Comparing diversity measures

Introduction

We are reccomending using Simpson's index where "no descriptions" are counted as unique responses. Simpson's index has the advantage of having a transparent definition: the probability that two observations taken at random from the sample are of the same type. Converting "no descriptions" to unique responses also means that Simpson's index and the Shannon index correlate very highly. It also has the advantage of producing numbers for stimuli where no participant provided a response.

Functions

```
di.shannon = function(labels){
  # counts for each label
  tx = table(labels)
  # convert to proportions
  tx = tx/sum(tx)
  -sum(tx * log(tx))
}
di.simpson = function(labels){
  # Full formula due to small datasets
  # (Hunter-Gaston index)
  n = table(labels)
  N = length(labels)
  sum(n * (n-1))/(N*(N-1))
}

di.BnL = function(labels){
  #CR-DR+20
  #where, DR is the number of different responses a stimulus item receives
  #and CR is the number of subjects who agree on the most common name
  DR = length(unique(labels))
  CR = max(table(labels))
  return(CR-DR+20)
}
```

Load data

```
a = read.csv("../data/AllData_LoP.csv",stringsAsFactors = F)

d = read.csv("../data/DiversityIndices.csv", stringsAsFactors = F)

d = d[!is.na(d$simpson.diversityIndex),]
d$id = paste(d$Language,d$Stimulus.code)

d.nd = read.csv("../data/DiversityIndices_ND.csv", stringsAsFactors = F)
```

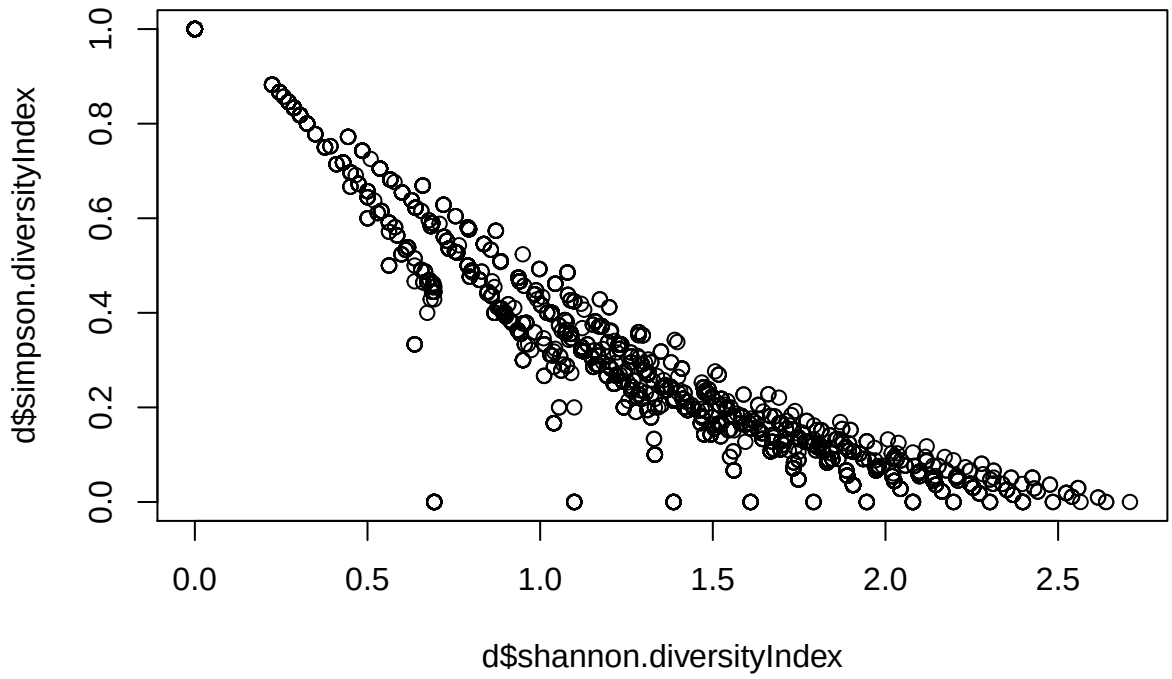
```
d.nd = d.nd[!is.na(d.nd$simpson.diversityIndex),]
d.nd$id = paste(d.nd$Language,d.nd$Stimulus.code)

l = read.delim("allCombs.txt", sep='', stringsAsFactors = F, header=F)
names(l) = c("N","shannon","simpson")
```

