Parameters evaluation of abstract web objects

Alessandro Pezzé, 182501
alessandro.pezze@studenti.unitn.it

***Abstract*—Starting from 10 sets of web objects, possibly representing advertisements linked to browsing generic web pages, and their associated metrics, which are their number of downloads and their revenue generated in a way we are not concerned with, analyse each set in order to find out the values of hidden parameters which rules how the system works, estimate the confidence interval about one of these parameters and plot significant data.**

I. Introduction

A system of revenue per download is defined by a known function with some hidden parameters. Those parameters are inferred by analysing the data and applying to it common formulas and calculations.

II. Fundamentals

We are presented with ten sets of web objects, each set comprises 2400 objects which are defined by two metrics which are: $x_i(t)$ the number of downloads of the i-th object at time t, $y_i(t)$ the real revenue gained by the i-th object at time t. $x_i(t)$ is a non-decreasing cumulative function affected by noise following $U(0, D_i^{max})$, which is a Uniform distribution between 0 and D_i^{max} which has to be inferred. Moreover, it is safe to say that we can derive the function of the estimated revenues from the two metrics mentioned above so that a revenue can be considered as $y_i(x_i)$. Such function is known a priori to be linear and we also know that there is some kind of noise which affects the real values of the revenue in a form of Gaussian distribution, that Gaussian noise is moreover known to be independent to the observed object. The estimated revenue so can be seen as the Equation (1)

$$r_i(t) = \alpha_i d_i(t-1) + \alpha_i U(0, D_i^{max}) + N(0, \sigma^2) \qquad (1)$$

Which says that the calculated revenue is governed by three unknown parameters which are α_i, D_i^{max} and σ^2. D_i^{max} will be estimated statistically by calculating the probability that a never happened event could occur after n iterations in a space of k variables (all the possible download increments from 0 to D_i^{max}) using the probability formula as shown in Equation (2)

$$p = \left(\frac{k-1}{k}\right)^n \qquad (2)$$

α_i can be considered from Equation (1) as the slope of the line derived from that equation and thus will be estimated by fitting the revenues over the downloads while minimizing the mean square error between real data points. σ^2 is the variance of the Gaussian distribution found in equation Equation (1) that can be approximated to a t-student distribution when the dataset is smaller than 30 individuals. The sample variance of a t-student distribution is equation Equation (3) which will thus be used to calculate this value.

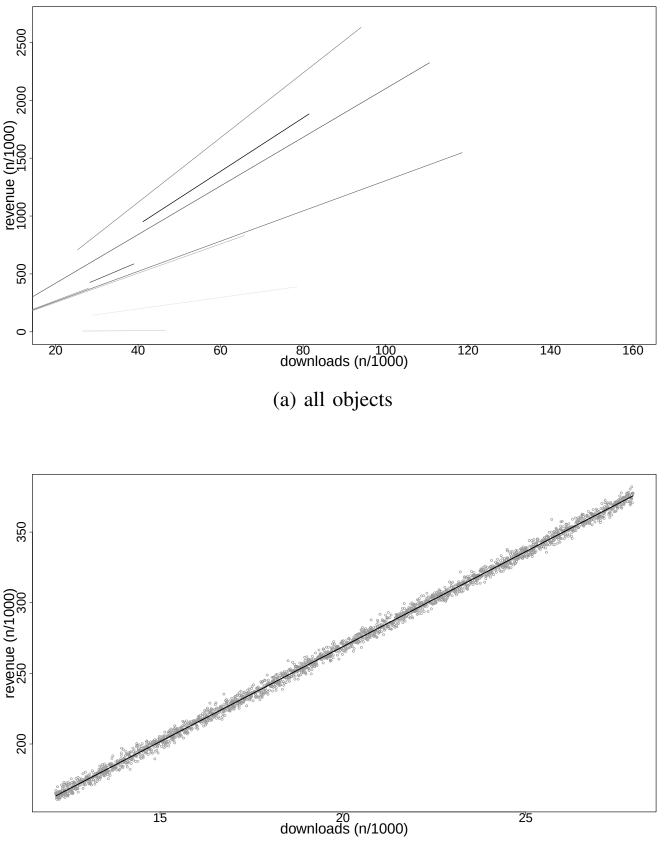
$$s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (3)$$

Moreover, the confidence interval of σ^2 must be calculated using the known mean's confidence interval for t-student distribution with a population on n values and $n-1$ degrees of freedom as depicted in Equation (4)

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}}, \quad \bar{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}}\right] \qquad (4)$$

III. Implementation

R was used to do all the necessary calculations to derive the unknown parameters since it provides a large group of primitives to aggregate data and permits to apply known built-in functions on it. Two steps of calculations were designed because of the conformation of the unknown parameters. The first two mentioned parameters are inherent to the sets analysed so in the end of the computation it will result in ten of both those parameters, one for each given set, the computation of these values is done firstly. The last parameter, σ^2, is known to be the same all over the ten sets so it will be calculated taking in consideration all the data in a second step of calculations. In order to start the first step of computations the data is opportunely divided into the ten sets. D_i^{max}, the number of maximum downloads per unit time, is derived by subtracting each pair of time-contiguous number of downloads, Δd, and taking the maximum value among the differences. To confirm the correctness of the result, Equation (2) was used to calculate the probability that $D_i^{max}+1$ could be the largest number of maximum downloads per unit time instead of the calculated D_i^{max}. α_i is the slope of the line that is formed by revenues over the number of downloads. Since these two metrics are not time continuous but a set of single pairs, linear fitting was used to derive the slope and the intercept of the hypothetical line that minimizes the error with the actual points by using the least square error function. The resulted line's slope is then α_i. To reinforce the calculated result, the Pearson correlation between the two metrics, revenues and number of downloads, is calculated, showing that the two are closely tightened and follow a line path. The estimation of σ^2, which is the second parameter of the Gaussian distribution which affects the real revenues, is calculated by looking at the real revenues data coming from the 10 datasets, each splitted in two in order to have more samples,



