Energy Usage of Imputation Methods in R

Dylan Loader

April 9, 2018

Preliminary Discussion

In this experiment, my objective was to determine the computational energy requirements for different imputation methods. My factors of interest were Imputation method, parent probability distribution, and percentage of the data missing.

To increase the applicability of my experiment I used my Macbook Pro (2015) which has fairly standard computational power at the time of writing and did not utilize computational optimization, unless it was default such as in the MI package. To measure the amount of energy consumed by each experimental run, I utilized Intel's Power Gadget which reads energy consumption directly from the computer's processor inputs the data directly into an Excel CSV file. To compute the energy used in each cycle the time before and after each imputation was taken, rounded to the nearest second and used to find the cumulative joules used over each imputation.

Packages Used

The R packages: mice, Amelia, and mi were chosen after initially attempting conventional mean and median imputation methods. In this case the length of time to run a simple mean or median imputation was less than a second for large data sets, which prevented a correct measurement of energy consumption being made.

Mice stands for Multivariate Imputation by Chained Equations, which uses multivariate imputation to estimate each missing value. In this experiment the data are pseudo-randomly generated from continuous distributions, so I chose to use the predictive mean matching which is a semi-parametric approach for numerical data. Mice does have packages for categorical data which uses a logistic regression to impute values. In the case of predictive mean modeling in the mice package the imputation is performed by assuming the data are approximately jointly multivariate normally distributed. During each imputation step the algorithm fits a linear regression using all other available data for m data sets and attempts to find the closest realized value in the data set which matches the predicted value. In the case of directly modeling the data set such as with a linear regression, the regression may be run over the m generated data sets and pooled together to give better estimates. In this experiment I randomly selected one of the m data sets to return because the simulated data are not of interest in this experiment other than for their structure.

The second package used is Amelia II which implements a Bayesian multiple imputation bootstrapping algorithm to estimate values to be imputed by sampling from the posterior distribution. With Amelia m imputed data sets are calculated with a modified Expectation Maximization algorithm. Like the mice package Amelia II requires the assumption of multivariate normality and that the data are missing at random. These assumptions are also utilized similarly to mice which allows for linear regression models to be fit to predict the values to be imputed. Similarly, with the mice package I randomly selected only one data set from the m datasets which are computed for accuracy.

Lastly, I chose to use the mi package which stands for Multiple Imputation with Diagnostics. Mi is another multiple imputation method for implementing predictive mean matching for imputation. Mi uses Markov Chain Monte Carlo to impute values in a similar fashion to Mice and Amelia by creating m imputed data sets. One major benefit of the Mi package is that it has parallel processing set to default, which is useful for reducing the time it takes for convergence in MCMC. In this experiment I had to limit the default number of chains from four to one and decreased the number of iterations allowed to five.

Design and Collection of Data

I began the experiment by generating a matrix of 10 variables of 1000 observations from the uniform, normal, and exponential distributions of the following form:

$$X \sim N\Big(100,\ \frac{2500}{3}\Big)$$
$$X \sim Uniform\Big(50,\ 150\Big)$$
$$X \sim Exp\Big(\frac{3}{2500}\Big)$$

The distributions were chosen in a way such that they have equal variance in an attempt to make the distributions comparable.

Statistical Model

The model of interest is:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \psi_k + \tau\beta_{ij} + \tau\psi_{ik} + \beta\psi_{jk} + \tau\beta\psi_{ijk} + \epsilon_{ijk}$$

For all

$$i, j, k \in \{0, 1, 2\}$$

Where:
μ: The grand mean irrespective of all other effects

τ_i: Is the effect of the ith missingness treatment

β_j Is the effect of the jth distribution treatment

ψ_k Is the effect of the kth imputation treatment

$\tau\beta_{ij}$ Is the interaction effect of the ith missingness treatment and jth distribution treatment

$\tau\psi_{ik}$ Is the interaction effect of the ith missingness treatment and kth imputation treatment

$\beta\psi_{jk}$ Is the interaction effect of the jth distribution treatment and kth missingness treatment.

