**Instructions**

**Literature review:**

**This includes several sub-topics that have to be explained. Use APA. Therefore you need the paper information to know what to do and the R code). Experience with missing data and R are needed.**

**Validity of ANOVA:**

There is another approach for assessing the validity of ANOVAs. We usually interpret the F-test by looking at the significance level of the test. This significance level is determined by the p-value, which reflects the probability of finding a less likely result under the null hypothesis - i.e. when there is no effect. As opposed to the F-test itself, its corresponding p-value behaves consistently under the null hypothesis. More specifically, for any ANOVA based on any given data set under the null hypothesis, the corresponding p-value has a uniform distribution, an expected value of .50 and requires the p-value to be confidence valid (a p-value of .05 should only occur 5 percent of the time).

**Explain how to examine the following about the p-value:**

- Has an expected value of .50 (theory + see the code to tell how I evaluated this point in my research; almost at the end of the code)

- Should be confidence valid (Neyman, 1934) in the sense that a p-value of .05 should only occur 5 percent of the time (theory and how did I test it in my research)

**Interpretation of the results: use the script on R (some of the results are included)**

**- Expected value of the p-values:** text with expected values and table + interpretation results

- **Proportion E(p<=.05):** interpretation result + table

- **Proportion correct** (i.e. how many p-values less than .05 are the same as the population)

-Summary of the results

**The main question is to discover which statistics (D1, D2, D3..) is valid to use when working with missing data of different missingness (1%, 25%, 50%, 75%)**

Those results (output) can be obtained by using the R script. I added some in the info document.

**Explain simulation:**

* Description of the structure of the simulation study (code r script) and reasons of the stepd (basically the multiple imputation steps and then testing the validity steps)
* Description of the analysis strategy and methods used

**Discussion and conclusion:**

* Which statistic is valid to use when working with missing data, and on which proportion of missingness + why
* Discussion points
* Implications of the results for theory formation
* Limitations of the current research

**Other questions that I have:**

* Why compare intercept only model with a model with predictors in the simulation
* How can I test the reliability?

**Request**:

* Add information when needed (see the research in this document)

The script:

library(mice)

library(miceadds)

library(magrittr)

library(dplyr)

library(purrr)

library(mvtnorm)

library(ggplot2)

#install.packages("ggplot2")

set.seed(123)

# Simulation parameters

nsim = 1000

rho = 0 # correlation set to zero to conform to null hypothesis (no effect)

# Generate data function

make.data <- function(n, correlation){

data <- rmvnorm(n = n, mean = c(0, 0),

sigma = matrix(c(1, correlation, correlation, 1),

nrow = 2, ncol = 2))

colnames(data) <- c("y", "x")

data %>% as\_tibble() %>% return()

}

# simulation function

simulate <- function(n, rho, prop.mis = .75){

# sample data from multivariate normal distribution

data <- make.data(n = n, correlation = rho)

# ampute data

missing <- ampute(data,

patterns = matrix(c(0, 1, 1, 0), ncol = 2, nrow = 2, byrow = TRUE),

prop = prop.mis,

mech = "MCAR")

# impute data

imp <- mice(missing$amp, method = "norm", m = 10, maxit = 10, print = FALSE)

return(list(data = data,

miss = missing,

imp = imp))

}

# run simulation

SIM01 <- replicate(nsim, simulate(n = 1000, rho = rho, prop.mis = .01), simplify = FALSE)

SIM25 <- replicate(nsim, simulate(n = 1000, rho = rho, prop.mis = .25), simplify = FALSE)

SIM50 <- replicate(nsim, simulate(n = 1000, rho = rho, prop.mis = .50), simplify = FALSE)

SIM75 <- replicate(nsim, simulate(n = 1000, rho = rho, prop.mis = .75), simplify = FALSE)

SIM75[[2]]

#evaluate

# Read in the simulation data

# evaluation function

evaluate <- function(sim){

# calculate true data statistics

fit <- sim$data %$% lm(y ~ x) #performe a lineaire fit between x and y

anova <- fit %>% anova

# calculate observed data statistics

fit.mis <- sim$miss$amp %$% lm(y ~ x)

anova.mis <- fit.mis %>% anova

# calculate imputed data statistics

fit.imp <- with(data = sim$imp, expr = lm(y ~ x))

fit.imp.empty <- with(data = sim$imp, expr = lm(y ~ 1))

fit.imp.pool <- pool(fit.imp)

fit.anova.D1 <- D1(fit.imp, fit.imp.empty)

fit.anova.D2 <- D2(fit.imp, fit.imp.empty) #cf. micombine.F from miceadds

fit.anova.D3 <- D3(fit.imp, fit.imp.empty)

F.vector <- unlist(with(sim$imp, anova(lm(y ~ x))$'F value'[1])$analyses)

micomb <- micombine.F(F.vector, anova$Df[[1]], display = FALSE)

avg.F.imp <- mean(F.vector)

avg.p.imp <- mean(unlist(with(sim$imp, anova(lm(y ~ x))$'Pr(>F)'[1])$analyses))

# prepare output for return

return(list(truefit = fit,

trueanova = anova,

missfit = fit.mis,

missanova = anova.mis,

fit = fit.imp,

emptyfit = fit.imp.empty,

pool = fit.imp.pool,

D1 = fit.anova.D1,

D2 = fit.anova.D2,

D3 = fit.anova.D3,

micomb = round(micomb, 4),

Fvector = F.vector,

Fbar = avg.F.imp,

pbar = avg.p.imp))

}

EVAL01 <- map(SIM01, evaluate)

EVAL25 <- map(SIM25, evaluate)

EVAL50 <- map(SIM50, evaluate)

EVAL75 <- map(SIM75, evaluate)

# Grab Anova's

grab.F <- function(x){

data.frame(true = x$trueanova$`F value`[1],

mis = x$missanova$`F value`[1],

D1 = x$D1$result[1],

D2 = x$D2$result[1],

D3 = x$D3$result[1],

micombine = x$micomb[1],

Fbar = x$Fbar)

}

F.out01 <- mapply(grab.F, EVAL01, SIMPLIFY = FALSE) %>% do.call(rbind, .)

