Analysing survey data in Public Health

Chris Shaw

12 May 2016

Introduction

Surveys are an essential part of policy making in Public Health. Limited resources have to be allocated to deliver the maximum outcomes. Conducting a survey across a wide spectrum of public health professionals is a valuable way to gather prevailing information and prioritise policy areas to address.

This paper examines effective ways to analyse and present survey data to support the decision making process about areas to concentrate on.

Raw survey data

The appendix shows a simulated survey with 10 questions and 19 responses. The raw data is captured in an excel spreadsheet in a tabulated manner. The topics are the column headings. Each row corresponds to one response. The participant's topic preferences are indicated by a number - 1 being the highest preference. Only three preferences are allowed per reponse in strict order.

The analysis will work on any number of topic colums, responses or preference choices.

Analysis

The first task is to calculate a score for each topic as follows:

$$\sum_{\{n_i\}} (4 - p_i)^2$$

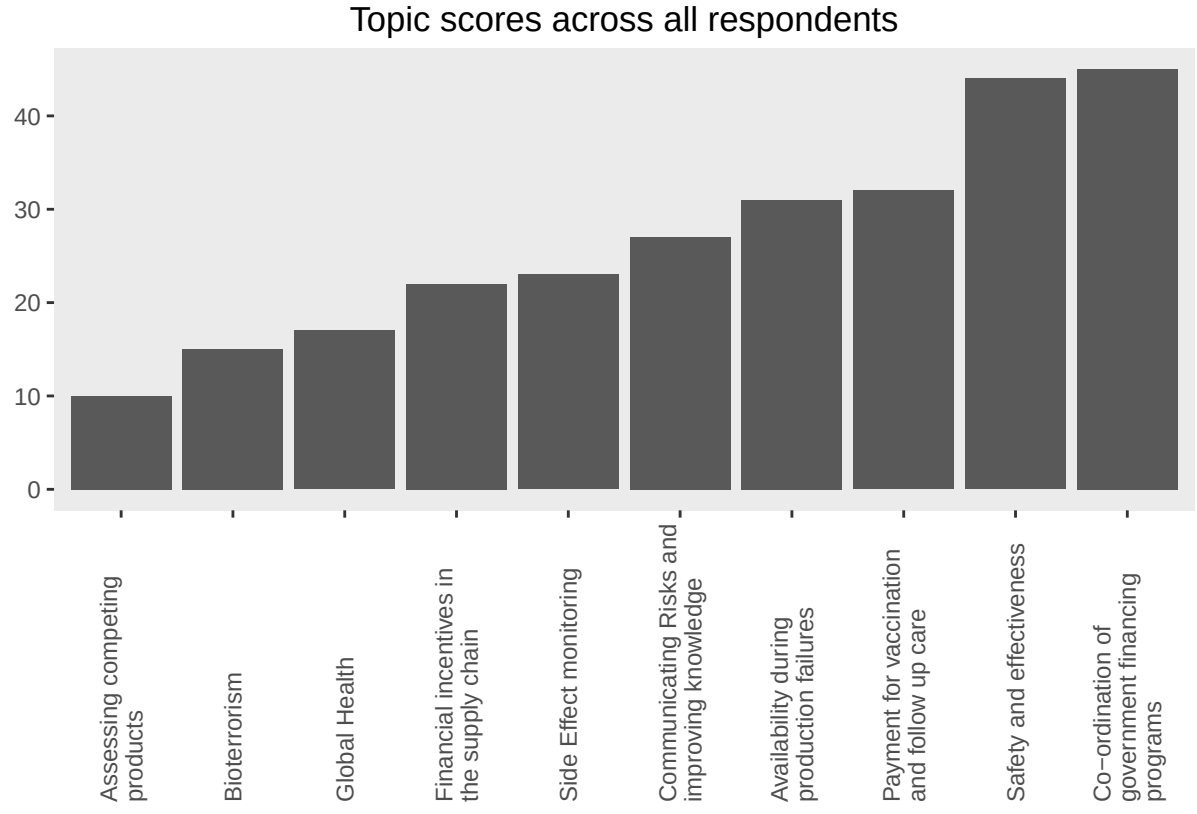
where $\{n_i\}$ is the subset of respondents who voted for that subject and p_i is their individual priority for each topic (1 being the highest and 3 the lowest).

