

Analyse der Benutzerdaten des Projekts “Verlustlisten des 1. WK” im CompGen

Michael Rose*

10. Februar 2014

1 Motivation

Zwischen dem Start des Projekts im Dezember 2011 und dem 17. Oktober 2013 machten 603 angemeldete Erfasser 5.759.202 Beiträge. Dies ist lediglich die Zahl derer, die mindestens einen Beitrag leisteten (Wer sich angemeldet hat und nie erfasst hat schaffte es nicht in den untersuchten Datensatz).

Ursprüngliche Anlaß dieser Untersuchung war und bleibt die Frage, warum die Beiträge pro Benutzer sinken. Die vom Projekt bereitgestellten Graphen¹ zeigen nämlich ein lineares Wachstum der Beiträge, bei linearem Wachstum der Benutzerzahlen (zumindest ab Februar 2012). Bei wachsenden Benutzerzahlen wäre jedoch ein nicht-lineares (exponentielles?) Wachstum der Beiträge zu erwarten. Tatsächlich aber sinkt offensichtlich der durchschnittliche Beitrag pro Benutzer im Zeitablauf. Dafür gibt es drei Erklärungen: Jeder Erfasser tippt weniger ab aber sie werden konstant mehr oder die Erfasser tippen gleich viel aber es steigen immer wieder welche aus, oder eine Mischung aus beidem: Jeder Erfasser tippt konstant weniger ab und Erfasser verlassen das Projekt (werden inaktiv).

Eine andere Intention war es, charakteristische Verhaltensmuster zu finden: um früh zu erkennen, ob jemand bald wieder inaktiv wird, ob sich jemand für das Administratorenamt qualifiziert, etc. Außerdem ist es immer mal gut, zu wissen, wie die Erfasser arbeiten.

Die dritte Intention war die Vorbereitung einer anonymen Umfrage unter den Erfassern. Die hier gewonnen Erkenntnisse dienen der Hypothesenbildung dieser Umfrage.

*Für Unstimmigkeiten bitte ich um Entschuldigung. Longitudinale Daten sind sonst nicht Hauptbetätigungsfeld.

¹<http://des.genealogy.net/eingabe-verlustlisten/statistics/index>

2 Umfang

Von der Analyse habe ich die Beiträge der Administratoren Marie-Luise, Jesper, Uwe, Susanne und Nick ausgenommen (ich war bis Dezember 2013 kein Admin in den Verlustlisten). Die Administratoren kommen zusammen auf 691.481 Beiträge, was 12.01% der Gesamtbeiträge entspricht.

Der Grund für die Selektion ist ein divergierendes Erfassungsmuster: Wenn ein Nicht-Admin 2 Stunden für das Projekt arbeiten möchte, so erfasst er die gesamte Zeit über; ein Admin dagegen teilt sein exogen gegebenes Zeitbudget auf die Tipparbeit und die administrative Arbeit auf. Besser wäre natürlich, die administrative Arbeit irgendwie hinzuzurechnen. Ebenso müsste man die Fehlersuche einiger Erfasser hinzurechnen. Leider geht aus dem Datensatz nicht hervor, ob jemand Fehler sucht und meldet. Ich nehme an, dass diese für das Projekt ebenso unverzichtbaren Erfasser ihr Zeitbudget zwischen Erfassung und Fehlersuche ähnlich den Admins aufteilen und deshalb in dieser Analyse unterbewertet werden.

Es blieben also übrig: 599 Erfasser und 5.067.275 Beiträge samt ihren Erfassungszeiten, Seiten und Fehlerstatus zurückgehen.²

Ich habe Erfasser drei Gruppen eingeteilt, auf deren typisches Verhalten ich abziele:

1. **Aktive:** Alle Personen, die in den zwei Wochen vor dem Stichtag noch einen Beiträge leisteten und mindestens zwei Tage mitmachten (106).
2. **Eintagsfliegen**³: Alle Personen, die sämtliche Beiträge innerhalb von 24 Stunden leisteten (121), sofern es nicht am Stichtag war. Eintagsfliegen erfassten insgesamt 9624 Einträge.
3. **Inaktive:** Alle übrigen Personen (= Personen, deren letzter Beitrag mind. zwei Wochen vor dem Stichtag war und mindestens an zwei Tage erfassten; 372). Die Zahl mag etwas niedriger sein, denn bis zum Ende des Projekts kann niemand sicher sagen, ob nun jemand wirklich inaktiv ist oder vielleicht in vier Wochen wiederkommt. Tatsächlich sind die 50 aktivsten Inaktiven während ihrer als aktiv gemessenen Zeit sporadisch aktiv gewesen, denn zwischen ihren Bearbeiten können ruhig mal Monate liegen.

Zunächst erkläre ich die Variablen, die ich danach tabellarisch und grafisch beschreibe. Anschließend nehme ich die Entwicklung der Benutzer genauer unter die Lupe. Es folgen danach Hypothesentests für bestimmte Aussage und ein kleiner Vergleich zwischen den produktivsten Aktiven und den produktivsten Inaktiven. Wer nicht alles durchlesen möchte, springt gleich zum zusammenfassenden Abschnitt 8.

3 Variablen

- **Dauer:** Zeitspanne zwischen dem ersten und der letzten Beitrag.

²Analyse, die in Zusammenhang mit den Seiten stehen, habe ich nicht untersucht.

³Ich entschuldige mich für den Ausdruck; ein schöneres Wort fiel mir nicht ein.

Datensatz erfasst. Dieser jemand zählt per Definition als Aktiver, was nicht stimmen muss.

Das durchschnittliche Alter beträgt 495 Tage, das Medianalter bei 595 Tagen. Der Median ist diejenige Merkmalsausprägung, die alle andern Ausprägungen in der Mitte teilt: 50% der Ausprägungen liegen davor, 50% liegen dahinter. Im konkreten Fall heißt das, dass 50% aller untersuchten Erfasser ihren ersten Beitrag vor dem 1. März 2012 tätigten. Die Inaktiven haben sich im Durchschnitt 531 Tage vor dem Stichtag (4. Mai 2012) zum ersten mal an DES versucht - sie sind also im Schnitt früher eingestiegen. Auch der Vergleich des Medianalters von 621 schlägt in diese Kerbe.