Compare measures

Plotting the Shannon index against Simpson's index, we see that there is a general correlation, but also several outliers.

```
plot(d$shannon.diversityIndex,d$simpson.diversityIndex)
```




```
cor.test(d$shannon.diversityIndex,d$simpson.diversityIndex)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d$shannon.diversityIndex and d$simpson.diversityIndex
## t = -150.45, df = 2838, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9466087 -0.9384030
## sample estimates:
##        cor
## -0.9426481
```

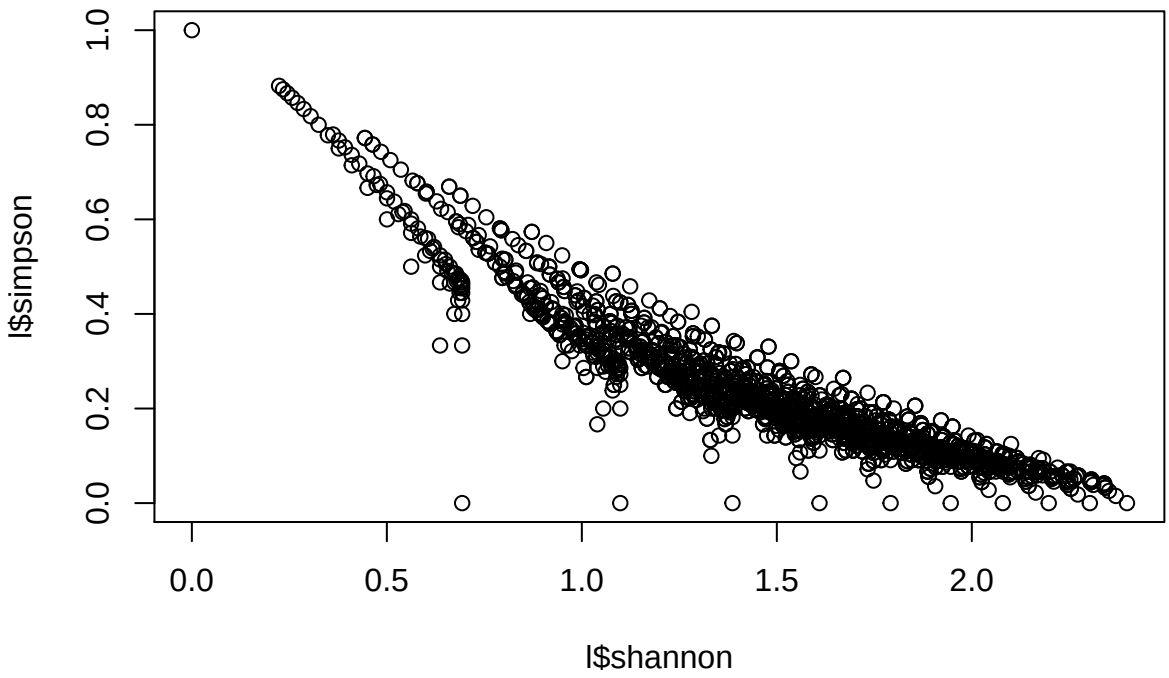
Taking into account the non-linear relationship, the variance explained in common is:

```
orig.cor = summary(lm(simpson.diversityIndex~
          shannon.diversityIndex +
          I(shannon.diversityIndex^2),
        data = d))
orig.cor$r.squared
```

```
## [1] 0.9442713
```

These outliers come close to covering the total space of possible relations between the two measures. The data from the plot below comes from `compareDiversityMeasures.py`, which generates all possible combinations of responses within the bounds of the experiment (between 2 and 14 categories, between 1 and 17 responses):

```
plot(l$shannon,l$simpson)
```



The outliers in the bottom left of the plot (where the two measures disagree) come from cases where there are many "no description" responses. e.g. colour 10G 8/6 for Umpila:

```
ax = a[a$Language=="Umpila" &          a$Stimulus.code=="colour.10G 8/6" &
         a$Response==1,]
table(ax$head)
```

```
##
##     kawithaman no description       paachala
##              1              9              1
```

```
tx = ax[ax$head!="no description",]$head
```

The original measures remove "no description" responses. That means that the example above above yeilds a Simpson index of 0 - there is no agreement between the two speakers who responded. The Shannon index is 0.6931472, since there are only two responses, and the Shannon index measures the information in that sequence. Had there been 14 completely different responses, then the Shannon index would be 2.64, which is closer to the simple relationship.

