AARHUS
UNIVERSITET

USER INTERFACES FOR EMBEDDED SYSTEMS

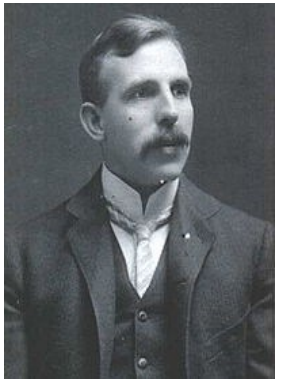
Lecture 9: Statistics

Stefan Wagner
sw@eng.au.dk

AGENDA

- Introduction
- Descriptive statistics
- Central tendency and variability
- Visualizing data with charts and graphs
- Histogram and boxplot
- t-test
- Correlations

STATISTICS AS A PHENOMENA



"If your experiment needs statistics, you ought to have done a better experiment"

–Ernest Rutherford (nuclear physicist)

"There are three kinds of lies: lies, damned lies, and statistics"

–Benjamin Disraeli (British statesman and literary figure)

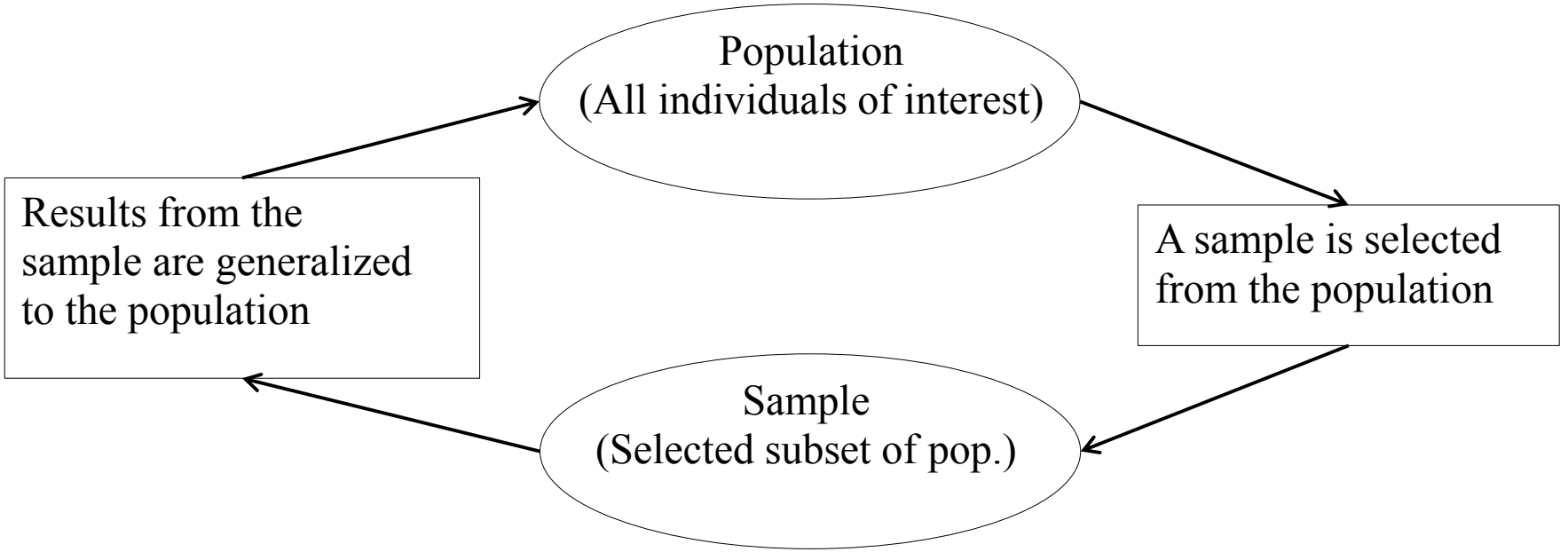
–Mark Twain (novelist)

DESCRIPTIVE STATISTICS

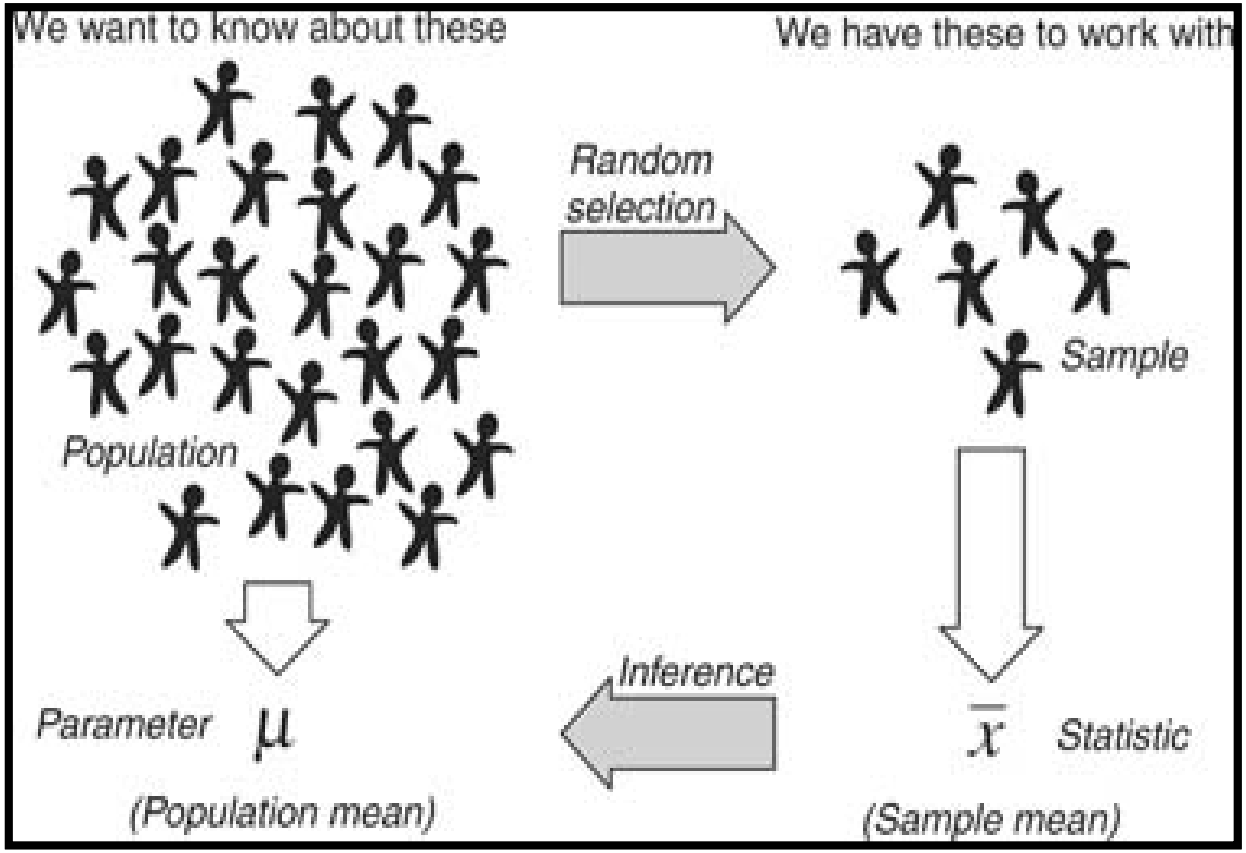
- Describes data in a **sample**
- Does not infer anything about the **population**
- Collect, classify, summarize, and present data
- Proceed to inferential statistics if
  - There is **enough data** to infer about the population
  - Your data is representative of the population
- Remember: With statistics
  - You do **not prove** a correlation or causality
  - You merely render a correlation **probable**

