

Introduction to meta-analysis

Outline for today

- Meta analysis compared with traditional review article
- Quantitative summaries vs. vote-counting
- How to carry out a meta-analysis
- Effect size
- Fixed and mixed-effects
- Associating effect sizes with relevant variables
- Publication bias
- Make your results accessible to meta-analysis

Scientific studies on a topic are often repeated

New studies improve/expand on previous studies, or examine the same issue in a different study system, or using different methods

- Schoener et al. (1983) found 164 published field experiments on interspecific competition.
- Gardner et al. (2003) obtained results from 51 separate studies reporting coral cover from 294 sites from across the Caribbean.
- Bell et al. (2009) found 759 published estimates of the repeatability of behavior, from 114 studies of 98 species.
- Vilà et al (2011) reviewed 199 articles reporting 1041 field studies describing the ecological impacts of 135 alien plant taxa.

A method is needed to summarize results from multiple studies

- Dr. Benjamin Spock sold 50 million copies of *Baby and Child Care* 1950s -- 1990s. In it he wrote “*I think it is preferable to accustom a baby to sleeping on his stomach from the beginning if he is willing*”. Other pediatricians made similar recommendations.
- From the 1950s into the 1990s, more than 100,000 babies died of sudden infant death syndrome (SIDS).
- In the early 1990s, researchers realized that the risk of SIDS decreased by at least 50% when babies were put to sleep on their backs rather than face down.
- Subsequent education campaigns led to a dramatic drop in the number of SIDS deaths.
- However, research was available from 1970 that sleeping on the stomach was hazardous to babies. An earlier synthesis of the data could have got the answer much sooner.

Traditional approach is the review article

An expert in the field assembles the studies published on a topic, thinks about them carefully and (hopefully) fairly, and then writes a review article summarizing the overall conclusions reached.

A first-rate review article advances a field far beyond a mere summary.

It reviews and comments on the current state of thought and knowledge about a particular topic.

Such a review will propose new hypotheses, uncover previously unnoticed relationships, and point to new paths of research.

The traditional review lacks a quantitative method

This might lead to two problems

- *Bias.*

In his 1986 book *How to Live Longer and Feel Better*, Linus Pauling (the only person to be awarded two unshared Nobel Prizes) cited 30 studies supporting his idea that large daily doses of vitamin C reduces the risk of contracting the common cold, but cited no studies opposing the idea, even though a number had been published. Not all reviews are **so** biased, but there are few rules regarding selection of studies for review.

- *Lack of a quantitative summary of research findings.*

Reviews don't tell us about how large the effect is.

“Vote-counting” is a step in the right direction

Divide studies into two categories: those that yielded a statistically significant result supporting the research hypothesis, and those that did not. The proportions of studies ‘voting’ for or against the hypothesis are then counted.

Vol. 122, No. 2

The American Naturalist

August 1983

FIELD EXPERIMENTS ON INTERSPECIFIC COMPETITION

THOMAS W. SCHOENER

Department of Zoology, University of California, Davis, California 95616

EXISTENCE OF COMPETITION

An overwhelming fraction of experimental attempts to detect interspecific competition in the field did so: 148 of 164 studies, or 90%, demonstrate some competition. One-hundred ten of the 148 studies record changes in numbers through local births and deaths or migration.

Limitations of vote-counting

- By counting only the statistically significant studies vote-counting ignores all the quantitative information about the magnitudes of effects.
- Too conservative. “Votes” are affected by the power of individual studies, which may be weak.
- Significance level by itself doesn’t indicate whether two or more studies obtained the same outcome.
- The magnitude of the effect is downplayed.
- It is difficult to quantify the effects of publication bias.
- Method is unable to weigh the effect of studies differing in sample size, and therefore power.

Limitations of vote-counting

- The Antiplatelet Trialists' Collaboration (1994) conducted a meta-analysis of 142 randomized experiments testing whether taking aspirin or other antiplatelet medication following a stroke or myocardial infarction ("heart attack") reduced the risk of future stroke. Total $N > 70,000$.
- The vote: 19 of 142 studies showed a statistically significantly better result for patients on antiplatelet therapy than for the control patients. Two of the 142 studies showed a significantly worse rate of vascular events with aspirin treatment.
- Yet 14.7% (5400/36,711) of patients in the control groups had subsequent vascular events, compared with 11.4% (4183/36,536) in the treated group. Small effect but real, according to meta-analysis methods. This conclusion saved many lives.

Meta-analysis, the “analysis of analyses”

Meta-analysis refers to the statistical synthesis of results from a series of studies (Borenstein et al 2009).