One solution is to count "no description" responses as unique labels. So in the example above, the table of responses would be:

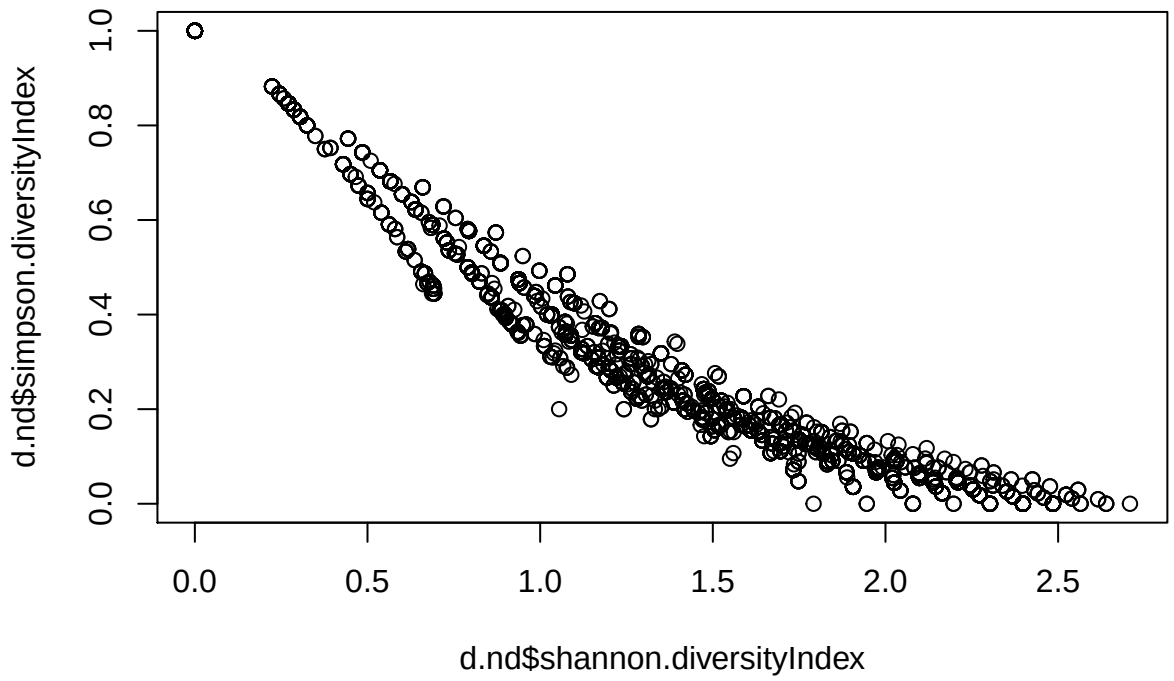
```
tx.nd = ax$head
tx.nd[tx.nd=='no description'] =
  paste0("R",1:sum(tx.nd=="no description"))
tx.nd
```

```
##  [1] "R1"         "R2"         "paachala"   "R3"         "R4"
##  [6] "R5"         "R6"         "kawithaman" "R7"         "R8"
## [11] "R9"
```

This does not affect the Simpson index, but raises the Shannon index to 2.4. It also has the advantage of producing a defined index for cases where no participants produced a label.

