Abstract

Agreement between methods for quantitative measurements are typically assessed by computing limits of agreement between the methods and/or by illustration through Bland-Altman plots. We consider the situation where measurement methods are considered randomly selected from a population of methods, and discuss how the underlying linear mixed effects model can be extended to this situation. This is relevant when, for example, the methods represent raters/judges that are used to score specific individuals or items.

In this case we will not want estimates pertaining to the specific methods, but consider the variation between the actually involved methods (i.e. raters) as an extra source of variation to take into account when generalizing to actual clinical performance of a method.

Estimation of limits of agreement are shown for a dataset of spatial perception of humans, and example code that present the implementation and analysis in R are provided.

Keywords: agreement, method comparison, random raters, mixed models.

1 Introduction

Comparison of methods for quantitative data is concerned with how well two methods agree on the measurement of an item. The interest is not on testing a hypothesis that the mean of the the two methods are identical, but on estimating the size and the standard deviation of their differences.

Considered a mixed effect model for computing the prediction limits (often termed the limits of agreement) for designs where replicate measurements
on each item may be present. This approach provides a full statistical model that can be used to compute the limits of agreement in the presence of replicate measurements instead of relying on summary measures. Moreover, it forces the investigator to focus on the nature of replicate measurements: are they exchangeable within each method by item stratum or only within items (i.e., the replicates are linked).

Here we consider the situation where we are not interested in determining the bias and agreement between two specific methods but regard the methods as a random sample of possible methods. This is relevant when a group of judges/raters are asked to rate a set of items using a predetermined scale. This is a common occurrence, for example when medical doctors are asked to give second or third opinions on measurements taken on patients. We are interested in how well medical doctors in general agree on the scoring of a particular condition, more so than determining how well specific doctors compare to each other.

Section ?? sets up the necessary mixed effect model that accommodates random raters, and section ?? extends the model to the situation where replicate measurements from the methods are available on each item. Section ?? discusses repeatability while the model is applied to a real dataset on random raters in Section ??.

2 Models for agreement among random raters

The classical setup for comparison of measurements methods is one where exactly one measurement is taken with each method on each item (here, item refers to for example individual, sample, or image). The limits of agreement for two methods are then computed as the prediction interval for the difference between future measurements taken by the two methods on a new individual. The focus of such studies is the comparison (and possibly prediction between) a few specific methods of interest.

When multiple raters are compared, the raters play the role of methods, but we are not interested in the difference between any two specific raters. Instead we consider each rater as a “random” rater from an (essentially infinite) population of raters. Thus, when we compute the prediction limits for the difference between measurements by two randomly chosen raters we should ensure that the variation between the random raters is taken into account in the modeling.

If we carry over the two-way analysis of variance model from the classical Bland-Altman set-up (??) we can model the value for a measurement taken on
item $i$ by rater $m$ as:

$$Y_{mi} = \mu_i + b_m + e_{mi},$$

$$b_m \sim N(0, \xi^2),$$

$$e_{mi} \sim N(0, \sigma^2_m),$$

(1)

where $\mu_i$ is the “true” value for item $i$, $b_m$ is a random effect that models a random bias for measurements taken by rater $m$, $\xi$ is the variation of biases between raters, and $\sigma_m$ is the individual variation of rater $m$.

The “true” value for item $i$ is of course arbitrary; a constant may be added to all $\mu_i$s provided it is subtracted from the $b_m$s; but since we have constrained the mean of the $b_m$s to be 0, the $\mu_i$s represent the average assessment of item $i$ by the set of raters at hand. Note that since the raters (or more precisely, the variation between them) is the focus of interest, the $\mu_i$s are essentially nuisance parameters.

Note that each rater is allowed to have his own precision such that some raters can be very precise (i.e., have low residual variation) while others can be less precise.

The limits of agreement for the difference between measurements taken by two random raters ($m$ and $m'$) on the same item corresponds to the prediction interval of the difference $Y_{mi} - Y_{m'i}$. If measurements by different raters are assumed to be independent we get that the 95% limits of agreement (prediction interval) for the difference between the two raters on a new item is

$$0 \pm z \sqrt{V(Y_{mi} - Y_{m'i})} = 0 \pm z \sqrt{V(Y_{mi}) + V(Y_{m'i})},$$

(2)

where $z$ is the quantile corresponding to the desired level of the prediction interval. Under model (??), the 95% limits of agreement simplifies to

$$0 \pm 1.96 \times \sqrt{2 \times \xi^2 + \sigma^2_m + \sigma^2_{m'}}.$$  

(3)

Even without replicate measurements it is possible to estimate the individual residual variance for each rater, because the model (??) by the very nature of the randomness of raters must impose an assumption of 0 average difference between raters. This means that the estimate of the single rater’s variation is strongly dependent on the other raters’ results, because it is essentially the variation around the common means ($\mu_i$).

