Help! I asked the wrong questions for my typing tool

W. Michael Conklin – Principal Consultant 56Stats LLC

Abstract

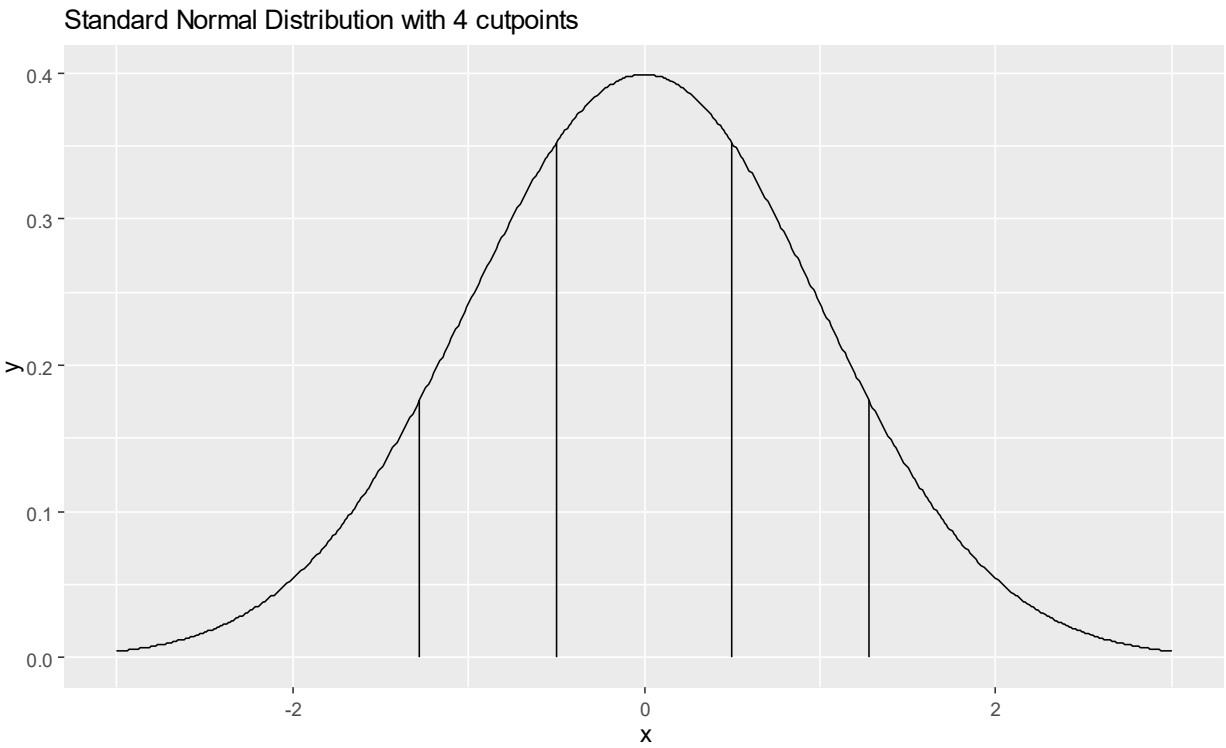
In this paper we explore methods for mapping elicited rating scale responses in surveys to underlying latent variables. This approach makes it possible to make survey answers that are from completely different surveys compatible for modeling by ensuring they come from the same latent variables. This approach has applications for adjusting tracking study data collected by different vendors as well as for utilizing questions from different surveys in segmentation typing tools.

Latent Variables and Survey Responses

Latent variables, in statistics, are variables that are hidden and can only be observed indirectly through some mathematical model from other observed variables[1] This can make them quite useful in our case. Consider a concept like "affordability" or "price sensitivity". We can conceive of a latent variable in the population that represents people's attitudes about "affordability". We can assume that this latent variable is normally distributed in the population, so we can treat it as a standard normal variable with a mean of 0 and a standard deviation of 1.

We can define a function that maps values of this latent variable to people's responses to a survey question about the underlying concept. Suppose, the survey question has 5 categories of response, from Very Affordable to Very Unaffordable. We can define a function that maps the latent variable to specific categories of the survey response by defining cut-points at specific values of the standard normal distribution. For a five point scale we need to define 4 cut-points that divide the normal distribution into 5 specific categories.

Suppose we define our 4 cut-points as -1.28, -0.5, 0.5, and 1.28.



The four cut-points divide the normal distribution into 5 areas. In this case, we would assume that anyone who has value on the latent standard normal distribution less than -1.28 would give a 1 (Very Unaffordable) on the survey. Anyone with a value between -1.28 and -.5 would give a value of 2 (Somewhat Unaffordable) on the survey. Given these cut-points we would expect that in the survey we would have approximately 10% of the population giving a 1 or Very Unaffordable rating in the survey because 10% of the area under the standard normal curve is at a value less than -1.28. The table below shows the mapping for the entire 5 point scale using the four cut-points we outlined above.