POPULATIONS AND SAMPLES (1/2)

- Statistics are calculated from the samples
- Inferences are made from the **sample** to the **population**
- In descriptive statistics we **only look at samples**



POPULATIONS AND SAMPLES (2/2)



NOTATION AND TERMINOLOGY

- A **statistical parameter** is a single measure of some attribute of a **population**
  - e.g. artithmetic mean
- A **statistic** is a single measure of some attribute of a **sample**
  - e.g. arithmetic mean
- E.g, the sample mean is a **statistic** which **estimates** the population mean, which is a **statistical parameter**

TYPES OF DATA

› Nominal data

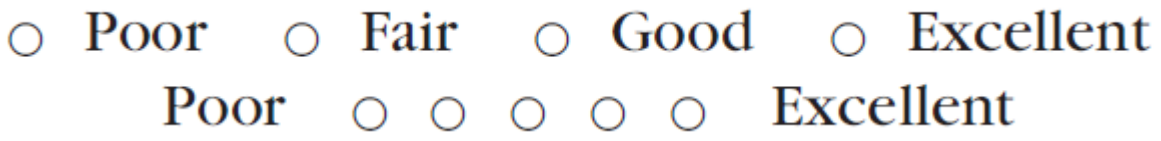
› unorderd data, cannot be compared, apples vs bananas

› Ordinal data

› Ordered data that can be compared, task performance 10 seconds vs 30 seconds
› Poor vs good, relative ranking, is number 3 is better than 6, but is it 2 times better?

› Interval data

› From a scale 1-10



› Range data

› Interval data with a specified offset (e.g. all tests starts with 0 seconds to infinity)

SAMPLE SIZE & CONFIDENCE INTERVALS

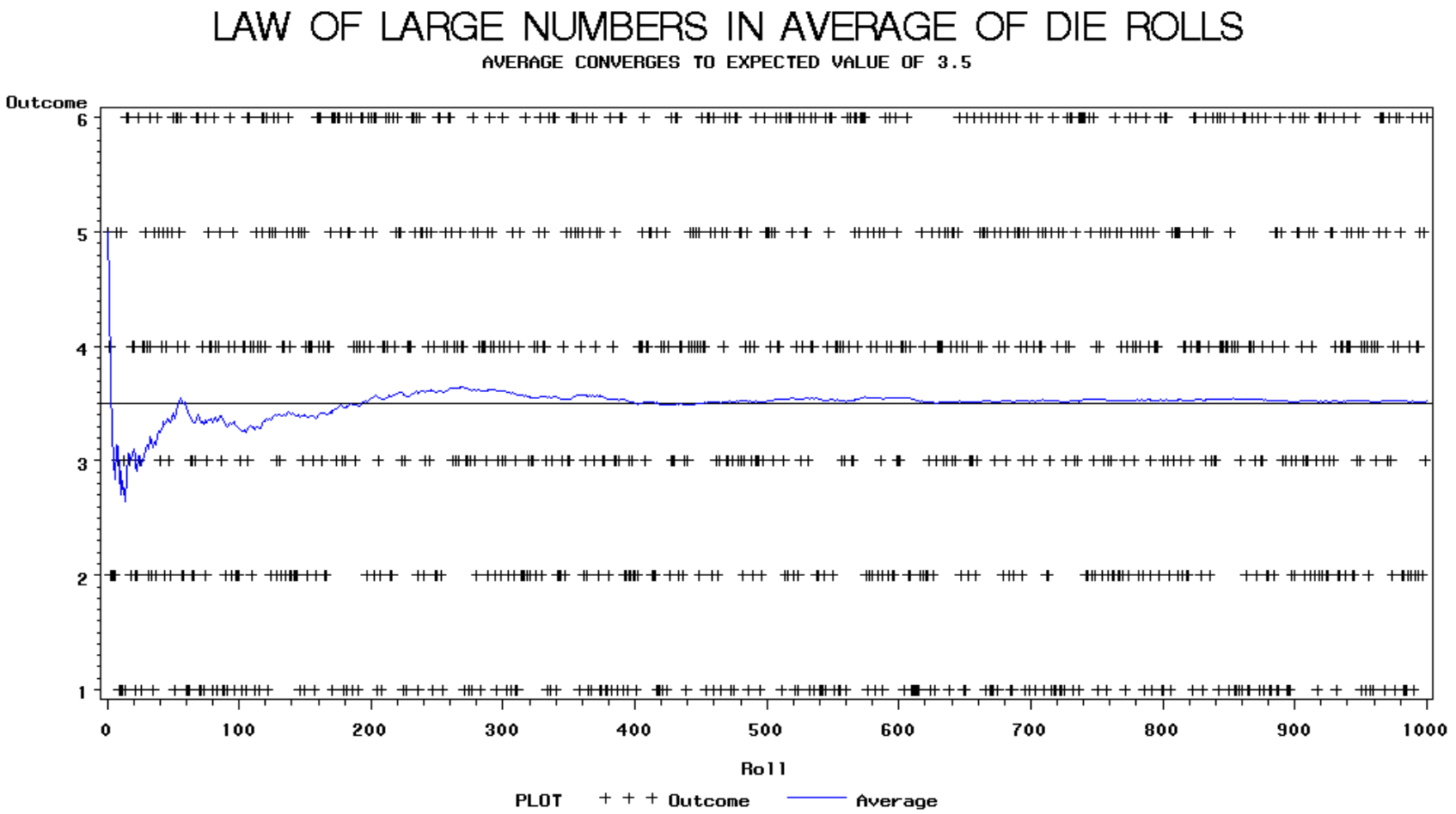
- Central question:
  - how many participants are needed"?
- Sample size:
  - Three to four participants might be enough for some studies
  - Other studies might need thousands of participants
- Confidence intervals indicates how results can be generalized
  - E.g. with 95 % confidence we can say that between 36 and 98 % of larger population will be able to successfully complete a given task when using a sample n=5
  - With n=100 the 95% confidence interval numbers are 71% to 86%. Larger samples, higher confidence.