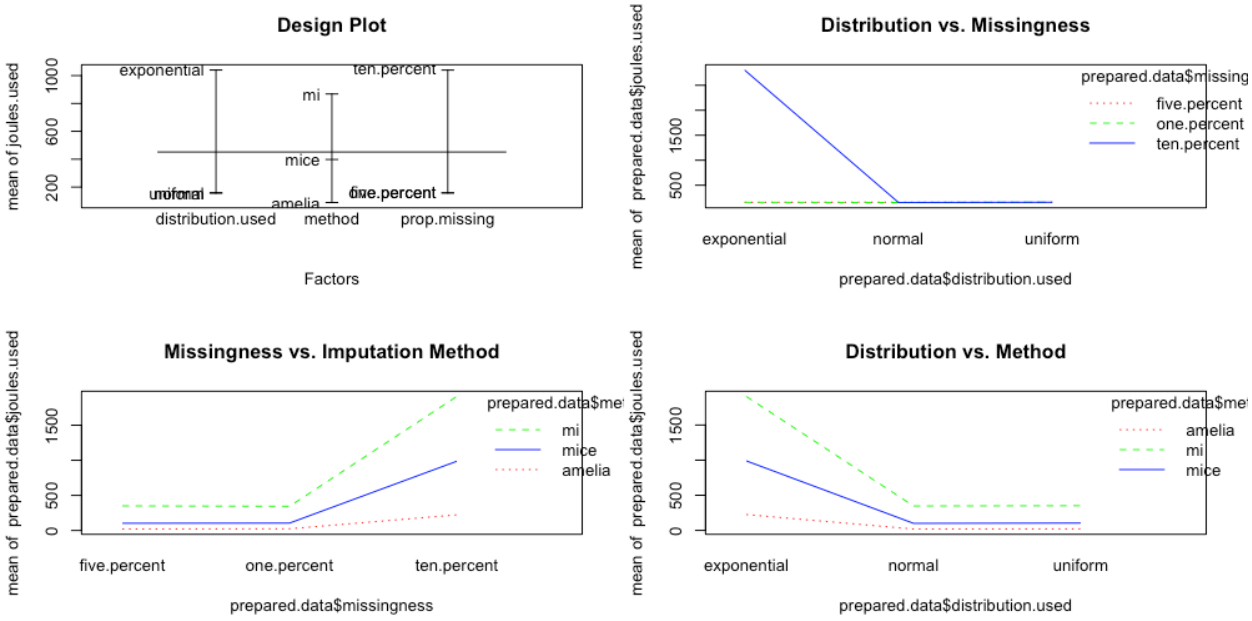
$\tau\beta\psi_{ijk}$ Is the three way interaction effect of the ith missingness treatment, jth distribution treatment and the kth imputation treatment

ϵ_{ijk} Is the random error which is

$$\epsilon \sim N\Big(0, \sigma^2\Big)$$

Data Analysis

To begin the analysis we consider the design plot of means by factor and pairwise interaction plots between factors. From the below figure for mean Joule requirement we can see that the means of uniform and normal are almost identical, where both differ significantly from the mean for exponential. For imputation level, Amelia appears to have the lowest mean energy usage followed by mice, with mi being the highest. From the intersection plot for Distribution type vs Missingness level there appears to be an interaction effect between the exponential distribution and a 10% level of missingness. For the Missingness vs. Imputation method there appears to be little evidence of an interaction effect as the lines appear fairly parallel. For the Distribution vs Imputation Method interaction plot the lines are also fairly similar suggesting little concern of interaction.



Using the above described model a full ANOVA was fit with two way and three way interaction effects with the hypotheses defined above at the $\alpha = 0.01$ level of significance:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
missingness	2	13989849	6994924.4	11.596089	0.0000646
distribution.used	2	14013809	7006904.3	11.615949	0.0000637
method	2	8315509	4157754.5	6.892668	0.0021576
missingness:distribution.used	4	28066093	7016523.3	11.631895	0.0000007
missingness:method	4	5564607	1391151.7	2.306232	0.0698758
distribution.used:method	4	5474497	1368624.2	2.268886	0.0736608
missingness:distribution.used:method	8	11113364	1389170.5	2.302948	0.0334140
Residuals	54	32573562	603214.1	NA	NA

To begin interpreting the results of our ANOVA we should verify that the assumptions of the model are correct.