The model (??) is a two-way analysis of variance model with random row (i.e. meth-) effect. The simple version with identical residual variances between methods can be fitted with either lmer or lme, whereas there is no facility for stratum-specific variances implemented in lmer, lme must be used.
```r
> library(MethComp)
> data(Ancona)
> Ancona <- Meth(Ancona, 1, 2, 3, 4)

The following variables from the dataframe "Ancona" are used as the Meth variables:
meth: rater
item: item
repl: repl
y: score

#Replicates

<table>
<thead>
<tr>
<th>Method</th>
<th>3</th>
<th>#Items</th>
<th>#Obs: 510 Values:</th>
<th>min</th>
<th>med</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>112345</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>18</td>
<td>58.5</td>
<td>102</td>
</tr>
<tr>
<td>1212</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>20</td>
<td>69.0</td>
<td>145</td>
</tr>
<tr>
<td>123</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>12</td>
<td>77.0</td>
<td>200</td>
</tr>
<tr>
<td>1234</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>24</td>
<td>68.5</td>
<td>130</td>
</tr>
<tr>
<td>12345</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>17</td>
<td>85.0</td>
<td>147</td>
</tr>
<tr>
<td>123456</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>19</td>
<td>72.0</td>
<td>120</td>
</tr>
<tr>
<td>1508</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>13</td>
<td>47.5</td>
<td>120</td>
</tr>
<tr>
<td>456</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>20</td>
<td>80.0</td>
<td>130</td>
</tr>
<tr>
<td>55555</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>15</td>
<td>79.0</td>
<td>160</td>
</tr>
<tr>
<td>A1268796D</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>23</td>
<td>43.5</td>
<td>130</td>
</tr>
<tr>
<td>butterscotch</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>23</td>
<td>78.5</td>
<td>200</td>
</tr>
<tr>
<td>BxC</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>14</td>
<td>74.5</td>
<td>130</td>
</tr>
<tr>
<td>delaunay</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>14</td>
<td>45.5</td>
<td>120</td>
</tr>
<tr>
<td>etabeta</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>18</td>
<td>73.0</td>
<td>150</td>
</tr>
<tr>
<td>prova1</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>21</td>
<td>73.0</td>
<td>120</td>
</tr>
<tr>
<td>r566f6688</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>20</td>
<td>62.5</td>
<td>140</td>
</tr>
<tr>
<td>vf</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>15</td>
<td>31.5</td>
<td>75</td>
</tr>
</tbody>
</table>

> library(lme4)
> print(lmer(y ~ item - 1 + (1 | meth), data = subset(Ancona, repl == +
|             |                       |                   | 1)), correlation = FALSE)

Linear mixed model fit by REML
Formula: y ~ item - 1 + (1 | meth)
Data: subset(Ancona, repl == 1)
AIC 1502  BIC 1539  logLik  -738.9  deviance 1530  REMLdev 1478
Random effects:
  Groups   Name     Variance Std.Dev.
  meth     (Intercept) 211.07   14.528
  Residual    420.76    20.512
Number of obs: 170, groups: meth, 17

Fixed effects:
  Estimate Std. Error  t value
  item24    22.176     6.096    3.638
```
item32  35.647  6.096  5.847
item48  48.765  6.096  7.999
item52  56.706  6.096  9.302
item72  76.706  6.096 12.583
item80  77.941  6.096 12.785
item88  84.941  6.096 13.933
item96  93.412  6.096 15.323
item100 104.000  6.096 17.060
item120 107.588  6.096 17.648

> print(lme(y ~ item - 1, random = ~1 | meth, data = subset(Ancona, + repl == 1)))

Linear mixed-effects model fit by REML
Data: subset(Ancona, repl == 1)
Log-restricted-likelihood: -738.9173
Fixed: y ~ item - 1
item24 item32 item48 item52 item72 item80 item88 item96
  22.17647  35.64706  48.76471  56.70588  76.70588  77.94118  84.94118  93.41176
item100 item120
  104.00000  107.58824

Random effects:
  Formula: ~1 | meth
    (Intercept) Residual
  StdDev:  14.52816 20.51246

Number of Observations: 170
Number of Groups:  17

> print(lme(y ~ item - 1, random = ~1 | meth, weights = varIdent(form = ~1 | + meth), data = subset(Ancona, repl == 1)))

Linear mixed-effects model fit by REML
Data: subset(Ancona, repl == 1)
Log-restricted-likelihood: -710.1988
Fixed: y ~ item - 1
item24 item32 item48 item52 item72 item80 item88 item96
  22.75546  30.82929  47.71897  53.76357  72.03878  73.15217  86.55763  92.56887
item100 item120
  103.14288 116.44149

Random effects:
  Formula: ~1 | meth
    (Intercept) Residual
  StdDev:  11.70443 32.64711

