Package 'reproducer'

April 30, 2019

Title Reproduce Statistical Analyses and Meta-Analyses

Version 0.3.0

Date 2019-04-30

Encoding UTF-8

Maintainer Lech Madeyski <lech.madeyski@gmail.com>

Description Includes data analysis functions (e.g., to calculate effect sizes and 95% Confidence Intervals (CI) on Standardised Effect Sizes (d) for ABBA cross-over repeated-measures experimental designs), data presentation functions (e.g., density curve overlaid on histogram), and the data sets analyzed in different research papers in software engineering (e.g., related to software defect prediction or multi-site experiment concerning the extent to which structured abstracts were clearer and more complete than conventional abstracts) to streamline reproducible research in software engineering.

Depends R (>= 3.5.0)

License CC BY 4.0

LazyData true

URL http://madeyski.e-informatyka.pl/reproducible-research/

Imports openxlsx (>= 2.4.0), ggplot2 (>= 2.0.0), gridExtra(>= 0.9.1), xtable(>= 1.7-4), metafor(>= 1.9-2), lme4(>= 1.1-10), MASS(>= 7.3-45), stats(>= 3.5.2), reshape(>= 0.8.8), GetoptLong(>= 0.1.7), dplyr(>= 0.8.0.1), httr(>= 1.4.0), jsonlite(>= 1.6), tidyr(>= 0.8.3), readr(>= 1.3.1), stringr(>= 1.4.0), magrittr(>= 1.5), tibble(>= 2.1.1)

RoxygenNote 6.1.1

Suggests testthat, assertthat

NeedsCompilation no

Author Lech Madeyski [cre, aut, ctb],
Barbara Kitchenham [ctb] (Data and code contributor),
Tomasz Lewowski [ctb] (Data and code contributor),
Marian Jureczko [ctb] (Data contributor),
David Budgen [ctb] (Data contributor),
Pearl Brereton [ctb] (Data contributor),

Jacky Keung [ctb] (Data contributor),
Stuart Charters [ctb] (Data contributor),
Shirley Gibbs [ctb] (Data contributor),
Amnart Pohthong [ctb] (Data contributor)

Repository CRAN

Date/Publication 2019-04-30 21:00:03 UTC

R topics documented:

aggregateIndividualDocumentStatistics . . . 3
boxplotAndDensityCurveOnHistogram . . . 4
boxplotHV . . . 5
calculateHg . . . 6
calculateSmallSampleSizeAdjustment . . . 7
Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR . . . 8
constructEffectSizes . . . 9
densityCurveOnHistogram . . . 10
effectSizeCI . . . 11
ExtractMAStatistics . . . 12
fmt . . . 13
getEffectSizesABBA . . . 13
getEffectSizesABBAIgnoringPeriodEffect . . . 15
getSimulationData . . . 16
getTheoreticalEffectSizeVariancesABBA . . . 17
KitchenhamMadeyski.SimulatedCrossoverDataSets . . . 18
KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults . . . 19
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes . . . 20
KitchenhamMadeyskiBrereton.DocData . . . 21
KitchenhamMadeyskiBrereton.ExpData . . . 22
KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults . . . 23
KitchenhamMadeyskiBrereton.ReportedEffectSizes . . . 24
KitchenhamMadeyskiBudgen16.COCOMO . . . 25
KitchenhamMadeyskiBudgen16.DiffInDiffData . . . 27
KitchenhamMadeyskiBudgen16.FINNISH . . . 28
KitchenhamMadeyskiBudgen16.PolishData . . . 29
KitchenhamMadeyskiBudgen16.PolishSubjects . . . 31
KitchenhamMadeyskiBudgen16.SubjectData . . . 33
Madeyski15EISEJ.OpenProjects . . . 34
Madeyski15EISEJ.PropProjects . . . 35
Madeyski15EISEJ.StudProjects . . . 36
Madeyski15SQJ.NDC . . . 37
MadeyskiKitchenham.EUBASdata . . . 38
MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR . . . 39
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324 . . . 41
percentageInaccuracyOfLargeSampleVarianceApproximation . . . 43
plotOutcomesForIndividualsInEachSequenceGroup . . . 44
PrepareForMetaAnalysisGtoR . . . 45

printXTable . . . 46
proportionOfSignificantTValuesUsingCorrectAnalysis . . . 47
proportionOfSignificantTValuesUsingIncorrectAnalysis . . . 47
readExcelSheet . . . 48
reproduceForestPlotRandomEffects . . . 49
reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator 49
reproduceMixedEffectsAnalysisWithExperimentalDesignModerator . . . 50
reproduceMixedEffectsForestPlotWithExperimentalDesignModerator . . . 50
reproduceSimulationResultsBasedOn500Reps1000Obs . . . 51
reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments . . . 51
reproduceTableWithEffectSizesBasedOnMeanDifferences . . . 52
reproduceTableWithPossibleModeratingFactors . . . 53
reproduceTableWithSourceDataByCiolkowski . . . 53
searchForIndustryRelevantGitHubProjects . . . 54
transformHgtoR . . . 54
transformHgtoZr . . . 55
transformRtoHg . . . 56
transformRtoZr . . . 57
transformZrtoHg . . . 57
transformZrtoHgapprox . . . 58
transformZrtoR . . . 59

Index **60**

`aggregateIndividualDocumentStatistics`
aggregateIndividualDocumentStatistics

Description

This function assumes an ABBA crossover experiment has reported means and variances for each technique in each time period. We calculate the weighted mean and pooled within group variance for the observations arising from the two different sets of materials for a specific technique.

Usage

```
aggregateIndividualDocumentStatistics(D1.M, D1.SD, D1.N, D2.M, D2.SD, D2.N)
```

Arguments

`D1.M`	is a vector of mean values from a set of experiments in a family reporting observations from participants using a specific document in the first time period with either the control or the treatment technique.
`D1.SD`	is a vector of results from the set of experiment in a family reporting the standard deviations of observations from participants using the same document in the first time period with the same technique.
`D1.N`	is a vector of the numbers of participants in each experiment in a family, using the same document for participants using either the same technique.

D2.M	is a vector of mean values of observations from participants using the alternative document in the second time period, but using the same technique.
D2.SD	is a vector of the standard deviations of observations from participants using the alternative document in the second time period with the same technique.
D2.N	is a vector of the numbers of participants using the same document in the second time period for participants using the same technique.

Value

data frame incl. the overall weighted mean and pooled standard deviation

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
aggregateIndividualDocumentStatistics(10, 2, 20, 15, 2, 20)
#     M SD
#1 12.5  2
```

`boxplotAndDensityCurveOnHistogram`
boxplotAndDensityCurveOnHistogram

Description

Boxplot and density curve overlaid on histogram

Usage

```
boxplotAndDensityCurveOnHistogram(df, colName, limLow, limHigh)
```

Arguments

df	Data frame with data to be displayed
colName	Name of the selected column in a given data frame
limLow	the limit on the lower side of the displayed range
limHigh	the limit on the higher side of the displayed range

Value

A figure being a density curve overlaid on histogram

Author(s)

Lech Madeyski

Examples

```
library(ggplot2)
library(grid)
library(gridExtra)
boxplotAndDensityCurveOnHistogram(Madeyski15EISEJ.PropProjects, "STUD", 0, 100)
boxplotAndDensityCurveOnHistogram(Madeyski15SQJ.NDC, "simple", 0, 100)
```

`boxplotHV` *boxplotHV*

Description

Box plot

Usage

```
boxplotHV(df, colName, limLow, limHigh, isHorizontal)
```

Arguments

`df`	Data frame with data to be displayed
`colName`	Name of the selected column in a given data frame
`limLow`	the limit on the lower side of the displayed range
`limHigh`	the limit on the higher side of the displayed range
`isHorizontal`	Boolean value to control whether the box plot should be horizontal or not (i.e., vertical)

Value

A box plot

Author(s)

Lech Madeyski

Examples

```
boxplotHV(Madeyski15EISEJ.PropProjects, "STUD", 0, 100, TRUE)
boxplotHV(Madeyski15EISEJ.PropProjects, "STUD", 0, 100, FALSE)
boxplotHV(Madeyski15SQJ.NDC, "simple", 0, 100, FALSE)
boxplotHV(Madeyski15SQJ.NDC, "simple", 0, 100, TRUE)
```

calculateHg *calculateHg*

Description

This function calculates Hedges g and Hedges g adjusted given the basic experimental statistics - the mean values for participants, number of observations (participants), and standard deviation in both the control group and the treatment group. . Hence, the function assumes the data is held as summary statistics including the control group mean, standard deviation and sample size and equivalent values for treatment group

Usage

```
calculateHg(Mc, Mt, Nc, Nt, SDc, SDt)
```

Arguments

Mc	is a vector containing the mean value of the control group for each experiment.
Mt	is a vector containing the mean value of the treatment group for each experiment.
Nc	is a vector containing the the number of observations (particpants) in the control group for each experiment.
Nt	is a vector of the number of observations (participants) in the treatment group for each experiment.
SDc	is a vector of the standard deviations of the control group for each experiment.
SDt	is a vector of the standard deviations of the the treatment group for each experiment.

Value

data frame composed of Hedges' g and Hedges' g adjusted effect sizes

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
calculateHg(10, 15, 20, 20, 2, 2)
#    Hg    HgAdjusted
# 1  2.5   2.450276
```

calculateSmallSampleSizeAdjustment
calculateSmallSampleSizeAdjustment

Description

Function calculates the small sample size adjustment for standardized mean effect sizes

Usage

```
calculateSmallSampleSizeAdjustment(df, exact = TRUE)
```

Arguments

df	A vector of degrees of freedom
exact	Default value=TRUE, If exact==TRUE the function returns the exact value of the adjustment(s) which is suitable for small values of df, if exact==FALSE the function returns the approximate version of the adjustment(s). See Hedges and Olkin 'Statistical methods for Meta-Analysis' Academic Press 1985.

