

Magnitude Matters: Effect Size in Research and Clinical Practice

Will G Hopkins
AUT University, Auckland, NZ

- Why Magnitude Matters in Research
- Why Magnitude Matters in Clinical Practice
- Magnitudes of Effects
 - Types of variables and models
 - Difference between means
 - "Slope"
 - Correlation * = innovation!
 - Difference of proportions
 - Number needed to treat
 - Risk, odds and hazard ratio
 - Difference in mean time to event

Background

- International Committee of Medical Journal Editors (icmje.org)
 - "Show specific effect sizes."
 - "Avoid relying solely on statistical hypothesis testing..., which fails to convey important information about effect size."
- Publication Manual of the American Psychological Association
 - A section on "Effect Size and Strength of Relationship"
 - 15 ways to express magnitudes.
- Meta-analysis
 - Emphasis on deriving average magnitude of an effect.

Why Magnitude Matters in Research

- Two reasons: estimating **sample size**, and making **inferences**.

Estimating Sample Size

- Research in our disciplines is all about **effects**.
 - An effect = a **relationship** between a predictor variable and a dependent variable.
 - Example: the effect of exercise on a measure of health.
- We want to know about the effect in a **population**.
- But we study a **sample** of the population.
- And the magnitude of an effect varies from sample to sample.
- For a big enough sample, the variation is acceptably small.
- How many is *big enough*?
 - Get via statistical, clinical/practical or mechanistic significance.
 - You need the smallest important **magnitude** of the effect. *
 - See MSSE 38(5), 2006: Abstract 2746. *

Making Inferences

- An inference is a statement about the effect in the population.
- Old approach: is the effect **real** (statistically significant)?
 - If it isn't, you apparently assume there is no effect.
 - Problem: no mention of magnitude, so depending on sample size...
 - A "real effect" could be clinically trivial.
 - "No effect" could be a clinically clear and useful effect.
- New approach: is the effect **clear**?
 - It's clear if it can't be substantially positive and negative.
 - That is, if the confidence interval doesn't overlap such values. *
- New approach: what are the **chances** the real effect is **important**?
 - ...in a clinical, practical or mechanistic sense. *
- Both new approaches need the smallest important **magnitude**.
- You should also make inferences about other magnitudes:
small, moderate, large, very large, awe-inspiring. *

Why Magnitude Matters in Clinical Practice

- What really matters is **cost-benefit**.
- Here I am addressing only the **benefit** (and harm).
- So need smallest important beneficial and harmful **magnitudes**.
 - Also known as **minimum clinically important difference**.
 - "A crock"?
 - In the absence of clinical consensus, need statistical defaults.
 - Also need to express in units the clinician, patient, client, athlete, coach or administrator can understand.
 - You should use these terms sometimes:
trivial, small, moderate, large, very large, awe-inspiring.
- The rest of this talk is about these magnitudes for different kinds of effect.

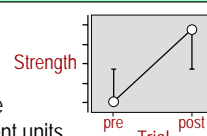
Magnitudes of Effects

- Magnitudes depend on **nature of variables**.
 - **Continuous**: mass, distance, time, current; measures derived therefrom, such as force, concentration, voltage.
 - **Counts**: such as number of injuries in a season.
 - **Nominal**: values are levels representing names, such as injured (no, yes), and type of sport (baseball, football, hockey).
 - **Ordinal**: values are levels with a sense of rank order, such as a 4-pt Likert scale for injury severity (none, mild, moderate, severe).
- Continuous, counts, ordinals can be treated as **numerics**, but...
 - As dependents, counts need generalized linear modeling.
 - If ordinal has only a few levels or subjects are stacked at one end, analyze as nominal. *
- **Nominals** with >2 levels are best dichotomized by comparing or combining levels appropriately. *
 - Hard to define magnitude when comparing >2 levels at once.

