ZecoByte

SQCT - Survey Quality Control Tool

technical presentation

Skeppargatan 8
114 52 Stockholm
08 410 216 40
www.zecobyte.com

Abstract

This is a technical presentation of survey quality control offered by ZB. Topics in this document present idealized demonstration models, example discussion, notation and insights into methodology.

1 Introduction

Survey quality control is a new branch of diversity in the field of market research. Traditionally surveys have been conducted in person, later by telephone and more recently in the Internet domain.

This move has created opportunities to conduct more comprehensive surveys and reach larger audiences. The anonymity aspect of the survey conducted online has a considerable impact on the obtained results, which broadly follows in the following two categories.

- People volunteer their true opinions, even if these opinions would have been shameful to admit to over the telephone or in person.
- People partake in the survey answering questions dishonestly, without interest, skim-reading, rushing through the survey or otherwise submitting undesired data.

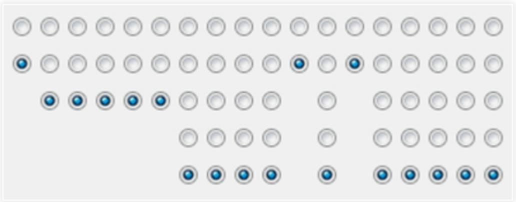

Figure 1: *A sample survey excerpt which demonstrates clearly otiose filled form data, as the user selected the last alternative for each question. It is of no value, and it may have an adverse effect on results obtained, if such statistic was considered as part of the survey.*

ZB's quality control service address the latter point by using mathematical methods herein discussed. Generally the problem of finding such inconsistencies in the data is nontrivial, however ZB is in the process of developing easy to use tools[1], that could be used even by non-professionals.

2 Information Content

Intuitively it's quite obvious that empty content exemplified in Figure 1 should be removed from the data sample. But the general *practical* case of identifying such empty content is nontrivial.

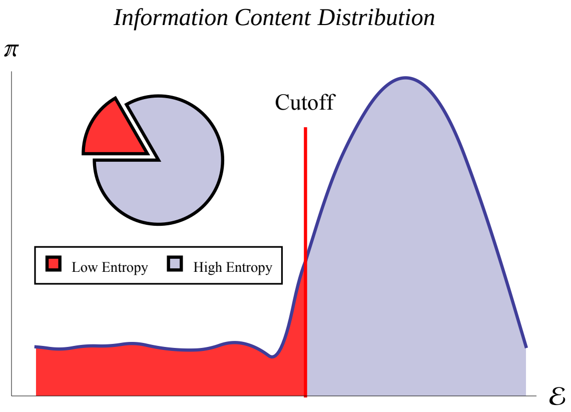

Figure 2: *A schematic representation of entropy distribution function. Here π is probability, and $\mathcal{E}$ is the entropy. The pie chart represents the relative portion of bad data (red) to the good data (blue).*

2.1 Entropy

ZB has developed methods which address this issue. Using combinatorics and information theory it's possible to associate with each individual a number $\mathcal{E}$ which measures the amount of information this individual has provided. We call this number

[1] The project is called *Survey Quality Control Tool*, or *SQCT* for short.

entropy and larger values of $\mathcal{E}$ have more information. Figure 2 provides an insight into a typical distribution of $\mathcal{E}$.

In a practical survey data sample around 10% of individuals provide low entropic content. Upon inspection it becomes clear that these individuals had no intention of answering survey questions, instead completing the survey by answering in repeat patterns. It may not even be the individuals fault, as sometimes surveys customized for a particular demographic reach another; however eliminating such data from the statistical sample is an important first step for further data analysis. *For more information see appendix A.1.*

2.2 Filtering

Entropic Filtering

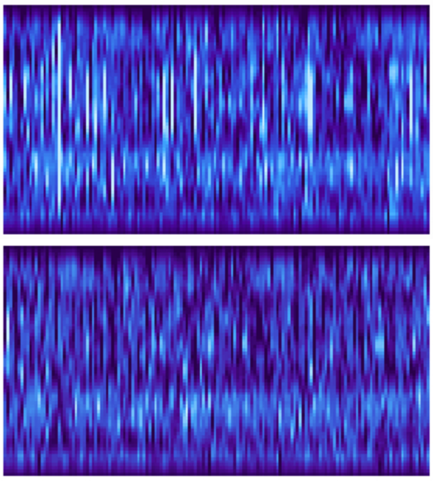

Figure 3: *These two images symbolize the filtering process. Each column represents an individual, while rows portray questions. Lighter portions represent low information content. The image above shows several bright vertical lines. These are people whose answers contribute nothing to the survey. The lower image shows these individuals removed.*

Figure 3 demonstrates filtering in a real data sample. Observe that entropy here is defined locally, in order to help identify poll flat-line effects. The overall effect for a typical poll may resemble Figure 5. This effect is more apparent for longer surveys, and implies less reliable answers toward the end of the survey.