Value

small sample size adjustment value

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
df <- 2
c <- calculateSmallSampleSizeAdjustment(df)

df=c(5,10,17)
adjexact=calculateSmallSampleSizeAdjustment(df)
# adjexact=0.8407487 0.9227456 0.9551115
# Hedges and Olkin values 0.8408, 0.9228,0.9551
adjapprox=calculateSmallSampleSizeAdjustment(df,FALSE)
# adjapprox=0.8421053 0.9230769 0.9552239
```

`Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR`

Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR data form a set of primary studies on reading methods for software inspections. They were reported and analysed by M. Ciolkowski ("What do we know about perspective-based reading? an approach for quantitative aggregation in software engineering", in Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, ESEM'09, pp. 133-144, IEEE Computer Society, 2009), corrected and re-analysed by Madeyski and Kitchenham ("How variations in experimental designs impact the construction of comparable effect sizes for meta-analysis" (to be submitted)).

Description

If you use this data set please cite: Lech Madeyski and Barbara Kitchenham, "How variations in experimental designs impact the construction of comparable effect sizes for meta-analysis", 2015.

Usage

```
Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR
```

Format

A data frame with 21 rows and 7 variables:

Study Name of empirical study

Ref. Reference to the paper reporting primary study or experimental run where data were originally reported

Control Control treatment: Check-Based Reading (CBR) or Ad-hoc Reading (AR)

Within-subjects Yes - if the primary study used the within-subjects experimental design, No - if the primary study did not use the within-subjects experimental design

Cross-over Yes - if the primary study used the cross-over experimental design, No - if the primary study did not use the cross-over experimental design

d_ByCiolkowski d effect size calculated by Ciolkowski

d_ByOriginalAuthors d effect size as reported by the original authors

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR
```

constructEffectSizes *constructEffectSizes*

Description

The function constructs various different d-style effect sizes for a set of different experiments given basic statistics from each experiment (the mean value of the control group Mc, the mean value of the treatment group Mt, the standard deviation of the control group SDc, standard deviation of the the treatment group SDt, the number of observations (particpants) in the control group Nc, and the number of observations (participants) in the treatment group Nt). The input variables can be vectors or individual numbers but all input vectors must be of the same length. The function returns Glass's Delta, Cohen's D, point bi-serial r (based on Hedges'g unadjusted), Hedges'g and Hegdes' g adjusted for small sample size.

Usage

```
constructEffectSizes(Mc, Mt, SDc, SDt, Nc, Nt)
```

Arguments

Mc	is a vector containing the mean value of the control group for each experiment.
Mt	is a vector containing the mean value of the treatment group for each experiment.
SDc	is a vector of the standard deviations of the control group for each experiment.
SDt	is a vector of the standard deviations of the the treatment group for each experiment.
Nc	is a vector containing the the number of observations (particpants) in the control group for each experiment.
Nt	is a vector of the number of observations (participants) in the treatment group for each experiment.

Value

data frame composed of five effect sizes (Glass delta, Cohen's d, Hedges' g, r, Hedges' g adjusted)

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
constructEffectSizes(10, 15, 0.3, 0.2, 15, 15)

Mt = c(0.633, 0.673, 0.423, 0.727, 0.631)
Mc = c(0.612, 0.526, 0.356, 0.618, 0.534)
SDt = c(0.198, 0.115, 0.172, 0.088, 0.122)
SDc = c(0.159, 0.089, 0.111, 0.166, 0.119)
Nt = c(12, 12, 14, 10, 8)
```

```
Nc= c(12, 12, 14, 10, 8)
EffectSizes=constructEffectSizes(Mc, Mt, SDc,SDt,Nt,Nc)
EffectSizes
# GlassDelta    Cohend   Hedgesg          r HedgesgAdjusted
# 1  0.1320755 0.1221516 0.1169513 0.05837591       0.1129107
# 2  1.6516854 1.4931812 1.4296121 0.58151846       1.3802200
# 3  0.6036036 0.4803405 0.4628677 0.22547423       0.4493641
# 4  0.6566265 0.8648343 0.8204538 0.37953300       0.7857047
# 5  0.8151261 0.8604924 0.8049169 0.37335594       0.7608781
```

`densityCurveOnHistogram` *densityCurveOnHistogram*

Description

Density curve overlaid on histogram

Usage

```
densityCurveOnHistogram(df, colName, limLow, limHigh)
```

Arguments

`df`	Data frame with data to be displayed
`colName`	Name of the selected column in a given data frame
`limLow`	the limit on the lower side of the displayed range
`limHigh`	the limit on the higher side of the displayed range

Value

A figure being a density curve overlaid on histogram

Author(s)

Lech Madeyski

Examples

```
densityCurveOnHistogram(Madeyski15EISEJ.PropProjects, "STUD", 0, 100)
densityCurveOnHistogram(data.frame(x<-rnorm(50, mean=50, sd=5)), "x", 0, 100)
```

effectSizeCI *effectSizeCI*

Description

95 The procedure is based on finding the upper and lower 0.025 bounds for the related t-variable. The t-variable needs to be adjusted for bias by multiplying by c The upper and lower bounds on the t-variable are then used to calculate to upper and lower bounds on the repeated measures effect size (d_RM) by multiplying the upper and lower bound of the t-variable by sqrt((n1+n2)/(2*(n1*n2))). Upper and lower bounds on the equivalent independent groups effect size (d_IG) are found by multiplying the upper and lower bounds on d_RM by sqrt(1-r).

Usage

```
effectSizeCI(expDesign, t, n1, n2, r = 0, epsilon = 1e-10,
  maxsteps = 1000, stepsize = 3)
```

Arguments

expDesign	Experimental design: 1) crossover repeated measures ("CrossOverRM"), 2) before-after repeated measures (expDesign=="BeforeAfterRM"), 3) independent groups ("IG)
t	t-statistics (t must be less than or equal to 37.62, the limit from the R function documentation)
n1	The number of observations in sequence group 1 (expDesign=="CrossOverRM"), the number of observations in group 1 (expDesign=="IG"), or the total number of observations (expDesign=="BeforeAfterRM")
n2	The number of observations in sequence group 2 (expDesign=="CrossOverRM") or the number of observations in group 2 (expDesign=="IG")
r	The correlation between outcomes for individual subject (the within subject correlation)
epsilon	The precision of the iterative procedure
maxsteps	The maximum number of steps of the iterative procedure (the procedure terminates at maxsteps or earlier if CI with enough precision have been calculated)
stepsize	The size of steps (influences the convergence of the calculations, i.e., the number of steps required to obtain the final result of precision defined by the epsilon)

Value

A list of Confidence Intervals for: t-statistic (t_LB and t_UB), repeated-measures effect size d_RM (d_RM_LB, d_RM_UB), independent groups effect size (d_IG_LB, d_IG_UB)

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
effectSizeCI(expDesign="CrossOverRM", t=14.4, n1=15, n2=15, r=0.6401)
effectSizeCI(expDesign = "BeforeAfterRM", t=14.16536, n1=15, n2=0, r=0.6146771)
effectSizeCI(expDesign = "IG", t=-6.344175, n1=15, n2=15)
effectSizeCI(expDesign="CrossOverRM", t=0.5581, n1=6, n2=6, r=0.36135)
effectSizeCI(expDesign = "CrossOverRM", r=0.855,t=4.33, n1=7, n2=6)
```

`ExtractMAStatistics` *ExtractMAStatistics*

Description

This function extracts summary statistics from meta-analysis results obtained from the rma function of the metafor R package. If required the function transform back to standardized mean difference (effect size type "d" i.e. Hg) or point biserial correlations (effect size type "r"). Warning: the 'ExtractMAStatistics' function works with 'metafor' version 2.0-0, but changes to metafor's method of providing access to its individual results may introduce errors into the function.

Usage

```
ExtractMAStatistics(maresults, Nc, Nt, Transform = TRUE, type = "d",
  sig = 4)
```

Arguments

`maresults`	is the output from the rma function.
`Nc`	is the number of participants in the control condition group.
`Nt`	is the number of participants in the treatment condition group.
`Transform`	is a boolean value indicating whether the outcome values need to be transformed back to standardized mean difference ("d" i.e. Hg) or point biserial correlations ("r"). It is defaulted to TRUE. If this parameter is set to FALSE, no transformation will be applied.
`type`	this indicates the type of transformation required - it defaults to "d" which requests transformation from Zr to Hg, using "r" requests transformation from Zr to r.
`sig`	indicates the number of significant digits requested in the output, the default is 4; it rounds the values of mean, pvalue, upper and lower bound to the specified number of significant digits.

Value

data frame incl. summary statistics from meta-analysis results: overall mean value for the effect sizes, the p-value of the mean, the upper and lower confidence interval bounds (UB and LB), QE which is the heterogeneity test statistic and QEp which the the p-value of the heterogeneity statistic

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
ExpData=reproducer::KitchenhamMadeyskiBrereton.ExpData
#Extract the experiment basic statics
S1data=subset(ExpData,ExpData=="S1")
#Use the descriptive data to construct effect size
S1EffectSizes = reproducer::PrepareForMetaAnalysisGtoR(
S1data$Mc,S1data$Mt,S1data$SDc,S1data$SDt,S1data$Nc,S1data$Nt)
# Do a random effect meta-analysis of the transformed r_pbs effect size
S1MA = metafor::rma(S1EffectSizes$zr, S1EffectSizes$vi)
# Extract summary statistics from meta-analysis results and transform back to Hg scale
S1MAStats=reproducer::ExtractMAStatistics(S1MA, sum(S1data$Nc),sum(S1data$Nt), TRUE, "d", 4)
#    mean   pvalue    UB     LB QE  QEp
#1 0.6658 0.002069 1.122 0.2384  4 0.41
```

---

fmt *fmt*

Description

Formatting function to set decimal precision in labels

Usage

```
fmt()
```

Author(s)

Lech Madeyski

---

getEffectSizesABBA *getEffectSizesABBA*

Description

Function to calculate both effect sizes (dIG, dRM), i.e., independent groups and repeated measures standardized effect sizes and variances, for AB/BA crossover design studies. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
getEffectSizesABBA(simulationData)
```

Arguments

`simulationData` - data set in a form required to calculate effect sizes in AB/BA crossover experimental designs