Example Mapping of Standard Normal to Survey Categories

Survey Value	Lower Value	Upper Value	Std Normal Area
Very Unaffordable	-Infinity	−1.28	10%
Somewhat Unaffordable	−1.28	−0.5	21%
Neutral	−0.5	0.5	38%
Somewhat Affordable	0.5	1.28	21%
Very Affordable	1.28	Infinity	10%

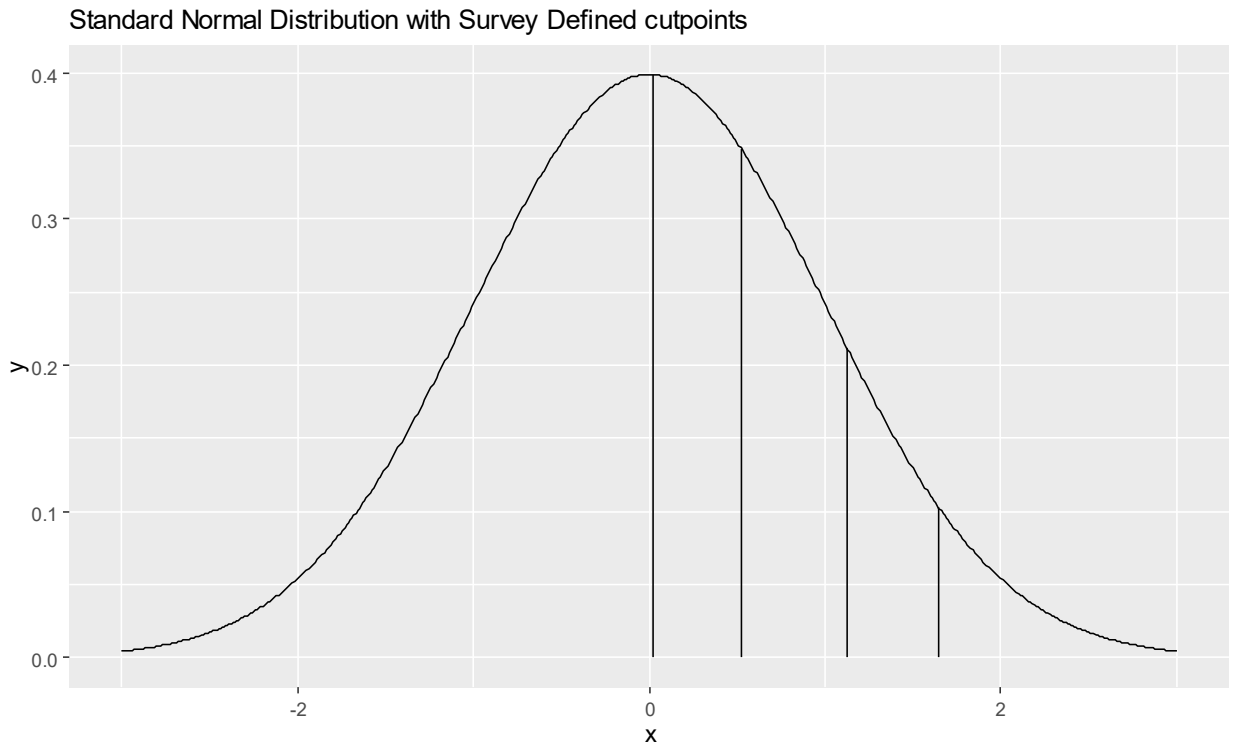
Of course we cannot simply arbitrarily define the four cut-points. What we want to do is map the observed survey responses to the unobserved latent variable. To do this we simply run the process in reverse. We find the estimate of the percent of the population that gives the Very

Unaffordable answer on the survey and calculate what the cutpoint would have to be to achieve that percentage of the population in that same category on the latent variable.

In an example we will use later in this paper, on a 5 point scale 51% of respondents to a survey had a response of 1 on a 5 point scale, 19% had a 2, 17% had a 3, 8% had a 4 and 5% had a 5. To calculate 4 cut-points we can use the following R code.

```
Cutpts<-function(pcts){
    Ptpct<-cumsum(pcts)
     Cutpts<-qnorm(c(Ptpct))
     return(Cutpts[-length(Cutpts)]) }
```

This function takes a vector of percentages and returns the cut-points needed. In this case the four cut-points are 0.02506891, 0.52440051, 1.12639113, 1.64485363.