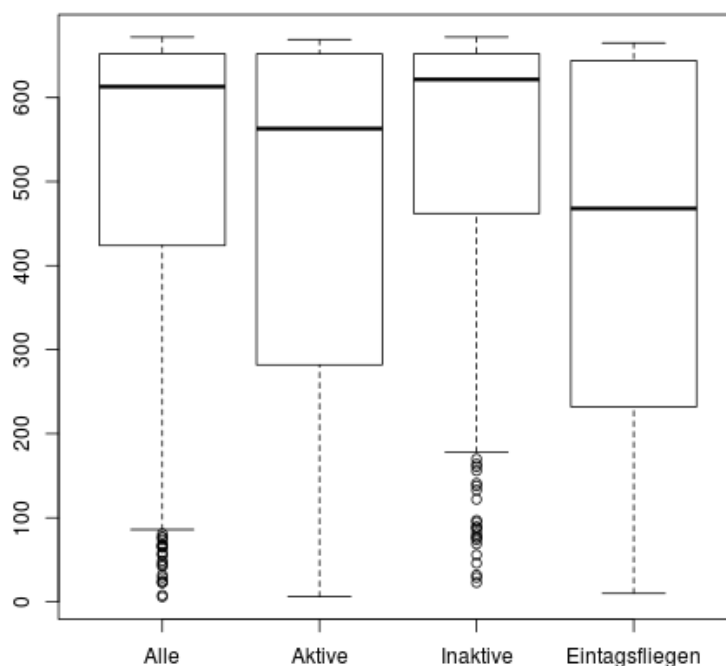


Abbildung 1: Kastengrafik des Alters aller Erfasser (außer Admins), nach Benutzergruppe

Die Abbildung 1 zeigt sehr schön, dass der Beitritt neuer Erfasser alles andere als gleichmäßig ist. Im Gegenteil, fast alle sind sehr früh eingestiegen.

4.2 Dauer

Dauer misst die Zeit, die jemand dabei war oder ist, unabhängig davon, wie aktiv er in der Zeit war. D.h., wer vom ersten bis zum letzten Tag der Analyse durchgearbeitet hat, besitzt dieselbe Dauer wie jemand, der nur an diesen beiden Tagen einen Datensatz erfasste.

Tabelle 2: Übersicht: Dauer

Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Alle	0.0	4.0	52.0	187.6	363.5	664.0
Aktive	1.0	283.0	552.5	447.3	647.0	664.0
Inaktive	2.0	15.0	72.5	174.5	293.8	641.0
Eintagsfliegen	0.0	0.0	0.0	0.2	0.0	1.0

Die Inaktiven waren im Durchschnitt nur 175 Tage tätig. Die Aktiven dagegen sind im Schnitt 447 Tage dabei Sitzungen. Die Schlußfolgerung: Wer ein gewisse Zeit dabei ist, der bleibt es wohl auch noch weiterhin noch ein Weilchen. Das wird weiter unten in Abschnitt 6 noch mal bestätigt.

Eine technische Randnotiz zu dem Maximum in der Zeile zu den Aktiven: Der als inaktiv definierte Erfasser, der seine Beiträge innerhalb von 641 tätigte, machte seinen letzten Beitrag am 18. September 2013. Also gut zwei Wochen vor dem Stichtag, ab dem ein Erfasser als aktiv gilt. Die Definition der Aktivität ist also relativ robust.

Abbildung 2 zeigt eine starke Umverteilung zwischen Aktiven und Inaktiven: Wer als aktiv zählt, ist sehr lange dabei, wer als inaktiv zeigt, war dagegen nicht so lange dabei. Man kann mutmaßen, dass die Wahrscheinlichkeit eines Ausstiegs sinkt je früher der Erfasser hinzukam. Diese Vermutung wird weiter unten in Abschnitt 5 bestätigt.

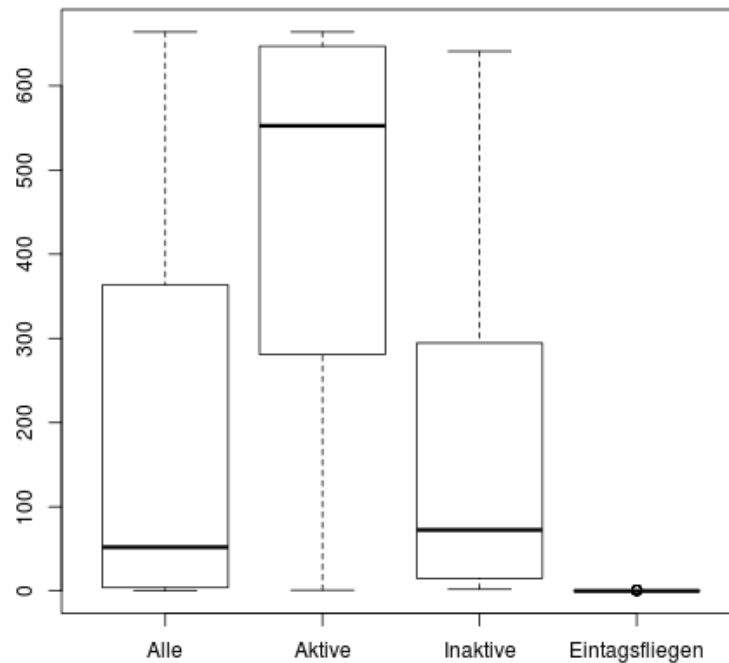


Abbildung 2: Kastengrafik des Dauer aller Erfasser (außer Admins), nach Benutzergruppe

Die Kastengrafik zu den Eintagsfliegen ganz rechts ähnelt naturgemäß einem Strich, denn sie waren höchstens einen Tag dabei.

4.3 Produktivität

Produktivität gibt die Anzahl der Beiträge normiert durch die Dauer wieder.

Tabelle 3: Übersicht: Produktivität

Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Alle	0	7	21	55.43	57	731
Aktive	0	10	31	72.00	92	498
Inaktive	0	7	20	42.87	43	731
Eintagsfliegen	1	5	20	79.54	81	568

Der höchste gemessene Produktivitätswert hat paradoxerweise ein als inaktiv definierter Erfasser. Aber auch die Eintagsfliegen sind stark dabei: Bis zu 568 Beiträge an einem Tag.

In der Abbildung 3 sind das interessanteste die vielen Ausreißer. Ausreißer sind absolut untypisch für ihre Gruppe. Je mehr Ausreißer, desto ungleicher arbeiten Gruppenmitglieder. Erstaunlicherweise sind die Aktiven noch die Homogensten. Das erlaubt die relativ präzise Vorhersage: Wer in den ersten Tagen seines Daseins als Erfasser konstant viel beiträgt, wird eine Weile dabei bleiben.