Once the score is calculated for each topic, the top three topics can be determined easily. However, as a second part of the analysis, we want to understand how well individual preferences are served by the top three. In other words, we want a sense of individual satisfaction with the outcome. This can be measured for each respondent by calculating

$$\sum_{i=1}^{i=3} T_i s_i^2 \qquad \text{where } s_i \text{ is the topic score and } T_i = \begin{cases} 1 & \text{if that topic is in the top three.} \\ 0 & \text{otherwise.} \end{cases}$$

Overall topic scores

The plot below shows the overall scores for each survey topic. There are two clear winners at the right hand side and clearly the topics on the far left are not so popular.

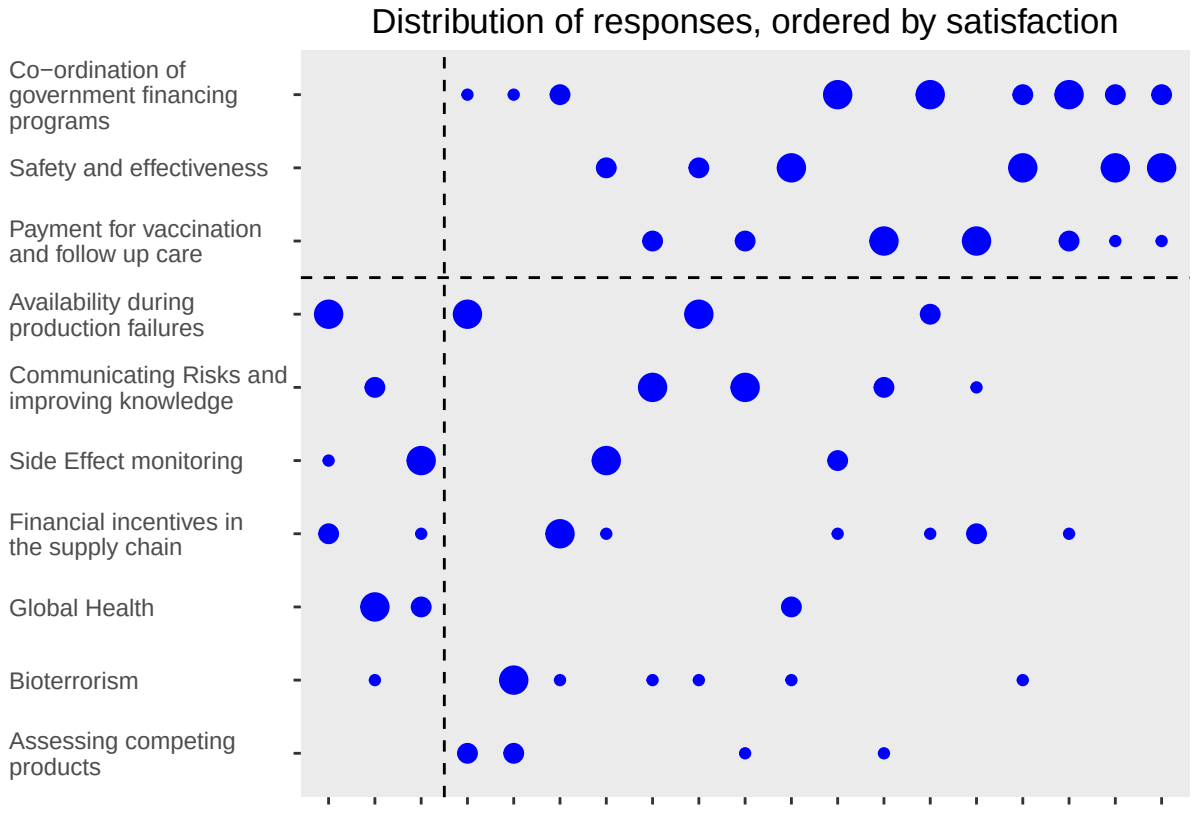


However it is not at all clear whether the fourth most popular topic *Availability during production failures* should be selected over the third - the scores are very close.