Independence:

Since the data are randomly simulated and because the previous run's energy consumption has a negligible effect on the next run's energy consumption, we may assume independence.

Normality:

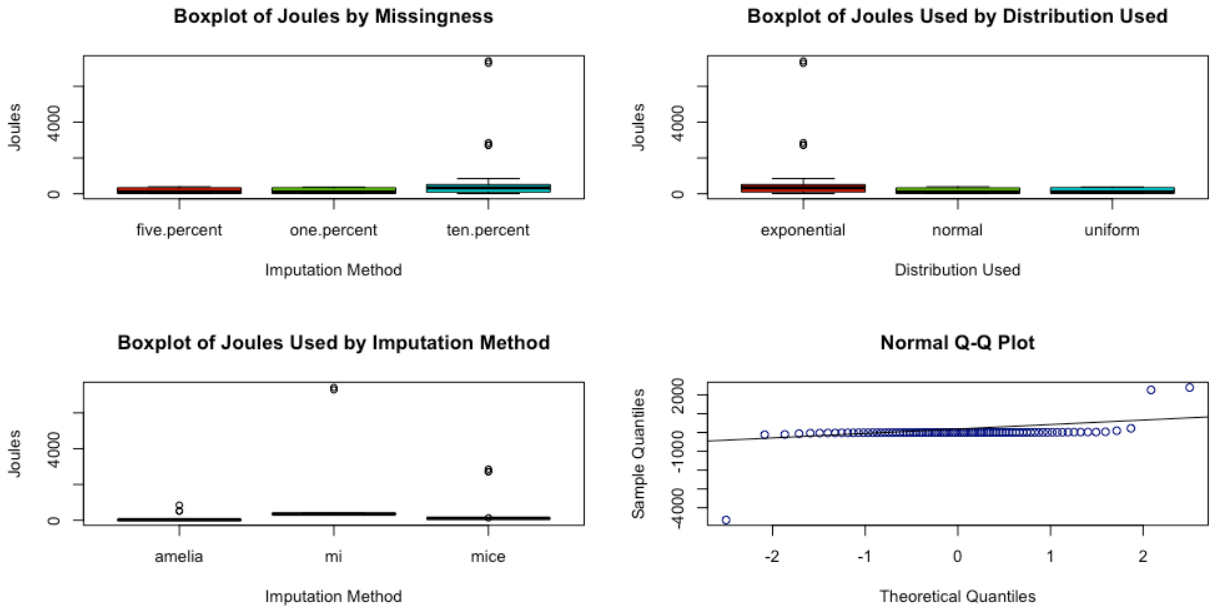


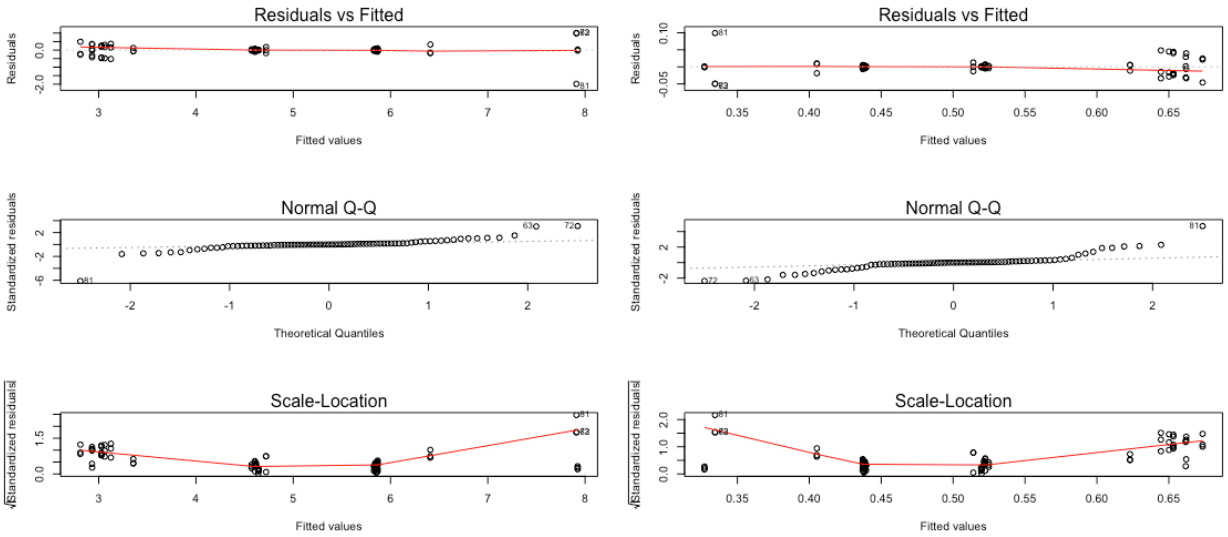
Figure 1: Normality Plots

From the above figure we can see there is a severe violation to the assumption of normality as residuals of the model are extremely right skewed as shown by the Normal QQ plot. Hence a transform should be performed on the data. From the box plots by factor, we can see there are outliers in the data. These outliers

are contained in the ten percent missingness level and the exponential distribution. This corresponds with general practice as imputation is usually suggested only for datasets with less than 5% missingness and the packages used assume multivariate normality. Given this violation we should attempt a transformation.

Transformation

Plotting the estimation for optimal power gives $\lambda = -0.1410263$ which is close to 0 in which case a log transform is recommend. I applied both transformations and refit the ANOVA model. The left log transform appears to give a worst result with more extreme outliers in the upper and lower quartiles, when compared to the Box-Cox on the right. In this case, I chose to utilize the natural log transformed data to retain interpretability.



Transformed ANOVA Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
