How many Stan users are there?

General

s.maskell   Dec '19

It seems we could look at a number of pre-existing data sources (eg discourse views and contributors, papers, StanCon attendance etc) to inform an inference of how many people use Stan (and/or use things that use Stan). We could also generate new data (eg via surveys etc). Do we know the answer and/or how best to work it out?

created	last reply	22	1.3k	8	19	3	
Dec '19	Jan '20	replies	views	users	likes	links	

lauren   Dec '19

I've thought a tonne about surveying the Stan community or potentially capture-recapture techniques, but generating new data in this way would be a lot of work and my current funding wouldn't cover it (that'll change next year though!). I'd be interested in collaborating if anyone is interested. :)

s.maskell   Dec '19

As is hopefully obvious, i am keen to help. I can probably deploy people on this who can help make this happen if we know what it is that we needed them to do.

avehtari Stan Developer   Dec '19

How about asking who has registered to discourse? From that we could extrapolate the population size for those who we could reach by survey?

s.maskell   Dec '19

@avehtari : who would we survey and how would we get the survey to them? I feel I am being dim.

lauren   Dec '19

From memory the hope was to utilized a snowball survey (once you finish you forward onto other people in the population), which is generally what you use when you have a hard to define population. One of the stats grads here, Jonathan Auerbach and I talked about starting multiple snowballs (Andrew's blog, Discourse, Twitter, StanCon mailing lists, etc.) and then tracking what snowball people were recruited by (and if multiples) as a way of measuring coverage.

1 Reply   2

andrewgelman   Dec '19

This is a fascinating statistics question. Maybe I could post a blog on this and see if there are any thoughts.

                                                                 6

jroon   lauren   Dec '19

@lauren I think you should name that design a snowball fight 😂

lauren   Dec '19

That's a fantastic name!!!

@andrewgelman you can, it'd be interesting to see what folks think.

Bob_Carpenter   Dec '19

What's your definition of a Stan user? Is it someone who once downloaded Stan and ran a model? Or is it someone who uses Stan regularly, and if so, how regularly? What about people who use packages like brms or rstanarm or prophet that are built on top of Stan?

There are over 3K users registered on Discourse, but that doesn't mean they're regulars. Many of them only showed up once.

1 ♡ 🔗

andrewgelman   Dec '19

Blog post on the topic scheduled for Monday.

♡ 🔗

s.maskell   Dec '19

For context, UK academic departments (eg the one I sit in) are assessed (as part of the "Research Excellence Framework", REF) on the basis of some criteria: the next assessment is imminent, but the one after that will be in 2026. When aggregated over each entire University, the outcome of that assessment process modulates the amount of (Quality-related Research, QR) income that the University receives from the UK government. So, it's important to inform the assessment process with pertinent information.

One of the three assessment criteria is "impact", which relates to the uptake of academic research outside of the academic discipline it came from and is metricated in terms of "reach and significance" during the census period. Unfortunately, "reach and significance" is not defined quantitatively. However, the notion of the census period is defined as people using the research during a specified period of time (eg 2021-2026).

The specific motivation for my question is that I'd like to understand how we (locally to my department) could quantify the "reach and significance" of any enhancements to Stan that (we hope!) might come out of our work between now and 2026. It seems natural to start by finding out what we (as the Stan community) know about how big we are now.

So, in answer to your question, I think I'd ideally like to know how many people and/or organisations are making use of specific subsets of Stan's code (including in any packages that use Stan) during a specified period. I'd also like to know where they are based geographically, whether people work in academic, industry or government, the demographic of applications they are working on, etc etc.

That's clearly hinting at scope creep, but hopefully helps explain the specific reason I asked the question, which I see as an important step towards quantifying the impact of our work to help enhance Stan.

♡ 🔗

lauren   Dec '19

I think this would be relevant outside of the REF as well - a useful addition to many grants/impact sections.

> 🖼 **s.maskell:**   ⌄ ⌃
>
> So, in answer to your question, I think I'd ideally like to know how many people and/or organisations are making use of specific subsets of Stan's code (including in any packages that use Stan) during a specified period. I'd also like to know where they are based geographically, whether people work in academic, industry or government, the demographic of applications they are working on, etc etc.

We were at the point where we were focussing on questions around this, plus an emphasis on barriers to entry. I can dig them up if there's renewed interest. I believe the conclusion last time was that the time-cost of doing it wasn't worth the expected benefit, but it could have changed! :)

♡ 🔗

andrewgelman   Dec '19

I posted my question here: https://statmodeling.stat.columbia.edu/2019/12/09/how-many-stan-users-are-there/ 48

2 ♡ 🔗

avehtari Stan Developer   Dec '19

This paper might have useful ideas Using an Online Sample to Estimate the Size of an Offline Population 5 how to estimate the number of those users who would be unlikely to be reached by survey.

♡ 🔗

mitzimorris 🛡   Dec '19

Stack Overflow has tag "Stan" and 254 users (if I'm interpreting their graphics correctly) have listed Stan as things they're interested in. about once a week someone asks a Stan question.

🟧 **stack overflow**   Products   🔍 Search…

Home

PUBLIC

🌐 Stack Overflow

Tags

Users

Jobs

TEAMS   What's this?

📇 First 25 Users Free

Tags

A tag is a keyword or label that categorizes your question
makes it easier for others to find and answer your questio

🔍 stan

stan × 254

open source software for Markov chain Monte Carlo
sampling, often used for multilevel Bayesian modeling.

