# Section I: Executive Summary

This document is a summary of the analysis performed on the SOA dataset. A description of this dataset can be found at [www.soa.org](http://www.soa.org). The objective of this analysis was to determine what issues, gaps, inconsistencies, problems, or outliers exist within the dataset. This dataset was too large (at over 30 million rows and 32 columns) to analyze in Microsoft Excel, so other analytical tools had to be used to manipulate and gain insights from the data. For our analysis, we imported the dataset into RStudio using the Amazon Web Services Simple Storage Service (AWS S3). The dataset, and multiple subsets of the dataset by plan type, were converted to R data objects to make data manipulation more efficient. Analysis was then performed at the individual policy (row) and aggregated levels. This 1-page Executive Summary highlights the key findings of our analysis and is followed by a 10-page analysis with more details from our work.

The most substantial problem uncovered was the vague or unknown categorizations found in 5 columns, 71% of entries by count, and 42% of entries by amount. Simply discarding or ignoring this much data is infeasible. Since unknowns were such a significant portion of the dataset, missing data was the most important problem to address in our analysis. More details can be found in section II below.

The next problem uncovered was the approximately 232,000 rows of data with zero exposure. All but three of these entries have zero in all other exposure and death columns. Three of these entries have values greater than zero for death amount. These are data gaps, and these entries have been flagged. More details can be found in section III below.

Once we completed a high-level overview of the dataset, we looked at trends in both given and feature-engineered columns. The primary trends analyzed were Actual to Expected death counts and amounts (A/E) by duration for Term policies. Guaranteed Periods of 10, 15, and 20 years all showed expected patterns of low A/E ratios during the term period followed by a spike in the year after the guaranteed period ends (shock mortality). This pattern was not as clear in the 5-year cohort. More details can be found in section VI below.

Throughout the analysis, a pattern of decreasing A/E by increasing face amount was observed. This was consistent with expectations. Average face amount was also summarized by attained age. It was determined that this pattern varied widely by Plan Type. Older UL policyholders tended to have larger face amounts while younger Perm policyholders had larger face amounts. It is recommended that further analysis is performed on these patterns to determine if this is consistent with expectations. More details can be found in section VII below.

The rest of this document summarizes our analysis. In addition to providing more details about our findings mentioned above, several other insights and validations are presented. In the appendix, ASOPs, including details about our analytical methods, are provided with a link to our team’s GitHub repository, which contains all of our code, analytical tables, and R Shiny dashboards.

Simon Hua

Alyssa Columbus

**Team Pacific Life**

Tommy Steed, FSA

# Section II: Unknown Categories

As discussed in the Executive Summary, five columns have vague category options such as “Unknown” or “Other” or have data points entered as “NA”. The five columns affected are summarized in Figure 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **CATEGORY** | **COUNT** | **% BY COUNT** | **% BY AMOUNT** |
| Anticipated Class “Unknown” | 5,655,138 | 41.4 | 37.7 |
| Preferred Class “NA” | 11,069,266 | 60.7 | 20.4 |
| Guaranteed Class “Unknown” | 2,079,337 | 12.7 | 10.1 |
| Smoker Type “Unknown” | 2,720,328 | 29.3 | 3.6 |
| Plan Type “Other” | 236,870 | 0.1 | 0.2 |
| **TOTAL** | **21,760,939** | **70.7** | **42.2** |

Figure 1: Unknown Category Summary

Judgment will have to be used when dealing with this data. Some factors are more significant than others, and that significance may vary from business question to business question. After accounting for overlap, 70.7% of policies by count and 42.2% of policies by amount are affected. This is a significant portion of the dataset, which makes this the most important problem to address.