Variance function:
Structure: Different standard deviations per stratum
Formula: ~1 | meth
Parameter estimates:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>112345</td>
<td>1.0000000</td>
</tr>
<tr>
<td>1212</td>
<td>0.4926918</td>
</tr>
<tr>
<td>123</td>
<td>0.9346331</td>
</tr>
<tr>
<td>1234</td>
<td>0.9888461</td>
</tr>
<tr>
<td>12345</td>
<td>0.4789892</td>
</tr>
<tr>
<td>123456</td>
<td>0.1904552</td>
</tr>
<tr>
<td>1508</td>
<td>0.4958086</td>
</tr>
<tr>
<td>456</td>
<td>0.7091460</td>
</tr>
<tr>
<td>5555</td>
<td>0.5035090</td>
</tr>
<tr>
<td>A1268796D</td>
<td>0.7425866</td>
</tr>
<tr>
<td>BxC</td>
<td>1.1235760</td>
</tr>
<tr>
<td>03B6f6688</td>
<td>0.2586849</td>
</tr>
<tr>
<td>0.2005935</td>
<td>0.2005935</td>
</tr>
<tr>
<td>0.3197186</td>
<td>0.3197186</td>
</tr>
<tr>
<td>r566f6688</td>
<td>0.5227227</td>
</tr>
<tr>
<td>vf</td>
<td>0.4354267</td>
</tr>
<tr>
<td>BxC</td>
<td>0.5533831</td>
</tr>
</tbody>
</table>

Number of Observations: 170
Number of Groups: 17

When we wish to compare the agreement between two random raters we need to take the variation in the individual rater precisions into account since we do not know which particular two raters we will use. The variance for rater \( m \) on item \( i \) is

\[
V(Y_{mi}) = \xi^2 + \sigma^2_m
\]

where we let the individual residual variances \( \sigma^2_m \) follow some distribution \( F \) which has support on the positive numbers to account for the heterogeneity of variances between raters. Thus, the distribution \( F \) is a central quantity in the prediction of differences between rates, it represents the distribution of “skills” among available raters. If prior knowledge about the precision distribution is available then that information can be used to model \( F \). In the following we assume that we have no prior knowledge about the precision and/or that we have too little information to be able to verify any assumptions we might have had.

The law of total variance provides the variance of a measurement from a randomly chosen rater from the population when applied to a fixed item:

\[
V(Y_{mi}) = V_{\sigma}(E(Y_{mi}|\sigma_m)) + E_{\sigma}(V(Y_{mi}|\sigma_m)) = V_{\sigma}(\mu_i) + E_{\sigma}(\xi^2 + \sigma^2_m) = \xi^2 + E_{\sigma}(\sigma^2_m)
\]

Now, the outer variances and expectations might as well be computed under the distribution of \( \sigma^2_m \) as under \( \sigma_m \). So by this token, the empirical counterpart of this results can be either of:

\[
\hat{\xi}^2 + E(\hat{\sigma}^2_m) \approx \hat{\xi}^2 + \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^2_m \\
\approx \hat{\xi}^2 + \left( \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}_m \right)^2
\]
where $M$ is the number of raters available in the dataset.

What the formula (??) provides us with is merely a statement that the variance of a measurement by a randomly chosen rater is the sum of between rater variance and some average within-rater variance. The calculation of the latter is pretty arbitrary, depending on the scale on which we do the calculation. Using an empirical mean (regardless of how shrewdly we choose the scale) may therefore not be the best bet of a robust estimate. The estimate is anyway going to be heavily reliant on the assumption that the sample of raters at hand is representative of the population of future raters.

It may therefore be recommendable to use the median instead, so that only the overall location of the distribution of actual raters influence the estimate.

Hence, the estimate for the 95% limits of agreement becomes

$$0 \pm 1.96 \times \sqrt{2 \times \left( \hat{\sigma}^2 + \text{median}(\hat{\sigma}^2_m) \right)}$$

The value 1.96 can of course be replaced by a suitable quantile from the $t$-distribution; the normal convention is to use 2 (which incidentally is the 97.5% quantile of the $t$-distribution with 60 d.f.).

### 3 Model for replicate measurements

As mentioned above, even if raters are measuring each item only once, we can still estimate the residual variation of each rater; it will simply be the variation around the common item-means. This is the point where the situation with random raters differs from comparing specific measurement methods. Because raters are considered random, we must necessarily assume that the mean difference between raters is 0, and by that token that any deviation from 0 is random. This means that the implicit assumption of randomly chosen raters is heavily exploited.

If we really want to assess the variability of the precision of the raters it is mandatory to have replicate measurements of each item by each rater (“method”); in that case observations are classified by replicate too, so we need a more elaborate model. However, replicates come in two guises: linked and exchangeable.

