

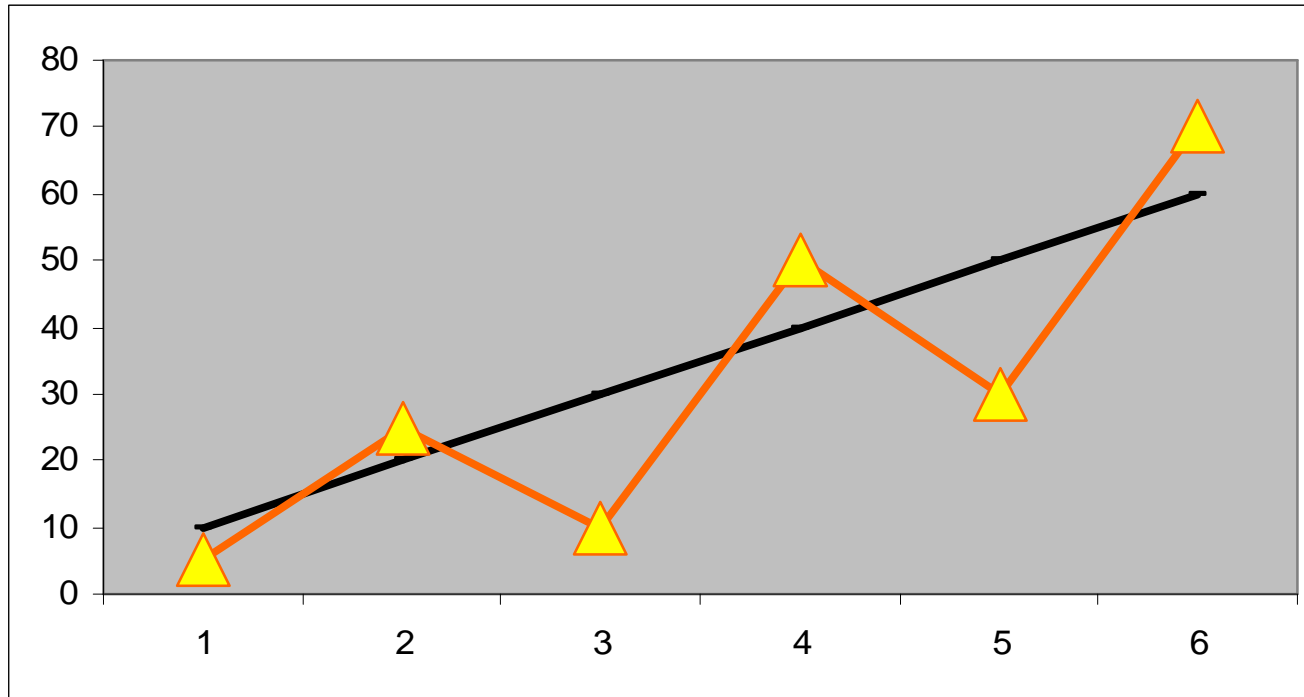
Background on statistics and research design

Broad design classes

- All designs seek to identify functions relating variables to one another.
- Correlational designs measure all the variables
- Experimental designs create the variance (manipulate) the independent variable(s) and measure the dependent variable(s)
- The two designs DO NOT differ in the statistics they employ
 - t-test comparing boys and girls in their need-for achievement = correlational design
 - Pearson correlation between the (manipulated) study time and memory = experimental design

Functions vs. measures of association

- Association measures tell us what proportion of the dependent variable's variance is explained by the independent variable(s) USING A GIVEN FUNCTION.
- Linear functions are commonly used but the reason is mathematical convenience, not psychological plausibility.



One may use a function-less measure of association (coloured line) which will show high association, or may alternatively use a function-dependent (linear, in this case) measure, which will show less association. Researchers usually prefer to sacrifice precision for generality and choose the function-dependent measure. Using a function allows one to interpolate, extrapolate, and to generate new functions based on known functions.