# Section III: Exposure

Looking at exposure by row reveals that 232,828 of the 30 million rows have zero exposure, or a value of zero for all exposure, death, and expected death columns. Our team recommends that these rows should be deleted from the dataset before any analysis is done, because the total population impacted is very small (0.76%) and the zero exposure associated with these policies would result in no change to a study’s results if they were taken out. The exceptions to this rule are three rows that have a non-zero death amount. Removing these policies from any study would have an impact on results (albeit a very small one), but the policies should still be removed due to the inconsistency between death amount and death count. Death count is zero, which is inconsistent with a non-zero death amount.

In addition, 42,195 rows were found where actual deaths exceeded exposure. This analysis assumes that all exposure (from both deaths and survivors) is accounted for in the exposure column. Under this assumption, the death total should never exceed exposure.

# Section IV: Plan Type



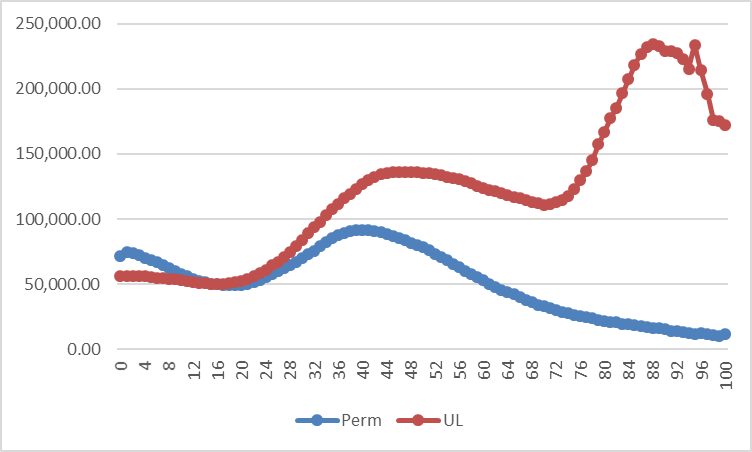
Figure 2: Summary by Plan Type

* **Throughout this document, heatmaps are formatted so larger values are green and smaller values are red.**
* **#: Indicates count-weighted values**
* **$: Indicates dollar-weighted values**
* **Δ: Indicates count-weighted minus dollar-weighted**

80% of exposure is in two plan types: Term and Perm. Although the split is relatively even by count, Term has much more exposure by amount since Term has a much higher average size ($402k compared to $53k). Both UL and VL are split by the option of a secondary guarantee, but most of the exposure for both is without secondary guarantee. The final category is Other, which was discussed in section II.

The distribution of expected deaths is not consistent with the distribution of exposure. Figure 2 shows actual and expected deaths for Perm policies are much greater than exposure. This suggests differences in underlying demographics. Also, Perm has a higher average age and a smaller size, which are consistent with higher deaths.

Plan type Perm has a higher count-weighted average age than dollar-weighted. This suggests younger policyholders have larger face amounts. On the other hand, UL has a higher dollar-weighted average age than count-weighted. This indicates older policyholders have larger face amounts. This is consistent with Figure 3.

As summarized above, these relationships all appear to be internally consistent. However, this should be compared to industry norms/expectations.

A/E ratios are shown for deaths using the 2015 VBT table. Count-weighted values show that every plan type has an A/E of at least one, which was unexpected. Count weighted A/E is greater than dollar-weighted for every plan type. This indicates that policyholders with larger face amounts live longer – and this is consistent with expectations. This difference is most pronounced in Term policies – which suggests face amount is a bigger driver of Term mortality than other plans – perhaps indicating more rigorous underwriting.

Figure : Average Size by Attained Age

# Section V: Observation Year



Figure 4: Summary by Observation Year

Figure 4 shows that all observations occurred in the expected timeframe of 2009 to 2015 with a very consistent pattern of moving from red to green (small to large) with time. There is only one column where this is the exception: dollar-weighted A/E deaths, which shows the opposite pattern – decreasing values by increasing observation year.