Replicates are called *linked* if the first replicate by all raters are made at the same time (or under similar circumstances), and if the second replicate by all raters are made at the same time too, etc. This means that the numbering of the replicates carries some information on some similar circumstance of the measurement.

Replicates are called exchangeable if this is not the case, that is if replicate numbering can be freely exchanged within a given $(m, i)$. 


This distinction has implications for the models used to describe data, and to derive predictions.

### 3.1 Linked replicates

If replicates are linked, we use the following extension of the simple model (7) for the measurement $y_{mir}$ by rater $m$, on item $i$, replicate $r$:

\[
Y_{mir} = \mu_i + b_m + a_{ir} + c_{mi} + e_{mir},
\]

\[
b_m \sim N(0, \xi^2)
\]

\[
a_{ir} \sim N(0, \omega^2)
\]

\[
c_{mi} \sim N(0, \tau_m^2)
\]

\[
e_{mir} \sim N(0, \sigma_m^2)
\]

(7)

Note that two new variance components have been added relative to the simple model (7).

The first, $\omega^2$, is the variation between replication instances; as such it is in principle irrelevant for the comparison of raters. The second, $\tau_m^2$, is the variation between items within each method — a method-specific interaction with the items. It represents the variability of a rater across items; that is how a specific rater’s measurements varies relative to the average measurement by all raters on a given item. Thus, this is a variance component whose size for the individual rater is very strongly tied to the concept of randomly chosen raters, in the sense that the estimates of $\tau_m^2$ will depend on the sample of raters to a much larger degree than will the estimates of $\sigma_m^2$ which are entirely estimates of the individual raters variation around his own measurements.

Model (7) is a standard variance component model as is the model defined by (7). When we derive the LoA in this case we use the same calculations as in the simple case; but we note that the measurements by different methods are no more independent because of the term $a_{ir}$. However these terms cancel when computing $V(Y_{mir} - Y_{mir'})$, because indices $i$ and $r$ are identical in the two terms.

By the same arguments that lead to (7), we get LoA between two randomly chosen raters as:

\[
0 \pm 1.96 \times \sqrt{2 \times \xi^2 + \tau_m^2 + \tau_{m'}^2 + \sigma_m^2 + \sigma_{m'}^2}.
\]

Thus when considering random raters we will not only need to estimate the average residual variation between raters, but also the average item by rater
variation between raters; which leads to limits of agreement estimated as:

\[
0 \pm 1.96 \times \sqrt{2 \times (\hat{\xi}^2 + \text{median}(\hat{\tau}_m^2 + \hat{\sigma}_m^2))}
\]

\[
0 \pm 1.96 \times \sqrt{2 \times (\hat{\xi}^2 + \text{median}(\hat{\tau}_m^2) + \text{median}(\hat{\sigma}_m^2))}
\]

(8)

3.2 Exchangeable replicates

When we have exchangeable replicates, we will use a model very similar to the one used above, except that the variance component \(\omega\) (corresponding to the random effects \(a_{ir}\)) is absent. However, this has no effect on the total variation, so formula (8) above for the limits of agreement will be the same in the case of exchangeable replicates. However, note that the model fitted to obtain the relevant estimates is different for exchangeable replicates compared to linked replicates.

4 Repeatability

The limits of agreement are not always the only issue of interest — the assessment of method specific repeatability and reproducibility may be of interest in their own right. In particular, if a method has large variation between replicates on the same item then the repeatability and agreement with other methods will be poor. Repeatability can only be assessed when replicate measurements by each method are available.

In classical assessment of specific measurement methods, the repeatability coefficient for a method is defined as the upper limit of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances. A small repeatability coefficient corresponds to good repeatability (i.e., little variation among replicates) while a large repeatability coefficient corresponds to poor repeatability.

If the standard deviation of a measurement is \(\sigma\) then the repeatability coefficient is \(1.96 \times \sqrt{2\sigma^2} \approx 2.8\sigma\).

In the case of random raters, the repeatability coefficient has a slightly different meaning, because we cannot hinge it on estimates of variances from any specific rater — they will just be a random set of variances. Instead the repeatability must refer to the average / median / expected repeatability, or even, if we cling to the classical definition, the variability of the repeatability as we may expect to see it in a sample of raters.
When replicates are exchangeable, the difference between two replicate measurements, \( r \) and \( r' \) taken by rater \( m \) on item \( i \) is

\[ Y_{mir} - Y_{mir'} = e_{mir} - e_{mir'} \]  \hspace{1cm} (9)

so the repeatability is based only on the residual standard deviation, \( i.e. \ 2.8\sigma_m \).

For linked replicates the difference becomes

\[ Y_{mir} - Y_{mir'} = a_{ir} - a_{ir'} + e_{mir} - e_{mir'} \]  \hspace{1cm} (10)

and the variation between replicates taken on the same item should be factored into the calculation of the repeatability coefficient which becomes \( 2.8\sqrt{\omega^2 + \sigma^2_m} \).

