# Medical Statistics - Seminar 4

*Group 1-2*

## Task 1: Measuring blood glucose

*To compare the laboratory facilities of two hospitals, blood samples from 48 patients were analysed for blood glucose with the equipment in the hospitals' laboratories for clinical chemistry. The glucose concentrations (in mmol/l) are given in glukosdata.txt.*

*What conclusions can be drawn from these numbers? How are they best presented and analysed?*

### 1.1 Data

The given data is continuous (measured in mmol/l) and two hospitals are compared, HospA and HospB (Figure 1.1).

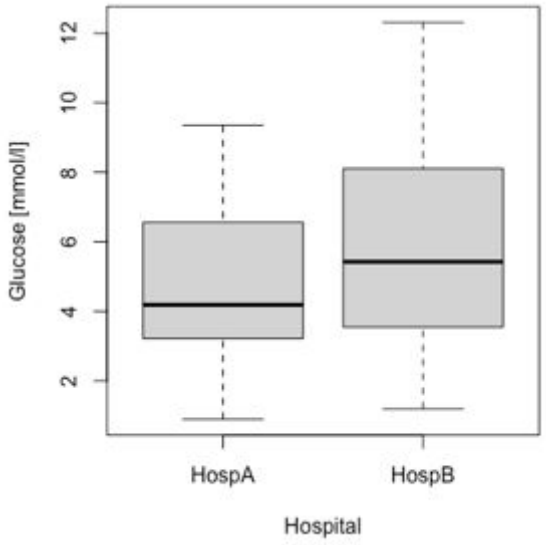


Figure 1.1: Boxplots of the raw data for measurements of glucose in two hospitals

The data can be analysed in different ways:

- Hypothesis testing
- Linear regression model
- Bland-Altman-Plot
- ICC

## 1.2 Hypothesis testing

To investigate if the data sets are different, we use the following hypothesis:

$H_0$: There is no difference between HospA and HospB.

$H_1$: There is a difference between HospA and HospB.

$\alpha = 0.05$

Since we have paired data, the differences between the measurements (Figure 1.2) are examined for normality. The p-value of the shapiro test is p=0.4553, which means that we can assume normality.

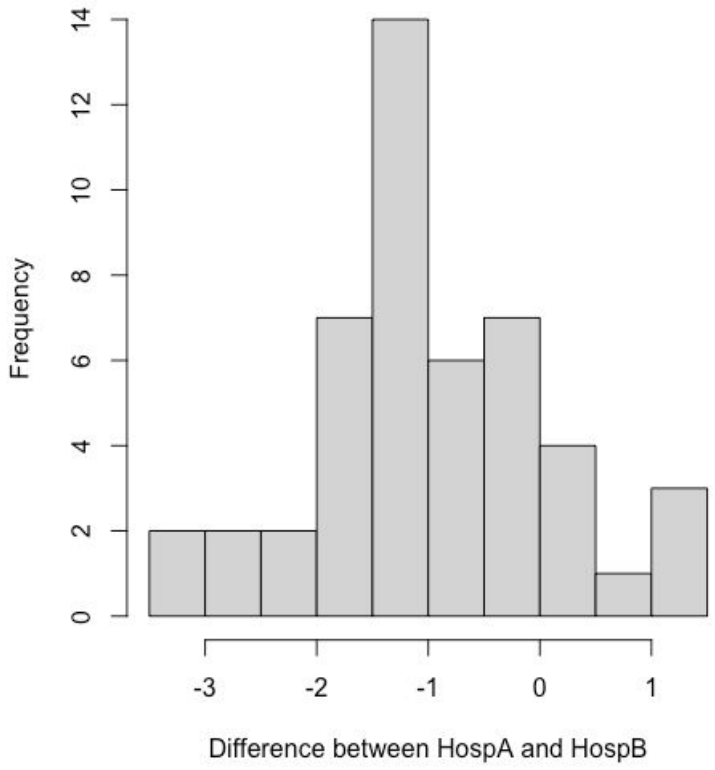


Figure 1.2: Histogram of the difference between the measurements

The hypothesis is tested with a paired t-test (Figure 1.3). The p-value is smaller than our chosen significance level and we can reject the null-hypothesis, that there is no difference between the measurements.

```
        Paired t-test

data:  HospA and HospB
t = -6.3736, df = 47, p-value = 7.304e-08
alternative hypothesis: true difference in means is not
 equal to 0
95 percent confidence interval:
 -1.2835684 -0.6676816
sample estimates:
mean of the differences
              -0.975625
```