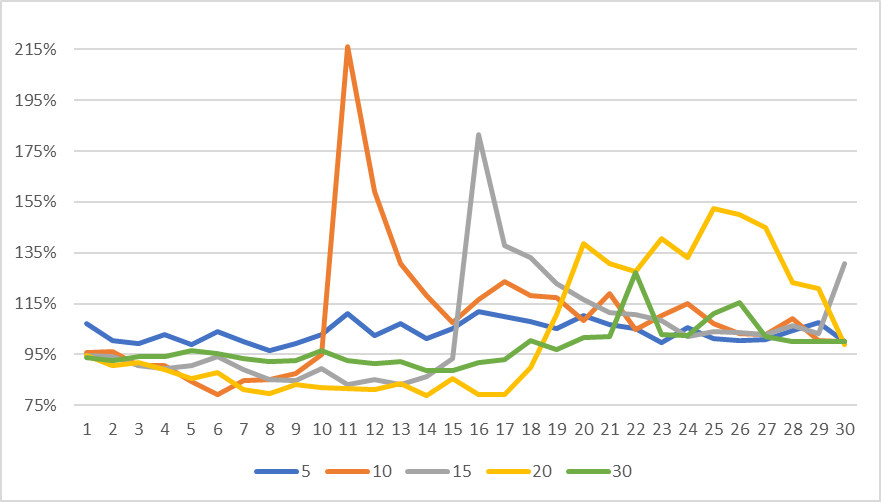
At first glance, this appears to be evidence of mortality improvement. Further analysis, however, suggests otherwise. For example, count-weighted A/E ratios don’t show a pattern that suggests mortality improvement. Also, the difference between count and dollar-weighted ratios increases by Observation Year. Instead of mortality improvement, this appears to be evidence of increased selection taking place on policies with high face amount. Further analysis would be required to more conclusively determine the cause of this anomaly.

# Section VI: Guaranteed Level Term Period

The following table shows results for Term policies grouped by guaranteed level term period.



Figure 5: Guaranteed Level Term Period

****Most of the exposure has a Guarantee Level Term Period of 20 years. A/E varies indirectly with average size, which is consistent with expectations. As we saw in Figure 2, Term policies have the highest difference between count and dollar-weighted A/E. In Figure 5, this is broken down by Guaranteed Level Term Period, and we can see these differences are consistent with expectations. Term mortality should exhibit a pattern of lower mortality during the guaranteed period with an increase immediately after. To check for consistency with this pattern, Figure 5 was further separated by duration, and the results are shown in Figure 6.

There are several distinct spikes in the figure. Cohorts 10, 15, and 20 all have spikes that occur at end of the guaranteed period. This is consistent with expectations. Cohort 5 does not have a distinct spike, which is inconsistent with the other cohorts. Cohort 30 does not have enough exposure in later durations to investigate shock mortality. Mortality shock rate appears to have a pattern of decreasing intensity with increasing guaranteed term period (cohort 10 shock > cohort 15 shock > cohort 20 shock). Cohorts 10 and 15 appear to drop off much more quickly than cohort 20. This inconsistency should be investigated further.

Figure : A/E by Duration and Guaranteed Term Period

Guaranteed Level Term Period was also analyzed along with Anticipated Level Term Period. Results are summarized in Figure 18 (Appendix). All anticipated level term periods are greater than or equal to the guaranteed level term period. The most common selection in any guaranteed level term period is the anticipated level term period of equal duration. This is consistent with expectations.

# Section VII: Face Amount



Figure 7: Face Amount Summary

In the above figure, the column “RATIO” has been added to indicate where the average size of a band fits within that band’s numerical range. For example, in row 1, $4,009 is 40.1% of the way between 1 and 10K. All “RATIO” values are between 0% and 100%, which indicates all average sizes fall within the appropriate band, and further inspection reveals that all ratios are below 50%. This indicates that, on average, policies are closer to the low end of their band than the high end. There is a very clear trend of decreasing A/E with increasing face amount, which has been a consistent theme in this report. There is also little difference between count and dollar-weighted A/E, since policies of similar sizes have been grouped. Most policies by count are in the 100K – 250K range, but looking at exposure by amount shows a concentration in a wider range.