Applications of the Latent Variable to Survey Response Mapping

The first application of this mapping that I encountered was in the arena of tracking studies. Typically, we would observe an announcement that there was "good news". Our firm had won a contract to take over a tracking study currently being performed by another vendor. In addition, we had promised the client that all of the trends they had historically observed in the tracking

study would be preserved. This last promise, to say the least, is a bit problematic. Anyone with experience in Marketing Research knows that many factors, such as, questionnaire layout, interviewing mode, sample source, and slight changes in wording can have dramatic impact on the way that respondents answer a question. The result is that we cannot simply promise to preserve trends because so many parts of the study will be different. If we simply start with the next wave of the tracking study, we will be unable to determine if the differences from the previous wave are due to changes in the true feelings of respondents or are just due to the changes in data collection.

The best way to understand the differences is expensive, but necessary. The previous vendor and the new vendor need to run a wave of the tracking study simultaneously. Differences in results then have to be due to the differences in the way that the data is collected (sample source, questionnaire layout, interviewing mode, wording changes etc.) as opposed to changes in the actual opinions of the respondent population.

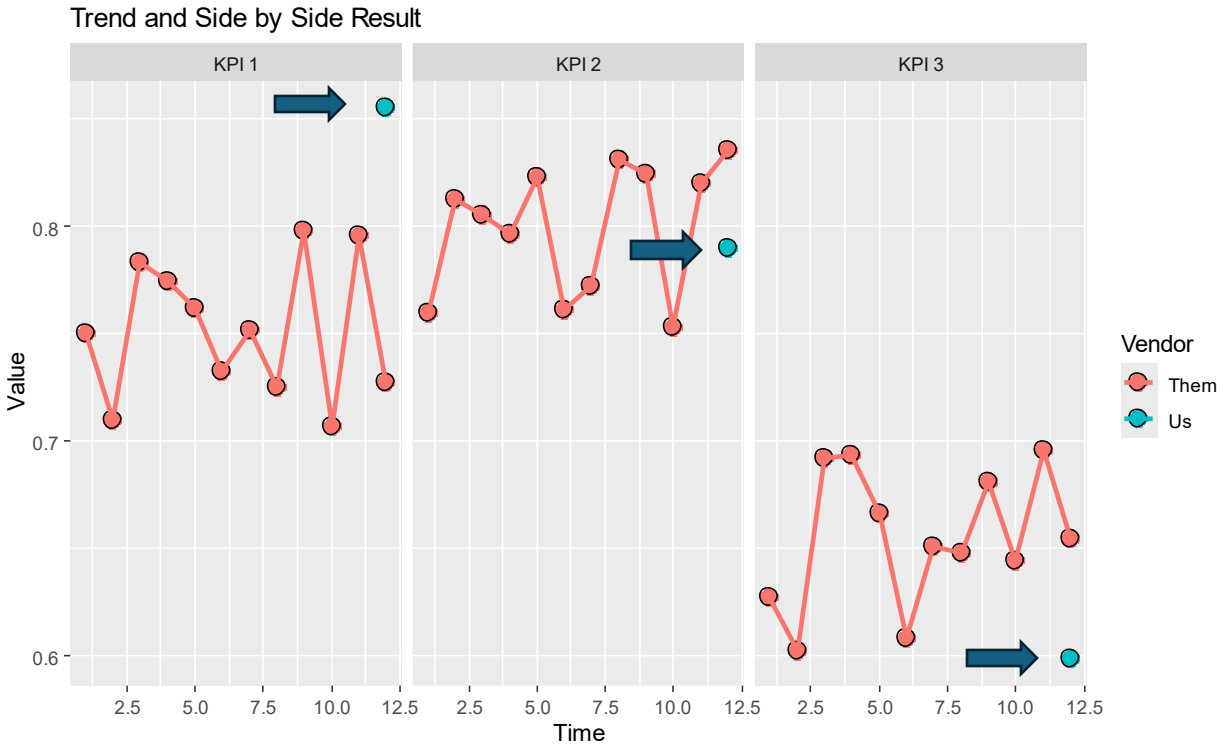
Once we measure the difference, the question becomes how to transform the responses from previous waves of the study to the level that we observe with the new vendor. A naïve approach would be to calculate the difference in a KPI in the current side by side wave and apply that difference to all previous waves reported by the previous vendor. This approach can be dangerous if we consider common KPIs like total brand awareness. Suppose in the side by side wave we observe 95% total brand awareness in the new vendor's survey and 90% total brand awareness in the previous vendors survey. The naïve approach would apply the 5 percentage point difference to all previous waves by the old vendor to create a series compatible with the new vendor's data collection. A problem occurs if any previous wave of the research showed total brand awareness above 95%. This would result in a transformed KPI that showed total brand awareness to be above 100%. Clearly this would impact the credibility of the procedure.

This is where the latent variable cutpoint model provides a cleaner and better solution.