Figure 1.3: Results of the t-test

## 1.3 Linear regression model

A linear regression model is used to check if there is a linear correlation between the two measurements (Figure 1.4). The results show that there is a linear relationship (p = 2e-16, R-squared = 0.8955). Figure 1.5 shows the fitted line and an “optimal” line, if the hospitals would have always measured the same. The lines are close for small measurement values, but the difference becomes larger the higher the values get.

```
Call:
lm(formula = HospA ~ HospB)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4125 -0.4828 -0.1314  0.4296  1.6347

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.69738    0.22986   3.034  0.00396 **
HospB        0.71088    0.03581  19.853  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6895 on 46 degrees of freedom
Multiple R-squared:  0.8955,    Adjusted R-squared:  0.8932
F-statistic: 394.1 on 1 and 46 DF,  p-value: < 2.2e-16
```

Figure 1.4: Results of linear regression model

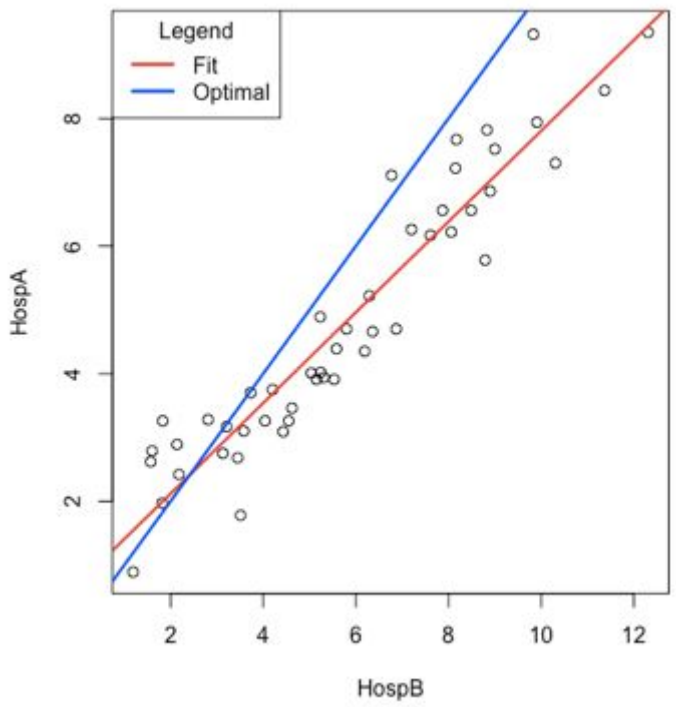


Figure 1.5: Comparison of the linear fit and an “optimal” fit
(both hospitals would have measured the same)

## 1.4 Bland-Altman-Plot

The Bland-Altman-Plot can be used to graphically compare two measurements (Figure 1.6). The mean of the difference (HospA-HospB) is negative, indicating that HospB measures higher values than HospB. Furthermore, a drifted deviation can be observed (Figure 1.7). The higher the measured values (higher average), the higher is the difference between the measurements. This confirms the observations of the linear regression model.

