Introduction to meta-analysis

Outline for today

- Meta analysis compared with traditional review article
- Quantitative summaries compared with vote-counting
- Defining the question and scope for a meta-analysis
- Gathering the data and calculating effect size
- Fixed and mixed-effects models to analyze effect sizes
- Associating effect sizes with study quality, other variables
- File-drawer problem and publication bias
- Make your results accessible to meta-analysis

Scientific studies on a topic are often repeated

Later studies improve/expand on previous studies, examine the same issue in a different study system

- White et al. (2006) found 24 studies on whether acupuncture helps to quit smoking
- 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.

Q: How best to summarize the results from multiple studies?

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 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

“Vote-counting” is an improvement

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
- Method unable to evaluate bias or weigh studies differing in size

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 sample size was more than 70,000 patients.

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.

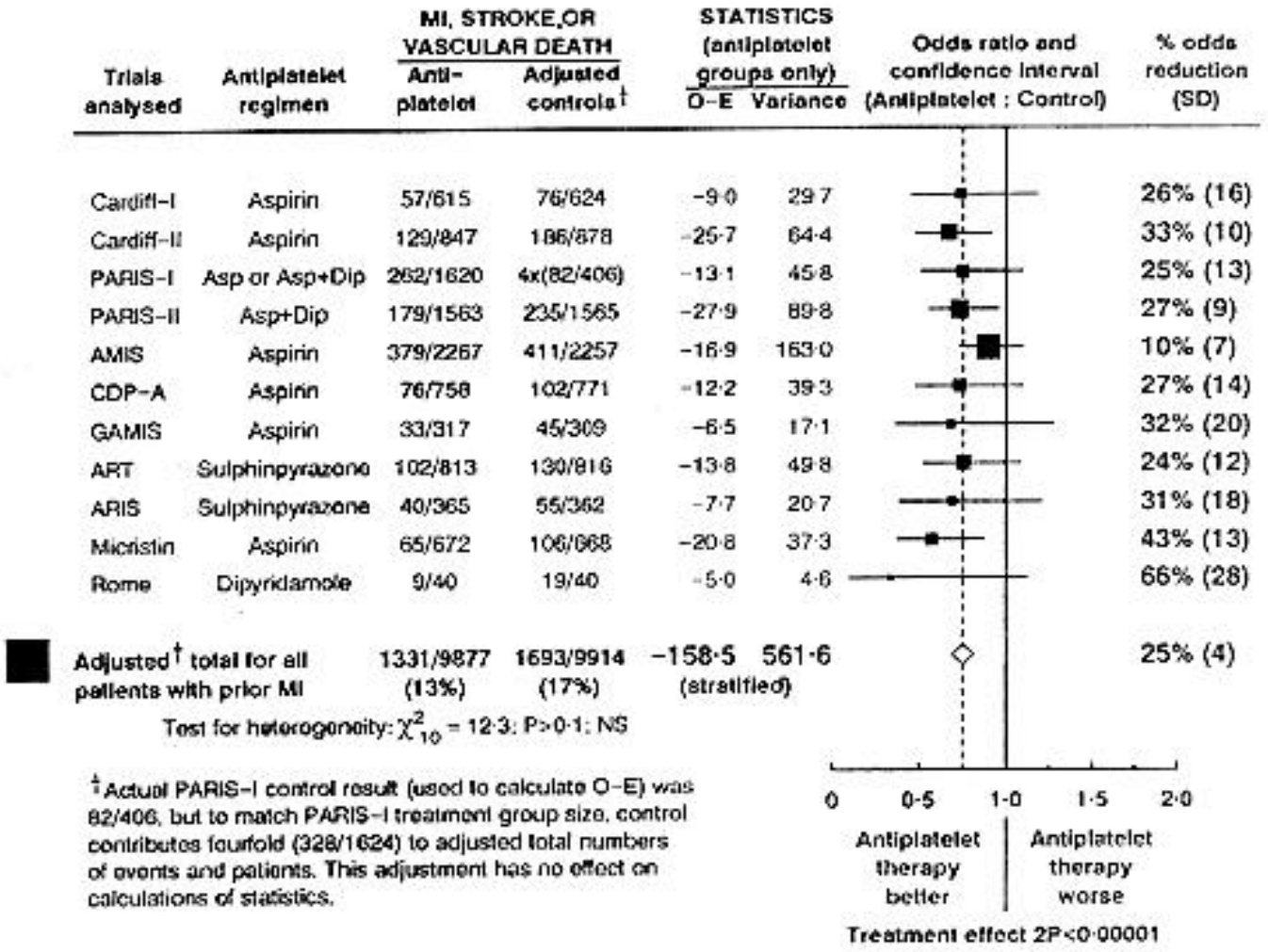
Meta-analysis, the “analysis of analyses”

