IaaFmeasure.Rmd

This R script does a little bit of descriptive analysis of the IAA and F1 data that we got.

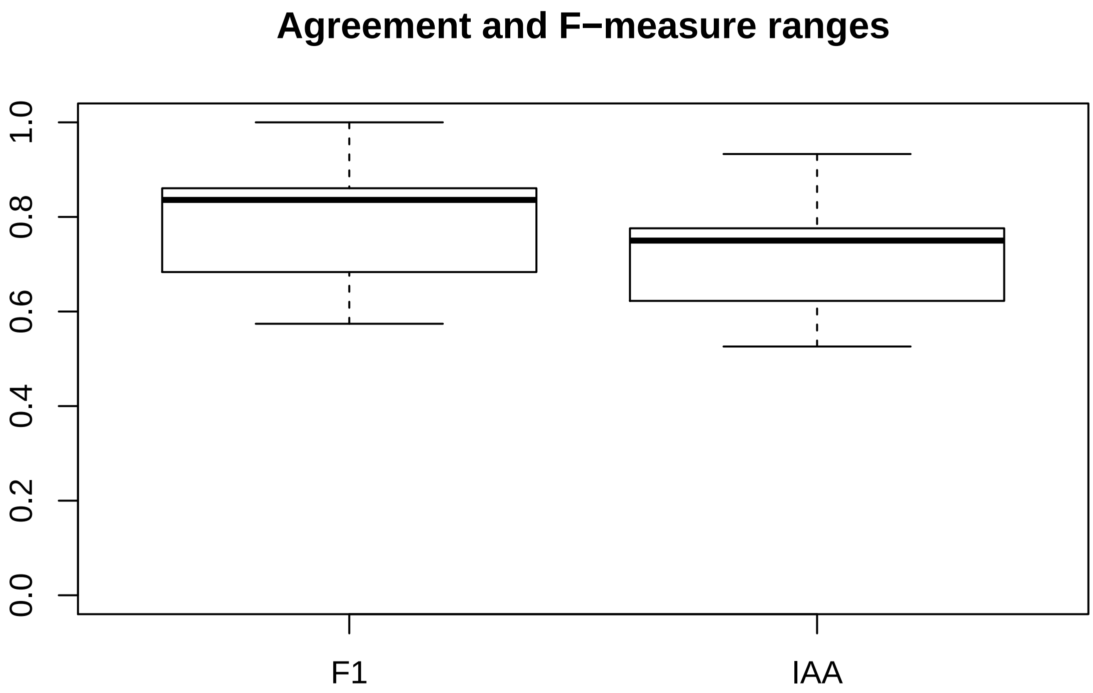
```
# kevin.cohen@gmail.com 303-916-2417

# if I were smarter I woulda pasted just the relevant columns and rows into a
# spreadsheet and saved it as a .csv file, but I'm not, sooo...

# the variable name is the column name and the cell column/row labels
iaa.e2.through.e21 <- c(0.5535, 0.5535, 0.5535, 0.617, 0.526,
                    0.69, 0.933, 0.819, 0.779, 0.815,
                    0.798, 0.773, 0.628, 0.75, 0.75,
                    0.75, 0.75, 0.75, 0.75, 0.75)
# F-measure, except two that are precisions and maybe I ought to remove
# them if I weren't so lazy---TODO
system.f2.through.f21 <- c(0.5741, 0.6504, 0.6649, 0.63, 0.664,
                           0.725, 0.978, 0.875, 0.824, 0.87,
                           0.857, 0.823, 0.702, 0.839, 0.839,
                           0.845, 0.8644, 1, 0.845, 0.833)
# I think this is F-measure minus IAA, if I'm reading the spreadsheet
# correctly, which is never a given
difference.g2.through.g21 <- c(0.0206, 0.0969, 0.1114, 0.013, 0.138,
                               0.035, 0.045, 0.056, 0.045, 0.055,
                               0.059, 0.05, 0.074, 0.089, 0.089,
                               0.095, 0.1144, 0.25, 0.095, 0.083)
```

