Some basic statistics for evidence-based practice

The statistics typically reported in research papers are not always the most useful for understanding the clinical applications of the findings. Fortunately, however, the reader can often transform these statistics through rather simple computations into other forms that have a more straightforward clinical interpretation. Below is a brief overview of some useful EBP statistics.

I. Intervention studies where the outcome of interest is expressed as a dichotomous variable (e.g., delayed/not delayed; walking/not walking; pass state-wide achievement test/fail test): the ***number needed to treat***.

> The ***number needed to treat (NNT)*** answers the question: how many people would I need to provide the new intervention to in order to see a benefit in ***one*** additional person (i.e. one person more than would be expected from providing the alternative intervention). (If I have to provide this intervention to 50 toddlers in order to see benefit in *one* additional toddler, this may not be a very useful intervention, especially if the intervention is expensive.)

NNT is calculated in a series of steps, which are worked out below using an example involving a comparison of two different intervention approaches. The two possible outcomes are passing and not-passing a kindergarten readiness test. My question is: how many children would need to receive the intervention in order to decrease the rate of failure by one additional child?.

	CER	**EER**	**RRR**	**ARR**	**NNT**
Definition of term	Control event rate	Experimental event rate	Relative risk reduction	Absolute risk reduction	Number needed to treat
Meaning of statistic	What proportion of Control Group failed?	What proportion of Exp. Group failed?	How much did the intervention reduce the risk for failure?	How much did the intervention reduce the risk for failure after taking into account "baseline" rates?	How many children would you have to provide the intervention to in order to see one additional child pass who otherwise would have failed?
Outcome Failing state test	0.20	0.10	50%	0.10	10
Calculation of statistic	c/c+d	a/a+b	$\frac{\text{CER} - \text{EER}}{\text{CER}}$	CER - EER	1/ARR

Schematic for statistics:

	Outcome	
Treatment	Fail	Pass
Experimental intervention group	a (10)	b (90)
Control intervention group	c (20)	d (80)

Some basic statistics for evidence-based practice

II. A statistic to express the magnitude of difference between groups on a continuous outcome variable: the *effect size "**d**".*

Evaluating the meaning of results from statistical comparisons such as t-tests and ANOVAs using only the 'p" value can be problematic for two reasons:

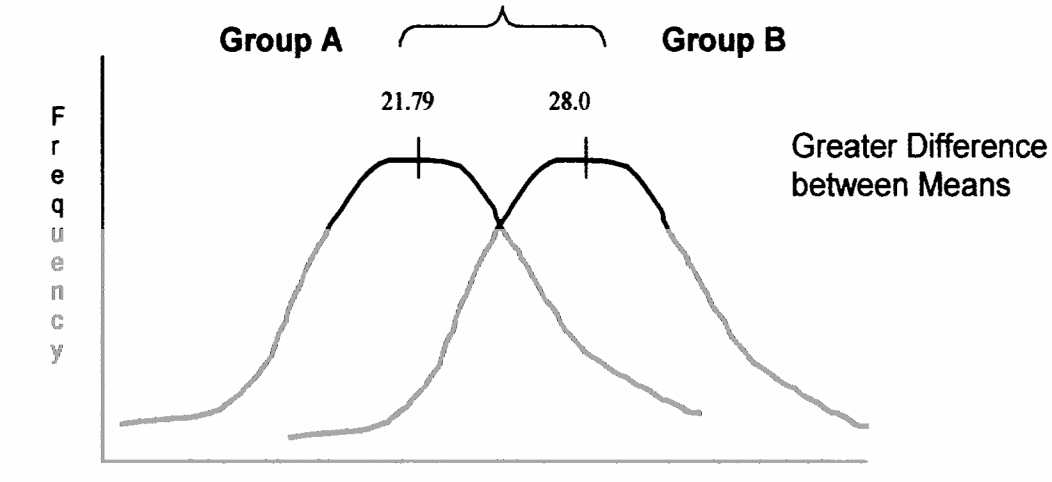
1. the "t" or "F" statistic by itself isn't easily interpreted
2. the value of "p" is very dependent on the sample size.

As an alternative, you can calculate the statistic "d", which provides an indicator of the ***magnitude*** of the difference between groups and is a standardized index that can be compared across studies. Two studies that seem to have "different" results based on the 'p' level may turn out to have very similar results when you compare the "d" calculated for each – the differences are probably due to differences in sample size.

