Imperial College London

Bland–Altman plots, rank parameters, and calibration ridit splines

Roger B. Newson
r.newson@imperial.ac.uk
http://www.rogernewsonresources.org.uk

Department of Primary Care and Public Health, Imperial College London

To be presented at the 2019 London Stata Conference,
05–06 September, 2019
To be downloadable from the conference website at
http://ideas.repec.org/s/boc/usug19.html

Statistical methods for method comparison

- Scientists frequently compare two methods for estimating the same quantity in the same things.
- *For example*, medics might compare two methods for estimating disease prevalences in primary–care practices, or viral loads in patients.
- Sometimes, the comparison aims to measure **components of disagreement** between two methods, such as discordance, bias, and scale difference.
- And sometimes, the comparison aims to predict (or **calibrate**) the result of one method from the result of the other method.

Statistical methods for method comparison

- Scientists frequently compare two methods for estimating the same quantity in the same things.
- *For example*, medics might compare two methods for estimating disease prevalences in primary–care practices, or viral loads in patients.
- Sometimes, the comparison aims to measure **components of disagreement** between two methods, such as discordance, bias, and scale difference.
- And sometimes, the comparison aims to predict (or **calibrate**) the result of one method from the result of the other method.

Statistical methods for method comparison

- Scientists frequently compare two methods for estimating the same quantity in the same things.
- *For example*, medics might compare two methods for estimating disease prevalences in primary–care practices, or viral loads in patients.
- Sometimes, the comparison aims to measure **components of disagreement** between two methods, such as discordance, bias, and scale difference.
- And sometimes, the comparison aims to predict (or **calibrate**) the result of one method from the result of the other method.

Statistical methods for method comparison

- Scientists frequently compare two methods for estimating the same quantity in the same things.
- *For example*, medics might compare two methods for estimating disease prevalences in primary–care practices, or viral loads in patients.
- Sometimes, the comparison aims to measure **components of disagreement** between two methods, such as discordance, bias, and scale difference.
- And sometimes, the comparison aims to predict (or **calibrate**) the result of one method from the result of the other method.

Statistical methods for method comparison

- Scientists frequently compare two methods for estimating the same quantity in the same things.
- *For example*, medics might compare two methods for estimating disease prevalences in primary–care practices, or viral loads in patients.
- Sometimes, the comparison aims to measure **components of disagreement** between two methods, such as discordance, bias, and scale difference.
- And sometimes, the comparison aims to predict (or **calibrate**) the result of one method from the result of the other method.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner (“the Mentor”) was the more experienced of the two.
- The second examiner (“the Mentee”) was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner ("the Mentor") was the more experienced of the two.
- The second examiner ("the Mentee") was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner ("the Mentor") was the more experienced of the two.
- The second examiner ("the Mentee") was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner ("the Mentor") was the more experienced of the two.
- The second examiner ("the Mentee") was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner ("the Mentor") was the more experienced of the two.
- The second examiner ("the Mentee") was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

Example dataset: 176 anonymised double–marked exam scripts in medical statistics

- Our example dataset comes from a first–year medical statistics course in a public–health department that no longer exists[2].
- 176 medical students sat the course examination, and their scripts were double–marked by 2 examiners.
- The first examiner ("the Mentor") was the more experienced of the two.
- The second examiner ("the Mentee") was marking exam scripts for the first time, and did this in an all–night session, dosed heavily with coffee.
- Marks awarded by each examiner had integer values up to a maximum of 50, and were averaged between the 2 examiners to give a final mark awarded to each student.

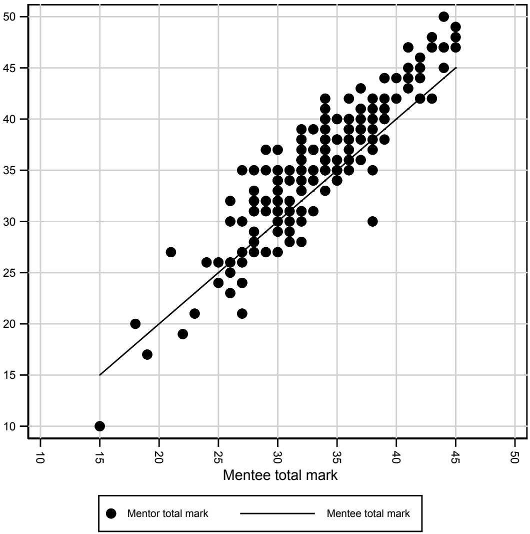
The dataset of students with pairwise marks

And here we `use` and `describe` the dataset, with 1 observation per exam script. The dataset is keyed by the variable `candno` (*anonymised* candidate number). The other variables are the mentor and mentee total marks, the mentor–mentee difference, and the mean of the mentor and mentee marks (awarded to the candidate).

```
. use candidate1, clear;

. desc, fu;

Contains data from candidate1.dta
  obs:           176
 vars:             5                          17 Jun 2019 18:01
 size:         1,584
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
candno          int     %9.0g                 Candidate number
atotmark        byte    %9.0g                 Mentor total mark
btotmark        byte    %9.0g                 Mentee total mark
dtotmark        byte    %9.0g                 Mentor-mentee difference in total mark
mtotmark        float   %9.0g                 Mean total mark (awarded)
--------------------------------------------------------------------------------
Sorted by: candno
```

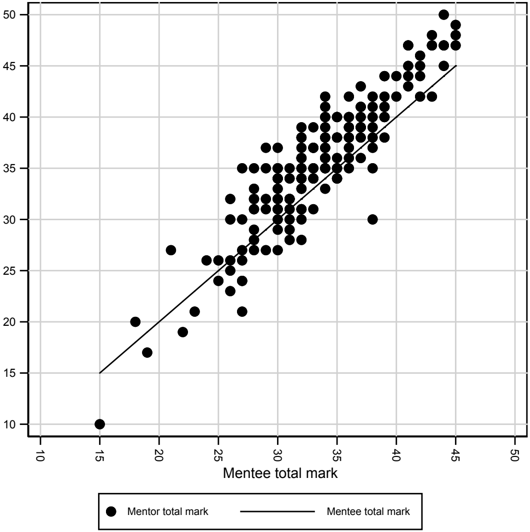
Scatter plot of mentor mark against mentee mark

- And here is a scatter plot of mentor mark against mentee mark, with a diagonal **equality line**.
- It *appears* that the mentor and mentee are *usually* concordant, and that the mentor *usually* awards the higher mark.
- *However...*