A meta-analysis compiles all known scientific studies estimating or testing an effect and quantitatively combines 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, the "analysis of analyses"

The method comes from medical research where all studies are all of the same species (us).



Trials analysed	Antiplatelet regimen	MI, STROKE,OR VASCULAR DEATH: Anti-platelet	MI, STROKE,OR VASCULAR DEATH: Adjusted controls†	STATISTICS (antiplatelet groups only): O-E	STATISTICS (antiplatelet groups only): Variance	Odds ratio and confidence interval (Antiplatelet : Control)	% odds reduction (SD)
Cardiff-I	Aspirin	57/615	76/624	-9·0	29·7		26% (16)
Cardiff-II	Aspirin	129/847	186/878	-25·7	64·4		33% (10)
PARIS-I	Asp or Asp+Dip	262/1620	4x(82/406)	-13·1	45·8		25% (13)
PARIS-II	Asp+Dip	179/1563	235/1565	-27·9	89·8		27% (9)
AMIS	Aspirin	379/2267	411/2257	-16·9	163·0		10% (7)
CDP-A	Aspirin	76/758	102/771	-12·2	39·3		27% (14)
GAMIS	Aspirin	33/317	45/309	-6·5	17·1		32% (20)
ART	Sulphinpyrazone	102/813	130/816	-13·8	49·8		24% (12)
ARIS	Sulphinpyrazone	40/365	55/362	-7·7	20·7		31% (18)
Micristin	Aspirin	65/672	106/668	-20·8	37·3		43% (13)
Rome	Dipyridamole	9/40	19/40	-5·0	4·6		66% (28)
Adjusted† total for all patients with prior MI		1331/9877 (13%)	1693/9914 (17%)	-158·5 (stratified)	561·6		25% (4)

Test for heterogeneity: $\chi^2_{10} = 12{\cdot}3$; P>0·1; NS

†Actual PARIS-I control result (used to calculate O-E) was 82/406, but to match PARIS-I treatment group size, control contributes fourfold (328/1624) to adjusted total numbers of events and patients. This adjustment has no effect on calculations of statistics.

Meta-analysis, the “analysis of analyses”

However, ecologists and evolutionary biologists attempt to generalize across a much wider range of species and 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.

