Interrater Reliability (Confidence Intervals)

```
# Define a function for formatting numbers
comma <- function(x, d = 2) format(x, digits = d, big.mark = ",")
```

```
library(lme4)
```

```
Warning in check_dep_version(): ABI version mismatch:
lme4 was built with Matrix ABI version 2
Current Matrix ABI version is 1
Please re-install lme4 from source or restore original 'Matrix' package
```

```
library(tidyverse)
```

```
slp_dat <- read.csv("https://osf.io/download/p9gqk/")
slp_vas_wide <- slp_dat |>
    select(slpID, Speaker, slp_VAS) |>
    pivot_wider(names_from = slpID, values_from = slp_VAS)
head(slp_vas_wide)
```

```
# A tibble: 6 x 22
  Speaker slp10 slp11 slp13 slp14 slp15 slp16  slp17 slp18 slp19  slp2 slp20
  <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 AF1      96.3 100   96.6  99.3  100   12.6  100     88.2 52.6   93.4 67.4
2 AF9      12.3  34.7  0.86  3.52  12.6  0      6.69   0    7.42  19.2  9.03
3 ALSF6    NA    30.8  5.73 73.9   73.9  4.52  45.1   41.9 26.1   39.8 17.4
4 ALSF7    94.0  96.3 61.3  64.4  100    7.74  69.9   88.2 37.4   43.3 59.4
5 ALSF9    29.5  52.2 43.6  46.8   60.6 23.9   25.2   91.6 37.1   58.7 19.4
6 ALSM1    96.3  NA   29.2  69.4   99.4 82.3   70.2   79.7 78.4   87.4 36.4
# i 10 more variables: slp21 <dbl>, slp22 <dbl>, slp23 <dbl>, slp25 <dbl>,
#   slp3 <dbl>, slp4 <dbl>, slp5 <dbl>, slp6 <dbl>, slp8 <dbl>, slp9 <dbl>
```

Derivation

Each score consists of grand mean + ratee effect + error. For $i = 1, \dots, k$ and $j = 1, \dots, N$,

$$Y_{ij} = \beta_j + e_{ij}$$

with $\beta_j \sim N(\mu, \sigma^2_{\text{ratee}})$ and $e_{ij} \sim N(0, \sigma^2_E)$

With N ratees, the mean rating for ratee j is

$$\bar{Y}_{.j} = \frac{\sum_i Y_{ij}}{k} = \beta_j + \frac{\sum_i e_{ij}}{k}$$

The grand mean is

$$\bar{Y}_{..} = \frac{\sum_j \sum_i Y_{ij}}{kN} = \frac{\sum_j \beta_j}{N} + \frac{\sum_j \sum_i e_{ij}}{kN}$$

The sum of squares for the ratee effect is

$$\begin{aligned}
SS_{\text{ratee}} &= \sum_j k(\bar{Y}_{.j} - \bar{Y}_{..})^2 \\
&= \sum_j k \left(\beta_j + \frac{\sum_i e_{ij}}{k} - \frac{\sum_j \beta_j}{N} - \frac{\sum_j \sum_i e_{ij}}{kN} \right)^2 \\
&= \sum_j k[(\beta_j - \bar{\beta}) + (\bar{e}_{.j} - \bar{e}_{..})]^2
\end{aligned}$$

Taking the expected value, we have

$$\begin{aligned}
E[\sum_j (\beta_j - \bar{\beta})^2] &= E[\sum_j (\beta_j - \mu + \mu - \bar{\beta})^2] \\
&= E[\sum_j (\beta_j - \mu)^2] - 2E[(\mu - \bar{\beta}) \sum_j (\beta_j - \mu)] + E[\sum_j (\mu - \bar{\beta})^2] \\
&= N\sigma^2_{\text{ratee}} - 2E[N(\mu - \bar{\beta})^2] + E[N(\bar{\beta} - \mu)^2] \\
&= N\sigma^2_{\text{ratee}} - N\sigma^2_{\text{ratee}}/N \\
&= (N-1)\sigma^2_{\text{ratee}}
\end{aligned}$$