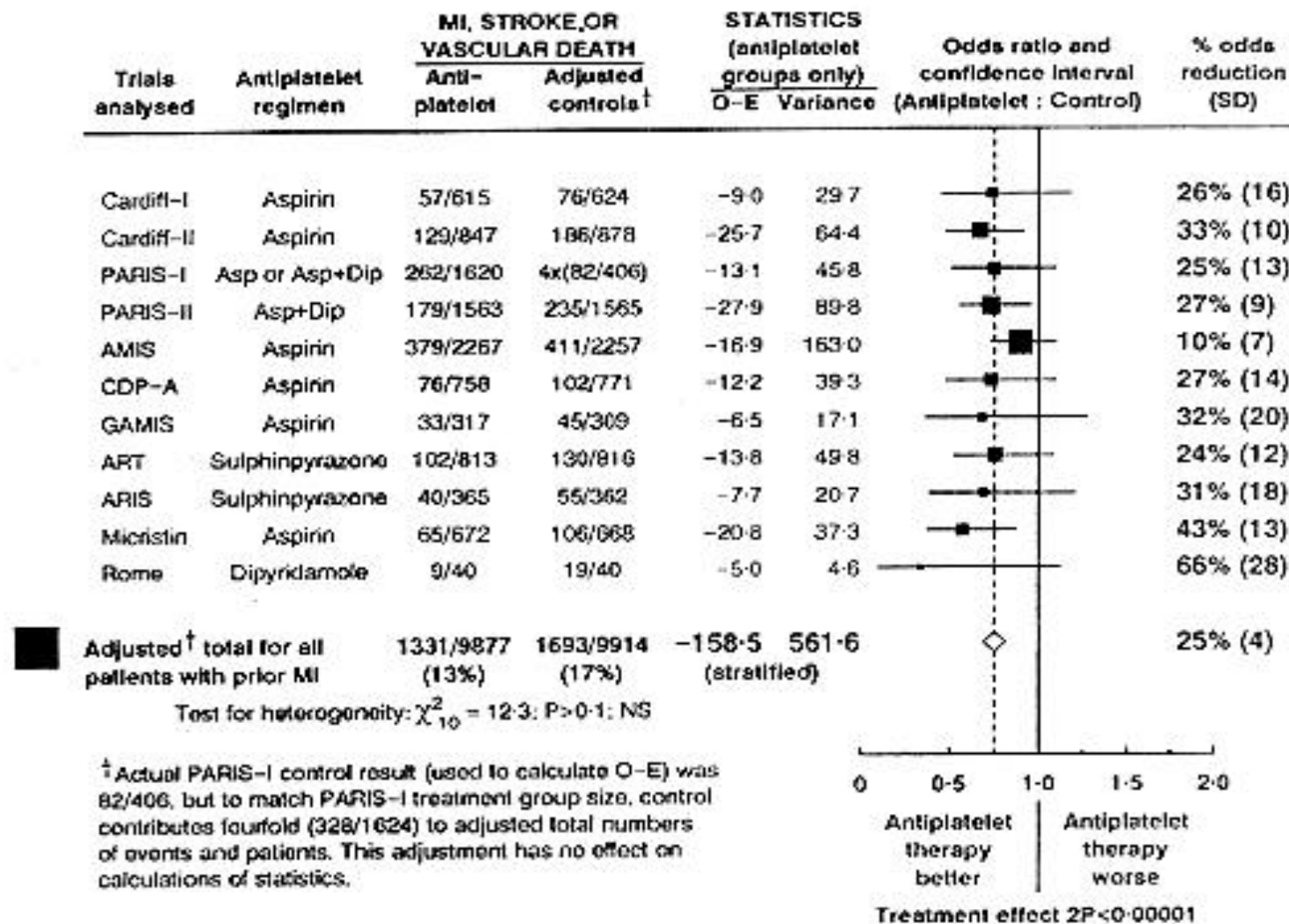
The method involves compiling all known scientific studies estimating an effect and quantitatively combining them to give an overall estimate of the effect.

Meta-analysis allows us to generalize. It lets us determine how frequent, how important, and how consistent effects are across a variety of systems.

Meta-analysis gets past the occasional sensational result (the one you read about in the newspaper) to an objective assessment of all the evidence.

Meta-analysis, the “analysis of analyses”

Came from medical research, in which all studies are all of the same species (humans). Here is a “forest plot” from the Antiplatelet Trialists’ Collaboration.



Meta-analysis, the “analysis of analyses”

Ecologists and evolutionary biologists attempt to generalize across a much wider range of species and systems.