Example 1: Meta-analysis of the Transylvania effect

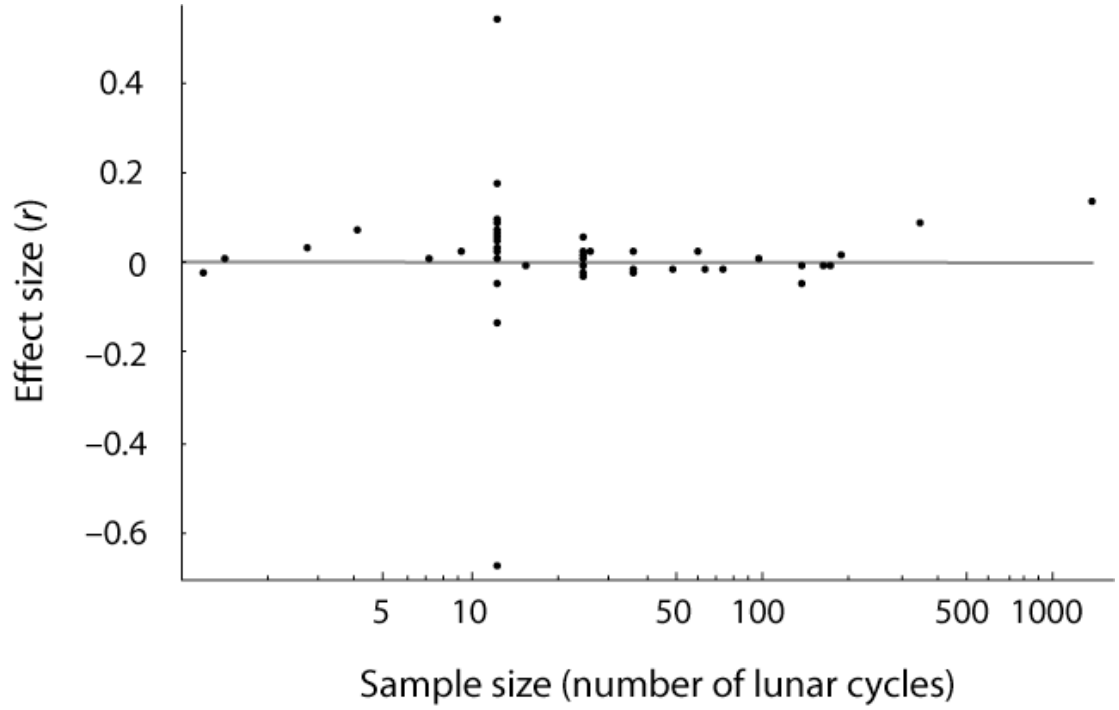
Many people believe that a full moon can affect human behavior. The word lunacy is derived from the Latin luna, moon, and legends of strange happenings, such as werewolves and vampires, have been connected to full moons for centuries.

In the 16th century, Paracelsus wrote that "*mania has the following symptoms: frantic behaviour, unreasonableness, constant restlessness and mischievousness. Some patients suffer from it depending on the phases of the moon*." 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*." During the 19th century, the German psychologist Ewald Hering observed in his textbook of psychiatry that "*with full moon, increasing mania*." At the Bethlehem (or Bedlam) Hospital in London, inmates were chained and flogged at certain phases of the moon "*to prevent violence*." This barbarous practice was abolished only in 1808 through the efforts of John Haslam, the hospital's apothecary. Benjamin Rush, the father of American psychiatry, kept accurate records of patients' conditions during the phases of the moon but observed a behavioural association in only "*few cases*".

A study by Rotton and Kelly in 1985 showed that 50% of university students believed that people act strangely during a full moon. In 1995, Vance 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 was r < 0.01.



[break]

Steps of a meta-analysis

1. Define the question and scope
    - Might be narrow question applied to a homogeneous group ("does aspirin reduce incidence of myocardial infarction").
    - Or might be a broad question applied to a heterogeneous set of studies or variables ("how much genetic variation exists in populations for behavioral traits").
    - Need to decide the scope of studies to be included. Only experiments with controls and randomization? Only replicated experiments? Only experiments with blinding?
    - Rather than define the question too narrowly, it may be better to adopt a reasonably wide scope and investigate later whether differences between methods lead to different effects overall.