The latter argument assumes that the variability between replication occasions can be consider representative of future scenarios. If the replicates are taken under substantially different circumstances, then the variance component \( \omega^2 \) may be considered irrelevant for the repeatability and the repeatability coefficient should be based on the measurement errors alone, \( i.e. \ use \ 2.8\sigma_m. \) However, if indeed the replicates are taken under substantially different circumstances it may be argued that we are not really measuring the same item repeatedly. Instead the effects of differing replication circumstances could be modeled by a systematic effect. Hence there is no subject-matter-free way of defining repeatability from the variance components in the models.

Thus, repeatability varies between raters too, and in the case of random raters essentially only the median of these makes sens as a summary estimate — half of the times one will have a rater with better repeatability, half of the times one with worse.
Worked example: spatial perception of point swarms

Some people have keen spatial perception and are able to almost instantaneously give a reasonable guess of, for example, the number of individuals in a crowd. We collected data on spatial perception from participants attending a course on comparison of measurement methods. Thus, we will use attendees of the course as a random selection of raters and will try to determine how well two random raters agree on the number of points in a point swarm.

The data set

10 pictures of scattered points were generated, and the purpose of the exercise was to estimate the number of points in each picture. Each picture were shown to the 17 raters three times (rotated and in random order to prevent recollection of the pictures), without telling them about the replication structure.

The dataset is available as the dataset Ancona in the MethComp package for R:

```R
> library( MethComp )
> data( Ancona )
> str( Ancona )
'data.frame': 510 obs. of 4 variables:
  $ rater: Factor w/ 17 levels "112345","1212",..: 3 3 3 3 3 3 3 3 3 3 ...
  $ item : int 120 48 88 32 24 100 52 80 72 96 ...
  $ repl : int 1 1 1 1 1 1 1 1 1 1 ...
  $ score: int 122 40 45 67 12 134 125 76 84 90 ...

> Ancona <- Meth( Ancona, meth=1, y=4 )
```

The following variables from the dataframe "Ancona" are used as the Meth variables:

- `meth`: rater
- `item`: item
- `repl`: repl
- `y`: score

<table>
<thead>
<tr>
<th>Method</th>
<th>#Replicates</th>
<th>#Items</th>
<th>#Obs</th>
<th>Values:</th>
<th>min</th>
<th>med</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>112345</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>18 58.5 102</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1212</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>20 69.0 145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>12 77.0 200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1234</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>24 68.5 130</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12345</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>17 85.0 147</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>123456</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>19 72.0 120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1508</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>13 47.5 120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>456</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>20 80.0 130</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>555555</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>15 79.0 160</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We can get an overview of the data, using the `plot` method:

```r
> plot( Ancona, cex=0.3, cex.name=0.8 )
```

which is shown in Figure ???. There might be a case for log-transformation of the data, although it may not be strong. An overview of the log-transformed data is obtained by the following command (output figure not shown).

```r
> plot( Ancona, cex=0.3, cex.name=0.8, Transform="log" )
```

The relatively large number of raters makes it difficult to get an overview of all raters simultaneously so an alternative way of summarizing the data graphically is to plot the methods (and replicates within methods) for each item as shown in Figure ???, separately for original scale and log-transformed data.

**Claus** Det er dig der har introduceret lattice, så kan du lige finde ud af hvordan man indlægger vandrette linjer: En sort svarende til det sande antal punkter, og tre farvede svarende til det empiriske gennemsnit over rates inden for replicates (de er jo linked...). Som du ser har jeg skukkeseret layout og fjernet sødsuppefarven fra overskrifterne. Det er sikkert bare et spørgsmål om at RTFM.

```r
> library( Epi )
> library(lattice)
> # Define colors for the replicates
> clrs <- c("red", "forestgreen", "blue")
> # Order the methods (i.e. random raters) by mean log-score
> tm <- with( Ancona, tapply( log(y), meth, mean ) )
> Ancona$meth <- Relevel( Ancona$meth, names(sort(tm)) )
> print(xyplot( y ~ meth | item, data=Ancona,
+           col = clrs,
+           ylab="Score", xlab="Rater", pch=20, group=repl,
+           scales = list(x = list(draw = FALSE)),
+           par.settings = list(strip.background=list(col=gray(0.95))),
+           layout=c(5,2),
+           panel = panel.superpose,
+           panel.groups = function(x, y, ..., lty, lwd, col.line)
```