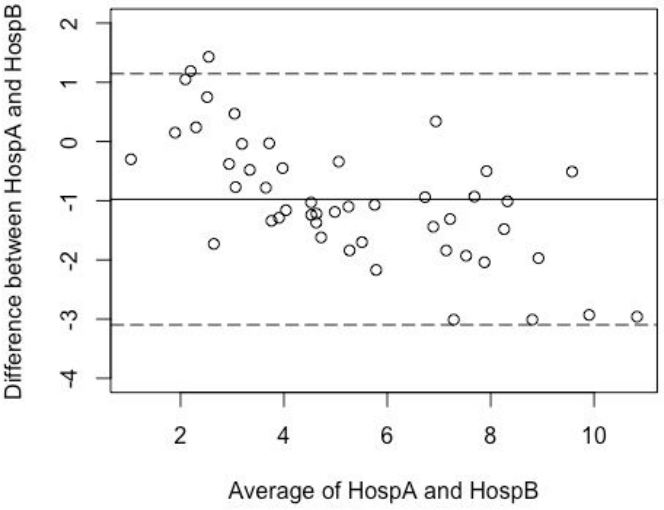


Figure 1.6: Bland-Altman-Plot

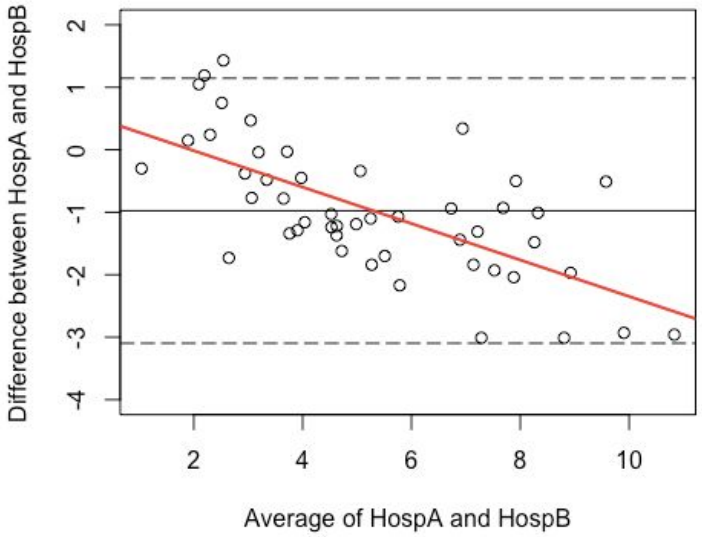


Figure 1.7: Drifted deviation

## 1.4 ICC with in-build function

Next, an ICC analysis is performed to investigate the agreement of the data sets (Figure 1.8). All values are measured by the same “raters”, so we use a two-way model. It can be seen that the ICC is very large (84,5%), but as well is the 95%-Confidence interval which also includes values below 50%.

```
> icc(samples, model="twoway", type=("agreement"), unit="single")
 Single Score Intraclass Correlation

   Model: twoway
   Type : agreement

   Subjects = 48
     Raters = 2
   ICC(A,1) = 0.845

 F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(47,4.96) = 20.9 , p = 0.00152

 95%-Confidence Interval for ICC Population Values:
  0.394 < ICC < 0.942
```

Figure 1.8: ICC analysis

## 1.5 ICC with mixed-effect model

Last, a mixed-effect model was created (Figure 1.9). The further analysis of the model (Figure 1.10) showed similar results than the first ICC analysis.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: glucose ~ 1 + (1 | case) + (1 | Hospital)

REML criterion at convergence: 366.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.76327 -0.45487  0.03413  0.43722  1.69128

Random effects:
 Groups   Name        Variance Std.Dev.
 case     (Intercept) 5.6079   2.3681
 Hospital (Intercept) 0.4642   0.6813
 Residual             0.5624   0.7499
Number of obs: 96, groups:  case, 48; Hospital, 2

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
(Intercept)   5.2986     0.5956 2.2103   8.896  0.00901 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1.9: Mixed-effect model

MODEL INFO:
*Observations:* 96
*Dependent Variable:* glucose
*Type:* Mixed effects linear regression

MODEL FIT:
*AIC* = 374.15, *BIC* = 384.41
*Pseudo-R² (fixed effects)* = 0.00
*Pseudo-R² (total)* = 0.92

FIXED EFFECTS:

| | Est. | S.E. | t val. | d.f. | p |
|---|---|---|---|---|---|
| (Intercept) | 5.30 | 0.60 | 8.90 | 2.21 | 0.01 |

*p values calculated using Satterthwaite d.f.*

RANDOM EFFECTS:

| Group | Parameter | Std. Dev. |
|---|---|---|
| case | (Intercept) | 2.37 |
| Hospital | (Intercept) | 0.68 |
| Residual | | 0.75 |

Grouping variables:

| Group | # groups | ICC |
|---|---|---|
| case | 48 | 0.85 |
| Hospital | 2 | 0.07 |

Figure 1.8: Analysis of the mixed-effect model

## 1.6 Conclusion

The data was analysed considering different statistical approaches. The t-test had a significant result indicating a systematic difference between both measurements. This observation was confirmed with the linear regression model and the Bland-Altman-Plot. Both analyses show that for higher measurements the difference between the hospitals is higher, while HospB measures higher values as HospA.

