Social Sampling

Marcel Mohler
mohlerm@student.ethz.ch

April 4, 2015

Abstract

This report provides the reader with an overview of the concept of *social sampling*. Social sampling is a class of methods where informants in a poll answer with a summary of their friends cumulative responses.

1 Introduction

Polling is a widely used method to achieve a public opinion on a certain subject gained from a sample. For example, one could think of polls before *Bundesrat* elections to estimate the outcome or polls to receive student's opinions on their visited lectures. Following the first example, pollers would like to ask only a subset of the population while still get significant results which party might win the vote. This is commonly known as sampling. The two fundamental quantities of interest in polling are *number of samples* (how many citizens to ask) and the *error rate* (error caused by observing a sample instead of the whole voters). They are related in a way, that if we wish to approximate a fraction within an additive error of ϵ, roughly $O(1/\epsilon^2)$ samples are both necessary and sufficient **[[TODO: cite]]**.

In practice we want to have a low error but still only require as little samples as possible. Since the sampling costs are usually a significant burden, research has proposed several alternatives to reduce the sample size. For instance instead of sampling uniformly one could choose their samples with a built in bias. However this method is vulnerable to introduce a so called *systematic bias* which can lead to systematic errors in the outcome.

Another, recently popular approach is to use "expectation polling", where voters are asked about their expectation about the outcome of the poll in contrast to the more classical "intent polling" (where voters are asked about their polling intent). In [3] David Rothschild and Justin Wolvers explored the value of expecation polling. Considering the focus of press and pollsters on intent polling they call it conventional wisdom that these type of polls are more accurate. However, executed on the example of Presidential Electoral College races they provide robust evidence that expectation based polling

yields to more accurate predicition of election outcomes. Unfortunately they fall short on providing any theoretical guarantees on either the sampling bias or sampling error.

In this paper we present an extended idea of "expectation polling" called "social sampling", suggested and described by Anirban Dasgupta, Ravi Kumar and D Sivakumar in [1]. Their idea is to reduce the sampling size by asking a member of a social network about the opinion of their friends. Even though this means that the actual structure of the network plays a role when analysing the error, they assume a saving in number of samples by a factor of around d, the average amount of friends of a member in a social network.

What follows are various social sampling strategies with the aim to observe sampling bias and sampling error. We will also provide precise characterizations of these errors and how they behave applied to large, real-world graphs.

2 Algorithms

Note that graph $G = (V, E)$ describes a social network with V denoting the set of nodes and E the set of edges. Further, $|V| = n$, $|E| = m$ and u, v are nodes of G. The set of neighbors of u is denoted by $N(u)$ and the degree of node u by $d_u = |N(u)|$. We will always assume $d_u \geq 1 \forall u$. Additionally, $f : V \to \{0, 1\}$ is a binary function which takes a node as input and outputs whether this node satisfies a certain property or not and $\bar{f} = \frac{1}{n}|\{u \mid f(u) = 1\}|$ is the fraction we wish to estimate. The algorithms use S as the set of nodes called "sample" with $|S| = r$ (sample size).

Our goal is to estimate the portion of nodes $v \in C$ such that $f(v) = 1$. We introduce the concept of a set of algorithms called sampler which is defined as follows:

Definition 2.1 (sampler) *A sampler $\hat{f}(n, \epsilon, \delta)$ is a randomized algorithm with input r (sample size), ϵ (sampler accuracy), δ (sampler error) and function f which outputs $\hat{f} = \frac{1}{r}\sum_{u \in S} f(u)$ with probability $1-\delta$ and $|\hat{f} - \bar{f} < \epsilon|$. Namely,*

$$\boldsymbol{Pr}[|\hat{f}(n, \epsilon, \delta) - \bar{f}| > \epsilon] < \delta$$

Note, that this definition of a sampler is covered in the definition of a statistical estimator. Therefore we can analyze our sampler for estimator properties like bias and we will use both expressions equivalently.

We will start by looking at an intuitive approach for estimating $\bar{f}$ where we sample a set of nodes and poll each node u to test if $f(u) = 1$. The algorithm will return the fraction of nodes that satisfy the condition and we call this approach the `Naive` estimator [2].

Note that in practice the drawing can be done by simply including a node u with probability $\frac{1}{n}$.

Alg. 1 Naive size estimator(G, f, r)

Input: Graph $G = (V, E)$, function $f : V \to \{0, 1\}$, sample size r
Output: $\hat{f} = \frac{1}{r} \sum_{u \in S} f(u)$

```
1    begin
2        initialize f* with 0
3        randomly draw a set S with r samples from V with replacement
4        for each element u ⊂ S do
5            f* = f* + f(u)
6        end
7    return f̂ = f*/r
8    end
```

Theorem 2.1 *The* `Naive` *estimator with accuracy* ϵ *and confidence* $1 - \delta$ *requires* $O(\frac{1}{\epsilon^2})$ *sample nodes*

Proof: Recall that we want to poll r uniformly chosen people independently and with replacement. The true fraction we want to approximate is $\bar{f}$. Let F_u be the random variable for $f(u) = 1$.