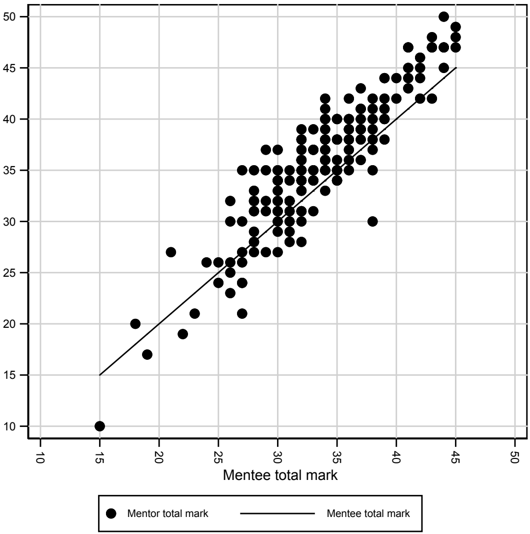
Scatter plot of mentor mark against mentee mark

- And here is a scatter plot of mentor mark against mentee mark, with a diagonal **equality line**.
- It *appears* that the mentor and mentee are *usually* concordant, and that the mentor *usually* awards the higher mark.
- *However…*



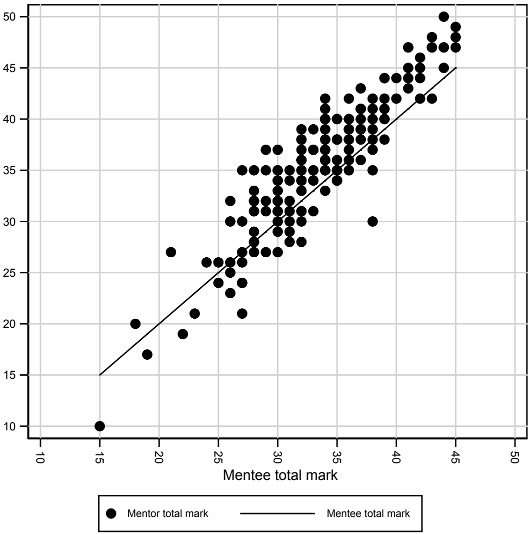
Scatter plot of mentor mark against mentee mark

- And here is a scatter plot of mentor mark against mentee mark, with a diagonal **equality line**.
- It *appears* that the mentor and mentee are *usually* concordant, and that the mentor *usually* awards the higher mark.
- *However…*



Scatter plot of mentor mark against mentee mark

- And here is a scatter plot of mentor mark against mentee mark, with a diagonal **equality line**.
- It *appears* that the mentor and mentee are *usually* concordant, and that the mentor *usually* awards the higher mark.
- *However*. . .



The Bland–Altman plot

- ...there is a more informative way of plotting these data, called the **Bland–Altman plot**[1].
- This is produced by rotating the scatterplot 45 degrees clockwise to produce a plot of the difference between measures (on the vertical axis) against the mean of the 2 measures (on the horizontal axis).
- This has the advantage of being space–efficient, as there is no empty dead space in the top left and bottom right corners of the graph.
- It is also more informative, as it visualises **bias** (represented by the difference) and **scale differential** (represented by mean–difference correlation).

The Bland–Altman plot

- …there is a more informative way of plotting these data, called the **Bland–Altman plot**[1].
- This is produced by rotating the scatterplot 45 degrees clockwise to produce a plot of the difference between measures (on the vertical axis) against the mean of the 2 measures (on the horizontal axis).
- This has the advantage of being space–efficient, as there is no empty dead space in the top left and bottom right corners of the graph.
- It is also more informative, as it visualises **bias** (represented by the difference) and **scale differential** (represented by mean–difference correlation).

The Bland–Altman plot

- ...there is a more informative way of plotting these data, called the **Bland–Altman plot**[1].
- This is produced by rotating the scatterplot 45 degrees clockwise to produce a plot of the difference between measures (on the vertical axis) against the mean of the 2 measures (on the horizontal axis).
- This has the advantage of being space–efficient, as there is no empty dead space in the top left and bottom right corners of the graph.
- It is also more informative, as it visualises **bias** (represented by the difference) and **scale differential** (represented by mean–difference correlation).

The Bland–Altman plot

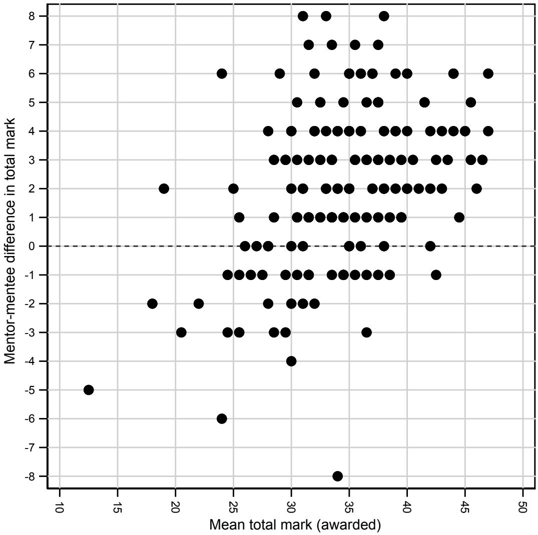
- ...there is a more informative way of plotting these data, called the **Bland–Altman plot**[1].
- This is produced by rotating the scatterplot 45 degrees clockwise to produce a plot of the difference between measures (on the vertical axis) against the mean of the 2 measures (on the horizontal axis).
- This has the advantage of being space–efficient, as there is no empty dead space in the top left and bottom right corners of the graph.
- It is also more informative, as it visualises **bias** (represented by the difference) and **scale differential** (represented by mean–difference correlation).

The Bland–Altman plot

- …there is a more informative way of plotting these data, called the **Bland–Altman plot**[1].
- This is produced by rotating the scatterplot 45 degrees clockwise to produce a plot of the difference between measures (on the vertical axis) against the mean of the 2 measures (on the horizontal axis).
- This has the advantage of being space–efficient, as there is no empty dead space in the top left and bottom right corners of the graph.
- It is also more informative, as it visualises **bias** (represented by the difference) and **scale differential** (represented by mean–difference correlation).