Effect size vs. statistical significance

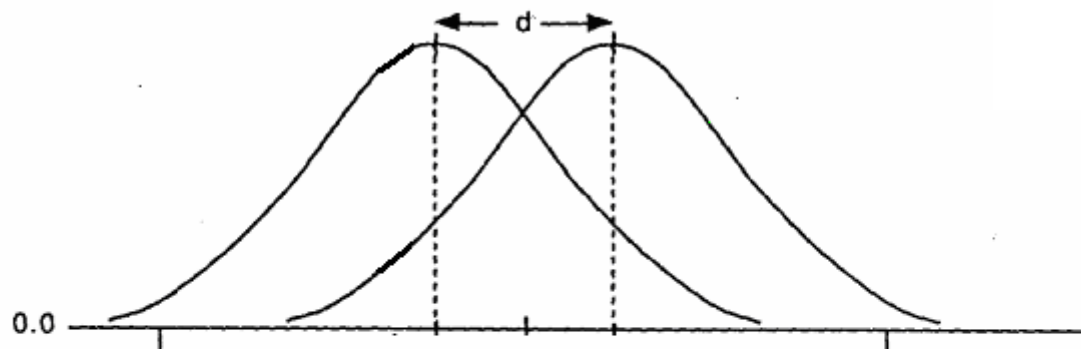
- Measures of association are also called effect sizes.
- Statistical significance depends on the effect size and on the sample size. Consequently, small effects can be significant if the sample is sufficiently large and large effects may not be significant if the sample size is small.
- Some authors therefore suggest that statistical significance may not be very important.

Effect size measures

$$\text{Cohen's } d = \frac{M_1 - M_2}{\sigma_{pooled}}$$

$$r = \sqrt{\frac{d^2}{d^2 + 4}}$$

$$d = \frac{2r}{\sqrt{1 - r^2}}$$



(expressed in units of variance)

Linking effect size to statistical significance

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

$$r = \sqrt{\frac{\chi^2(1)}{N}}$$

$$r = \sqrt{\frac{F}{F + df_{error}}}$$

$$r = \frac{Z}{\sqrt{N}}$$

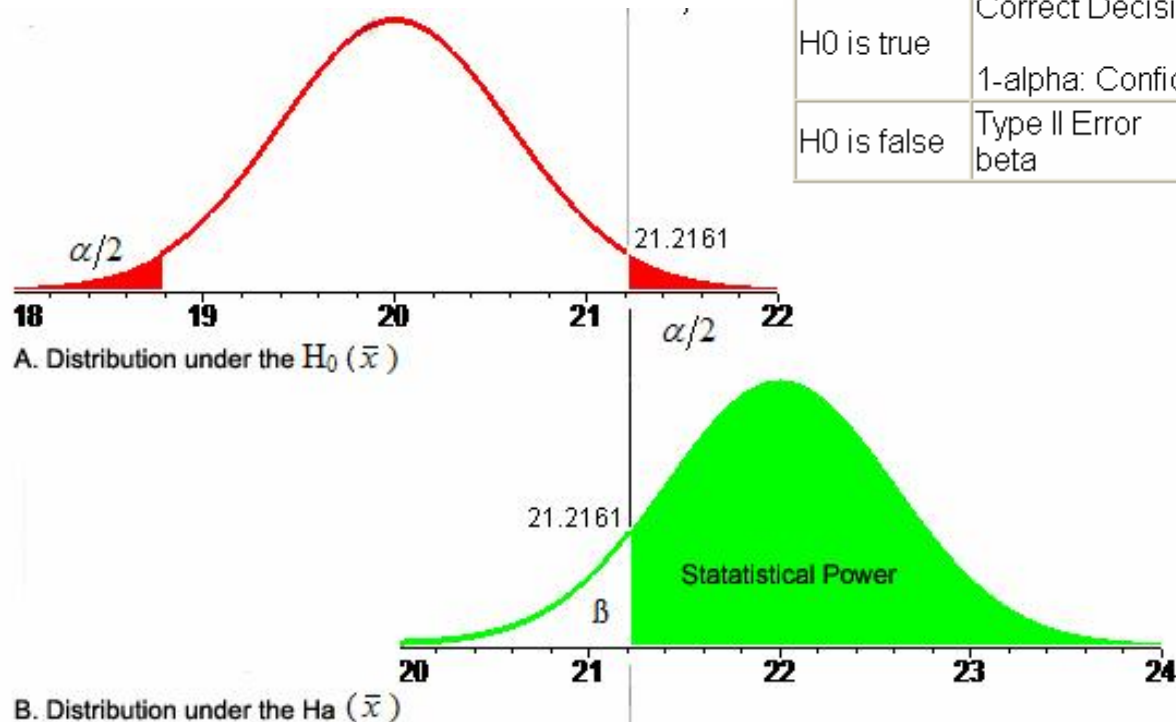
It is easy to see from the last equation that