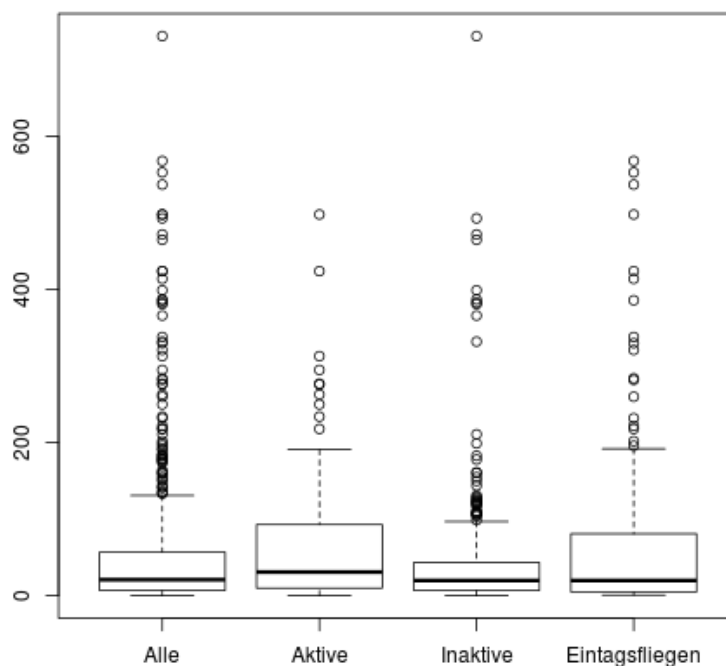


Abbildung 3: Kastengrafik der Produktivität aller Erfasser (außer Admins), nach Benutzergruppe

Die zahlreichen Ausreißer bei den Eintagsfliegen machen mich skeptisch. Wie kann jemand, der so viel auf einmal erfasst, so schnell die Lust verlieren? Vermutlich sind darunter Personen mit mehreren Konten, die sich also einfach mehrfach anmeldeten.

4.4 Fehlerquote

Ein Eintrag, der mindestens einmal als falsch markiert wurde, gilt in der Analyse als Fehler. Alle Fehler eines Erfasser dividiert durch seine gesamten Beiträge ergibt die Fehlerquote.

Es scheint, als wenn unter den Beiträgen der Aktiven tatsächlich weniger Fehler gefunden werden als unter den Inaktiven. Den Zusammenhang zwischen Fehlerquote und Beiträgen untersuche ich weiter unten in Abschnitt 6 ausführlicher.

Tabelle 4: Übersicht: Fehlerquote

Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Alle	0.0000	0.0220	0.0450	0.1182	0.0995	1.0000
Aktive	0.0000	0.0113	0.0220	0.0356	0.0385	0.3760
Inaktive	0.0000	0.0260	0.0460	0.0860	0.0875	1.0000
Eintagsfliegen	0.0000	0.0310	0.1390	0.2892	0.4200	1.0000

Dass die Gruppe der Eintagsfliegen eine relativ hohe Fehlerquote haben, verwundert nicht. Sie machen wenige Einträge und haben keine Erfahrung. Es gab Eintagsfliegen, die sieben von sieben Einträge falsch erfasst haben. Das zeigt auch die sehr lange obere Antenne in der rechten Kastengrafik von Abbildung 4:

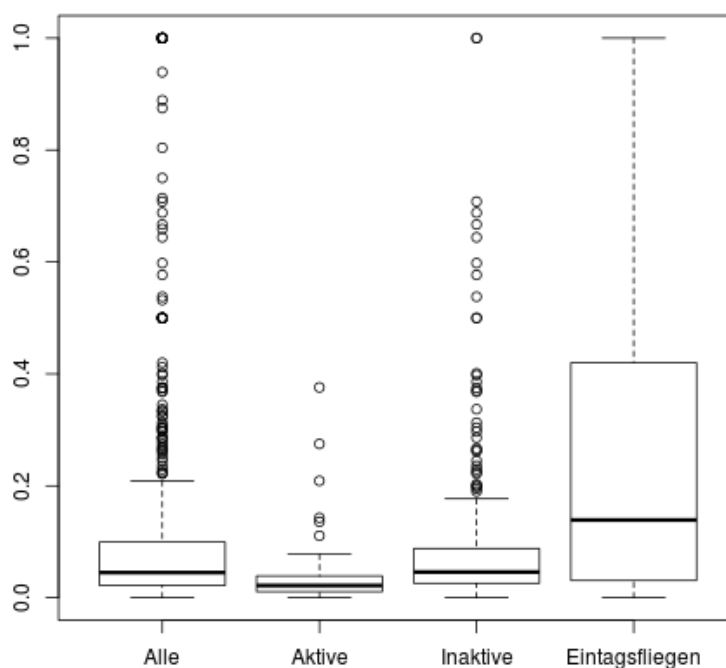


Abbildung 4: Kastengrafik der Fehlerquote aller Erfasser (außer Admins), nach Benutzergruppe

4.5 Sitzungen

Interessant hier ist die hohe Zahl der Sitzungen bei den Eintagsfliegen: Innerhalb eines Tages 11 Sitzungen. Der Wert zum 3. Quartil zeigt aber, dass 75% höchstens 2 Sitzungen

Tabelle 5: Übersicht: Sitzungen

Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Alle	1.00	3.00	14.00	85.55	69.50	1945.00
Aktive	2.0	52.5	200.5	302.6	422.0	1945.0
Inaktive	2.00	6.00	17.00	50.99	55.25	847.00
Eintagsfliegen	1.000	1.000	1.000	1.669	2.000	11.000

hatten, was wiederum völlig verständlich ist.

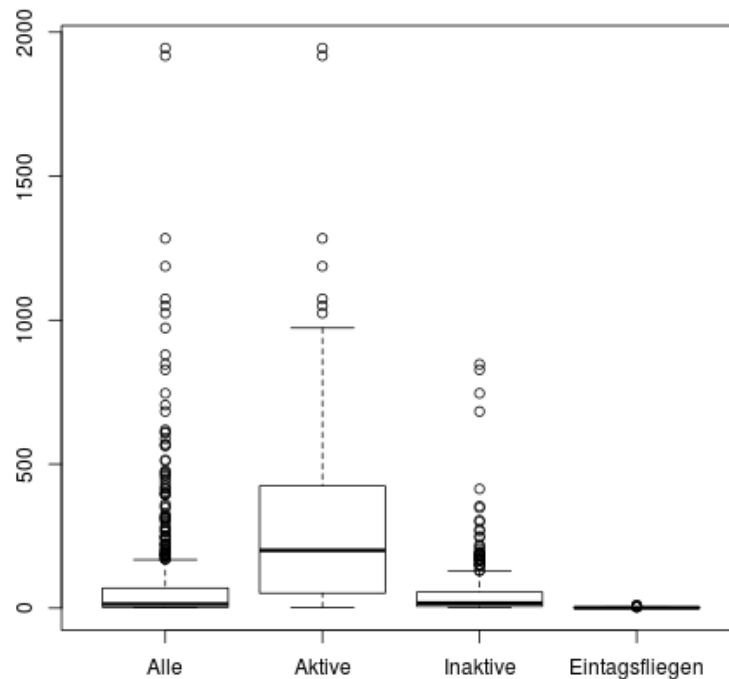


Abbildung 5: Kastengrafik der Sitzungen aller Erfasser (außer Admins), nach Benutzergruppe

5 Entwicklung der Eintritte

Es lohnt ein gesonderter Blick auf die Entwicklung der Eintritte im Zeitablauf.