# Section VIII: Preferred Class and Smoker Status



Figure 8: Preferred Class Summary

Figure 8 summarizes policies grouped by 3 fields: Preferred Class, Number of Preferred Classes, and Smoker Status. There are several policies that have Preferred Class and Number of Preferred Classes set to “NA”. There are also several policies with Smoker Status set to “UNKNOWN”. These policies are addressed in section II. There are no instances of preferred class > number of preferred classes, which is consistent with expectations. There are distinct patterns of increasing A/E and decreasing average size as preferred class increases, which are also consistent with expectations. The cohort with the largest A/E is a non-smoker cohort (preferred class 2 out of 2 preferred classes). This is inconsistent with the overall pattern of smoker policies having larger A/E than non-smoker policies.



Figure 9: Smoker Status Summary

The “UNKNOWN” cohort has an extremely low average size, which would suggest a much higher A/E than observed. This appears to be an inconsistency, as the observed trend is that a lower face amount is consistent with a higher A/E.

# Section IX: Age Basis



Figure 10: Age Basis Summary

Policies using ALB have a higher A/E than ANB policies. This difference is substantial (~10%). Average size is smaller for ALB policies, which is consistent with the higher A/E.

To confirm that the face amount pattern we saw earlier wasn’t a result of an ANB/ALB mix, the data were sorted by face and age basis. The results are shown to the right in Figure 11.

As the table shows, the pattern of decreasing A/E by increasing face amount appears in both the ANB and ALB groups. This indicates the pattern is not a result of changing the age basis mix and supports the claim that A/E decreases with increasing size.

Figure : Face Amount Band Summary

# Section X: Gender, Preferred Indicator, Common Company Indicator



Figure 12: Common Company Indicator Summary

Over 98% of the business is Common Company Indicator 1. This means nearly all of the data is from companies that participated in at least 5 of the 7 observation years (2009-2015).



Figure 13: Preferred Indicator Summary

Policies that were submitted into a preferred class structure have much larger average sizes and lower A/E ratios. Policies not submitted into a preferred class structure make up 61% of exposure by count but 90% of actual and expected deaths by count. This is consistent with the expectation that preferred policies will be healthier.

The difference between count and dollar-weighted A/E ratios is 0.17, which is much greater than the .06 difference in the Standard Cohort. This suggests more rigorous underwriting on high-dollar Preferred policies than high-dollar Standard policies.

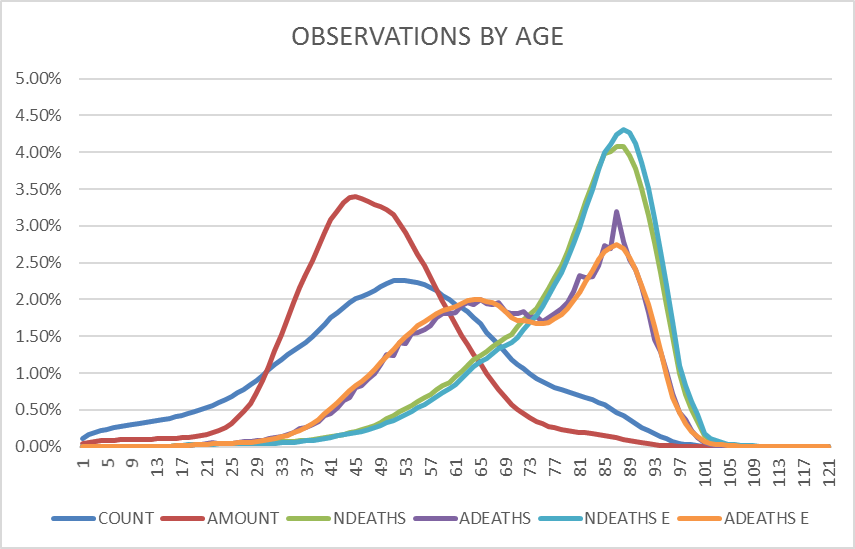


Figure 14: Gender Summary

Distribution by gender is even by count. Males have a larger average size, which is consistent with the higher proportion by amount. Group A/E ratios are nearly identical, as expected with gender-specific tables. Males had a higher portion of (expected) deaths than exposure while females had a lower portion. This is consistent with expectations.