Eventually, the agreement analysis shows a high ICC (85%) indicating a high agreement between both hospitals. However, the confidence interval is quite large which probably results due to the systematic effect.

# Task 2: Hip osteoarthrosis

*In order to study the reliability of osteoarthrosis diagnosis using MRI, two radiologists independently reviewed MRI exams of the hips in 38 patients and graded of the degree of osteoarthrosis according to the following scale:*

• *grade 0: normal*

• *grade 1: inhomogeneous high signal intensity in cartilage (T2WI)*

• *grade 2: inhomogeneity with areas of high signal intensity in articular cartilage (T2WI); indistinct trabeculae or signal intensity loss in femoral head and neck (T1WI)*

• *grade 3: criteria of grade 1 and 2 plus indistinct zone between femoral head and acetabulum; subchondral signal loss due to bone sclerosis*

• *grade 4: above criteria plus femoral head deformity*

*The gradings are found in the file hipdata.txt.*

*Analyse the data and discuss whether the agreement between observers is good or not.*

## 1.1 The data

The data consists of two different observers for radiological images of 38 patients. The aim is to conclude if the agreement between observers are good. Figure 2.1 shows the histogram of the classifications and the frequency for both of the observers.

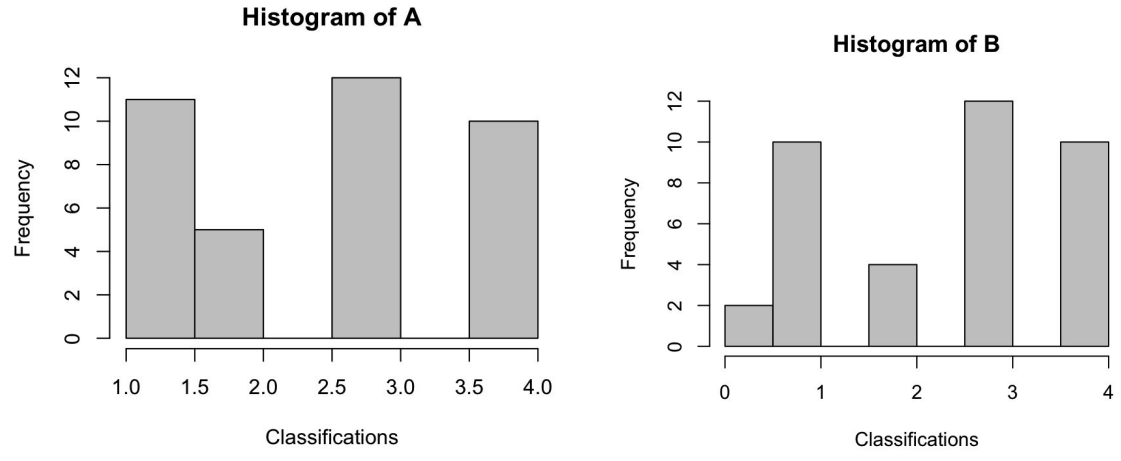


Figure 2.1. Histogram of classifications of observer A(left) and observer B(right)

In order to determine the agreement we can use:

- Bland Altman plot
- ICC score
- Random effects model

Because this is quite simple data, the bland altman plot and ICC score will be enough, rather than creating a random effect model.