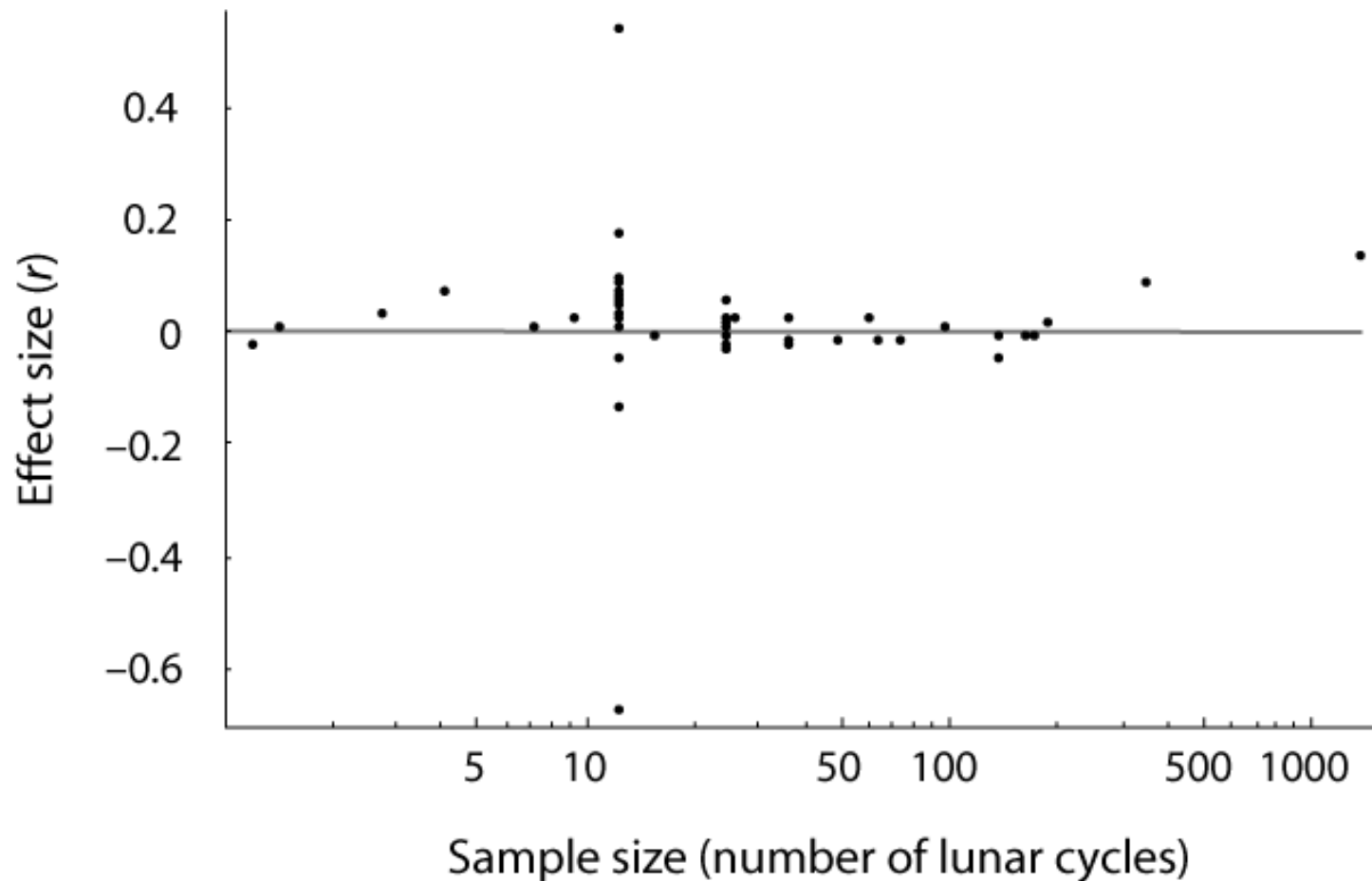
This is more challenging than studies carried out on a single species (e.g., humans).

Example 1: Meta-analysis of the Transylvania effect

- Many people believe that a full moon can affect human behavior. The word lunacy is derived from the Latin luna, moon.
- Legends of strange happenings, such as werewolves and vampires, have been connected to full moons for centuries.
- Lord Blackstone, an 18th-century English jurist, was the first to define a condition of madness exacerbated by the lunar cycle: *“A lunatic, or non compos mentis, is properly one who hath lucid intervals, sometimes enjoying his senses and sometimes not and that frequently depending upon the changes of the moon.”*
- Rotton and Kelly (1985) showed that 50% of university students believed that people act strangely during a full moon.
- Vance (1995) reported that as many as 81% of mental health professionals believed that the full moon alters individual behaviour.

Example 1: Meta-analysis of the Transylvania effect

Rotton and Kelly (1985) carried out a meta-analysis of studies correlating homicide rates, psychiatric hospital admissions, suicide rates, crisis calls, etc. The average effect size r was smaller than 0.01.



How to carry out a meta-analysis

1. Define the question and scope.

- A narrow question applied to a homogeneous group?
“Does aspirin reduce incidence of myocardial infarction?”
- Or a heterogeneous set of studies or variables?
“How much genetic variation exists in populations for behavioral traits?”
- Only experiments with controls and randomization? Only replicated experiments? Only experiments with blinding?
- It may be best to adopt a reasonably wide scope and investigate later whether differences between methods lead to different effects overall.

How to carry out a meta-analysis

2. Literature search, gather data.

- Make it exhaustive to avoid bias.
- Easily-found studies are *different* from those that we cannot find easily. Studies finding large, statistically significant effects are more likely to be published, more likely to be in “first-rate” journals, and more likely to be referenced in other articles.
- Statistical techniques exist to account partially for publication bias (funnel plots) but they do not replace an exhaustive survey.
- Decide whether to (hold your nose and) include studies of apparently poor quality. Failure to have well-defined criteria can lead to bias (we are more likely to discard a poor study if it disagrees with our pet hypothesis).

