Tables for planning the number of subjects

Kjell Brunnström

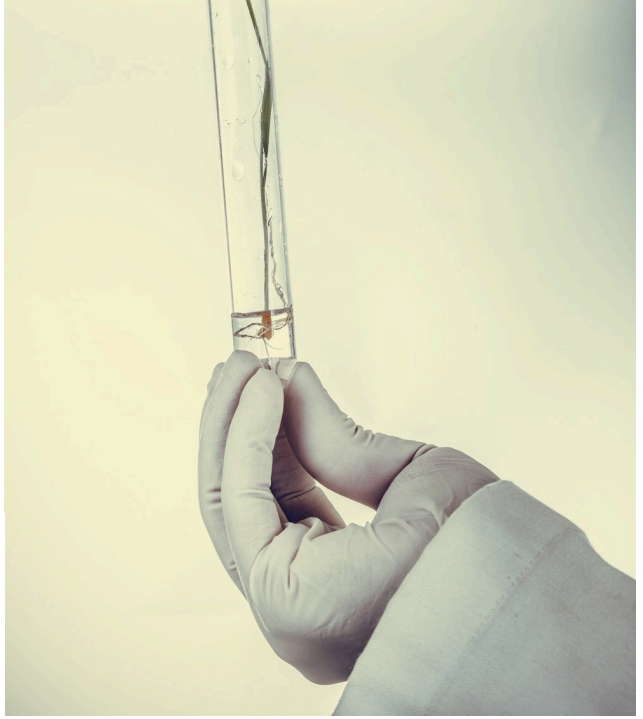
Research Institutes of Sweden

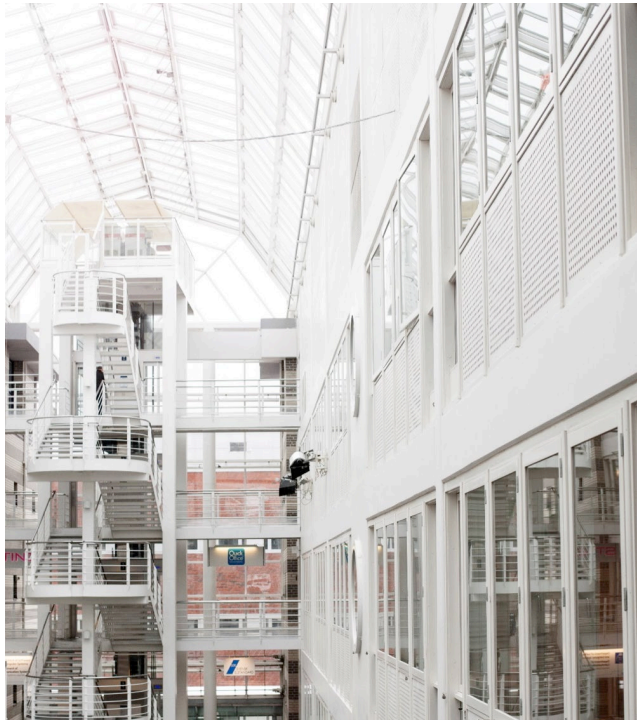
RISE ICT

Acreo

Mittuniversitetet

MID SWEDEN UNIVERSITY





Planning number of test subjects

- How can we plan in QoE experiments so that we can find the effect we would like to see?

Brunnström, K. and M. Barkowsky, *Statistical quality of experience analysis: on planning the sample size and statistical significance testing.* Journal of Electronic Imaging, 2018. **27**(5): p. 11.

Mittuniversitetet
MID SWEDEN UNIVERSITY

RI.SE

Introduction

- Type I error: to claim that there is an effect while there is none
  - The risk depends on confidence level, typically $\alpha = 0.05$ (5%)
- Type II error: to miss an effect if it is there
  - Larger risk if we safeguard more against Type I error
  - The risk to miss an effect if is there ($\beta = 0.20$)
  - This gives 4 times higher risk for Type II than Type I, which is common

Introduction

- Power (more common to use)
  - The probability to correctly conclude that there is an effect if it is there ($1 - \beta = 0.80$)
- We need to balance the experiments between Power and Type I error
- Factor influencing are expected effect size, number of samples and significance level