## 2.1 Bland Altman Plot

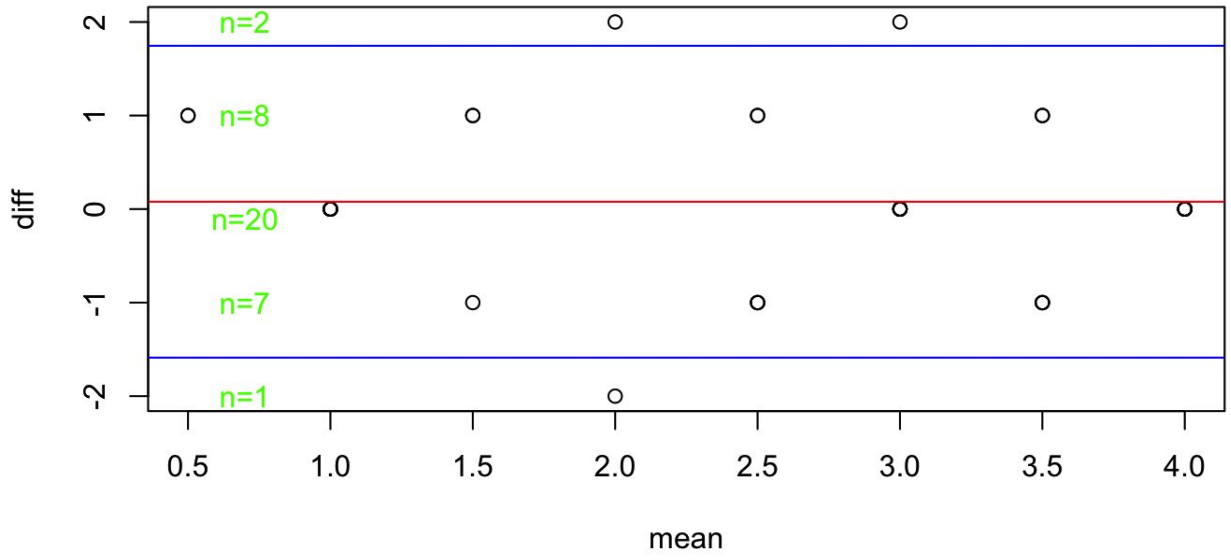


Figure 2.2. Bland Altman plot of the observers A and B

35 observations between A and B are within the limits of agreements. 20 of them are closest to the error line, 15 are further away from the error line, and 3 are outside of the limits.

## 2.2 Intraclass Correlation Coefficient (ICC)

Next we perform an ICC test of the rawdata to quantify the agreements(figure2.3) and we can see an ICC score of 76,6% with a confidence interval below 50%, which indicated a good agreement.

```
Single Score Intraclass Correlation

  Model: oneway
  Type : consistency

  Subjects = 38
    Raters = 2
    ICC(1) = 0.767

F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(37,38) = 7.57 , p = 4.45e-09

95%-Confidence Interval for ICC Population Values:
 0.597 < ICC < 0.871
```

Figure 2.3 ICC test results of raw data.

## 2.3 Conclusion

There is a quite high realibility between the two radiologists which can observed both in the Bland Altman plot as well as in the results of ICC test.

# Task3: Lung function test equipment

*A producer of equipment for lung function measurements has introduced a new model of Peak Expiratory Flow (PEF) meter, PEF Super, to replace the older PEF Standard. One of the aims was to increase the reproducibility of the measurements.*

*In 20 patients, the new equipment was used for repeated measurements on two consecutive days. The results were then compared to similar data obtained previously in 20 other patients with the older equipment. The measurements (in l/min) are found in PEFdata.txt.*

*Select a suitable method for analysis and use the results of that for answering the question: Has reproducibility been improved by the introduction of the new model or not?*