missingness	2	15.8709260	7.9354630	50.2466913	0.0000000
distribution.used	2	18.9647465	9.4823733	60.0415982	0.0000000
method	2	97.4614539	48.7307269	308.5589070	0.0000000
missingness:distribution.used	4	33.1199366	8.2799842	52.4281705	0.0000000
missingness:method	4	0.6219011	0.1554753	0.9844565	0.4238572
distribution.used:method	4	1.2441575	0.3110394	1.9694754	0.1122810
missingness:distribution.used:method	8	1.4339055	0.1792382	1.1349213	0.3554969
Residuals	54	8.5282233	0.1579301	NA	NA

Removing the non-significant high order terms I find the ANOVA model to be:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
missingness	2	15.87093	7.9354630	46.96260	0
distribution.used	2	18.96475	9.4823733	56.11732	0
method	2	97.46145	48.7307269	288.39168	0
missingness:distribution.used	4	33.11994	8.2799842	49.00150	0
Residuals	70	11.82819	0.1689741	NA	NA

To verify this model was significant to model the data I ran a Chi-squared test since the proposed final model is nested in the log model, which resulted in a p-value of 0.2278 . Thus we do not have significant evidence to suggest that they are significantly different. Therefore we choose the reduced model:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \psi_k + \beta\psi_{jk} + \epsilon_{ijk}$$

Using the selected models we may calculate our least squares estimates for mean effect of the form (shown for factor i)

$$\hat{Y} = \sum_i c_i \sum_j \sum_k y_{ijk}$$

:

$$\hat{\mu} = 4.827659 \tag{1}$$

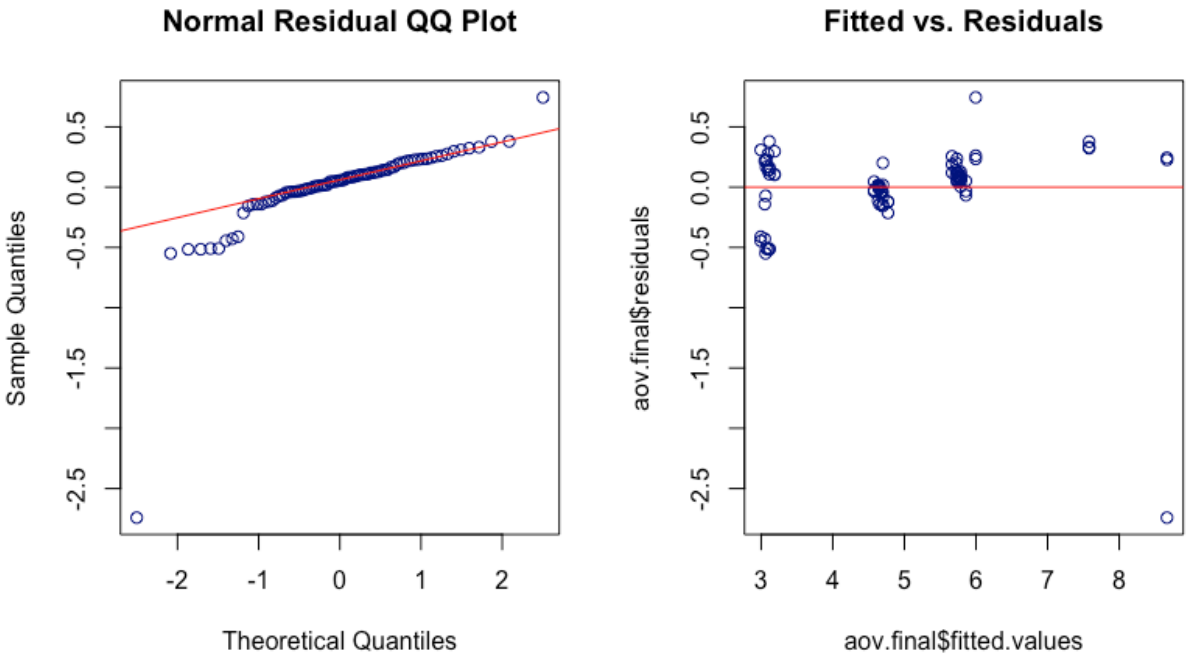
$$\hat{\tau}_{OnePercent} = 4.542; \hat{\tau}_{FivePercent} = 4.488; \hat{\tau}_{TenPercent} = 5.453 \tag{2}$$

$$\hat{\beta}_{Exponential} = 5.511; \hat{\beta}_{Normal} = 4.460; \hat{\beta}_{Uniform} = 4.511 \tag{3}$$

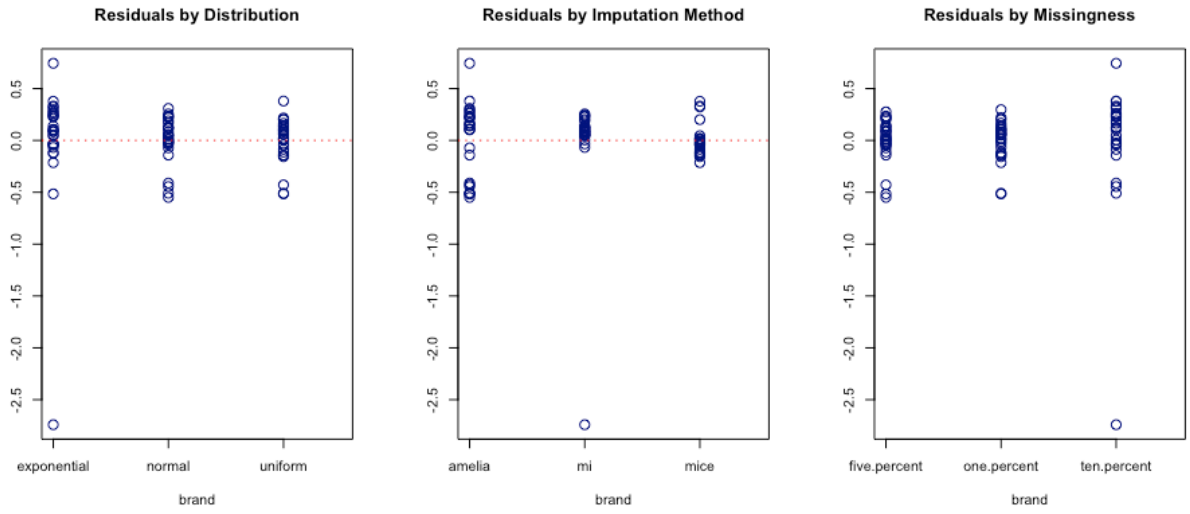
$$\hat{\psi}_{Amelia} = 3.409; \hat{\psi}_{Mi} = 6.081; \hat{\psi}_{Mice} = 4.993 \tag{4}$$

Before continuing with this analysis we should check that the residuals of our model are normally distributed and homoscedastic.