Introduction – Multiple comparisons

- One comparison – α (5%) risk of Type I error
- Each comparison same risk
- n comparisons: $1 - (1 - \alpha)^n$
- For 100 comparisons: 99.4 % risk of at least one Type I error
- Preplanned testing – only fixed number of comparisons
- Post-Hoc testing – all possible comparisons

Method

- Assume parametric statistical methods and underlying probability distribution is Normal
- **Within subject design**: test subjects used more than once giving a dependency between votes
  - Common case for video quality tests
  - Dependent T-test for paired samples: $t_{obs} = \frac{\mu_D - \mu_o}{\sigma_D}\sqrt{n}$
- **Between subject design**: test subjects are only used once giving independent votes
  - Could occur if a subjective experiment is repeated with the same video clips with a different panel of observers
  - Student T-test for independent samples: $t_{obs} = \frac{\mu_1 - \mu_2}{\sqrt{2}\sigma}\sqrt{n}$

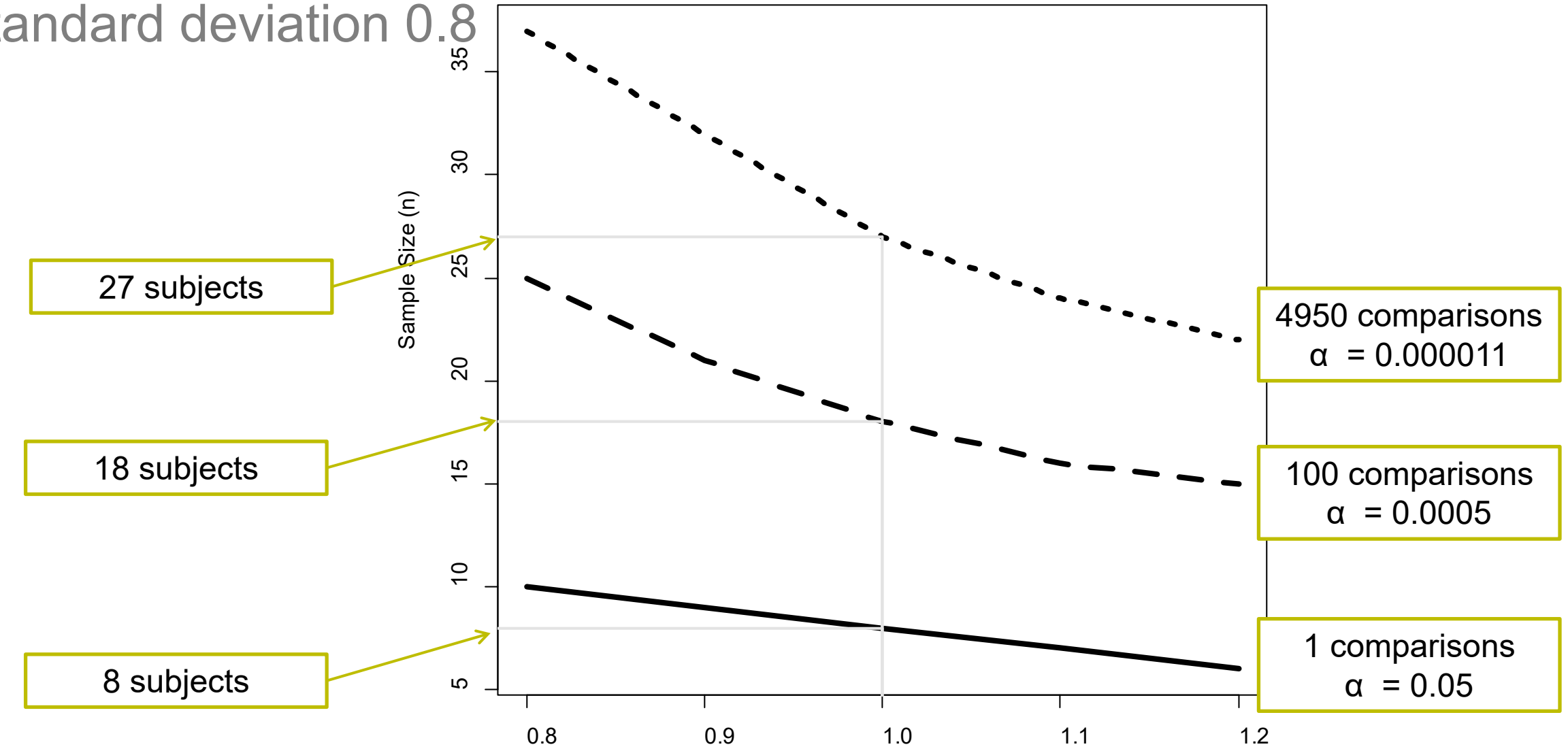
Method: Simulate influence

- Number of video clips in experiment: n_PVS = 100
- Assume pre-planned gives 100 comparisons
- Post-hoc gives 100*99/2 = 4950
- Method to compensate for multiple comparisons: Bonferroni (α/n)
- Bonferroni significance levels:
  - 0.05 (1 comparison)
  - 0.0005 (100 comparisons)
  - 0.000011 (4950 comparisons)

Method: Simulate influence

- Compute probability of significance i.e. p-value of the two T-tests
- Interesting difference MOS = 0.5 and 1.0 on five graded scale
- We have used $\sigma = 0.8$ (Typical experimental value e.g. VQEG HDTV)
- Planning the number of test subjects based on power of 0.8 i.e. $\beta = 0.20$, $\alpha = 0.05$, $\beta/\alpha = 4$

Estimation of sample size for power 0.8, MOS difference 1.0 and standard deviation 0.8



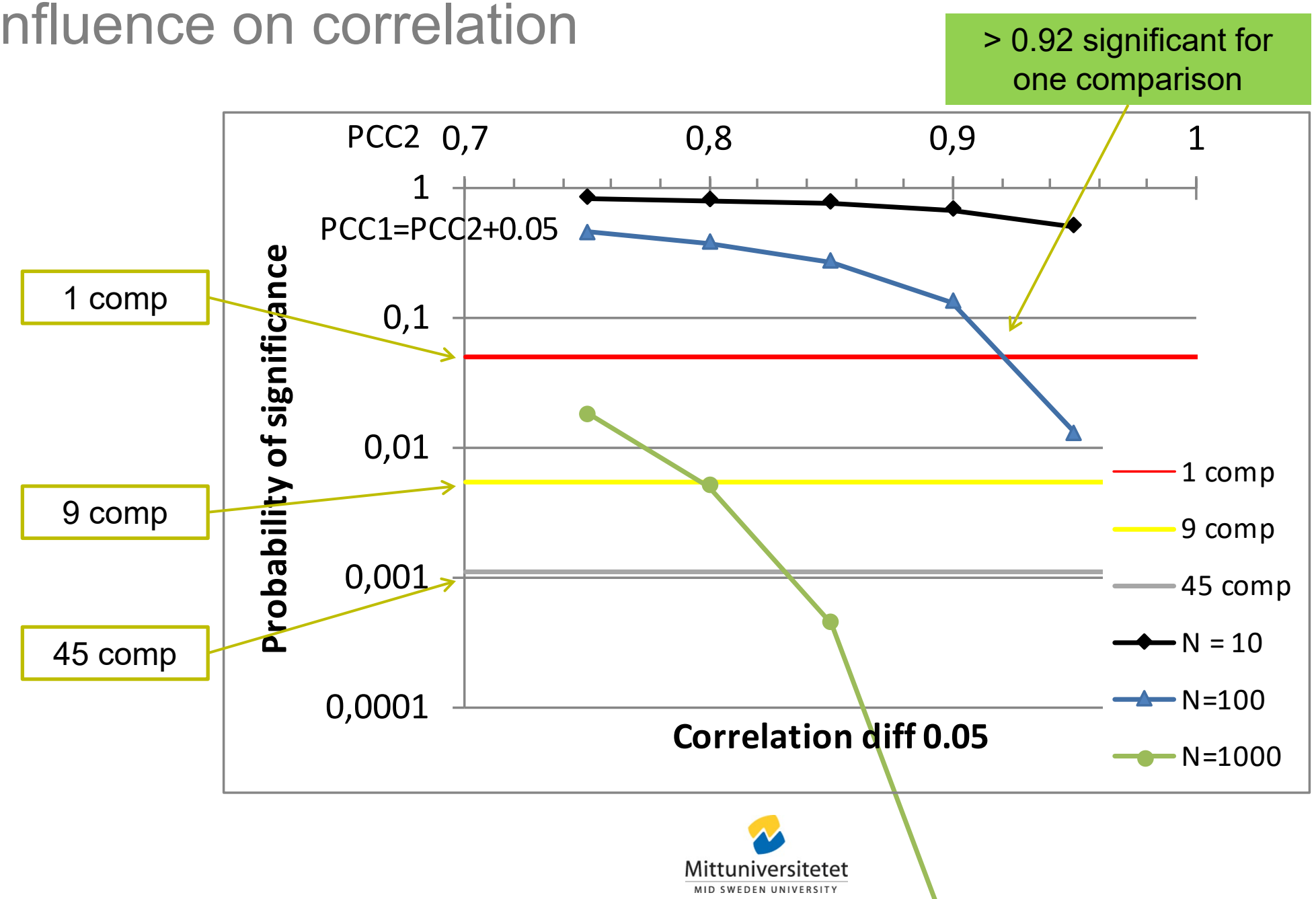
Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