Significance = Effect size * sample size

$$Z = r \sqrt{N}$$

Hypothesis testing and statistical power

	Do not reject H0	Reject H0
H0 is true	Correct Decision 1-alpha: Confidence Level	Type I Error Size of a test alpha: Significance level
H0 is false	Type II Error beta	Correct Decision 1-beta: Power of a test

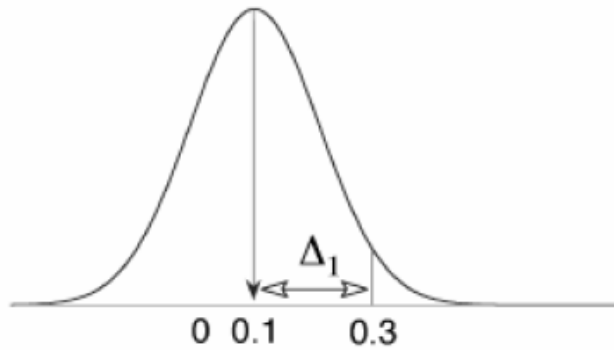


Why do we need to know the statistical power

- A-priori: When we plan an experiment and need to know the required sample size (to obtain Power=.80, usually)
- A-posteriori, when we failed rejecting the null hypothesis and we wish to endorse this hypothesis
- Power is a function of reliability and sample size.
 - In correlational designs, power is a positive function of sample heterogeneity.
 - In experimental designs, power is a negative function of sample heterogeneity
 - Reliability refers to BOTH the consistency in which the manipulation has been applied and the consistency of the dependent measure
 - In experimental designs, within subjects designs usually provide more power (=fewer subjects for the same power) as compared to between subjects designs.
 - Discuss potential problems of within subjects designs

A-posteriori p-value vs. p-rep

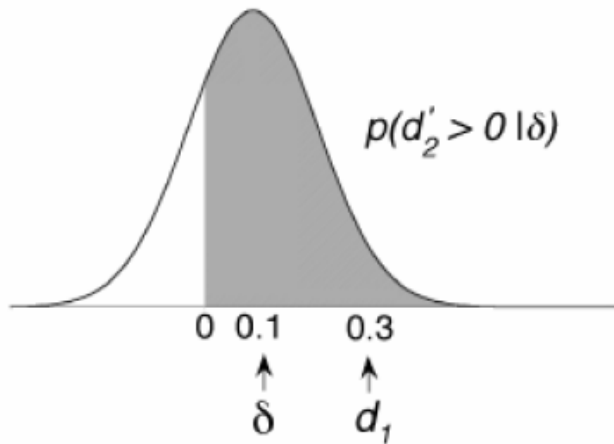
- A-posteriori p-value is the probability of finding a sample showing an effect of size X , given that the population effect is zero (unless H_0 is different which is rare).
- P-rep is based on the same data but it reflects the more intuitive notion of the probability that the next sample will also show a non-zero effect in the same direction as found in the present sample.



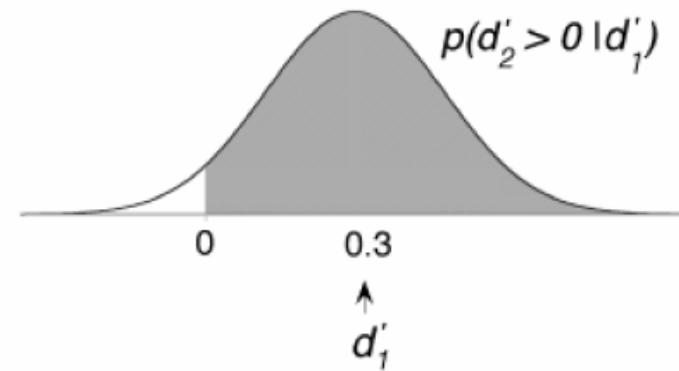
The population effect size is $d=0.01$

The sample has $d=0.03$

Probability Density



A-priori p-rep, knowing that the population effect size is $d=0.01$



A posteriori p-rep, assuming that the population effect size is the same as that in the sample, i.e. $d=0.03$