Similarly,

$$E[\sum_j (\bar{e}_{\cdot j} - \bar{e}_{\cdot\cdot})^2] = (N-1)\sigma_E^2/k$$

and given β_j and e_{ij} are independent, we have

$$E(SS_{\text{ratee}}) = k(N-1)\sigma_{\text{ratee}}^2 + (N-1)\sigma_E^2.$$

The mean square is the sum of squares divided by the degrees of freedom, which is $N-1$ for the rater effect. Therefore, the expected mean square is

$$E(MS_{\text{ratee}}) = k\sigma_{\text{ratee}}^2 + \sigma_E^2.$$

For sum of squares error, we have

$$\begin{aligned} SS_E &= \sum_j \sum_i k(Y_{ij} - \bar{Y}_{\cdot j})^2 \\ &= \sum_j \sum_i \left(\beta_j + \frac{\sum_i e_{ij}}{k} - \beta_j - e_{ij} \right)^2 \\ &= \sum_j \sum_i (e_{ij} - \bar{e}_{\cdot})^2 \end{aligned}$$

And the expected sum of squares (and mean squares) error can be shown as

$$\begin{aligned} E(SS_E) &= N(k-1)\sigma_E^2 \\ E(MS_E) &= \sigma_E^2 \end{aligned}$$

Therefore, using method of moments, the estimators for σ_E^2 and σ_{ratee}^2 are

$$\begin{aligned} \tilde{\sigma}_E^2 &= MS_E \\ \tilde{\sigma}_{\text{ratee}}^2 &= (MS_{\text{ratee}} - MS_E)/k \end{aligned}$$

Using the above, you can obtain the ICC formulas for "one-way random" of Table 9.5 of your text.

Confidence intervals

From standard ANOVA results Cochran's theorem, with a balanced design, one gets

$$SS_E \sim \sigma_E^2 \chi^2_{N(k-1)}$$
$$SS_{\text{ratee}} \sim (k\sigma^2_{\text{ratee}} + \sigma_E^2)\chi^2_{N-1}$$

and the two sum of squares are independently distributed, and

$$\frac{\sigma_E^2}{k\sigma^2_{\text{ratee}} + \sigma_E^2} \frac{MS_{\text{ratee}}}{MS_E} \sim F_{N-1,N(k-1)}$$

Note that the distribution of $\frac{\sigma_E^2}{k\sigma^2_{\text{ratee}}+\sigma_E^2} \frac{MS_{\text{ratee}}}{MS_E}$ does not depend on the parameters, σ^2_{ratee} and σ_E^2. Therefore, we can invert a two-tailed F test to consider values that are "non-significant" to obtain a confidence interval. Let $F_{\frac{\alpha}{2},N-1,N(k-1)}$ and $F_{1-\frac{\alpha}{2},N-1,N(k-1)}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of a central F distribution with df1 $= N-1$ and df2 $= N(k-1)$, and $F_0 = MS_{\text{ratee}}/MS_E$, then because ICC(1, k) for average rating is

$$\begin{aligned}\frac{\sigma^2_{\text{ratee}}}{\sigma^2_{\text{ratee}} + \sigma_E^2/k} &= \frac{k\sigma^2_{\text{ratee}}}{k\sigma^2_{\text{ratee}} + \sigma_E^2} \\ &= 1 - \frac{\sigma_E^2}{k\sigma^2_{\text{ratee}} + \sigma_E^2},\end{aligned}$$

We have

$$\frac{\sigma_E^2}{k\sigma^2_{\text{ratee}} + \sigma_E^2} \frac{MS_{\text{ratee}}}{MS_E} = [1-\rho(1,k)]F_0 \sim F_{N-1,N(k-1)}$$

Define $F = [1-\rho(1,k)]F_0$, which follows an F distribution. Because $\rho(F) = 1 - F/F_0$ is a monotonic decreasing function with known F_0, we can obtain 95% CI by transforming $F_{\frac{\alpha}{2},N-1,N(k-1)}$ and $F_{1-\frac{\alpha}{2},N-1,N(k-1)}$, so an analytic 95% CI for ICC(1, k) is

$$\left\{1 - \frac{F_{1-\frac{\alpha}{2},N-1,N(k-1)}}{F_0} \le \rho(1,k) \le 1 - \frac{F_{\frac{\alpha}{2},N-1,N(k-1)}}{F_0}\right\}$$

Similarly, for ICC(1, 1) $= \sigma^2_{\text{ratee}}/(\sigma^2_{\text{ratee}} + \sigma_E^2)$, as $F_0/F - 1 = k\sigma^2_{\text{ratee}}/\sigma_E^2 = k\rho(1,1)/[1-\rho(1,1)]$, we have

$$\rho(F) = \frac{F_0/F - 1}{F_0/F - 1 + k},$$

which is monotonic decreasing in F. Therefore, we can again obtain a 95% CI by transforming $F_{\frac{\alpha}{2},N-1,N(k-1)}$ and $F_{1-\frac{\alpha}{2},N-1,N(k-1)}$

$$\left\{ \frac{\frac{F_0}{F_{1-\frac{\alpha}{2},N-1,N(k-1)}}-1}{\frac{F_0}{F_{1-\frac{\alpha}{2},N-1,N(k-1)}}-1+k} \leq \rho(1,1) \leq \frac{\frac{F_0}{F_{\frac{\alpha}{2},N-1,N(k-1)}}-1}{\frac{F_0}{F_{\frac{\alpha}{2},N-1,N(k-1)}}-1+k} \right\}$$