How to carry out a meta-analysis

- Ideally, the data obtained should all be independent, but non-independence of various sorts creeps in (e.g. multiple studies by the same lab).
- A single study may provide measurements on multiple species, or measurements of multiple responses on the same species. Include them all or take a summary measure?
- One or a small number of species (e.g., great tit) or systems (e.g., intertidal zone) may be overrepresented in the literature. Treat them all as independent?
- It may be worse to leave data out, or take summary measures, than to throw every data point into the analysis.

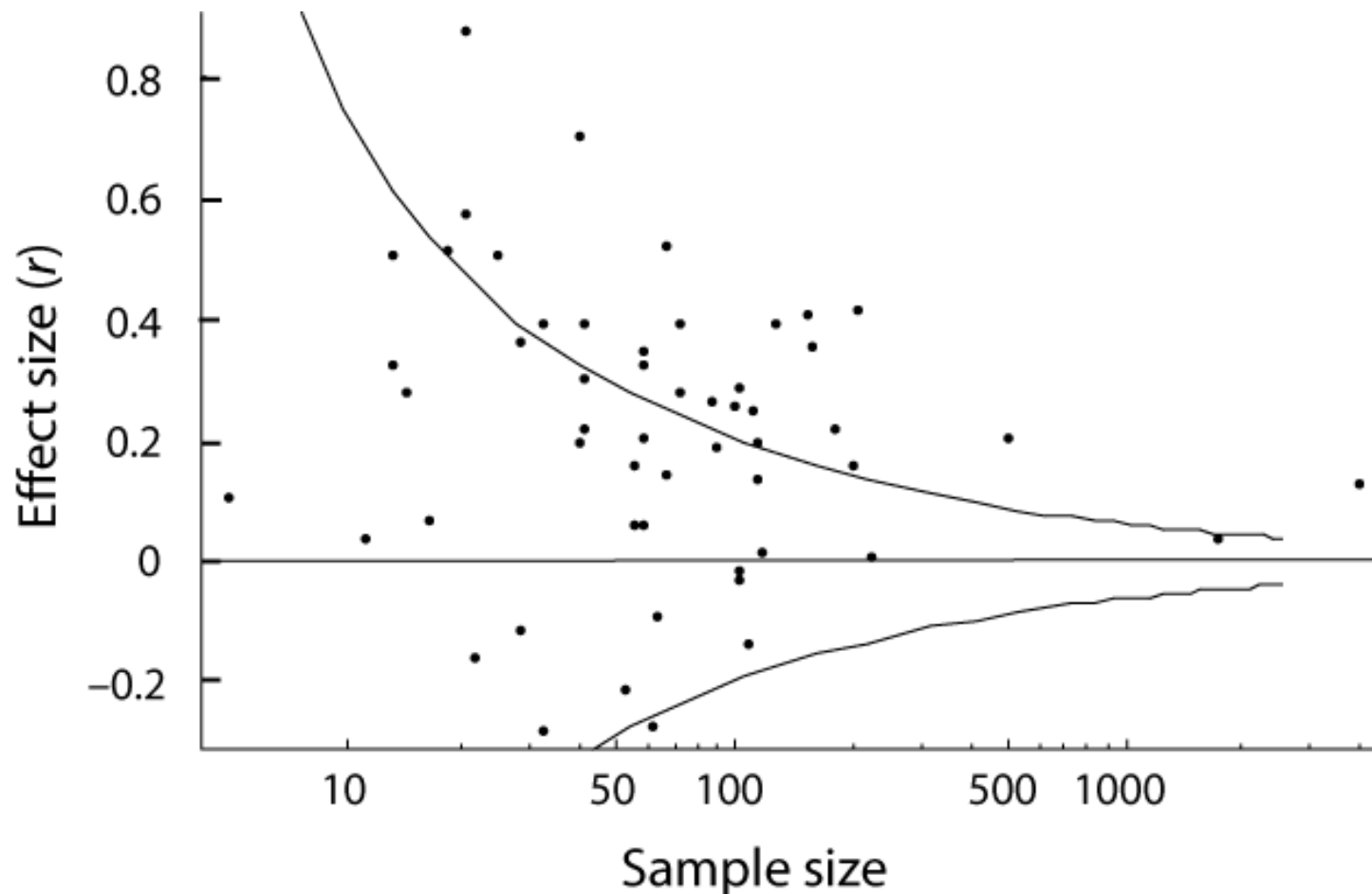
Example 2: Testosterone vs aggression

Book et al. (2001) asked “Are testosterone levels and aggression correlated in human males?” It included a huge diversity of types of studies:

- levels of testosterone in prisoners convicted of violent crimes compared to those of prisoners convicted of property crimes
- levels of testosterone in university students compared with their answers to questionnaires that asked them for levels of agreement to statements like “If somebody hits me, I hit back.”
- levels of aggression in !Kung San males as determined by counting “their scars and sometimes still open wounds in the head region.”
- drunken Finnish spouse-abusers compared to drunken Finns drinking quietly in a bar.
- members of “rambunctious” fraternities compared to “responsible” fraternities

Example 2: Testosterone vs aggression

Below is the “funnel plot” of studies comparing human aggression to levels of testosterone. The curves show the approximate boundaries of the critical regions that would reject the null hypothesis in any one study with $\alpha = 0.05$.