Let’s look at the basic distributional aspects of this stuff:

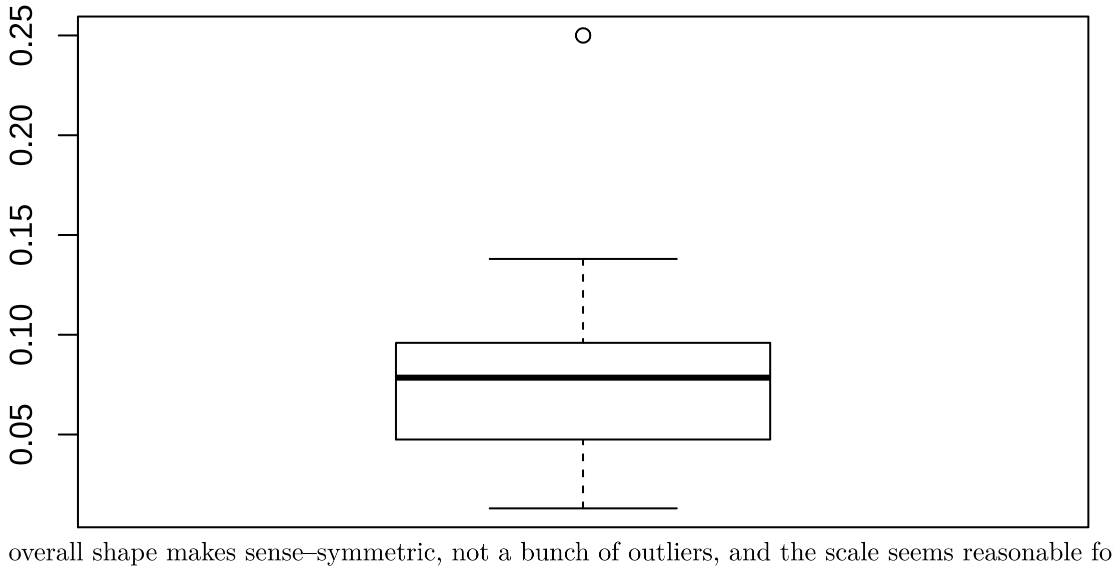
```
labels <- c(rep("IAA", 20), rep("F1", 20))
iaa.and.f1 <- c(iaa.e2.through.e21, system.f2.through.f21)
boxplot(iaa.and.f1~labels, ylim=c(0, 1.0), main="Agreement and F-measure ranges")
```



A little skewed right, but otherwise looks pretty reasonable. Median system performance is a bit higher than median IAA, which it oughta be, given our sample.

The difference is on a totally different scale, so we'll plot it all by its lonesome:

```
boxplot(difference.g2.through.g21)
```



Again, the overall shape makes sense–symmetric, not a bunch of outliers, and the scale seems reasonable for what we're looking at.

Quick summaries of the data:

```
summary(iaa.e2.through.e21)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5260  0.6252  0.7500  0.7144  0.7745  0.9330
```

```r
summary(system.f2.through.f21)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5741  0.6927  0.8360  0.7951  0.8588  1.0000
```

```r
summary(difference.g2.through.g21)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01300 0.04875 0.07850 0.08071 0.09548 0.25000
```

…and tests for normality of distribution:

```r
shapiro.test(iaa.e2.through.e21)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  iaa.e2.through.e21
## W = 0.90252, p-value = 0.046
```

```r
shapiro.test(system.f2.through.f21)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  system.f2.through.f21
## W = 0.92209, p-value = 0.1087
```

```r
shapiro.test(difference.g2.through.g21)
```