# Section XI: Attained Age

Distribution by attained age is summarized in Figure 15 and is consistent with expectations.

Consistency examples:

* Distribution of Actual Deaths is nearly identical to distribution of Expected Deaths (Count and Amount).
* Exposure is skewed younger than deaths (Count and Amount).
* Exposure by Count is more widely dispersed than exposure by Amount.

It is interesting to note the two peaks in the distribution of (expected) number of deaths. This is consistent with some of the results from the analysis by plan type where we saw certain plan types had very large face amounts at these older ages.

Figure : Age Distribution

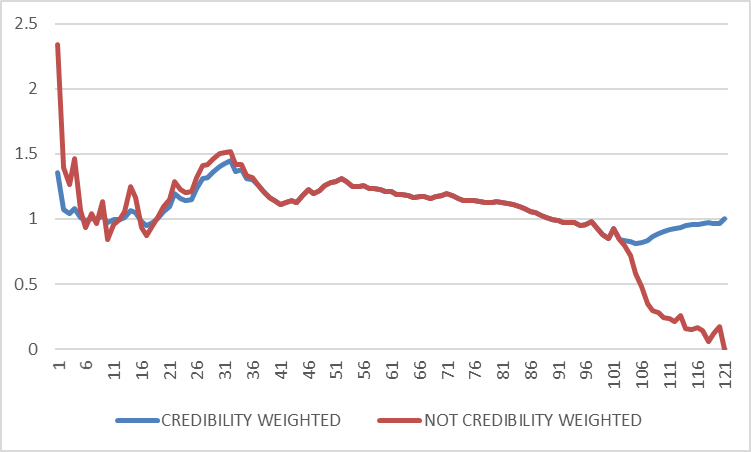
Figure 16 summarizes A/E (pre and post-credibility weighting) by attained age. Extreme ages (both high and low) appear to be inconsistent with other ages and expectations. These extreme ages are also the areas of least credibility since this is where the 2 lines diverge. The patterns, however, suggest something more than just high variance associated with low frequency. Older ages are all very low. This may suggest some deaths that have not been reported. The reason behind the large ratios at lower ages is more difficult to explain without further investigation.

Figure 16: Attained Age Summary

# Section XII: Policy Duration

Figure 17 summarizes exposure of several key values by duration.

Exposure by amount is heavily concentrated in early durations. Some of this can be explained by higher face Term policies lapsing after the guaranteed period.

Actual and expected deaths have nearly identical distributions for both amount and count.

Number of deaths (both actual and expected) are both concentrated in later policy years, with a substantial portion of exposure occurring in durations 50+. This exposure seems abnormally high.

Figure : Duration Summary

# Section XIII: Smoker Status Prevalence

In addition to the Smoker Status inconsistences discussed in section VIII, there are issues with smoking prevalence in the dataset’s policyholder population. 495,392 juvenile policyholders (issue age < 18) were labeled as smokers, which implies that at least 1.6% of smoker status data points are incorrect. This issue can most likely be attributed to errors in data entry. Also, the prevalence of smokers in the dataset had an increase from 0.000% in the first issue year, 1906, to 0.186% in the most recent issue year, 2015, with the first recorded smokers in 1925. With the surge in legislation and public awareness about the harms of tobacco in the last century, we would expect this trend to fall, not rise. One possible explanation is that insurance companies didn’t start recording the smoking status of their policyholders until 1925, and until enough evidence of tobacco’s adverse effects on life expectancy was widely accepted, life insurance companies weren’t very concerned about knowing their policyholder’s smoker status until around 1981, when the prevalence of smokers in this dataset spiked.