How to carry out a meta-analysis

3. Calculate an effect size that can be combined across studies to produce a quantitative summary of the findings.

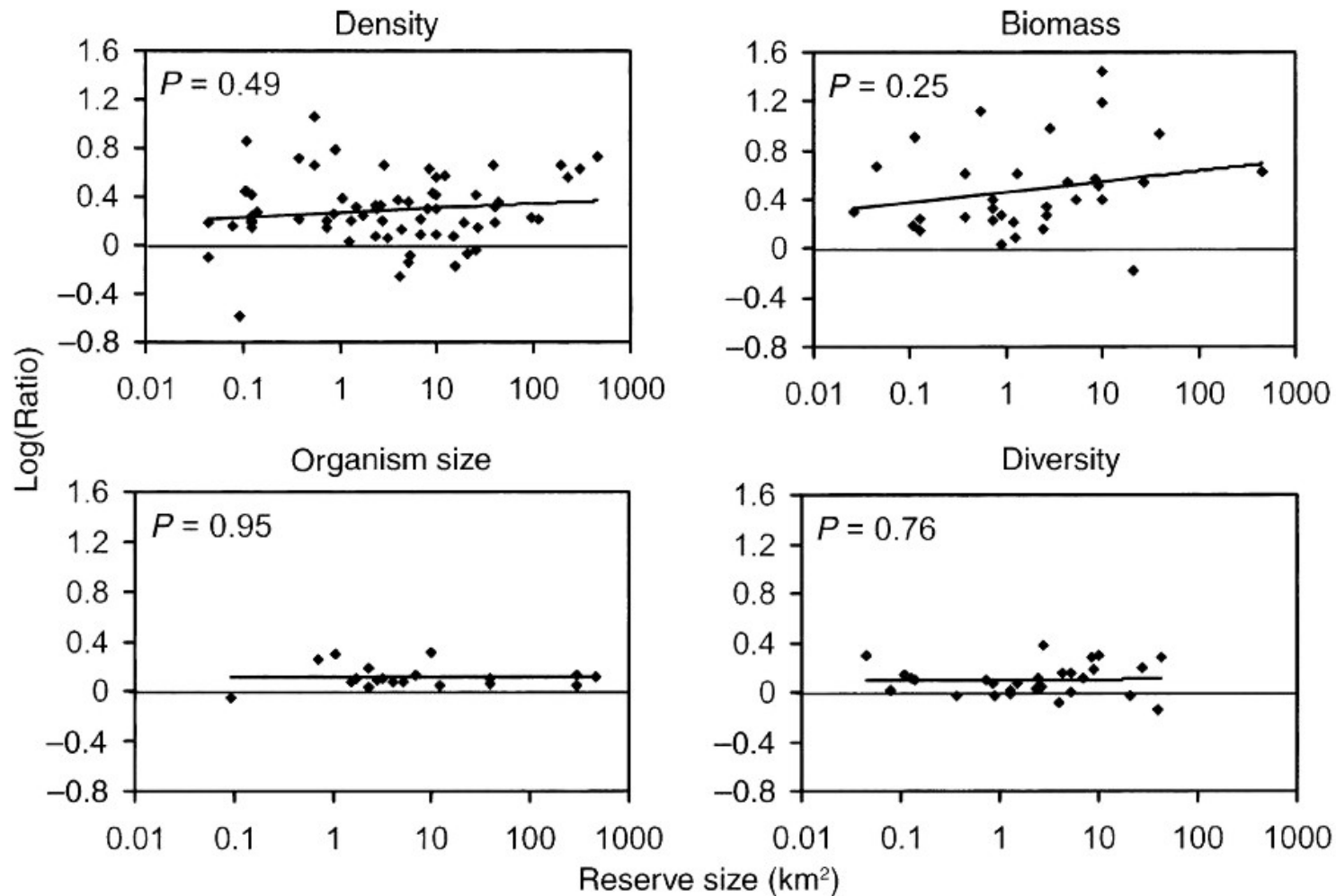
- Correlation coefficient r is commonly used though not always ideal, because effect size depends on the range of the data.
- Odds ratio – used in highly homogeneous studies (odds ratios in tests of aspirin and myocardial infarctions).
- Response ratio: $R = \bar{Y}_E / \bar{Y}_C$ or log of response ratio: $\ln(R)$
- Standardized mean difference, Cohen's d or Hedges' g :

$$g = \frac{\bar{Y}_E - \bar{Y}_C}{s_{\text{pooled}}} J(m)$$

s is the pooled sample variance and $J(m)$ is a small-sample bias correction.

Example 3: Effectiveness of marine reserves

Halpern (2003) used the log of response ratio to compare marine reserves to comparison areas (or the same area before reserve establishment) in abundance and diversity of fish and/or invertebrates



How to carry out a meta-analysis

4. Statistical inference on average effect size.

Fixed effects models

- Most commonly used in medical studies.
- Assumes that the multiple studies have the same mean, differing only because of sampling error. If every study were infinitely large, every study would yield an identical result. No heterogeneity among the studies.
- Perhaps never justified unless all studies conducted similarly and on the same species. This is rarely the case in ecology and evolution.

