Can we tell the difference between text, non-text and random data?

Motivation

I read the *Why 83?* document and wanted to try a different statistical analysis.

Sub-questions

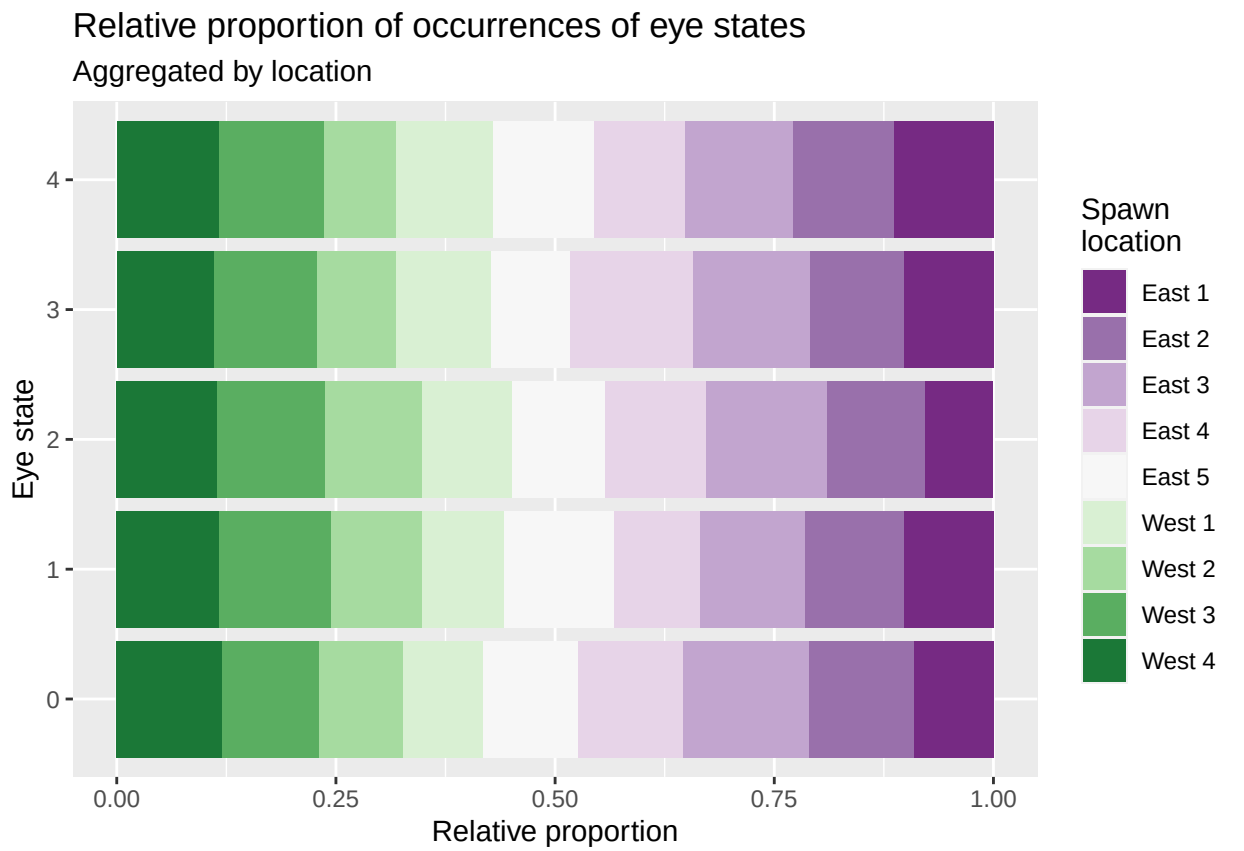
Are messages correlated to the location they spawn in?

The underlying assumption in this section is that *eye_state* is a categorical variable. We'll be looking at the χ^2 and $p_i - p_j$ statistics.