Effect Size

Hypothesis testing vs. confidence intervals

“... hypothesis testing is primarily designed to obliquely address a restricted, convoluted, and usually uninteresting question—Is it not true that some set of population means are all equal to one another?—whereas confidence intervals are designed to directly address a simpler and more general question—*What are the population means?*” (Loftus & Masson, 1994)

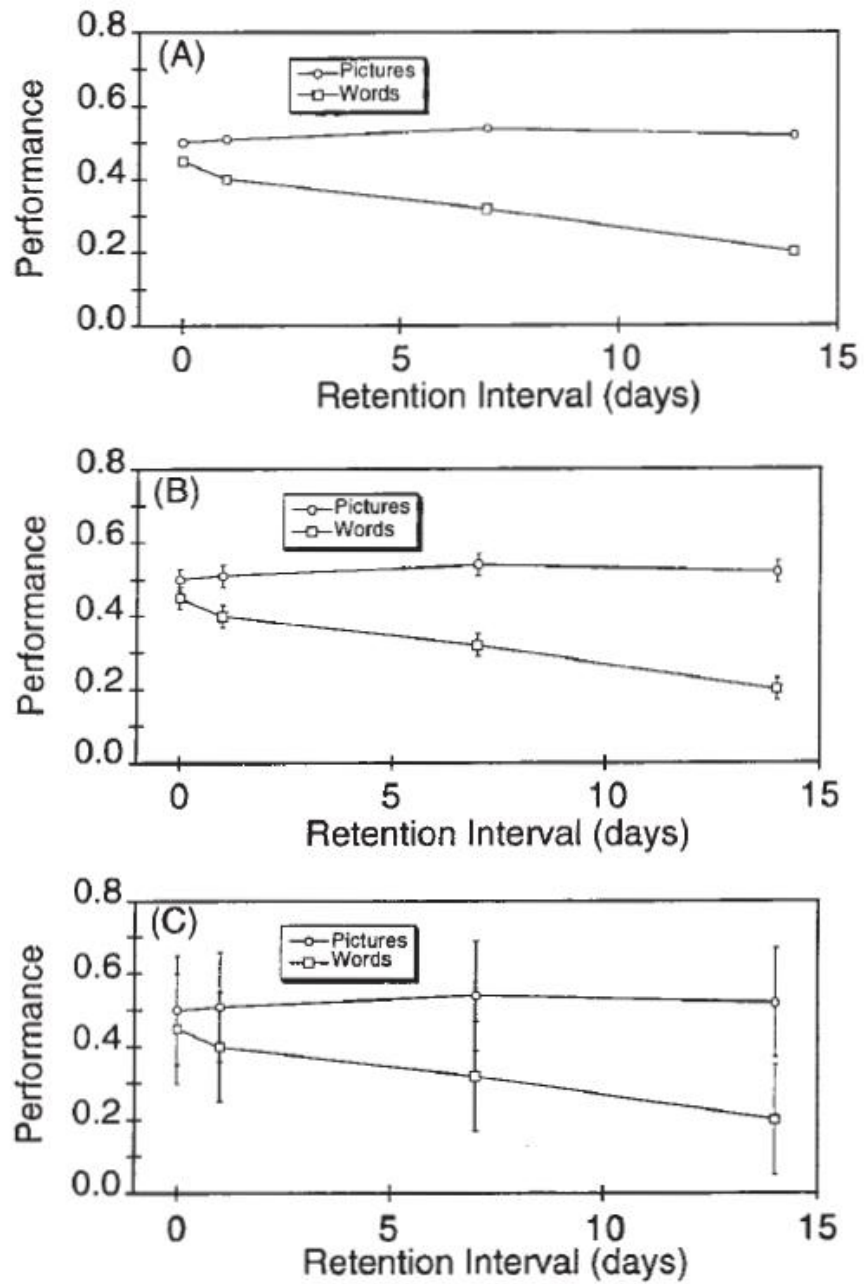