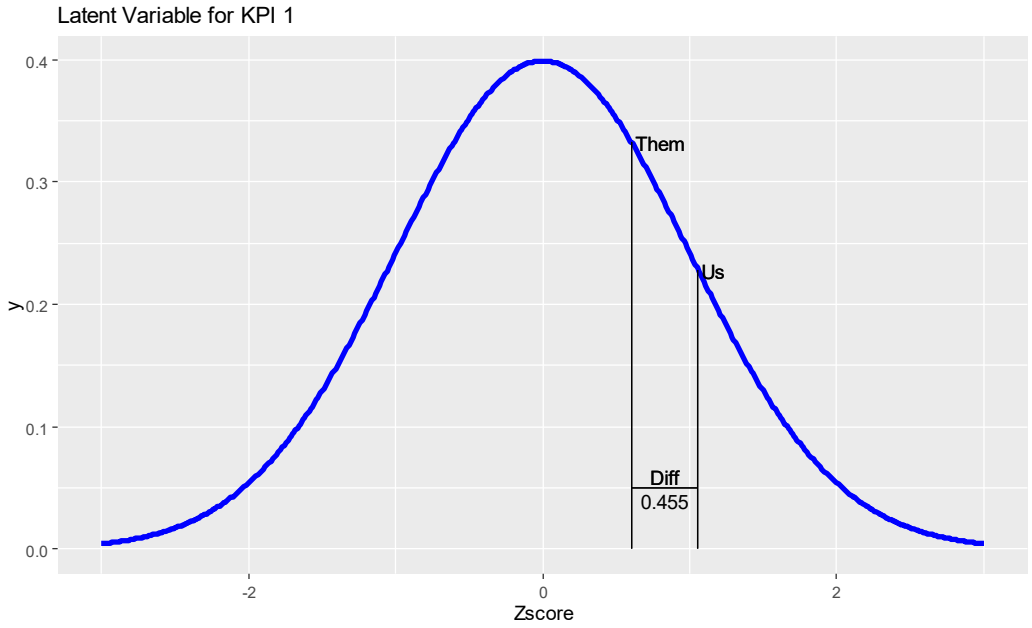
A better approach

- Treat the two side by side observations as measures of the same latent variable.
- The differences between the two observations represent different cut-points in the latent variable.
- If we calculate the shift in cut-points we can ensure that the KPIs cannot go out of range.
- This approach also allows transformations of entire scales between the studies.

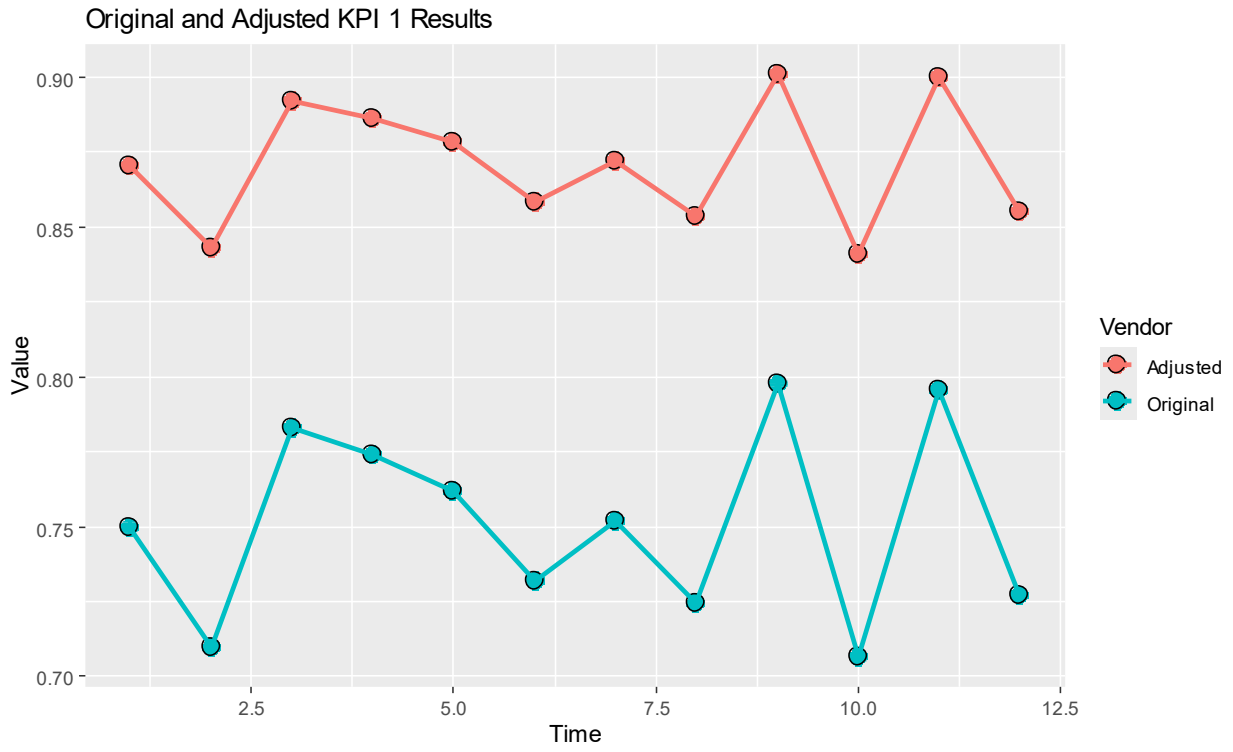
As an example, let's assume we are measuring several KPIs over multiple waves of a tracking study and now as the new vendor we perform the side by side wave.