- Magnitude also depends on the **relationship** you model between the dependent and predictor.
 - The model is almost always **linear** or can be made so.
 - Linear model: sum of predictors and/or their products, plus error.
 - Well developed procedures for estimating effects in linear models.
- Effects for linear models:

| Dependent | Predictor | Effect | Statistical model |
|-----------------------|--------------------|--|---|
| numeric Strength | nominal Trial | difference in means | regression; general linear; mixed; generalized linear |
| numeric Activity | numeric Age | "slope" (difference per unit of predictor); correlation | generalized linear |
| nominal InjuredNY | nominal Sex | diffs or ratios of proportions, odds, rates, mean event time | logistic regression; generalized linear; proportional hazards |
| nominal SelectedNY | numeric Fitness | "slope" (difference or ratio per unit of predictor) | logistic regression; generalized linear; proportional hazards |

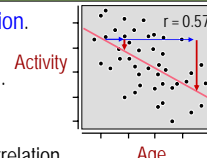
| Dependent | Predictor | Effect |
|---------------------|------------------|-------------------------------|
| numeric Strength | nominal Trial | difference or change in means |

- You consider the **difference or change in the mean** for pairwise comparisons of levels of the predictor.
 
- Clinical or practical **experience** may give smallest important effect in raw or percent units.
- Otherwise use the **standardized difference or change**.
 - Also known as **Cohen's effect size** or Cohen's *d* statistic.
 - You express the difference or change in the mean as a fraction of the **between-subject standard deviation** ($\Delta\text{mean}/\text{SD}$).
 - For many measures use the log-transformed dependent variable.
 - It's biased high for small sample size.
 - Correction factor is $1-3/(4v-1)$, where v =deg. freedom for the SD.
 - The smallest important effect is ± 0.2 .

Measures of Athletic Performance

- For **team-sport** athletes, use standardized differences in mean to get smallest important and other magnitudes. *
- For **solo** athletes, smallest important effect is about half of a top athlete's typical event-to-event variation. *
 - Currently no values for moderate, large, very large, awesome.
 - Beware: smallest effect depends on how it's measured, because...
 - A percent change in an athlete's ability to output power results in different percent changes in performance in different tests.
 - Example: a 1% change in endurance power output produces the following changes...
 - 1% in running time-trial speed or time;
 - ~0.4% in road-cycling time-trial time;
 - 0.3% in rowing-ergometer time-trial time;
 - ~15% in time to exhaustion in a constant-power test.
 - An indeterminate change in any test following a pre-load. *

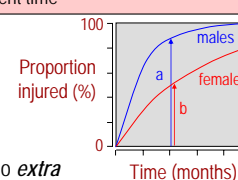
| Dependent | Predictor | Effect |
|---------------------|----------------|---|
| numeric Activity | numeric Age | "slope" (difference per unit of predictor); correlation |

- A **slope** is more practical than a **correlation**.
 
- But unit of predictor is **arbitrary**, so it's hard to define smallest effect for a slope.
 - Example: -2% per year may seem trivial, yet -20% per decade may seem large.
 - For consistency with interpretation of correlation, better to express slope as difference per **two SDs** of predictor. *
 - Fits with smallest important effect of 0.2 SD for the dependent.
 - But underestimates magnitude of larger effects. For an explanation, see newstats.org/efectmag.html
- Easier to interpret the correlation, using **Cohen's scale**.
 - Smallest important correlation is ± 0.1 . Complete scale:

| | | | | | |
|---------|---------|----------|---------|------------|---------|
| <0.1 | 0.1-0.3 | 0.3-0.5 | 0.5-0.7 | 0.7-0.9 | >0.9 |
| trivial | small | moderate | large | very large | awesome |