The number test subjects a power of 0.8

Design type	# comparisons	α	Simulated MOS difference	Sample size Std dev 0.8	Sample size Std dev 1.0
Within	1	0.05	0.5	23	34
			1.0	8	10
	100	0.0005	0.5	54	81
			1.0	18	25
	4950	0.00001	0.5	81	121
			1.0	27	37
Between	1	0.05	0.5	42	64
			1.0	12	17
	100	0.0005	0.5	99	153
			1.0	27	41
	4950	0.00001	0.5	147	227
			1.0	41	61

Influence on correlation



Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

Effect of method for controlling Type-I error

- Many methods have been proposed in the litterature.
- Bonferroni is simple, but usually accused for beeing too conservative
- We compared a few methods

Method/experiment	No control (%)	Bonferroni (%)	Holm (%)	Tukey HSD (%)	Benjamini–Hochberg (%)	Benjamini–Yekutieli (%)
VQEG HDTV C2	41	5 (18)	5 (19)	6 (21)	31	11
VQEG HDTV C3	82	50	52	60	80	70
Adaptive streaming C2	17	1 (4)	1 (4)	0	2	0 (1)
Adaptive streaming C3	69	20	21	29	66	50

Conclusions

- Multiple comparisons should be considered and compensated for in planning and analysis of subjective QoE experiments as well as in comparison of objective quality methods.
- The number of test subjects currently recommended are too low considering this (possibly with the exception of P.913)

ITU input

- We have (RISE, Ericsson and TU Ilmenau) prepared in input to ITU SG12
- We propose to harmonize the text in BT.500, P.910 and P.913, when it comes to number of test subjects. Basically using P.913 as a starting point, but separating two cases.
  - No pre-planned number of comparison => 27 controlled env. and 37 in public env.
  - Pre-planned number of comparison (<100) => 18 and 25
- Compute a table for different conditions instead

Paired

MOSdiff/std dev	1 0.05	5 0.01	10 0.005	50 0.001	100 0.0005	500 0.0001	1000 0.00005	5000 0.00001
0.2	199	296	337	433	474	568	608	702
0.3	90	134	152	196	214	257	275	317
0.4	52	77	88	113	123	148	159	183
0.5	34	51	58	74	81	98	105	121
0.6	24	36	41	53	58	70	75	87
0.7	19	28	32	41	45	54	58	66
0.8	15	22	25	33	36	43	46	53
0.9	12	18	21	27	30	36	38	44
1	10	16	18	23	25	30	32	37
1.1	9	14	15	20	22	26	28	33
1.2	8	12	14	18	19	23	25	29
1.3	7	11	12	16	17	21	22	26
1.4	7	10	11	14	16	19	20	24
1.5	6	9	10	13	15	17	19	22
1.6	6	8	10	12	14	16	17	20
1.7	5	8	9	12	13	15	16	19
1.8	5	8	9	11	12	14	15	18
1.9	5	7	8	10	11	14	14	17
2	5	7	8	10	11	13	14	16
2.1	5	7	7	9	10	12	13	15
2.2	4	6	7	9	10	12	13	14
2.3	4	6	7	9	9	11	12	14
2.4	4	6	7	8	9	11	12	13

Number of comparisons
alpha/Number of comparisons)