Table 2.1 Example of How Confidence Intervals Change as a Function of Sample Size

Number Successful	Number of Participants	Lower 95% Confidence	Upper 95% Confidence
4	5	36%	98%
8	10	48%	95%
16	20	58%	95%
24	30	62%	91%
40	50	67%	89%
80	100	71%	86%

Note: These numbers indicate how many participants in a usability test successfully completed task and the confidence interval for that mean completion rate in the larger population.

LAW OF LARGE NUMBERS: DIE ROLL EXAMPLE



COUNTER BALANCING

› Avoid or negate a learning effect by using counter balancing

Table 2.2 Example of How to Counterbalance Task Order for Four Participants and Four Tasks

Participant	First Task	Second Task	Third Task	Fourth Task
P1	T1	T2	T3	T4
P2	T3	T1	T4	T2
P3	T2	T4	T1	T3
P4	T4	T3	T2	T1

APPLIED DESCRIPTIVE STATISTICS

- Descriptive statistics:
  - are essential for any interval or ratio-level data. Descriptive statistics, as the name implies, describe the data without saying anything about the larger population.
- Inferential statistics:
  - let you draw some conclusions or infer something about a larger population above and beyond your sample.
- Descriptive statistics are very easy to calculate using most statistical software packages.
- Inferential statistics are more difficult and error prone.

APPLIED DESCRIPTIVE STATISTICS: COMMON MEASURES

	A	B	C	D	E	F
1	**Participant**	**Task time**		**Task time**		
2	P1	34		Mean	35.0833333	Average of Sample (S)
3	P2	33		Standard error (sample std)	3.24611267	Stdev/Sqrt(Count) of S
4	P3	28		Median	33.5	Midway point in S
5	P4	44		Mode	22	Most common value in S
6	P5	46		Standard deviation	11.2448641	Spread relative to mean of S
7	P6	21		Sample variance	126.446970	Spread relative to mean of S
8	P7	22		Kurtosis	-1.32152597	Peakedness of PD of S
9	P8	53		Skewness	0.25144172	Asymmetry of PD of S
10	P9	22		Range	32	Max – Min of S
11	P10	29		Minimum	21	Min value of elements in S
12	P11	39		Maximum	53	Max value of elements in S
13	P12	50		Sum	421	Sum of element values in S
14				Count	12	Number of elements in S
15				Confidence level (95%)	6.36226393	Reliability of an estimate about P based on S

Central tendency measures | Variability measures | Other measures

[Tullis08]

CENTRAL TENDENCY

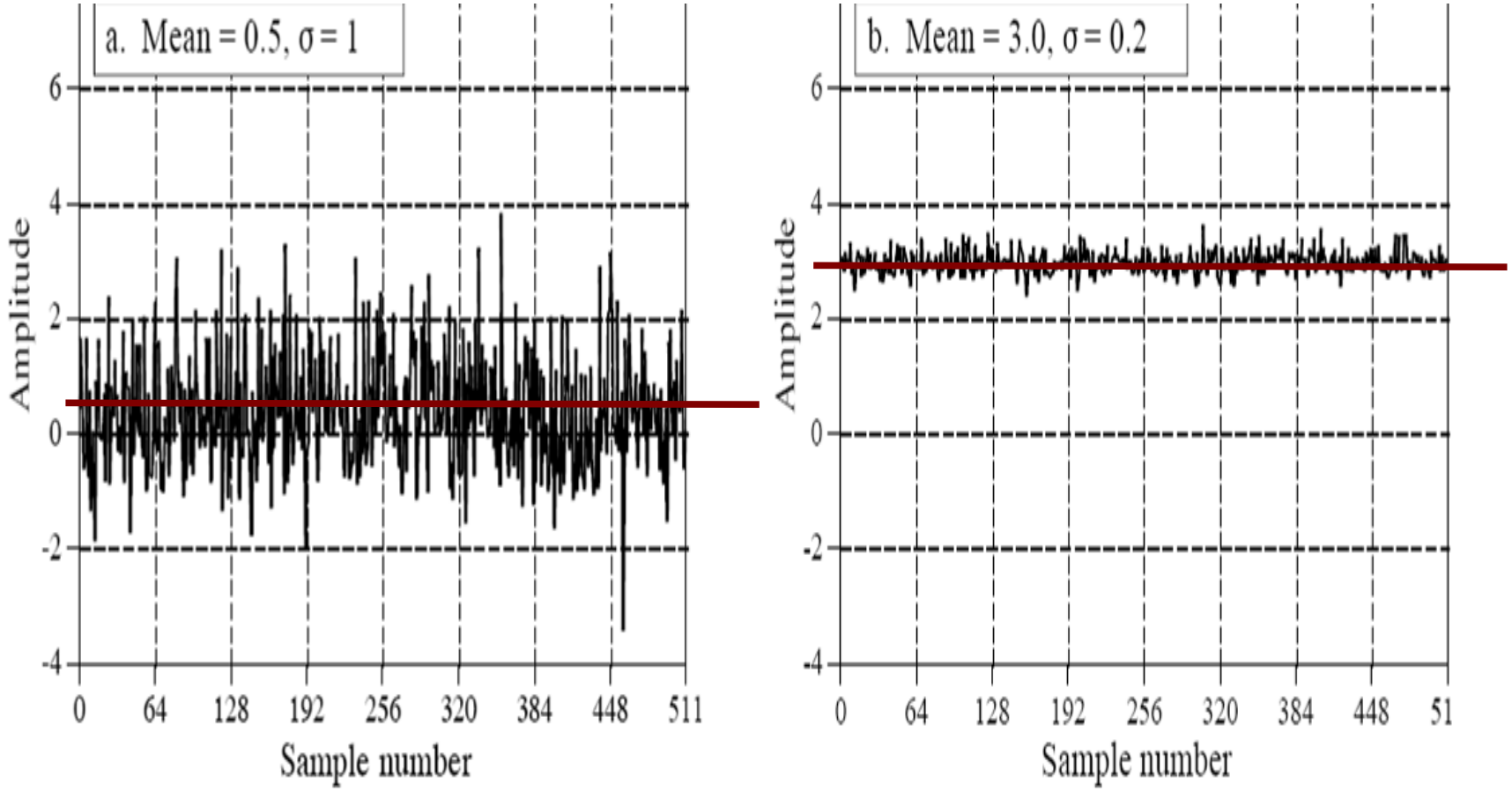
- Clustering around a central measure
  - Mean, Median, and Mode
- Median location = 50th percentile
- Mode, i.e. the most commonly occuring score