"d" is calculated as: $\frac{M_1 - M_2}{(SD_{control})}$

i.e. the difference between the group means, divided by the standard deviation of the control (or comparison) group.

To understand *d*, you need to remember that statistics like t and ANOVA are actually comparing the distributions of scores in the samples being compared.



A larger *d* means there is less overlap between the distributions of scores in the two groups, and there is less likelihood that there will be many people in both groups who get the same score (i.e., their performance is more consistently different).

With the effect sizes often seen in intervention research (.4 to .6), this means that there will be a fairly large group of control participants and experimental participants who get the same score (i.e. perform the same).This is probably one reason why "clinical experience" can be mistaken: we see the people who improve with our intervention of choice, but don't have the information that would show us that as many (or more) people improve without the intervention – or with a different intervention.

As a general rule: a "small effect" =. 20; a "medium effect" = .50; a "large effect" = .80

Some basic statistics for evidence-based practice

III. A statistic that translates d into success rates: BESD

Although the effect size (d) is a helpful statistic, NNT has the advantage of expressing results in a way that is easier to understand and to communicate to patients. Fortunately, there is a way to express results from studies with continuous outcome measures in a somewhat similar way: the **Binomial Effect Size Display or BESD**. The BESD expresses the difference in outcomes in terms of differences in "success rates", a contrast that has more practical meaning.

Steps to calculate a BESD:

1. Transform your *d* to an *r*.

   This transformation is possible because, mathematically, statistics like the "t-test", the effect size "d", and the correlation being calculated for the BESD are interrelated.

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

2. Use the "r" to calculate "success" rates for both groups:

   The Experimental group success rate is calculated as = .50 + r/2 x 100
   The Control group success rate is calculated as .50 – r/2. x 100.
   Now you can use these rates to construct a 2 x 2 table for easy display.

   For example: if you calculated r = .30, then your table would look like this

	Outcome = + (success)	Outcome = - (no success)
Experimental Treatment group	65	35
Control group	35	65

You could then present this finding to a client or colleague by saying "The experimental treatment increased the rate of successful outcome from 35% to 65%.

You will not generally see a BESD reported in a journal article, but it is easy to compute, and often makes it easier to talk to clients about the meaning of the research evidence. For a useful discussion of the application of BESD, see:

McCartney, K. & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. Child Development, 71, 173-180.

Some basic statistics for evidence-based practice

Practice examples

1. Outcomes of a study with a continuous outcome measure

A study comparing two types of intervention for young children with CP reported the following changes in scores on an assessment of functional movement skills. Calculate the effect size "d" for this study.

	Pre test mean & s.d.	Follow up mean & s.d.
Control. Group	75.8 (11.6)	81.2 (7.5)
Experimental Group	78.2 (11.3)	88.1 (10.2)

Write a sentence "in English" that summarizes the meaning of this result.

Hint: there are two ways to calculate a "*d"* here:

1. using the follow-up means for the two groups in the equation or
2. computing the "mean change" for each group, and using those two means in your equation.

Try the calculation both ways to see if the results are the same. Interpret your findings.

2. Calculate a BESD

Use one of the *d*'s calculated above to calculate a BESD.

d = _____ r = _____

	Outcome = + (success)	Outcome = - (no success)
Experimental Treatment group		
Control group		

Some basic statistics for evidence-based practice

3. Outcome calculations: dichotomous variables.

You find a CAT for a study investigating the efficacy of a new type of positioning equipment that reduces risk of hip subluxation in young children with spasticity. It reports the rates given below.

Treatment	Subluxation	No subluxation
New (Exp. Group)	a = 5	b = 50
Usual care (Control group)	c = 9	d= 60

Calculate the NNT.

		CER	**EER**	**RRR**	**ARR**	**NNT**
Definition of term		Control event rate	Experimental event rate	Relative risk reduction	Absolute risk reduction	Number needed to treat
Calculation of statistic		c/c+d	a/a+b	$\frac{\text{CER} - \text{EER}}{\text{CER}}$	CER - EER	1/ARR

Write a sentence "in English" that expresses the result.

WHAT YOU REALLY NEED TO KNOW ABOUT STATISTICS

A. Measures of Central Tendency

Measures of central tendency are measures of the "average" or "most typical", and are the most widely used statistical description of data. Measures of central tendency include:

1. Mean – the arithmetic average – the mean of a set of observations is simply their sum, divided by the number of observations.

