# Validity

*Andrew Johnson*

## Introduction

Most of the validity analyses that you will undertake are based on a simple correlation coefficient, but there are two analyses that require a bit more analytic sophisticiation: (1) the construction of Bland-Altman plots; and (2) the construction of a "confusion matrix" (used in the calculation of sensitivity, specificity, and predictive value). We will discuss each of these methods in this demonstration.

## Bland-Altman Plots

The Bland-Altman plot is sometimes called a *Tukey Mean-Difference Plot*, owing to the fact that the plot was first described by John Tukey. It is intended to see if there is a marked difference between two measures, in terms of their prediction of a common construct. The plot is quite straightforward to create - it's simply a plot of the average of the two measures (on the x-axis) versus the difference between the two measures (on the y-axis).

Our baseline assumption when undertaking to use this graphical technique is that both measures are effective in their assessment of the construct. In other words, we assume that each of the measures are "reasonably valid." Thus, by computing the mean between the two measures, we are coming up with our best estimate of the "true score" for the construct. We can then consider the difference between scores on the two measures to be an estimate of the "bias" that is represented by the different methods of measurement at all measured levels of the construct.

Let's take a look at some example data. In this experiment, we wanted to evaluate the usefulness of a Wii balance board, in assessing the length of the centre of pressure pathway during quiet stance (eyes open, feet apart). The gold standard for this assessment is a fixed-mount forceplate, that is embedded into concrete. We want to see if there is a significant difference between these two measures, when they are used within a sample of individuals with Parkinson's disease.

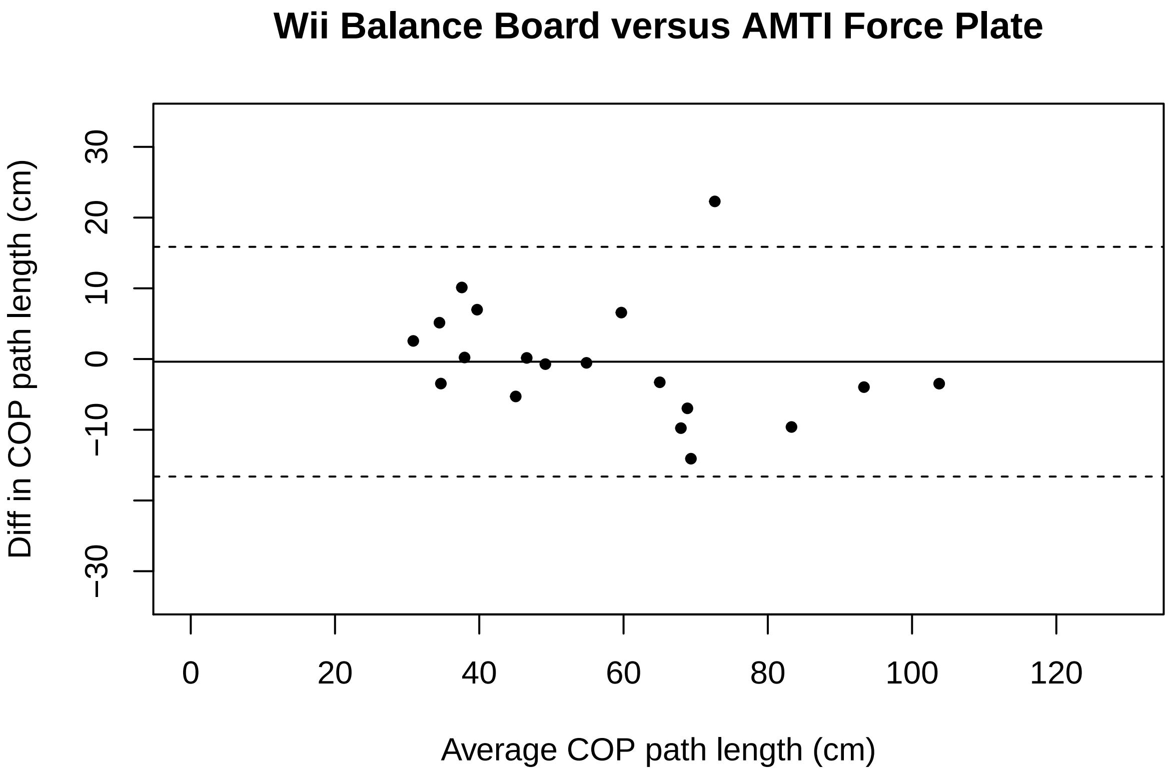
```
FP <- c(46.66, 38.07, 78.48, 43.19, 91.35, 42.64, 42.41, 37.05,
        63.38, 32.13, 62.30, 32.94, 102.02, 54.59, 62.98, 48.8,
        63.07, 83.8, 65.37)
Wii <- c(46.50, 37.85, 88.09, 36.21, 95.32, 32.52, 47.70, 31.91,
         66.67, 29.57, 76.39, 36.41, 105.50, 55.13, 56.41, 49.52,
         72.84, 61.51, 72.34)
```

There is no built-in function for the creation of Bland-Altman plots, and so I have written one. This function automates the creation of the plot, the labelling of the axes, and the creation of the reference lines. I have also built in functionality for plotting a standardized plot (i.e., a plot of the z-scores, rather than the raw scores). The function ("baplot.R") is posted to the module website. You may load it into your R environment by placing it in your working directory and typing:

```
source("baplot.R")
```

We can now create a Bland-Altman plot from our data.

```
baplot(Wii, FP,
       main="Wii Balance Board versus AMTI Force Plate",
       xlab="Average COP path length (cm)",
       ylab="Diff in COP path length (cm)",
       std=FALSE)
```



We are looking for two things within this plot.