F.out25 <- mapply(grab.F, EVAL25, SIMPLIFY = FALSE) %>% do.call(rbind, .)

F.out50 <- mapply(grab.F, EVAL50, SIMPLIFY = FALSE) %>% do.call(rbind, .)

F.out75 <- mapply(grab.F, EVAL75, SIMPLIFY = FALSE) %>% do.call(rbind, .)

# Grab p-values

grab.p <- function(x){

data.frame(true = x$trueanova$`Pr(>F)`[1],

mis = x$missanova$`Pr(>F)`[1],

D1 = x$D1$result[4],

D2 = x$D2$result[4],

D3 = x$D3$result[4],

micombine = x$micomb[2],

pbar = x$pbar)

}

p.out01 <- mapply(grab.p, EVAL01, SIMPLIFY = FALSE) %>% do.call(rbind, .)

p.out25 <- mapply(grab.p, EVAL25, SIMPLIFY = FALSE) %>% do.call(rbind, .)

p.out50 <- mapply(grab.p, EVAL50, SIMPLIFY = FALSE) %>% do.call(rbind, .)

p.out75 <- mapply(grab.p, EVAL75, SIMPLIFY = FALSE) %>% do.call(rbind, .)

# Save the evaluations and the resulting p-value and F-value objects

#save(list = c("EVAL01", "EVAL25", "EVAL50", "EVAL75"),

#file = "Evaluations\_objects.RData")

#save(list = c("F.out01", "F.out25", "F.out50", "F.out75",

#"p.out01", "p.out25", "p.out50", "p.out75"),

#file = "Evaluations\_processed.RData")

# Test whether the p-values follow a uniform distribution

apply(p.out01, 2, function(x) ks.test(x, "punif", 0, 1))

apply(p.out25, 2, function(x) ks.test(x, "punif", 0, 1))

apply(p.out50, 2, function(x) ks.test(x, "punif", 0, 1))

apply(p.out75, 2, function(x) ks.test(x, "punif", 0, 1))

# Test how many p=values <.05 (we expect .05)

# apply(p.out25, 2, function(x) mean(x < .05))

data.frame(mis01 = apply(p.out01, 2, function(x) mean(x < .05)),

mis25 = apply(p.out25, 2, function(x) mean(x < .05)),

mis50 = apply(p.out50, 2, function(x) mean(x < .05)),

mis75 = apply(p.out75, 2, function(x) mean(x < .05)))

# What is the mean p-value

# apply(p.out, 2, function(x) mean(x))

data.frame(mis01 = apply(p.out01, 2, function(x) mean(x)),

mis25 = apply(p.out25, 2, function(x) mean(x)),

mis50 = apply(p.out50, 2, function(x) mean(x)),

mis75 = apply(p.out75, 2, function(x) mean(x)))

# Bias in p-value with respect to the population p=value

# population = .5 under null hypothesis

# apply(p.out, 2, function(x) mean(x - .5))

data.frame(mis01 = apply(p.out01, 2, function(x) mean(x - .5)),

mis25 = apply(p.out25, 2, function(x) mean(x - .5)),

mis50 = apply(p.out50, 2, function(x) mean(x - .5)),

mis75 = apply(p.out75, 2, function(x) mean(x - .5)))

# Bias in p-value with respect to the sampled complete p=value

# population = .5 under null hypothesis

# truth <- p.out$true

# apply(p.out, 2, function(x) mean(x - truth))

truth01 <- p.out01$true

truth25 <- p.out25$true

truth50 <- p.out50$true

truth75 <- p.out75$true

data.frame(mis01 = apply(p.out01, 2, function(x) mean(x - truth01)),

mis25 = apply(p.out25, 2, function(x) mean(x - truth25)),

mis50 = apply(p.out50, 2, function(x) mean(x - truth50)),

mis75 = apply(p.out75, 2, function(x) mean(x - truth75)))

# Bias in p-value with respect to the sampled incomplete p=value

# population = .5 under null hypothesis

# mis <- p.out$mis

# apply(p.out, 2, function(x) mean(x - mis))

mis01 <- p.out01$mis

mis25 <- p.out25$mis

mis50 <- p.out50$mis

mis75 <- p.out75$mis

data.frame(mis01 = apply(p.out01, 2, function(x) mean(x - mis01)),

mis25 = apply(p.out25, 2, function(x) mean(x - mis25)),

mis50 = apply(p.out50, 2, function(x) mean(x - mis50)),

mis75 = apply(p.out75, 2, function(x) mean(x - mis75)))

# Correlations with truth

cor(p.out01)[1, ]

cor(p.out25)[1, ]

cor(p.out50)[1, ]

cor(p.out75)[1, ]

# proportions correct

sign01 <- p.out01 < .05

sign25 <- p.out25 < .05

sign50 <- p.out50 < .05

sign75 <- p.out75 < .05

p.correct01 <- apply(sign01, 2, function(x) mean(x == sign01[, 1]))

p.correct01

p.correct25 <- apply(sign25, 2, function(x) mean(x == sign25[, 1]))

p.correct50 <- apply(sign50, 2, function(x) mean(x == sign50[, 1]))

p.correct75 <- apply(sign75, 2, function(x) mean(x == sign75[, 1]))

#Plots

library(ggplot2)

library(gridExtra)