How to carry out a meta-analysis

4. Statistical inference on average effect size.

Random (mixed) effects models

- Random variation is present among means of studies in addition to sampling error.
- Individual studies are therefore estimating different treatment effects.
- Most interest is focused on the central value, or mean, of the distribution of effects.
- But the idea of a random effects meta-analysis is to understand the distribution of effects across different studies.

How to carry out a meta-analysis

Fixed effect model

Effect size of each study i is $Y_i = \Theta + \varepsilon_i$

where Θ is the one “true” effect size.

Random effect model

Effect size of study i is $Y_i = \mu + \zeta_i + \varepsilon_i$

where μ is the grand mean and ζ_i is the deviation of the “true” effect size of study i from the grand mean.

The difference affects how each study is weighed when calculating the average effect size over all studies. We’ll do this in the workshop.

Unfortunately, `lme` can’t be used for random effects meta-analysis in R, because it won’t calculate the necessary weights. Other packages (e.g., `meta`) are available.

How to carry out a meta-analysis

5. Look for effects of study quality. For example, are effect sizes different on average between studies that included blinding and those that did not?
6. Look for associations with variables that might explain heterogeneity of effect sizes among studies. For example, does the average effect size differ between studies carried out on women subjects and those on male subjects?

Example 4: Meta-analysis of competition in field experiments

Gurevitch et al (1992) study of inter- and intra-specific competition, looking only at studies published in 1980's

Vol. 140, No. 4

The American Naturalist

October 1992

A META-ANALYSIS OF COMPETITION IN FIELD EXPERIMENTS

JESSICA GUREVITCH, LAURA L. MORROW, ALISON WALLACE,
AND JOSEPH S. WALSH*

Department of Ecology and Evolution, State University of New York,
Stony Brook, New York 11794-5245

Example 4: Meta-analysis of competition in field experiments

They looked for effects of study quality

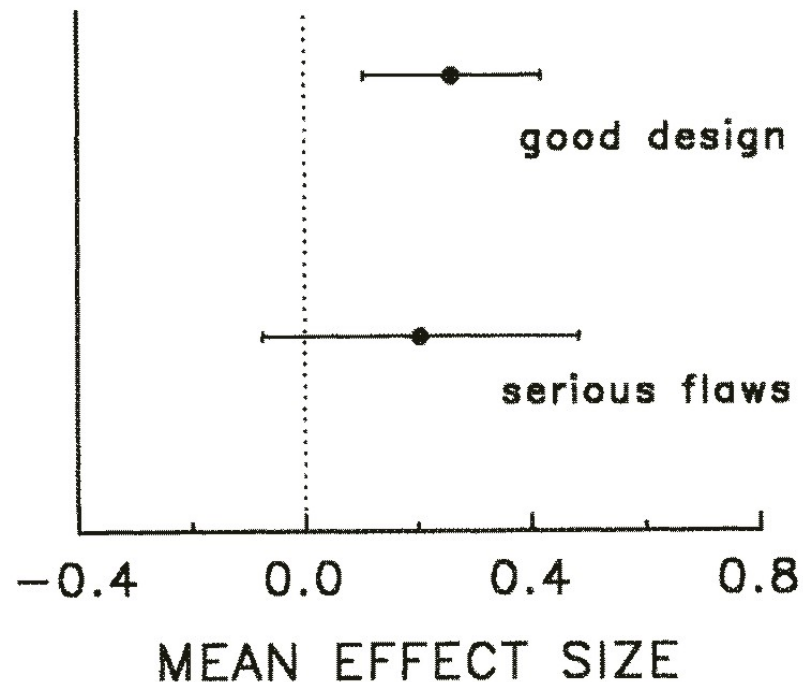


FIG. 12.—Mean effect size (d_+) and 95% CI for carnivores in experiments with good experimental designs or only minor design problems in contrast with those in experiments with serious problems in experimental design.

Example 4: Meta-analysis of competition in field experiments

They looked for associations with variables that might explain variation in effect size