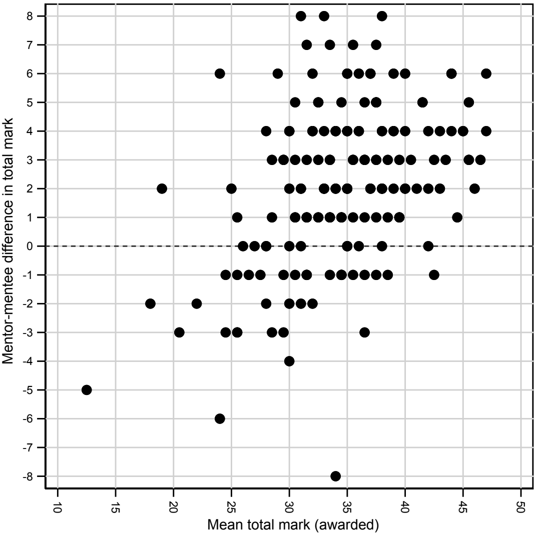
Bland–Altman plot of mentor–mentee difference against mean mark

- In this plot, the diagonal equality line has been rotated 45 degrees to a horizontal Y–axis **reference line** at zero.
- As most points *seem* to be above the reference line, the mentor *seems* to be "Mr Nice".
- And there is a *hint* of an upwards trend in difference with rising mean, *suggesting* that the mentor's mark varies on a larger scale than the mentee's mark.



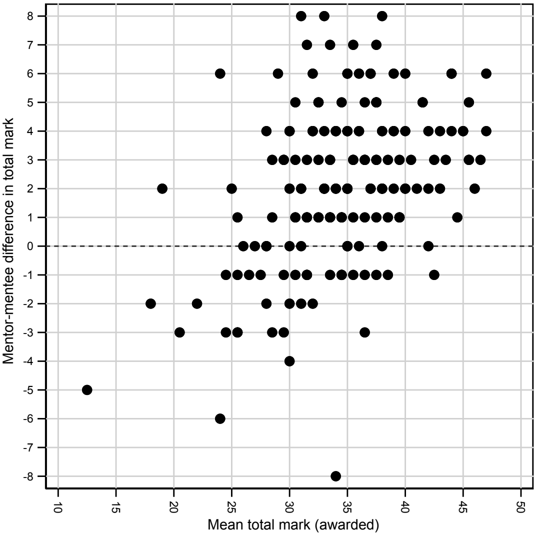
Bland–Altman plot of mentor–mentee difference against mean mark

- In this plot, the diagonal equality line has been rotated 45 degrees to a horizontal Y–axis **reference line** at zero.
- As most points *seem* to be above the reference line, the mentor *seems* to be "Mr Nice".
- And there is a *hint* of an upwards trend in difference with rising mean, *suggesting* that the mentor's mark varies on a larger scale than the mentee's mark.



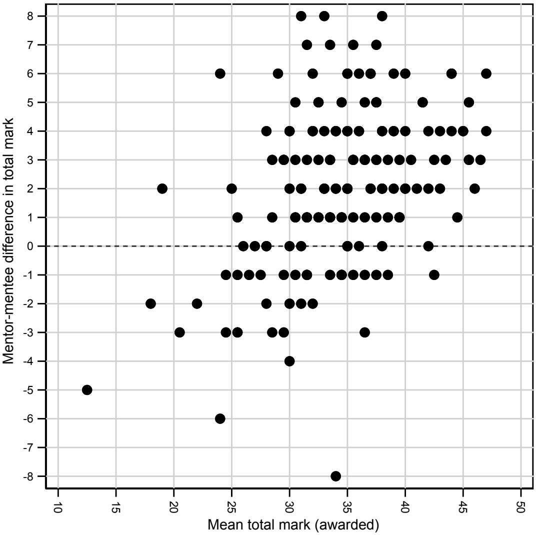
Bland–Altman plot of mentor–mentee difference against mean mark

- In this plot, the diagonal equality line has been rotated 45 degrees to a horizontal Y–axis **reference line** at zero.
- As most points *seem* to be above the reference line, the mentor *seems* to be "Mr Nice".
- And there is a *hint* of an upwards trend in difference with rising mean, *suggesting* that the mentor's mark varies on a larger scale than the mentee's mark.



Bland–Altman plot of mentor–mentee difference against mean mark

- In this plot, the diagonal equality line has been rotated 45 degrees to a horizontal Y–axis **reference line** at zero.
- As most points *seem* to be above the reference line, the mentor *seems* to be "Mr Nice".
- And there is a *hint* of an upwards trend in difference with rising mean, *suggesting* that the mentor's mark varies on a larger scale than the mentee's mark.



But where are the parameters?

- A Bland–Altman plot is a stroke of genius as a visualisation tool, but we would really like to see **parameters** (with confidence limits and *P*–values) to *quantify* the disagreement.
- Van Belle (2008)[6] proposed measuring 3 **principal components** of disagreement, reparameterizing the bivariate Normal model to measure discordance, bias and scale differential.
- I would agree with Van Belle about the 3 principal components, but would prefer to measure them using **rank parameters**, which are less prone to being over–influenced by outliers.
- SSC packages for estimating rank parameters include `somersd`[4][5], `scsomersd`, and `rcentile`[3].

But where are the parameters?

- A Bland–Altman plot is a stroke of genius as a visualisation tool, but we would really like to see **parameters** (with confidence limits and *P*–values) to *quantify* the disagreement.
- Van Belle (2008)[6] proposed measuring 3 **principal components** of disagreement, reparameterizing the bivariate Normal model to measure discordance, bias and scale differential.
- I would agree with Van Belle about the 3 principal components, but would prefer to measure them using **rank parameters**, which are less prone to being over–influenced by outliers.
- SSC packages for estimating rank parameters include `somersd`[4][5], `scsomersd`, and `rcentile`[3].

But where are the parameters?

- A Bland–Altman plot is a stroke of genius as a visualisation tool, but we would really like to see **parameters** (with confidence limits and *P*–values) to *quantify* the disagreement.
- Van Belle (2008)[6] proposed measuring 3 **principal components** of disagreement, reparameterizing the bivariate Normal model to measure discordance, bias and scale differential.
- I would agree with Van Belle about the 3 principal components, but would prefer to measure them using **rank parameters**, which are less prone to being over–influenced by outliers.
- SSC packages for estimating rank parameters include `somersd`[4][5], `scsomersd`, and `rcentile`[3].

But where are the parameters?

- A Bland–Altman plot is a stroke of genius as a visualisation tool, but we would really like to see **parameters** (with confidence limits and *P*–values) to *quantify* the disagreement.
- Van Belle (2008)[6] proposed measuring 3 **principal components** of disagreement, reparameterizing the bivariate Normal model to measure discordance, bias and scale differential.
- I would agree with Van Belle about the 3 principal components, but would prefer to measure them using **rank parameters**, which are less prone to being over–influenced by outliers.
- SSC packages for estimating rank parameters include `somersd`[4][5], `scsomersd`, and `rcentile`[3].

But where are the parameters?