The new observations from the new vendor are highlighted by the arrows. As we can see each KPI shows results that differ between vendors in the side by side wave. As an example let us focus on KPI 1. If we assume that both vendors are trying to measure the same latent variable then the cut-points for each vendor must be different due to the differences in data collection. If we plot the differences between the cut-points we can see that the difference is 0.455 on the scale of the latent variable.



The procedure is then to add that difference to the cut-points derived from the previous waves of research and retransform them into equivalent KPI values as if the previous waves of research had been conducted under the new survey conditions.



The result is that the trends are preserved but there is no danger that the transformation will result in non-credible results. You can note that the adjusted values are of the same shape as the original values but the changes from wave to wave are slightly smaller. This is because under the new vendor the changes are closer to the boundary (100%) so the changes from wave to wave have to be less extreme.

There are some caveats for this approach.

- The transform is still based on only one period of side by side surveys. Future waves could also use a different mix of sample source, sometimes without you knowing.
- This is inherently an aggregate analysis. If KPIs are reported for subgroups, the procedure would need to be repeated for each subgroup. Smaller subgroup sample sizes would mean that the estimated cut-points would be less reliable.

A more complex application: Segmentation Typing Tools

Typically, when clients perform a segmentation study they request a "Typing Tool". A typing tool is a model that predicts which segment from the segmentation study a respondent belongs to, based on a small set of questions collected in another survey. The typing tool questions are designed to be added to other surveys without causing a large increase in interview time and cost.

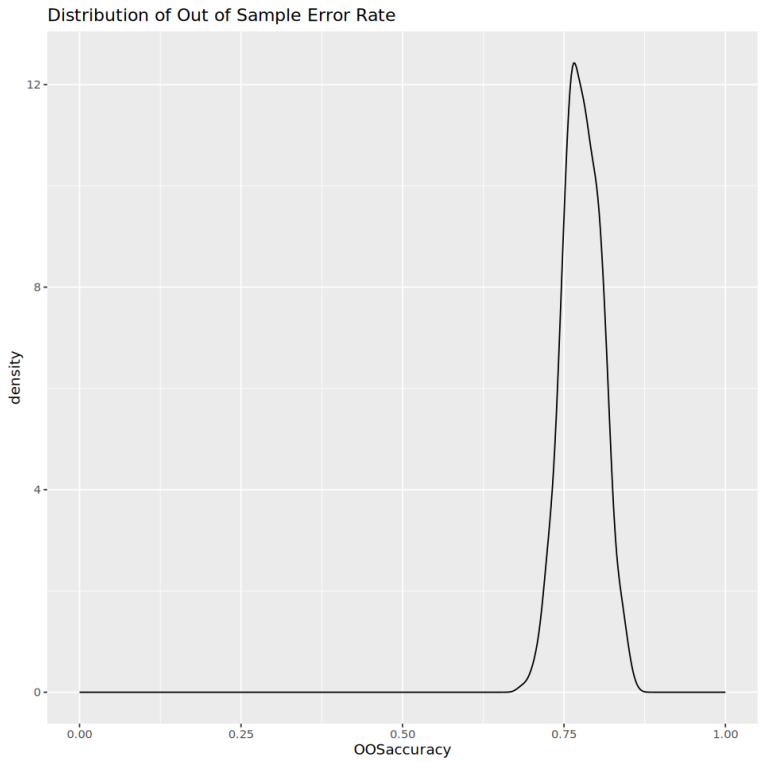
By applying the typing tool model a client can get an idea of how different segments of interest respond in any new survey.

Typing tool models are never perfectly accurate because they use fewer questions than the original segmentation model used to define the segments. In addition, models typically used for typing tools may be sensitive to the distribution of scale answers if they differ between the segmentation study and the study to which the typing tool is applied.