	A	B	C	D	E
1	**Participant**	**Task time**		**Task time**	
2	P1	34		Mean	35.0833333
3	P2	33			
4	P3	28		Median	33.5
5	P4	44		Mode	22

VARIABILITY

- Spread of data “measured” as “distance” to central tendency
  - The larger the spread - the more data you need for higher confidence
- Sample range
  - distance between max and minimum
  - Identifies outliers and coding errors (e.g. if scale is 1-5, a zero value is wrong)
- Sample variance
  - estimator for population variance relative to the mean value
- Sample standard deviation –
  - estimator for population standard deviation – e.g. an average of 10 seconds
- Confidence interval (in Excel)
  - = confidence (alpha, standard deviation, sample size)
  - alpha = significance level (often 0,05 - 0,01): meaning how certain do we want to be

EXAMPLE: SAMPLE MEAN AND SAMPLE STANDARD DEVIATION



[Smith97]

EXERCISE

- 45 minutes Exercise
- 1) Recruit four(+) participants from your peer group
- 2) Have them solve a specified task from your scenario using your two prototypes
- 3) Record for each participant:
  - 3.1) Time to solve task using prototype 1
  - 3.2) Time to solve task using prototype 2
- 4) Enter the data into excel
- 5) Calculate the mean, median, mode for each column (=AVERAGE, =MEDIAN, =MODE)
- 6) Calculate the standard deviation, and the confidence interval (=STDEV.S(C5:C8))
- 7) Calculate the 95% Confidence
- 8) Calculate the Confidence Interval

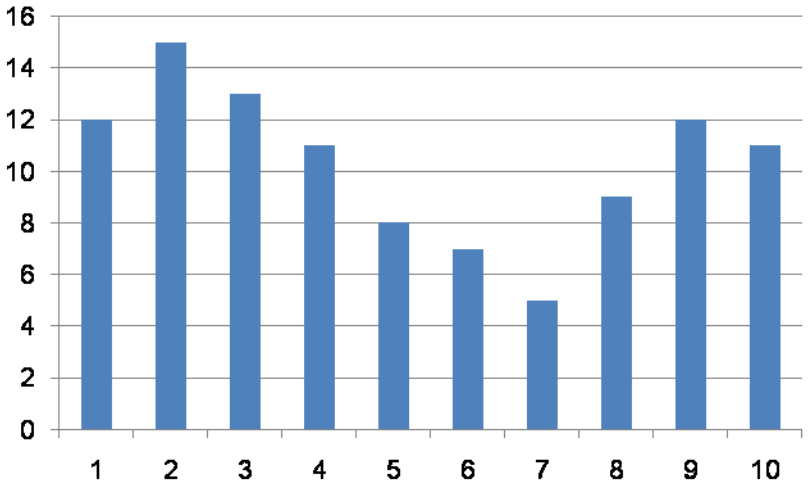
	Time to perform task 1 (s)	
Participant	**Prototype 1**	**Prototype 2**
P1	20	15
P2	23	23
P3	19	18
P4	21	15
Maximum	23	23
Minimum	19	15
Mean/Average	20,75	17,75
Median	20,50	16,50
Mode	#N/A	15,00
Std. Deviation	1,71	3,77
Confidence 95% *)	1,67	3,70
***) meaning that:**		
with 95% certainty		
data are in the range	22,42	21,45
to max	24,67	26,70