- A Bland–Altman plot is a stroke of genius as a visualisation tool, but we would really like to see **parameters** (with confidence limits and *P*–values) to *quantify* the disagreement.
- Van Belle (2008)[6] proposed measuring 3 **principal components** of disagreement, reparameterizing the bivariate Normal model to measure discordance, bias and scale differential.
- I would agree with Van Belle about the 3 principal components, but would prefer to measure them using **rank parameters**, which are less prone to being over–influenced by outliers.
- SSC packages for estimating rank parameters include `somersd`[4][5], `scsomersd`, and `rcentile`[3].

Measuring discordance: Kendall's τ_a between A and B

- Given pairs of bivariate data points (A_i, B_i) and (A_j, B_j), Kendall's τ_a is defined as

$$\tau_a(A, B) = E[\mathrm{sign}(A_i - A_j)\mathrm{sign}(B_i - B_j)],$$

or (alternatively) as the difference between the probabilities of **concordance** and **discordance** between the A–values and the B–values.

- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and Kendall's τ_a is the difference between the probabilities of agreement and disagreement between the mentor and the mentee, when asked which of 2 random exam scripts is better.

Measuring discordance: Kendall's τ_a between A and B

- Given pairs of bivariate data points (A_i, B_i) and (A_j, B_j), Kendall's τ_a is defined as

$$\tau_a(A, B) = E[\text{sign}(A_i - A_j)\text{sign}(B_i - B_j)],$$

  or (alternatively) as the difference between the probabilities of **concordance** and **discordance** between the A–values and the B–values.

- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and Kendall's τ_a is the difference between the probabilities of agreement and disagreement between the mentor and the mentee, when asked which of 2 random exam scripts is better.

Measuring discordance: Kendall's τ_a between A and B

- Given pairs of bivariate data points (A_i, B_i) and (A_j, B_j), Kendall's τ_a is defined as

$$\tau_a(A,B) = E[\mathrm{sign}(A_i - A_j)\mathrm{sign}(B_i - B_j)],$$

  or (alternatively) as the difference between the probabilities of **concordance** and **discordance** between the A–values and the B–values.
- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and Kendall's τ_a is the difference between the probabilities of agreement and disagreement between the mentor and the mentee, when asked which of 2 random exam scripts is better.

Kendall's τ_a between mentor and mentee marks

We use the `somersd` command, with a `taua` option to specify Kendall's τ_a and a `transf(z)` option to specify the z–transform:

```
. somersd atotmark btotmark, taua transf(z) tdist;
Kendall's tau-a with variable: atotmark
Transformation: Fisher's z
Valid observations: 176
Degrees of freedom: 175

Symmetric 95% CI for transformed Kendall's tau-a
------------------------------------------------------------------------------
             |              Jackknife
    atotmark |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    atotmark |   1.883532   .0451456    41.72   0.000     1.794432    1.972632
    btotmark |   .8824856   .0548829    16.08   0.000      .774168    .9908032
------------------------------------------------------------------------------

Asymmetric 95% CI for untransformed Kendall's tau-a
                 Tau_a      Minimum      Maximum
    atotmark  .95480519   .94622635    .9620421
    btotmark  .70766234   .64934653   .75770458
```

The first confidence interval is for the τ_a of mentor mark with itself (the probability of non–tied mentor marks). The second confidence interval is for the mentor–mentee τ_a, indicating that the mentor and mentee are 65 to 76 percent more likely to agree than to disagree, given 2 random exam scripts and asked which is best.

Measuring bias: The mean sign of $A - B$

- Given bivariate data points (A_i, B_i), the mean sign $E[\text{sign}(A_i - B_i)]$ is the difference between the probabilities $\Pr(A_i > B_i)$ and $\Pr(A_i < B_i)$.
- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and the mean sign is the difference between the probability that the mentor is more generous than the mentee and the probability that the mentee is more generous than the mentor, given one random exam script to mark.

Measuring bias: The mean sign of $A - B$

- Given bivariate data points (A_i, B_i), the mean sign $E[\mathrm{sign}(A_i - B_i)]$ is the difference between the probabilities $\Pr(A_i > B_i)$ and $\Pr(A_i < B_i)$.
- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and the mean sign is the difference between the probability that the mentor is more generous than the mentee and the probability that the mentee is more generous than the mentor, given one random exam script to mark.

Measuring bias: The mean sign of $A - B$

- Given bivariate data points (A_i, B_i), the mean sign $E[\text{sign}(A_i - B_i)]$ is the difference between the probabilities $\Pr(A_i > B_i)$ and $\Pr(A_i < B_i)$.
- *So*, in our example, the A–values are mentor marks, the B–values are mentee marks, and the mean sign is the difference between the probability that the mentor is more generous than the mentee and the probability that the mentee is more generous than the mentor, given one random exam script to mark.

The mean sign of the mentor–mentee difference

We use the `scsomersd` command, with a `transf(z)` option again:

```
. scsomersd dtotmark 0, transf(z) tdist;
Von Mises Somers' D with variable: _scen0
Transformation: Fisher's z
Valid observations: 352
Number of clusters: 176
Degrees of freedom: 175

Symmetric 95% CI for transformed Somers' D
                                 (Std. Err. adjusted for 176 clusters in _obs)
------------------------------------------------------------------------------
             |               Jackknife
      _scen0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _yvar |   .5958514   .0850423     7.01   0.000     .4280109    .7636918
------------------------------------------------------------------------------

Asymmetric 95% CI for untransformed Somers' D
               Somers_D     Minimum     Maximum
       _yvar   .53409091   .40365763   .64324638
```

The bottom confidence interval is for the untransformed mean sign of the difference between mentor and mentee marks. The mentor is 40 to 64 percent more likely than the mentee to be “Mr Nice”, when given one random script from the total population.

Measuring scale differential: The Kendall τ_a between $A+B$ and $A-B$

- Given bivariate data points (A_i, B_i) and (A_j, B_j), the Kendall's τ_a between the sum and the difference (or, equivalently, between the mean and the difference) is $\tau_a(A+B, A-B)$.
- This can be shown (Newson, 2018)[2] to be equal to another difference between probabilities, namely $\Pr(|A_i - A_j| > |B_i - B_j|)$ and $\Pr(|A_i - A_j| < |B_i - B_j|)$.
- *So*, in our example, $\tau_a(A+B, A-B)$ is the difference between the probability that the mentor is more discriminating and the probability that the mentee is more discriminating, when both are asked to mark 2 random scripts and give the difference between the best and the worst.