If we calculate all indices while counting "no description" responses as unique responses, then we get the following relationship:

```
plot(d.nd$shannon.diversityIndex,
     d.nd$simpson.diversityIndex)
```




```
cor.test(d.nd$shannon.diversityIndex,
     d.nd$simpson.diversityIndex)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d.nd$shannon.diversityIndex and d.nd$simpson.diversityIndex
## t = -206.66, df = 2848, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9704554 -0.9658590
## sample estimates:
##        cor
## -0.9682389
```

```
# Proportion of values unchanged
pvu = sum(d$simpson.diversityIndex ==
     d.nd$simpson.diversityIndex[match(d$id, d.nd$id)]) / nrow(d)
```

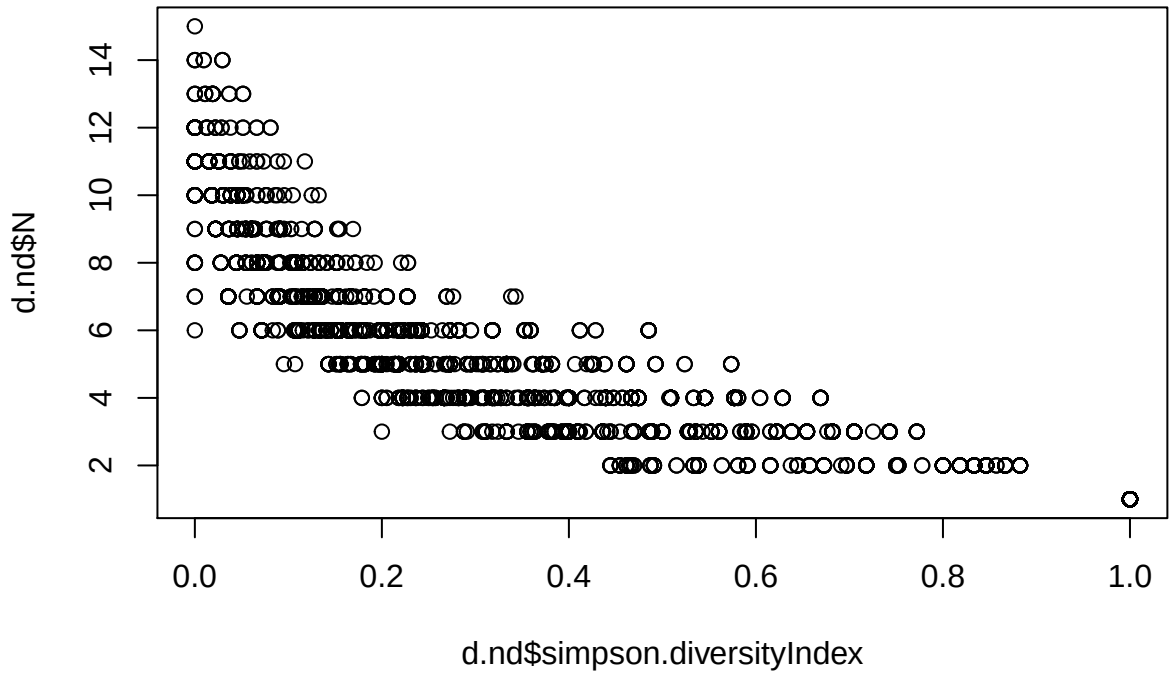
63.17% of values are unchanged. The new values are also more highly correlated:

```
nd.cor = summary(lm(simpson.diversityIndex~
          shannon.diversityIndex +
          I(shannon.diversityIndex^2),
        data = d.nd))
nd.cor$r.squared
```

