Developing, running and evaluating an experiment

A case study:
Evaluating implicit judgements from Image search interactions

Background: Case study topic

- Click-through data
- Multiple types of search
- Click-through data as implicit judgement

Overview: Developing the Experiment

- Inception of idea for the experiment
- The importance of evaluation - the notion of statistically significant
- Experimental design
  - The importance of evaluation
  - Ethics consideration
  - Justification of participant numbers (on power, chicken and eggs)
- Pilot Study setup (and the importance of evaluation)
- Predicting participant numbers

Overview: Running an experiment

- Participant perception
- Completion rates
- Availability of the system
- Considerations for a system in “the wild”
- Some problems we faced and what (not) to do

Overview: Evaluating the experiment

- Step 3 should be easy right..... (experiences from MSN experiment)
- Its just some simple statistics, it'll only take a week....
- The importance of evaluation
- Interpreting the statistics and explaining "why"

Developing the experiment

The most important part

Inception of the idea

Inception of the idea

Motivation - a research question

Inception of the idea

Motivation - a research question

“How reliable is click-though data as implicit user judgement as a proxy for relevance in Internet image search?”

Inception of the idea

Motivation - a research question

“How reliable is click-though data as implicit user judgement as a proxy for relevance in Internet image search?”

What I care about:

- The accuracy of click-through information

Inception of the idea

Motivation - a research question

“How reliable is click-though data as implicit user judgement as a proxy for relevance in Internet image search?”

What I care about:

- The accuracy of click-through information

What I’m measuring:

- The proportion of clicks where the query and image are relevant

Inception of the idea - motivation

"... users' clicking decisions... are biased by

the trust they [the users] have in the

retrieval function, and by the overall quality

of the result set. This makes it difficult to

interpret clicks as absolute feedback."

Joachims et al. 2005

Inception of the idea - motivation

"When users clicked on a page, they were

satisfied 39% of the time."

Fox et al. 2005

Inception of the idea - motivation

"...the proportion of clicked documents that

are actually considered to be relevant is only

52%. This is surprisingly low, and indicates

that it is not safe to infer relevance directly

from recorded click information."

Scholer et al. 2008

Experimental design

- Vague experiments lead to vague, results that are not statistically significant
- Won’t be accepted as research in conferences, journals or by your peers
- Things to consider:
  - Hypothesis
  - Dependent/Independent variables
  - Statistical evaluation methodology
  - Doing this: “Yes! I can collect a lot of data” = fail

Experiment design

- Worst nightmare:

- Experiment is considered invalid or flawed
    - You influenced the participants
    - You presume that topics are known or unknown when they are not

- Experiment does not show anything
    - Variation among subjects is higher than the variation among variables

- Experiment shows something trivial
    - E.g. peoples notion of ambiguity is not the same

- Experiment does not answer the question

Experiment design

- If possible base on a previous (accepted) experiment:

- Falk Scholer, Milad Shokouhi, Bodo Billerbeck, and Andrew Turpin. Using clicks as implicit judgments: Expectations versus observations. In ECIR, volume 4956, pages 28–39. Springer, 2008.

- Fixed system precision to evaluate:
    - ordering bias
    - quality bias
- Evaluate accuracy as a side note (proportion of clicks)

Experiment design: Hypothesis testing

- Important: a basic knowledge of statistics
  Recommend: http://onlinestatbook.com/

- Step one: Develop a hypothesis test

- What is a hypothesis test?
  - Statistical procedure
  - Tests whether chance is a plausible explanation

- A basic hypothesis:
  Does the precision of the system affect the click-through accuracy?

Experiment design Recommend: http://onlinestatbook.com/

- A basic hypothesis:
  Does the precision of the system affect the click-through accuracy?

- Generic form:
  If I change X and measure Y, does Y change?
  Does changing the independent variable affect the dependent variable?

- Independent variable (factor): System precision

- Dependent variable: Click-through accuracy (measured as a proportion)

Experiment design

Does the precision of the system affect the click-through accuracy?

- Basic hypothesis:
  - 1 independent variable [aka factor] (system precision)
  - 1 dependent variable (accuracy = accurate clicks/total clicks)

- Formulate the hypothesis as a **null hypothesis (H_0)**
  H_0: avg(accuracy@lowPrecision) = avg(accuracy@highPrecision)

- Previous hypothesis is then called the **alternate hypothesis (H_1)**

- Simple statistical process to verify if findings by chance or not
  - e.g. t-test, one-way ANOVA. Today I will focus on ANOVA designs.