For the purposes of this section *eye messages* are the equivalent of a 5-choice response to a poll.

```
options(warn=FALSE)
suppressMessages(library(ggplot2))

messages = read.csv('./single_message.csv')
ggplot(messages, aes(y=factor(eye_state), fill=factor(location))) +
  geom_bar(position='fill') +
  scale_fill_brewer(palette='PRGn') +
  labs(title='Relative proportion of occurrences of eye states',
       subtitle='Aggregated by location',
       y='Eye state', x='Relative proportion', fill='Spawn\nlocation')
```



Qualitative observation: the relative proportion of occurrences of eye states suggests that all messages have a common structure.

```
contingency_table = t(table(messages))
contingency_table
```

```
##          eye_state
## location   0   1   2   3   4
##   East 1  70  76  55  50  46
##   East 2  93  83  78  53  47
##   East 3 111  89  96  65  50
##   East 4  93  72  81  69  42
##   East 5  84  93  74  44  47
##   West 1  70  69  72  53  45
##   West 2  75  77  77  44  33
##   West 3  85  94  86  58  49
##   West 4  93  86  80  54  47
```

H_0 : Aggregated counts of eye states are independent of the spawn location of the message

H_A : Aggregated counts of eye states depend on the spawn location of the message

```
chisq.test(contingency_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
```

```
## X-squared = 21.917, df = 32, p-value = 0.9096
```

The R output above reports a p-value of 0.91: fail to reject H_0, there is no relation between location and relative proportions of eye states. This result supports what was highlighted by the stacked bar chart earlier. It's then *reasonable* to think all messages share a common structure even if this structure is a random variable in the probabilistic sense.

Difference of sample proportions: eye mesasges

For any given message, the difference in relative proportions gives us information about the prevalence of states, and more importantly an edge on the question *is my sample just a fluke?*

Pairwise difference of proportions,

$$\begin{cases} p_0 - p_1,\ p_0 - p_2,\ p_0 - p_3,\ p_0 - p_4, \\ p_1 - p_2,\ p_1 - p_3,\ p_1 - p_4, \\ p_2 - p_3,\ p_2 - p_4, \\ p_3 - p_4 \end{cases}$$

```
# This code is re-used in the following sections
# Run it to define the functions
suppressMessages(library(latex2exp))

compute_difference_of_sample_props = function(n, N, baseSample) {
  bsprops = array(dim=c(N, 10))
  for (i in 1:N) {
    props = as.vector(unlist(table(sample(baseSample,
                                          n,
                                          replace=TRUE))/n, use.names=FALSE))
    assertthat::assert_that(sum(props) == 1)
    bsprops[i, 1]  = props[1] - props[2]
    bsprops[i, 2]  = props[1] - props[3]
    bsprops[i, 3]  = props[1] - props[4]
    bsprops[i, 4]  = props[1] - props[5]
    bsprops[i, 5]  = props[2] - props[3]
    bsprops[i, 6]  = props[2] - props[4]
    bsprops[i, 7]  = props[2] - props[5]
    bsprops[i, 8]  = props[3] - props[4]
    bsprops[i, 9]  = props[3] - props[5]
    bsprops[i, 10] = props[4] - props[5]
  }

  value = NULL
  statistic = NULL
  labels = c('p0 -p1', 'p0 - p2', 'p0 - p3', 'p0 - p4', 'p1 - p2',
             'p1 - p3', 'p1 - p4', 'p2 - p3', 'p2 - p4', 'p3 - p4')
  for (i in 1:10) {
    value = c(value, bsprops[ , i])
    statistic = c(statistic, rep(labels[i], N))
  }

  list(plotdf=data.frame(value=value, statistic=statistic),
       propdf=data.frame(bsprops))
}
```

```
suppressMessages(library(stringr))
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
suppressMessages(library(numform))
```

```
## Warning: package 'numform' was built under R version 4.2.2
```