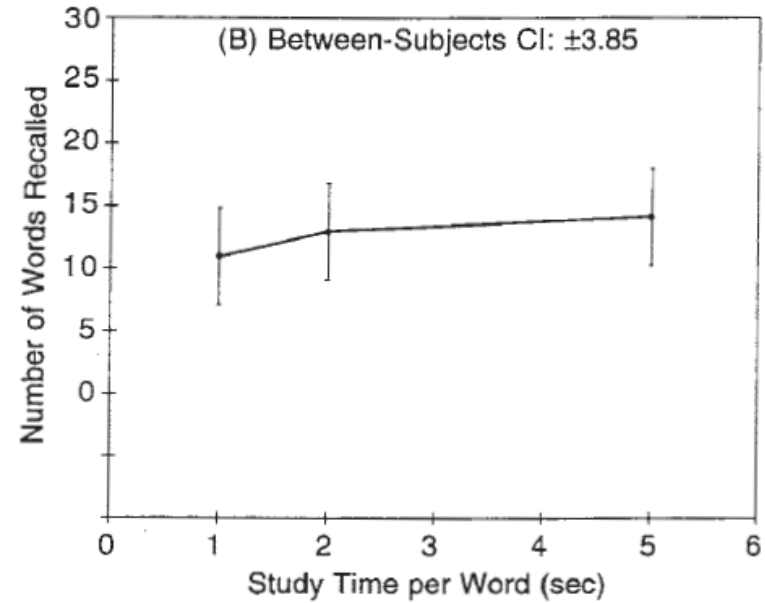
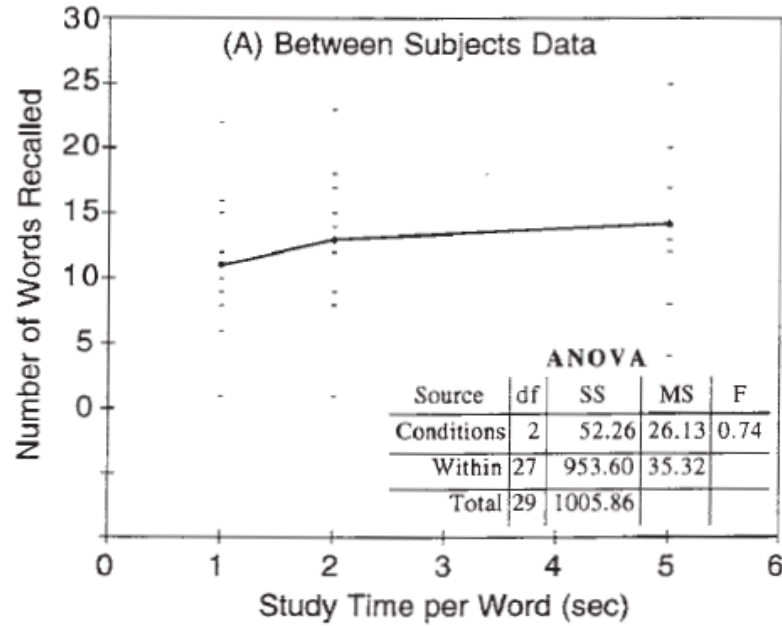
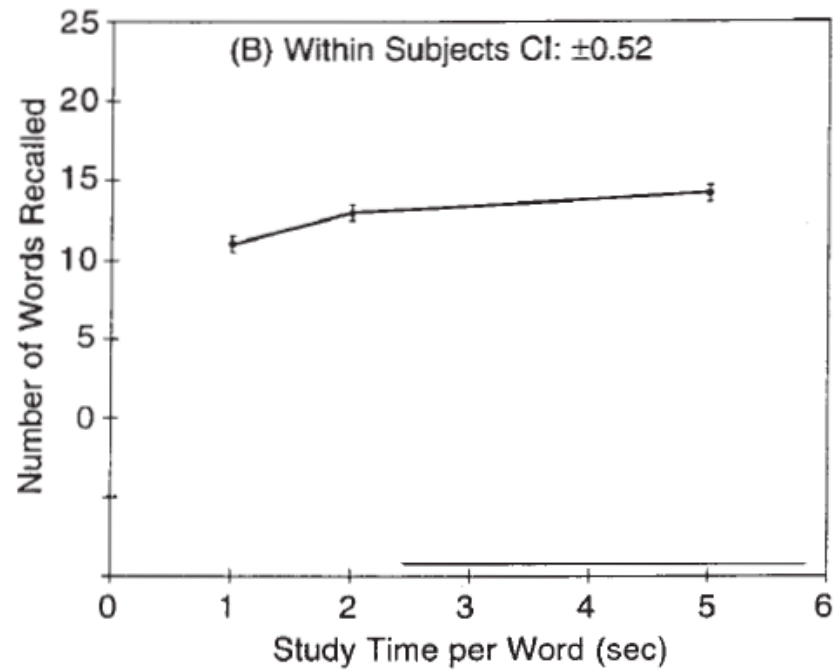
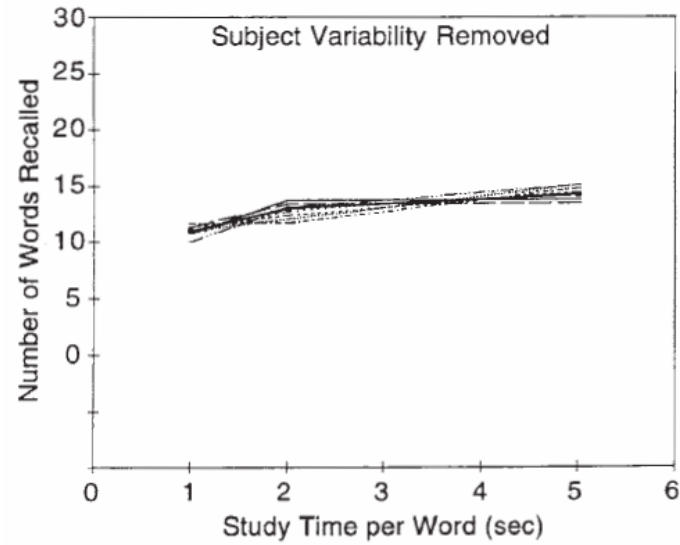
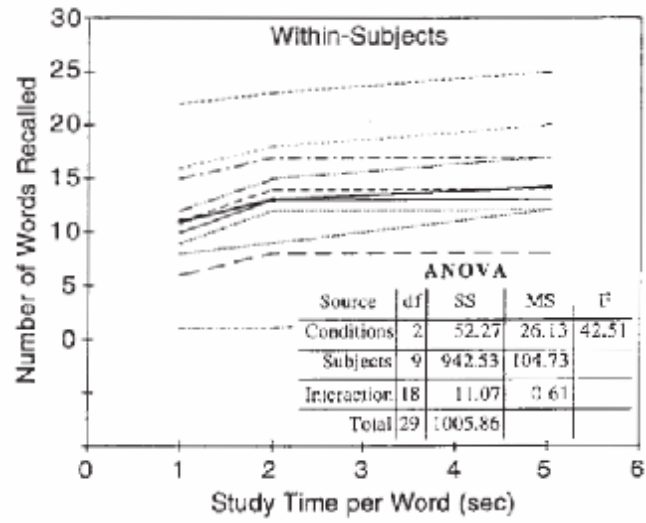