Residual Analysis



The normal QQ plot looks to have outlying data in the lower quantiles and thus I suspect the assumption of normality may be violated. Running the Shaprio-Wilk test of normality on the log of joules used and we find a p-value of 3.009e-13 suggesting extremely strong evidence to reject the H_0, that the residuals are normally distributed. The ANOVA model has only an approximate assumption of normality so I continued by testing the assumption of homoscedastic variance of residuals.



At the $\alpha = 0.05$ level of significance we test for homoscedasticity of variance using Bartlett's test and Levene's test:

$H_0 : The\ variance\ of\ the\ factor\ are\ homoscedastic\ vs.\ H_1 : The\ variance\ of\ the\ factor\ are\ heteroscedastic$

Bartlett's Test	Factor	Statistic	DF	P-value
Imputation Method	1	7.381231	2	0.0249566
Distribution	2	4.947412	2	0.0842720
Missingness	3	2.899623	2	0.2346145

Here we see that the violation of the assumption of homoscedastic variances are violated for the Imputation and Distribution factors. Since I have found a violation to the assumption of normality, I chose to run Levene's test because the median based statistic is robust against violations to normality.

Levene's Test	Factor	Statistic	DF	P-value
Imputation Method	1	2.9815	2	0.05653
Distribution	2	1.9593	2	0.1478
Missingness	3	0.913	2	0.4056

At the $\alpha = 0.05$ level of significance we fail to reject the null for all factors. Using Levene's test, the data do suggest that the assumption of homogeneous variance is valid.

Since the interaction effect for missingness and distribution is statistically significant, we may not directly say which distribution requires the most energy at all levels of missingness. In this case we may construct Tukey's confidence intervals by fixing the factor of interest and taking the marginal mean.

Constructing the Tukey HSD confidence intervals at the $\alpha = 0.05$ level of significance level provided in the appendix, we find significant differences in the Missingness factor pairs: (ten percent, one percent) and (ten percent, five percent).
In the table for the distribution factor, I found a significant difference in the factor level pairs: (Normal, Exponential) and (Uniform, Exponential).
In the table for the Imputation Method factor, I found a significant difference between all imputation methods.

Examining the Tukey HSD confidence intervals for I found all terms pairings of the form (Missingness:Exponential,Missingness:Uniform) and (Missingness:Exponential,Missingness:Normal), with all other pairings being statistically insignificant.

Conclusion

Looking at the Tukey HSD confidence intervals we find the log of joule used is consistently higher for the factor of missingness of the percent with no statistically significant difference between the one percent and five percent levels. This suggests that given the option to impute five percent of values over one percent we should not expect to see a significant increase in energy consumption. For the factor of distribution chosen we see that values from exponential distribution are significantly more energy intensive to impute. Where as there appears to be no statistically significant difference between imputing values from the normal and uniform distributions. This is most likely do to the imputation methods requiring multivariate normality. For the imputation methods we find that all the methods require significantly different with the Mi package consuming the most power, followed by Mice, with Amelia requiring the least. To further explore this result we could compare the quality of estimation done by the models and attempt the finely turn the parameters of the algorithms that the packages use to make them more comparable. For the two-way interaction we find

that larger required proportions of missingness with exponentially distributed data requires more energy to impute.

Appendix

Tukey's (Missingness)	diff	lwr	upr	p adj
one.percent-five.percent	0.0548301	-0.2130677	0.3227279	0.8762497
ten.percent-five.percent	0.9652121	0.6973143	1.2331099	0.0000000
ten.percent-one.percent	0.9103820	0.6424842	1.1782799	0.0000000

Tukey's (Distribution)	diff	lwr	upr	p adj
normal-exponential	-1.0511374	-1.3190352	-0.7832396	0.0000000
uniform-exponential	-0.9998362	-1.2677340	-0.7319384	0.0000000
uniform-normal	0.0513011	-0.2165967	0.3191989	0.8907591

Tukey's (Imputation Method)	diff	lwr	upr	p adj
mi-amelia	2.671609	2.403711	2.9395070	0
mice-amelia	1.583602	1.315704	1.8514999	0
mice-mi	-1.088007	-1.355905	-0.8201093	0