Die Tabelle 6 vergleicht die Einstiege und Ausstiege in das Projekt in den 23 Monaten, für die Daten vorliegen. Die letzte Spalte gibt den Saldo dieser beiden Werte an. Leider sind im Jahr 2013 bis zum 17. Oktober mehr Erfasser ausgestiegen also hinzukamen.

Tabelle 6: Einsteiger und Aussteiger pro Jahr

Jahr	Einsteiger	Aussteiger	Saldo
Dez. 2011	89	3	+86
2012	396	203	+325
2012 (bis Oktober)	357	180	+177
2013 (bis Oktober)	114	189	-75

Korreakterweise müsste man zu den Aussteigern noch die Eintagsfliegen hinzuzählen, die ja nur besondere Aussteiger sind. Nach dieser Lesart wäre das Jahr 2013 insgesamt das schlechteste Jahr was die Erfasserentwicklung angeht. Daher in der Tabelle 7 ein genauerer Blick auf die Eintagsfliegen. Die Tabelle zeigt pro Jahr die Anzahl der neu eingestiegenen Erfasser, sowie die Anzahl und den Anteil der Eintagsfliegen unter ihnen:

Tabelle 7: Einsteiger und Eintagsfliegen pro Jahr

Jahr	Anzahl der Einsteiger	Eintagsfliegen	Anteil
Dez. 2011	89	11	12%
2012	396	71	18%
2012 (bis Oktober)	357	62	17%
2013 (bis Oktober)	114	40	35%

Wir sehen im Vergleich der Zahlen nicht nur, dass 2013 interessanterweise weniger Leute hängen geblieben sind (größerer Anteil Eintagsfliegen), sondern auch, dass es von 2012 zu 2013 insgesamt etwa $(357-114)/357 = 68\%$ weniger Neuversuche gab. Eher keine gute Entwicklung. Es kann auch sein, dass alle, die sich für das Projekt begeistern, einfach schon im Jahr 2012 eingestiegen sind (unter der Annahme, dass der Pool der potentiellen Erfasser konstant ist).

Die folgende Tabelle 8 zeigt, in welchem Jahr die späteren Inaktiven eingestiegen sind (Eintagsfliegen wurden nicht mitgezählt):

Tabelle 8: Eintrittsjahr der späteren Inaktiven

Jahr	Spätere Aussteiger	Einsteiger	Anteil
Dez. 2011	59	89	66%
2012	265	396	67%
2012 (bis Oktober)	243	357	68%
2013 (bis Oktober)	70	114	61%

Die letzte Spalte gibt den Anteil der späteren Inaktiven (=Aussteiger) in ihrer Einstiegskohorte an. Die Unterschiede sind nicht groß und wahrscheinlich nicht signifikant (hab ich nicht überprüft). Mit 65% ist der Anteil späterer Aussteiger dagegen sehr hoch und sollte uns zu denken geben, warum so viele Erfasser das Projekt wieder verlassen, obwohl sie einige Tage dabei waren.

6 Hypothesentest

Ich habe eine Reihe von Hypothesen im Laufe der Zeit formuliert. Wenn nötig, habe ich dazu einen Test ausgeführt:

H1: Es gibt etwa 20% der Bearbeiter, die 80% der Bearbeitungen erledigen (Pareto-Regel).

Leichte Abweichung: 20% der Bearbeiter erledigten zum Stichtag 85% der Beiträge. Allerdings ist die Verteilung stark linkssteil/rechtsschief: Der Erfasser mit den meisten Bearbeitungen kommt auf 6,3%, der zweite auf knapp 5%.

Nehme ich nun die Produktivität als Maß, so ist der höchste Wert 498, der niedrigste 0. Der Mittelwert liegt bei 49, der Median bei 21. Die Verteilung ist augenscheinlich linkssteil/rechtsschief, wie auch die folgende Abbildung 6 nahelegt:

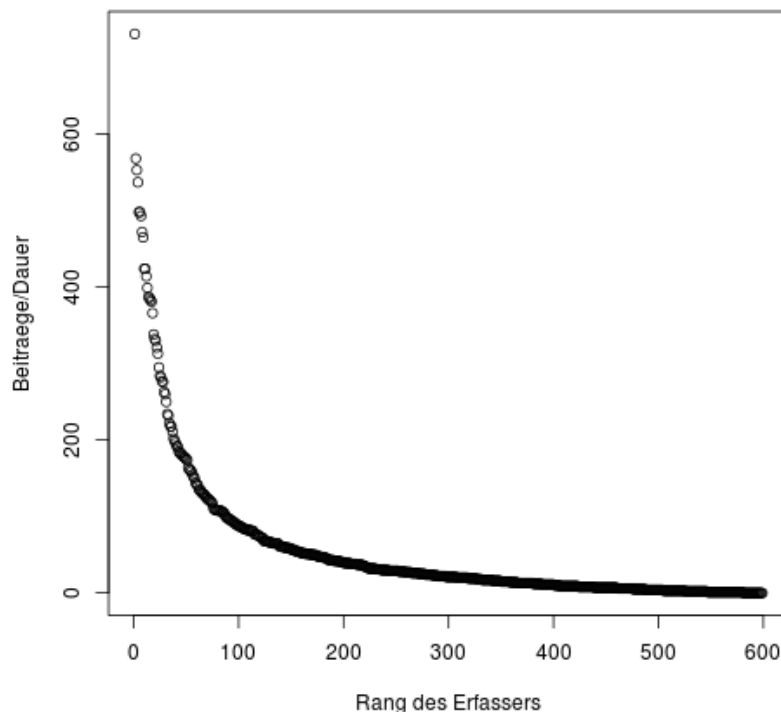


Abbildung 6: Verteilung der Produktivität, **H1**

Ein Kolmogorov-Smirnov-Test auf Gleichverteilung bestätigt mit einem P-Wert $< 2.2e - 16$ die extreme Ungleichverteilung.

H2a: Abends werden mehr Beiträge geleistet.

Für diese und die folgenden beiden Hypothese habe ich die Beiträge der Admins wieder mit einbezogen. Das lohnt, um Belastungszeiten für Server, Fehlersucher und administrative Aufgaben zu ermitteln.