2. Median – the median is the 50th percentile of a distribution – the point below which half of the observations fall

3. Mode – the mode is the most frequently occurring observation – the most popular score of a class of scores.

B. Measures of Variability or Dispersion

Measures of variability reflect the degree of spread or dispersion that characterizes a group of scores and the degree to which a set of scores differs from some measure of central tendency.

1. Range – the range is the difference between the highest and lowest scores in a distribution.

2. Standard deviation – the standard deviation is the most commonly used measure of variability. The standard deviation is the average amount that each of the individual scores varies from the mean of the set of scores.

C. The most commonly used statistical procedures

1. Chi-square ($?^2$): a statistic that can be used to analyze nominal (categorical) data. It compares the *observed* frequency of a particular category to the *expected* frequency of that category.

Example: Is ADHD diagnosed more frequently in boys than girls of kindergarten age?
(ADHD diagnosis & gender are both nominal data)

Result is written as: $?^2$ (df) = 289.3, p<.05

Result is reported as: A chi-square analysis found, in children of kindergarten age, ADHD was diagnosed significantly more frequently in boys than girls ($?^2$ (150) = 289.3, p<.05).

2. t-test: a statistical analysis that is used to compare the means of two groups.

Example: Do 2-year-old children born premature and VLBW score lower than children born full term on the Motor Scale of the BSID-II?

Result is written as: t(df) = 3.86, p<.05.

Result is reported as: The mean Motor Scale score of the two groups was compared with a t-test and found to be significantly different ($t(df) = 3.86$, $p<.05$).

Note: you must look at the descriptive statistics (means) to tell which group had the higher score.

3. Analysis of variance (ANOVA): a more complex statistical procedure that can be used to compare more than two groups on a dependent variable. ANOVA can also be used when the design has more than 1 independent variable. (There are several different "types" of ANOVA).

Example: For two-year-old children with developmental delay, is weekly home-based consultation to parents associated with greater improvement in children's functional skills compared to weekly therapist-provided direct service or no intervention at all?
IV = Intervention/group; DV = Community Participation score

Result is written as: $F (df, df) = 9.82$, $p<.01$.

Result is reported as: The amount of change in children's functional skill performance across a one-year period was compared using analysis of variance. There was a significant difference between groups ($F (df, df) = 9.82$, $p<.01$).

Note: The "F" value only tells you that there is a difference between groups. It does not necessarily mean that each pair-wise comparison between groups will be significant. You will need to look at the means (and the results of further tests, referred to as post-hoc analyses) to determine which groups performed significantly higher (or lower) than the others.

4. Correlation: A measure of the extent to which two variables tend to change together; i.e. a measure of the degree of association between them. Since correlational designs do not involve manipulation, they do not have an IV or DV.

An "r" may vary between -1.0 and +1.0:

negative correlation = as one measure increases, the other decreases; e.g., air temperature and amount of clothing worn are negatively correlated.

positive correlation = the measures tend to increase or decrease together; e.g., age and height are positively correlated (through childhood)

Result is written as: $r = .42$, $p<.05$.

Result is reported as: The two tests of hand function were only moderately correlated ($r = .42$, $p<.05$), suggesting that they do not measure the exact same skills.

Note: Correlations are particularly sensitive to variations in sample size. When interpreting a correlation, the size of the correlation should be considered as well, not just the "p" level. When samples are in the hundreds, even a correlation of $r = .10$ may be "significant". However $r = .10$ is still quite small, and an association of this magnitude may not have "real life" value.

5. Regression: Regression is a type of analysis in which one or more variables (IV's) are used to try to predict (statistically) levels of another variable (DV). There are several different types of regression, but they all have essentially the same goal of statistical prediction. The program will typically select from the whole set of IV's a smaller set of those that, as a set, do the best job of predicting the outcome variable.

> Example: In a set of variables that includes age, gross motor development, general health, and level of cognitive function (IV's), which variables best predict the current ADL skill performance (DV) of a child with cerebral palsy?
>
> Results are written in a variety of ways, depending on the study. One general approach is to report the overall amount of variance "accounted for" (predicted) by the regression model, e.g. $R^2 = .27$, and to provide additional statistics (referred to as Beta-weights) for each significant independent (predictor) variable in a table.
>
> Result is reported as: Level of cognitive function and gross motor development were the only significant predictors, accounting for 23% of the variance in ADL skill level.
>
> Note: A "significant" regression analysis only tells you that the set of selected variables can statistically predict an individual's score on the outcome variable (DV) to some degree better than chance. The closer the R^2 is to 1, the better the prediction (so in the example above, the prediction wasn't terrific). It does not tell you (1) that there is a causal relationship between the IV's and DV; (2) that variables that were not "significant" (i.e. were not selected) had no relation to the DV – just that the variables selected by the analysis program could create a good statistical predictive model without them.