Tukey's (Missing*Distribution)	diff	lwr	upr	p adj
one.percent:exponential-five.percent:exponential	0.0919301	-0.5283200	0.7121801	0.9999238
ten.percent:exponential-five.percent:exponential	2.9007009	2.2804508	3.5209509	0.0000000
five.percent:normal-five.percent:exponential	-0.0332954	-0.6535455	0.5869546	1.0000000
one.percent:normal-five.percent:exponential	-0.0261657	-0.6464158	0.5940843	1.0000000
ten.percent:normal-five.percent:exponential	-0.1013200	-0.7215700	0.5189301	0.9998410
five.percent:uniform-five.percent:exponential	-0.0450894	-0.6653395	0.5751607	0.9999997
one.percent:uniform-five.percent:exponential	0.0203410	-0.5999091	0.6405911	1.0000000
ten.percent:uniform-five.percent:exponential	0.0178706	-0.6023795	0.6381207	1.0000000
ten.percent:exponential-one.percent:exponential	2.8087708	2.1885207	3.4290209	0.0000000
five.percent:normal-one.percent:exponential	-0.1252255	-0.7454756	0.4950246	0.9992373
one.percent:normal-one.percent:exponential	-0.1180958	-0.7383459	0.5021543	0.9995032
ten.percent:normal-one.percent:exponential	-0.1932500	-0.8135001	0.4270000	0.9849621
five.percent:uniform-one.percent:exponential	-0.1370195	-0.7572695	0.4832306	0.9985412
one.percent:uniform-one.percent:exponential	-0.0715890	-0.6918391	0.5486610	0.9999888
ten.percent:uniform-one.percent:exponential	-0.0740595	-0.6943095	0.5461906	0.9999855
five.percent:normal-ten.percent:exponential	-2.9339963	-3.5542464	-2.3137462	0.0000000
one.percent:normal-ten.percent:exponential	-2.9268666	-3.5471167	-2.3066165	0.0000000
ten.percent:normal-ten.percent:exponential	-3.0020208	-3.6222709	-2.3817708	0.0000000
five.percent:uniform-ten.percent:exponential	-2.9457903	-3.5660403	-2.3255402	0.0000000
one.percent:uniform-ten.percent:exponential	-2.8803598	-3.5006099	-2.2601098	0.0000000
ten.percent:uniform-ten.percent:exponential	-2.8828303	-3.5030803	-2.2625802	0.0000000
one.percent:normal-five.percent:normal	0.0071297	-0.6131204	0.6273798	1.0000000
ten.percent:normal-five.percent:normal	-0.0680245	-0.6882746	0.5522256	0.9999925
five.percent:uniform-five.percent:normal	-0.0117939	-0.6320440	0.6084561	1.0000000
one.percent:uniform-five.percent:normal	0.0536365	-0.5666136	0.6738865	0.9999988
ten.percent:uniform-five.percent:normal	0.0511660	-0.5690840	0.6714161	0.9999992
ten.percent:normal-one.percent:normal	-0.0751542	-0.6954043	0.5450958	0.9999837
five.percent:uniform-one.percent:normal	-0.0189237	-0.6391737	0.6013264	1.0000000
one.percent:uniform-one.percent:normal	0.0465068	-0.5737433	0.6667568	0.9999996

Tukey's (Missing*Distribution)	diff	lwr	upr	p adj
ten.percent:uniform-one.percent:normal	0.0440363	-0.5762137	0.6642864	0.9999998
five.percent:uniform-ten.percent:normal	0.0562306	-0.5640195	0.6764807	0.9999983
one.percent:uniform-ten.percent:normal	0.1216610	-0.4985891	0.7419111	0.9993821
ten.percent:uniform-ten.percent:normal	0.1191906	-0.5010595	0.7394406	0.9994683
one.percent:uniform-five.percent:uniform	0.0654304	-0.5548197	0.6856805	0.9999945
ten.percent:uniform-five.percent:uniform	0.0629600	-0.5572901	0.6832101	0.9999959
ten.percent:uniform-one.percent:uniform	-0.0024704	-0.6227205	0.6177797	1.0000000

References

http://www.stat.purdue.edu/ bacraig/notes1/topic15.pdf
http://www.mathstat.ualberta.ca/ wiens
Design and Analysis of Experiments, 8th Edition - Montgomery
https://sites.ualberta.ca/ ccwj/teaching/stats/doe