Die folgende Häufigkeitstabelle zeigt, in welchen Stunden des Tages die 5.759.218 untersuchten Einträge erstellt worden sind. In der ersten Zeile die Stunde, darunter die absoluten Beiträge in dieser Stunde, in dritter Reihe die relativen Werte.

00	01	02	03	04	05	06	07	08	09	10
182223	90032	37123	19204	17576	14894	41079	106087	146229	201168	253045
3.16%	1.56%	0.64%	0.33%	0.31%	0.26%	0.71%	1.84%	2.54%	3.49%	4.39%
11	12	13	14	15	16	17	18	19	20	21
306468	350401	320580	331965	328104	346839	414771	407985	353060	379981	417722
5.32%	6.08%	5.57%	5.76%	5.70%	6.02%	7.20%	7.08%	6.13%	6.60%	7.25%
22	23									
313008	249396									
6.60%	5.42%									

Gleichverteilung würde bedeuten, dass jede Stunde einen Anteil von $100/24 \approx 4.17$ (%) hätte. Das ist eindeutig nicht der Fall: Zum Abend hin wird mehr gearbeitet. Diese offensichtliche Ungleichverteilung muss man nicht eigens testen.

Etwas anschaulicher wird sie durch die Abbildung [7](#), die die absolute Beiträge als Säule gegen

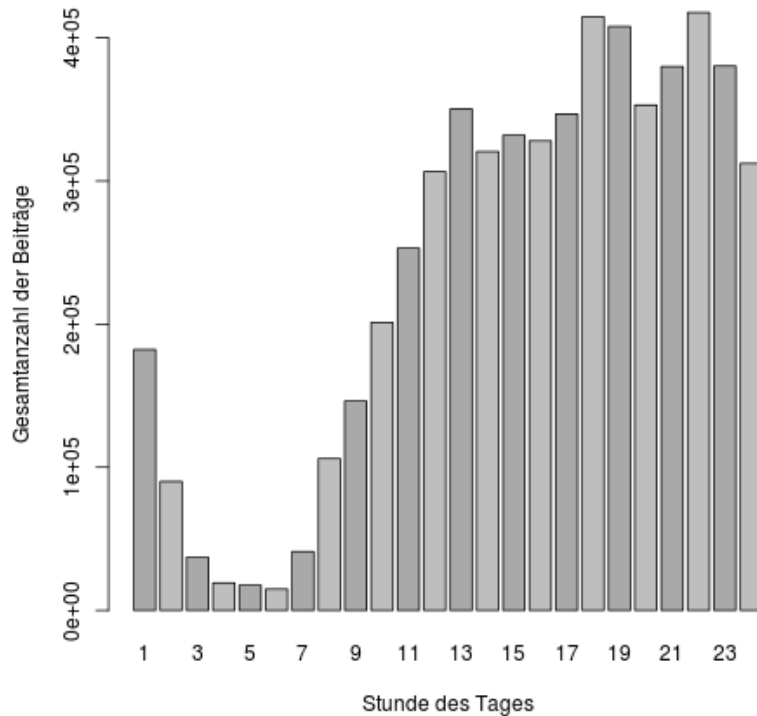


Abbildung 7: Verteilung der Beiträge im Tagesablauf, **H2a**

Man beachte auch die kleine Spitze zur Mittagspause (zwischen 12 und 13 Uhr).

H2b: Am Wochenende werden mehr Beiträge geleistet. (= Es gibt Häufungen an bestimmten Wochentagen.)

Die folgende Häufigkeitstabelle zeigt, in welchen Tagen der Woche die 5 759 218 Einträge erstellt worden sind. In der ersten Zeile die Stunde, darunter die absoluten Beiträge in dieser Stunde, in dritter Reihe die relativen Werte.

Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag	Sonntag
858592	818051	825932	817044	843956	826829	768814
14.91%	14.20%	14.34%	14.19%	14.65%	14.36%	13.35%

Diese Verteilung sieht sehr homogen aus. Darum wagte ich mich wieder eine an einen Kolmogorov-Smirnov-Test auf Gleichverteilung gewagt. Er gibt auch tatsächlich einen p-Wert von $0.1711 > 0.05$ aus: Die Nullhypothese, dass an allen Tagen gleich viel beigetragen wird, kann bei konventionellen Signifikanzniveaus nicht abgelehnt werden. Es gibt also keinen wöchentlichen Zyklus, wie ich erst annahm.

H2c: In den dunklen Monaten wird mehr gearbeitet. (= Es gibt jahreszeitliche Schwankungen.)

Das Balkendiagramm in Abbildung 8 stellt grafisch die Arbeitsleistung pro Monat dar:

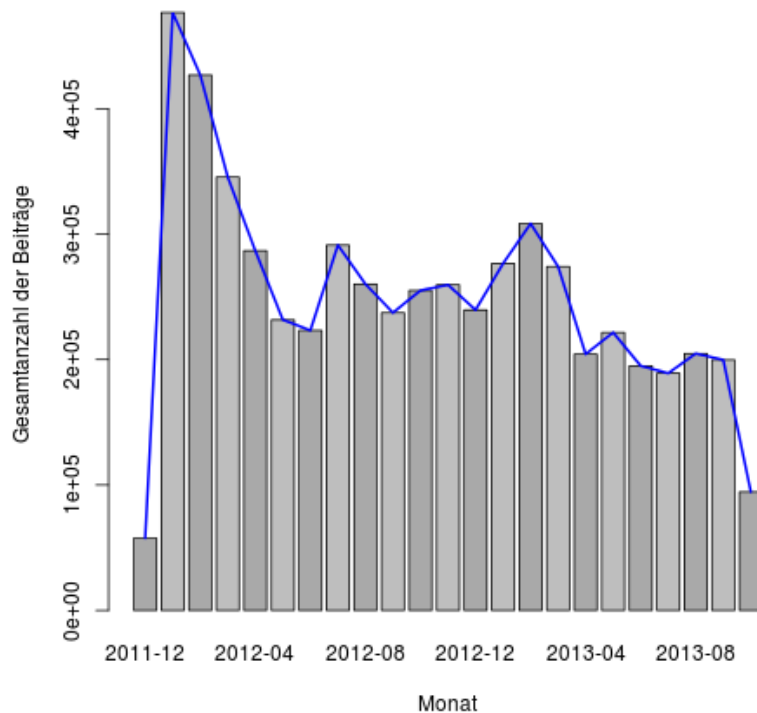


Abbildung 8: Monatliche Beiträge, **H2c**

Ob saisonale Trends vorliegen lässt sich meiner Meinung nach mit 23 Monaten (nicht mal zwei ganze Jahreszeitzyklen) nicht angemessen beurteilen. Die Abbildung 8 legt jedoch nicht den Schluß starker saisonaler Zyklen nahe.