Value

data frame incl. calculated effect sizes and variances: # dIG - independent groups standardized effect size # var.dIG - variance of independent groups standardized effect size # dRM - repeated measures (within-subjects) standardized effect size # var.dRM - variance of repeated measures (within-subjects) standardized effect size # dIG.Fromt - independent groups standardized effect size calculated from t: dIG.Fromt=t*sqrt(1-r)*sqrt((N1+N2)/(2*N1*N2)) # var.dIG.Fromt - variance of independent groups standardized effect size calculated from t: var.dIG.Fromt=var.t*(1-r)*((N1+N2)/(2*N1*N2)) # dRM.Fromt - dRM calculated from t: dRM.Fromt=t*sqrt((N1+N2)/(2*N1*N2)) # var.dRM.Fromt - var.dRM calculated from t: var.dRM.Fromt = var.t*((N1+N2)/(2*N1*N2)) # var.dRM.Fromt2 - var.dRM calculated from t or rather dRM.Fromt: var.dRM.Fromt2=(df/(df-2))*((N1+N2)/(2*N1*N2)+dRM dRM.Fromt^2/c^2 # var.dRM.Approx - var.dRM calculated on a basis of Johnson and Welch (1940) report an approximate formulate for the variance of a t variable: var.dRM.Approx=((N1+N2)/(2*N1*N2)) + (dRM^2)/(2*(N1+N2-2)) #see paper and Equation 49 # var.dIG.Approx - var.dIG calculated on a basis of Johnson and Welch (1940) report an approximate formulate for the variance of a t variable: var.dIG.Approx=(((N1+N2)*(1-r))/(2*N1*N2)) + (dIG^2)/(2*(N1+N2-2)) #see paper and Equation 50 # unstandardizedES - estimated unstandardized technique effect size # periodES - estimated period effect # var.sig - sum of within-subjects variance and between-subjects variance # var.within - within-subjects variance # var.between - between-subjects variance # t - t-value # var.t - variance of t-variable # gRM - Hedges and Olkin (1985) unbiased estimator of the repeated measures effect size gRM=dRM*c # var.gRM - variance of gRM calculated as follows: var.gRM=(df/(df-2))*(((N1+N2)/(2*N1*N2))*c^2+gRM^2)- gRM^2/c^2 #Equation 56 # var.gRM2 - variance of gRM calculated as follows: var.gRM2=var.dRM*c^2 # gIG - Hedges and Olkin (1985) unbiased estimator of the independent groups effect size gIG=dIG*c # var.gIG - variance of gIG calculated as follows: var.gIG=(df/(df-2))*(((N1+N2)/(2*N1*N2))*c^2+gIG^2)- gIG^2/c^2 #Equation 57 # var.gIG2 - variance of gRM calculated as follows: var.gIG2=var.dIG*c^2 # r - the correlation between the values observed for the same subject

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
simulationData<-getSimulationData(25, 18.75, 50, 10, 5, 500) #generate simulated data set
es<-getEffectSizesABBA(simulationData) #return effect sizes and variances
#OR
simulationData<-getSimulationData(25, 18.75,50,10,5,15)
es<-getEffectSizesABBA(simulationData) #return effect sizes and variances
```

getEffectSizesABBAIgnoringPeriodEffect
getEffectSizesABBAIgnoringPeriodEffect

Description

Function to calculate both effect sizes (dIG.ipe, dRM.ipe), i.e., independent groups and repeated measures standardized effect sizes and variances, for AB/BA crossover design studies ignoring period effect (thus wrong). Function was removed in the revision of the paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
getEffectSizesABBAIgnoringPeriodEffect(simulationData)
```

Arguments

`simulationData`	- data set in a form required to calculate effect sizes in AB/BA crossover experimental designs

Value

data frame incl. calculated effect sizes and variances: # dIG.ipe - independent groups standardized effect size # var.dIG.ipe - variance of independent groups standardized effect size # dRM.ipe - repeated measures (within-subjects) standardized effect size # var.dRM.ipe - variance of repeated measures (within-subjects) standardized effect size # dIG.Fromt.ipe - independent groups standardized effect size calculated from t: dIG.Fromt=t*sqrt(1-r)*sqrt((N1+N2)/(2*N1*N2)) # var.dIG.Fromt.ipe - variance of independent groups standardized effect size calculated from t: var.dIG.Fromt=var.t*(1-r)*((N1+N2)/(2*N1*N2)) # dRM.Fromt.ipe - dRM calculated from t: dRM.Fromt=t*sqrt((N1+N2)/(2*N1*N2)) # var.dRM.Fromt.ipe - var.dRM calculated from t: var.dRM.Fromt = var.t*((N1+N2)/(2*N1*N2)) # var.dRM.Fromt2.ipe - var.dRM calculated from t or rather dRM.Fromt: var.dRM.Fromt2=(df/(df-2))*((N1+N2)/(2*N1*N2)+dRM.Fromt^2)- dRM.Fromt^2/c^2 # unstandardizedES.ipe - estimated unstandardized technique effect size # var.sig.ipe - sum of within-subjects variance and between-subjects variance # var.within.ipe - within-subjects variance # var.between.ipe - between-subjects variance # t.ipe - t-value # var.t.ipe - variance of t-variable

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
simulationData<-getSimulationData(25, 18.75, 50, 10, 5, 500) #generate simulated data set
es.ipe<-getEffectSizesABBAIgnoringPeriodEffect(simulationData) #return effect sizes and variances
```

getSimulationData *getSimulationData*

Description

Function to generate the simulated data set used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham

Usage

```
getSimulationData(var, covar, meanA1, treatmentDiff, periodEffect,
  numOfSamples)
```

Arguments

var	Variance among subjects is a sum of the between subjects variance and the within subjects variance
covar	Covariance equal to the between subjects variance
meanA1	Mean for treatment sequence A1
treatmentDiff	technique effect which is the difference between the effect of technique A and technique B
periodEffect	Period effect which is the difference between period 1 and period 2
numOfSamples	Number of samples ("rows" of data) required for each technique and period

Details

———————————————————————————————————- Functions related to a paper "Effect sizes and their variance for AB/BA crossover design studies" by Lech Madeyski and Barbara Kitchenham ———————————————————————————————————-

Value

Data frame: 'data.frame': 4*numOfSamples obs. of 5 variables: $ pid : int 1 2 3 4 5 6 7 8 9 10 ... $ technique: Factor w/ 2 levels "T1","T2": ... $ period : Factor w/ 2 levels "P1","P2": ... $ sequence : Factor w/ 2 levels "S1","S2": ... $ result : num ...

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
data<-getSimulationData(25, 18.75, 50, 10, 5, 500) # generate the simulated data set from the paper
data<-getSimulationData(25, 18.75, 50, 10, 5, 15)
```

getTheoreticalEffectSizeVariancesABBA
getTheoreticalEffectSizeVariancesABBA

Description

Function provides the theoretical value of the t-statistic, variance of t, and variance of the effect sizes based on the parameters built into crossover model data simulated by the getSimilationData() function. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
getTheoreticalEffectSizeVariancesABBA(theoreticalvarW,
  theoreticalTechniqueEffect, theoreticalrho, N1, N2)
```

Arguments

theoreticalvarW	- The within subject variance used to construct the simulation, i.e., the built-in Variance - the built-in Covariance
theoreticalTechniqueEffect	- The technique effect built into the crossover model data
theoreticalrho	- The between subject correlation built into the crossover model simulation data
N1	- The number of subjects in sequence group 1 in the crossover model simulation
N2	- The number of subjects in sequence group 2 in the crossover model simulation

Value

data frame incl. calculated: theoreticalt - the theoretical value of the t-statistic theoreticalvart - variance of t theoreticalvardIG - variance of the effect size dIG based on the parameters built into crossover model data simulated by the getSimilationData function theoreticalvardRM - variance of the effect size dRM based on the parameters built into crossover model data simulated by the getSimilationData function

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
# Generates data used in Table 15 of the paper
theoreticalEffectSizeVariances <- getTheoreticalEffectSizeVariancesABBA(6.25,-10,0.75,15,15)
```

KitchenhamMadeyski.SimulatedCrossoverDataSets
KitchenhamMadeyski.SimulatedCrossoverDataSets data

Description

If you use this data set please cite this R package and the following paper: Lech Madeyski and Barbara Kitchenham, "Effect Sizes and their Variance for AB/BA Crossover Design Studies", Empirical Software Engineering, vol. 24, no.4, p. 1982-2017, 2018. DOI: 10.1007/s10664-017-9574-5

Usage

```
KitchenhamMadeyski.SimulatedCrossoverDataSets
```

Format

A data frame with variables:

actualSampleSize Sample size

SSFull Sample Size

CFull Correlation

ESFull Effect Size

Accuracy Accuracy

PropSig ...

WrongTSig ...

Details

This is simulated normally distributed data from 30 subjects, with technique A being 10 units more effective than technique B, and there is a period effect equaling 5 units. Subject 1 to 15 used technique B first while subjects 16 to 30 used technique A first.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyski.SimulatedCrossoverDataSets
```

KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults *KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults data*

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This data set reports the meta-analysis results reported by the authors of the primary studies included in the systematic review that reported results on a per document basis which for S7 and S11 was equivalent to reporting the results for each time period.

Usage

```
KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults
```

Format

A text file file with variables:

Study This field includes the study identifier of each of the the 3 primary studies which reported results per document.

Type This identifies the type of effect size used by the study authors. d or g refer to d_IG and g_IG, P is the aggregated p values, if the repeated measures (RM) estimate was obtained it is appropropriately specified.

Source Always set to Rep. This identifies that the data was as reported by the primary study authors.

mean The overall mean effect size reported by the study authors

pvalue The one-sided p-value associated with the overall mean reported by the study authors. NA means the authors did not report this statistic.

UB The upper bound of the confidence interval of the overall mean as reported by the primary study authors. NA means the authors did not report this statistic.

LB The lower bound of the confidence interval of the overall mean as reported by the primary study authors. NA means the authors did not report this statistic.

QE The heterogeneity statistic associated with the meta-analysis as reported by the study authors. NA means the authors did not report this statistic.

Qep The p-value of the heterogeneity statistic associated with the meta-analysis as reported by the study authors. NA means the authors did not report this statistic.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults
```

`KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes`
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes data

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This file holds the individual effect sizes for the first time period (or equivalently the first document), as reported by the 3 primary studies in the systematic review that reported results for each document/time period separately.

Usage

```
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes
```

Format

A text file file with variables:

Study This field includes the study identifer of each of the 3 primary studies which were included in the systematic review. The studies are S3, S7 and S11.

Type This identifies the type of effect size used by the study authors. d or g refer to dIG and gIG.

Source Always set to Rep. This identifies that the data was as reported by the primary study authors.

Design Mixed means different experiments in a particular family used different methods (onlyS3 used mixed methods and 4 experiments used the 4 goup crossover and one used an independent groups design). ABBACO is the standard 2-group crossover design.

Exp1 This is the reported standardised effect size for the first time period and the first experiment in the family.

Exp2 This is the reported standardised effect size for the first time period and second experiment in the family.

Exp3 This is the reported standardised effect size for the first time period and the third experiment in the family.

Exp4 This is the reported standardised effect size for the first time period and the fourth experiment in the family. NA means there was no fourth experiment in the family.