12
> print(xyplot(log(y) ~ meth | item, data=Ancona, 
+           col = clrs,
+           
Figure 1: Pairwise comparison of the raters. Note that replicate measurements are represented by three different points. Original count scale.
We can get estimates of the variance components using the `BA.est` function on the `Meth` object `Ancona`, with the `random.raters` argument set to `TRUE`:

```r
> system.time(
+ RR <- BA.est( Ancona, linked = TRUE,
+ random.raters = TRUE ) )

user  system elapsed
101.01   0.15  105.22

> RR

Variance components (sd):

<table>
<thead>
<tr>
<th></th>
<th>IxR</th>
<th>MxI</th>
<th>M</th>
<th>res</th>
</tr>
</thead>
<tbody>
<tr>
<td>vf</td>
<td>1.455</td>
<td>23.703</td>
<td>10.762</td>
<td>18.727</td>
</tr>
<tr>
<td>delaunay</td>
<td>1.455</td>
<td>2.435</td>
<td>10.762</td>
<td>18.903</td>
</tr>
<tr>
<td>A1268796D</td>
<td>1.455</td>
<td>3.479</td>
<td>10.762</td>
<td>43.707</td>
</tr>
<tr>
<td>1508</td>
<td>1.455</td>
<td>0.002</td>
<td>10.762</td>
<td>17.143</td>
</tr>
<tr>
<td>112345</td>
<td>1.455</td>
<td>16.076</td>
<td>10.762</td>
<td>15.641</td>
</tr>
<tr>
<td>r566f6688</td>
<td>1.455</td>
<td>0.013</td>
<td>10.762</td>
<td>10.524</td>
</tr>
<tr>
<td>123456</td>
<td>1.455</td>
<td>1.491</td>
<td>10.762</td>
<td>12.426</td>
</tr>
<tr>
<td>1234</td>
<td>1.455</td>
<td>8.159</td>
<td>10.762</td>
<td>16.650</td>
</tr>
<tr>
<td>prova1</td>
<td>1.455</td>
<td>4.509</td>
<td>10.762</td>
<td>19.170</td>
</tr>
<tr>
<td>1212</td>
<td>1.455</td>
<td>0.013</td>
<td>10.762</td>
<td>20.616</td>
</tr>
<tr>
<td>BxC</td>
<td>1.455</td>
<td>6.226</td>
<td>10.762</td>
<td>31.419</td>
</tr>
<tr>
<td>etabeta</td>
<td>1.455</td>
<td>8.242</td>
<td>10.762</td>
<td>8.431</td>
</tr>
<tr>
<td>456</td>
<td>1.455</td>
<td>9.659</td>
<td>10.762</td>
<td>11.026</td>
</tr>
<tr>
<td>12345</td>
<td>1.455</td>
<td>6.367</td>
<td>10.762</td>
<td>14.805</td>
</tr>
<tr>
<td>123</td>
<td>1.455</td>
<td>12.652</td>
<td>10.762</td>
<td>12.922</td>
</tr>
<tr>
<td>55555</td>
<td>1.455</td>
<td>7.640</td>
<td>10.762</td>
<td>16.502</td>
</tr>
<tr>
<td>butterscotch</td>
<td>1.455</td>
<td>0.007</td>
<td>10.762</td>
<td>12.104</td>
</tr>
</tbody>
</table>
```
What is printed by the last command is merely the estimated variance components using the model where the conversion between methods have intercept is 0 and slope 1. The limits of agreement between two random raters is stored in the LoA element returned by \texttt{BA.est}, so the limits are

\begin{verbatim}
> RR$LoA

               Mean   Lower   Upper     SD
Rand. rater - rand. rater 0  -68.0205  68.0205 34.01025
\end{verbatim}

If we wish to use the median of the variance components instead of the mean to get a more robust estimate of the limits of agreement then we have to extract the median of the relevant variance components and use them together with formula (??). We also make the computations using the mean, illustrating the calculation of the LoA embedded in \texttt{BA.est}:

\begin{verbatim}
> ( RRmd <- apply( RR$VarComp, 2, median ) )

     MxI    M    res
     1.455358 6.225722 10.762222 16.502186

> sqmean <- function( x ) sqrt(mean(x^2))
> ( RRmn <- apply( RR$VarComp, 2, sqmean ) )

     MxI    M    res
     1.455358 9.030594 10.762222 19.518490
\end{verbatim}

According to formula (??) we get the LoA as plus/minus:

\begin{verbatim}
> ( LoA.md <- 2 * sqrt( 2 * ( RRmd["M"]^2 +
+ RRmd["MxI"]^2 +
+ RRmd["res"]^2 ) ) )

     M
     58.4402

> ( LoA.mn <- 2 * sqrt( 2 * ( RRmn["M"]^2 +
+ RRmn["MxI"]^2 +
+ RRmn["res"]^2 ) ) )

     M
     68.0205
\end{verbatim}

We get slightly smaller limits of agreement when we use the median of the estimated variance components instead of the mean.
5.1.1 Log-transformed data