H3: Die Häufigkeit der Bearbeitungen derer, die das Projekt nicht sofort wieder verlassen, steigt zunächst stark an, und sinkt dann wieder gegen ein langfristiges Gleichgewicht.

Diese Hypothese fragt, ob das individuelle Tätigkeitsprofil Schwankungen unterliegt. Ein bei vielen Freiwilligen und Ehrenamtlichen beobachtetes Phänomen ist anfangs eine hohe Begeisterung, der eine Ernüchterung folgt. Manchmal steht dann der Ausstieg an.

Die folgende Abbildung 9 habe ich aus technischen Gründen zunächst mit dem aktivsten Quantil (siehe unten im Abschnitt 7) erstellt. Ich denke dennoch, dass die Daten auch repräsentativ für die anderen 75% der Aktiven sind.

Leider sind die Daten etwas dadurch verzerrt, dass sich am Anfang viele anmeldeten und deswegen das gleiche Alter haben. Deswegen führen sie alle zur selben Zeit in den Sommerurlaub (oder andere Gründe). Vielleicht fällt mir ein pfiffiger Trick ein, etwaige saisonale Schwankungen im individuellen Erfassungsverhalten herauszufischen.

Jedenfalls erkennbar ist ein deutlicher Abfall der Beiträge von einem sehr hohen Niveau. Dieses hohe Niveau erklärt auch die extreme Spitze in der Abbildung 8 im Januar

2012. Abbildung 9 veranschaulicht das Beitragsprofil im Zeitablauf: Ganz links sind die durchschnittlichen Beiträge am ersten Tag, ganz rechts die durchschnittlichen Beiträge des 600. Tages. Das heißt, wenn es einen repräsentativen aktiven Erfasser gäbe: Wie er tippen würde zeigt die schwarze Linie. Die rote Linie gibt dagegen den Trend an.

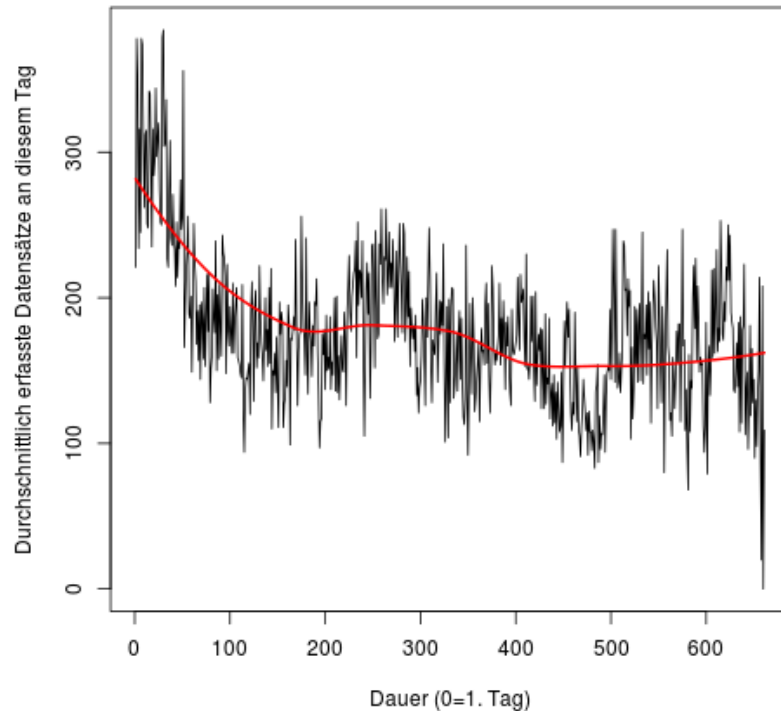


Abbildung 9: Beiträge nach Dauer, **H3**

Einen hohen Anstieg zu Beginn lässt der Graph nicht erkennen, dafür gibt es zu viele Einträge. Tatsächlich aber erfassen 72% der hier untersuchten aktiven Erfasser am zweiten Tag mehr als am ersten. Und der Sprung ist ziemlich hoch: Die durchschnittliche Erfassungsleistung am ersten Tag liegt bei 221 Einträgen, am zweiten Tag bei 378 und am dritten Tag immer noch 353. Erst nach über einem Monat(!) unterschritt die durchschnittliche Erfassungsleistung das Niveau des ersten Tages. Auch haben 56% der untersuchten Erfasser ihr Gesamtmaximum innerhalb der ersten 30 Tage.

Dennoch ist dies ohne statistischen Test kein Beweis, sondern nur ein Indiz für anfängliche Begeisterung.

H4: Die Wahrscheinlichkeit, dass unter den Beiträgen eines Erfassers Fehler gemeldet werden, sinkt mit der Anzahl seiner Bearbeitungen.

Das zu schätzende Modell lautet also: Fehlerquote = $\ln(\text{Anzahl der Beiträge}) + \epsilon$, wobei ϵ für den Fehler steht. Ich verwende eine logarithmische Skala, um nicht zu große Abstände zwischen den Beiträgen zu haben.

Natürlich würde ein Panel-Model besser zur Überprüfung der Hypothese dienen. Dabei würde für jeden Erfasser eine Zeitreihe gebildet, die seine Fehlerquote wiedergibt. Also die gemeldeten Fehler des ersten Arbeitstages dividiert durch die Gesamtzahl seiner Beiträge dieses Tages, usw. für jeden Erfasser und jeweils die Anzahl der Tage, die dieser Erfasser erfasst hat.