Exp5 This is the reported standardised effect size for the first time period and the fifth experiment in the family. NA means there was no fifth experiment in the family.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes
```

`KitchenhamMadeyskiBrereton.DocData`

KitchenhamMadeyskiBrereton.DocData data

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This file holds the descriptive data for each document and each experiment for studies 3, 7 and 11 which include the mean, standard deviation and sample size for the control and treatment techniques. These studies performed ABBA crossover experiments and reported data for each document separately. Note Study 3 also undertook an independent groups study but data from that experiment is held in the ExpData file.

Usage

```
KitchenhamMadeyskiBrereton.DocData
```

Format

A text file file with variables:

Study This field includes the study identifier of each of the 3 primary studies which reported their basic statistics on a time period & document basis.

Exp This identifies the experiment to which the descriptive data belongs.

Doc This identifies whether the data arose from the document used in the first or second time period. The value "Doc1" identifies the data as coming from the first document or first time period. The value "Doc2" identifies the data as comning from the second time period or document. Note for Study 3 we used the analysis of a specific document that was used in all 4 ABBA experiments. For studies 7 and 11, the authors identified which we used in r=each time period and Doc1 refers to data from the first time period.

Mc The mean value of the observations obtained using the control technique for the identified document.

SDc The standard deviation of the observations obtained using the control technique for the identified document.

Nc The number of participants using the control technique in the first time period for the identified document.

Mt The mean value of the observations obtained using the treatment technique for the identified document.

SDt The standard deviation of the observations obtained using the treatment technique for the identified document.

Nt The number of participants using the treatment technique in the first time period for the identified document.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.DocData
```

`KitchenhamMadeyskiBrereton.ExpData`

KitchenhamMadeyskiBrereton.ExpData data

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This file holds the descriptive data for each experiment which include the mean, standard deviation and sample size for the control and treatment techniques. Note in the case of studies 3, 7 and 11, which reported descriptive data for each time period (or equivalently each document) separately, the values for of the desciptive data were obtained by analysing the data reported in the DocData file.

Usage

```
KitchenhamMadeyskiBrereton.ExpData
```

Format

A text file file with variables:

Study This field includes the study identifier of each of the 13 primary studies which were included in the systematic review.

Exp This identifies the experiment to which the descriptive data belongs.

Source Always set to Rep. This identifies that the data was as reported by the primary study authors.

Mc The mean value of the observations obtained using the control technique.

SDc The standard deviation of the observations obtained using the control technique.

Nc The number of participants using the control technique in the first time period.

Mt The mean value of the observations obtained using the treatment technique.

SDt The standard deviation of the observations obtained using the treatment technique.

Nt The number of participants using the treatment technique in the first time period.

r The correlation between repeated measures. NA if not reported. Note only study 13 reported this correlation.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.ExpData
```

`KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults` *KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults data*

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This data set reports the meta-analysis results reported by the authors of the 13 primary studies included in the systematic review.

Usage

```
KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults
```

Format

A text file file with variables:

Study This field includes the study identifier of each of the 13 primary studies which were included in the systematic review.

Type This identifies the type of effect size used by the study authors. d or g refer to d_IG and g_IG, P is the aggregated p values, if the repeated measures estimate was obtained it is appropriately specified, r refers to the point bi-serial correlation.

Source Always set to Rep. This identifies that the data was as reported by the primary study authors.

mean The overall mean effect size reported by the study authors

pvalue The one-sided p-value associated with the overall mean reported by the study authors. NA means the authors did not report this statistic.

UB The upper bound of the confidence interval of the overall mean as reported by the primary study authors. NA means the authors did not report this statistic.

LB The lower bound of the confidence interval of the overall mean as reported by the primary study authors. NA means the authors did not report this statistic.

QE The heterogeneity statistic associated with the meta-analysis as reported by the study authors. NA means the authors did not report this statistic.

Qep The p-value of the heterogeneity statistic associated with the meta-analysis as reported by the study authors. NA means the authors did not report this statistic.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults
```

`KitchenhamMadeyskiBrereton.ReportedEffectSizes`
KitchenhamMadeyskiBrereton.ReportedEffectSizes data

Description

This data is used in the paper: Barbara Kitchenham, Lech Madeyski and Pearl Brereton. Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment (to be submitted). This file holds the individual effect sizes for each experiment, as reported by 13 primary studies in the systematic review.

Usage

```
KitchenhamMadeyskiBrereton.ReportedEffectSizes
```

Format

A text file file with variables:

Study This field includes the study identifier of each of the 13 primary studies which were included in the systematic review.

Type This identifies the type of effect size used by the study authors. d or g refer to dIG and gIG, p is the p-value used for aggregation, if the repeated measures estimate was obtained it is appropropriately specified as gRM, r refers to the point bi-serial correlation.

Source Always set to Rep. This identifies that the data was as reported by the primary study authors.

Design The refers to the design method used by the study author. 4GroupCO is a 4-group crossover design. Mixed means different experiments in a particular family used different methods (only S3 used mixed methods and 4 experiments used the 4 goup crossover and one used an independent groups design). ABBACO is the standard 2-group crossover design. IndGroups is the independent groups design also called between groups design or a randomised design. PrePost is pretest and posttest design with a post test control.

Exp1 This is the reported standardised effect size for the first experiment in the family.

Exp2 This is the reported standardised effect size for the second experiment in the family.

Exp3 This is the reported standardised effect size for the third experiment in the family.

Exp4 This is the reported standardised effect size for the fourth experiment in the family. NA means there was no fourth experiment in the family.

Exp5 This is the reported standardised effect size for the fifth experiment in the family. NA means there was no fifth experiment in the family.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBrereton.ReportedEffectSizes
```

`KitchenhamMadeyskiBudgen16.COCOMO`

KitchenhamMadeyskiBudgen16.COCOMO data

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no.2, p. 579-630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiE

Usage

```
KitchenhamMadeyskiBudgen16.COCOMO
```

Format

A data frame with variables:

Project Project ID

Type A categorical variable describing the type of the project

Year The year the project was completed

Lang A categorical variable describing the development language used

Rely Ordinal value defining the required software reliability

Data Ordinal value defining the data complexity / Data base size

Cplx Ordinal value defining the complexity of the software / Process complexity

Aaf ??

Time Ordinal value defining the stringency of timing constraints / Time constraint for cpu

Stor Ordinal value defining the stringency of the data storage requirements / Main memory constraint

Virt Virtual Machine volatility

Turn Turnaround time

Type2 A categorical variable defining the hardware type: mini, max=mainframe, midi

Acap Ordinal value defining the analyst capability

Aexp Ordinal value defining the analyst experience / application experience

Pcap Ordinal value defining the programming capability of the team / Programmers capability

Vexp Ordinal value defining the virtual machine experience of the team

Lexp Ordinal value defining the programming language experience of the team

Cont ??

Modp / Modern programming practices

Tool Ordinal value defining the extent of tool use / Use of software tools

ToolCat Recoding of Tool to labelled ordinal scale

Sced Ordinal value defining the stringency of the schedule requirements / Schedule constraint

Rvol Ordinal value defining the requirements volatility of the project

Select Categorical value calculated by BAK for an analysis example

Rvolcat Recoding of Rvol to a labelled ordinal scale

Modecat Mode of the projects: O=Organic, E=Embedded, SD-Semi-Detached

Mode1 Dummy variable calculated by BAK: 1 if the project is Organic, 0 otherwise

Mode2 Dummy variable calculated by BAK: 1 if the project is Semi-detached, 0 otherwise

Mode3 Dummy variable calculated by BAK: 1 if the project is Embedded, 0 otherwise

KDSI Product Size Thousand of Source Instructions

AKDSI Adjusted Product Size for Project in Thousand Source Instructions - differs from KDSI for enhancement projects

Effort Project Effort in Man months

Duration Duration in months

Productivity Productivity of project calculated by BAK as AKDSI/Effort, so the the larger the value the better the productivity

Details

Data set collected at TRW by Barry Boehm see: B.W. Boehm. 1981. Software Engineering Economics. Prentice-Hall.

Explanations by Barbara Kitchehnam / https://terapromise.csc.ncsu.edu:8443/!/#repo/view/head/effort/cocomo/cocomo1/nas

COCOMO.txt: pro type year Lang Rely Data CPLX aaf time store virt turn type2 acap aexp pcap vexp lexp cont modp TOOL TOOLcat SCED RVOL Select rvolcat Modecat Mode1 Mode2 Mode3 KDSI AKDSI Effort Dur Productivity

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.COCOMO
```

`KitchenhamMadeyskiBudgen16.DiffInDiffData`
KitchenhamMadeyskiBudgen16.DiffInDiffData data

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no.2, p. 579-630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiE

Usage

```
KitchenhamMadeyskiBudgen16.DiffInDiffData
```

Format

A data frame with variables:

Abstract The abstract identifier

Site A numeric identifier of the site

Treatment A three character alphanumeric identifying the jounral and time period of the abstract

Journal The journal in which the abstract was published: IST or JSS

Timeperiod The time period in which the abstract: 1 or 2

J1 The identifer for the judge who made the next 2 assessments

J1Completeness The average completeness made by judge J1 based on the 8 completeness questions

J1Clarity The clarity assessment made by judge J1

J2 The identifer for the judge who made the next 2 assessments

J2Completeness The average completeness made by judge J2 based on the 8 completeness questions

J2Clarity The clarity assessment made by judge J2

J3 The identifer for the judge who made the next 2 assessments

J3Completeness The average completeness made by judge J3 based on the 8 completeness questions

J3Clarity The clarity assessment made by judge J3

J4 The identifer for the judge who made the next 2 assessments

J4Completeness The average completeness made by judge J4 based on the 8 completeness questions

J4Clarity The clarity assessment made by judge J4

MeanCompleteness The mean of J1Completeness, J2Completeness, J3Completeness, J4Completeness

MedianCompleteness The median of J1Completeness, J2Completeness, J3Completeness, J4Completeness

MedianClarity The median clarity of J1Clarity, J2Clarity, J3Clarity, J4Clarity

MeanClarity The mean clarity of J1Clarity, J2Clarity, J3Clarity, J4Clarity

VarCompleteness The variance of J1Completeness, J2Completeness, J3Completeness, J4Completeness

VarClarity The variance clarity of J1Clarity, J2Clarity, J3Clarity, J4Clarity

Details

Data set was derived from the data reported in the SubjectData data set (subjectdata.txt). It contains the summary completeness and clarity data from 4 judges who assessed the same abstract. Only the initial 5 sites are included.

dinddata.txt

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.DiffInDiffData
```

`KitchenhamMadeyskiBudgen16.FINNISH`
KitchenhamMadeyskiBudgen16.FINNISH data

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no.2, p. 579-630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiE

Usage

```
KitchenhamMadeyskiBudgen16.FINNISH
```

Format

A data frame with variables:

Project Project ID

DevEffort Development Effort measured in hours

UserEffort Effort provided by the customer/user organisation measured in hours

Duration Project duration measurted in months

HWType A catagorical variable defining the hardware type

AppType A categorical variable definiting the application type

FP Function Points measured using the TIEKE organisation method

Co A categorical variable defining the company

Details

Data set collected from 9 Finish companies by Mr Hanna M\"aki from the TIEKE organisation see Barbara Kitchneham and Kari K\"ans\"al\"a, Inter-item correlations among function points, Proceedings ICSE 15, 1983, pp 477-480

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.FINNISH
```