# plot function

qqplot <- function(x, title = "ECDF and theoretical CDF"){

ggplot(aes(x), data = data.frame(x)) +

stat\_function(fun=punif,

args=list(0, 1),

col = "orange",

lwd = 1.1) +

stat\_ecdf() +

labs(title = title) +

xlab("Empirical CDF") +

ylab("CDF") +

xlim(0, 1)

}

# 01% missingness

qqplot(p.out01$true)

qqplot(p.out01$mis)

qqplot(p.out01$D1)

qqplot(p.out01$D2)

qqplot(p.out01$D3)

qqplot(p.out01$micombine)

qqplot(p.out01$pbar)

# 25% missingness

qqplot(p.out25$true)

qqplot(p.out25$mis)

qqplot(p.out25$D1)

qqplot(p.out25$D2)

qqplot(p.out25$D3)

qqplot(p.out25$micombine)

qqplot(p.out25$pbar)

# 50% missingness

qqplot(p.out50$true)

qqplot(p.out50$mis)

qqplot(p.out50$D1)

qqplot(p.out50$D2)

qqplot(p.out50$D3)

qqplot(p.out50$micombine)

qqplot(p.out50$pbar)

# 75% missingness

qqplot(p.out75$true)

qqplot(p.out75$mis)

qqplot(p.out75$D1)

qqplot(p.out75$D2)

qqplot(p.out75$D3)

qqplot(p.out75$micombine)

qqplot(p.out75$pbar)

#side by side

p1 <- qqplot(p.out75$D1, title = "D1")

p2 <- qqplot(p.out75$D2, title = "D2")

p3 <- qqplot(p.out75$D3, title = "D3")

grid.arrange(p1, p2, p3,

nrow = 1, respect=TRUE,

top = "ECDF and theoretical CDF")

**Introduction**

Missing data is a common problem in social sciences where personal data are collected. The missingness results in statistics not being defined, which means that standard analyses cannot be carried out on incomplete data sets. Simply analyzing the observed cases may be too limited.

When the missingness in the data is simply ignored, the researcher is taking a great risk. If the missingness is not truly random, estimates may be biased. Moreover, not treating a missingness problem results in invalid inference: variances of estimates are wrong because the number of **effective cases** in the data set is smaller than the sampled set. This smaller set of fully observed cases may also lead to lower statistical power. Ultimately, the missing data may lead to severely biased results and may pose a threat to the validity of the inference (Kang, 2013).

Multiple imputation (MI) is one of the popular methods used to obtain valid inference on incomplete data (Rubin, 1987, p. 2). In MI, missing values are filled in (imputed) *m* times and the resulting *m* data sets are analyzed as if they were completely observed. The *m* analyses are then pooled into a single inference. With this procedure the missing data problem and the analyses problem are clearly separated.

With this separation of problems and the subsequent pooling of analyses comes a new issue: there may be more than one way to pool inferences. For example, with a common analysis of variance (ANOVA), there are multiple approaches for drawing inference after multiple imputation - i.e. D2-statistics which computes the statistics of the repeated analyses and compares two nested models.

In this manuscript, I investigate the validity of the currently known approaches for pooling ANOVA’s after imputation and I present some guidelines for pooling ANOVA’s in practice.

**Literature review**

**Analysis of variance**

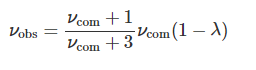
ANOVA is a family of statistical models that is originally aimed at determining differences in a scale-level outcome variable by nominal-level variable with at least two categories (ref Fisher). The aim of an ANOVA is to determine whether groups means differ on the outcome variable, by looking whether between groups significantly differs from the differences of the within groups. The statistical null hypothesis of Anova for two groups is noted as follows:

H0: u1=u2 (1).

The null hypothesis is tested by

Another important component when testing hypothesis, probability distribution and ANOVA analysis are the degree of freedom and the F-test. Sampeling distributions are based on the sample size (population). The sample size is not used directly, instead, the degrees of freedom is used. The DF of an estimate indicates the number of independent values of the sampling that can vary in a dataset without breaking any constraints. So is it the number of observations after accounting for the number of parameters in the model. The degree of freedom of missing data is different from the DF of the complete data.

To account the missing values to the estimated observed data degrees of freedom, the following formula can be used



Where *v com* is equal to the sample size *n*, minus the *k* parameters on data

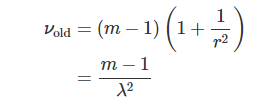


The formula of the adjusted degrees of freedom that can be used, for testing in multiple imputation, can be written as



(van Buuren, 20120, p.42).

Where the *v old* is



(Rubin, 1987b eq. 3.1.6)

The F-test is a parametric test that requires a normal distribution. The F-test can be used in ANOVA to determine whether the variability between group means are equal to the variability of the observations within the group. The F-test compares the variances of two normal distributions, S1 and S2, to determine whether the variances of the distributions differ from each other.

The F-statistics is the test statistic for the F- test and it is defined as

F= S^2-1/S^2\_2

The F-statistic does not depend on the unit of the measurement and therefore it is a ratio.

The F ratio of two measurements are expected to be equal under the null hypothesis, which produces a positive F-ratio that can be up to approximately 1. So if the variance are equal, then the F ratio would be equal to 1. The distribution of the ratio of the estimated variances is called the F-distribution (Field, 2017).

…..

Aside from determining group means, the F-test is often used in regression analysis to assess overall model significance - i.e. does the model under investigation explain more variance in the outcome than a model where the outcome is only modeled based on the intercept. In that case the F-test can be written as

**F = MS model / MS residual**

**Performing ANOVAs on incomplete data**

Multiple imputation is a popular technique to solve missing data problems in practice. This popularity can be explained by the excellent implementation of the technique in statistical software, such as R (ref R), SPSS (ref SPSS), SAS, (ref SAS), SPLUS (ref SPLUS) and Mplus (Ref Mplus). The other reason for its popularity lies in the intuive approach to solving the missingness problem at hand: The missingness problem is solved before analyses can take place, such that standard statistical analyses can be performed on the imputed data set.