## 3.1 Data

- Basic analysis with PEF data

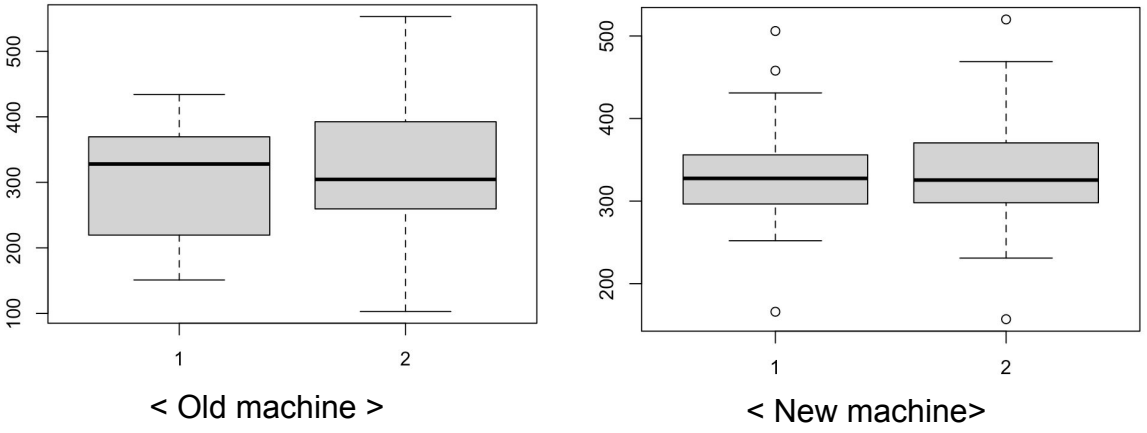


Figure 1.1 Boxplots of each machine by days

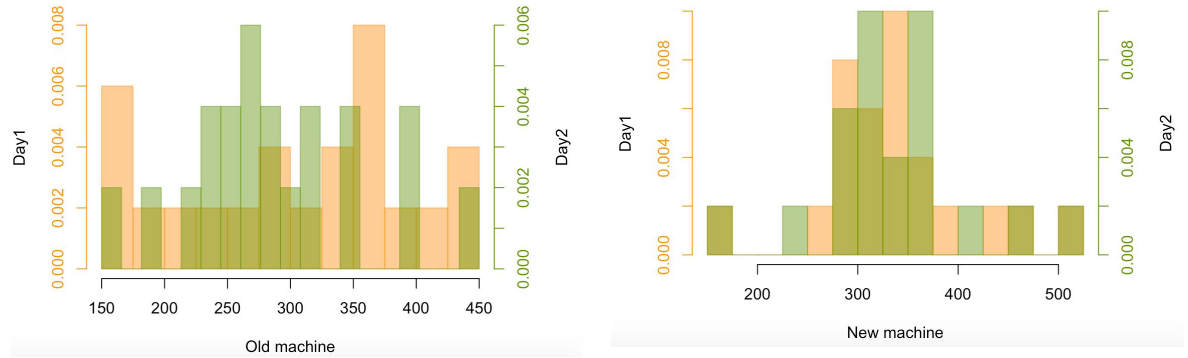


Figure 1.2 Density chart of each machine

Figure 1.1 and 1.2 shows schematically which machine produces the same output for two consecutive days. The mean values in boxplot and density chart results suggest that the new machine is producing similar results over two days.

- Select suitable method for analysis

## 3.2 Intraclass Correlation Coefficient and Mean Absolute Error

ICC test was performed with two-way and agreement parameters. The results of PEF measurement are continuous and we want to confirm if each machine on different days assigns the same score to the same subject. And there are two random effects that randomly chosen 2 consecutive days and 20 subjects.

Additionally, MAE(Mean Absolute Error) is also measured because it is a good way to compare the consistency of responses in time series.

- Compare results

```
Single Score Intraclass Correlation

  Model: twoway
  Type : agreement

  Subjects = 20
    Raters = 2
  ICC(A,1) = 0.705

F-Test, H0: r0 = 0 ; H1: r0 > 0
  F(19,20) = 5.81 , p = 0.000129

95%-Confidence Interval for ICC Population Values:
 0.4 < ICC < 0.871
```

Figure 1.3 ICC test results of old machine

```
Single Score Intraclass Correlation

  Model: twoway
  Type : agreement

  Subjects = 20
    Raters = 2
  ICC(A,1) = 0.964

F-Test, H0: r0 = 0 ; H1: r0 > 0
F(19,19.4) = 52.5 , p = 9.35e-13

95%-Confidence Interval for ICC Population Values:
 0.912 < ICC < 0.986
```

Figure 1.4 ICC test results of new machine

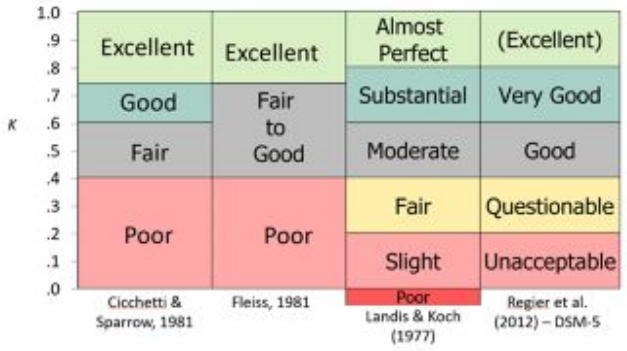


Figure 1.5 ICC grade (Source: http://www.wikiwand.com/en/Inter-rater_reliability)