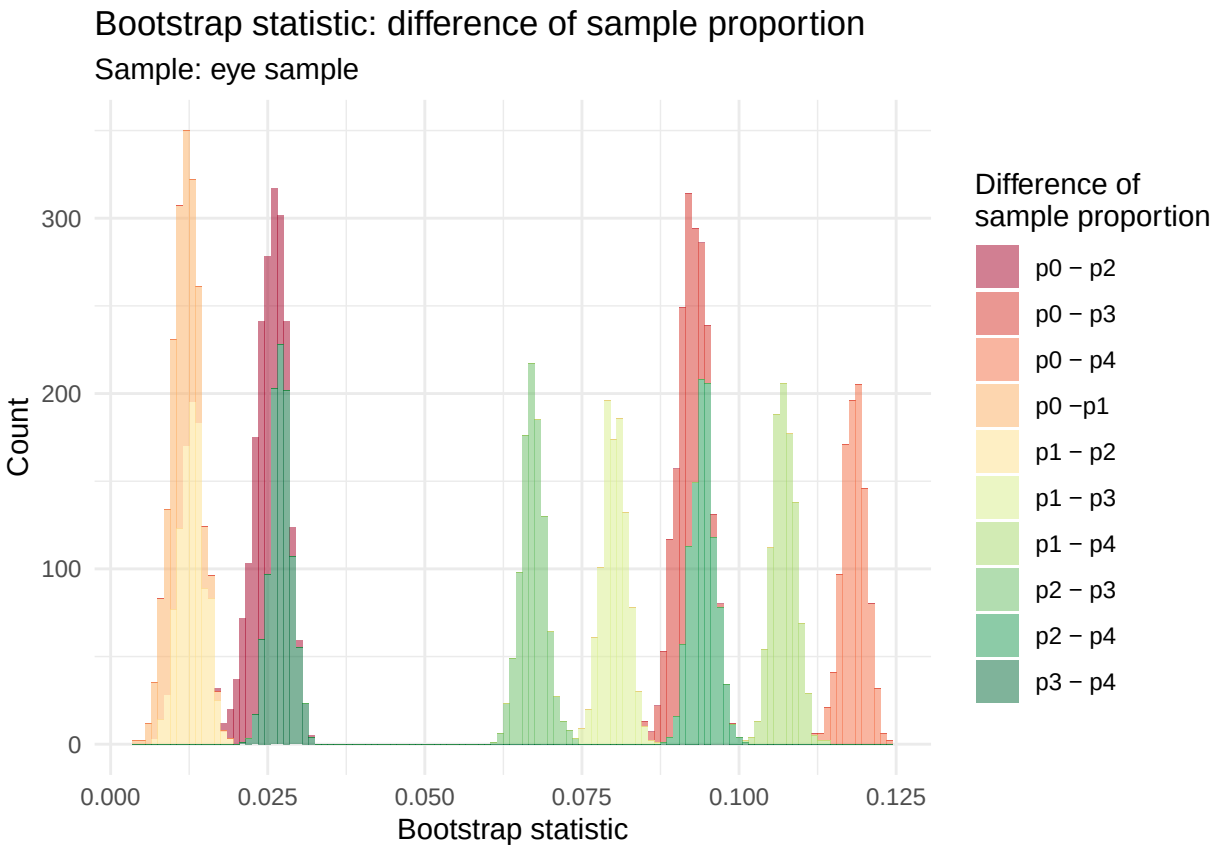
```
plot_and_CI95 = function(plotdf, propdf, sampleType) {
  p = ggplot(plotdf, aes(x=value, fill=statistic)) +
        geom_histogram(binwidth=0.001, alpha=0.5) +
        scale_fill_brewer(palette='RdYlGn') +
        labs(title='Bootstrap statistic: difference of sample proportion',
             subtitle=str_interp('Sample: ${sampleType}'),
             fill='Difference of\nsample proportion',
             x='Bootstrap statistic', y='Count') +
        theme_minimal()

  CI = c(0.025, 0.975)
  for (i in 1:10) {
    interval = as.vector(unlist(quantile(propdf[[str_interp('X${i}')]], CI)))
    width = abs(interval[1] - interval[2])
    print(paste('Width:', f_pad_right(round(width, 3),pad.char='0', width=5),
                ' | CI 95 contains 0:', interval[1] <= 0 & interval[2] >= 0, ' | Mean:',
                round(mean(propdf[[str_interp('X${i}')]]), 4)))
  }

  return(p)
}
```

```
res = compute_difference_of_sample_props(n=100000,
                                         N=1000,
                                         baseSample=messages$eye_state)
plot_and_CI95(res$plotdf, res$propdf, 'eye sample')
```

```
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0111"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0241"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0913"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.1183"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0129"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0802"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.1072"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0672"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0943"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.027"
```