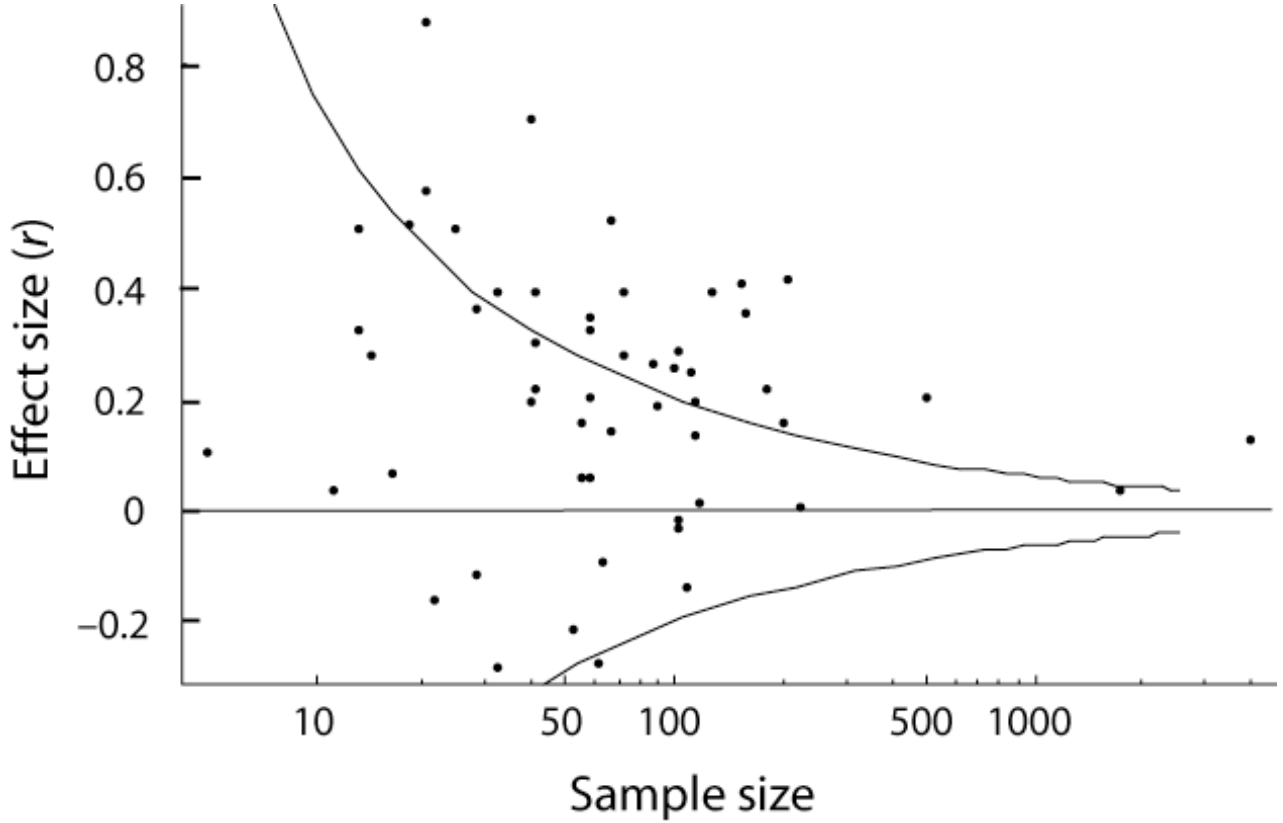
Example 2: Testosterone vs aggression

Book et al. (2001) asked “Are testosterone levels and aggression correlated in humans?” The studies reviewed for this meta-analysis were all done on humans, but 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 $\alpha = 0.05$.



Steps of a meta-analysis

2. Literature search, gather data

   - Want it to be exhaustive to avoid bias.
   - The studies that find large and 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. The studies that we can find easily are *different* from those that we cannot so easily find.
   - 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 also lead to bias (we are more likely to discard a poor study if it disagrees with our pet hypothesis).

- Ideally, the data obtained should all be independent, but non-independence of various sorts creeps in.
- 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 simply throw every data point into the analysis.