MOSdiff 1 MOSdiff/std dev:1.25
Std dev 0.8

MOSdiff 0.5 MOSdiff/std dev:0.625
Std dev 0.8

Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

Paired

```
library(pwr)
mosdelta <- seq(.3,1.2,.1)
dimMOSdelta <- c("0.3","0.4","0.5", "0.6", "0.7", "0.8", "0.9", "1.0", "1.1", "1.2")
r <- length(mosdelta)
stddev <- seq(0.5,1.5,0.1)
nstddev <- length(stddev)
dimStdDev <- c("0.5", "0.6", "0.7", "0.8", "0.9", "1.0", "1.1", "1.2", "1.3", "1.4", "1.5")
alphas <- c(0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001)
dimAlphas <- c("0.05","0.0005","0.00001")


dvals <- seq(0.2,2.4,0.1)
ndvals <- length(dvals)



p <- seq(.8,.8,.1)
nalphas <- length(alphas)
samsize <- matrix(1:ndvals*nalphas, nrow=ndvals, ncol=nalphas)
for (j in 1:ndvals) {
            for (i in 1:nalphas) {
                        result <- pwr.t.test(n=NULL, d=dvals[j], sig.level=alphas[i], power=p[1], type="paired")
                        samsize[j,i] <- ceiling(result$n)
            }
}
write.csv(samsize, file="sample_size.csv")
```


Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

Independent

MOSdiff/std dev	1 0.05	5 0.01	10 0.005	50 0.001	100 0.0005	500 0.0001	1000 0.00005	5000 0.00001
0.2	394	586	668	857	938	1124	1204	1388
0.3	176	262	298	383	419	502	538	620
0.4	100	148	169	217	237	284	304	351
0.5	64	96	109	140	153	183	196	227
0.6	45	67	76	98	107	129	138	159
0.7	34	50	57	73	80	96	103	118
0.8	26	39	44	57	62	74	80	92
0.9	21	31	35	45	50	60	64	74
1	17	26	29	37	41	49	53	61
1.1	15	22	25	31	34	41	44	51
1.2	12	18	21	27	30	35	38	44
1.3	11	16	18	23	26	31	33	38
1.4	10	14	16	21	23	27	29	34
1.5	9	13	14	18	20	24	26	30
1.6	8	11	13	17	18	22	23	27
1.7	7	10	12	15	17	20	21	24
1.8	6	10	11	14	15	18	19	22
1.9	6	9	10	13	14	17	18	21
2	6	8	9	12	13	15	17	19
2.1	5	8	9	11	12	14	15	18
2.2	5	7	8	10	11	14	14	17
2.3	5	7	8	10	11	13	14	16
2.4	4	6	7	9	10	12	13	15

Number of comparisons
alpha/Number of comparisons)

MOSdiff 1 MOSdiff/std dev:1.25
Std dev 0.8

MOSdiff 0.5 MOSdiff/std dev:0.625
Std dev 0.8

Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

Independent

```
library(pwr)
mosdelta <- seq(.3,1.2,.1)
dimMOSdelta <- c("0.3","0.4","0.5", "0.6", "0.7", "0.8", "0.9", "1.0", "1.1", "1.2")
nr <- length(mosdelta)
stddev <- seq(0.5,1.5,0.1)
nstddev <- length(stddev)
dimStdDev <- c("0.5", "0.6", "0.7", "0.8", "0.9", "1.0", "1.1", "1.2", "1.3", "1.4", "1.5")
alphas <- c(0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005, 0.00001)
dimAlphas <- c("0.05","0.0005","0.00001")


dvals <- seq(0.2,2.4,0.1)
ndvals <- length(dvals)



p <- seq(.8,.8,.1)
nalphas <- length(alphas)
samsize <- matrix(1:ndvals*nalphas, nrow=ndvals, ncol=nalphas)
for (j in 1:ndvals) {
            for (i in 1:nalphas) {
                                result <- pwr.t.test(n=NULL, d=dvals[j], sig.level=alphas[i], power=p[1],  type="two.sample")
                                samsize[j,i] <- ceiling(result$n)
            }
}
write.csv(samsize, file="sample_size_2-tailed.csv")
```


Mittuniversitetet
MID SWEDEN UNIVERSITY

RI. SE

THANK YOU!

Kjell Brunnström

kjell.brunnstrom@ri.se

Research Institutes of Sweden

RISE Acreo

Mittuniversitetet

MID SWEDEN UNIVERSITY