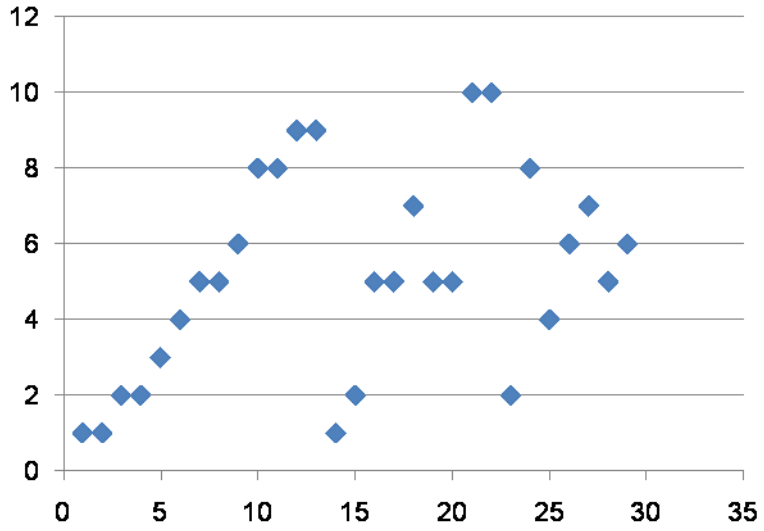
VISUALIZING DATA IN GRAPHS

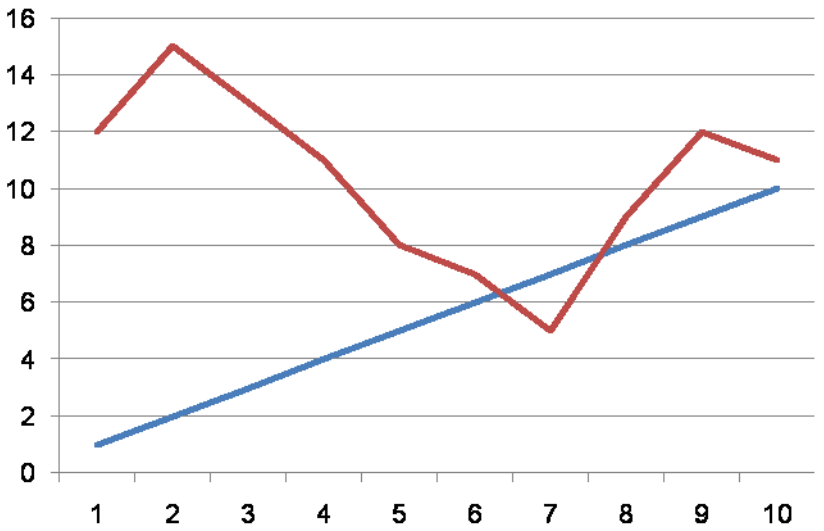
- Precision and clarity is a **virtue** and **necessity** in any field of science

- Please remember
  - Label diagrams and axes
  - Always remember units (SI)
  - Do not imply more precision than data allows, e.g. decimals
  - Careful if using colors and fancy diagram types, e.g. 3D
  - Do not overload diagrams with information
  - Use confidence intervals whenever possible