1) The extent to which there is a consistent over- or under-estimation of the construct, by either measure. This would be demonstrated by a distribution with points that were predominantly above the mean reference line.

2) The range of values that the difference between the measures takes on - sometimes called the "margin of error". This can be evaluated by looking at the separation between the two dotted lines. The best way to identify whether or not the margin of error is acceptable is to consider the clinical / practical applications of the measurement instrument - if the disparity between the two instruments is acceptable, then the measures may be considered to be roughly equivalent.

For our example, we can see that the Wii Balance Board does not consistently over- or under-estimate the centre of pressure pathway. Furthermore, the differences between the measures falls within a confidence interval that extends approximately 16cm on either side of zero. In other words, the margin of error is approximately 16cm. We would consider this to be an acceptable margin of error, and thus we can conclude that the Wii Balance Board is an acceptable substitute measurement tool for a "professional-grade" force-plate.

# Constructing and Using a Confusion Matrix

The establishment of predictive validity requires a simple correlation coefficient when the criterion variable is a continuous variable, but what about the case when we are evaluating the ability of a measure to replicate (or approximate) a diagnosis? When the criterion variable is a dichotomous variable (as is the case when we are predicting a diagnosis), we are better to use statistics that capture the ability of a measure to approximate this classification. For this, we are better served with epidemiological statistics such as *sensitivity* and *specificity*.

Both of these indicators can be estimated by a very specific frequency table called a *confusion matrix*. In a confusion matrix, we identify the number of individuals that are *true positives* (i.e., individuals that are identified as "positive cases" by both the measure being validated, and by the gold standard), and individuals that are *true negatives* (i.e., individuals that are identified as "negative cases" by both the measure being validated and by the gold standard). All other cases are, by definition, *false* (be they *false positives* or *false negatives*). We can use the *confusion matrix* to estimate sensitivity and specificity.

Although it would be a simple matter to construct a confusion matrix using the standard `table` function, the `confusionMatrix` function within the `caret` package automates much of the drudgery within the sensitivity and specificity calculations, and so it is worthwhile introducing here.

We will need the `caret` package for the `confusionMatrix` function, and the `MASS` package for the `lda` function (used to fit the model that we are using for our classification function).

```r
library(caret)
library(MASS)
```

We will be using the famous Anderson iris data (sometimes attributed to Fisher, as he used the data to describe the use of statistics in taxonomy). This data presents the measurements, in centimeters, of the variables sepal length and width, and petal length and width, for 50 flowers for each of 3 species of iris. The species are iris setosa, iris versicolor, and iris virginica. We will use all four of the sepal and petal characteristics, in an effort to re-capture the classification specified within the data.

```r
fit <- lda(Species ~ ., data = iris)
predicted <- predict(fit)$class
```

Note that this syntax has `Species` predicted by ".". in this context, the "." means that the criterion variable will be predicted by all other variables within the dataset. If we had viewed the `fit` object, we would have seen the output from the linear discriminant analysis that we ran with the `lda` function. We don't really need the coefficients of the linear discriminants - just the classification results. In other words, we want to look at the classifications that the model would make, on the basis of the four sepal and petal characteristics.

We can use the `table` function to cross-tabulate the predicted values with the actual values from the original dataset.

```r
irisCrossTabs <- table(predicted, iris$Species)
irisCrossTabs
```

```
##
## predicted    setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         1
##   virginica       0          2        49
```

This tells us that our model is excellent at predicting iris species - particularly with regards to the *setosa* varietal.

You'll note that we are looking at a 3x3 matrix, rather than the typical 2x2 matrix used in confusion matrix demonstrations. This is where the `confusionMatrix` shines - it computes sensitivity and specificity by comparing each factor level to the remaining levels, thereby automating a few additional steps that you would normally need to perform by hand.

```
confusionMatrix(irisCrossTabs)
```

```
## Confusion Matrix and Statistics
##
##
## predicted    setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         1
##   virginica       0          2        49
##
## Overall Statistics
##
##                Accuracy : 0.98
##                  95% CI : (0.9427, 0.9959)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.97
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.9600           0.9800
## Specificity                 1.0000            0.9900           0.9800
## Pos Pred Value              1.0000            0.9796           0.9608
## Neg Pred Value              1.0000            0.9802           0.9899
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3200           0.3267
## Detection Prevalence        0.3333            0.3267           0.3400
## Balanced Accuracy           1.0000            0.9750           0.9800
```

In addition to the standard epidemiological information that you could have calculated by hand (i.e., sensitivity, specificity, positive predictive value, and negative predictive value), this function also provides us with a confidence interval around the overall accuracy rate. Thus, we can say that we are 95% certain that we can classify an iris into one of these three species using only their petal and sepal characteristics, with 94.27% to 99.59% accuracy.