Figure 1. Hypothetical data without confidence intervals (Panel A) and with confidence intervals (Panels B and C).

Confidence intervals for within-subjects designs





Advantage of within subjects designs: lower error rates and therefore more statistical power (less subjects, less work...)

Problem: practice effects, order effects (A after B behaves differently than A after C).

Overcoming this problem is by counterbalancing: testing all the possible orders.

When there are too many levels, the number of orders (N!) becomes very large. Solution: Latin square: each level appears once in each row and once in each column.

Orthogonal Latin Squares further add the constraint that for any given left-right row order there is a corresponding top-bottom column order

	X ₁	X ₂	X ₃	X ₄	In this table, each subscripted entry represents a measure of task performance for one particular subject under one particular combination of the levels of X, Y, and Z. Thus, A ₁ represents the measure for one subject on level 1 of IV-X, level 1 of IV-Y, and level A of IV-Z; B ₃ represents the measure for one subject on level 4 of IV-X, level 3 of IV-Y, and level B of IV-Z; and so on.
Y ₁	A ₁	B ₁	C ₁	D ₁	
Y ₂	B ₂	C ₂	D ₂	A ₂	
Y ₃	C ₃	D ₃	A ₃	B ₃	
Y ₄	D ₄	A ₄	B ₄	C ₄	

Between-subjects and within subjects counterbalancing:

AB-BA

BA-AB

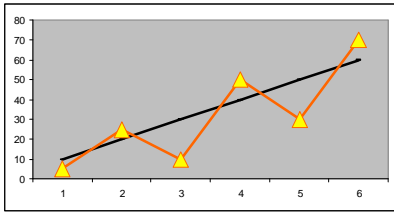
Background on statistics and research design

Broad design classes

- All designs seek to identify functions relating variables to one another.
- Correlational designs measure all the variables
- Experimental designs create the variance (manipulate) the independent variable(s) and measure the dependent variable(s)
- The two designs DO NOT differ in the statistics they employ
 - t-test comparing boys and girls in their need-for achievement = correlational design
 - Pearson correlation between the (manipulated) study time and memory = experimental design

Functions vs. measures of association

- Association measures tell us what proportion of the dependent variable's variance is explained by the independent variable(s) USING A GIVEN FUNCTION.
- Linear functions are commonly used but the reason is mathematical convenience, not psychological plausibility.



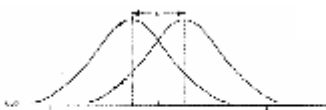
One may use a function-less measure of association (coloured line) which will show high association, or may alternatively use a function-dependent (line, in this case) measure, which will show less association. Researchers usually prefer to sacrifice precision for generality and choose the function-dependent measure. Using a function allows one to interpolate, extrapolate, and to generate new functions based on known functions.

Effect size vs. statistical significance

- Measures of association are also called effect sizes.
- Statistical significance depends on the effect size and on the sample size. Consequently, small effects can be significant if the sample is sufficiently large and large effects may not be significant if the sample size is small.
- Some authors therefore suggest that statistical significance may not be very important.

Effect size measures

$$\text{Cohen's } d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$



(expressed in units of variance)

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

Linking effect size to statistical significance

$$r = \frac{1}{\sqrt{1 + \frac{N-2}{F}}}$$

$$r = \frac{1}{\sqrt{1 + \frac{N-2}{F}}}$$

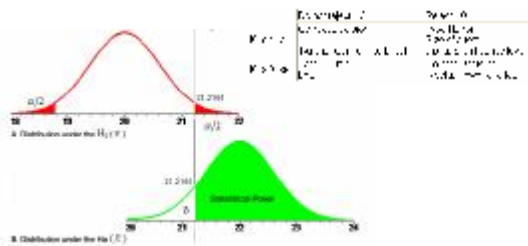
$$r = \frac{1}{\sqrt{1 + \frac{N-2}{F}}}$$

$$r = \frac{1}{\sqrt{1 + \frac{N-2}{F}}}$$

It is easy to see from the last equation that
Significance = Effect size * sample size

$$Z = r \sqrt{N}$$

Hypothesis testing and statistical power

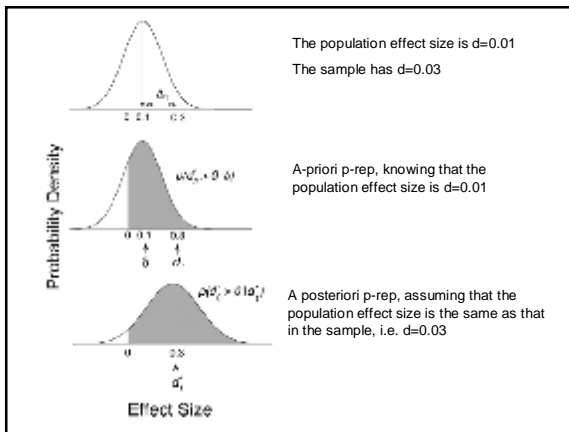


Why do we need to know the statistical power

- A-priori: When we plan an experiment and need to know the required sample size (to obtain Power=.80, usually)
- A-posteriori, when we failed rejecting the null hypothesis and we wish to endorse this hypothesis
- Power is a function of reliability and sample size.
 - In correlational designs, power is a positive function of sample heterogeneity.
 - In experimental designs, power is a negative function of sample heterogeneity
 - Reliability refers to BOTH the consistency in which the manipulation has been applied and the consistency of the dependent measure
 - In experimental designs, within subjects designs usually provide more power (=fewer subjects for the same power) as compared to between subjects designs.
 - Discuss potential problems of within subjects designs