In the plots in figure ?? there are pretty clear indications that a log-transform might provide a better fit to the data. This is easily accommodated using the Transform argument to BA.est

```r
> system.time(
+ RRl <- BA.est( Ancona, linked = TRUE,
+ random.raters = TRUE,
+ Transform = "log" ) )
```

```
user  system elapsed
59.55   0.31  62.04
```

```r
> RRl$LoA
```

```
                        Mean    Lower    Upper         SD
Rand. rater - rand. rater 0  -1.01247  1.01247  0.5062349
```

```r
> ( RRlmd <- apply( RRl$VarComp, 2, median ) )
```

```
                        IxR   MxI    M   res
0.02760046  0.05507546 0.16001402 0.21470780
```

```r
> ( LoA.log.md <- 2 * sqrt( 2 * ( RRlmd["M" ]^2 +
+ RRlmd["MxI"]^2 +
+ RRlmd["res"]^2 ) ) )
```

```
                        M
0.7732386
```

These figures are based on s.d.s of the log-transformed data. Thus, the variance components represent coefficients of variation, see e.g ?, chapter 9. To convert this to a sensible number applicable on the original (count) scale, we must take the exponential of these figures to get a multiplicative factor. Thus in the case of the log-transformed data the LoA will be multiplicative factor:

```r
> LoA.log.mn <- RRl$LoA[3]
> mLoA <- exp( c(LoA.log.mn,LoA.log.md) )
> names( mLoA ) <- c("mult(mean)","mult(med)")
> mLoA
```

```
mult(mean)  mult(med)
2.752390   2.166772
```
5.1.2 Visualization of the LoA

We can illustrate the relationship between the two approaches (mean and median) and the two transformations (identity or log) by making a graph for converting between two random raters. The mean conversion line will necessarily be the identity line, so the plot will basically show envelopes of where we can expect to find points.

```r
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, las=1 )
> plot( NA, xlim=c(0,200), ylim=c(0,200),
+ xlab="Count", ylab="Count",
+ xaxs="i", yaxs="i" )
> abline( v=seq(0,200,10), h=seq(0,200,10), col=gray(0.8) )
> for( pm in c(-1,1) )
+ {
+ abline( pm*LoA.mn, 1, lwd=2, lty=1, col="black" )
+ abline( pm*LoA.md, 1, lwd=2, lty=3, col="black" )
+ abline( 0, exp( pm*LoA.log.mn), lwd=2, lty=1, col="blue" )
+ abline( 0, exp( pm*LoA.log.md), lwd=2, lty=3, col="blue" )
+ }
> # Plot points from random pairs of raters
> for( i in sample( levels(Ancona$meth) , 16 ) )
+ for( j in sample( setdiff(levels(Ancona$meth),i), 1 ) )
+ {
+ wd <- to.wide( subset( Ancona, meth %in% c(i,j) ) )
+ # random order
+ ro <- sample( 0:1, 1 )
+ points( wd[,3+ro], wd[,4-ro], col=clrs[wd$repl], pch=16, cex=0.5 )
+ }
```

Also note that the envelope based on the log-transformation is with straight lines because we have a model where the conversion between methods (raters) is forced to go through 0. We can of course show the same conversion on a plot of the differences versus the averages; it is merely a rotation of the plot. The conversion could be done using the y2DA function:

```r
> y2DA

function (A = 0, B = 1, S = NA)
{
  if (length(A) > 1) {
    S <- A[3]
    B <- A[2]
    A <- A[1]
  }
  res <- c(2 * A, 2 * (B - 1), 2 * S)/(B + 1)
  names(res) <- c("DA-int", "DA-slope", "DA-sd")
}
```
invisible(res)
}  
<environment: namespace:MethComp>

> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, las=1 )
> plot( NA, xlim=c(0,200), ylim=c(-100,100),
+ xlab="Average count", ylab="Count difference",
+ xaxs="i", yaxs="i" )
> abline( v=seq(0,200,10), h=seq(-100,100,10), col=gray(0.8) )
> for( pm in c(-1,1) )
+ {
+ abline( pm*LoA.mn, 0, lwd=2, lty=1, col="black" )
+ abline( pm*LoA.md, 0, lwd=2, lty=3, col="black" )
+ abline( 0, y2DA(0,exp( pm*LoA.log.mn))[2], lwd=2, lty=1, col="blue" )
+ abline( 0, y2DA(0,exp( pm*LoA.log.md))[2], lwd=2, lty=3, col="blue" )
+ }
> # Plot points from random pairs of raters
> for( i in sample( levels(Ancona$meth) , 16 ) )
+ for( j in sample( setdiff(levels(Ancona$meth),i), 1 ) )
+ {
+ wd <- to.wide( subset( Ancona, meth %in% c(i,j) ) )
+ # random order
+ ro <- sample( 0:1, 1 )
+ points( (wd[,3+ro]+wd[,4-ro])/2,
+ wd[,3+ro]-wd[,4-ro],
+ col=clrs[wd$repl], pch=16, cex=0.5 )
+ }