In the case of ANOVA, such a procedure would be as follows:

1. Multiple imputation software is used to obtain more than 1 imputed version of the data.
2. Each imputed data set is seperately analyzed with ANOVA
3. The seperate ANOVAs can be interpreted
4. The seperate ANOVAs are pooled into a single, overarching ANOVA that describes the average effect

The last step is crucial in this procedure.

* Some cells are observed. Some cells are imputed.
* Imputed cells do not carry the same weight as observed cells
* Imputed cells are uncertain.

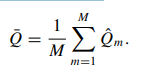
The **MI** uses all the observed data, including the covariate x, to generate the replacements for the missing value(s) *y*, which takes the relation of x into account/independent variable. There is a different between the observed and the imputed data/cells.

Altho the difference is not that extreme, the imputed data tend to have more ..

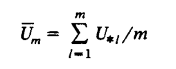
Naturally, when drawing imputations it is crucial that good software is used and that the model that is used for imputation is congenial with the analysis model (Meng 1994 https://projecteuclid.org/euclid.ss/1177010269)

**Pooling inferences after imputation**

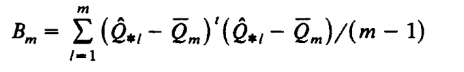
Before I can detail the strategies for pooling ANOVA’s, we first need to establish how inferences are pooled after imputation. Rubin (1987) proposed the following procedure for combining the m repeated complete data estimates and variances



by defining ^Q as the estimate of parameter Q of a complete dataset, where ^Q is the mean of the m complete data estimates. And by using the ^ U (^Um) as the within-imputation variance/the standard error of the complete data



And by using B as the between-imputation variance of the *m* complete data estimates



Further, Tm can be used as the total variance of the average of the m complete data estimates and variances



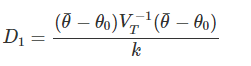
Rubin’s rules assume that the estimand has a normal sampling distribution. The F-distribution is not normal. Moreover the F-distribution does not reflect a population parameter, but rather considers a model parameter.

**How to pool ANOVA after imputation.**

Further, Van Ginkel and Kroonenberg (2011) proposed a way for applying the pooling procedures when using multiple imputation for ANOVA.

Furthermore, there are different strategies to pool ANOVA after multiple imputation. The multivariate Wald test, D1, the combining test statistics, D2, and the likelihood ratio test, namely D3 can be used to perform multivariate parameter tests on the m repeated analysis of the multiple imputed data.

D1-statistics is the pooled version of the F-test, also called the multivariate Wald test. The requirements for D2 function are the variance-covariance matrix estimates U of Q (Van buuren, 2012). The equation of Wald statistics is computed as:

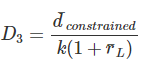
****

D2 statistic is a function that computes the statistics of the repeated analyses. This function compares two nested models. The equation of D2 method is as follows:



D2 statistics can be used for any general scenario because it only requires a vector of chi-squared and Df (k) from each imputation for the test statistic.

D3 statistics is the likelihood ratio test that is developed to manage occasions/cases where it is not possible for the researcher to run covariance matrices of the complete-data estimates. The formula is as followed:



1. Every parameter θ two nested models get fitted in each imputed dataset, hereby includes is the full model included and
2. Average likelihood ratio test statistic dL

The function *micombine* takes *mids* objects as input. It can be used to combine chi-squared test (X2) and the F-test from multiply imputed datasets. The function micombine.chisquare can be used for the X2 distribution. These function uses the vectors calculated on each imputed dataset and then pools them (Enders, 2010, p.239; Alisson, 2002). The approximation of chisquare is used to combine the F statistics for MI dataset. It corresponding function in R is *micombine.F*. Hereby, consists *micombine. F*  a vector containing F values, degree of freedom of the numerator and the version of the formula that should be used. Micombine.chisquare and micombine.F give us the same output value (Allison, 2002; Eders, 2010; Grund, Luendtke & Robitzsch, 2016). The Micombine does the same as D2 statistics (...)

**Comparison D1, D2 and D3**

1. D1-statistics is the pooled version of the F-test, also called the multivariate Wald test. D1 compares two nested models. It uses the pooled parameter estimates and the pooled sampling variances to construct a test that closely resembles the multivariate Wald statistic. D1 statistic uses a total parameter covariance matrix based on the average relative increase in variance. Assuming that the relative increase in variance is the same for all parameters (i.e., ARIV is representative of each parameter’s RIV value). This assumption is unlikely to hold in practice because it essentially requires that the analysis variables have the same missing data rates and the same correlations.
2. D1-Statistic and D3-statistic pools are more powerful than D2-statistic pools. In contrast to D1 and D3, D2-statistic obtains less information from the data.

* D2 Is approximately normally distributed

1. D2 and D3 pool significance tests from the analysis phase;

* mncombine = D2

In data analysis and statistics, the issue of consist and accuracy is highly affected by missing values. Therefore understanding the concept of missing values will help in ensuring consistency and accurate results. The BLUE property of statistical estimators can therefore not be achieved in the presence of missing data points the analysis data. Biasedness as commonly referred to in statistics is a situation whereby the calculated estimate parameter differs from the actual population parameter. This might result from bias in the data or missing values.

# **When are pooled ANOVAs valid?**

We have seen that there are several techniques for combining ANOVAs after multiple imputation. But before we can answer the question which technique is best, we need to consider what a good ANOVA pooling procedure would result in. In this sense, pooling an ANOVA and assessing its performance is less straightforward than pooling a population parameter, because the ANOVAs F-test is designed to change based on the observed information. This observed information is very fragile in the case of incomplete data sets. Simply pooling the ANOVAs by taking the average may be too limited and could potentially result in biased estimates and invalid conclusions (referentie).