D. Further notes on the interpretation of statistical results

Degrees of freedom: the (df) in parentheses following $?^2$, t, or F reflect the size of your sample and the number of variables in your analyses. Each statistical test has a formula for calculating the appropriate degrees of freedom (e.g. for t, df = n - 2). The df are important because they determine the "p" level of a given value obtained for $?^2$, t, or F.

Example: Using the appropriate formula, I calculated t = 2.20. When I look this number up in the table, I find that if my *df* were 10, this result would not be significant at $p<.05$. However, if I had a large sample, and my *df* = 30, the result would be significant.

A note on "p": As seen above, all of these statistical analyses yield a "p level". The "p" is a measure of the probability that the particular result obtained could have occurred by chance. Some examples of the correct way to interpret "p<.05" are:

> a). There is less than a 5% likelihood that a difference of this size between the means of the two groups occurred by chance (i.e. because of random events or by fluke, rather than due to the effect you are examining).
>
> b). There is less than a 5% likelihood that a correlation of this size would have occurred by chance (i.e. occurred randomly, rather than because there is some solid or true basis for the association).

E. Checking interpretations for accuracy

1. When interpreting statistical results in a research report, it is not appropriate to say that statistically significant results prove a hypothesis was correct or prove that two groups were really different. "Statistically significant" results mean that the results that are "not very likely" to be due to chance alone (but there is always a small chance that one could be wrong...). Significant results "lend support" to a hypothesis, or "provide evidence" that a hypothesis may be correct, but (except in very extraordinary circumstances) a single study never proves anything.

2. When interpreting results from a study that uses "t" or "ANOVA", you cannot assume simply because groups are being compared that this is a true experimental design from which causal implications can be drawn. The study design must meet other requirements (e.g., random selection and assignment to groups) in order for causal interpretations to be appropriate.

3. Correlational designs do not establish causality, so interpretations of "r" should not use language that implies a causal relation between the two variables, regardless of what the author's favorite theory suggests. A "statistically significant" correlation means that the variables change together in a predictable way more than would be likely because of chance (but there is always a small possibility that one could be wrong...). A significant correlation does not demonstrate "the effect of A on B" or the "impact of intensity of treatment A on functional assessment score B". Similarly, results of regression analyses (which are a variation of correlational design) show "the extent to which variance in outcome A can be predicted by cognitive status measure B", NOT "the effect of cognitive status on outcome".

Can I Apply this Evidence?

Example 1

Your question: Are there data on when and in what sequence children with spina bifida typically acquire functional skills that I could use to help set appropriate goals?

Focus of the study you found: descriptive study of the acquisition of functional skills (ADL)

Study features: One-time assessment of a convenience sample of 50 young children with spina bifida; assessed during a routine follow-up visit at a regional clinic in Sweden; age range 2-7 years (mean age = 3.8 years); 65% female.

Your client: A 2-year-old girl from an urban Hispanic family.

Can you/should you apply these findings? How? What specific factors would you weigh?

Example 2

Your question: Does participation in center-based programs result in greater developmental progress for children with general developmental delay (compared to home-based services)?

Focus of the study you found: Comparison of outcomes for children receiving early intervention services through different service models.

Study features: 10 early intervention programs in 2 Midwestern states; programs differed in extent to which children were served through center-based versus home-based intervention; data were gathered at time of entry into the program, and at 3 month intervals afterwards; results based on a total of 240 children – mean age 2.3 years; 97% Caucasian; 25% rural; largest diagnostic group = speech/language delay (40%).

Your client: a child with Down syndrome from an African-American family; parents are both professionals.

Can you/should you apply these findings? How? What specific factors would you weigh?

Example 3

Your question: Does providing very early intervention to children born VLBW improve neurodevelopmental outcomes for these children born at-risk?

Focus of the study you found: Effects of SI and NDT based OT on neurodevelopment of children born <1000 grams.

Study features: 104 infants, followed prospectively; divided into matched intervention and control groups; intervention children had a weekly session of 60 minutes of occupational therapy from the corrected age of 6 months up to 12 months; children were compared at 6, 12, 18, and 24 months of age; neurodevelopment of the two groups did not differ at any point.

Your client: a 4-month old infant, born weighing 1500 grams

Can/should you apply these findings? How? What specific factors would you weigh?