# StackOverflow data dump

## Annie Ying

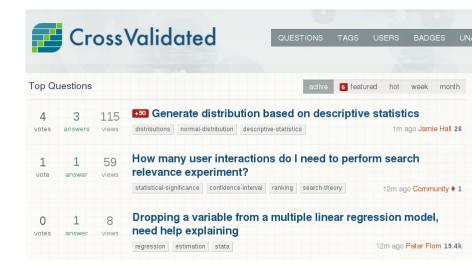
PhD candidate, McGill School of Computer Science

April 3, 2013

ever had

a burning stats question?

## StackOverflow!



## StackOverflow!



QUESTIONS

## How many user interactions do I need to perform search relevance experiment?



I know similar questions have been asked many times here. But I still don't get simple formula or idea how estimate this



So I have site with traffic and I want to split portion of it to test new search relevance algorithms i.e. result sorting.

I will measure abandonment rate (searches without click/total amount of searches). I want to get this statistics with -/+ 2% accuracy with 95% confidence. Is there simple heuristic(something practical) I can apply to get required number of experiments(10^3,10^6,...)?

share improve this question

statistical-significance confidence-interval ranking search-theory

edited Sep 26 '12 at 23:16 Peter Flom

9.4k 2 2 21 47

asked Sep 26 '12 at 23:05

active oldest

RN Wound Care

Blake Medical - Bradenton FL

Specialist Clinical Outcomes...

Trauma PI & Injury Prevention Lawnwood... - Fort Pierce. FL

RN-Nurse Practitioner... Alegent Health - Council Bluffs, IA

Your question relates to the idea of statistical power. There are a lot of threads on CV that provide information to help you think about this, search under: power and power-analysis. In particular, you

1 Answer

may want to read through this answer of mine: How to report general precision in estimating correlations within a context of justifying cample size, which discusses the idea of Accuracy in

tagged

statistical-significance × 553 confidence-interval × 443

6 months ago

ranking × 68

search-theory × 7

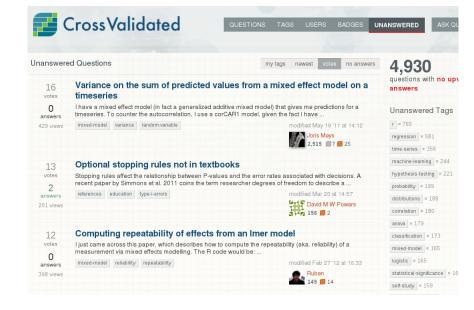
viewed 59 times active today

Related Jobs

Texas Health... - Stephenville, TX

ever asked one but unanswered?

# Many questions unanswered



why doesn't a stat question get answered on StackOverflow?

## We can look at the data!



## Stack Overflow Creative Commons Data Dump

06-04-09 by Jeff Atwood, 94 comments

We decided early on that all user-generated content on Stack Overflow would be under a Creative Commons license.

All those great Stack Overflow questions, answers, and comments, so generously contributed by all of you, are licensed under cc-wiki (also known as cc-by-sa):



### cc-wiki license

#### You are free

- to Share to copy, distribute, and transmit the work
- to Remix to adapt the work

### Under the following conditions

- Attribution You must attribute the work in the manner specified by the author or licensor (but not in
  any way that suggests that they endorse you or your use of the work).
- Share Alike If you alter, transform, or build upon this work, you may distribute the resulting work only
  under the same or similar license to this one.

The community has selflessly provided all this content in the spirit of sharing and helping each other. In that very same spirit, we are happy to return the favor by providing a database dump of public data.

some ideas for the hackathon:

- what factors associate with an answered question?

- a classifier on whether a question will be answered

# Data dump stats.stackexchange.com 2011/09

```
drwx----- 2 aying1 nogroup
                                4096 Jun 13
                                             2011 062011 Statistical Analysis
                             637074 Jun 13
            aying1 nogroup
                                             2011 badges.xml
            aying1 nogroup
                                             2011 comments.xml
                             5685153 Jun 13
rw-r--r-- 1 aving1 nogroup
                                1786 Jun 13
                                             2011 license.txt
rw-r--r-- 1 aying1 nogroup 25767045 Jun 13
                                             2011 posthistory.xml
           aying1 nogroup
                                             2011 posts.xml
                            15393713 Jun 13
           aying1 nogroup
                                4675 Jun 13
                                             2011 readme.txt
            aving1 nogroup
                             8885685 Aug 4
            aving1 nogroup
                             1390687 Jun 13
                                             2011 users.xml
rw-r--r-- 1 aving1 nogroup
```

## Schema

```
stats.stackexchange.com : more
      - Text, e.g.: "@Stu Thompson: Seems possible to me - why not try it?"
      - CommentCount
```