There is another approach for assessing the validity of ANOVAs. We usually interpret the F-test by looking at the significance level of the test. This significance level is determined by the p-value, which reflects the probability of finding a *less likely* result under the null hypothesis - i.e. when there is no effect. As opposed to the F-test itself, its corresponding p-value *behaves* consistently under the null hypothesis. More specifically, for any ANOVA based on any given data set under the null hypothesis, the corresponding p-value has a uniform distribution, an expected value of .50 and requires the p-value to be confidence valid (a p-value of .05 should only occur 5 percent of the time).

**Uniform distribution**

* Follows a uniform distribution where 0 <= p <= 1 (Murdoch, Tsai & Adcock, 2008).
* Has an expected value of .50 (Murdoch et al., 2008).
* Should be confidence valid (Neyman, 1934) in the sense that a p-value of .05 should only occur 5 percent of the time.

This means that, under the null hypothesis, 100 ANOVAs performed on 100 random sets would result in 100 p-values, where no p-value has a probability of occurrence that is larger than any of the others and the average probability would amount to E(p-value) = .50.

**How to evaluate the validity of ANOVA inference after imputation**

**Introdoction to how to evaluate the above 3 points:**

1. Follows a uniform distribution where 0 <= p <= 1 (Murdoch, Tsai & Adcock, 2008).
2. Has an expected value of .50 (Murdoch et al., 2008?).
3. Should be confidence valid (Neyman, 1934) in the sense that a p-value of .05 should only occur 5 percent of the time.

1.Uniformity of the p-value will be evaluated using Kolmogorov Smirnov test and the qqplot. The Kolmogorov Smirnov test is a nonparametric test that is based on a measure of differences in two distributions. The KS can be used to compare a sample with a reference probability distribution or to compare two samples. Kolmogorov Smirnov determines whether the sample obtained from a population deviates from a specific distribution, as the normal distribution and the uniform distribution (Chakravart, Laha & Roy, 1967).



The Kolmogorov Smirnov test is based on the empirical distribution



**The p-value and D statistic in KS test**

The KS test gives us a D statistic and p-value corresponding to the D statistic. Thereby, the D statistic is the absolute maximum distance between the (CDF) of two samples. The closer the D statistic is to 0, the more likely it is that the two compared samples are from the same distribution (...).

However, there are 7 steps to run the KS test.

1. Create an EDF for your sample data (see [*Empirical Distribution Function*](https://www.statisticshowto.com/empirical-distribution-function/) for steps),
2. Specify a parent distribution (i.e. one that you want to compare your EDF to),
3. Graph the two distributions together
4. **Measure the greatest vertical distance between the two graphs.**
5. Calculate the test statistic.
6. **Find the critical value in the** [**KS table**](https://www.statisticshowto.com/kolmogorov-smirnov-test/#table)**.**
7. **Compare to the critical value.**

2. How to examine: Has an expected value of .50 (Murdoch et al., 2008).??

**Raw bias (x-true)?**

The raw bias of the estimate also called RB, can be defined as follows



Literatuur: Where the is the expected value of the estimate and the is the truth value.

If the RB value is close to zero, then we can conclude that there is no bias. For acceptable performance we use an upper limit for PB of 5%. (Demirtas, Freels, and Yucel [2008](https://stefvanbuuren.name/fimd/references.html#ref-DEMIRTAS2008C)).

To evaluate the bias a defined in section ..(raw bias), we have to look at the bias in the p value with respect to the population E=( .50) under the null hypothesis. We start by specifying a ANOVA model of the complete data set and an ANOVA model with incomplete data set, where a part of the observations are deleted, resulting in an incomplete data set to generate approximately 1%, 25%, 50% and 75 % data sets with random missing data mechanism. The missing values have to be filled using multiple imputation procedure and the pooling techniques of Rubin’s rules have to be applied to compute the estimate  and the *T*. Therefore D1,D2 and D3 will be used. After performing 1000 simulations.

ANOVA multiple imputation by method … …

(x-mis) in code

(x-.5) in code

3. How to examine: Should be confidence valid (Neyman, 1934) in the sense that a p-value of .05 should only occur 5 percent of the time.

…

# **The concept of Bias**

There are several types of bias that can arise in statistical data analysis such as selection bias, self-selection bias, recall bias, cognitive bias, and observational bias. Selection bias occurs whenever the researcher, performs a random sample or any other kind of sampling, such that the elements in the data set do not have equal chances of being selected. This results in a predefined element distribution, for instance having a certain class of data having too much or too few data points (McNeish, 2017).

Randomization is one of the most applied concepts in dealing with bias, this notion gives data elements equal chances of appearing or missing in the sampled data. The benefit of randomization is that it helps in obtaining a representative sample that is a data section that entirely represents the population of interest. Conducting any type of estimation in such kind of data will result in unbiased population estimators and the estimates will be BLUE (best linear unbiased estimates) most of the time (Schouten, Lugtig &, Gerko, 2018).

Data imputation is a technique used in statistical analysis in dealing with missing numbers by introducing and estimating the missing values with a function of the available data. Additionally, imputation helps in dealing with the pitfalls which result from the listwise deletion. Listwise deletion involves deleting the data rows which have missing data. This results in a decreased sample size which finally might not be representing the population due to decreased sample size.