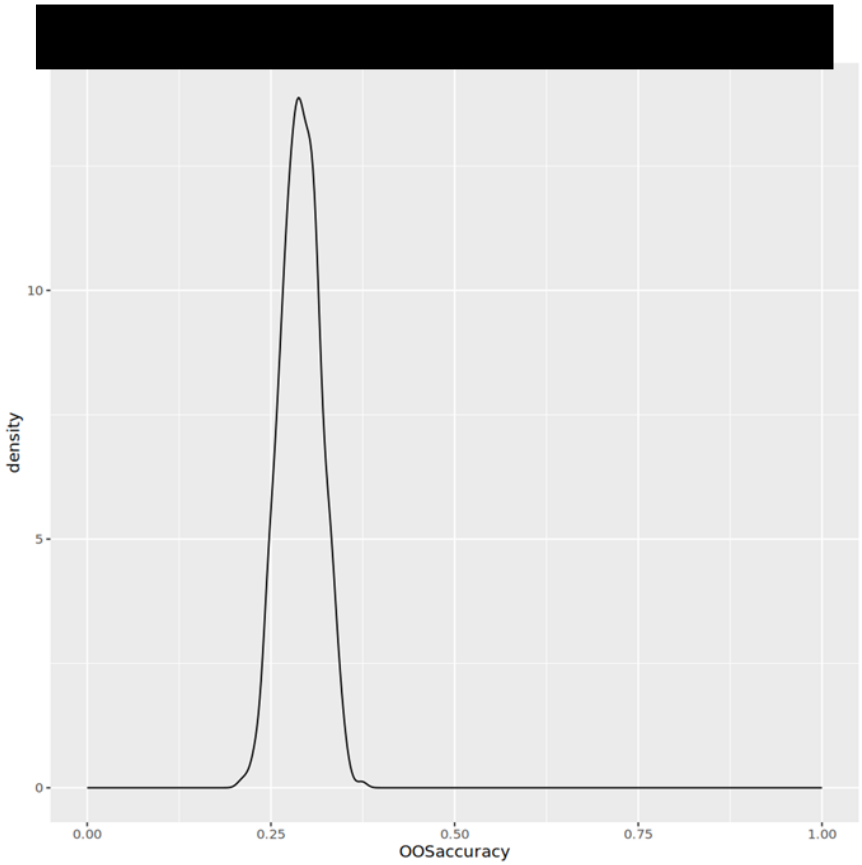
To explore some of the issues and demonstrate how this can work, I will take you through a case study.

- Client had conducted a large segmentation study among people who had the potential to be their customers.
- Questionnaire covered a wide variety of attitudes, behaviors and demographics.
- Client wished to tag people in their customer database to understand the mix of segments among their customers to help in developing marketing and messaging plans.
- Database contained mostly demographics and behavioral data.

If the database had contained all of the variables used to construct the segmentation we could create a typing tool model that would be about 80% correct. (The 80% is an out of sample measurement that takes into account overfitting that would happen if we used the same sample to fit the model as we used to evaluate it).