Measuring scale differential: The Kendall τ_a between $A+B$ and $A-B$

- Given bivariate data points (A_i, B_i) and (A_j, B_j), the Kendall's τ_a between the sum and the difference (or, equivalently, between the mean and the difference) is $\tau_a(A+B, A-B)$.
- This can be shown (Newson, 2018)[2] to be equal to another difference between probabilities, namely $\Pr(|A_i - A_j| > |B_i - B_j|)$ and $\Pr(|A_i - A_j| < |B_i - B_j|)$.
- *So*, in our example, $\tau_a(A+B, A-B)$ is the difference between the probability that the mentor is more discriminating and the probability that the mentee is more discriminating, when both are asked to mark 2 random scripts and give the difference between the best and the worst.

Measuring scale differential: The Kendall τ_a between $A+B$ and $A-B$

- Given bivariate data points (A_i, B_i) and (A_j, B_j), the Kendall's τ_a between the sum and the difference (or, equivalently, between the mean and the difference) is $\tau_a(A+B, A-B)$.
- This can be shown (Newson, 2018)[2] to be equal to another difference between probabilities, namely $\Pr(|A_i - A_j| > |B_i - B_j|)$ and $\Pr(|A_i - A_j| < |B_i - B_j|)$.
- *So*, in our example, $\tau_a(A+B, A-B)$ is the difference between the probability that the mentor is more discriminating and the probability that the mentee is more discriminating, when both are asked to mark 2 random scripts and give the difference between the best and the worst.

Measuring scale differential: The Kendall τ_a between $A+B$ and $A-B$

- Given bivariate data points (A_i, B_i) and (A_j, B_j), the Kendall's τ_a between the sum and the difference (or, equivalently, between the mean and the difference) is $\tau_a(A+B, A-B)$.
- This can be shown (Newson, 2018)[2] to be equal to another difference between probabilities, namely $\Pr(|A_i - A_j| > |B_i - B_j|)$ and $\Pr(|A_i - A_j| < |B_i - B_j|)$.
- *So*, in our example, $\tau_a(A+B, A-B)$ is the difference between the probability that the mentor is more discriminating and the probability that the mentee is more discriminating, when both are asked to mark 2 random scripts and give the difference between the best and the worst.

Kendall's τ_a between mean mark and mentor–mentee difference

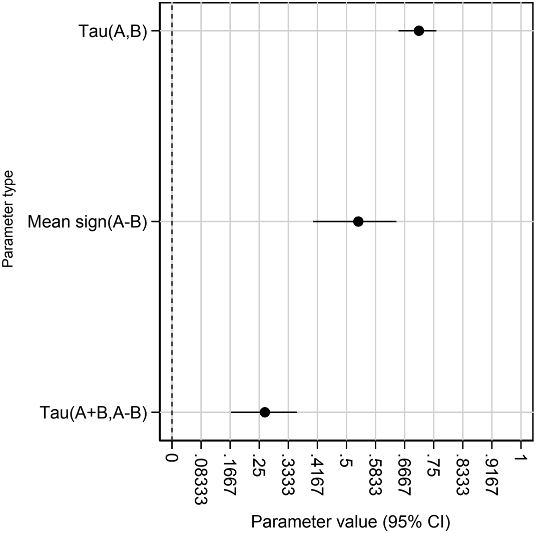
We use the `somersd` command again:

```
. somersd mtotmark dtotmark, taua transf(z) tdist;
Kendall's tau-a with variable: mtotmark
Transformation: Fisher's z
Valid observations: 176
Degrees of freedom: 175

Symmetric 95% CI for transformed Kendall's tau-a
------------------------------------------------------------------------------
             |              Jackknife
    mtotmark |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    mtotmark |   2.210341   .0510751    43.28   0.000     2.109539    2.311144
    dtotmark |   .2728059   .0516663     5.28   0.000     .1708365    .3747752
------------------------------------------------------------------------------

Asymmetric 95% CI for untransformed Kendall's tau-a
                 Tau_a     Minimum     Maximum
    mtotmark  .97623377   .9710022  .98053082
    dtotmark  .26623377  .16919376  .35816145
```

This time, the final confidence interval is for the τ_a between the mean mark and the mentor–mentee difference. The mentor is 17 to 36 percent more likely than the mentee to be the more discriminating of the two.

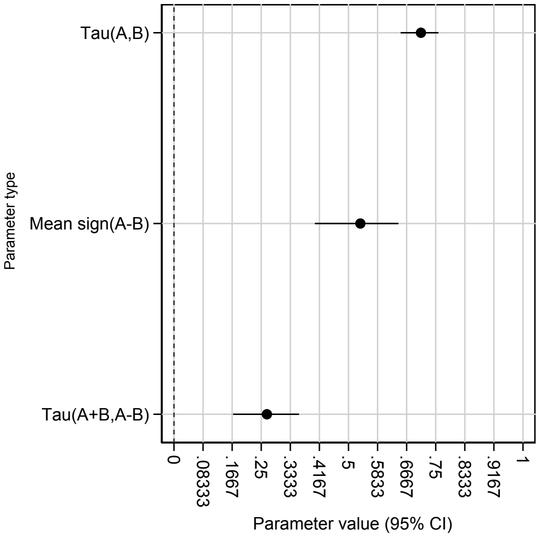
Rank parameters (with confidence limits) for the double–marking data

- The mentor and mentee are 71% more likely to be concordant than to be discordant.
- And the mentor is 53% more likely to be the more generous of the two.
- And the mentor is 27% more likely to be the more discriminating of the two.
- This *may* be because the mentee's brain was dosed with too much coffee!



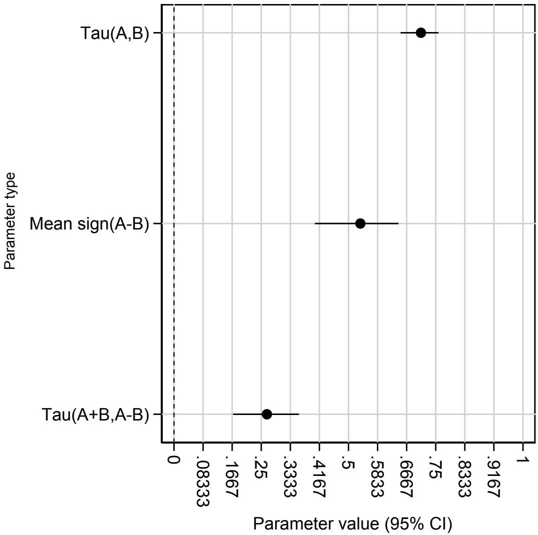
Rank parameters (with confidence limits) for the double–marking data

- The mentor and mentee are 71% more likely to be concordant than to be discordant.
- And the mentor is 53% more likely to be the more generous of the two.
- And the mentor is 27% more likely to be the more discriminating of the two.
- This *may* be because the mentee's brain was dosed with too much coffee!