Experimental Design: ANOVA

- ANOVA ANalysis Of VAriance
- Compares two or more means
- Concludes that at least one population mean is different from at least one other mean, but NOT between which ones.
- **Balanced ANOVA:** The same number of observation per factor combination
- **Unbalanced ANOVA:** More complicated (not covered today)

Experiment design: People make things harder

Does the precision of the system affect the click-through accuracy?

- Basic hypothesis:
  - 1 independent variable (system precision)
  - 1 dependent variable (accuracy = accurate clicks/total clicks)

- More complex: When people are involved.
  - Variance between people!
  - Must be taken into account

  - **Important:** If variance between people is to great it will obscure the effect of the variable trying to be observed!

  - Tasks involving human interpretation - be VERY CAREFUL!!!!

Experiment design: With people, some thoughts...

Does the precision of the system affect the click-through accuracy?

- Say I choose to evaluate 2 levels of system precision, high & low

- Will each participant give measurements for both? (high **and** low)?
  - known as **within-subject factor design**

- Only for one level? (high **or** low)?
  - known as **between-subject factor design**

- Easiest: Between subject. Drawback, need more participants.

- Within-subject: Must take into count replication. Less participants required.

Experimental Design: Recap

- Question: *Does the precision of the system affect the click-through accuracy?*

- Alternative hypothesis: System precision effects click-though accuracy
  - ANOVA selected. **Single factor** (system precision), with 2 levels (high, low)
  - We will use a **Balanced ANOVA design** (same num observations per level)

  H_1: avg(accuracy@lowPrecision) != avg(accuracy@highPrecision)

- Formulated as the null hypothesis
  H_0: avg(accuracy@lowPrecision) = avg(accuracy@highPrecision)

- Experiment involves people - **within subject** experiment chosen.

- Question: How many people do I need?

Experiment design

Does the precision of the system affect the click-through accuracy?

- Foreseeable outcomes:
  I) An observed differences in the means
    - statistically significant result (good - reject null hypothesis)
    - not statistically significant result (not so good)

- Misconception: A non-significant outcome means that the null hypothesis is probably true.

  Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

- Clearly a non-significant result is BAD!!! How can we try and prevent this?

Experiment design - The notion of power

- Clearly a non-significant result is BAD!!! How can we try and prevent this?
- Power is defined as the probability of correctly getting a significant result
- More formally: Power is defined as the probability of correctly rejecting a false null hypothesis (i.e. accepting the alternate hypothesis, aka getting a result).

Calculating Power - The chicken and egg problem

- A simple demo based on our simple example:

- Using some software: Piface, http://www.stat.uiowa.edu/~rlenth/Power/

- The design properties we are assuming:
  - ANOVA assumptions hold
  - Balanced, single factor ANOVA design
  - Within-subject design

- Extra data we will need to know:
  - Variance between the measurements of participants for each level (high and low). For ANOVA design these are **presumed equal!**
  - Variance between levels (high and low)
  - An acceptable power level

Calculating Power - The chicken and egg problem

- Extra data is hard to predict!

- Either guess from past studies
  OR....

- Pilot study!

Demo: http://www.stat.uiowa.edu/~rlenth/Power/

design properties	ANOVA applicable Balanced design Single factor design Within-subject design
s.d. between participants	0.093
s.d. between levels	0.119
acceptable power level	0.8

Estimating participant numbers: Who cares?

- You - ultimately you have to find willing participants!

- Ethics committees - will not allow you to waste participants time
  - You must justify your choice
  - To few, won't get a result and waste everyone's time
  - To many and you waste the time of the extras you didn't need

Experimental design: This case study in reality

- Much more complex.

- Multiple factors:
  - System precision
  - Generic vs. Specific search topics
  - Object vs. Scene search topics
  - Known vs. Unknown search topics

- Factors at multiple levels (mix of both within and between subject factors)
  - System precision (between subject)
  - Rest (within subject)

- Multi-level model design. Get help!

Extra design consideration in practice

- Goal: Evaluating the accuracy of click-through data in image search
  - Evaluation metrics of “accuracy”
  - Giving images ground truths

- Goal: Evaluating numerous factors,
  - System precision
  - Generic vs. Specific search topics
  - Object vs. Scene search topics
  - Known vs. Unknown search topics

- Task description to limit participant interpretation (aka variance)

Thinking it through: potential “hard” questions

- Inevitably you will be asked questions about your work

- In experimental design must think though questions to ensure you have a good answer

- Examples with respect to the case study

  “how were *generic, specific* topics selected?”
  “do you think you can actually define *generic*, *specific* etc?”
  “why those factors? What about abstract concepts like *happy?*”

What we covered

- Formulating a hypothesis test
- Transforming a question into a null hypothesis
- Dependent and Independent variables
- The idea of ANOVA as to perform statistical analysis
- The terminology of factors and levels
- Two types of designs when people are involved (within vs between)
- Power to predict required number of participants

What to remember from today

- Have I scared you? I hope so.
- Experimental design is complex
- Requires in-depth planning
- Poorly designed experiments lead to:
    - poor or unpublishable results
    - wasting participant time
    - wasting your time
    - failing to pass ethics committees

What to remember from today

- Statistical analysis is important
- Requires pre-planning so you collect the right data
- Get help, however:
  - You still need a basic knowledge of statistics
    - e.g. hypothesis testing
  - Recommended: http://onlinestatbook.com/
  - Performing the statistical analysis will take longer than you think!