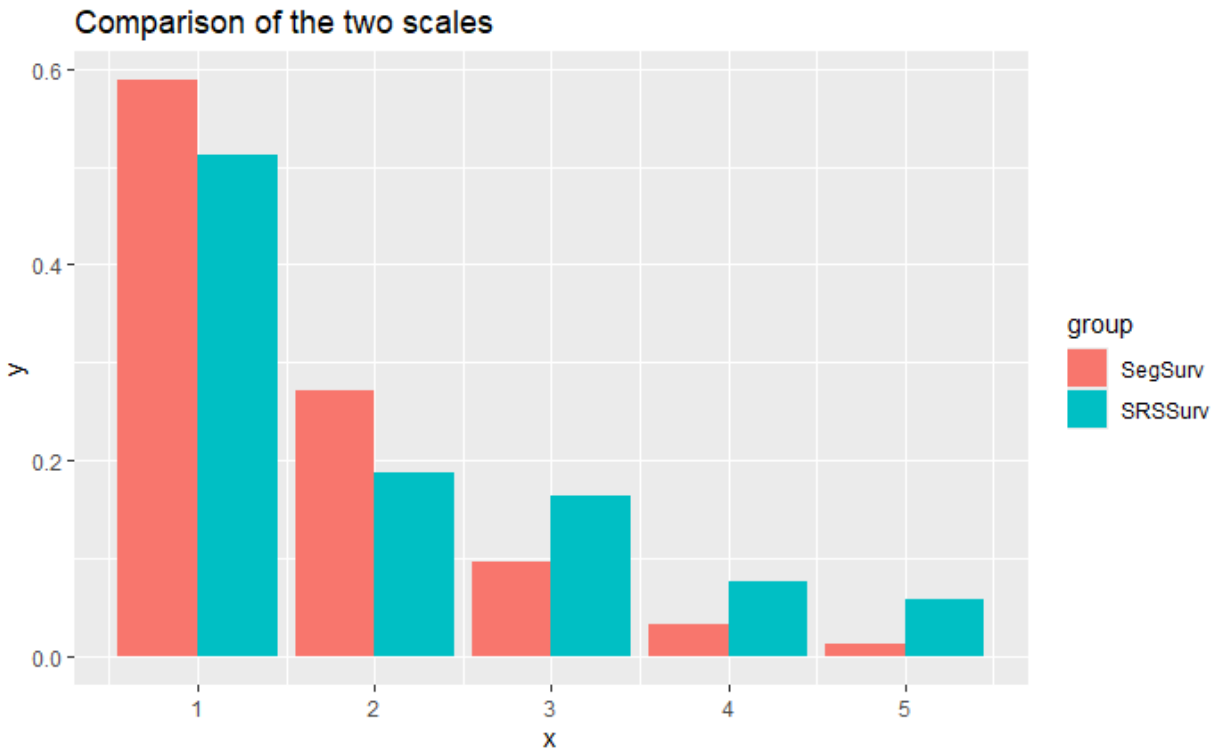


However, if we limited our model to the variables that were actually available in the database the likely performance of the model deteriorates substantially. The estimated out of sample accuracy drops to 29% on average. We can also measure the "No information accuracy rate". This is estimated by simply assigning every respondent to the largest segment. In this case, the largest segment had 36% membership. Therefore our model actually performs worse on an accuracy basis than simply assigning everyone to the largest segment.

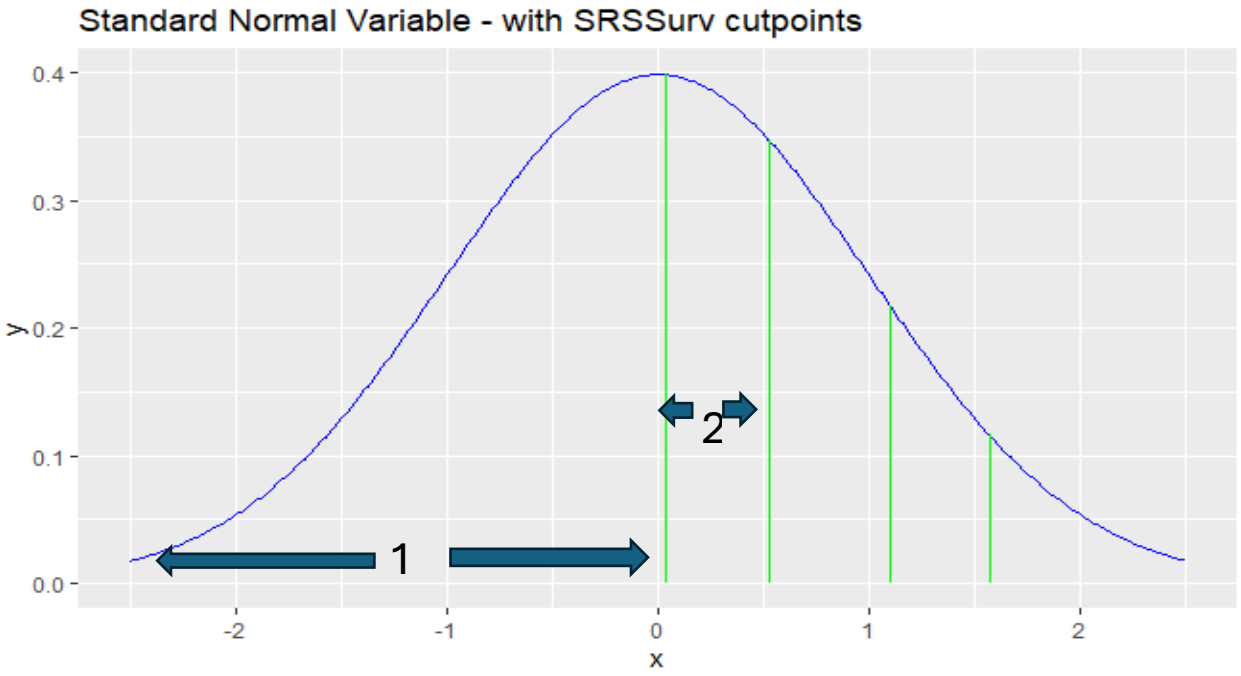


- Discussions with the client revealed that a number of people in the database had been administered an "intake" survey that included a small number of attitudinal questions that covered similar concepts to items in the segmentation survey.
- Using the concepts we covered in the tracking study example we set out to make these survey responses compatible with a typing tool.

We start by comparing the raw distributions from the two surveys. In the chart below, the intake survey is labeled SRSSurv while the segmentation survey is labeled SegSurv. As you can see, there are some clear differences in how the respondents use the scales in the two surveys.