It follows that $F_u \sim Bernoulli(\bar{f})$ and $F_1, F_2, \ldots, F_r$ are independent. Our `Naive` estimator is defined with $F = \sum_{u=1}^{r} F_u$ as $\hat{F} = F/r$. We want our estimate to have accuracy ϵ and confidence $1 - \delta$:

$$\mathbf{Pr}[|\hat{F} - \bar{f}| \leq \epsilon] \quad \geq \quad 1 - \delta$$

Since $F \sim Binomial(r, \bar{f})$, $\mathbf{E}[F] = r\bar{f}$ follows by definition. Using the Chernoff Bounds the following inequality holds for any $\gamma \geq 0$

$$\mathbf{Pr}[|F - r\bar{f}| \geq \gamma r\bar{f}] \quad \leq \quad 2\exp(-\frac{\gamma^2}{2+\gamma} \cdot r\bar{f})$$

$$\Leftrightarrow \quad \mathbf{Pr}[|\hat{F} - \bar{f}| \geq \bar{f}\gamma] \quad \leq \quad 2\exp(-\frac{\gamma^2}{2+\gamma} \cdot r\bar{f})$$

To achieve $\hat{F}$ bounded by ϵ we set $\epsilon = \gamma\bar{f}$, so $\gamma = \epsilon/\bar{f}$. Inserted in the formula above:

$$\mathbf{Pr}[|\hat{F} - \bar{f}| \geq \epsilon] \quad \leq \quad 2\exp(-\frac{\epsilon^2/\bar{f}^2}{2+\epsilon/\bar{f}} \cdot r\bar{f}) \quad = \quad 2\exp(-\frac{\epsilon^2}{2\bar{f}+\epsilon} \cdot r)$$

Since the largest possible value of $\bar{f}$ is 1, because all nodes u have $f(u) = 1$, the term in $\exp(\cdot)$ is an upper bound

$$\frac{\epsilon^2}{2\bar{f}+\epsilon} \quad \geq \quad \frac{\epsilon^2}{2+\epsilon}$$

and therefore

$$\mathbf{Pr}[|\hat{F} - \bar{f}| \geq \epsilon] \quad \leq \quad 2\exp(-\frac{\epsilon^2}{2+\epsilon} \cdot r)$$

Remember we want the confidence of the estimator bounded by $1 - \delta$, this means

$$\begin{aligned} & 2\exp(-\frac{\epsilon^2}{2+\epsilon} \cdot r) \leq \delta \\ \Leftrightarrow \quad & \exp(\frac{\epsilon^2}{2+\epsilon} \cdot r) \quad \geq \quad \frac{2}{\delta} \\ \Leftrightarrow \quad & \frac{\epsilon^2}{2+\epsilon} \cdot r \quad \geq \quad \ln\frac{2}{\delta} \\ \Leftrightarrow \quad & r \quad \geq \quad \frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta} \end{aligned}$$

Since $\frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta}$ lies within $O(\epsilon^2)$ this concludes the proof. □

As briefly mentioned in the introduction the network structure is of great importance to the sampling method. We will illustrate this on a small example. Consider a graph structure as shown in Figure 1 where a single `RED` node is connected to several `BLUE` nodes. Even though the portion of `RED` nodes is way smaller, using the `Naive` estimator the result is very biased towards the decision of RED.

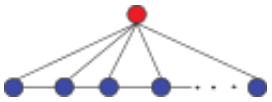

Figure 1: neighbor scaling

To circumvent this behaviour we introduce an improved sampler called `Ideal` estimator that divides each $f(v)$ by d_v. The pseudocode is shown in Alg. 2.

3 Results

References

[1] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 235–243. ACM, 2012.

[2] O. Goldreich. A sample of samplers: A computational perspective on sampling. *def*, 1:2n, 1997.

Alg. 2 Ideal size estimator(G, f, r)

Input: Graph $G = (V, E)$, function $f : V \to \{0, 1\}$, sample size r, distribution p
Output: $\hat{f} = \frac{1}{r} \sum_{u \in S} f(u)$

```
1    begin
2        initialize f* with 0
3        randomly draw a set S with r samples from V with replacement
4        for each element u ⊂ S do
5            f* = f* + f(u)
6        end
7    return f̂ = f*/r
8    end
```

[3] D. Rothschild. Forecasting elections comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73(5):895–916, 2009.