Some potentially useful resources

- Introduction to research statistic course http://onlinestatbook.com/
- Power calculation software: Piface, http://www.stat.uiowa.edu/~rlenth/Power/
- Statistic packages available free via uni: SPSS, Minitab
http://www.unisa.edu.au/ists/staff/purchasing/software/Licensing/SoftwareLicensing.asp

Web **Images** Maps News Shopping Mail

Search Eiffel Tower Search Images Search the Web Advanced Image Search Preferences

Images Showing: All image sizes Any content Results **1 - 12** of about **50,204,000** (0.97 seconds)

Back to index page

I have finished this task. I have found six, or as many images as I can.

You have saved **0/6** images for this topic.

Topic: *Eiffel Tower*

Description: Relevant images will include any photo for which the real Eiffel Tower appears. The whole tower need not appear and it does not have to be the major focus. Toys, replicas, drawings, paintings etc. are not relevant.




Running an experiment

With an online system example

Demo of the system | http://sl.etc

Running an anonymous online experiment

- Large amount of testing
- Embedded sites breaking out of frames
- User interface (look and feel)
- Support for returning user
- Recording of prize details in an anonymous fashion
- Regular backups (hot or close to)

Running an anonymous online experiment

- Overwhelming response - however, many who didn't complete, need to balance respondents

- Securing the system (particularly if release to computer science students)
  - ID encoding
  - Secure web forms (escaping input, particularly SQL)
  - etc.

- Graceful and timely shutdown of study

- Concurrent users and time-stamps as unique IDs in databases (not so good)
  - SQL error, not so good

- Hot testing for small updates (you do need test users :-))

Evaluating the experiment

Evaluating the experiment

- Step 3 should be easy right..... (experiences from MSN experiment)
  - If you've set the experiment up correctly
  - Chosen the correct statistical analysis
  - Recorded to correct data for the chosen analysis
  - Chosen the correct questions to ask to answer your question

6. Results: Human Evaluation

Method	Average coherence of clusters
Method 0	0.94779547
Method 1	0.86140836
Method 2	0.90143668
Method 3	0.91061948
Method 4	0.91401892

Evaluating the experiment

- Its just some simple statistics, it'll only take a week....
  - Recall the idea of statistical significance
  - Simple aggregates etc do not tell us anything about the general case

- Real statistics take time
  - Particularly if you are not familiar with the methods or statistical packages
  - Have partially complete users you potentially want to consider

- Data cleaning (removing test users, dealing with partially complete users)

Evaluating the experiment

- Interpreting the statistics and explaining “why”

Results cont.

System precision: 16.67%

Category	P(U)	P(G)
Specific of scene (unknown)	0.836	0.587
Specific of object (unknown)	0.804	0.610
Specific of scene (known)	0.859	0.738
Generic of scene	0.902	0.752
Generic of object	0.886	0.835
Specific of object (known)	0.890	0.836

Table: Mean click-through relevance proportions

$$P(U) = \frac{\textit{total "saved"}}{\textit{total clicks}} \quad P(G) = \frac{\textit{total relevant}}{\textit{total clicks}}$$

What to remember from today

- Statistical analysis is important and requires pre-planning so you collect the right data
- Get help, however:
- You still need a basic knowledge of statistics, e.g. hypothesis testing
- Performing the statistical analysis will take longer than you think!
- Introduction to research statistic course http://onlinestatbook.com/
- You can not waste participants time
- Experiments/studies must be very well thought through
- Evaluation is more time consuming than you think, esp. 1st time
- Statistics packages help, but are require correct use and the **results require correct interpretation**
- Finding correct explanations can take time

Some potentially useful resources

- Introduction to research statistic course http://onlinestatbook.com/

- Power calculation software: Piface, http://www.stat.uiowa.edu/~rlenth/Power/

- Statistic packages available free via uni: SPSS, Minitab http://www.unisa.edu.au/ists/staff/purchasing/software/Licensing/SoftwareLicensing.asp