Tip

For 95% CIs with two-way crossed designs, see the following papers:
Fleiss, J.L., Shrout, P.E. Approximate interval estimation for a certain intraclass correlation coefficient. Psychometrika 43, 259–262 (1978). https://doi.org/10.1007/BF02293867
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420
See also this paper for some updated guidelines and recent development on ICC:
ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. Psychological Methods. Advance online publication. https://doi.org/10.1037/met0000516

- Direct link: https://pure.uva.nl/ws/files/73290493/TenHove.etal_2022_GuidelinesSelectingICCforIRR_AuthorVersion.pdf

Demonstrating Analytic CIs in R

Let's look at the results by treating the raters as different for each ratee (which is incorrect, but just to show the calculation for CIs):

```
m1_full <- lmer(slp_VAS ~ 1 + (1 | Speaker), data = slp_dat)
vc_m1f <- as.data.frame(VarCorr(m1_full))
# Obtain mean squares from the variance components
k <- 21
msr <- k * vc_m1f$vcov[1] + vc_m1f$vcov[2]
mse <- vc_m1f$vcov[2]
f0 <- msr / mse
# 95% CI (df1 = 19, df2 = 20 * 20 = 400)
```

```r
# For ICC(1, k)
c("LL" = 1 - 1 / f0 * qf(.975, 19, 400),
  "UL" = 1 - 1 / f0 * qf(.025, 19, 400))
```

```
       LL        UL
0.9410646 0.9844892
```

```r
# For ICC(1, 1)
c("LL" = (f0 / qf(.975, 19, 400) - 1) / (f0 / qf(.975, 19, 400) - 1 + k),
  "UL" = (f0 / qf(.025, 19, 400) - 1) / (f0 / qf(.025, 19, 400) - 1 + k))
```

```
       LL        UL
0.4319368 0.7513950
```

Bootstrap Confidence Intervals

Because the models we fitted are random-effect or mixed-effect models (i.e., multilevel models), we'll need to do multilevel bootstrap. We could use the `bootMer()` available in `lme4` for *parametric* bootstrap; the bootmlm package provides alternative bootstrap methods that are useful with assumption violations.

From my observation, the bootstrap CIs seem wider than the analytic CIs for ICCs, at least for this example. This observation is not unique for bootstrap CIs.

One-Way Random

```r
# One-way random-effect
icc_11 <- function(x) {
    # Function for ICC(1, 1)
    vc_x <- as.data.frame(VarCorr(x))
    vc_x$vcov[1] / sum(vc_x$vcov)
}
boo <- bootMer(m1_full, icc_11, nsim = 1999,
               # change to .progress = "txt" for progress bar
               .progress = "none")
boot::boot.ci(boo, index = 1, type = c("norm", "basic", "perc"))
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates

CALL :
boot::boot.ci(boot.out = boo, type = c("norm", "basic", "perc"),
    index = 1)

Intervals :
Level      Normal              Basic              Percentile
95%   ( 0.4293,  0.7645 )   ( 0.4467,  0.7852 )   ( 0.3734,  0.7119 )
Calculations and Intervals on Original Scale
```

Two-Way Random

```
m2 <- lmer(slp_VAS ~ 1 + (1 | Speaker) + (1 | slpID), data = slp_dat)
# Two-way random-effect
icc_2k <- function(x) {
    # Function for ICC(2, k), which is for consistency
    vc_x <- as.data.frame(VarCorr(x))
    vc_x$vcov[2] / (vc_x$vcov[2] + (vc_x$vcov[1] + vc_x$vcov[3]) / 21)
}
boo2 <- bootMer(m2, icc_2k, nsim = 1999,
                # change to .progress = "txt" for progress bar
                .progress = "none")
boot::boot.ci(boo2, index = 1, type = c("norm", "basic", "perc"))
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates

CALL :
boot::boot.ci(boot.out = boo2, type = c("norm", "basic", "perc"),
    index = 1)

Intervals :
Level      Normal              Basic              Percentile
95%   ( 0.9435,  0.9975 )   ( 0.9518,  1.0033 )   ( 0.9301,  0.9815 )
Calculations and Intervals on Original Scale
```