Imputation exists in different forms such as mean imputation, multiple imputations, regression imputation, stochastic regression imputation, Cold Dec and Hot Deck imputation, interpolation and extrapolation imputation, etc. all these techniques can be used in obtaining new data elements in the missing cases (Twisk, de Boer, de Vente, & Heymans, 2013). Understanding these techniques is important to understand where best to apply each of the techniques. While dealing with missing values, the crucial concept is to understand the structure at which the missing values occur to comprehend the appropriate techniques for the imputation. The MCAR is a situation whereby the missing data are completely missing randomly such as the probability of a missing number occurring in the data is not related to any other value or values as well the data variables (Schouten et al., 2018). The MAR missing data technique, the likelihood of data value missing is related to the data variables, in other words, there is a predetermined relationship between the missing data values and the data variables.

Aside from the above criteria it would also be important to evaluate how often an pooling approach would yield the same conclusion as the true (or population) data. Even if all of the above criteria are successfully satisfied, a method cannot be deemed sufficient if it never identifies the effect in the true data set.

**Simulation setup**

We use statistical simulation to evaluate the performance of the p-values for the following ANOVA combination approaches for multiply imputed data:

* D1
* D2
* D3
* Function micombine.F (Allison, 2002; Eders, 2010; Grund, Luendtke & Robitzsch, 2016).
* Taking the average p-value

To set a baseline performance to which all of the approaches are compared, we also calculate the p-values for the true data set and for the incomplete data set - i.e. the set where the missingness is ignored and only complete rows are analyzed.

To generate imputations under the null hypothesis, we simulate a 1000 complete data sets independently from a bivariate normal distribution with means … and variance covariance matrix …

**Give a description of the simulation:**

For every simulated data set, we ampute the data according to a *missing completely at random* missingness mechanism and subsequently impute the data with mice (Van buren, 2011) in R (R Core Team, 2020). We calculate the linear intercept-only model, which contains the slope intercept, and the linear model with the intercept and the predictors. Furthermore, we used D1 statistics, D2 Statistics and D3 statistics to compare the intercept-only model and the model with intercept and predictors.

Afterwards, we used the function *micombine.F* from the package miceadds to multiply the F-value with the degrees of freedom (Allison, 2001), which conforms to Eders (2010).

We also calculated the p-values of the true data, data with missing, D1, D2, D3, micombine and pbar with varying degrees of missingness.

and .... We used the average p-value as the expected p-values of the above statistics. In addition we computed the proportions correct (p <=.05).

The following pseudo code details the full scope of the simulations:

# pseudo-code of simulation

for (number of simulation runs from 1 to 1000)

1. simulate data

for (missingness proportions 1%, 25%, 50% and 75%)

2. create missingness

3. impute missingness

4. Run regressions models

5. Compare regression models

6. Compute performance measures

7. Aggregate outcomes over all simulations

**Results**

**-1. unifrom +ggplot**

**- 2.expected value of p-values**

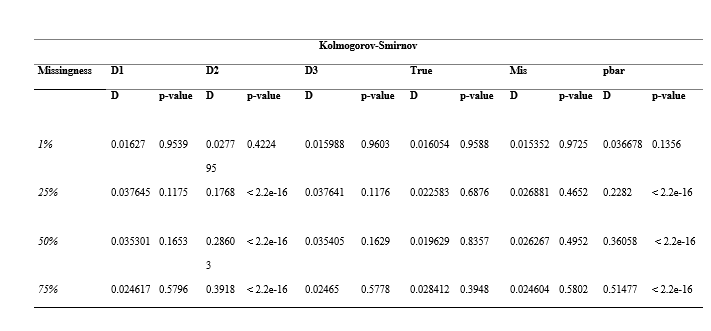
**- 3.Proportion E(p<=.05)**

**- 4.Proportion correct**

The results of the simulation are discussed below. D2-statistics and micombine.F are identical to each other. Therefore the results of D2-statistics are discussed in the results section.

***Uniformity of p-values***

Tabel 1

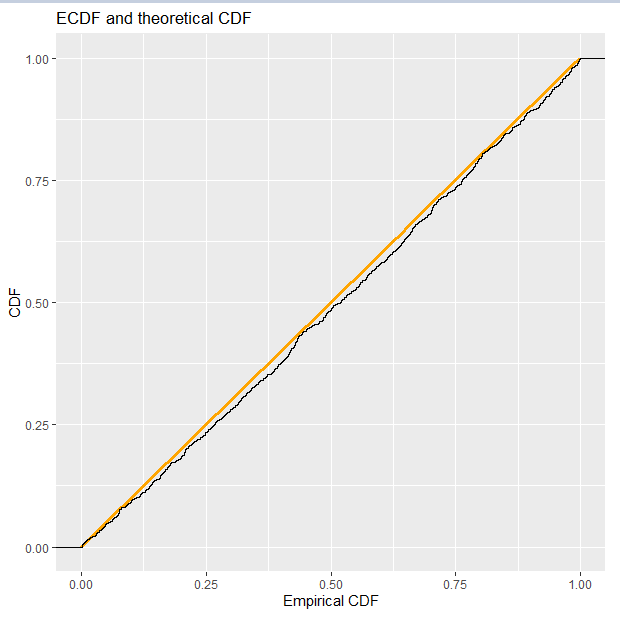
*Uniformity of P-Values of D1, D2, D3, Mis, True and Pbar with different Missigness Proportion* ******Results for the Kolmogorov-Smirnov test are shown in Table 1. Hereby we tested the null hypothesis whether the p-values follow a uniform distribution. Thus, a significance less than 0.05 indicates that it is unlikely that the data is normally distributed. The rows contain missingness proportions (1%, 25%, 50% and 75%) and the columns contain the used techniques D1, D2, D3, True, Mis and pbar. Looking at the missingness proportion of 1%, we **see that** D1-statistics ***D***(999)= 0.016 ***p*** = .95, the D2-statistics***D***(999)= .028 ***p=*** .42 and the D3-statistics **D**(999)= .015 ***p=*** .96 are significantly normal distributed. Furthermore, the true data **D**(999)= .996 ***p=***  .96, the mis data **D**(999)= .015 ***p=*** .97 and pbar **D**(999)= .057 ***p=*** .14 are also significantly normal distributed.

