

Interrater agreement for continuous measurements - ICC

Introduction

To evaluate different instruments, different ways to measure or the effect of different raters it is often interesting to compare the outputs with agreement measures. Depending on the type of data agreement measures can be correlations, especially interrater correlation coefficients, or kappa measures. Kappa measures are used if the measurements are made on a categorical scale and are not discussed here. Correlations measures are used for continuous measurements. In the following example, two different types of instruments (digital and manual) are used, two different joints are measured in animals (joint A and joint B), four raters make assessments with both methods on both joints and repeat them (in total three measurements for each combination of instrument, joint and animal). Ten animals are part of the experiment and are identified by different ID numbers.

This example is based on a real dataset and data can be found in the Excelfile “`Interrater_agreement.xlsx`”. To work in R with this type of problem we need to have the following packages installed: `tidyverse` for base programming, `readxl` to read Excel files and `ICC` for interrater correlations.

```
library(tidyverse)
library(readxl)
library(ICC)
```

Read the dataset

The dataset is read using the function `read_excel()`. Here the Excel file contains only one sheet and no further statement is needed. If your Excel file contains several sheets specify which one should be read using `sheet=`, e.g. `sheet="Measurement horses"` if the name of the sheet is “Measurement horses”.

```
data_animals<-read_excel("Interrater_agreement.xlsx")
```

Check if the data is read correctly to R, e.g. by

```
str(data_animals)
```

In this file `Instrument`, `Joint` and `Rater` are character (text) variables, while `ID`, `measurement` and `Value` are numeric.

Data handling

`measurement` and `ID` need to be coded as categorical or factor variables for the statistical computations. Therefore these and the three character variables are changed into factors.

```
data_animals$ID<-as.factor(data_animals$ID)
data_animals$Rater<-as.factor(data_animals$Rater)
data_animals$Joint<-as.factor(data_animals$Joint)
data_animals$Instrument<-as.factor(data_animals$Instrument)
data_animals$measurement<-as.factor(data_animals$measurement)
```

Intrarater agreement

First we might be interested to determine if the measurements that are made by the same rater on the same joint on the same animal using the same instrument agree with each other. If intrarater agreement is high we

can say that it is easy for a rater to reproduce his or her own measurement. If intrarater agreement is low the method might be difficult to use or the exact area on the animal could be difficult to determine. To be able to make an assessment we need to filter out one of the joints, one of the instruments and one of the raters. For a more effective way to do this for all joints, instruments and raters at once see further below. We start with quantifying if Therese can reliably use the manual instrument on joint A.

```
data_animals%>%  
  filter(Joint=="Joint A", Instrument=="Manual", Rater=="Theresa")->data_A_manual_intra  
  
ICCest(ID, Value, data=data_A_manual_intra)
```

The output should show that the intrarater agreement for Theresa is 0.732 (\$ICC).

Exercise: Compute the intrarater agreement for Theresa, using the digital method on joint A (should be 0.786) and the intrarater agreement for David using a digital instrument on joint B (should be 0.396).

Interrater agreement

If we instead are interested how well the measurements agree between raters a high correlations would mean that different raters end up with similar measurement values, while low correlations indicate that raters potentially measure different areas or using slightly different approaches. To be able to compute interrater agreement for this data set we need first to handle the multiple measurements on the same objects by a rater. There are two possibilities: either we select one measurement (per rater, joint, instrument and animal) or we use the mean value of the three replicated measurements. We should expect better agreement if mean values are used but it might be more realistic to use single measurement. We give both options here