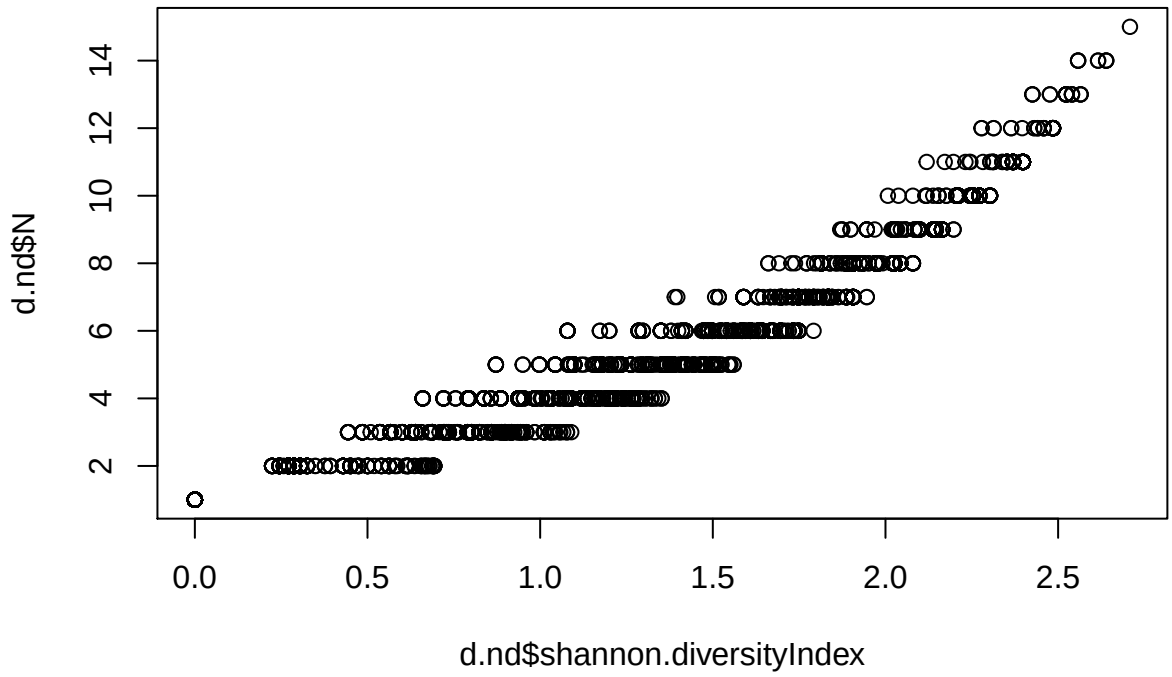
```
## [1] 0.9855785
```

Diversity and number of types

```
plot(d.nd$simpson.diversityIndex, d.nd$N)
```




```
plot(d.nd$shannon.diversityIndex, d.nd$N)
```



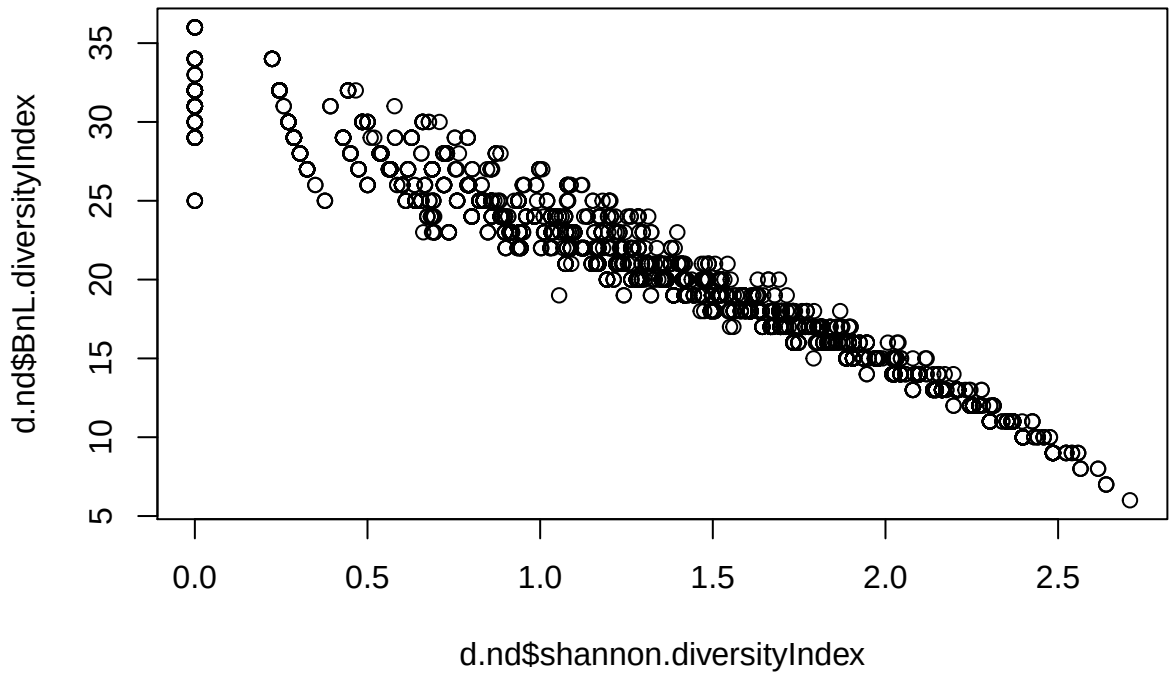
Brown and Lenneberg measure

The measure from Brown and Lenneberg (1954) of "interpersonal agreement" is caluclated as

CR-DR+20

Where, CR is the number of subjects who agree on the most common name and DR is the number of different responses a stimulus item receives (the +20 is so that the values remain positive). This does not adjust for the number of participants/responses. Still, the values are very highly correlated with both Simpson and Shannon indices, albeit non-linearly.