Call:

```
lm(formula = Fehlerquote ~ Beitraege)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.22795	-0.04775	-0.01399	0.02410	0.78612

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.277301	0.016284	17.03	<2e-16 ***
Beitraege	-0.027543	0.002119	-13.00	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 478 degrees of freedom

Multiple R-squared: 0.2612, Adjusted R-squared: 0.2596

F-statistic: 169 on 1 and 478 DF, p-value: < 2.2e-16

Das Modell hat zwar nur schwachen Erklärungsgehalt ($R^2 = 0.2612$), aber es läßt den Schluß zu, dass die Fehlerwahrscheinlichkeit mit der Anzahl der Beiträge sinkt. Der signifikante Ordinatenabschnitt (Intercept) und der sehr geringe Koeffizient (-0.027543) legen allerdings nahe, dass individuelle Effekte viel stärker sind.

Nicht-technisch ausgedrückt heißt dies: Zwar führt mehr Erfahrung mit der Schrift und dem Projekt zu weniger Problemen, aber nicht jeder Erfasser lernt in derselben Weise/Geschwindigkeit noch starten oder enden sie auf demselben Niveau.

Zur Veranschaulichung des großen Messfehlers die Abbildung 10. Die blaue Linie ist die Regressionskurve. Da die Steigung negativ ist, führen mehr Beiträge zu einer geringeren Fehlerquote. Die schwarzen Punkte geben eine Kombination von Fehlerquote (linke Achse) und Anzahl der Beiträge an (untere Achse). Die untere Achse ist logarithmisiert.

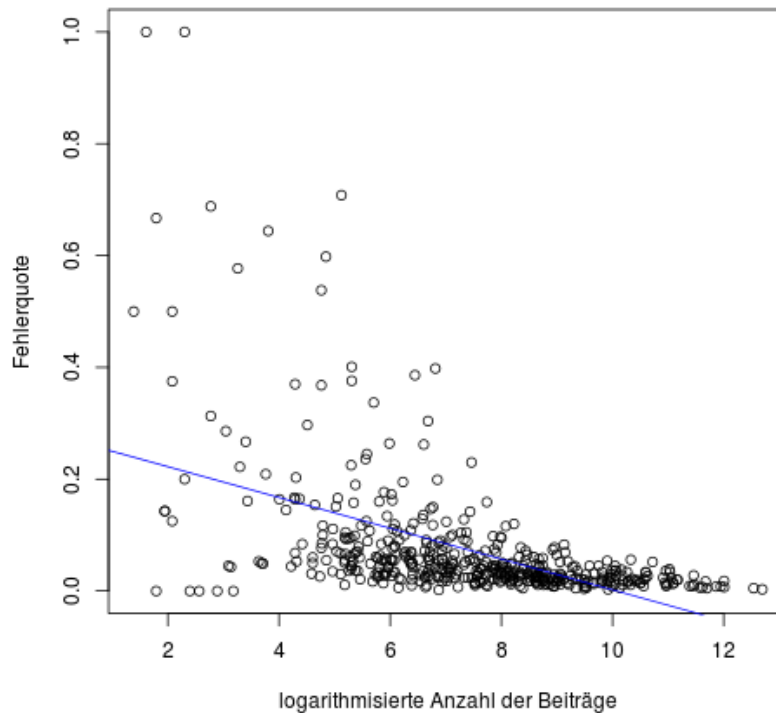


Abbildung 10: Ergebnis der Regression zu **H4**

7 Vergleich 1. Quartil Aktive mit 1. Quartil Inaktive

Aus technischen Gründen musste ich mich bei der Analyse des unterschiedlichen Benutzerverhaltens auf eine Stichprobe beschränken. Darum entschied ich mich für das jeweils 1. Quartil der Produktivität (= Beiträge/Dauer).

Das 1. Quartil der Aktiven umfasst naturgemäß 25% der Aktiven. Das entspricht 27 Personen, denn die Anzahl der Aktiven beträgt ja 106. Die Spanne der Produktivität dieser Erfasser reicht von 498 bis 93. Diese 25% haben zusammen genommen knapp 50% aller untersuchten Beiträge geleistet.

Das 1. Quartil der Inaktiven umfasst 93 von insgesamt 372 Inaktiven. Sie haben zusammen 16,7% aller untersuchten Beiträge getätigt.

Ich wollte nun wissen, wie sich diese unterscheiden. Dazu als erstes ein Vergleich der Beiträge im Zeitablauf. Beide Säulendiagramme in [Abbildung 11](#) zeigen die Gesamtzahl der Beiträge aller Quartilsmitglieder für diesen Tag dividiert durch die Anzahl der Mitglieder, die zu diesem Zeitpunkt noch aktiv waren. So kann man sehen, wie viele Datensätze die Aktiven und Inaktiven durchschnittlich am 1. Tag, am 2. Tag, etc. ihres Daseins als Erfasser erstellen. Insgesamt gibt es so 661 Balken:

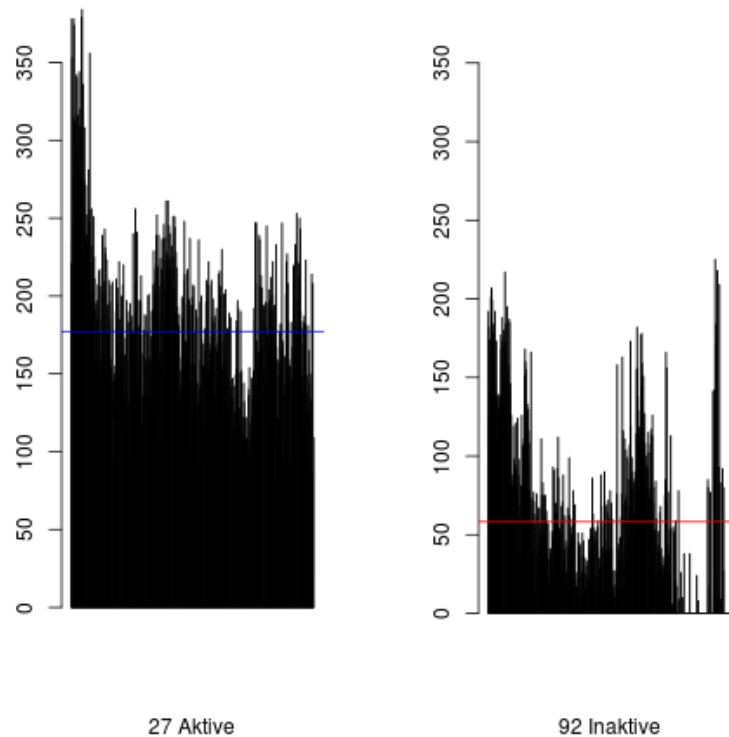


Abbildung 11: Vergleich der durchschnittlichen Beiträge nach Dauer

In beiden Säulendiagrammen habe ich zusätzlich die Mittelwerte eingezeichnet. Die Zahlen zu den Grafiken gibt es in der Tabelle 9:

Tabelle 9: Einstige und Ausstiege pro Jahr

Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Aktive	0.0	144.0	172.0	177.2	205.0	384.0
Inaktive	0.00	14.00	46.00	58.26	85.00	225.00

Im Datensatz erkennt man, dass die Inaktiven immer mal wieder Phasen haben, wo sie lange nichts machen. Das birgt ein Risiko: Eingangs erwähnte ich ja schon, dass die Definition der Inaktiven Fehler bergen kann: Personen, die nur sporadisch aktiv sind, sind hier nämlich als inaktiv gezählt, sofern sie bis nach dem 3. Oktober 2013 keinen Eintrag mehr vornahmen. Von 92 untersuchten Personen verlassen aber 50% das Projekt spätestens nach 17 Tagen; die letzten 200 Säulen gehen auf das Konto von gerade mal 10 Personen. Den Fehler, den ich bei der Definition mache, umfasst also höchstens diese wenigen als inaktiv gezählten sporadischen Aktiven.