Based on single measurements

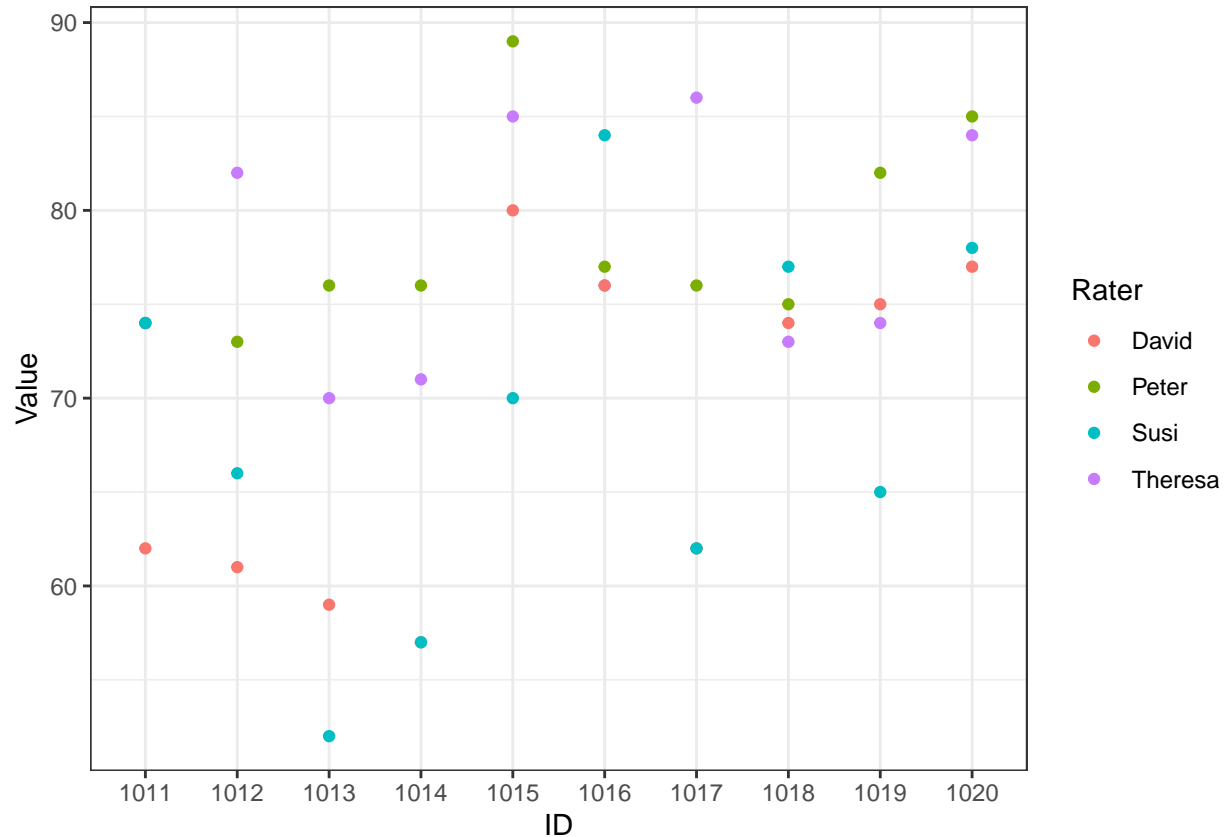
To compute ICC on single measurements per rater we first need to extract one of the measurements, e.g. measurement 1. The interrater correlation is then computed with the function ICC using Rater as first input:

```
data_animals%>%  
  filter(Joint=="Joint A", Instrument=="Manual", measurement==1)->data_A_manual_inter  
  
ICCest(ID, Value, data=data_A_manual_inter)
```

The inter correlation coefficient is 0.23. i.e. not especially high.

This agreement is not very good and we can check the validity for example with a plot, which shows that there is a lot of variation between measurements and especially Susi deviates from the others.

```
data_A_manual_inter%>%  
  ggplot(aes(x=ID, y=Value, col=Rater))+  
  geom_point()+  
  theme_bw()
```



Exercise: compute the inter correlation coefficient for joint B, the manual instrument and measurement number 3 (should be -0.23).

Based on mean measurements

To base the inter-rater correlation on the mean of the three measurements made on each individual we first need to compute this mean. We use again joint A and the manual instrument and compute a rater mean for each animal (ID). The mean is denoted by `meanV` and computed by the function `summarise`.

```
data_animals%>%
  filter(Joint=="Joint A", Instrument=="Manual")%>%
  group_by(ID, Rater)%>%
  summarise(meanV=mean(Value))>data_A_manual_inter_by_mean

ICCest(ID, meanV, data=data_A_manual_inter_by_mean)
```

The interrater correlation is 0.33.