# Section XIV: Overall Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **FIELD** | **SECTION** | **VALIDATIONS** | **FLAGS** |
| Observation Year | V | * All observations occurred 2009-2015. | * Review potential mortality improvement. |
| Common Company Indicator 57 | X | * All values are 0 or 1. | * Very unbalanced data - ~1.5% labeled 0. |
| Preferred Indicator | X | * All values are 0 or 1. * Much larger average size for policies listed as 1. | * Disproportionately high deaths in preferred indicator 0 cohort. |
| Gender | X | * All values are Male or Female. * Similar A/E ratios. * Males have more deaths than females. | * None. |
| Smoker Status | VIII, XIII | * Overall A/E > Non-Smoker. | * Combined with preferred class – cohort with the largest A/E is a non-smoker cohort. * Some policies are classified “Unknown.” * Some juveniles (issue age < 18) are labeled as smokers. * The prevalence of smokers increased from 1906 to 2015. |
| Insurance Plan | IV | * None. | * None. |
| Issue Age | XI | * None, see Attained Age. | * None, see Attained Age. |
| Duration | XII | * A/E by duration in level term period groups. | * None. |
| Attained Age | XI | * All policies are between 0 and 120. * Distribution of deaths is older than exposure. | * Extremely low A/E in extremely low ages – indicating policies may be forgotten. * Dual peaks in distribution of death amount by attained age. |
| Age Basis | IX | * ANB A/E < ALB A/E. * All policies classified as ANB or ALB. | * None. |
| Face Amount Band | VII | * Decreasing A/E by increasing face band. * Overall average size within face band for all groups. * Exposure concentrated in the central bands: 100K – 2.5M. | * Group average size typically on the low end of the band. * Some rows had 0 for amount yet had a face band listed. |
| Issue Year | XIII | * None. | * None. |
| Number of Preferred Classes | VIII | * None. | * Preferred Class and Number of Preferred Classes are both NA for some entries. |
| Preferred Class | VIII | * None. | * Preferred Class and Number of Preferred Classes are both NA for some entries. |
| SOA Anticipated Level Term Period | VI | * All anticipated level term periods greater than or equal to guaranteed level term period. * Most common anticipated level term period was the same duration as the guaranteed level term period. | * Large portion categorized as unknown. |
| SOA Guaranteed Level Term Period | VI | * 10, 15, and 20YT show decreasing A/E by increasing term within level term period. * Large drop in average size coincides with shock mortality. | * Over 10% categorized as unknown. * 5YT and 30YT shock mortality inconsistent with end of term period. |
| SOA Post level term indicator | NA | * None. | * None. |
| Select\_Ultimate\_Indicator | NA | * None. | * None. |

# Section XV: Appendix

## Actuarial Standards of Practice (ASOPs)

ASOP 23 Data Selection

* All data was provided to us by the SOA.
* The only intended use of this analysis is the SOA contest. This data is reasonable for this purpose.
* Data Definitions: provided. Some vague entries – and little feedback.
* Identify questionable values – this was the purpose of the report.
* In my professional judgment – this data is sufficient to perform the analysis.

ASOP 25

* Classical Credibility (3,007) was used.
* This is expected to produce reasonable results.
* This is appropriate for the intended use and purpose.
* This is practical to implement when considering cost and benefit of employing a procedure.

ASOP 41: Communications and Disclosures

* Responsible Actuary: Tommy Steed
* Actuarial Documents: See GitHub repository
* Intended Users: judges of the SOA Contest
* Scope and Intended Purpose: All analysis was done in support of the SOA data contest.
* Information on which we relied: SOA contest dataset.
* Reliance on others: All data was provided by the SOA.

## Links

[Our GitHub Repository (github.com/acolum/soa-data-analysis-contest)](https://github.com/acolum/SOA-Data-Analysis-Contest)

## Additional Tables



Figure : Guaranteed Term by Anticipated Term period