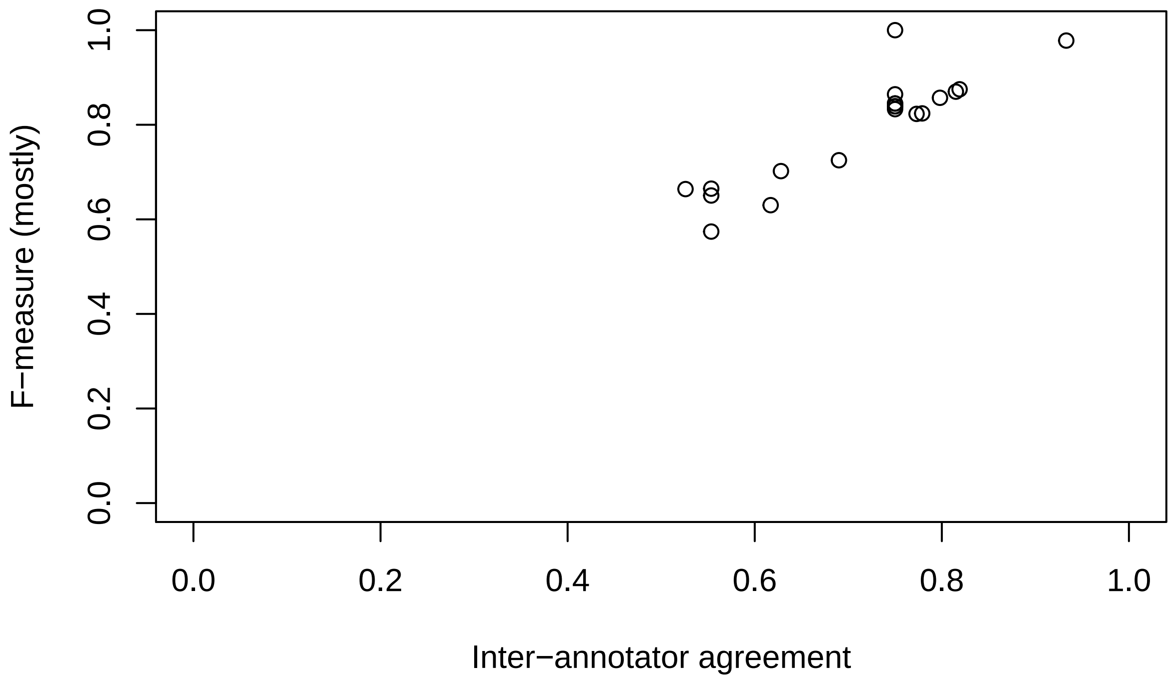
```
## 
##  Shapiro-Wilk normality test
## 
## data:  difference.g2.through.g21
## W = 0.84741, p-value = 0.004825
```

…so, system performance is normally distributed ($p > 0.05$), but the other two aren't ($p < 0.05$). (This is probably counter-intuitive, so let me remind you how tests of fit to distributions work. Here the null hypothesis is that the data is NOT different from a null distribution. So, if you have a HIGH value, i.e. a non-significant p-value, then the data IS normally distributed.)

Can you calculate Pearson's correlation coefficient for data that isn't normally distributed? Dr. Google suggests Spearman instead.

First let's graph it, then we'll find out what the correlation is:

```r
plot(iaa.e2.through.e21, system.f2.through.f21,
     xlim=c(0,1.0), ylim=c(0,1.0),
     xlab="Inter-annotator agreement",
     ylab="F-measure (mostly)")
```



```
cor(iaa.e2.through.e21, system.f2.through.f21, method="spearman")
```

```
## [1] 0.8069756
```

. . . and, that is a REALLY strong positive correlation! I need to find a package that will give me a p-value for the correlation, but there's no way that that's not going to be significant. Other than that, I think what this tells us is that it's actually true that F-measure goes up with IAA, even in the weird case where F-measure is higher than IAA. Or, maybe ONLY in that case–who knows! Doubtful, but we can check it out. I guess that the other thing that it tells us is that in the aggregate, all of this stuff suggests that this is non-weird data–it's just the F > IAA that's bizarre. Everything else looks typical.