Exercise: Compute the interrater correlation for joint B and the manual instrument using the mean per rater (should be -0.25)

Using `map` to compute inter- and intra correlations for several subgroups.

With the approach above we need to make several subgroups and compute the ICC for each of them separately. Using the function `map` we can easily do the computations for all joints and all instruments (and all raters) at the same time

Intrarater correlations

To use `map` we need to define the groups for which we want to compute the ICC. When the interest is to compare the measurements within a rater (intra) we make groups that are defined by instrument, joint and the rater. The function `nest()` creates subset-dataset for these groups, which then can be analysed separately. The function `map()` or here `map_dbl()` runs the function we define (`ICCest()`) on each of the subsets (which are called `data`) and returns the value of ICC (`$ICC`) from each.

The results are saved in a variable called `ICC`. Subgroups are removed and only ICC is kept for each (`unnest(ICC)`) and the grouping is forgotten (`ungroup()`). In the last row we select the variables that we want to keep (`select()`)

```
data_animals %>%
  group_by(Instrument, Joint, Rater)%>%
  nest()%>%
  mutate(ICC= map_dbl(data, ~ICCest(.$ID, .$Value)$ICC))%>%
  unnest(ICC)%>%
  ungroup() %>%
  select(Instrument, Joint, Rater, ICC)
```

```
## # A tibble: 16 x 4
##   Instrument Joint   Rater    ICC
##   <fct>      <fct> <fct>  <dbl>
## 1 Manual    Joint A Theresa 0.732
## 2 Digital   Joint A Theresa 0.786
## 3 Manual    Joint B Theresa 0.727
## 4 Digital   Joint B Theresa 0.355
## 5 Manual    Joint A Peter   0.774
## 6 Digital   Joint A Peter   0.821
## 7 Manual    Joint B Peter   0.410
## 8 Digital   Joint B Peter   0.300
## 9 Manual    Joint A David   0.625
## 10 Digital   Joint A David   0.804
## 11 Manual    Joint B David   0.777
## 12 Digital   Joint B David   0.396
## 13 Manual    Joint A Susi    0.719
## 14 Digital   Joint A Susi    0.895
## 15 Manual    Joint B Susi    0.614
## 16 Digital   Joint B Susi    0.567
```

We see that the intrarater agreement is quite high for all measurements, except for the digital method on joint B.

Interrater correlations

Similarly, we can do the computations for the interrater correlations, which only contain the extra step of computing the mean value for the three measurements made by the different raters (`summarise()`). In the next step we do no longer group by rater but only using instrument and joint.

```
data_animals %>%
  group_by(Instrument, Joint, ID, Rater)%>%
  summarise(medel=mean(Value))%>%
  ungroup() %>%
  group_by(Instrument, Joint)%>%
  nest() %>%
  mutate(ICC= map_dbl(data, ~ICCest(.$ID, .$medel)$ICC))%>%
  unnest(ICC)%>%
```

```
ungroup()%>%
select(Instrument, Joint, ICC)
```