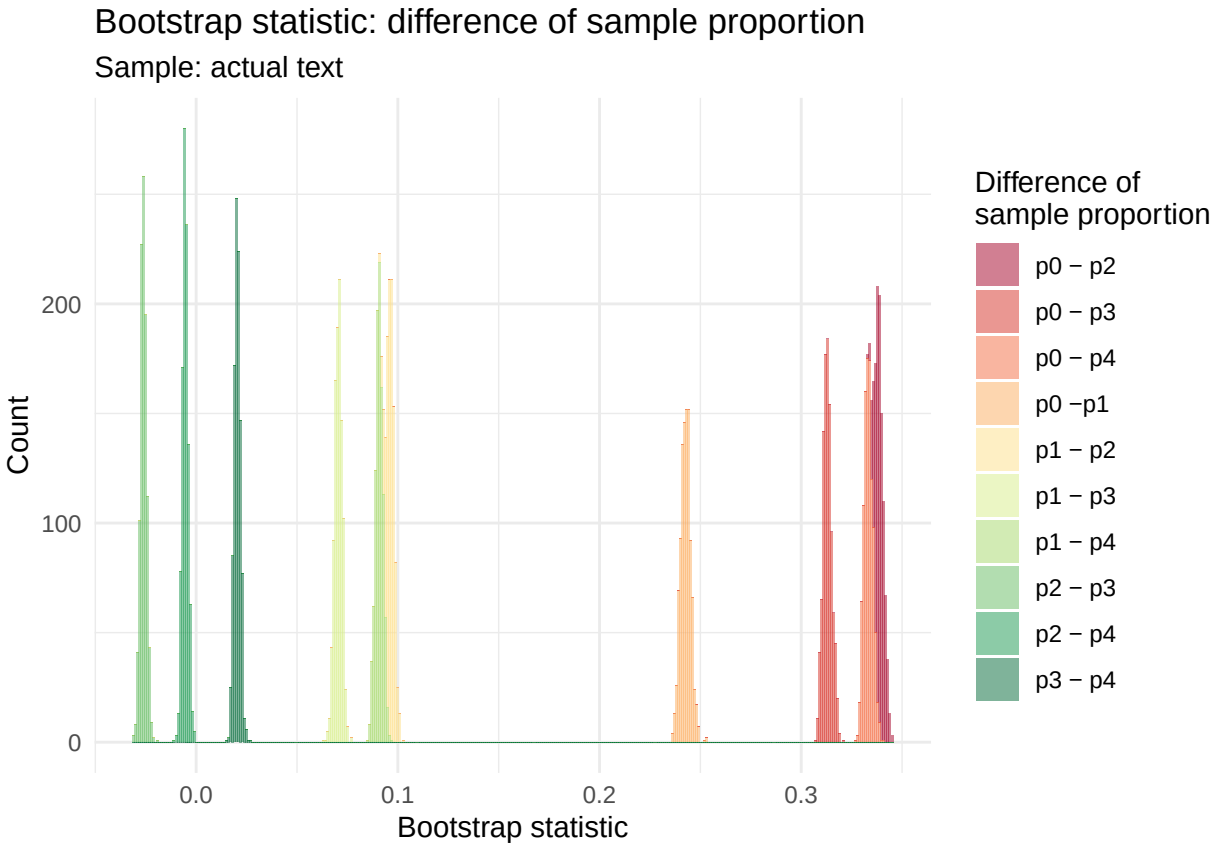


Difference of sample proportions: actual text

Let's take a sample of text and run it by the same machinery. Refer to the data_wrangling notebook for details regarding the choices made when converting to a 5-state system.

```
res = compute_difference_of_sample_props(n=100000,
                                         N=1000,
                                         baseSample=strsplit(
                                           readLines('./encoded_text3.txt'),
                                           '')[[1]]
                                         )
plot_and_CI95(res$plotdf, res$propdf, 'actual text')
```

```
## [1] "Width: 0.010  | CI 95 contains 0: FALSE  | Mean: 0.2426"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.339"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.313"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.3334"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.0964"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.0704"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.0908"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: -0.026"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: -0.0056"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: 0.0204"
```