Satisfaction

The plot below shows every response across all topics. Each response is shown along the horizontal axis. Every response has three topics selected. The larger the dot, the higher the priority is for that respondent.

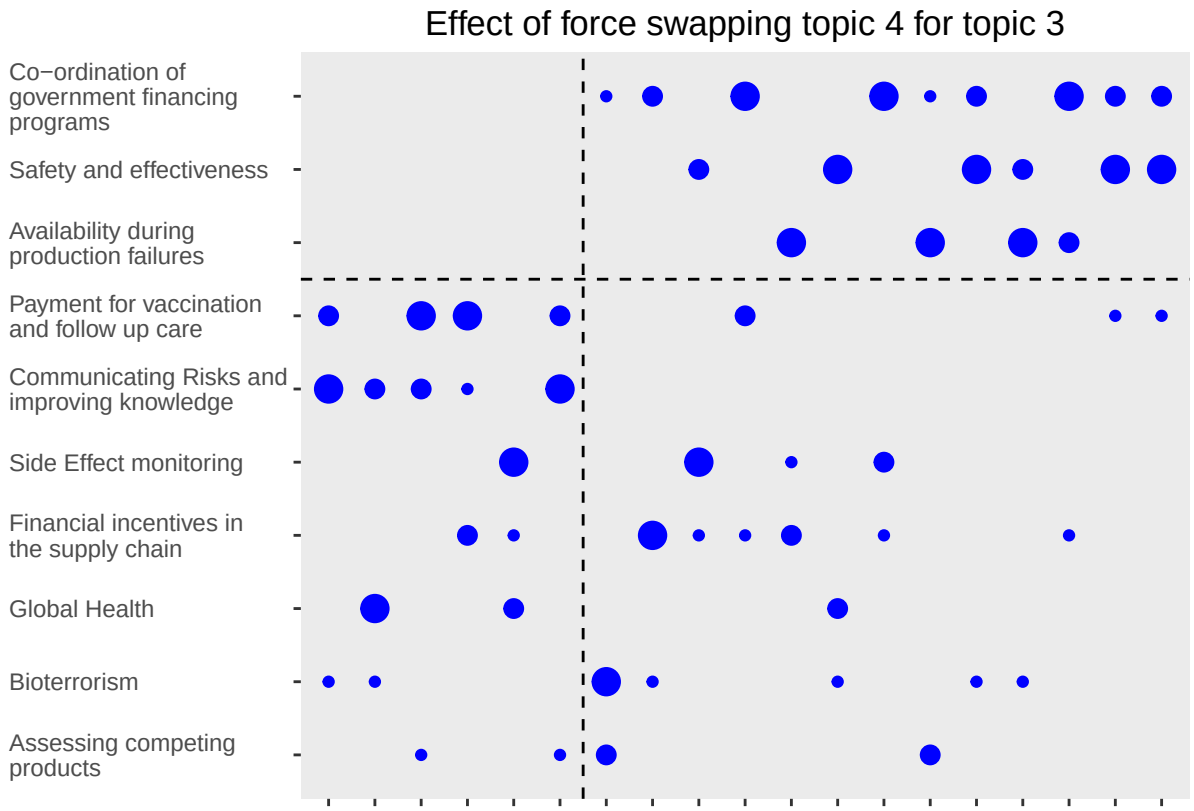
The responses are arranged in order of satisfaction. Those responses to the far right are in general more satisfied with the outcome.



We can see that nine respondents had one of the top three topics as their top priority, and a further five had their second choice. Only three respondents had none of their topics at all selected.

A judgement call can be made whether the score for *Payment for vaccination and follow up care* is really justified over the next one. For example, *Availability during production failures* has three first choice preferences over two for the other (but chosen less overall). The final choice rests with the policy analyst who will use her skill and judgement to make the recommendation.