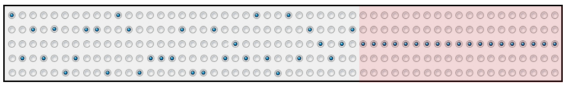

Figure 4: *An example of filled questionnaire, with high information density in the beginning and a sudden flat line more than halfway toward the end. This is indicative of poll fatigue. Questions answered in good faith can still be taken along for further statistical analysis, however.*

The discovered bad data is removed from the sample. This causes a mild disruption to the sample base size, as typically between 5% to 15% of the total population has to be removed. This need not always be the case, however. As Figure 4 would attest, sometimes partial information can be recovered from the questionnaire before the so-called flat-line effect occurs.

Poll Flatline

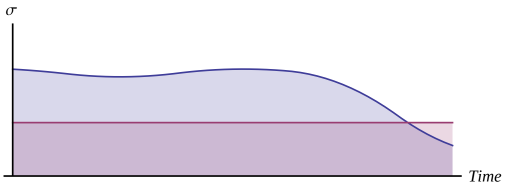

Figure 5: *Overall flat-line indicator helps to identify surveys that take too long time to complete. Additionally conclusions about validity of latter questions may be drawn. Here the schematic red line represents dangerously low level of variance σ, and we see the blue variance dipping below the red line toward the end.*

3 Question Quality Control

Another aspect of quality control is improvement, and in current context that means verifying validity and necessity of asked questions. In this section we consider a sample survey with 20 questions in total.

3.1 Question Dependency Network

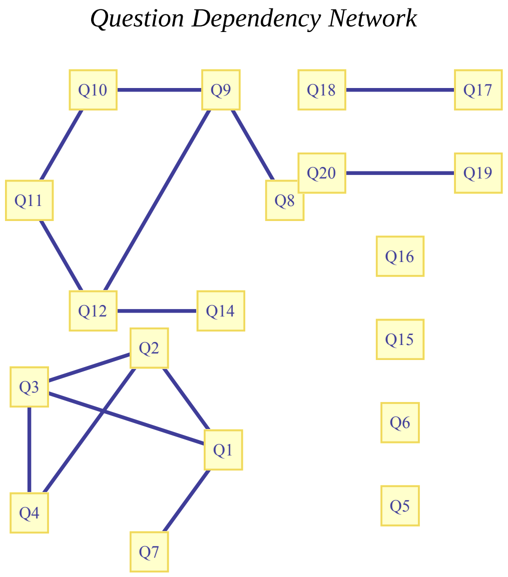

Figure 6: *This graph represents dependencies between questions. Edges between drawn nodes signify a dependence, a lack of such edge means that questions are largely independent.*

By comparing relative change to distribution, based on all possible permutations of questions, it's conceivable to create a question dependency network, as illustrated in Figure 6. This graph helps to identify hidden relationships between questions. Those questions that have many dependencies, such as questions 2, 3 and 4 in the graph above should be considered for removal or reformulation in future survey revisions. This is doubly true if the survey has demonstrated bad flatline fitness.

An example of dependent set of questions could be along the lines of the classic chocolate ice cream example.

1. Do you like chocolate?
2. Do you like ice cream?
3. Do you like chocolate ice cream?

While this seems self evident, often these dependencies go unnoticed in practical surveys. While the ice cream example could even have potential use, survey makers oftentimes ask the same question with slightly different wording.

Eliminating unnecessary questions can create potential space for other questions, or improve the result quality of a survey which takes too much time or is perceived as repetitive and dull. The question dependency network provides means to identify like questions and help improve surveys.

