Quantitative Research Syntheses: Meta-Analysis

Advanced Biostatistics

Dean Adams

Lecture 14

EEOB 590C

-Methods for quantitative research synthesis
-Brief history of methods for combining results from prior studies
-Vote-counting
-Combined probability method
-Meta-analysis

For further information on these approaches see:
Cooper and Hedges (1994). *Handbook of Research Synthesis.*
Hedges and Olkin (2000). *Statistical Methods for Meta-Analysis.*
Rosenberg, Adams, and Gurevitch (2000). *MetaWin: Statistical Software for Meta-Analysis.* Vsn 2.

-One important goal of science is synthesizing existing knowledge

-What does a body of literature say about a particular topic?

-Does existing published evidence support a particular hypothesis?

-Is there a general ‘consensus’ about the importance of a hypothesis?

-This is an obvious question to ask (what do we already know?)

-Literature reviews are common approach: usually narrative

-Other more quantitative methods exist

-Three main approaches:

-Vote-counting

-Combined probability methods

-Meta-analysis

-Quantitative research synthesis as old as modern statistics

-First QRS: Pearson (1904) calculated average correlation from several studies on effectiveness of typhoid vaccine

-Early 20th century: narrative reviews most common (and still are)

-1930's: several methods for combining probabilities developed (but infrequently used)

-1970's: 'modern' meta-analytic methods for combining effect sizes from independent studies developed by Glass (1976), Rosenthal etc.

-Currently, meta-analytic methods common in social sciences and medicine; use in ecology and evolutionary biology is increasing

-ANY research synthesis begins with a hypothesis (e.g., does smoking significantly increase cancer rates?)

-Published studies* are then obtained via a literature search (e.g., keyword search on Web of Science, Scholar.Google, Biological Abstracts, etc.)

-Unusable articles are discarded based on certain criteria (e.g., incomplete information)

-Remaining articles are reviewed and summarized in some way

*Note: unpublished studies that can be obtained from authors can also be included

-Begin with hypothesis and set of published studies

-Results from each study classified as 1 of 3 outcomes

- -Significant in expected direction
- -Significant in unexpected direction
- -Not significant

-Calculate proportion of each class, and that class with highest proportion represents the 'support' (for, against, equivocal)

-**Advantages**: quick and easy to calculate, intuitive

-**Disadvantages**: overly conservative, low statistical power (# non-significant findings > expected # significant findings), ignores magnitude of effects of studies, not sensitive to sample sizes (all studies treated equally)

-Begin with hypothesis and set of published studies (with significance levels)

-Combine probabilities in some way

-Many methods exist for various distributions (uniform, normal, t, X^2, etc: see Becker, 1994 in *Handbook of Research Synthesis*: Cooper & Hedges)

-**Advantages**: relatively easy to calculate, sample sizes taken into account (b/c use exact probabilities), general approach (can almost always obtain p-value from a study)

-**Disadvantages**: don't directly assess magnitude of study effects, cannot assess direction of effects, cannot assess whether effects are homogeneous

-Often called *omnibus* tests (only depend on exact probabilities of each study)

-Minimum P method (Tippet, 1931): uses uniform distribution, significant if any study is significant at α-level:

$$\alpha = 1 - \left(1 - \alpha^{*}\right)^{1/n}$$

-Sum of logs method (Fisher, 1932): uses inverse X2 distribution, significant if < 0.05 from X^2 with 2n df (*n* is # studies and *pi* are study significance levels)

$$P = -2\sum_{1}^{n} \log\left(p_i\right)$$

-Sum of Z method (Stouffer et al., 1949): use normal distribution, significant if probability of < 0.05 (*Z(pi)* are Z-scores for study p-values)

$$Z = \sum_{1}^{n} Z\left(p_i\right) \Big/ \sqrt{n}$$

-Sum of p method (Edgington, 1972): uses uniform distribution, significant if < 0.05 (n is # studies and pi are study significance levels)

$$P = \left(\sum_{1}^{n} p_i\right)^{n} \Big/ n!$$

-**Sum of logs method** (Fisher, 1932): uses inverse X2 distribution, significant if < 0.05 from X2 with 2n df (n is # studies and pi are study significance levels)

Pi: 0.06; 0.02; 0.035; 0.001; 0.24

Log(pi): -1.22; -1.70; -1.46; -3; -0.62

$-2\Sigma(pi) = 16$; $P_X2 = 0.096$ NS

-Approach that combines weighted effect sizes for each study to assess overall significance

-Allows the interpretation of the ***strength*** of the statistical finding, not just whether or not there is significance

-M-A model can be generalized to address more complicated synthesis questions

Requires calculating an ***effect size*** and ***weight*** for each study

-Meta-analysis has two steps:

-Calculate effect sizes (and weights) for each study

-Summarize effect sizes to address hypothesis (m-a model)

-Effect size: statistical measure of the magnitude of factor in the data (*how much* does smoking increase cancer rates?)

