

# StackOverflow data dump

Annie Ying

PhD candidate, McGill School of Computer Science

April 3, 2013

ever had

a burning stats question?

# StackOverflow!



## CrossValidated

[QUESTIONS](#)[TAGS](#)[USERS](#)[BADGES](#)[UNANSWERED](#)

### Top Questions

[active](#)[6 featured](#)[hot](#)[week](#)[month](#)

4

votes

3

answers

115

views

+50

**Generate distribution based on descriptive statistics**[distributions](#)[normal-distribution](#)[descriptive-statistics](#)1m ago [Jamie Hall](#) 26

1

vote

1

answer

59

views

**How many user interactions do I need to perform search relevance experiment?**[statistical-significance](#)[confidence-interval](#)[ranking](#)[search-theory](#)12m ago [Community](#) ♦ 1

0

votes

1

answer

8

views

**Dropping a variable from a multiple linear regression model, need help explaining**[regression](#)[estimation](#)[stata](#)12m ago [Peter Flom](#) 19.4k



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



1



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  $\pm 2\%$  accuracy with 95% confidence. Is there simple heuristic (something practical) I can apply to get required number of experiments ( $10^3, 10^6, \dots$ )?

[statistical-significance](#)[confidence-interval](#)[ranking](#)[search-theory](#)

share improve this question

edited Sep 26 '12 at 23:16



Peter Flom

19.4k 2 21 47

asked Sep 26 '12 at 23:05



yura

173 5

tagged

[statistical-significance](#) × 553

[confidence-interval](#) × 443

[ranking](#) × 68

[search-theory](#) × 7

asked 6 months ago

viewed 59 times

active today

#### Related Jobs

[RN Wound Care](#)

Blake Medical... - Bradenton, FL

[Specialist Clinical Outcomes...](#)

Texas Health... - Stephenville, TX

[Trauma PI & Injury Prevention](#)

Lawnwood... - Fort Pierce, FL

[RN-Nurse Practitioner...](#)

Alegent Health - Council Bluffs, IA

1 Answer

[active](#)[oldest](#)[votes](#)

0




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

may want to read through this answer of mine: [How to report general precision in estimating correlations within a context of justifying sample size](#), which discusses the idea of *Accuracy in*

ever asked one  
but unanswered?

# Many questions unanswered

 CrossValidated

QUESTIONS TAGS USERS BADGES **UNANSWERED** ASK Q

Unanswered Questions

my tags newest votes no answers

16  
votes

0  
answers


429 views

**Variance on the sum of predicted values from a mixed effect model on a timeseries**

I have a mixed effect model (in fact a generalized additive mixed model) that gives me predictions for a timeseries. To counter the autocorrelation, I use a corCAR1 model, given the fact I have ...

mixed-model variance random-variable

modified May 19 '11 at 14:12

 **Joris Meys**  
2,515 7 25

13  
votes

2  
answers


201 views

**Optional stopping rules not in textbooks**

Stopping rules affect the relationship between P-values and the error rates associated with decisions. A recent paper by Simmons et al. 2011 coins the term researcher degrees of freedom to describe a ...

references education type-i-errors

modified Mar 20 at 14:57

 **David M W Powers**  
156 2

12  
votes

0  
answers


398 views

**Computing repeatability of effects from an lmer model**

I just came across this paper, which describes how to compute the repeatability (aka. reliability) of a measurement via mixed effects modelling. The R code would be: ...

mixed-model reliability repeatability

modified Feb 27 '12 at 16:33

 **Ruben**  
145 14

4,930

questions with no up  
answers

Unanswered Tags

r × 765

regression × 581

time-series × 356

machine-learning × 244

hypothesis-testing × 221

probability × 199

distributions × 188

correlation × 180

anova × 179

classification × 173

mixed-model × 165

logistic × 165

statistical-significance × 16

self-study × 159

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

```
total 59910
drwx----- 2 aying1 nogroup    4096 Jun 13  2011 062011 Statistical Analysis
-rw-r--r--  1 aying1 nogroup  637074 Jun 13  2011 badges.xml
-rw-r--r--  1 aying1 nogroup  5685153 Jun 13  2011 comments.xml
-rw-r--r--  1 aying1 nogroup    1786 Jun 13  2011 license.txt
-rw-r--r--  1 aying1 nogroup 25767045 Jun 13  2011 posthistory.xml
-rw-r--r--  1 aying1 nogroup 15393713 Jun 13  2011 posts.xml
-rw-r--r--  1 aying1 nogroup    4675 Jun 13  2011 readme.txt
-rw-r--r--  1 aying1 nogroup  8885685 Aug  4  2011 stats.stackexchange.com.7z
-rw-r--r--  1 aying1 nogroup  1390687 Jun 13  2011 users.xml
-rw-r--r--  1 aying1 nogroup  3564146 Jun 13  2011 votes.xml
```

# Schema

stats.stackexchange.com : more

Statistical Analysis - Data Dump: June 2011

- Format: 7zipped
- Files:
  - \*\*badges\*\*.xml
    - UserId, e.g.: "420"
    - Name, e.g.: "Teacher"
    - Date, e.g.: "2008-09-15T08:55:03.923"
  - \*\*comments\*\*.xml
    - Id
    - PostId
    - Score
    - Text, e.g.: "@Stu Thompson: Seems possible to me - why not try it?"
    - CreationDate, e.g.: "2008-09-06T08:07:10.730"
    - UserId
  - \*\*posts\*\*.xml
    - Id
    - PostTypeId
      - 1: Question
      - 2: Answer
    - ParentId (only present if PostTypeId is 2)
    - AcceptedAnswerId (only present if PostTypeId is 1)
    - CreationDate
    - Score
    - ViewCount
    - Body
    - OwnerUserId
    - LastEditorUserId
    - LastEditorDisplayName="Jeff Atwood"
    - LastEditDate="2009-03-05T22:28:34.823"
    - LastActivityDate="2009-03-11T12:51:01.480"
    - CommunityOwnedDate="2009-03-11T12:51:01.480"
    - ClosedDate="2009-03-11T12:51:01.480"
    - Title=
    - Tags=
    - AnswerCount
    - CommentCount
    - FavoriteCount