KitchenhamMadeyskiBudgen16.PolishData

KitchenhamMadeyskiBudgen16.PolishData data

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no.2, p. 579-630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiE

Usage

```
KitchenhamMadeyskiBudgen16.PolishData
```

Format

A data frame with variables:

Abstract The abstract identifier

Site Numeric identifier for the site

Treatment The first three characters of the Abstract field which identies the journal and time period of the abstract

Journal An acronym for the journal from which the abstarct was obtained: IST or JSS

Timeperiod The Time period in which the abstarct was found: 1 or 2

J1 The identifier for the judge who made the next 2 assessments

J1Completeness The average completeness made by judge J1 based on the 8 completeness questions

J1Clarity The clarity assessment made by judge J1

J2 The identifer for the judge who made the next 2 assessments

J2Completeness The average completeness made by judge J2 based on the 8 completeness questions

J2Clarity The clarity assessment made by judge J2

J3 The identifer for the judge who made the next 2 assessments

J3Completeness The average completeness made by judge J3 based on the 8 completeness questions

J3Clarity The clarity assessment made by judge J3

J4 The identifer for the judge who made the next 2 assessments

J4Completeness The average completeness made by judge J4 based on the 8 completeness questions

J4Clarity The clarity assessment made by judge J4

MedianCompleteness The median of J1Completeness, J2Completeness, J3Completeness, J4Completeness

MedianClarity The median of J1Clarity, J2Clarity, J3Clarity, J4Clarity

Details

Data set derived from PolishSubjects data set collected at Wroclaw University. It summarizes the completeness and clarity data collected from 4 judges about the same abstract.

PolishData.txt

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.PolishData
```

KitchenhamMadeyskiBudgen16.PolishSubjects
KitchenhamMadeyskiBudgen16.PolishSubjects data

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no.2, p. 579-630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMadeyskiE

Usage

```
KitchenhamMadeyskiBudgen16.PolishSubjects
```

Format

A data frame with variables:

Judge The identifier for each subject

Abstract The identifier for each abstract - the code starts with a three alphanumeric string that defines the source of the abstract

OrderViewed Each judge assessed 4 abstracts in sequence, this data item identifies the order in which the subject viewed the specified abstract

Completness1 Assessment by judge of question 1:Is the reason for the project clear? Can take values: Yes/No/Partly

Completness2 Assessment by judge of question 2: Is the specific aim/purpose of the study clear? Can take values: Yes/No/Partly

Completness3 Assessment by judge of question 3: If the aim is to describe a new or enhanced software technology (e.g. method, tool, procedure or process) is the method used to develop this technology defined? Can take values: Yes/No/Partly/NA

Completness4 Assessment by judge of question 4: Is the form (e.g. experiment, general empirical study, data mining, case study, survey, simulation etc.) that was used to evaluate the technology made clear? Can take values: Yes/No/Partly

Completness5 Assessment by judge of question 5: Is there a description of how the evaluation process was organised? Can take values: Yes/No/Partly

Completness6 Assessment by judge of question 6: Are the results of the evaluation clearly described? Can take values: Yes/No/Partly

Completness7 Assessment by judge of question 7: Are any limitations of the study reported?: Yes/No/Partly

Completness8 Assessment by judge of question 8: Are any ideas for future research presented?: Yes/No/Partly

Clarity Assessment by judge of question regarding the overall understandability of the abstract: Please give an assessment of the clarity of this abstract by circling a number on the scale of 1-10 below, where a value of 1 represents Very Obscure and 10 represents Extremely Clearly Written.

Completness1NumValue A numerical value for completeness question 1 where 0=No, Partly=0.5, yes =1

Completness2NumValue A numerical value for completeness question 2 where 0=No, Partly=0.5, yes =1, NA means not applicable

Completness3NumValue A numerical value for completeness question 3 where 0=No, Partly=0.5, yes =1, NA means not applicable or not answered

Completness4NumValue A numerical value for completeness question 4 where 0=No, Partly=0.5, yes =1, NA means not applicable

Completness5NumValue A numerical value for completeness question 5 where 0=No, Partly=0.5, yes =1, NA means not applicable

Completness6NumValue A numerical value for completeness question 6 where 0=No, Partly=0.5, yes =1, NA means not applicable

Completness7NumValue A numerical value for completeness question 7 where 0=No, Partly=0.5, yes =1, NA means not applicable

Completness8NumValue A numerical value for completeness question 8 where 0=No, Partly=0.5, yes =1, NA means not applicable

Sum The sum of the numerical completeness questions excluding those labelled NA

TotalQuestions The count of the number of question related to completeness excluding questions considered not applicable

Completeness Sum/TotalQuestions

Details

Data set collected at Wroclaw University of Technology (POLAND) by Lech Madeyski includes separate entries for each abstract assessed by a judge, that is 4 entries for each judge. Data collected from 16 subjects recruited from Wroclaw University of Technology who were each asked to assess 4 abstracts.

Note Only completeness question 2 was expected to be context dependent and have a NA (not applicable) answer, if other completeness answers were left blank, BAK coded the answer as NA

polishsubjects.txt

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.PolishSubjects
```

KitchenhamMadeyskiBudgen16.SubjectData

KitchenhamMadeyskiBudgen16.SubjectData

Description

If you use this data set please cite this R package and the following paper when accepted: Barbara Kitchenham, Lech Madeyski, David Budgen, Jacky Keung, Pearl Brereton, Stuart Charters, Shirley Gibbs, and Amnart Pohthong, "Robust Statistical Methods for Empirical Software Engineering", Empirical Software Engineering, vol. 22, no. 2, pp. 579–630, 2017. DOI: 10.1007/s10664-016-9437-5 (http://dx.doi.org/10.1007/s10664-016-9437-5), URL: http://madeyski.e-informatyka.pl/download/KitchenhamMade

Usage

```
KitchenhamMadeyskiBudgen16.SubjectData
```

Format

A data frame with variables:

Judge Alphanumeric identifier for each judge

Institution Numerical value identifying each site from which data was collected

JudgeID Numerical value idetifying each judge

Age Age of the judge in years

Eng1st Whether the judge's first langauage was Enlish: Yes/No

YearsStudy The number of years have student been studying computing at University: 1, 2, 3, 4

AbstractsRead Number of abstracts the judge had read prior to the study" 0, 1 to 10, 10+

AbstractsWritten Whether the judge had ever written an abstract for a scientific report/article

AbstractID Alphanumeric identifier for an abstract. The first character identifies the journal, I=IST, J=JSS, the third digit identifies the time period as 1 or 2, the remaining digits identify the abstract number within the set of abstracts found for the specified journal and time period

Treat The initial 3 characters of AbstractID

TreatID A numeric identifier for the journal and time period, 1=IB1, 2=IB2, 3=JB1, 4=JB2

Order The order in which the judge should have viewed the specified abstract

Completness1NumValue The numeric answer to completeness question 1

Completness2NumValue The numeric answer to completeness question 2

Completness3NumValue The numeric answer to completeness question 3

Completness4NumValue The numeric answer to completeness question 4

Completness5NumValue The numeric answer to completeness question 5

Completness6NumValue The numeric answer to completeness question 6

Completness7NumValue The numeric answer to completeness question 7

Completness8NumValue The numeric answer to completeness question 8

Clarity The response to the clarity question or NA if not answered

NumberOfAnsweredCompletnessQuestions The number of completeness questions excluding those with NA

TotalScore Sum of the numeric values of the 8 completeness questions

MeanScore Sum of the completeness questions 1 to 8 divided by TotalScore

Site The name of the site which provided the data. HongKong refers to the Polytechnic University, HongKong.2 refers to the City University

Details

Data set collected from 16 judges assessing 4 abstracts at 6 sites: Lincoln University NZ=1, Hong Kong Polytechnic University=2, PSu Thailand=3, Durham=4, Keele=5, Hong Kong City University=6

subjectdata.txt: Judge Institution JudgeID age eng1st years.study abs.read Absid Treat TreatID Order Com.1 Com.2 Com.3 Com.4 Com.5 Com.6 Com.7 Com.8 Clarity num.questions total.score av.score Site

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
KitchenhamMadeyskiBudgen16.SubjectData
```

`Madeyski15EISEJ.OpenProjects`

Madeyski15EISEJ.OpenProjects data

Description

If you use this data set please cite: Marian Jureczko and Lech Madeyski, "Cross-project defect prediction with respect to code ownership model: An empirical study", e-Informatica Software Engineering Journal, vol. 9, no. 1, pp. 21-35, 2015. DOI: 10.5277/e-Inf150102 (http://dx.doi.org/10.5277/e-Inf150102) URL: http://madeyski.e-informatyka.pl/download/JureczkoMadeyski15.pdf)

Usage

```
Madeyski15EISEJ.OpenProjects
```

Format

A data frame with variables:

PROP The percentage of classes of proprietary (i.e., industrial) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on open source projects.

NOTOPEN The percentage of classes of projects which are not open source projects that must be tested in order to find 80% of defects in case of software defect prediction models built on open source projects.

STUD The percentage of classes of student (i.e., academic) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on open source projects.

OPEN The percentage of classes of open source projects that must be tested in order to find 80% of defects in case of software defect prediction models built on open source projects.

Details

This paper presents an analysis of 84 versions of industrial, open-source and academic projects. We have empirically evaluated whether those project types constitute separate classes of projects with regard to defect prediction. The predictions obtained from the models trained on the data from the open source projects were compared with the predictions from the other models (built on proprietary, i.e. industrial, student, open source, and not open source projects).

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
Madeyski15EISEJ.OpenProjects
```

`Madeyski15EISEJ.PropProjects`

Madeyski15EISEJ.PropProjects data

Description

If you use this data set please cite: Marian Jureczko and Lech Madeyski, "Cross-project defect prediction with respect to code ownership model: An empirical study", e-Informatica Software Engineering Journal, vol. 9, no. 1, pp. 21-35, 2015. DOI: 10.5277/e-Inf150102 (http://dx.doi.org/10.5277/e-Inf150102) URL: http://madeyski.e-informatyka.pl/download/JureczkoMadeyski15.pdf)

Usage

```
Madeyski15EISEJ.PropProjects
```

Format

A data frame with variables:

NOTPROP The percentage of classes of non-proprietary (i.e., non-industrial) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on proprietary (i.e., industrial) projects.

OPEN The percentage of classes of open source projects that must be tested in order to find 80% of defects in case of software defect prediction models built on proprietary (i.e., industrial) projects.

STUD The percentage of classes of student (i.e., academic) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on proprietary (i.e., industrial) projects.