**Tabe 1.1** MAE (Mean Absolute Error) values

| | Old Machine | New Machine |
|---|---|---|
| MAE | 70.55 | 18.22 |

## 3.3 Conclusion

The results of the ICC test shows that both machines have reproducibility for days because the p-value is smaller than 5%, but the p-value of the new machine is significantly smaller than the old machine's result. Moreover, the ICC score of the old machine is 0.705 and the new machine is 0.986. According to Figure 1.5, the grade of the old machine is around good or substantial level but the new machine is in excellent grade. And the confidence level of each result also indicates that the ICC score of the new machine is in excellent range.

MAE value is also a good indicator to compare reproducibility. Since the MAE value of the new machine is smaller than the old one, it is possible to analyze that the new machine has more consistency on time series.

Therefore, reproducibility has been improved by the introduction of the new model.

# Task 4: Diagnostic certainty of tumour diagnosis

*For a study of ultrasonic diagnosis of kidney disease, a physician performed a pilot study where she judged the probability of a tumour being present in the kidneys in 40 patients according to the following scale:*

*1 – tumour certainly not present*

*2 – tumour probably not present*

*3 – inconclusive*

*4 – tumour probably present*

*5 – tumour certainly present*

*This assessment was performed both before and after injection of contrast medium.*

*A relevant research question is whether the addition of contrast medium reduced the diagnostic uncertainty or not, i.e. if the examiner could give a clearer answer to the diagnostic question or not when contrast was given.*

*A) Try to answer the research question using the results in kidneydata.txt. Consider both possible ways to modify the data and techniques for the statistical analysis.*

*B) After the pilot study, a larger study of the same research question in 100 patients is planned, where two different contrast media are to be compared, so that each patient is examined before contrast medium injection and then again after injection of one (randomly chosen) contrast medium. Each examination is to be recorded and reviewed by three reviewers, using the same scale as above, separately for the left and right kidney. Propose a statistical method for analysing the data from this larger study.*

## 4.1 Task A

In this task we reordered the data to describe confidence in diagnosis. As one can observe in the task description, rank 1 and 5 correspond to equal levels of confidence in diagnosis. Similarly, rank 2 and 4 do the same.

So what we did was to set all the diagnoses that were rank 5, to rank 1. In this way, the total number of diagnoses with rank 1 describes all the diagnoses of which the physician was confident. After that we set all the diagnoses that were rank 4 to rank 2 which meant that the total number of diagnoses with rank 2, describes all the diagnoses of which the physician was somewhat uncertain. And finally, rank 3 remained the same which is an inconclusive diagnosis. The ordinal scale now goes from rank 1 which is certain to rank 3 which is inconclusive. This was done for both before and after contrast injection and the results were displayed in the histograms which can be seen in figure 1.6 and figure 1.7 below.