Looking at the missingness proportion of 25%, we **see that** D1-statistics ***D***(999)= 0.038 ***p*** = .12, the D3-statistics **D**(999)= .015 ***p =*** .96 , the true data **D**(999) = .023 ***p =***  .69, the mis data **D**(999)= .027 ***p=*** .47 are significantly normal distributed. The D2-statistics ***D***(999)= .177 ***p =*** .00 and and pbar **D**(999)= .228 ***p=*** .00 are significantly not normal distributed (Table 1.).

At 50% missingness proportion, **the** D1-statistics ***D***(999)= 0.035 ***p*** = .17, the D3-statistics **D**(999)= .038 ***p =*** .16, the true data **D**(999) = .020 ***p =***  .84, the mis data **D**(999)= .026 ***p=*** .36 are significantly normal distributed. The D2-statistics ***D***(999)= .177 ***p =*** .00 and pbar **D**(999)= .228 ***p=*** .00 are not significantly normal distributed (Table 1.).

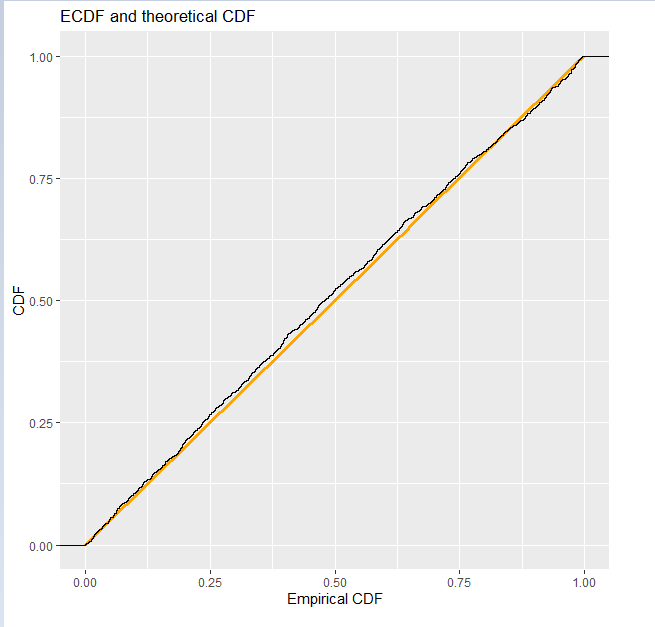
At 75% of missingness, **the** D1-statistics ***D***(999)= 0.025 ***p*** = .58, the D3-statistics **D**(999)= .025 ***p =*** .58, the true data **D**(999) = .028 ***p =***  .39, the mis data **D**(999)= .025 ***p=*** .58 are significantly normal distributed. The D2-statistics ***D***(999)= .39 ***p =*** .00 and pbar **D**(999)= .515 ***p=*** .00 are not significantly normal distributed (Table 1.).

**2. Plaatje 1:** window van 2 vensters: **qqplot** van de **true data** en **mis data (75%)** daaronder in tekst de beschrijving + interpretatie van dit plaatje.

****

*Figure 1. Empirical Cumulative Distribution and theoretical cumulative distribution of 75 % missigness of true data.*

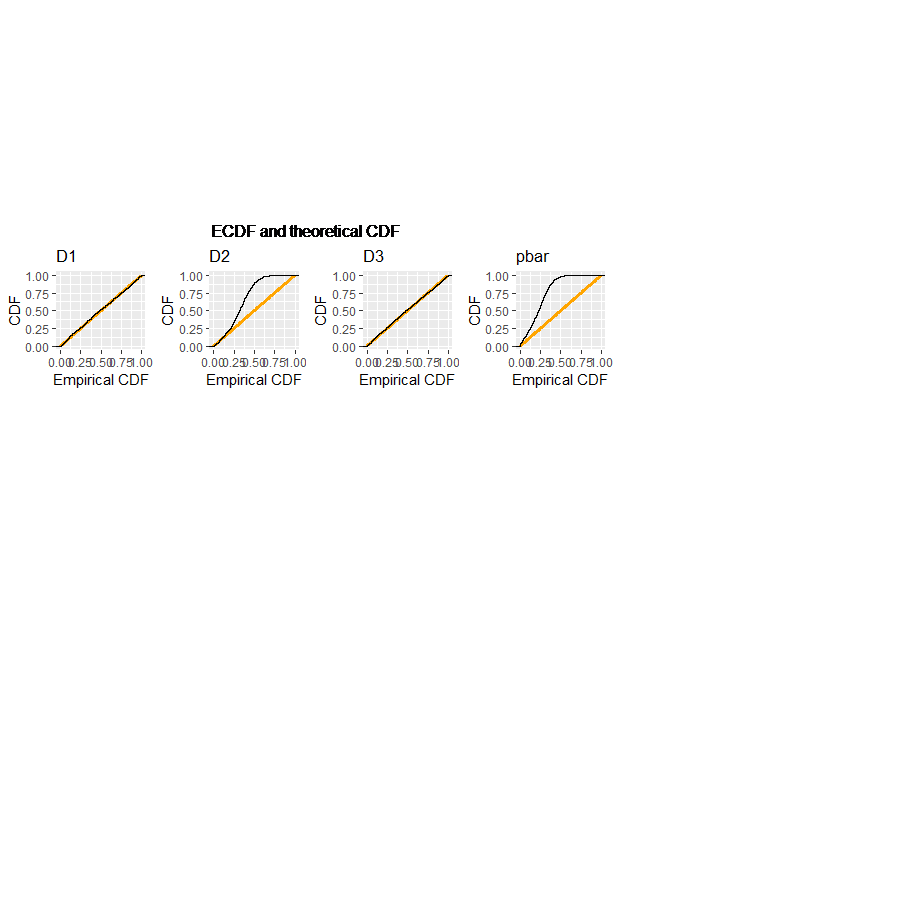
The Empirical Cumulative Distribution Function (ECDF) is a summary statistic visualization that is an estimator of the Cumulative Distribution Function (CDF) (ref). Figure 1 shows the similarity between the ECDF and the theoretical CDF. This confirms the KS test decision that true data at 75% of missingness is significantly normal distributed.