PROP The percentage of classes of proprietary (i.e., industrial) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on proprietary (i.e., industrial) projects.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
Madeyski15EISEJ.PropProjects
```

Madeyski15EISEJ.StudProjects

Madeyski15EISEJ.StudProjects data

Description

If you use this data set please cite: Marian Jureczko and Lech Madeyski, "Cross-project defect prediction with respect to code ownership model: An empirical study", e-Informatica Software Engineering Journal, vol. 9, no. 1, pp. 21-35, 2015. DOI: 10.5277/e-Inf150102 (http://dx.doi.org/10.5277/e-Inf150102) URL: http://madeyski.e-informatyka.pl/download/JureczkoMadeyski15.pdf)

Usage

```
Madeyski15EISEJ.StudProjects
```

Format

A data frame with variables:

PROP The percentage of classes of proprietary (i.e., industrial) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on student (i.e., academic) projects.

NOTSTUD The percentage of classes of projects which are not student projects that must be tested in order to find 80% of defects in case of software defect prediction models built on student (i.e., academic) projects.

STUD The percentage of classes of student (i.e., academic) projects that must be tested in order to find 80% of defects in case of software defect prediction models built on student (i.e., academic) projects.

OPEN The percentage of classes of open source projects that must be tested in order to find 80% of defects in case of software defect prediction models built on student (i.e., academic) projects.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
Madeyski15EISEJ.StudProjects
```

`Madeyski15SQJ.NDC` *Madeyski15SQJ.NDC data*

Description

If you use this data set please cite: Lech Madeyski and Marian Jureczko, "Which Process Metrics Can Significantly Improve Defect Prediction Models? An Empirical Study," Software Quality Journal, vol. 23, no. 3, pp.393-422, 2015. DOI: 10.1007/s11219-014-9241-7

Usage

```
Madeyski15SQJ.NDC
```

Format

A data frame with variables:

Project In case of open source projects this field includes the name of the project as well as its version. In case of industrial projects this field includes the string "proprietary" (we were not allowed to disclose the names of the analyzed industrial software projects developed by Capgemini Polska).

simple The percentage of classes that must be tested in order to find 80% of defects in case of simple defect prediction models, i.e., using only software product metrics as predictors.

advanced The percentage of classes that must be tested in order to find 80% of defects in case of advanced defect prediction models, using not only software product metrics but also the NDC (Number of distinct committers) process metric.

Details

"This paper presents an empirical evaluation in which several process metrics were investigated in order to identify the ones which significantly improve the defect prediction models based on product metrics. Data from a wide range of software projects (both, industrial and open source) were collected. The predictions of the models that use only product metrics (simple models) were compared with the predictions of the models which used product metrics, as well as one of the process metrics under scrutiny (advanced models). To decide whether the improvements were significant or not, statistical tests were performed and effect sizes were calculated. The advanced defect prediction models trained on a data set containing product metrics and additionally Number of Distinct Committers (NDC) were significantly better than the simple models without NDC, while the effect size was medium and the probability of superiority (PS) of the advanced models over simple ones was high (p=.016, r=-.29, PS=.76), which is a substantial finding useful in defect prediction. A similar result with slightly smaller PS was achieved by the advanced models trained on a data set containing product metrics and additionally all of the investigated process metrics (p=.038, r=-.29, PS=.68). The advanced models trained on a data set containing product metrics and additionally Number of Modified Lines (NML) were significantly better than the simple models without NML, but the effect size was small (p=.038, r=.06). Hence, it is reasonable to recommend the NDC process metric in building the defect prediction models." [http://dx.doi.org/10.1007/s11219-014-9241-7]

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
Madeyski15SQJ.NDC
```

`MadeyskiKitchenham.EUBASdata`

MadeyskiKitchenham.EUBASdata data

Description

If you use this data set please cite this R package and the paper where we analyze the data set: Lech Madeyski and Barbara Kitchenham, "Effect Sizes and their Variance for AB/BA Crossover Design Studies", Empirical Software Engineering, vol. 24, no.4, p. 1982-2017, 2018. DOI: 10.1007/s10664-017-9574-5

Usage

```
MadeyskiKitchenham.EUBASdata
```

Format

A data frame with variables:

ID Project ID

TimePeriod Period of time (run): R1, R2

SequenceGroup Sequence group: G1, G2, G3, G4

System Software system identifier indicates the system (i.e., S1 or S2) used as the experimental object: S1. A software system to sell and manage CDs/DVDs in a music shop, S2. A software system to book and buy theater tickets

Technique The independent variable. It is a nominal variable that can assume the following two values: AM (analysis models plus source code) and SC (source code alone)

Comp_Level This denotes the comprehension level of the source code achieved by a software engineer

Modi_Level This denotes the capability of a maintainer to modify source code

Details

Data set comes from an experiment conducted in Italy at the University of Basilicata (with 24 first-year students from the Master's Program in Computer Science) to answer the question "Do the software models produced in the requirements analysis process aid in the comprehensibility and modifiability of source code?", see G. Scanniello, C. Gravino, M. Genero, J. A. Cruz-Lemus, and G. Tortora, "On the Impact of UML Analysis Models on Source-code Comprehensibility and Modifiability," ACM Transactions on Software Engineering and Methodology, vol. 23, pp. 13:1-13:26, Apr. 2014. However, the inconsistent subject data for subject 2 was removed, see the aforementioned paper by Madeyski and Kitchenham.

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
MadeyskiKitchenham.EUBASdata
```

`MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR`

MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR data form a set of primary studies on reading methods for software inspections. They were analysed by Lech Madeyski and Barbara Kitchenham, "How variations in experimental designs impact the construction of comparable effect sizes for meta-analysis", 2015.

Description

If you use this data set please cite: Lech Madeyski and Barbara Kitchenham, "How variations in experimental designs impact the construction of comparable effect sizes for meta-analysis", 2015.

Usage

```
MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR
```

Format

A data frame with 17 rows and 26 variables:

Study Name of empirical study

Ref. Reference to the paper reporting primary study or experimental run where data were originally reported

Teams The number of teams including both, PBR and Control teams

DesignDesc Experimental design description: Before-after, Between-groups, Cross-over

ExpDesign Experimental design: between-groups (BG), within-subjects cross-over (WSCO), within-subjects before-after (WSBA)

M_PBR The average proportion of defects found by teams using PBR

M_C The average proportion of defects found by teams using Control treatment: Check-Based Reading (CBR) or Ad-Hoc Reading (AR)

Diff The difference between M_PBR and M_C, i.e. Diff = M_PBR - M_C

Inc The percentage increase in defect rate detection, i.e. Inc=100*[(M_PBR-M_C)/M_C]

SD_C_ByAuthors The standard deviation of the control group values reported by the original Authors, i.e., obtained from the papers/raw data

SD_C The standard deviation of the control group values equals SD_C_ByAuthors for studies for which the data was available OR the weighted average of SD_C_ByAuthors (i.e., 0.169) for studies where SD_C_ByAuthors is missing.

V_C The variance of the Control group observations, i.e., the variance obtained from the teams using the Control method V_C=SD_C^2

V_D The variance of the unstandardized mean difference D (between the mean value for the treatment group and the mean value for the Control group)

SD_C_Alt This is the equivalent of SD_C (the standard deviation of the control group) based on a different variance for the student studies or the practitioner studies depending on the subject type of the study with the missing value.

V_Alt The variance of the mean difference in the meta-analysis based on SD_C_Alt

SS_C The sum of squares of the Control group values. For within subjects studies SS=V_C*(n-1). For between subjects studies SS=V_C*(n_C-1)

n_PBR The number of PBR teams

n_C The number of Control (CBR or AR) teams

ControlType Type of Control treatment: CRB or AR

ParticipantsType Type of participants: Engineers or Students

TeamType Type of team: Nominal or Real

TwoPersonTeamVsLargerTeam Reflects size of the teams: 2-PersonTeam or LargerTeam

ArtefactType The type of artefact: Requirements or Other

AssociatedWithBasili Whether study is associated with Basili (the forerunner): Yes or No

ControlType_Basili Combined ControlType and AssociatedWithBasili: AH_AssociatedWithBasili, CBR_AssociatedWithBasili, CBR_NotAssociatedWithBasili

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR
```

`MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324`
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324 data

---

Description

This data is used in the paper: Tomasz Lewowski and Lech Madeyski, "Creating evolving project data sets in software engineering", 2019 (submitted). This file holds the descriptive data for each document and each experiment for studies 3, 7 and 11 which include the mean, standard deviation and sample size for the control and treatment techniques. These studies performed ABBA crossover experiments and reported data for each document separately. Note Study 3 also undertook an independent groups study but data from that experiment is held in the ExpData file.

Usage

```
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324
```

Format

A text file file with variables:

rowID unique id assigned to projects before filtering (source: API)

id GitHub repository ID (source: API)

repository owner the organization or user owning the repository (source: API)

project name name of the project (source: API)

manual link to best found project documentation - wiki, webpage, documentation directory or readme. Projects with limited documentation were marked with (limited) and ones that had documentation in Chinese - (Chinese) (source: manual)

installation the recommended installation medium(s) for the project. Some mediums may be missing for projects with multiple recommendations. (source: manual)

support channel(s) that can be used to get support and/or report bugs. Some channels may be missing for projects with multiple ones. Abbreviations used (source: manual): GH GitHub Issues SO Stack Overflow GG Google Groups ML Mailing list FB Facebook MM Mattermost LI LinkedIn ? not found

is not sample/playground/docs/... 1 if the project is an actual application or library, 0 if it is a set of samples, only documentation or some experimental area (source: manual)

is industrial whether the project can be treated as industrial quality one. Values and their meanings: 1 the repository can be classified as industrial grade; 0,5 the repository can sometimes be classified as industrial grade, but it is either a minor project or its documentation or support may be lacking the depth; 0 the repository cannot be classified as industrial-grade; -1 the repository is no longer actively maintained as of the date of data acquisition; -2 the repository is no longer in Java as of the date of data acquisition. (source: manual)

createdAt the date at which the repository was created (source: API)

updatedAt the date of last repository update - including changes in projects, watchers, issues etc. (source: API)

pushedAt the date of last push to the repository - NOT the date of last pushed commit (source: API)

diskUsage total number of bytes on disk that are needed to store the repository (source: API)

forkCount number of existing repository forks (independent copies managed by other entities) (source: API)

isArchived true if the repository is archived (no longer maintained), false otherwise (source: API)

isFork true if the repository is a fork (not the main repository), false otherwise (source: API)

isMirror true if the repository is a mirror, false otherwise (source: API)

sshUrlOfRepository URL that can be used to immediately clone the repository (source: API)

licenseInfo.name name of license under which the project is distributed. Names are the same as in https://choosealicense.com/appendix/ (source: API)

commitSHA unique Git identifier of commit that was top of the main branch at the time of data acquisition (source: API)

defaultBranchRef.target.history.totalCount number of commits on the default branch in the repository (usually master) at the time of data acquisition (source: API)

stargazers.totalCount number of stargazers for the repository at the time of data acquisition (source: API)

watchers.totalCount number of watchers for the repository at the time of data acquisition (source: API)

languages.totalSize total size of all source code files (source: API)

Java.byte.count total size of Java files (source: API)

Language main programming language used in the repository, i.e. one that the most code is written in (source: API)

searchQuery query used during search that obtained this project (source: API)

Source

http://madeyski.e-informatyka.pl/reproducible-research/

Examples