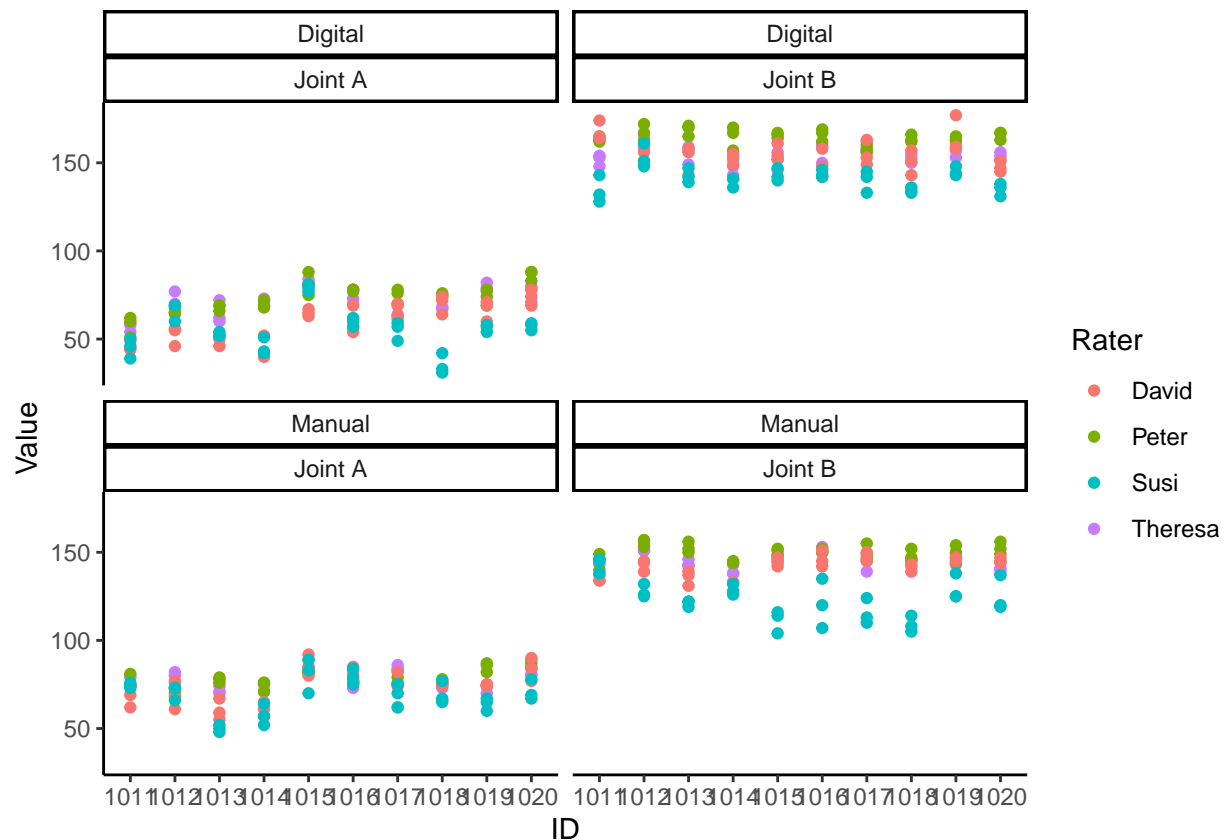
```
## # A tibble: 4 x 3
##   Instrument Joint      ICC
##   <fct>      <fct>    <dbl>
## 1 Digital    Joint A    0.192
## 2 Digital    Joint B   -0.225
## 3 Manual     Joint A    0.331
## 4 Manual     Joint B   -0.254
```

When comparing between raters for any of the methods and joins we get low correlations, which could indicate that the raters measure on different places or with different approaches, which lets them repeat their own measurements (high intrarater agreement) but is not comparable between raters (low intrarater agreement).

Visualisations

To get a plot over how well the concordance of the raters are we could make a scatter plot

```
data_animals%>%
  ggplot(aes(x=ID, y=Value, col=Rater))+
  geom_point()+
  facet_wrap(~Instrument+Joint)+
  theme_classic()
```



Even here we see that Susi often differs and that the variation is largest for manual measurements of joint B,