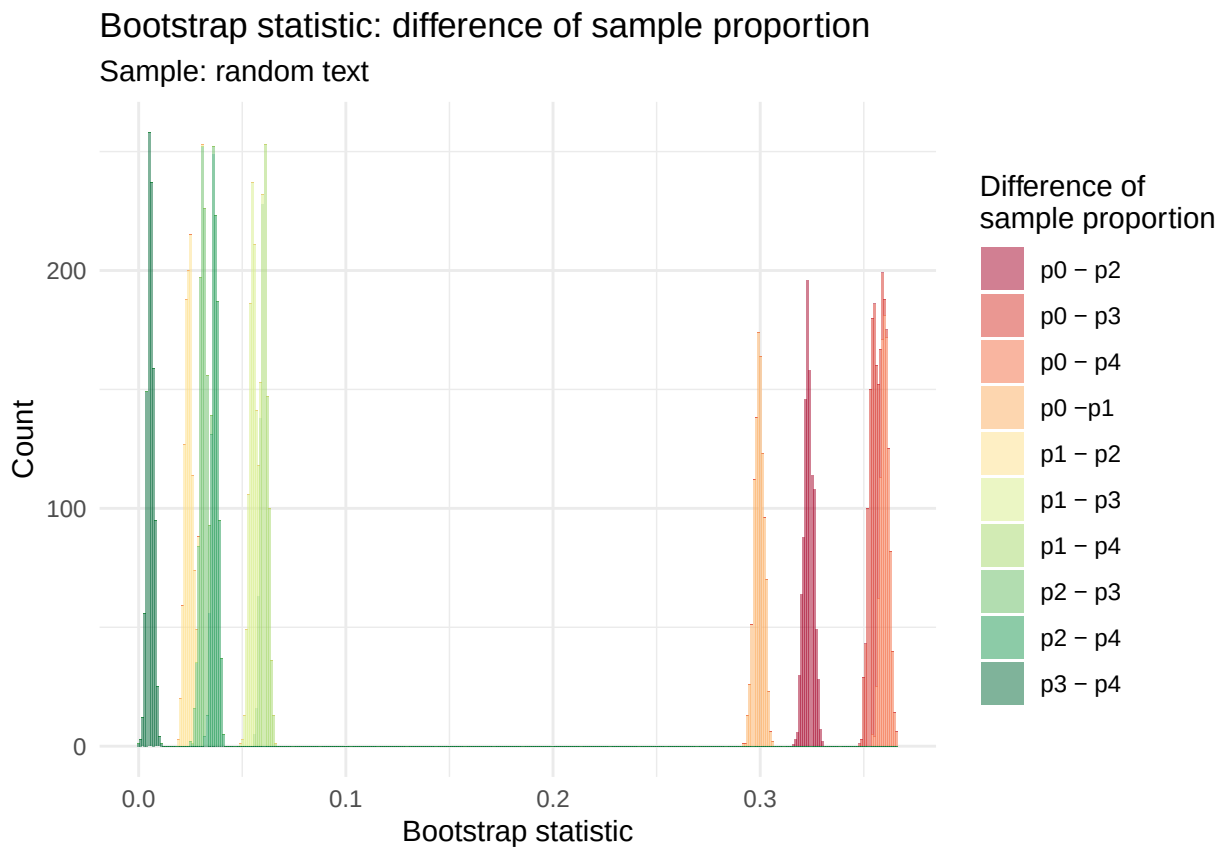


- `encoded_text1.txt` is English in a 5-state representation,
- `encoded_text2.txt` is Latin encoded the same way.
- `encoded_text3.txt` is a much larger block of English text encoded the same way

Difference of sample proportions: uniformly distributed random numbers

```
res = compute_difference_of_sample_props(n=100000,
                                         N=1000,
                                         baseSample=strsplit(
                                           readLines('./random_text.txt'),
                                           '')[[1]]
                                         )
plot_and_CI95(res$plotdf, res$propdf, 'random text')
```