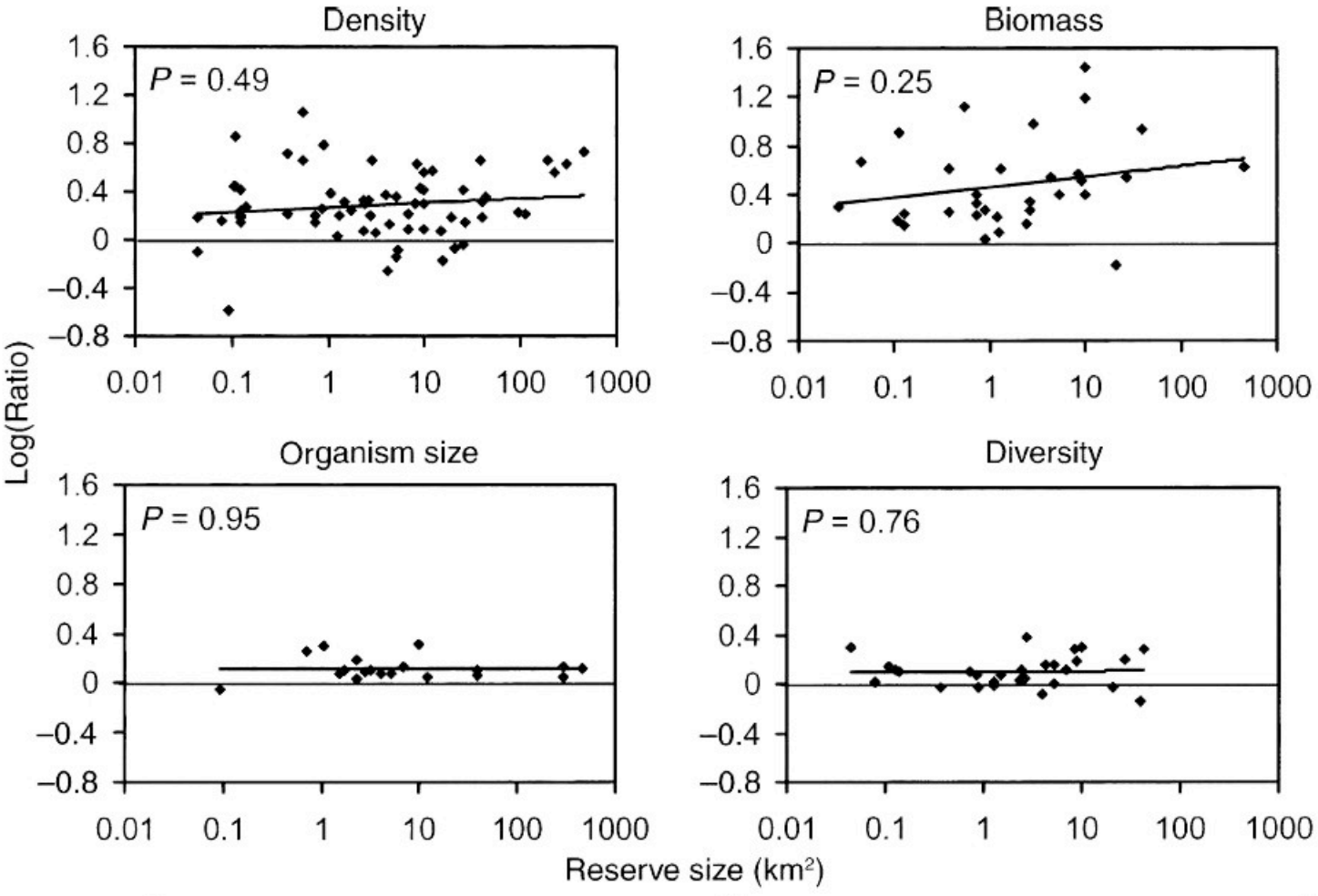
Steps of a meta-analysis

3. Effect size. The result of each independent experiment must be expressed as an index of effect that can be combined across studies to produce a summary of the findings.

   - Correlation coefficient – commonly used though not always ideal, because effect size *r* depends on the range of the data.
   - Odds ratio - useful in narrow, highly homogeneous studies (odds ratios in tests of aspirin and myocardial infarctions).
   - Response ratio: $R = \overline{Y}_E / \overline{Y}_C$, or $\ln(R)$
   - Hedges' g or standardized mean difference:
     $g = \frac{(\overline{Y}_E - \overline{Y}_C)}{s} J(m)$, where *s* is the pooled sample variance and J(*m*) is a small-sample bias correction
   - Many studies fail to provide enough info to calculate *R*, *g*