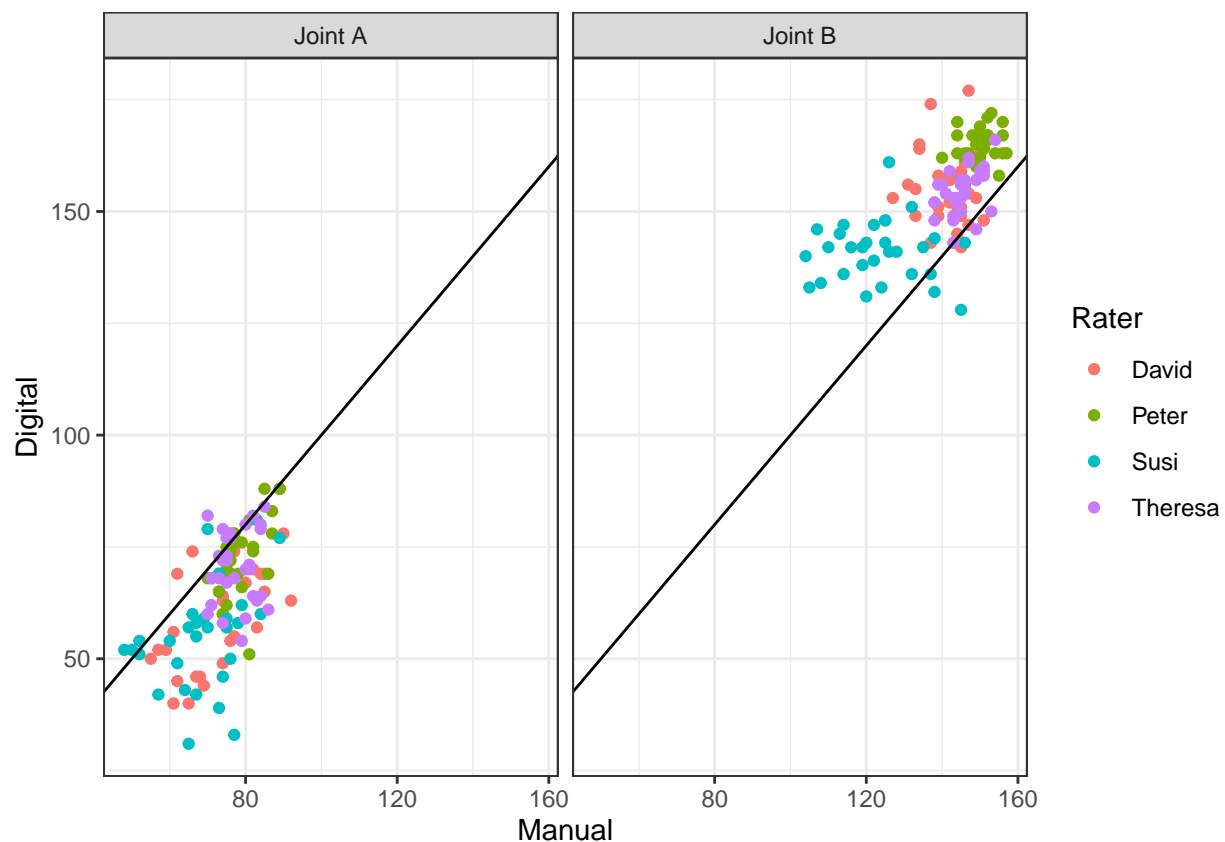
while joint A is easier to determine.

We could also want to compare the different instruments, while keeping the two different joints apart. We start by changing the data structure in our data set to create two different columns: one containing the measurement value for the digital method and one for with measurements for the manual method:

```
data_animals%>%  
  spread(key= Instrument, value=Value)->data_spread
```

The new data frame `data_spread` can now be used to plot the two instrument results against each other. On the x-axis we use the manual instrument (`x=Manual`) and on the y-axis the digital (`y=Digital`). We also use a color to code the raters (`col=Rater`). `geom_point()` draws dots for each measurement and `abline(intercept=0, slope=1)` draws a identity line. If there is a perfect agreement between the methods all dots should be on this line. We make the same type of plot for both joints (`facet_wrap(Joint)`) and use a neutral white background (`theme_bw()`).

```
data_spread%>%  
  ggplot(aes(x=Manual, y=Digital, col=Rater))+  
  geom_point()+  
  geom_abline(intercept = 0, slope = 1)+  
  facet_wrap(~Joint)+  
  theme_bw()
```



We see that the digital instrument underestimates the value compared to the manual instrument on joint A. Meanwhile, for joint B the opposite is true - the digital method generally gives higher values than the manual. We also see that especially for joint B rater Susi gives substantially lower measurements than the other raters and also with more variation.