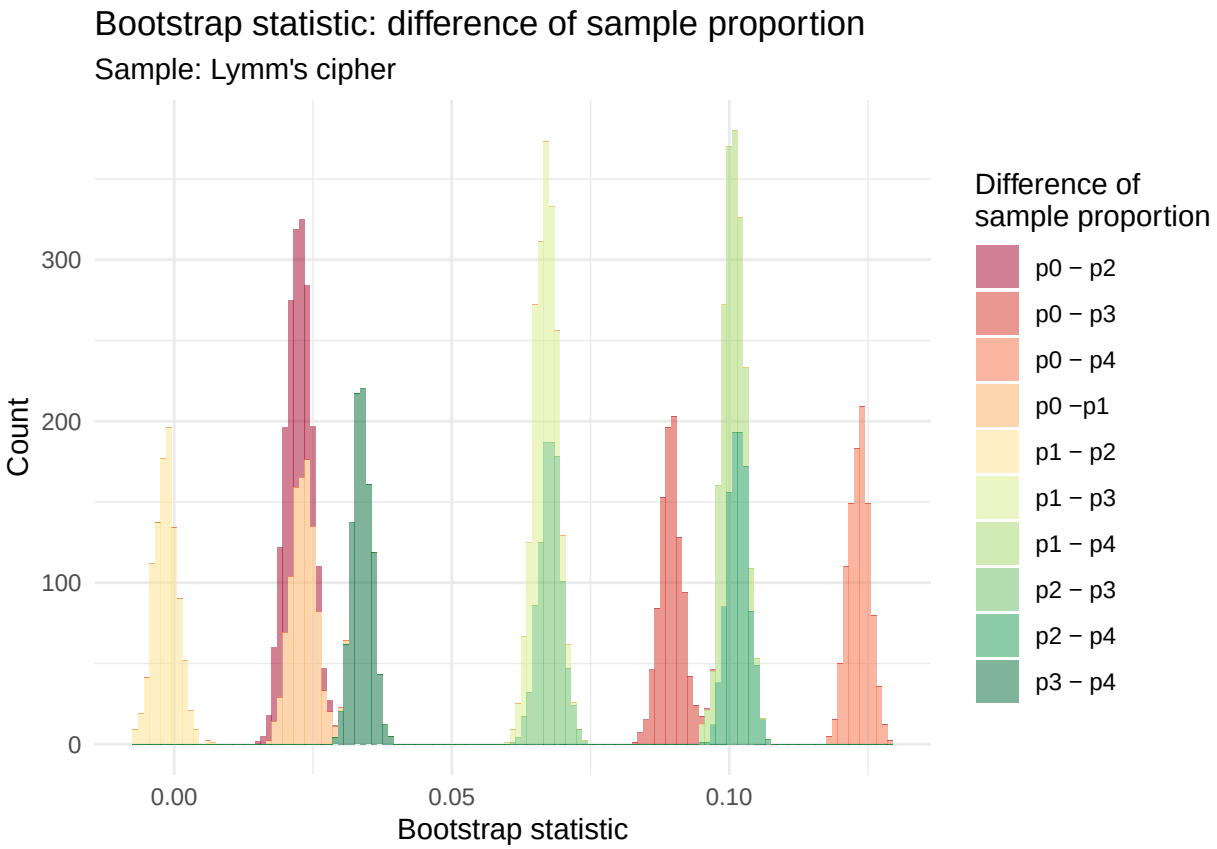
```
## [1] "Width: 0.009  | CI 95 contains 0: FALSE  | Mean: 0.2994"
## [1] "Width: 0.009  | CI 95 contains 0: FALSE  | Mean: 0.3234"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.3545"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.3602"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.024"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.0551"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: 0.0607"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: 0.0311"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: 0.0368"
## [1] "Width: 0.006  | CI 95 contains 0: FALSE  | Mean: 0.0057"
```



Difference of sample proportions: discord user Lymm's cipher

```
res = compute_difference_of_sample_props(n=100000,
                                         N=1000,
                                         baseSample=strsplit(
                                           readLines('./lymm_cipher.txt'),
                                           '')[[1]]
                                         )
plot_and_CI95(res$plotdf, res$propdf, 'Lymm\'s cipher')
```

```
## [1] "Width: 0.009  | CI 95 contains 0: FALSE  | Mean: 0.0232"
## [1] "Width: 0.009  | CI 95 contains 0: FALSE  | Mean: 0.0218"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0895"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.1234"
## [1] "Width: 0.008  | CI 95 contains 0: TRUE  | Mean: -0.0015"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0663"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.1001"
## [1] "Width: 0.008  | CI 95 contains 0: FALSE  | Mean: 0.0677"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.1016"
## [1] "Width: 0.007  | CI 95 contains 0: FALSE  | Mean: 0.0338"
```


Bootstrap statistic: difference of sample proportion
Sample: Lymm's cipher
Count
300
200
100
0
0.00
0.05
0.10
Bootstrap statistic
Difference of sample proportion
p0 − p2
p0 − p3
p0 − p4
p0 −p1
p1 − p2
p1 − p3
p1 − p4
p2 − p3
p2 − p4
p3 − p4