Conversion Benchmark Report Analysis Methodology

This methodology was prepared by the data team at Unbounce and lays out the research process in well-considered detail. If you’re looking for a simpler guide to understanding how to apply our findings to your marketing campaigns, read the section called “How We Created This Report.”

1. Industry Classification

Using the machine-generated topic mixture for the 2020 Conversion Benchmark Report, we began the manual (i.e., not machine) process of assigning topics to industries. We started with the set of machine-generated topics, the Global Industry Classification Standard, and the North American Industry Classification System, and began to map a relationship between topics and industries.

Our goal was to assign each topic to a single industry and to generate a final list that described the landing pages in our dataset. The priority was to define industries that fit the data, rather than imposing an existing classification system as-is. We found we needed more granular classification to capture the conversion rate variation within the industry, and so we created a second layer of hierarchy using subcategories. Since each page in our dataset received a topic mixture, it was also assigned a normalized industry and subcategory mixture; thus each page could be grouped under multiple topics, industries, and subcategories.

For the 2021 Conversion Benchmark Report, we used the same classification to make sure we were looking at apples to apples when comparing year-over-year changes.

2. Industry Inclusion

Our analysis showed that the majority of pages in the dataset were assigned a mixture of 1 to 4 industries. When we included pages whose industry mixture was at least 16%, the majority of pages were assigned 2 industries. Given our set of industries and pages, we decided it was an appropriate description of the data to allow a page to be split between 2 industries. We also decided to only include pages with industries or subcategories that had at least 16% inclusion in the given mixture element.
3. Topic Exclusion

We found that some topics had a much higher conversion rate than others, and resulted in a conversion rate distribution that wasn’t representative of the whole. Some of these topics were manually removed from the analysis to ensure an accurate representation of the dataset. The excluded topics were:

- Politics
- Religion
- Rewards and Affiliate Programs
- Sales Vernacular
- Terms and Conditions
- Testimonials
4. Sample Size

We haven’t included analysis on any data subsets with fewer than 400 pages for the benchmarks. We explored a number of options to decide how large a subset should be to ensure statistically significant results.

Again, this methodology was carried over from the 2020 Conversion Benchmark Report. The exploration included an examination of the standard error on the distribution of conversion rate, a Chi-Square test to check the statistical significance of categorical factors on conversion rate, and a two-sample Kolgomorov-Smirnov test to compare distributions of subsets with varying sample sizes. We chose a conservative value of 400 to ensure we had enough data in each analysis to confidently report on the results.
5. Conversion Type
For our analysis of conversion types, we defined call-to-action types based on conversion goals set in the Unbounce builder. There were five possible types:

- **Click**: All conversion goals on the page are set up as click conversions.
- **Form**: All conversion goals on the page are set up as form conversions.
- **Phone**: All conversion goals on the page are set up as phone conversions.
- **Mixed**: The conversion goals on the page are set up with at least two conversion types (listed above).
- **Tracking**: We infer that pages without an explicit conversion goal that still receive conversions must have external tracking in place. This includes mechanisms like Google Analytics, which are added to the page using JavaScript or third-party integrations.

6. Mean vs Median
Conversion rates, by nature, have a wide range of values depending on the amount of traffic the page naturally receives. We found that the **mean** conversion rate was much higher than the **median**, due to skew from a handful of very high-converting pages.

To account for both outliers and skew, we used median across the majority of the report. We believe that the median better represents conversion rate benchmarks because it describes where a typical page will fall within subsets of the data.
7. Year-Over-Year Analysis

To look at how the benchmarks have changed between the two years, we look at the relative difference of the medians. For example, Finance & Insurance saw an increase in reading ease of 12.9%, the median values for 2020 and 2021 are 51.58 and 58.21 respectively, i.e. $0.129 = (58.21 - 51.58)/51.58$
8. Traffic Analysis

The traffic analysis looks at a period of data from **August 1st to November 30th 2020**. Channel performance is evaluated by looking at the distribution of conversion rates (using box plots) for each channel. This is useful to give an indication of the low- and top-performing pages, as well as the median. It is designed to give the user more insight into where their pages may sit. Proportional traffic volume is then used to show areas of opportunity. Proportional volume is used in this case rather than raw numbers to make sure the data is relevant for any time period.

9. COVID-19 Research

A comparison of traffic data from **1st August to 30th November 2019** with the **same time period in 2020** was used in this study. Note that this is not sufficient data to draw any concrete conclusions or trends, coupled with the fact that any difference between 2019 and 2020 is a correlation and by no means attributable solely to COVID-19, this analysis is to be used for signals only.

This study looked at differences at the conversion type, mobile/desktop, and ad channel level.

Both industry aggregations and page-level statistics were used here to give a rounded view of the traffic behavior year-over-year. Box plots are used to visualize the distribution of traffic at the page level to show the variance across pages in each industry.
10. Box Plots

We have used box plots (technical name ‘box and whisker plot’) to compare both the conversion rates year over year and the traffic channel performance for each industry. This visualization gives a better sense of the distribution of data, especially when the shape of said distribution is not a simple bell curve.

The ‘box’ here represents the first quartile (25th percentile) to the third quartile (75th percentile) of the dataset. The median, or second quartile (50th percentile), is the middle value of the dataset and is represented by a line through the box. The whiskers represent the range of the remaining data, excluding any extreme outliers.

The range here is defined at 1.5 x the interquartile range (the difference between the first and third quartile). Specifically IQR = q75 - q25. The upper whisker extends to the highest point in the data set that is within 1.5 x IQR above the 75th quantile (q75) and the lower whisker extends to the lowest point in the data set that is within 1.5 x IQR below the 25th quantile (q25). Outliers are any data points that fall outside this range.
11. Regression Analysis

We used a linear regression model to determine the extent to which each factor in the analysis (conversion type, number of words, and reading ease) affected conversion rate. For these factors, we used a nonparametric lowess model (locally weighted linear regression) to approximate the trend line, as we felt it described the data better than a linear model.

We can view the relationship between conversion rate and these factors using the overall trend produced by the regression. For readability, we further bucketed the data points into 10 groups, visualizing the median of each group, to show the concentration of data within the subset.
12. Sentiment Analysis

We expressed sentiment on a page by computing the number of words that the EmoLex algorithm identified by the total number of stemmed words on the page. In this way, each page was assigned a percentage for each sentiment.

We investigated the relationship between sentiment and conversion rate using regression analysis, as with the other factors on a page. While it may have seemed natural and intuitive that sentiment could be a good predictor of conversion rate, we didn’t find this to always be true in our research and analysis.

The data we saw in our analysis is generally noisy and not clearly patterned with a clear correlation. While some pages may convert better, and show an increased percentage of sentiment, this potentially is random, and shouldn’t be relied upon to alter conversion rates in a predictable way.
The visualization of sentiment is based on a linear regression with a small number of averaged data points. We choose this for visual clarity, and to illustrate the limited predictability of sentiment on conversion rates.