HISTOGRAMS, GRAPHS, AND PLOTS

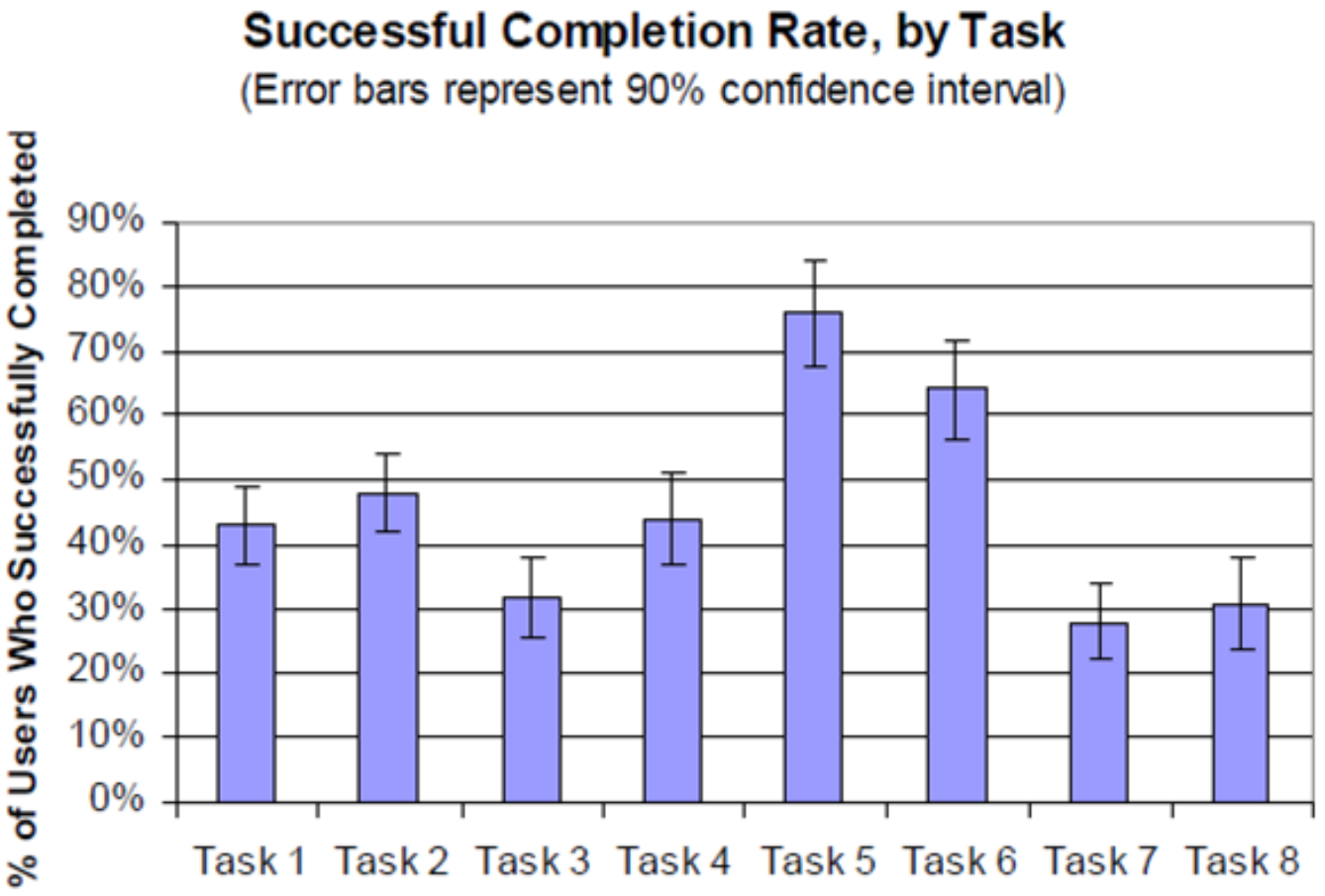






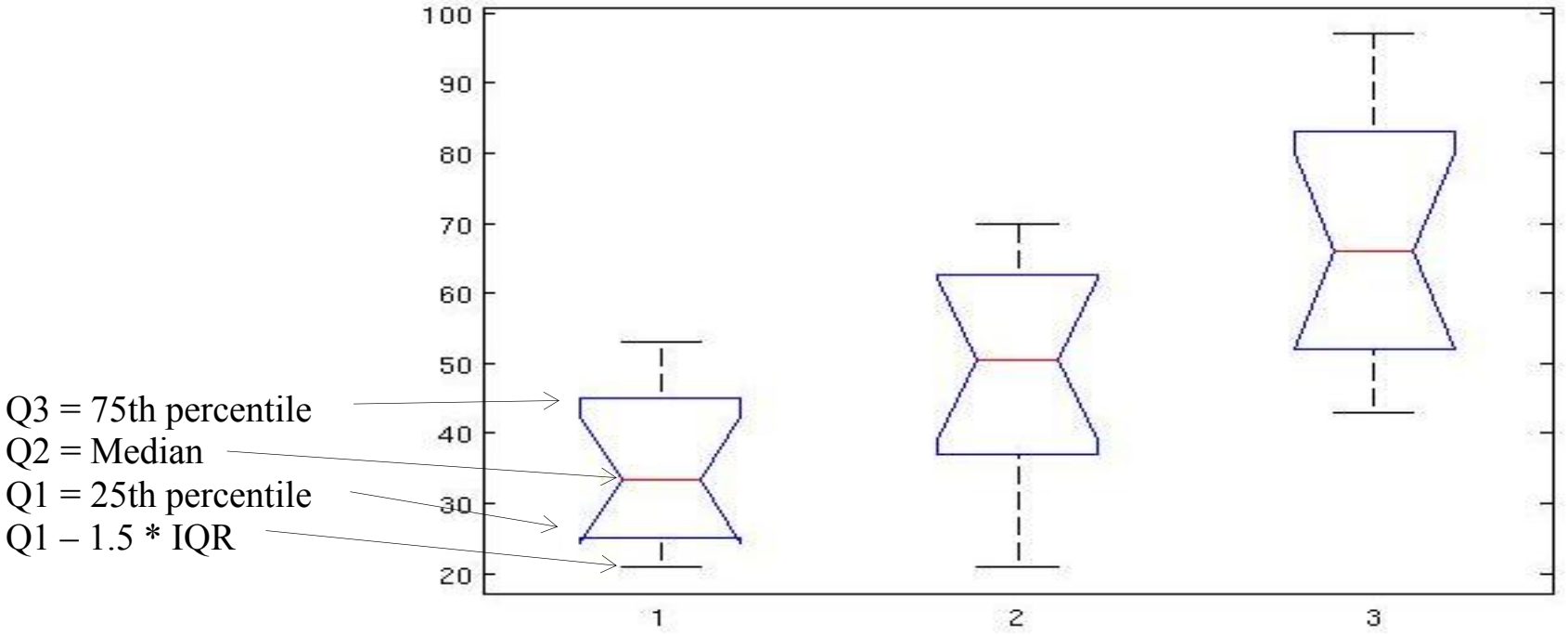
What is wrong with these?

GRAPHS WITH CONFIDENCE INTERVALS



BOXPLOT

- E.g. comparing means of three independent samples
- Boxplot (overview): Design A Design B Design C



TOP 10 MISTAKES IN DATA GRAPHS

1. Not labeling axes or units
2. Implying more precision in your data than it deserves
3. Not showing confidence intervals when you can
4. Not starting a bar graph at lowest possible y-axis value (usually 0)
5. Using a line graph when it should be a bar graph
6. Using 3D when it doesn’t add any value
7. Trying to include too much
8. Poor labeling of pie charts
9. Using colors as the only way to convey information
10. Not knowing when to use stacked bar graphs

[Tullis08]

SUGGESTED SOFTWARE

- Excel, Open Office Calc and other spreadsheets
- Matlab (ENG and ASE uses this)
- SPSS (Statistical Package for the Social Sciences)
- STATA (Aarhus University Hospital uses this)
- STATISTICA
- R (free)
- GNU Octave (free)

COMPARING SAMPLES

› Is prototype 1 more efficient than 2?

› Just compare means and medians for simple comparison
› Use statistical tests for comparing means with greater power
› For small populations (n < 30) e.g. use t-test
› For larger populations use z-test (n > 29)
› Comparing two samples use t-test, more than two use ANNOVA

› Independent samples

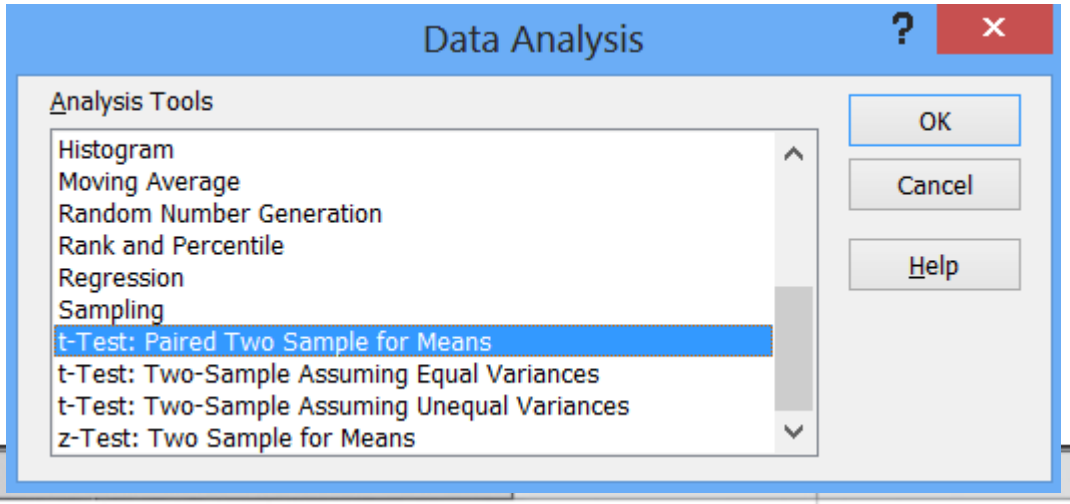
› If you compare two different groups (e.g. novice users to experienced users)
› Use an unpaired t-test