3.2 Identifying Leading Questions.

ZB has developed a metric that helps to identify leading questions. Its simple output is demonstrated in Figure 7. There are several ways to detect leading questions, and here we'll discuss only the most important one - how well do answers to immediate previous questions predict the answers given by the same people to the question being considered.

The answer lies in a predictor functor, which has to be constructed ZB's quality service considers several possible such predictors, and selects the best one. If the question is easily predicted by previous questions, it's either contained within answers to those questions or it's being led. If upon inspection questions seem dissimilar, consider changing their order when re-issuing the survey.

Note that a much better analysis model exists for randomly permutated questions, i.e. questions that don't appear in order to poll takers, assuming the order is stored with the survey. It's then possible to detect leading questions with great accuracy, however this approach isn't often available. For more information see appendix A.2

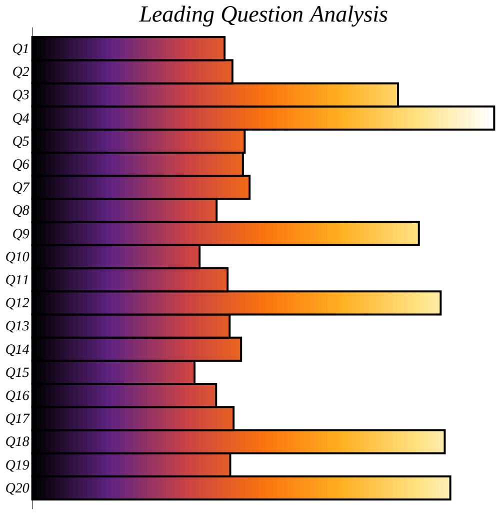

Figure 7: *A sample output of leading question detector. The high values represent the questions that are easily predicted from answers given to previous questions. Here larger values mean that the question is being lead. While this metric isn't fool proof, it offers a simple insight into redundancy and possible detection of leading questions. Some questions with high values should be removed, while others should be moved to a different position in the survey.*

3.3 Other Useful Information

Along with all of the above, the quality report contains rudimentary statistics about the survey data and each question in particular. Some of these are exemplified in this section.

Figure 9 demonstrates agreement measure and agreement type between four different questions. This is part of the quality control suite for two important reasons. Firstly this has to be computed in an intermediate step during entropic coding, and second it gives a good insight into questions at a glance guiding further analysis.

For questions prompting for answers within an interval, distribution inconsistencies are detected and highlighted. Consider for example Figure 8.

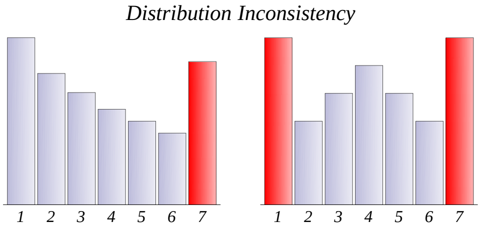

Figure 8: *The quality control report contains information about questions that have inconsistently distributed interval values. Typically this effect is minimized by entropic coding, sometimes however this has to be handled manually.*

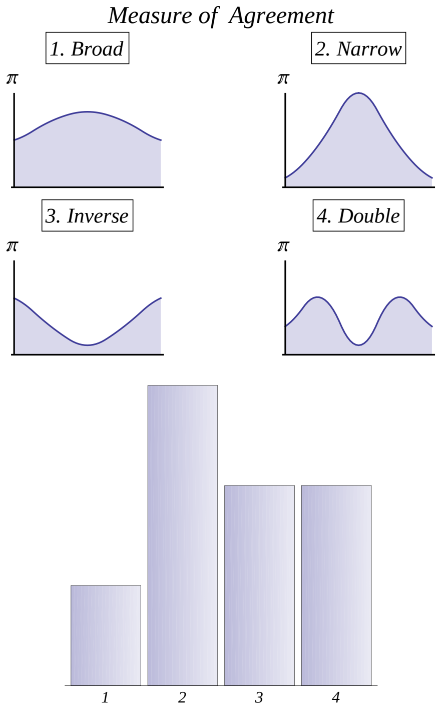

Figure 9: *A simple measure of agreement and the type of agreement is provided alongside each question in the report.*

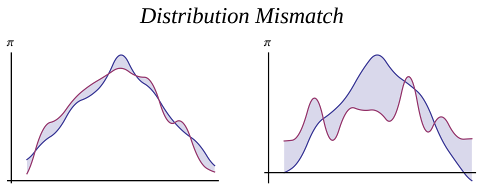

Figure 10: In this example two surveys are compared. The mismatch that occurs in the leftmost image is small compared to that in the rightmost image. This is a solid metric to compare successive surveys or surveys performed on different media.