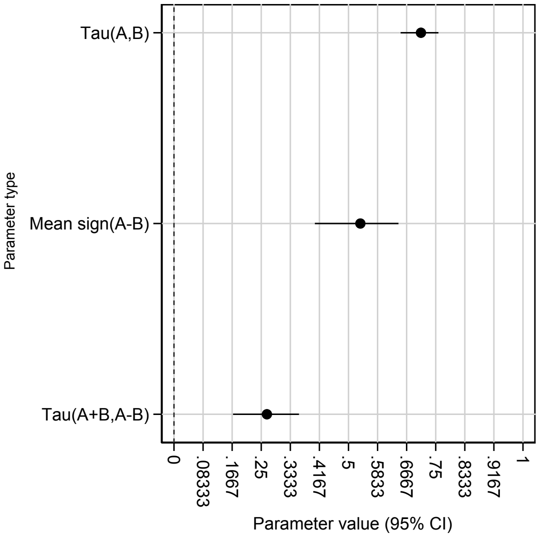
Rank parameters (with confidence limits) for the double–marking data

- The mentor and mentee are 71% more likely to be concordant than to be discordant.
- And the mentor is 53% more likely to be the more generous of the two.
- And the mentor is 27% more likely to be the more discriminating of the two.
- This *may* be because the mentee's brain was dosed with too much coffee!



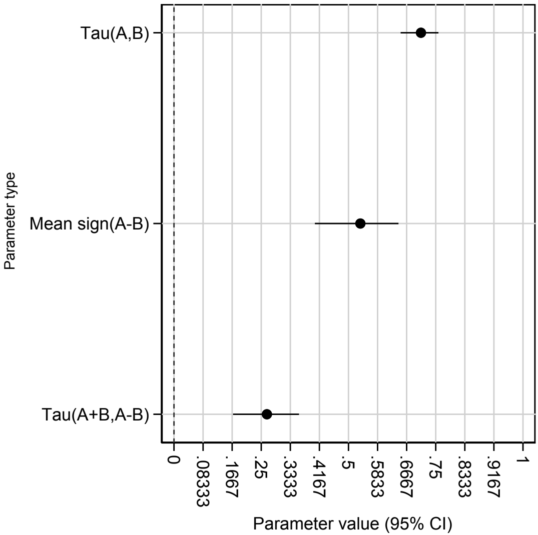
Rank parameters (with confidence limits) for the double–marking data

- The mentor and mentee are 71% more likely to be concordant than to be discordant.
- And the mentor is 53% more likely to be the more generous of the two.
- And the mentor is 27% more likely to be the more discriminating of the two.
- This *may* be because the mentee's brain was dosed with too much coffee!



Rank parameters (with confidence limits) for the double–marking data

- The mentor and mentee are 71% more likely to be concordant than to be discordant.
- And the mentor is 53% more likely to be the more generous of the two.
- And the mentor is 27% more likely to be the more discriminating of the two.
- This *may* be because the mentee's brain was dosed with too much coffee!



Percentile differences

- Re–focussing on bias, we might like to know the *size* distribution for the mentor–mentee differences, as well as their mean direction.
- The SSC package `rcentile`[3] is a "robust" version of `centile`, and saves its confidence intervals in a matrix.

Percentile differences

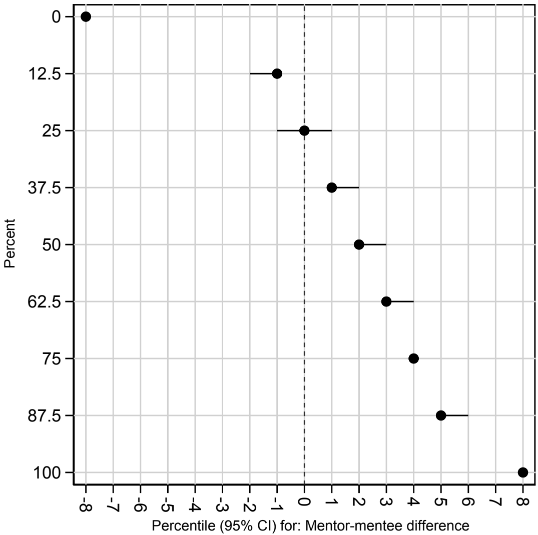
- Re–focussing on bias, we might like to know the *size* distribution for the mentor–mentee differences, as well as their mean direction.
- The SSC package `rcentile`[3] is a "robust" version of `centile`, and saves its confidence intervals in a matrix.

Percentile differences

- Re–focussing on bias, we might like to know the *size* distribution for the mentor–mentee differences, as well as their mean direction.
- The SSC package `rcentile`[3] is a “robust” version of `centile`, and saves its confidence intervals in a matrix.

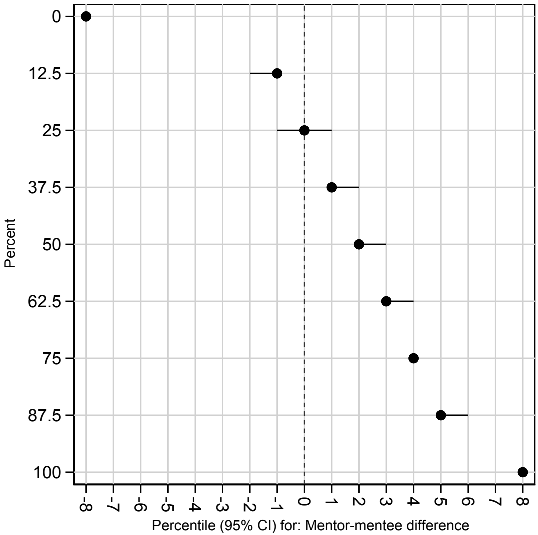
Percentiles of the difference between mentor and mentee marks

- The median difference is 2 marks (out of 50).
- The inter–quartile range is from 0 to 4 marks.
- And the full range is only from -8 to 8 marks.
- Note that these marks are integer-valued!



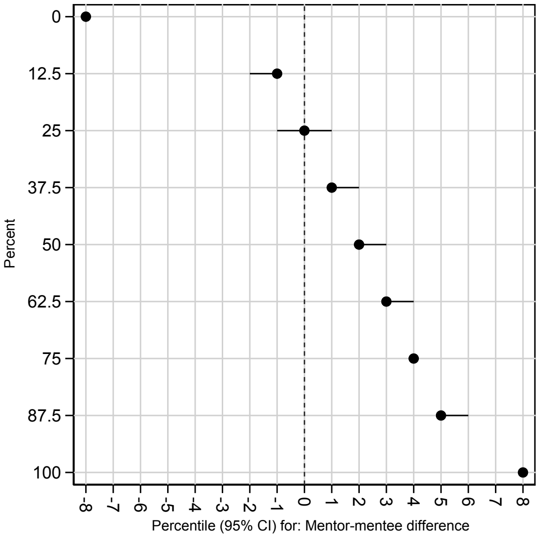
Percentiles of the difference between mentor and mentee marks

- The median difference is 2 marks (out of 50).
- The inter–quartile range is from 0 to 4 marks.
- And the full range is only from -8 to 8 marks.
- Note that these marks are integer-valued!