- You **can** use correlation to assess nominal predictors.
 - For a two-level predictor, the scales match up.
 - For >2 levels, the correlation doesn't apply to an individual.
- Magnitudes when **controlling for something**...
 - Control for = hold it equal or constant or **adjust** for it.
 - Example: the effect of age on activity adjusted for sex.
 - Control for something by adding it to the model as a predictor.
 - Effect of original predictor changes.
 - No problem for a difference in means or a slope.
 - But correlations are a challenge.
 - The correlation is either **partial** or **semi-partial** (SPSS: "part").
 - Partial = effect of the predictor within a virtual subgroup of subjects who all have the same values of the other predictors.
 - Semi-partial = unique effect of the predictor with **all** subjects.
 - Partial is probably more appropriate for the individual.
 - Confidence limits may be a problem in some stats packages.

| Dependent | Predictor | Effect |
|----------------------|----------------|---|
| nominal InjuredNY | nominal Sex | differences or ratios of proportions, odds, rates; difference in mean event time |

- Subjects all start off "N", but different proportions end up "Y".
 
- Risk difference** = $a - b$.
 - Good measure for an individual, but time dependent.
 - Example: $a - b = 83\% - 50\% = 33\%$, so **extra** chance of one in three of injury if you are a male.
 - Smallest effect: $\pm 5\%$?
- Number needed to treat (NNT)** = $100/(a - b)$.
 - Number of subjects you would have to treat or sample for one subject to have an outcome attributable to the effect.
 - Example: for every 3 people ($=100/33$), one **extra** person would be injured if the people were males.
 - NNT < 20 is clinically important?

- **Population attributable fraction**
= $(a - b) \times (\text{fraction population exposed})$.
- Smallest important effect for policymakers = ?
- **Relative risk** = a/b .
- Good measure for public health, but time dependent.
- Smallest effect: 1.1 (or 1/1.1).
 - Based on smallest effect of hazard ratio.
 - Corresponds to risk difference of 55 - 50 = 5%.
 - But relative risk = 6.0 for risk difference = 6 - 1 = 5%.
 - So smallest relative risk for individual is hard to define.
- **Odds ratio** = $(a/c)/(b/d)$.
 - Used for logistic regression and some case-control designs.
 - Hard to interpret, but it approximates relative risk when $a < 10\%$ and $b < 10\%$ (which is often).
 - Can convert exactly to relative risk if know a or b.

- **Hazard or incidence rate ratio** = e/f .
- Hazard = instantaneous risk rate = proportion per infinitesimal of time.
- Hazard ratio is best statistical measure.
 - Hazard ratio = risk ratio = odds ratio for low risks (short times).
 - Not dependent on time if incident rates are constant.
 - And even if both rates change, often OK to assume their ratio is constant.
 - Basis of proportional hazards modeling.
- Smallest effect: 1.1 or 1/1.1.
 - This effect would produce a 10% increase or decrease in the workload of a hospital ward, which would impact personnel and budgets.

- **Difference in mean time to event** = $t_2 - t_1$.
- Best measure for individual when events occur gradually.
- Can standardize with SD of time to event.
 - Therefore can use default standardized thresholds of 0.2, 0.6, 1.2, 2.0.
- Bonus: if hazard is constant over time, SD of $\log(\text{time to event})$ is independent of hazard.
 - Hence this scale for hazard ratios, derived from standardized thresholds applied to time to event:

<1.28 1.28-2.0 2.0-4.5 4.5-13 13-400 >400
trivial small moderate large very large awesome

| Dependent | Predictor | Effect |
|-----------------------|--------------------|--|
| nominal SelectedNY | numeric Fitness | "slope" (difference or ratio per unit of predictor) |

- Researchers derive and interpret the slope, not a correlation.
- Has to be modeled as odds ratio per unit of predictor via logistic regression.
 - Example: odds ratio for selection = 8.1 per unit of fitness.
- Otherwise same issues as for numeric dependent variable.
 - Need to express as effect of 2 SDs of predictor.
 - When controlling for other predictors, interpret effect as "for subjects all with equal values of the other predictors".

This presentation was downloaded from:

SPORTSCIENCE sportsci.org
A Peer-Reviewed Site for Sport Research

See SportsScience 10, 2006