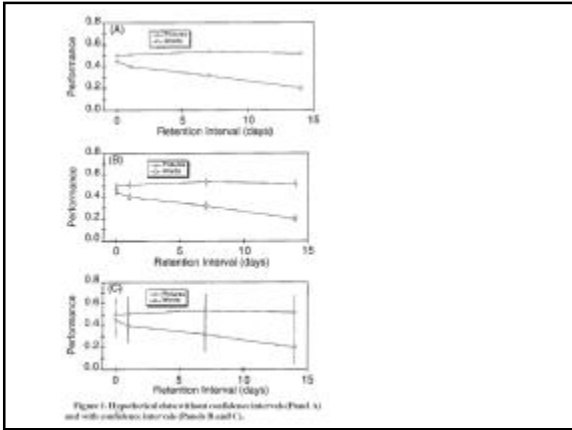
A-posteriori p-value vs. p-rep

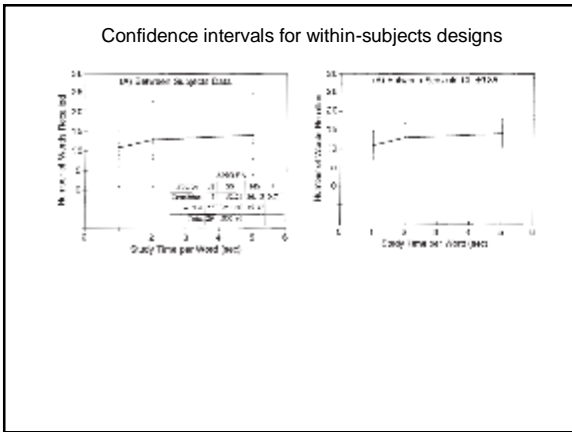
- A-posteriori p-value is the probability of finding a sample showing an effect of size X , given that the population effect is zero (unless H_0 is different which is rare).
- P-rep is based on the same data but it reflects the more intuitive notion of the probability that the next sample will also show a non-zero effect in the same direction as found in the present sample.

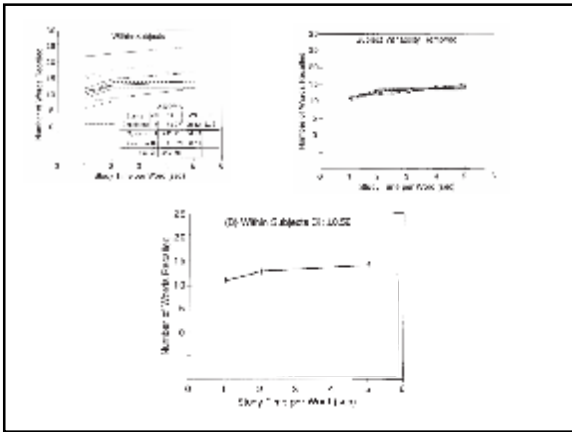


Hypothesis testing vs. confidence intervals

“... hypothesis testing is primarily designed to obliquely address a restricted, convoluted, and usually uninteresting question—Is it not true that some set of population means are all equal to one another?—whereas confidence intervals are designed to directly address a simpler and more general question—What are the population means?” (Loftus & Masson, 1994)







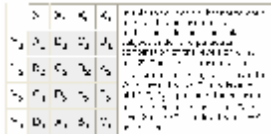
Advantage of within subjects designs: lower error rates and therefore more statistical power (less subjects, less work...)

Problem: practice effects, order effects (A after B behaves differently than A after C).

Overcoming this problem is by counterbalancing: testing all the possible orders.

When there are too many levels, the number of orders (N!) becomes very large.
Solution: Latin square: each level appears once in each row and once in each column.

Orthogonal Latin Squares further add the constraint that for any given left-right row order there is a corresponding top-bottom column order



Between-subjects and within subjects counterbalancing:

AB-BA

BA-AB