› Dependent samples

› If you compare the same group of participants – using two different prototypes
› Use a paried t-test

AARHUS
UNIVERSITET

PAIRED T-TEST



	A	B	C	D
1	Participant	Design A_SUS	Design B_SUS	
2	P1	80	48	
3	P2	88	55	
4	P3	76	53	
5	P4	90	80	
6	P5	93	81	
7	P6	67	51	
8	P7	68	61	
9	P8	55	41	
10	P9	77	55	
11	P10	71	57	
12	P11	88	59	
13	P12	80	44	
14				

t-Test: Two Samples for Means

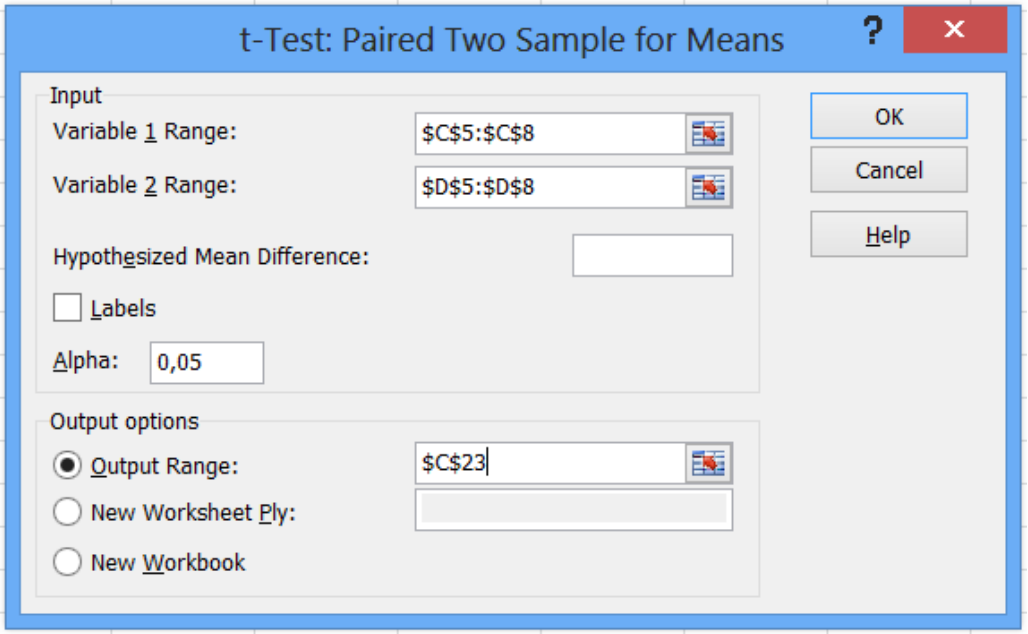
	Design A_SUS	*Design B_SUS*
Mean	77.7500000	57.0833333
Variance	125.4772727	153.7196970
Observations	12.0000000	12.0000000
Pearson Correlation	0.6521213	
Hypothesized Mean Difference	0.0000000	
df	11.0000000	
t Stat	7.2295917	
P(T<=t) one-tail	0.0000084	
t Critical one-tail	1.7958848	
P(T<=t) two-tail	0.0000169	
t Critical two-tail	2.2009852	

UNPAIRED T-TEST

	A	B	C	D	E	F
1	Expert_time	Novice_time		t-Test: Two Samples Assuming Equal Variance		
2	34	45				
3	33	48			*Expert_time*	*Novice_time*
4	28	53		Mean	35.08333333	49.33333333
5	44	66		Variance	126.4469697	229.6969697
6	46	67		Observations	12	12
7	21	35		Pooled Variance	178.0719697	
8	22	39		Hypothesized Mean Difference	0	
9	53	21		df	22	
10	22	34		t Stat	-2.61572876	
11	29	55		P(T<=t) one-tail	0.007892632	
12	39	59		t Critical one-tail	1.717144335	
13	50	70		P(T<=t) two-tail	0.015785265	
14				t Critical two-tail	2.073873058	

T-TEST IN EXCEL

	Time to perform task 1 (s)	
Participant	**Prototype 1**	**Prototype 2**
P1	20	15
P2	23	23
P3	19	18
P4	21	15
Maximum	23	23
Minimum	19	15
Mean/Average	20,75	17,75
Median	20,50	16,50
Mode	#N/A	15,00
Std. Deviation	1,71	3,77
Confidence 95% *)	1,67	3,70
***) meaning that:**		
with 95% certainty		
data are in the range	22,42	21,45
to max	24,67	26,70
t-test		



T-TEST

So given the assumption (hypothesis) – that there is no difference between tasks times between prototype 1 and 2, there is p = 0.07 that this is the case.
p should be lower than 0.05 for significance (95%) , and p < 0.01 (99%) for high significance.
Thus, data are not significantly different … but close

	Time to perform task 1 (s)	
Participant	**Prototype 1**	**Prototype 2**
P1	20	15
P2	23	23
P3	19	18
P4	21	15
Maximum	23	23
Minimum	19	15
Mean/Average	20,75	17,75
Median	20,50	16,50
Mode	#N/A	15,00
Std. Deviation	1,71	3,77
Confidence 95% *)	1,67	3,70
***) meaning that:**		
with 95% certainty		
data are in the range	22,42	21,45
to max	24,67	26,70