```
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324
```

`percentageInaccuracyOfLargeSampleVarianceApproximation` *percentageInaccuracyOfLargeSampleVarianceApproximation*

Description

Plot the extent of inaccuracy using the large sample approximate effect size variance on 4 related graphs corresponding to the four different correlation values. Plot visualizes the relationship between sample size and effect size and the percentage inaccuracy of the large sample variance approximation. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
percentageInaccuracyOfLargeSampleVarianceApproximation(data)
```

Arguments

`data`	- data behind the plot returned by getSimulatedCrossoverDataSets() or stored in reproducer::KitchenhamMadeyski.SimulatedCrossoverDataSets

Value

plot described in description

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
data <- KitchenhamMadeyski.SimulatedCrossoverDataSets
myPlot <- percentageInaccuracyOfLargeSampleVarianceApproximation(data)
```

plotOutcomesForIndividualsInEachSequenceGroup
plotOutcomesForIndividualsInEachSequenceGroup

Description

Function to plot a figure on the outcomes for individuals in each sequence group used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham

Usage

```
plotOutcomesForIndividualsInEachSequenceGroup(var, covar, meanA1,
  treatmentDiff, periodEffect, numOfSamples)
```

Arguments

var	Variance among subjects is a sum of the between subjects variance and the within subjects variance
covar	Covariance equal to the between subjects variance
meanA1	Mean for treatment sequence A1
treatmentDiff	technique effect which is the difference between the effect of technique A and technique B
periodEffect	Period effect which is the difference between period 1 and period 2
numOfSamples	Number of samples ("rows" of data) required for each technique and period

Value

plot

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
myPlot<-plotOutcomesForIndividualsInEachSequenceGroup(25, 18.75, 50, 10, 5, 15)
```

PrepareForMetaAnalysisGtoR *PrepareForMetaAnalysisGtoR*

Description

This function calculates the standardized effect sizes and their confidence intervals, the equivalence point biserial effect size and the Zr and var(Zr) needed for input into the metafor rma function (meta analysis). In this function the point bi-serial effect size is based on the adjusted Hedges g value. The function uses the Hedges g to r transformation to prepare for meta-analysing the data where the mean values, the standard deviations, and the number of observations are available.

Usage

```
PrepareForMetaAnalysisGtoR(Mc, Mt, SDc, SDt, Nc, Nt)
```

Arguments

Mc	is a vector containing the mean value of the control group for each experiment.
Mt	is a vector containing the mean value of the treatment group for each experiment.
SDc	is a vector of the standard deviations of the control group for each experiment.
SDt	is a vector of the standard deviations of the the treatment group for each experiment.
Nc	is a vector containing the the number of observations (particpants) in the control group for each experiment.
Nt	is a vector of the number of observations (participants) in the treatment group for each experiment.

Value

data frame incl. calculated effect sizes (Hedges' g, Hedges' g adjusted), upper and lower confidence bounds on Hedges' g, zr, vi - variance of zr, r and pvalue

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
PrepareForMetaAnalysisGtoR(c(10,10), c(12,14), c(4,4), c(4,4), c(20,20), c(40,40))
#HGvalues.Hg HGvalues.HgAdjusted  Hgupper    Hglower      zr       vi       r     pvalue
#      0.5         0.4935018 1.082017 -0.06156572 0.2305901 0.01754386 0.2265882 0.0816981743
#      1.0         0.9870036 1.634701  0.40620071 0.4499419 0.01754386 0.4218513 0.0006813222
```

printXTable *printXTable*

Description

print data table using xtable R package

Usage

```
printXTable(data, selectedColumns, tableType = "latex", alignCells,
  digits, caption, label, fontSize, captionPlacement = "bottom",
  alignHeader)
```

Arguments

data	Data structure including columns to be printed.
selectedColumns	Columns selected to be printed.
tableType	Type of table to produce. Possible values are "latex" or "html". Default value is "latex".
alignCells	Defines how to align data cells.
digits	Defines the number of decimal points in each column.
caption	Caption of the table.
label	Label of the table.
fontSize	Size of the font used to produce a table.
captionPlacement	The caption will be have placed at the bottom of the table if captionPlacement is "bottom" and at the top of the table if it equals "top". Default value is "bottom".
alignHeader	Defines how to align column headers of a table.

Value

A table generated on the fly on a basis of passed data (data, selectedColumns etc.).

Author(s)

Lech Madeyski

Examples

```
d <- reproducer::MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR
reproducer::printXTable(d, '"Study"', "latex", '"cc"', 0, "C", "L", "tiny", "top", "l")
```

proportionOfSignificantTValuesUsingCorrectAnalysis *proportionOfSignificantTValuesUsingCorrectAnalysis*

Description

Plots visualize the relationship between sample size, effect size and the proportion of significant t-values using the correct analysis. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
proportionOfSignificantTValuesUsingCorrectAnalysis(data)
```

Arguments

data	- data behind the plot returned by getSimulatedCrossoverDataSets() or stored in reproducer::KitchenhamMadeyski.SimulatedCrossoverDataSets

Value

plot described in description

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
data <- KitchenhamMadeyski.SimulatedCrossoverDataSets
myPlot <- proportionOfSignificantTValuesUsingCorrectAnalysis(data)
```

proportionOfSignificantTValuesUsingIncorrectAnalysis *proportionOfSignificantTValuesUsingIncorrectAnalysis*

Description

Plots visualize the relationship between sample size, effect size and the proportion of significant t-values using the incorrect analysis. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
proportionOfSignificantTValuesUsingIncorrectAnalysis(data)
```

Arguments

data	- data behind the plot returned by getSimulatedCrossoverDataSets() or stored in reproducer::KitchenhamMadeyski.SimulatedCrossoverDataSets

Value

plot described in description

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
data  <- KitchenhamMadeyski.SimulatedCrossoverDataSets
myPlot <- proportionOfSignificantTValuesUsingIncorrectAnalysis(data)
```

---

readExcelSheet *readExcelSheet*

Description

Function reads data from an Excel file from a specified sheet

Usage

```
readExcelSheet(path, sheet, colNames)
```

Arguments

path	Path to an Excel file, e.g. /User/lma/datasets/MyDataSet.xls
sheet	Name of a sheet within an Excel file we want to read
colNames	If TRUE, first row of data will be used as column names.

Author(s)

Lech Madeyski

Examples

```
myPath=system.file("extdata", "DataSet.xlsx", package = "reproducer")
Madeyski15SQJ.NDC<-readExcelSheet(path=myPath, sheet="Madeyski15SQJ.NDC", colNames=TRUE)
```

reproduceForestPlotRandomEffects *reproduceForestPlotRandomEffects()*

Description

Function reproduces Forest Plot of a Random-Effects Meta-analysis of Mean Differences.

Usage

```
reproduceForestPlotRandomEffects()
```

Author(s)

Lech Madeyski

Examples

```
reproduceForestPlotRandomEffects()
```

reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator *reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator()*

Description

Function reproduces Mixed-Effects Analysis using Subject Specific Estimated Variance with Experimental Design as a Moderator.

Usage

```
reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator()
```

Author(s)

Lech Madeyski

Examples

```
reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator()
```

reproduceMixedEffectsAnalysisWithExperimentalDesignModerator *reproduceMixedEffectsAnalysisWithExperimentalDesignModerator()*

Description

Function reproduces Mixed-Effects Analysis with Experimental Design as a Moderator.

Usage

```
reproduceMixedEffectsAnalysisWithExperimentalDesignModerator()
```

Author(s)

Lech Madeyski

Examples

```
reproduceMixedEffectsAnalysisWithExperimentalDesignModerator()
```

reproduceMixedEffectsForestPlotWithExperimentalDesignModerator *reproduceMixedEffectsForestPlotWithExperimentalDesignModerator()*

Description

Function reproduces Forest Plot of a Mixed Effects Meta-analysis of Mean Differences with Experimental Design as a Moderator Variable.

Usage

```
reproduceMixedEffectsForestPlotWithExperimentalDesignModerator()
```

Author(s)

Lech Madeyski

Examples

```
reproduceMixedEffectsForestPlotWithExperimentalDesignModerator()
```

reproduceSimulationResultsBasedOn500Reps1000Obs
reproduceSimulationResultsBasedOn500Reps1000Obs

Description

Function to calculate simulation results based on 500 repetitions of 1000 observation samples. Function is used in a paper "Effect Sizes and their Variance for AB/BA Crossover Design Studies" by Lech Madeyski and Barbara Kitchenham.

Usage

```
reproduceSimulationResultsBasedOn500Reps1000Obs()
```

Value

data frame including the following simulation results: # treatmentEffect.Ave - Average Technique Effect # dRM.Ave - Average dRM # dRM.Var - Variance of dRM # dRM.Var.Ave - Average of var(dRM) # dRM.Var.ModerateSampleSizeApprox - # dIG.Ave - Average dIG # dIG.Var - Variance of dIG # dIG.Var.Ave - Average of var(dIG) # dIG.Var.ModerateSampleSizeApprox -

Author(s)

Lech Madeyski and Barbara Kitchenham

Examples

```
# return simulation results based on 500 repetitions of 1000 observation samples
simulationResultsTable500x1000<-reproduceSimulationResultsBasedOn500Reps1000Obs()
```

---

reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments
reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments

Description

This function reproduces five of the output tables used in the systematic review paper "Meta-analysis for Families of Experiments: A Systematic Review and Reproducibility Assessment". It extracts the reported values for effect sizes, meta-analysis and descriptive statistics in the primary studies. It uses the descriptive statistics to re-calculate effect sizes and then performs a meta-analyses using the constructed effect sizes and compares the calculated values with the reported values.

Usage

```
reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments()
```

Value

list incl. the data presented in five of the tables presented in the paper.

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
rrData = reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments()
# Reproduce Table "Overall Mean Values of Effect Sizes Reported and Calculated":
xtable::xtable(rrData$MAStats)
# Reproduce Table "Calculated and Reported Effect Sizes":
xtable::xtable(rrData$ESdata)
# Report values for 3 papers that reported per document
rrData$MAStatsTP1=data.frame(rrData$MAStatsTP1,row.names=NULL)
rrData$ESTP1res=data.frame(rrData$ESTP1res,row.names=NULL)
xtable::xtable(rrData$MAStatsTP1)
xtable::xtable(rrData$ESTP1res)
# Report extra results for Study 8
# Reproduce Table "Calculating r_PB Effect Size from Probabilities"
xtable::xtable(rrData$GH2015extra)
```

---

reproduceTableWithEffectSizesBasedOnMeanDifferences

reproduceTableWithEffectSizesBasedOnMeanDifferences()

Description

Function reproduces Table, which shows the effect sizes based on mean differences.