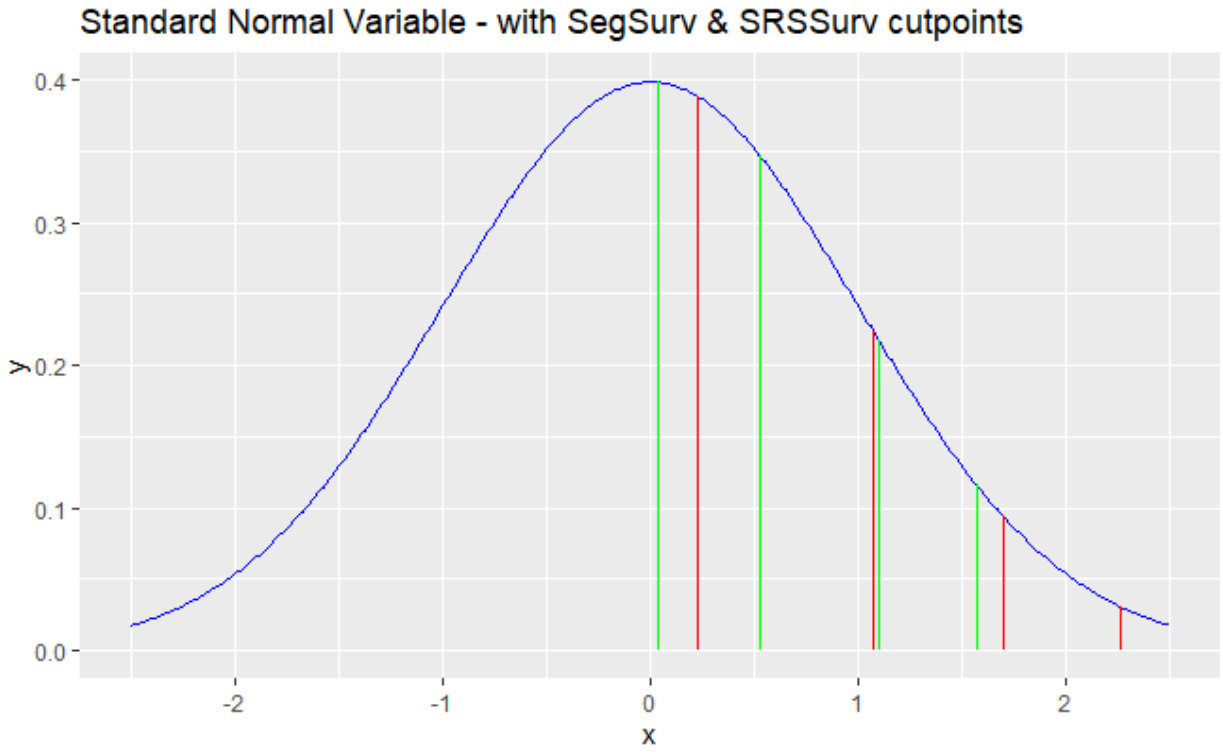


We start by mapping the SRS Survey on to the latent variable. This is the same example we showed earlier in the example of calculating the cut-points.



If a respondent gave a 2 in the SRS Survey we randomly select a value from the portion of the standard normal distribution between 0.025 and 0.524 which are the cut-points corresponding to the region labeled 2 on the chart above. This creates a value for the respondent on the latent scale that corresponds to their response of 2 on the actual scale.

The cut-points calculated for the segmentation survey question are slightly different from the cut-points in the SRS Survey. If we overlay the two sets of cut-points we get the following chart.



Recall that we selected a random value for the latent between the two cut-points that define a 2 on the SRS Survey for each respondent who gave a 2 on that survey. We then apply the cut-points from the segmentation survey question to these latent variable values that we created. This will change some responses on the SRS Survey from a 2 to a 1 depending on the actual random value that was chosen for the latent variable. This makes the overall distribution of responses to the SRS Survey compatible with the overall distribution of responses from the segmentation survey. This enable us to use the data from the SRS Survey to create a typing tool that was more accurate than the original tool that only used items from the original database.

Does this work?

There is no way to definitively know if this works. In the case study, the overall results showing the segment distributions among the customers in the database were credible to the client. Some

segments were different in size compared to the segmentation study but the differences were explainable to the client and likely due to past customer acquisition efforts.

There are a number of caveats however.

- Evaluation is primarily based on aggregate results. We look at the characteristics of people assigned to each segment and ask ourselves if the resulting distributions make sense.
- The procedure depends on the analyst identifying items from the two surveys that "seem" to measure the same concept. There were only five or six attitudinal statements in the SRS survey that seemed to measure the same attitudes in the segmentation survey. There is a lot of judgement involved in determining if the items in the two surveys are really measuring the same latent variable.
- It is clear from the procedure that an individual response of the transformed variable has a random component in it.
- Fortunately, we know exactly how that random component is introduced so we can, if necessary, understand how much it affects the assigned segment.
- In many ways, this is similar to AI generated synthetic data. The evaluation looks good on an aggregate basis but confidence in the individual respondent data is unknown. In our case though, we know the exact process used to create the data. Therefore we can repeat the procedure multiple times and evaluate how consistent the segment assignments are.

In conclusion, It is possible, without AI, to convert attitudinal items from a one survey and make them useful for inferring insights in another.

1. Dodge, Y. (2003) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9