Such information helps to identify insufficient answer space, and sometimes it's even intended by the creators of the survey. When these inconsistencies become undesirable users should be provided with more options, or the question should be reformulated.

3.4 Comparing Survey Results

Two or more surveys can easily be compared with each other in terms of poll fatigue, information density value, leading questions, inconsistencies, etc. And, as Figure 10 attests, also in terms of results. Distributions are compared and a mismatch coefficient is calculated for each question.

4 Conclusion

Quality control is a self evident necessity when performing online surveys. ZB has a three part approach, which roughly follows into these categories

- **Clean up.** Identify and eliminate otiose answers from the survey. Remove (at least in part) answers that have flatlined.
- **Improve.** Get all the information you need about your questions. Which questions aren't necessary, which are leading other questions and which questions are predictable.
- **Analyze.** Ready made useful statistics about the survey help you get a head start on analysis. Clean data makes analysis go much smoother.

For more information about quality control and other services visit `www.zecobyte.com`

A Appendix

This section provides a technical insight into some topics discussed earlier.

A.1 Shannon Entropy and Entropic Coding

Shannon entropy is a measure of information content in a given message, defined as

$$\mathcal{E} = -\sum_{i \in \mathcal{I}(A)} p_i \log p_i$$

Here, $\mathcal{I}$ is the indexing set on an alphabet A, and p is the probability distribution. However, there's a problem - what is the correct choice for the alphabet A?

It turns out that this isn't a trivial question, and A should be adapted to the problem specification. Given the domain of online surveys, A needs to take into account different data types such as interval questions, categorical questions, special answers, etc.

The clever choice for the alphabet A is made by permutation on predefined and detected alphabetical constructs and operators. The exact configuration of these constructs is ZB's trade secret. The applicable definition of entropy $\mathbf{E}$ thusly becomes

$$\mathbf{E} = \min \left\{ \sum_{i \in I} \min((1 - \mathcal{E}_n)^n, |\hat{A}|) \qquad \forall \hat{A} \in \hat{\mathbf{S}} \right\}$$

Where $\hat{\mathbf{S}}$, is a collection of possible solutions for A. This deceptively simple definition gives rise to ZB's quality control model. More information about entropy and information theory in general can be found at `http://en.wikipedia.org/wiki/Entropy_(information_theory)`.

A.2 On Identifying Leading Questions

Given a set of individuals $\mathbf{S}$ and a question indexing set I, define $\mathcal{F}(I)_n$ as an ordered set of contiguous subsets of I of length n. Let $\mathbf{R}$ be the I-indexed question-range set, then define a E_i as

$$E_i \equiv \frac{1}{\prod_{e \in i \in \mathcal{F}(\mathcal{I})_n} R_e}$$

Then E_i is an a-priori expectation of any independent $i \in \mathrm{UDF}(\max R, 0)$ Following this definition, let $\hat{\mathbf{S}}_m$ be sequences that occur with frequency mE, or

$$\hat{\mathbf{S}}_m \equiv \{ \left\{ s \in S \;\middle|\; \frac{\#\{s \in \mathbf{S}_i\}}{mE_i} \right\} \qquad \forall i \in I \}$$

Let the neighboring set $\mathbf{N}_m$ be a similarly defined covariant of I under R. Then using the long pass entropy filter, define

$$\mathbf{A} = \{ \sum_{i \in I} f(\mathcal{E}_n, |\hat{\mathbf{s}}|) \;\middle|\; \#\hat{\mathbf{n}} > m \qquad \forall \hat{\mathbf{s}} \in \hat{\mathbf{S}} \quad \forall \hat{\mathbf{n}} \in \hat{\mathbf{N}} \}$$

Here $f : [0,1] \times \mathbb{Z}^+ \to \mathbb{R}^+$ is any monotone score function. $\mathbf{A}$ is the score tensor. Higher values mean high covariance and lower entropy. Therefore higher values of $\mathbf{A}$ indicate direct neighbor dependence.

Other predictive functors include neural network a priori training, and histogram a posteriori analysis. The best measure is selected as a final predictor.