5 asked this month, 60 this year

---

2 Replies ⌄   1 ♡   🔗

---

mcol Stan Developer   Dec '19

254 is the number of questions with that tag, the number of users will be less than that.

                                                       1 ♡   🔗

---

mitzimorris 🛡   Dec '19

doh!

users do list their interests, but no way to scrape that out of SO.

                                                       ♡   🔗

Ⓛ **lauren**   Dec '19

It's an interesting question! We might want to think about targeting this question using a
statistical/surveying method and then benchmarking our population estimates against other potential
indicators as another coverage check (number of downloads of Stan in R packages relative to Python
etc.).

One of the challenges for this is agreeing upon what the definition of a Stan user is.

                                                       ♡   🔗

Ⓛ **lauren**   Dec '19

I think we're going to have a hangouts meeting on this topic to see if we can make some sort of plan.
I've emailed folks who's email I have but for those interested we're picking a time here, and if you put
your full name in I can probably google your email for an invite. :) https://www.when2meet.com/?
8493674-nr5Ai 1

                                                       1 ♡   🔗

Bob_Carpenter 🛡   ↩ 👤 mitzimorris   Dec '19

There are also questions tagged rstan and pystan and brms, etc. on StackOverflow. We have fewer
questions asked on StackOverflow than we would if we didn't have the dedicated Stan forums here.

                                                       ♡   🔗

---

avehtari Stan Developer   Dec '19

From July 2015 to this week, RStan for **@mcmc_stan** has been downloaded from RStudio CRAN mirror
one million times (loo 630K times). Python users love much more downloading as PyStan gets 700K+
downloads just in the last month.

We don't know how many **@mcmc_stan** users these imply, but during the last month RStan downloads
were 6% of ggplot2 downloads (200% of R tensorflow) and PyStan downloads were 1.6% of numpy
downloads (9% of matplotlib and 14% of tensorflow downloads).

In 2019 it was estimated that there are 8 million Python developers
(https://www.zdnet.com/article/programming-languages-python-developers-now-outnumber-java-ones/).
If half of them use numpy, then we get an estimate of 64K PyStan users.

In 2014 it was estimated that there are 2+ million R users
(https://blog.revolutionanalytics.com/2014/04/a-world-map-of-r-user-activity.html). Assuming
conservatively the same number in 2019 and half of them download ggplot2, then we get estimate of
60K RStan users.

Combined estimate would be more than 120K **@mcmc_stan** users plus CmdStan, StatStan, JuliaStan
and MatlabStan users. Of course it is possible that Stan users download much more often than average
users which would inflate these estimates.

Also it's difficult to estimate how many users use pre-installed Stan in cloud services or how many
"users" are students who are forced to use Stan in a course (e.g.
https://github.com/avehtari/BDA_course_Aalto/) and then never use it after that. But the order of

magnitude of the above estimates is likely to be correct.

5 ♡ 🔗

1 MONTH LATER

Bob_Carpenter ⬚   Jan '20

> avehtari: ∨ ↑
>
> Python users love much more downloading as PyStan gets 700K+ downloads just in the last month.

But that's just the RStudio mirror. I don't use it as I don't use RStudio.

> avehtari: ∨ ↑
>
> But the order of magnitude of the above estimates is likely to be correct.

I've been telling people "roughly 100K". There are about 4K Google Scholar citations now I can find and about 3K users registered in the forums.

♡ 🔗

↩ Reply

Suggested Topics

Topic	Replies	Views	Activity
Hello,I am new user and I need help.Inside E[i, m] what will it add each time?! The first time and with (j in 1:1) it will add E [1, m] (I'm not sure). The next ones?! ■ General	4	76	Nov '20
Creation of a bunch of small paid jobs to help Stan operate ■ General	5	523	Dec '19
Issues with building a package with Rstan ■ General ■ rstan	1	172	May '20
Stan for multi-omics data integration ■ General ■ rstan	8	213	Oct '20
Constraints on parameters/data ■ General	3	74	Dec '20

Want to read more? Browse other topics in ■ **General or view latest topics.**

WinBUGS and Friends Downloads

These count downloads from the R packages that either are interfacees to Stan, winBUGS, JAGS, Nimble or OpenBUGS or are directly dependent on these packages.

```
library(tidyverse)
library(knitr)
library(cranlogs)
```

Specify packages and get download statistics.

```
packages <- c("rstan","rstanarm","MetaStan","StanHeaders","R2WinBUGS","R2jags","rjags","runjags","rstantools","jag
sUI","nimble","nimbleEcology", "nimbleSCR","nimbleSMC","R2OpenBUGS","BRugs")
dat <- cran_downloads(package = packages, from = "2013-08-01", to = "2020-12-31")
```

Downloads by package:

```
kable(dat %>% group_by(package) %>% summarise(downloads=sum(count)))
```

package	downloads
BRugs	141688
jagsUI	109241
MetaStan	13784
nimble	64031
nimbleEcology	8559
nimbleSCR	1738
nimbleSMC	663
R2jags	326454
R2OpenBUGS	166839
R2WinBUGS	342887
rjags	929072
rstan	2040196
rstanarm	528787
rstantools	695631
runjags	166738
StanHeaders	1766344

Total downloads:

```
kable(dat %>% summarise(downloads=sum(count)))
```

downloads
7302652