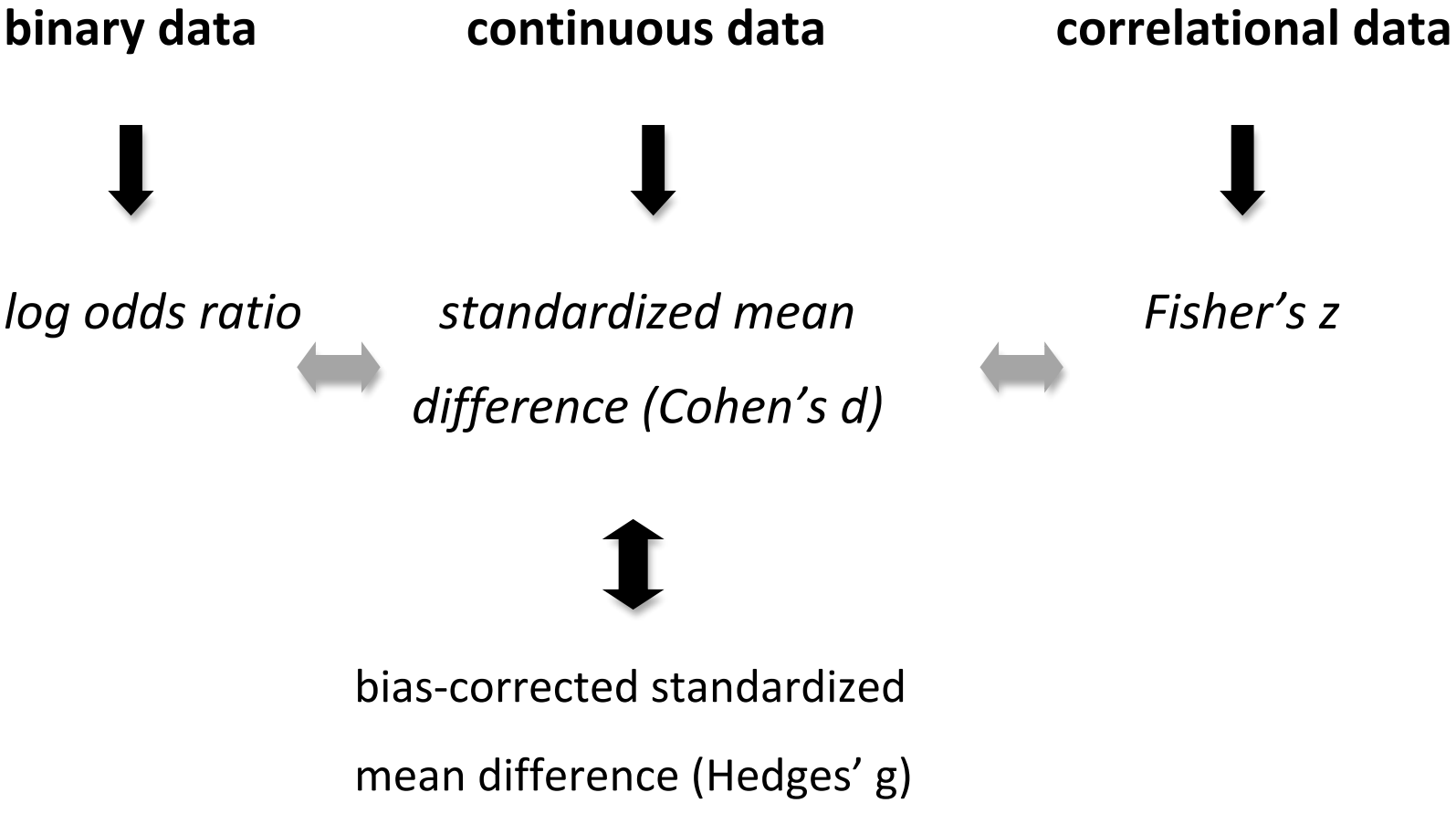
Example 3: Effectiveness of marine reserves

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



What if the effects sizes are in different measures?

Convert them!



Steps of a meta-analysis

4. Statistical inference on average effect size -- confidence limits and tests

Fixed effects models

- Most commonly used in medical studies.
- Assumes that the multiple studies of a given category (“plants”) have the same mean, differing only because of sampling error. If every study were infinitely large, every study would yield an identical result. There is no (statistical) 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.

Steps of a meta-analysis – statistical inference

Random (mixed) effects models

- Random variation is present among means of studies within a category, 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.

Steps of a meta-analysis – statistical inference

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* and the grand mean.