(a) all objects

(b) object 7

Figure 1. The ten sets derived lines, a single set with scatter real data points and the linear fit

and the theoretical revenue values which can be found on the fitted line derived while estimating α_i. In order to have more samples and thus effectively use a t-student distribution instead of a Gaussian distribution, the 10 initial sets were divided in 20 disjoint sets. For each value of the revenues in the dataset was calculated the difference with the theoretical corresponding one and then with the twenty sets derived (10 sets divided in two) were formed their standard deviation was calculated. In order to calculate the confidence interval for σ^2, its mean and its variance were calculated and fed, together with the appropriate $z_{\frac{\alpha}{2}}$, into Equation (4) giving us the confidence intervals for the desired confidence level required (0.9, 0.95, 0.98, 0.99, 0.998, 0.999).

IV. Evaluation

Table I shows for each of the ten sets of web objects the calculated maximum number of downloads, D_i^{max}, which rules the uniform distribution they follow. Through Equation (2) was calculated the error probability of the existence of a D_i^{max} equal to the calculated $D_i^{max}+1$. The error ranges around 10^{-13} and is considerably low to calculate the probability that a higher value of $D_i^{max}+1$ could be the real D_i^{max} value. Table I shows also the related $\alpha_i$of the ten sets with the Pearson correlation between the revenues and the number of downloads. The different α_is can be seen across the datasets by looking at Figure 1(a), in particular at the different values of steepness of the ten lines (one for each set). Figure 1(b) analyses one single set and shows the theoretical fitted line derived from the least square error function been applied to the data points representing the real values of the revenues over the downloads. Table II shows the found σ^2 and its confidence intervals for different confidence levels with $n = 20$ degrees of freedom (similar results with $n = 10$ can be found in Table III), it can be seen that even for the confidence level of 0.999 the σ^2 confidence interval is very close to the σ^2 calculated value with a drift ranging from -5.4 to $+5.4$. Figure 2 shows the noise affecting the real values of the revenues where bars shows the real Gaussian noise distribution

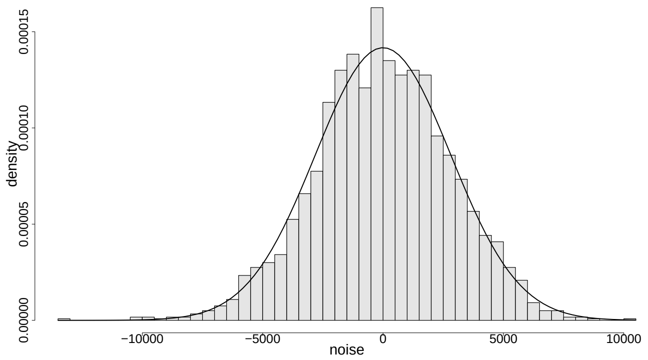


Figure 2. Gaussian distribution obtained from object 6. The bars describe the real data grouped in bins large 280 elements, the curved line describe the theoretical line derived from calculations

Table I
D_i^{max} AND α_i RESULTS

object	D_i^{max}	error probability	α_i	correlation
1	82	$6.290477 \cdot 10^{-31}$	23.094	0.999 946
2	80	$4.99975 \cdot 10^{-100}$	15.043	0.998 311
3	79	$4.6774 \cdot 10^{-13}$	20.999	0.999 990
4	59	$5.450097 \cdot 10^{-12}$	13.046	0.999 977
5	13	$3.084084 \cdot 10^{-18}$	27.947	0.999 988
6	86	$1.312886 \cdot 10^{-72}$	13.451	0.998 952
7	99	$1.261195 \cdot 10^{-23}$	12.646	0.999 899
8	24	$4.664994 \cdot 10^{-57}$	0.267 37	0.492 021
9	84	$1.116488 \cdot 10^{-24}$	4.9389	0.999 267
10	97	$1.927963 \cdot 10^{-123}$	7.8592	0.990 403

Table II
σ^2 RESULTS ($n = 20$)

σ^2	σ^2 interval	confidence
2755.3	[2752.6, 2758.0]	0.900
	[2752.1, 2758.5]	0.950
	[2751.5, 2759.1]	0.980
	[2751.1, 2759.6]	0.990
	[2750.2, 2760.4]	0.998
	[2749.9, 2760.7]	0.999

Table III
σ^2 RESULTS ($n = 10$)

σ^2	σ^2 interval	confidence
2755.1	[2751.9, 2758.3]	0.900
	[2751.3, 2758.9]	0.950
	[2750.6, 2759.6]	0.980
	[2750.1, 2760.1]	0.990
	[2749.1, 2761.0]	0.998
	[2748.7, 2761.4]	0.999

and the curved line the calculated Gaussian noise line with σ^2 equal to 2755.3. It can be seen that the bars and the curved line follows the same trend.

V. Conclusion

A simple model of ad revenues was provided with hard constraints like the linear correlation between the revenues and the number of downloads, and with some unknown parameters. Such a model could have been studied thanks to those hard constraints which brought significant simpler deductions about how the model works. In the end the system could be studied as a system in which the revenues of a web object is dependent on how well that object performed in the past so that, if the number of downloads is steady (Uniform), the revenue grows linearly; while if suddenly the number of downloads drops nearly to zero the revenue will maintain constant without ever decreasing. This conformation of ad revenues system could be considered as a system in which a user does not buy a single product when visiting the proposed ads, but instead a subscription which generates a fixed amount of money every fixed time length in favor of the subscripstion seller.