5.2 Repeatability

The repeatability coefficient for rater $m$ is calculated as $2 \times \sqrt{2\sigma_m^2} \approx 2.8\sigma_m$ provided we assume exchangeability for the replicates. For the Ancona data the individual repeatability coefficients are theM

> 2.8 * RR$VarComp[,"res"]

<table>
<thead>
<tr>
<th></th>
<th>vf</th>
<th>delaunay</th>
<th>A1268796D</th>
<th>1508</th>
<th>112345</th>
<th>r566f6688</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.43486</td>
<td>52.92740</td>
<td>122.38057</td>
<td>48.00165</td>
<td>43.79397</td>
<td>29.46593</td>
<td></td>
</tr>
<tr>
<td>123456</td>
<td>123456</td>
<td>prova1</td>
<td>1212</td>
<td>BxC</td>
<td>etabeta</td>
<td></td>
</tr>
<tr>
<td>34.79169</td>
<td>46.61890</td>
<td>53.67547</td>
<td>57.72389</td>
<td>87.97357</td>
<td>23.60766</td>
<td></td>
</tr>
<tr>
<td>456</td>
<td>456</td>
<td>123</td>
<td>55555</td>
<td>butterscotch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.87328</td>
<td>41.45524</td>
<td>36.18039</td>
<td>46.20612</td>
<td>33.89031</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and the mean / median repeatability coefficient are then:

> 2.8 * mean( RR$VarComp[,"res"] )

18
So broadly speaking we can say that the repeatability of academic point-counters (such as represented by the course participants at SISMEC in Ancona) is about 45 points, meaning that assessing the same picture twice by the same rater will produce two guesses closer than 45 with probability 95%.

If replicates are linked we should include the between-replicates variation in the calculation of the repeatability, if this is a variation which is considered unavoidable between replicates.

In the case of the Ancona experiment, it is debatable whether replicates consisting of showing a rotated version of the picture are replicates in the repeatability sense. But if this is considered so, then we should calculate the individual repeatability coefficients as:

\[
RC <- 2*\sqrt{2*(RR$VarComp[,"IxR"]^2 + RR$VarComp[,"res"]^2)}
\]

and the mean and median repeatability coefficient among raters are then:

\[
> c( mean(RC), median(RC) )
\]

5.2.1 Multiplicative effects

We saw that there was some indication that data was better described by a constant variance model for the log-transformed counts. In that case we should compute the repeatability on the multiplicative scale, i.e. we should basically do the same calculations on the log-transformed data:

\[
> mean( 2*\sqrt{2*(RRl$VarComp[,"IxR"]^{-2} + RRl$VarComp[,"res"]^{-2})} )
\]

\[
> median( 2*\sqrt{2*(RRl$VarComp[,"IxR"]^{-2} + RRl$VarComp[,"res"]^{-2})} )
\]

But these are (“expected”) upper limits of absolute differences of natural-log transformed data, so not readily interpretable. However if we take the exponential of these, we will get upper limit of a 95% prediction interval for the ratio of the larger to the smaller rating between two replicates:

\[
> \exp( c( mean( 2*\sqrt{2*(RRl$VarComp[,"IxR"]^{-2} + RRl$VarComp[,"res"]^{-2})} ),
+ median( 2*\sqrt{2*(RRl$VarComp[,"IxR"]^{-2} + RRl$VarComp[,"res"]^{-2})} ) ) )
\]

\[
[1] 2.081381 1.844637
\]
Figure 2: Illustration of the Ancona data. Each panel shows a specific item and the columns of dots within each panel represent the scores given by each rater. The colouring corresponds to replicates. Raters are sorted by their mean log-score, and are in the same order in all plots. The upper graph is counts on the original scale, the lower after log-transform. It appears that variances are reasonably homogeneous in the lower panels, but not in the upper. The full horizontal lines correspond to the mean score within replicates, the dashed line to the true number of points.
Figure 3: Visualization of the LoA from the Ancona data. The black lines are from a model using original count data, the blue lines from a model for log-transformed data. Dotted lines are from using the median to summarize the variances of the raters instead of the mean (full lines). The points are from 16 randomly chosen pairs of raters.
Figure 4: Visualization of the LoA from the Ancona data. The black lines are from a model using original count data, the blue lines from a model for log-transformed data. Dotted lines are from using the median to summarize the variances of the raters instead of the mean (full lines).
6 Discussion

Unlike the modeling approach of \textit{?}, we have addressed the question of precision of a single method of measurement which requires persons (raters) to produce measurements. In that vein the prediction from one rater to another will inevitably be a prediction that has zero bias, and where only the precision is involved.

Note that the approximation (\textit{??}) may be rather crude when the number of available raters is low. In that situation, it may be more reasonable to use a measure like the median or possibly the geometric mean of each variance component to obtain a more robust (albeit possibly slightly biased) estimate of variance.