Usage

```
reproduceTableWithEffectSizesBasedOnMeanDifferences()
```

Author(s)

Lech Madeyski

Examples

```
reproduceTableWithEffectSizesBasedOnMeanDifferences()
```

reproduceTableWithPossibleModeratingFactors *reproduceTableWithPossibleModeratingFactors()*

Description

Function reproduces Table with possible moderating factors.

Usage

```
reproduceTableWithPossibleModeratingFactors()
```

Author(s)

Lech Madeyski

Examples

```
reproduceTableWithPossibleModeratingFactors()
```

reproduceTableWithSourceDataByCiolkowski *reproduceTableWithSourceDataByCiolkowski*

Description

Function reproduces Table, which shows the effect sizes reported by Ciolkowski identifying the type of design used in each study.

Usage

```
reproduceTableWithSourceDataByCiolkowski()
```

Author(s)

Lech Madeyski

Examples

```
reproduceTableWithSourceDataByCiolkowski()
```

searchForIndustryRelevantGitHubProjects *searchForIndustryRelevantGitHubProjects*

Description

Function searches for industry relevant software projects available from GitHub. The function was used to deliver data set of software projects in an NCBiR project. More details are desribed in a report: Lech Madeyski, "Training data preparation method," tech. rep., code quest (research project NCBiR POIR.01.01.01-00-0792/16), 2019, as well as a paper: Tomasz Lewowski and Lech Madeyski, "Creating evolving project data sets in software engineering", 2019. If you use this function or the returned data set than please cite: Tomasz Lewowski and Lech Madeyski, "Creating evolving project data sets in software engineering", 2019

Usage

```
searchForIndustryRelevantGitHubProjects(myToken)
```

Arguments

myToken	A private token used to access GitHub

Value

selected GitHub projects

Author(s)

Lech Madeyski and Tomasz Lewowski

Examples

```
#to run this function you need to use your own token as a parameter of the function
#calculateSmallSampleSizeAdjustment("...") #use your own token as a parameter of the function
```

transformHgtoR *transformHgtoR*

Description

The functions transforms a vector of Hedges g values to their equivalent point bi-serial values.

Usage

```
transformHgtoR(g, Nc, Nt)
```

Arguments

g	A vector of Hegdes g values.
Nc	A vector of numbers identifying the number of control condition participants in each group
Nt	A vector of numbers identifying the number of treatment condition participants in each group

Value

value of point biserial r

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformHgtoR(0.4, 20, 20)
# [1] 0.1961161
```

---

transformHgtoZr *transformHgtoZr*

Description

The functions transforms a vector of Hedges g values to their normal approximation of point biserial values.

Usage

```
transformHgtoZr(g, Nc, Nt)
```

Arguments

g	value of Hedges' g
Nc	the number of observations (particpants) in the first (control) group
Nt	the number of observations (participants) in the second (treatment) group

Value

value of normal approximation of point biserial r

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformHgtoZr(0.5, 20, 20)
# [1] 0.2474665
```

`transformRtoHg` *transformRtoHg*

Description

This function coverts a vector of point bi-serial r values with associated sample size information back to the mean difference effect size Hedges g.

Usage

```
transformRtoHg(r, Nc, Nt)
```

Arguments

`r`	A vector of point bi-serial correlation values.
`Nc`	A vector of the number of observations in the control condition for the related experiments.
`Nt`	A vector of the number of observations in the treatment condition for the related experiments.

Value

value of Hedges' g

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformRtoHg(c(0.4,0.2), c(20,20), c(20,20))
# [1] 0.8728716 0.4082483
```

transformRtoZr *transformRtoZr*

Description

The function transforms a vector of point biserial r values to their normal approximation. It also works for the correlation r.

Usage

```
transformRtoZr(r)
```

Arguments

r	A vector of r-values

Value

value of normal approximation of point biserial r

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformRtoZr(0.4)
# [1] 0.4236489
Zr=transformRtoZr(c(0.4,0.2))
Zr
# [1] 0.4236489 0.2027326
```

transformZrtoHg *transformZrtoHg*

Description

Transforms Zr to Hedge's g.

Usage

```
transformZrtoHg(Zr, Nc, Nt)
```

Arguments

Zr	the normal variate
Nc	the number of observations (particpants) in the first (control) group
Nt	the number of observations (participants) in the second (treatment) group

Value

value of Hedges' g

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformZrtoHg(0.5, 20, 20)
#[1] 1.042191
```

---

transformZrtoHgapprox *transformZrtoHgapprox*

Description

This function provides an approximate transformation from Zr to Hedges g when the number of observations in the treatment and control group are unknown. It is also used to allow the forest plots to display Hedge's g when they are based on r. It is necessary because the transformation function in the forest plot function does not allow any parameters other than effect size used. The function assumes that Nc=Nt and gives the same results as transformZrtoHg when Nc=Nt.

Usage

```
transformZrtoHgapprox(Zr)
```

Arguments

Zr	A vector of normalised point bi-serial values

Value

approx. value of Hedges' g

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformZrtoHgapprox(c(0.4,0.2))
# [1] 0.8215047 0.4026720
```

transformZrtoR *transformZrtoR*

Description

The function transforms a vector of standardized normal variates to their equivalent r-values.

Usage

```
transformZrtoR(zr)
```

Arguments

zr	A vector of standard normal variates.

Value

value of point biserial r

Author(s)

Barbara Kitchenham and Lech Madeyski

Examples

```
transformZrtoR(0.4236489)
# [1] 0.4
transformZrtoR(c(0.4236489, 0.2027326))
# [1] 0.4 0.2
```

Index

∗Topic **datasets**
Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR, 8
KitchenhamMadeyski.SimulatedCrossoverDataSets, 18
KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults, 19
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes, 20
KitchenhamMadeyskiBrereton.DocData, 21
KitchenhamMadeyskiBrereton.ExpData, 22
KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults, 23
KitchenhamMadeyskiBrereton.ReportedEffectSizes, 24
KitchenhamMadeyskiBudgen16.COCOMO, 25
KitchenhamMadeyskiBudgen16.DiffInDiffData, 27
KitchenhamMadeyskiBudgen16.FINNISH, 28
KitchenhamMadeyskiBudgen16.PolishData, 29
KitchenhamMadeyskiBudgen16.PolishSubjects, 31
KitchenhamMadeyskiBudgen16.SubjectData, 33
Madeyski15EISEJ.OpenProjects, 34
Madeyski15EISEJ.PropProjects, 35
Madeyski15EISEJ.StudProjects, 36
Madeyski15SQJ.NDC, 37
MadeyskiKitchenham.EUBASdata, 38
MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR, 39
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects2019v324, 41

aggregateIndividualDocumentStatistics, 3
boxplotAndDensityCurveOnHistogram, 4
boxplotHV, 5

calculateHg, 6
calculateHgResults, 6
calculateSmallSampleSizeAdjustment, 7
Ciolkowski09ESEM.MetaAnalysis.PBRvsCBRorAR, 8
constructEffectSizes, 9

densityCurveOnHistogram, 10

effectSizeCI, 11
ExtractMAStatistics, 12

rint, 13

getEffectSizesABBA, 13
getEffectSizesABBAIgnoringPeriodEffect, 15
getSimulationData, 16
getTheoreticalEffectSizeVariancesABBA, 17

KitchenhamMadeyski.SimulatedCrossoverDataSets, 18
KitchenhamMadeyskiBrereton.ABBAMetaAnalysisReportedResults, 19
KitchenhamMadeyskiBrereton.ABBAReportedEffectSizes, 20
KitchenhamMadeyskiBrereton.DocData, 21
KitchenhamMadeyskiBrereton.ExpData, 22
KitchenhamMadeyskiBrereton.MetaAnalysisReportedResults, 23
KitchenhamMadeyskiBrereton.ReportedEffectSizes, 24
KitchenhamMadeyskiBudgen16.COCOMO, 25
KitchenhamMadeyskiBudgen16.DiffInDiffData, 27
KitchenhamMadeyskiBudgen16.FINNISH, 28

KitchenhamMadeyskiBudgen16.PolishData, 29
KitchenhamMadeyskiBudgen16.PolishSubjects, 31
KitchenhamMadeyskiBudgen16.SubjectData, 33

Madeyski15EISEJ.OpenProjects, 34
Madeyski15EISEJ.PropProjects, 35
Madeyski15EISEJ.StudProjects, 36
Madeyski15SQJ.NDC, 37
MadeyskiKitchenham.EUBASdata, 38
MadeyskiKitchenham.MetaAnalysis.PBRvsCBRorAR, 39
MadeyskiLewowski.IndustryRelevantGitHubJavaProjects20190324, 41

percentageInaccuracyOfLargeSampleVarianceApproximation, 43
plotOutcomesForIndividualsInEachSequenceGroup, 44
PrepareForMetaAnalysisGtoR, 45
printXTable, 46
proportionOfSignificantTValuesUsingCorrectAnalysis, 47
proportionOfSignificantTValuesUsingIncorrectAnalysis, 47

readExcelSheet, 48
reproduceForestPlotRandomEffects, 49
reproduceMixedEffectsAnalysisWithEstimatedVarianceAndExperimentalDesignModerator, 49
reproduceMixedEffectsAnalysisWithExperimentalDesignModerator, 50
reproduceMixedEffectsForestPlotWithExperimentalDesignModerator, 50
reproduceSimulationResultsBasedOn500Reps1000Obs, 51
reproduceTablesOfPaperMetaAnalysisForFamiliesOfExperiments, 51
reproduceTableWithEffectSizesBasedOnMeanDifferences, 52
reproduceTableWithPossibleModeratingFactors, 53
reproduceTableWithSourceDataByCiolkowski, 53

searchForIndustryRelevantGitHubProjects, 54

transformHgtoR, 54
transformHgtoZr, 55
transformRtoHg, 56
transformRtoZr, 57
transformZrtoHg, 57
transformZrtoHgapprox, 58
transformZrtoR, 59