task Completion Times. *Journal of Usability Studies*, 9(4), 152-172

Rummel, B. (2015): Does Your Website Make People Want To Stay? SAP SE, Experience.sap.com. Zugriff Juni 2015 von https://experience.sap.com/skillup/does-your-website-make-people-want-to-stay/

Sauro, J. (2011). 10 things to know about task times. *Measuring Usability*. Zugriff Dezember 2013 von http://www.measuringusability.com/blog/task-times.php