Dies veranlasste mich aber, herauszufinden, wie oft die Erfasser tippen. Die Abbildung 12 stellt den Anteil der Tage ohne Aktivität an der gesamten Dauer dar.

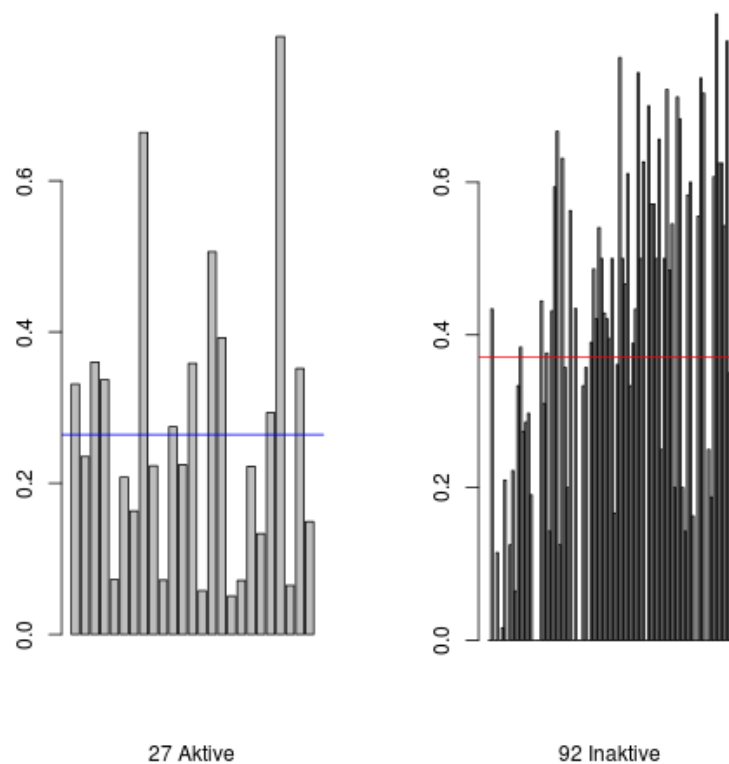


Abbildung 12: Vergleich der durchschnittlichen Beiträge nach Dauer

Die Zahlen dazu in der Tabelle 10:

Tabelle 10: Anteil der Tage ohne Aktivität						
Benutzergruppe	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Aktive	0.0506	0.1332	0.2244	0.2642	0.3517	0.7909
Inaktive	0.000	0.1875	0.3902	0.3708	0.5628	0.8202

Die 0 beim Minimum stammt von den 14 Erfassern, die nur bis zu fünf Tage dabei waren und an jedem dieser Tage gearbeitet haben.

Zwar finden sich in beiden Gruppen Erfasser, die nahezu täglich erfasst haben, doch sind diese im 1. Quartil der Aktiven viel häufiger. Unter den 92 Inaktiven finden sich nur 28 Personen (30%), die mehr als die Hälfte aller Tage, die sie aktiv waren, nicht erfasst haben. Sie haben also im Durchschnitt höchstens jeden zweiten Tag beim Projekt vorbeigeschaut. Bei den 27 Aktiven sind es nur 3 Erfasser (11%).

8 Zusammenfassung

- Die fünf Admins erbrachten alleine 12,01%.

- 50% aller untersuchten Erfasser erbrachten ihren ersten Beitrag vor dem 1. März 2012.
- Je später ein Erfasser dazu kam, desto häufiger wurde er inaktiv.
- Je mehr ein Erfasser konstant zu Beginn leistet, desto niedriger die Wahrscheinlichkeit, dass er in naher Zukunft aussteigt.
- Sowohl bei den Eintagsfliegen als auch bei den Inaktiven gibt es hinsichtlich der Produktivität extreme Ungleichgewichte derart, dass wenige viel machen und viele wenig.
- Vermutlich haben einige der Eintagsfliegen ein weiteres Konto.
- Im Jahr 2013 sind (bis 17. Oktober) 75 Erfasser mehr ausgestiegen als eingestiegen.
- 35% aller Neuerfasser im Jahr 2013 (bis 17. Oktober) waren Eintagsfliegen, im Vorjahreszeitraum dagegen nur 17% der Neuerfasser.
- 2012 gab es (bis 17. Oktober) 68% weniger Neuversuche als im Vorjahreszeitraum.
- Etwa 65% aller Neuerfasser pro Jahr waren zum 17. Oktober 2013 als inaktiv definiert.
- 20% der untersuchten Bearbeiter erledigten 85% der untersuchten Beiträge.
- Es gibt einen täglichen Zyklus: 40% der Bearbeitungen zwischen 17 und 23 Uhr mit kleiner Spitze zur Mittagspause.
- Es gibt keinen wöchentlichen Zyklus.
- Es gibt vermutlich keinen jahreszeitlichen Zyklus.
- Etwa die Hälfte aller Erfasser sind im ersten Monat am produktivsten.
- Mehr Erfahrung führt zu einer geringeren Fehlerquote, aber individuelle Effekte sind viel stärker.
- Unter den als inaktiv definierten Erfasser gibt es etliche, die nur sporadisch aktiv waren und deshalb nicht zwingend inaktiv sind.
- Die produktivsten 25% der Erfasser tippen im Schnitt mindestens jeden zweiten Tag.

9 Ausblick

Die Analyse ist noch nicht fertig. Die Eingangsfrage ist jedenfalls leicht beantwortet: Der Rückgang der durchschnittlichen Erfassungsleistung ist sowohl auf zurückgehende individuelle Beitragsleistung als auch auf zurückgehende Aktiven-Zahlen zurückzuführen. Ich möchte aber weitermachen und genau klären, welcher Effekt wie stark eingeht.

Auch wäre es interessant, die Benutzer mit weiteren Daten zu analysieren: Wie viel Zeit verging zwischen der Anmeldung und dem ersten Beitrag? Haben sie sich gleich auf der Mailingliste angemeldet? Haben sich die Inaktiven von der Mailingliste abgemeldet? Haben sich manche Eintagsfliegen nochmal angemeldet? Waren die Erfasser vorher schon in anderen CompGen-Projekten (inkl. Forum) angemeldet? Wie viele Erfasser suchen systematisch Fehler?

Vielleicht habt ihr auch noch Ideen für weitere Analyse und Vermutung, die als Hypothese getestet werden könnten? Dann immer her damit!