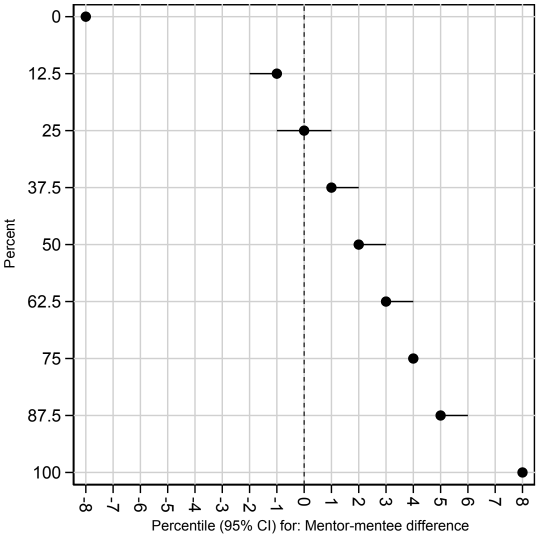
Percentiles of the difference between mentor and mentee marks

- The median difference is 2 marks (out of 50).
- The inter–quartile range is from 0 to 4 marks.
- And the full range is only from -8 to 8 marks.
- Note that these marks are integer-valued!



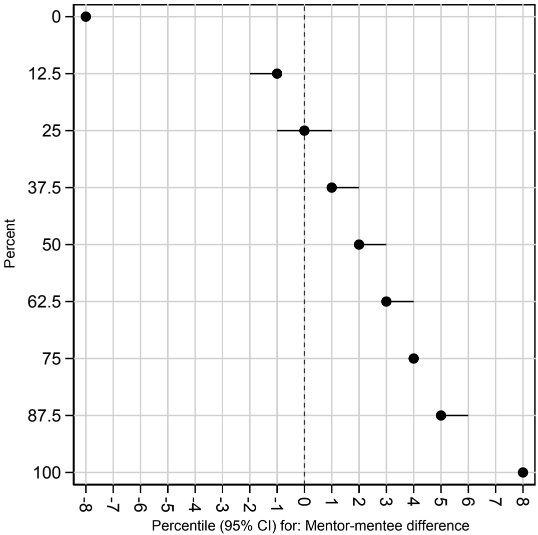
Percentiles of the difference between mentor and mentee marks

- The median difference is 2 marks (out of 50).
- The inter–quartile range is from 0 to 4 marks.
- And the full range is only from -8 to 8 marks.
- Note that these marks are integer-valued!



Percentiles of the difference between mentor and mentee marks

- The median difference is 2 marks (out of 50).
- The inter–quartile range is from 0 to 4 marks.
- And the full range is only from -8 to 8 marks.
- Note that these marks are integer-valued!



Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating "gold–standard" mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating “gold–standard” mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating "gold–standard" mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating "gold–standard" mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating “gold–standard” mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating "gold–standard" mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

Calibration: Estimating the mentor mark from the mentee mark

- We might want to define a **calibration model** to predict one mark from the other.
- *For instance*, the mentee might want to single–mark exam scripts in the future, and to correct his mark to estimate what his more generous and discriminating "gold–standard" mentor would have given.
- He might do this using a linear regression model of mentor mark with respect to mentee mark, with an intercept to correct for bias and a slope to correct for scale differential.
- *However*, it might be better to calibrate non–linearly, correcting for other components of disagreement.
- A common non–linear model is a **decile plot**, with decile of mentee mark on the horizontal axis, and mean mentor mark for that mentee decile on the vertical axis.
- *However*, a possible improvement on both these methods might be a **reference spline**, which might ideally be a **ridit spline**.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \ = \ 100 \times \left[\Pr(X < x) \ + \ \frac{1}{2} \Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) = 100 \times \left[\Pr(X < x) + \frac{1}{2} \Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \; = \; 100 \times \left[\Pr(X < x) \; + \; \frac{1}{2} \Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \;=\; 100 \times \left[\Pr(X < x) \;+\; \frac{1}{2} \Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \;=\; 100 \times \left[\Pr(X < x) \;+\; \frac{1}{2} \Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \;=\; 100 \times \left[\Pr(X < x) \;+\; \frac{1}{2}\Pr(X = x) \right],$$

meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

What are reference splines and ridit splines?

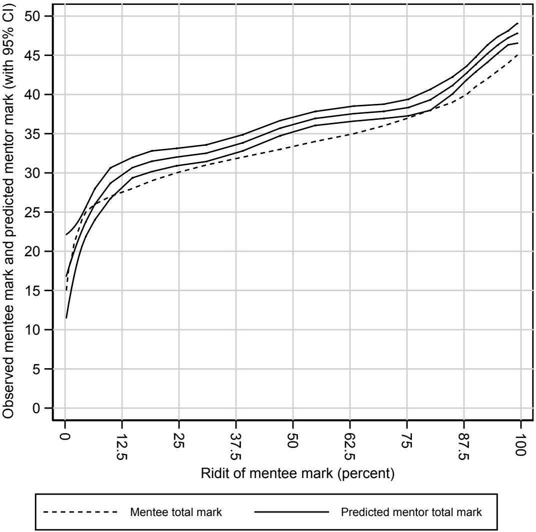
- A **reference spline**[3] is a spline whose parameters are values of the spline at reference points on the X–axis.
- And, given a random variable X, the **percentage ridit function** of X is defined by the formula

$$R_X(x) \;=\; 100 \times \left[\Pr(X < x) \;+\; \frac{1}{2}\Pr(X = x) \right],$$

  meaning that ridits are sample–size–invariant ranks (on a scale from 0 to 100), and percentiles are generalized–inverse ridits.
- *So*, a **ridit spline** in X is a spline in $R_X(X)$.
- In our example do–file, we model (and plot) the mentor marks as a cubic **calibration ridit spline** in the mentee marks.
- This is better than a linear model, as it is non–linear.
- And it is better than a decile plot, as it is continuous.