If a decision is made to swap out topic 4 for topic 3, we can re-plot the data to see the effect of this:

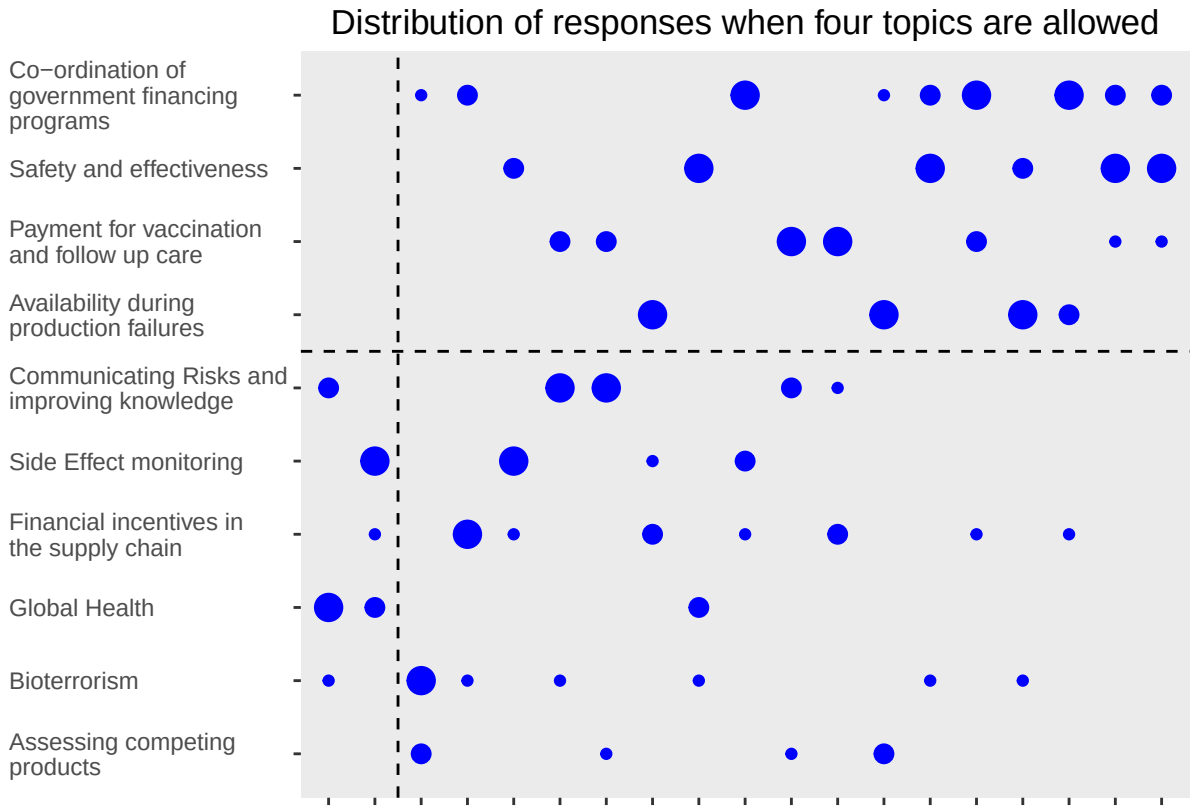


We can see that now 6 people have not had their preferences satisfied, even though 10 first choices are now included. So, although the overall topic scores are close, the algorithm selected the approach which maximises the satisfaction across all respondents.

It would be a brave policy analyst who swims against the tide of data analysis.

Allowing more topics to be selected

Finally, an analysis can be made to see what would happen if four topics were to be selected rather than three. The plot below shows how satisfaction changes by allowing one more topic to be included.



Now 12 people have their first priorities met and only two have none at all. Of course the extra time and cost in allowing four selected topics has to be balanced against the improved satisfaction metrics.

Appendix 1 - Raw Data

The raw data is summarised in the following table. Each respondent provided a priority 1 to 3 for each topic. The subject of each topic is defined in a separate table.

RespondentID	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
1				1	2			3		
2		3			1	2				
3	1		2		3					
4		1			2			3		
5			2		3			1		
6	1	2					3			
7				1				3		2
8		3		2			1			
9						2		3	1	
10		3			1		2			
11								3	2	1
12	1			2				3		
13			3			1			2	
14	2	3			1					
15		2				1			3	
16				1	2	3				
17		3					1			2
18			3			2			1	
19				1	2	3				

topic	subject
Topic1	Availability during production failures
Topic2	Financial incentives in the supply chain
Topic3	Assessing competing products
Topic4	Safety and effectiveness
Topic5	Co-ordination of government financing programs
Topic6	Payment for vaccination and follow up care
Topic7	Side Effect monitoring
Topic8	Bioterrorism
Topic9	Communicating Risks and improving knowledge
Topic10	Global Health

This data is stored in an excel file with two sheets. The data can be edited in the excel sheet and the analysis re-run.