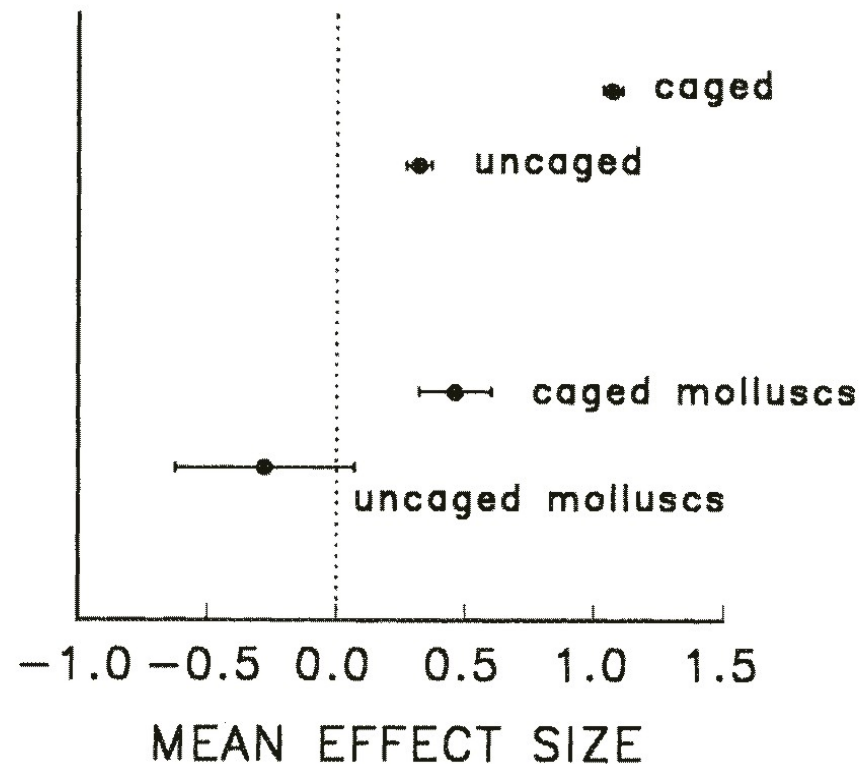


FIG. 7.—Mean effect size (d_+) and 95% CI of competition for all caged or enclosed organisms in contrast with all uncaged organisms (*top*) and mean effect size (d_+) of interspecific competition for marine mollusks in caged vs. uncaged trials (*bottom*).

File-drawer problem

In meta-analysis, the difficulties caused by publication bias are called the *file-drawer problem*, in reference to the unknown studies sitting unavailable in researchers' file drawers or hidden in obscure journals.

The *file-drawer problem* is the possible bias in estimates and tests caused by publication bias.

Funnel plot

Funnel plots can give an indication of the bias resulting from small studies.

Soma and Garamszegi (2011) used the Trimfill algorithm to fill in hypothetical missing studies in the funnel plot to achieve theoretical symmetry.