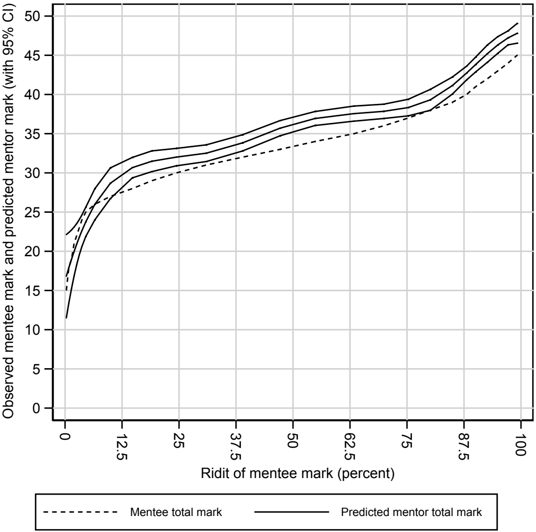
Observed mentee marks and predicted mentor marks

- The horizontal axis gives the percentage ridits, from 0 to 100.
- The dashed line gives the corresponding percentiles of the **observed mentee marks**.
- And the solid line (with solid confidence limits) gives the corresponding **predicted mentor marks**.
- The mentor still appears to be "Mr Nice", but *not* to the lowest–ranking students!



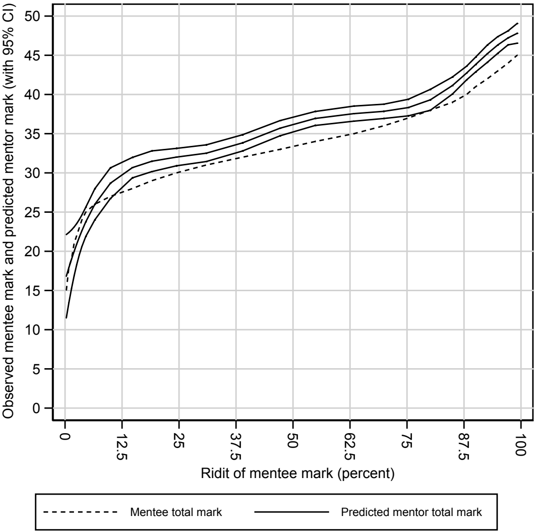
Observed mentee marks and predicted mentor marks

- The horizontal axis gives the percentage ridits, from 0 to 100.
- The dashed line gives the corresponding percentiles of the **observed mentee marks**.
- And the solid line (with solid confidence limits) gives the corresponding **predicted mentor marks**.
- The mentor still appears to be "Mr Nice", but *not* to the lowest–ranking students!



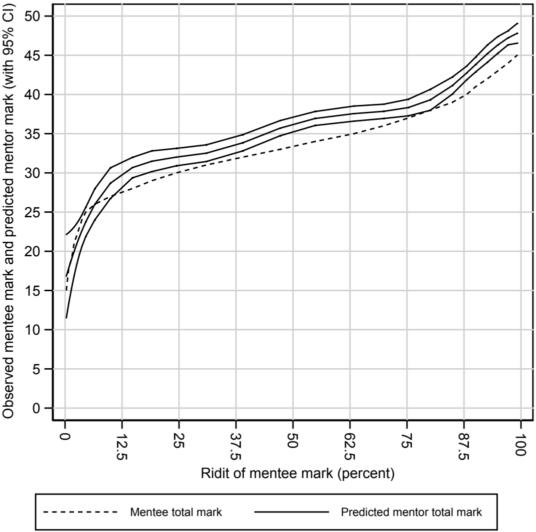
Observed mentee marks and predicted mentor marks

- The horizontal axis gives the percentage ridits, from 0 to 100.
- The dashed line gives the corresponding percentiles of the **observed mentee marks**.
- And the solid line (with solid confidence limits) gives the corresponding **predicted mentor marks**.
- The mentor still appears to be "Mr Nice", but *not* to the lowest–ranking students!



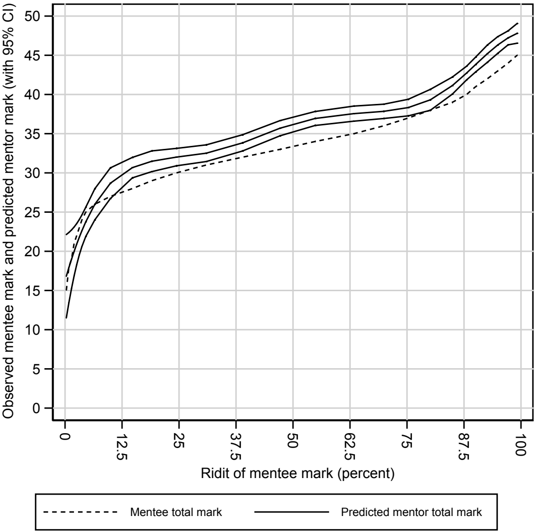
Observed mentee marks and predicted mentor marks

- The horizontal axis gives the percentage ridits, from 0 to 100.
- The dashed line gives the corresponding percentiles of the **observed mentee marks**.
- And the solid line (with solid confidence limits) gives the corresponding **predicted mentor marks**.
- The mentor still appears to be "Mr Nice", but *not* to the lowest–ranking students!



Observed mentee marks and predicted mentor marks

- The horizontal axis gives the percentage ridits, from 0 to 100.
- The dashed line gives the corresponding percentiles of the **observed mentee marks**.
- And the solid line (with solid confidence limits) gives the corresponding **predicted mentor marks**.
- The mentor still appears to be "Mr Nice", but *not* to the lowest–ranking students!



References

[1] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i(8476)**: 307–310.

[2] Newson RB. Rank parameters for Bland–Altman plots. Downloaded on 11 June 2019 from the author's website at *http://www.rogernewsonresources.org.uk/papers.htm#miscellaneous_documents*

[3] Newson RB. Easy-to-use packages for estimating rank and spline parameters. Presented at the *17th UK Stata User Meeting*, 11–12 September, 2014. Downloadable from the conference website at *http://ideas.repec.org/p/boc/usug14/01.html*

[4] Newson R. Confidence intervals for rank statistics: Percentile slopes, differences, and ratios. *The Stata Journal* 2006; **6(4)**: 497–520. Download from *http://www.stata-journal.com/article.html?article=snp15_7*

[5] Newson R. Confidence intervals for rank statistics: Somers' *D* and extensions. *The Stata Journal* 2006; **6(3)**: 309–334. Download from *http://www.stata-journal.com/article.html?article=snp15_6*

[6] van Belle G. *Statistical Rules of Thumb. Second Edition.* Hoboken, NJ: John Wiley & Sons, Inc.; 2008.

The presentation, and the example dataset and do–files, can be downloaded from the conference website, and the packages used can be downloaded from SSC.

And special thanks are due to the late Professor Ken MacRae for mentoring me in marking exam scripts in the 1990s.