Sentiment Distribution

Jason Cairns

May 3, 2019

1 Introduction

Once we have a distribution to work with, there are many properties that arise from one that can be useful to us, such as contextual summary statistics, modelling, etc. As an example, we will start with a sentiment distribution, but the concept (if useful) can be expanded to any other data with numerical variance.

2 Creating the Data

First, we will load in the nzqhs data and run a simple sentiment analysis on it

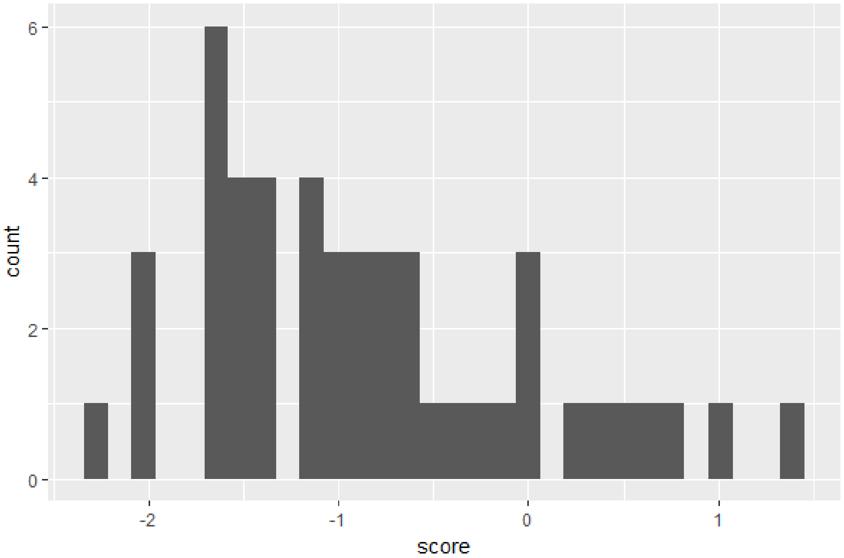
```
library(tidyverse)
library(tidytext)
library(readxl)

nzqhs <- read_excel("../data/raw/Schonlau1.xls")

sents <- nzqhs %>%
     unnest_tokens(word, `expert clinical summary`) %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(`record ID`) %>%
    summarise(score = mean(score))
```

3 Handling the Distribution

We can view a histogram of the distribution



A further question is what to do now that we have a distribution? We can get a summary of the distribution:

```
summary(sents$score)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-2.3333 -1.5000 -1.0000 -0.8449 -0.3839  1.3333 
```

Some potential uses follow:

3.1 Conditioning

Something that may be of use is to condition the data, answering the question, given some range of values in one dimension, what is the data like in another? In this case, given some range of values in sentiment, what are the keywords? Here we will assess for sentiment > 1, and sentiment < -2

```
data(stop_words)

id_conds1 <- sents %>%
  filter(score > 1)

cond1 <- nzqhs %>%
  inner_join(id_conds1) %>%
  unnest_tokens(word, `expert clinical summary`) %>%
  anti_join(stop_words)

id_conds2 <- sents %>%
  filter(score < -2)

cond2 <- nzqhs %>%
  inner_join(id_conds2) %>%
  unnest_tokens(word, `expert clinical summary`) %>%
  anti_join(stop_words)
```

Keywords for sentiment > 1:

```
head(textrank::textrank_keywords(cond1$word)$keywords)
```

```
    keyword ngram freq
1      turp     1    2
2 operation     1    2
3   voiding     1    2
4   patient     1    2
5     urine     1    2
6 operative     1    2
```

Keywords for sentiment < -2

```
head(textrank::textrank_keywords(cond2$word)$keywords %>% filter(ngram == 1))
```

```
       keyword ngram freq
1         line     1    3
2        staph     1    2
3           98     1    2
4 hypertension     1    1
5     catheter     1    1
6      central     1    1
```

This data is all fairly morbid, so the split is not entirely obvious to me. Perhaps it is worth attaining a more diverse dataset.