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

Unfortunately, `lme` can't be used to carry out a random effects meta-analysis in R, as it won't calculate the necessary weights. Other packages (e.g., `meta`) are available.

Steps of 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

Look for effects of study quality

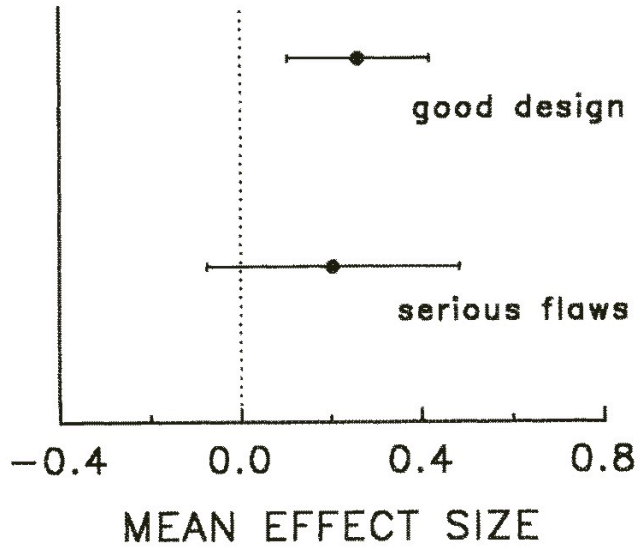


FIG. 12.—Mean effect size (d_+) and 95% CI for carnivores in experiments with good experimental designs or only minor design problems in contrast with those in experiments with serious problems in experimental design.

Example 4: Meta-analysis of competition in field experiments

Look for associations with variables that might explain variation in effect size

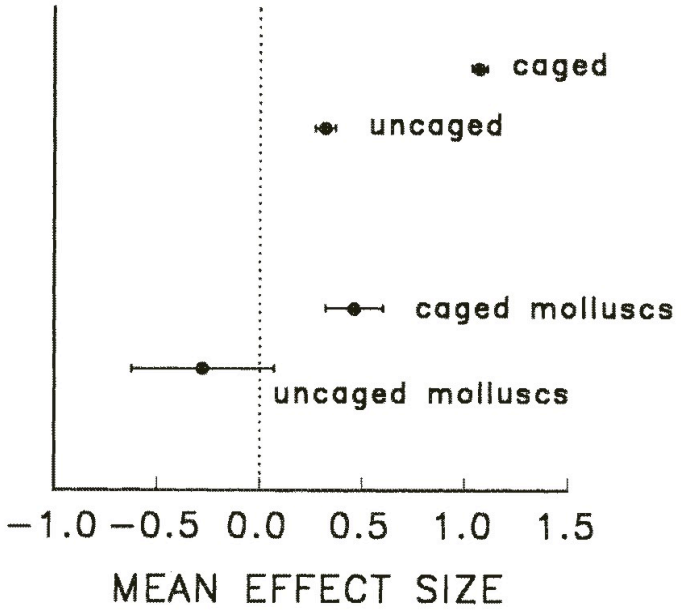


FIG. 7.—Mean effect size (d_+) and 95% CI of competition for all caged or enclosed organisms in contrast with all uncaged organisms (*top*) and mean effect size (d_+) of interspecific competition for marine mollusks in caged vs. uncaged trials (*bottom*).

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

File-drawer problem

A few methods are available to partially address the problem

- Funnel plots can give some indication of the bias resulting from small studies.
- The fail-safe number calculates how many missing studies would be needed to change the overall result of the meta-analysis. If the fail-safe number is small (i.e., roughly the same as the number of published studies included in the initial meta-analysis), then the results of that meta-analysis would be regarded as unreliable. If the fail-safe number is very large (e.g., in the millions), then we can be more certain that the meta-analysis is giving us the right answer—it is simply too unlikely that there are a million unpublished studies out there on the subject.

Make <u>your</u> 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. This difficulty can be avoided by a few simple changes in the way information is presented.

- Always give sizes of effects and their standard errors. A *P*-value by itself is useless.
- Give estimates of the mean and standard deviation 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 on an online archive.

Consider a meta-analysis for your first 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:

Surprise…

Last assignment has been put online: reproducibility presentation