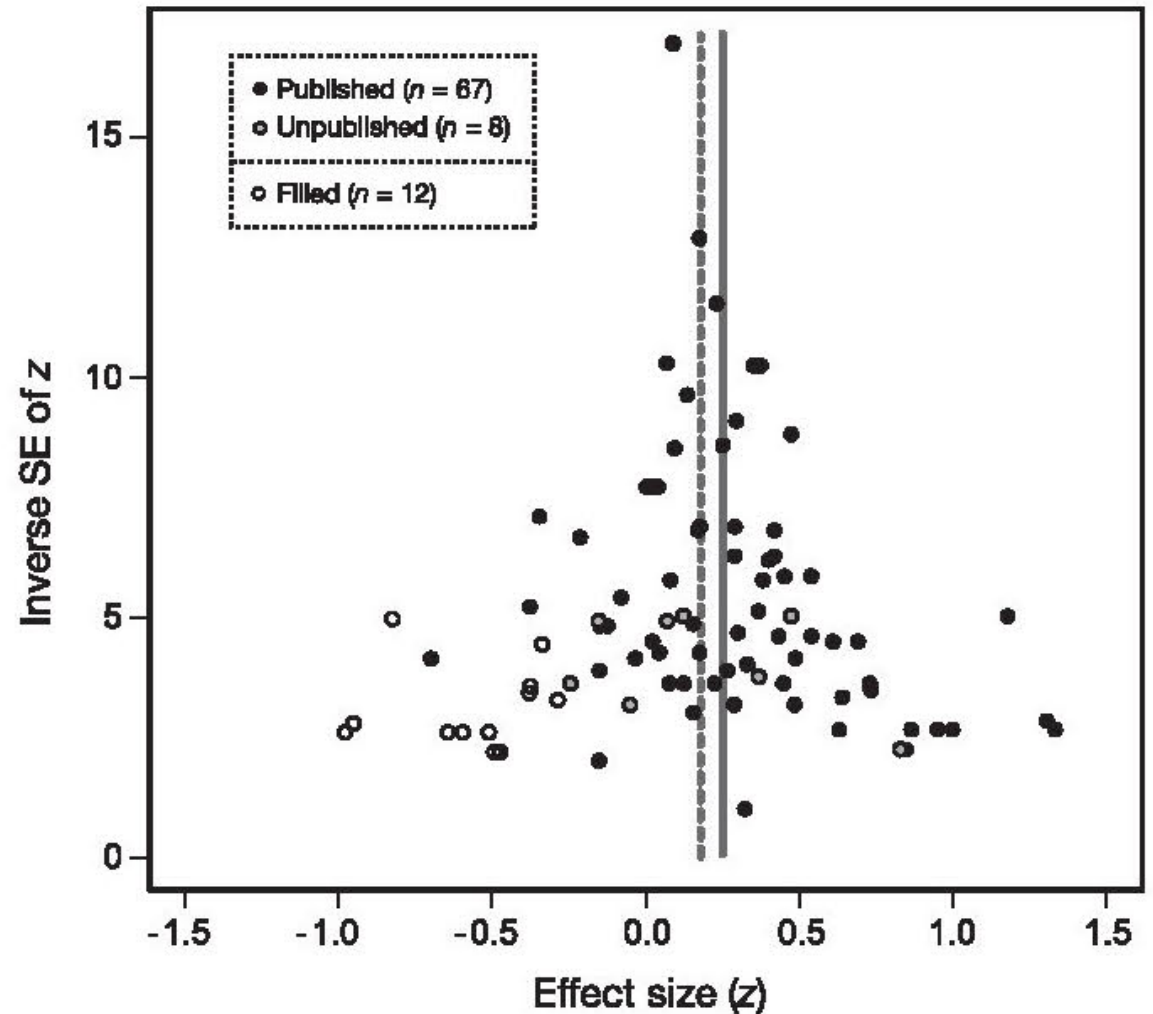


Figure 1

Funnel plot of effect sizes for the relationship between song complexity and reproductive success when using the multiple effect size data set (data set A). Black and gray circles show published and unpublished effect sizes, respectively. Solid and dotted lines show mean effect sizes before and after controlling for publication bias, in which theoretical missing data points (open circles, $n = 12$) were added to adjust funnel plot asymmetry.

Fail-safe number

The *fail-safe number* calculates how many missing studies would be needed to change the overall result of the meta-analysis.

Vilà et al (2011) estimated the number of studies that would have to be added to change the results of their invasive plant meta-analysis from significant to non-significant as 37,689. This was too implausible, so they concluded that their estimates were reliable.

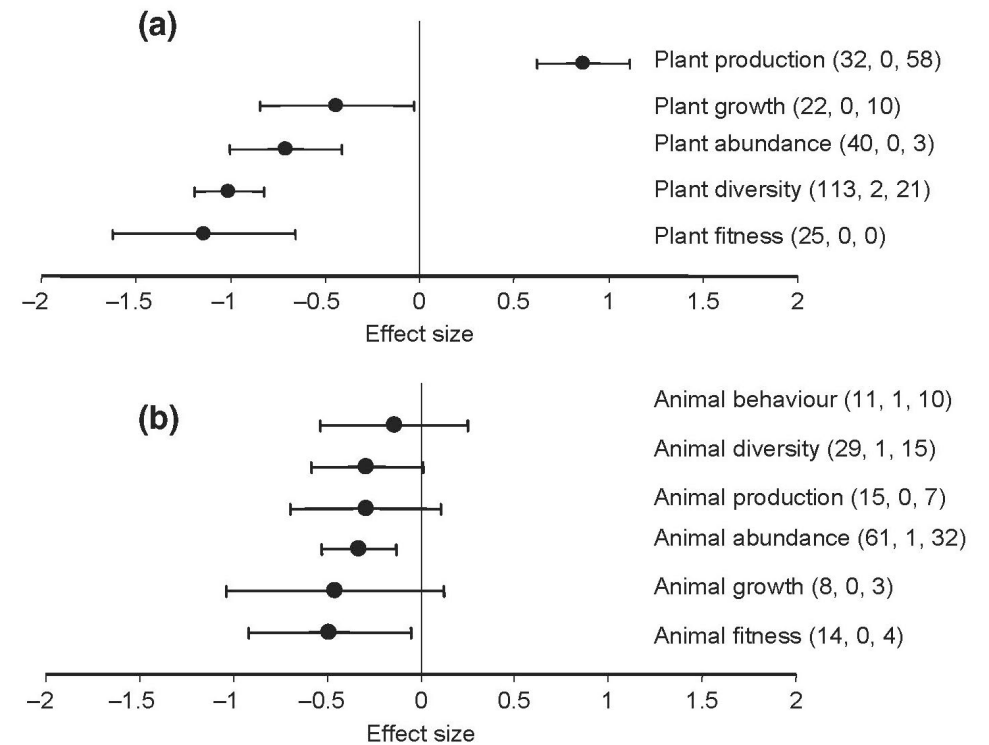


Figure 1 Mean effect size (Hedges' d) of differences between alien plant species impacts to (a) plant species and communities and (b) animal species and communities. The bars around the means denote bias-corrected 95%-bootstrap confidence intervals. A mean effect size is significantly different from zero when its 95% confidence interval do not bracket zero. Positive mean effect sizes indicate that the invaded plots had on average greater values for variables describing a particular impact type. The sample sizes with Hedges' $d < 0$, Hedges' $d = 0$ and Hedges' $d > 0$ are given next to the bars.

Make your results accessible to meta-analysis

Many published papers do not report enough information for meta-analysts to extract the numbers that they need. As a result, many otherwise relevant papers have to be discarded. Don't let this happen to your work.

- Always give sizes of effects and their standard errors. A *P*-value by itself is useless.
- Give estimates of the means and standard deviations of the important variables.
- Always indicate your sample sizes or degrees of freedom.
- Make the data accessible. Publish the raw data in the paper or deposit to an online archive such as Dryad.

Consider a meta-analysis for your first thesis chapter

Often, the first chapter of a thesis is a review of the literature. If your review is thorough, and you kept track of the important quantities and feature of each study, you may have enough for a quantitative review – your own meta-analysis.

Discussion paper next week:

Richman and Price: Evolution of ecological differences in warblers (multivariate analysis)

Download from “**Handouts**” tab on course web site.

Presenters: Mackenzie and Stephanie

Moderators: Sam J and Ellika