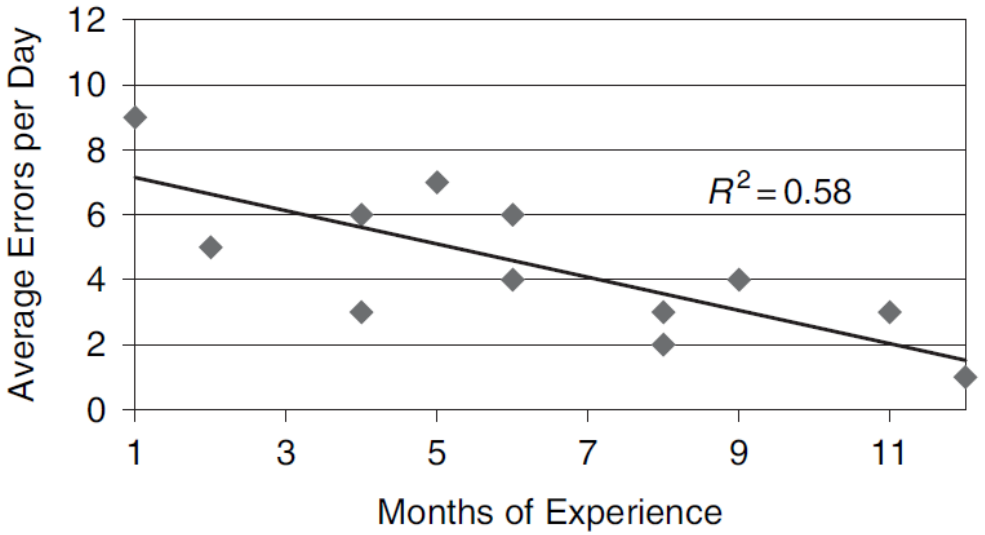
t-Test: Paired Two Sample for Means

	Prototype 1	*Prototype 2*
Mean	20,75	17,75
Variance	2,92	14,25
Observations	4,00	4,00
Pearson Correlation	0,66	
Hypothesized Mean Difference	0,00	
df	3,00	
t Stat	2,04	
P(T<=t) one-tail	0,07	
t Critical one-tail	2,35	
P(T<=t) two-tail	0,13	
t Critical two-tail	3,18	

CORRELATIONS

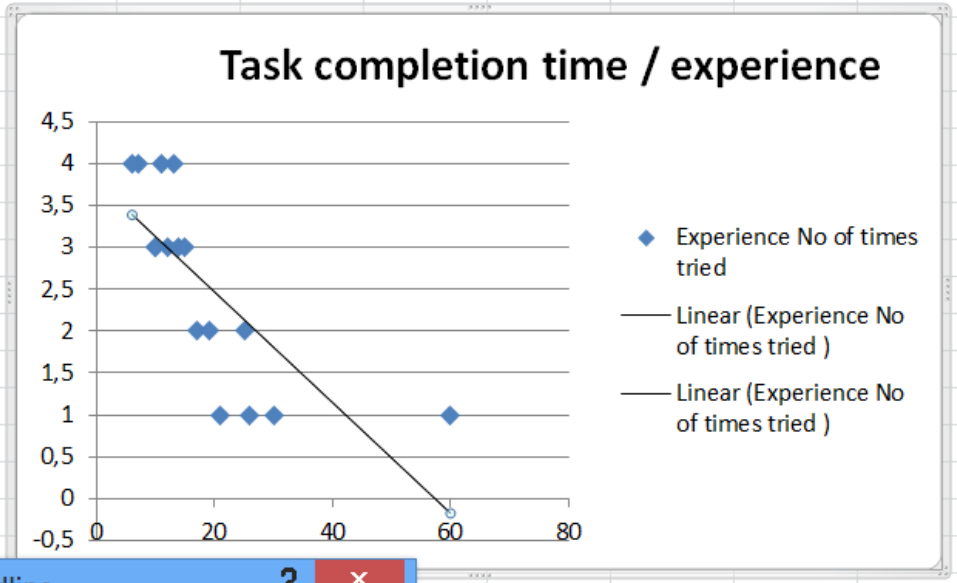
› Investigating whether parameters has a relationship

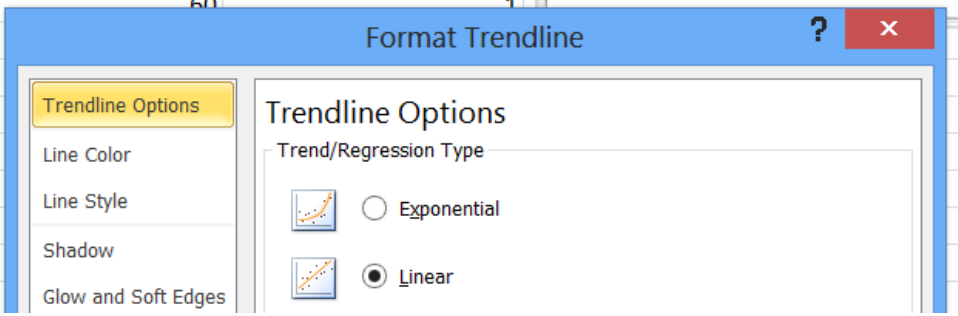
	A	B	C	D	E	F	G
1	**Participant**	**Months of Experience**	**Errors**			*Months of Experience*	*Errors*
2	P1	6	4		Months of Exp	1	
3	P2	6	6		Errors	−0.76160463	1
4	P3	8	3				
5	P4	5	7				
6	P5	4	3				
7	P6	12	1				
8	P7	11	3				
9	P8	1	9				
10	P9	9	4				
11	P10	8	2				
12	P11	4	6				
13	P12	2	5				



ADDING TREND LINES / REGRESSION ANALYSIS

	Task completion time	Experience
Particpant	Time in [s]	No of times tried
P1	30	1
P2	25	2
P3	12	3
P4	10	3
P5	13	4
P6	14	3
P7	15	3
P8	17	2
P9	19	2
P10	21	1
P11	11	4
P12	10	3
P13	7	4
P14	6	4
P15	26	1
P16	60	1





REFERENCES

- [Tullis08] Tullis and Albert, Measuring the User Experience, 2008
- [Sharp02] Sharp, Rogers, and Preece, Interaction Design, 2002
- [Nielsen94] Jakob Nielsen, Usability Engineering, 1994
- [Rubin94] Jeffrey Rubin, Handbook of Usability Testing, 1994
- [Garret02] J.J. Garret, The Elements of User Experience, 2002
- [IDBook11] www.id-book.com
- [Popper68] Popper, K. R. The Logic of Scientific Discovery. London: Hutchinson
- [Smith97] S.W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, 1997