Appendix 2 - Analysis code

In the code below, the variable *number_to_select* refers to how many topics will be chosen to be taken forward for policy analysis. For the bulk of the analysis, this is set at 3, however some effects are shown for a value of 4.

The following code provides the main calculations and analysis:

```
calculate_scores <- function(number_to_select=4, force_swap=NULL) {
        # Get the data from excel and flatten into a data set that can
        # be manipulated more easily
        response_data <<- read.xlsx(xlsfile,1)
        topic_data <<- read.xlsx(xlsfile,2)

        respondent <- names(response_data)[1]
        survey<-melt(response_data, id.vars = respondent, variable.name="topic",
                        na.rm=TRUE, value.name="priority")
        survey$RespondentID <- as.factor(survey$RespondentID)
        number_respondents=nrow(response_data)

        # Calculate the score for each response
        lowest_priority=max(survey$priority)
        survey <- survey %>% mutate(score=(lowest_priority + 1 - priority)^2)

        # Calculate the total score per topic
        topic_scores <- survey %>% group_by(topic) %>%
                        summarise(topic_score=sum(score)) %>%
                        arrange(topic_score)

        # Put a 1 in the weight columm for the last number of topics determined by
        # number_to_select variable.  Other topics have weight zero.  This selects
        # the topics which are the most popular
        topic_scores$weight <- c(rep(0, nrow(topic_scores)-number_to_select),
                                 rep(1,number_to_select))

        if(!is.null(force_swap)){
                        topic_scores$weight[force_swap[1]] <- 0
                        topic_scores$weight[force_swap[2]] <- 1
                        tmp <- topic_scores$topic_score[force_swap[1]]
                        topic_scores$topic_score[force_swap[1]] <- topic_scores$topic_score[force_swap[2]]
                        topic_scores$topic_score[force_swap[2]] <- tmp
        }

        # Add the topic descriptions
        topic_scores <- merge(topic_scores, topic_data, by="topic")
        topic_scores <- arrange(topic_scores, topic_score)

        # Mix the weights into the survey data
        survey <- merge(survey, topic_scores, by="topic")

        # Calculate the satisfaction for each respondent against the selected most
        # popular topics
        survey <- survey %>% mutate(satisfaction=score^2*weight)
```

```
        resp <- survey %>% group_by(RespondentID) %>%
                summarise(satisfaction=sum(satisfaction)) %>%
                arrange(satisfaction)

        resp<<-resp
        topic_scores<<-topic_scores
        survey<<-survey
}
```

The bar plot showing the overall topic scores is produced with the following:

```
plot_topic_score <- function(){
        ggplot(topic_scores, aes(x=topic, y=topic_score)) +
        geom_bar(stat="identity") +
        scale_x_discrete(limits=topic_scores$topic,
                         labels=ylabels) +
        ylab("") + xlab("") +
        theme(panel.grid.minor=element_blank(),
              panel.grid.major=element_blank(),
              axis.text.x = element_text(angle = 90,  hjust=0, vjust=0.5))
}
```

The plot of each respondent's preference, ordered by satisfaction is:

```
plot_responses <- function (number_to_select) {

ggplot(survey, aes(x=RespondentID, y=topic)) +
        geom_point(size=sqrt(survey$score)*1.5, colour="blue") +
        scale_y_discrete(limits=topic_scores$topic,
                         labels=ylabels) +
        ylab("") + xlab("") +
        scale_x_discrete(limits=resp$RespondentID) +
        theme(panel.grid.minor=element_blank(),
              panel.grid.major=element_blank(),
              axis.text.x = element_blank(),
              axis.text.y = element_text(hjust=0)) +
        geom_hline(yintercept = nrow(topic_scores) - number_to_select + 0.5,
                   linetype="dashed") +
        geom_vline(xintercept = nrow(resp[resp$satisfaction==0,])+0.5,
                   linetype="dashed")
}
```