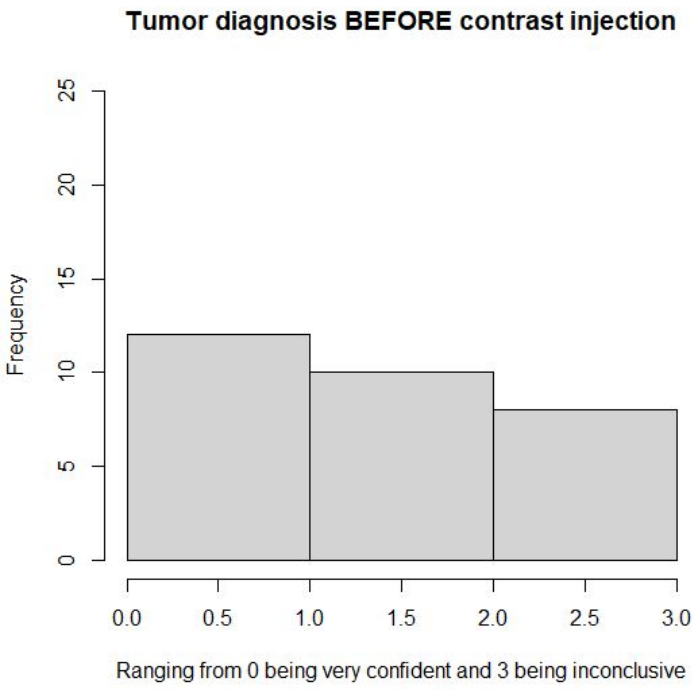


Figure 1.6 = Histogram of confidence in diagnosis before contrast injection

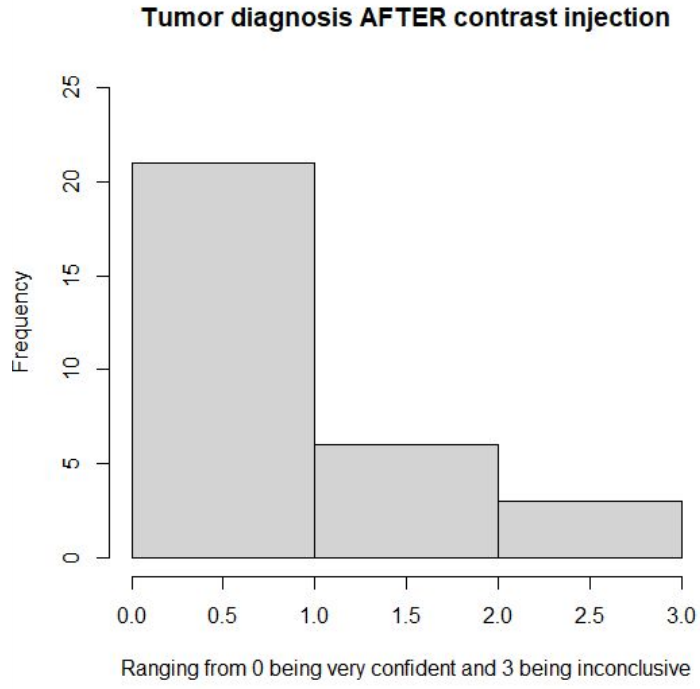


Figure 1.7 = Histogram of confidence in diagnosis after contrast injection

From observing these histograms, we could infer that the number of confident diagnoses had increased after contrast injection. The next step was to do a statistical test on this data in order to confirm, as it was a part of the task.

The test we chose was a Wilcoxon signed-rank test as we had paired and non-parametric data. The altered data table can be seen in figure 1.8 and the results from the test can be seen in figure 1.9. As we can observe, the result was significant with a p-value of 0.00893, which means that there is a difference in confidence of diagnosis between before and after contrast injection.

| | X | before | after |
|---|---|---|---|
| 1 | 1 | 3 | 1 |
| 2 | 2 | 3 | 1 |
| 3 | 3 | 1 | 1 |
| 4 | 4 | 2 | 1 |
| 5 | 5 | 2 | 1 |
| 6 | 6 | 1 | 1 |
| 7 | 7 | 2 | 2 |
| 8 | 8 | 2 | 3 |
| 9 | 9 | 1 | 1 |
| 10 | 10 | 3 | 2 |
| 11 | 11 | 1 | 1 |
| 12 | 12 | 1 | 1 |

Figure 1.8 = Snippet of the data table which has been altered

```
> wilcox.test(kidneydata_altered$after, kidneydata_altered$before, paired=TRUE, data=kidneydata_altered)

        Wilcoxon signed rank test with continuity correction

data:  kidneydata_altered$after and kidneydata_altered$before
V = 19.5, p-value = 0.00893
alternative hypothesis: true location shift is not equal to 0
```

Figure 1.9 = Results from Wilcoxon signed-rank test

## 4.2 Task B

In this task we would have two different contrast media reviewed by three reviewers for each case. As is the contrast media that is to be compared we would use the same technique as in A in order to figure out if the contrast media in themselves has improved the certainty in diagnosis. Once that is established, we would use a two-sided intraclass reliability (ICC) measure to compare the two different contrast media. I.e. taking the ICC of the “after” measurements which has been converted to represent confidence in diagnosis. The reason behind choosing a two-sided ICC is that we have the same set of raters that are evaluating each patient.