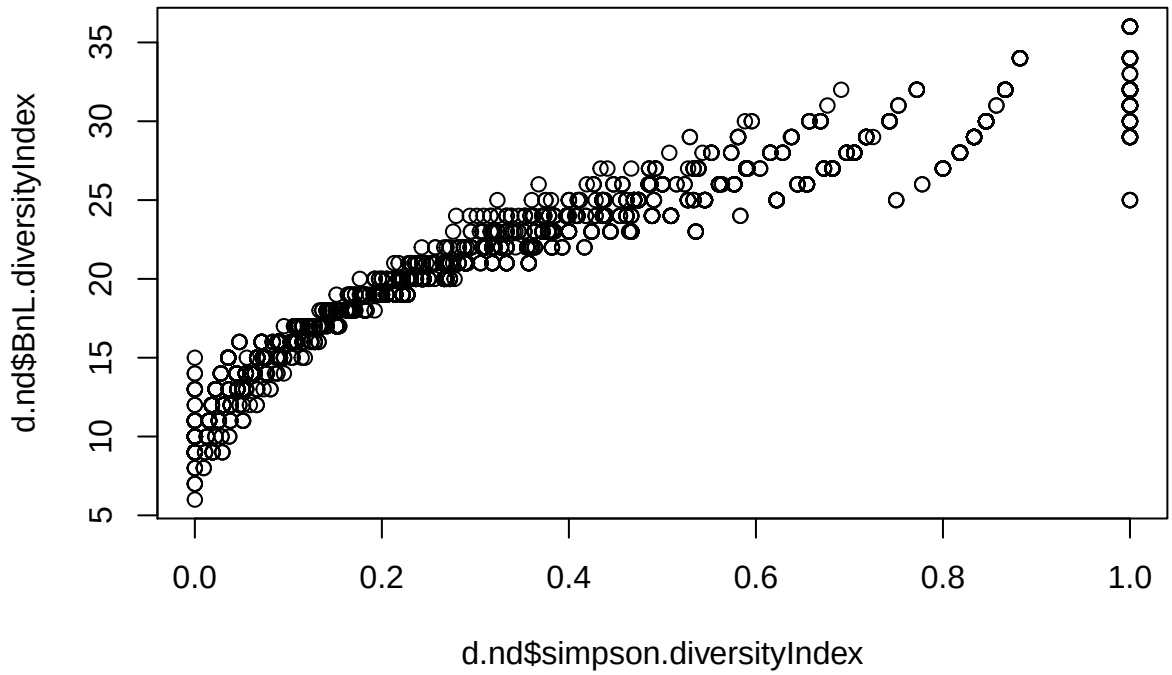
```
plot(d.nd$shannon.diversityIndex,
     d.nd$BnL.diversityIndex)
```




```
cor.test(d.nd$shannon.diversityIndex,
     d.nd$BnL.diversityIndex)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d.nd$shannon.diversityIndex and d.nd$BnL.diversityIndex
## t = -272.42, df = 2848, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9826563 -0.9799387
## sample estimates:
##        cor
## -0.9813465
```

```
plot(d.nd$simpson.diversityIndex,
     d.nd$BnL.diversityIndex)
```




```
cor.test(d.nd$simpson.diversityIndex,
     d.nd$BnL.diversityIndex)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d.nd$simpson.diversityIndex and d.nd$BnL.diversityIndex
## t = 166.92, df = 2848, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9489745 0.9557929
## sample estimates:
##       cor
## 0.9525029
```

In particular, the correlation with the Shannon index makes sense, because it is related to the 'surprisal' of encountering a label that is not your own. However, this measure is not theoretically motivated, so the other two are preferred.