****

*Figure 2. Empirical Cumulative Distribution and theoretical cumulative distribution of 75 % missingness of mis data.*

By visually comparing the ECDF and the theoretical CDF of mis data at 75 % missingness, we confirm the decision of the KS test that miss data is significantly normal distributed.

**3.Plaatje 2:**  wind**ow van 2x2 vensters voor 75% missingness waarin je de qqplot van D1, D2, D3 en pbar laat zien.**

****

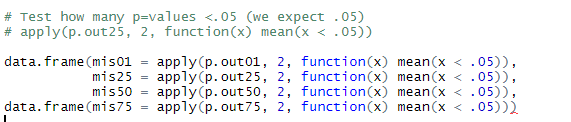
*Figure 2. Empirical Cumulative Distribution and theoretical cumulative distribution of 75 % missingness of D1, D2, D3 and pbar.*

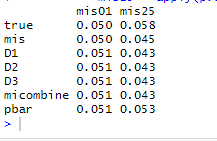
By comparing the ECDF and the theoretical CDF of D1, D2, D3-statistics and pbar at 75 % missingness, we confirm the decision of the KS test that D1 and D3 are significantly normal distributed, and that D2 and pbar are not significantly normal distributed.

***I am not sure that I used the right output (see the script):***

***→ Expected value of the p-values***

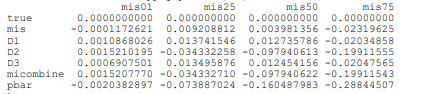
Tekst en tabel met de expected values

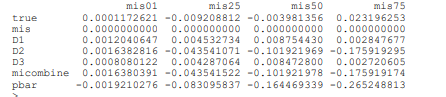
***→ Proportion E(p<=.05): ***



Tekst en tabel met de expected values

*Table 2. Bias complete data*

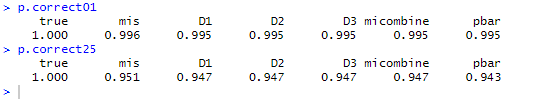
Table 2. Shows the bias in p-value with respect to the sampled complete population (***p*** = .5, under null hypothesis). The true condition is the complete data set and therefore it is unbiased ( ***p*** = .00). The mis condition with  ***p*** = -.00 at 1% missingness***, p =*** .01 at 25% missingness, ***p =*** .00 at 50% missingness and ***p =*** -.02 at 75% missingness are close to zero and therefore they are unbiased. The D1 condition with  ***p*** = .00 at 1% missingness***, p =*** .01 at 25% missingness, ***p =*** .01 at 50% missingness and ***p =*** -.02 at 75% missingness are close to zero and therefore they are unbiased. The D2 condition with  ***p*** = .00 at 1% missingness***, p = -***.03 at 25% missingness, ***p =*** -.10 at 50% missingness and ***p =*** -.20 at 75% missingness are close to zero and therefore they are unbiased. The D3 condition with  ***p*** = .00 at 1% missingness***, p =*** .01 at 25% missingness, ***p =*** .01 at 50% missingness and ***p =*** -.02 at 75% missingness are close to zero and therefore they are unbiased. The pbar condition with  ***p*** = -.00 at 1% missingness***, p = -***.07 at 25% missingness, ***p =*** -.16 at 50% missingness and ***p =*** -.29 at 75% missingness are close to zero and therefore they are unbiased.



x-mis??

***→ Proportion correct*** (i.e. hoeveel p-waarden **kleiner** dan .05 hetzelfde als de populatie)

**3. Text and tabele with proportons**

****

Discussion and conclussion:

References:

Allison, P. D. (2002). Missing data. Newbury Park, CA: Sag

Chakravart, Laha & Roy, 1967/ Marsaglia, G., Tsang, W. W., Wang, J. (2003) "Evaluating Kolmogorov’s Distribution", *Journal of Statistical Software*, 8 (18), 1-4. [Jstor](http://www.jstatsoft.org/v08/i18/paper)

Enders, C.K. (2010). *Applied missing data analysis*. New York, United States: Guilford Press.

Grund, S., Luedtke, O., & Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. Methodology, 12(3), 75-88. doi: 10.1027/1614-2241/a000111

McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, *44*, pp.24-39. https://doi.org/10.1080/02664763.2016.1158246

Murdoch, D. J., Tsai, Y. L., & Adcock, J. (2008). P-values are random variables. *The American Statistician, 62*, 242-245. Doi: <https://doi.org/10.1198/000313008X332421>

Schouten, R.M., Lugtig, P., & Vink, G., 2018. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*, pp.2909-2930. <https://doi.org/10.1080/00949655.2018.1491577>

Twisk, J., de Boer, M., de Vente, W., & Heymans, M. (2013). Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *Journal of clinical epidemiology*, *66*, pp.1022-1028. doi: 10.1016/j.jclinepi.2013.03.017

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

van Buuren, S., Groothuis-Oudshoorn, K.(2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Softwar*e, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. United States, Cambridge: Department of statistics Harvard University

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*, 402. doi: [10.4097/kjae.2013.64.5.402](https://dx.doi.org/10.4097%2Fkjae.2013.64.5.402)

Appendix

Description of notation