-Different types of primary data require different effect size estimates (some data types have several possible effect sizes)

-Many test statistics are a form of effect size (e.g., $t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma}$ is a standardized mean difference effect size)

-Use of effect sizes in QRS is desirable because they 'standardize' results from independent studies and express them in a common way (i.e., all results expressed as t-values)

-Weights are inverse of effect size variance: $w = \frac{1}{v}$

Effect sizes are often transformed so range is - ∞ to +∞

-Powerful effect sizes, but require much data from studies

-Require means, sample sizes, and std from experimental and control group

($\overline{X}_C$ & $\overline{X}_E$, s_C & s_E, N_C & N_E)

-Many are variants on *standardized mean difference* (like t-test)

Name/s	Equation	Variance
Glass' Δ	$\Delta = \frac{\left(\overline{X}^E - \overline{X}^C\right)}{s^C}$	$v_\Delta = \frac{N^C + N^E}{N^C N^E} + \frac{\Delta^2}{2\left(N^C - 1\right)}$
Hedges' g	$g = \frac{\left(\overline{X}^E - \overline{X}^C\right)}{S}$	$v_g = \frac{N^C + N^E}{N^C N^E} + \frac{g^2}{2\left(N^C + N^E - 2\right)}$
Cohen's d	$d_{Cohen} = \frac{\left(\overline{X}^E - \overline{X}^C\right)}{\sigma}$	$v_{d-Cohen} = \left(\frac{N^C + N^E}{N^C N^E} + \frac{d^2}{2\left(N^C + N^E - 2\right)}\right)\left(\frac{N^C + N^E}{N^C + N^E - 2}\right)$
Hedges' d	$d = \frac{\left(\overline{X}^E - \overline{X}^C\right)}{S} J$	$v_d = \frac{N^C + N^E}{N^C N^E} + \frac{d^2}{2\left(N^C + N^E\right)}$
response ratio	$\ln R = \ln\left(\frac{\overline{X}^E}{\overline{X}^C}\right)$	$v_{\ln R} = \frac{\left(s^E\right)^2}{N^E\left(\overline{X}^E\right)^2} + \frac{\left(s^C\right)^2}{N^C\left(\overline{X}^C\right)^2}$

-Common in medicine: for data summarized by 2 X 2 table

	Treatment	Control	Total
Response	A	B	A + B
No Response	C	D	C + D
Total	n_t = A + C	n_c = B + D	N= A + B + C + D

-From table calculate $P_t = \frac{A}{n_t}$ and $P_c = \frac{B}{n_c}$: used for effect sizes

Name/s	Equation	Variance
rate difference *risk difference*	$RD = P_t - P_c$	$v_{RD} = \frac{P_t(1-P_t)}{n_t} + \frac{P_c(1-P_c)}{n_c}$
relative rate *risk ratio* *rate ratio* 	$RR = \frac{P_t}{P_c}$	$v_{\ln RR} = \frac{(1-P_t)}{n_t P_t} + \frac{(1-P_c)}{n_c P_c}$
odds ratio *relative odds* 	$OR = \frac{P_t(1-P_c)}{P_t(1-P_t)}$	$v_{\ln OR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$

-Useful when only summary statistics are available

-Convert all test-statistics to correlations, then convert these to Fisher's Z-transform: $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ variance: $v_z = \frac{1}{n-3}$

-Common transformations

statistic	conversion
Z^*	$r = \frac{Z}{\sqrt{N}}$
t	$r = \sqrt{\frac{t^2}{t^2 + df}}$
F	$r = \sqrt{\frac{F}{F + df}}$
χ^2	$r = \sqrt{\frac{\chi^2_{(1)}}{N}}$

*Probabilities can be converted to Z as standard normal deviates

-Summarize effect sizes to assess significance
-Standard statistical summary variables: mean, variance
-Cumulative Effect Size: weighted mean of effect sizes
-Homogeneity Statistic: Quantifies variation in effect sizes (analogous to SS) Are effect sizes homogeneous?

-Method of summary depends upon model for effect size variation
    -No structure: all studies belong to one 'population'
    -Categorical structure: studies belong to groups
    -Continuous structure: studies covary with continuous variable

-For models with structure (categorical, continuous), variables are often called moderator variables (groups, covariate, etc.)
-All models are actually special cases of same model

Model: All studies belong to same group

Example H_0: **Is there an effect of competition on plant communities?**

Cumulative effect size: $\bar{\bar{E}} = \sum_{i=1}^{n} w_i E_i \Big/ \sum_{i=1}^{n} w_i$ variance: $s^2_{\bar{\bar{E}}} = \frac{1}{\sum_{i=1}^{n} w_i}$

$CI = \bar{\bar{E}} \pm t_{\alpha/2[n-1]} * s_{\bar{\bar{E}}}$: $\bar{\bar{E}}$ significant if it CI does not bracket 0.0

Homogeneity: $Q_T = \sum_{i=1}^{n} w_i E_i^2 - \frac{\left(\sum_{i=1}^{n} w_i E_i\right)^2}{\sum_{i=1}^{n} w_i}$ or $Q_T = \sum_{i=1}^{n} w_i \left(E_i - \bar{\bar{E}}\right)^2$

Test against X^2 (n-1 df)

Significant Q_T implies samples are NOT homogeneous

Implies structure in data: may be captured by a moderator variable

Model: Studies belong to different groups

Example H_0: **Does competition differ among habitats (terrestrial, marine, etc.)?**

For each group calculate: $\overline{E}_j = \frac{\sum_{i=1}^{k_j} w_{ij} E_{ij}}{\sum_{i=1}^{k_j} w_{ij}}$ $\quad s^2_{\overline{E}_j} = \frac{1}{\sum_{i=1}^{k_j} w_{ij}}$ $\quad CI = \overline{E}_j \pm t_{\alpha/2[k_j-1]} * s_{\overline{E}_j}$

$Q_{Wj} = \sum_{i=1}^{k_j} w_{ij} \left(E_{ij} - \overline{E}_j\right)^2$ Test if each group is different from zero

Test if groups differ: $Q_T = Q_M + Q_E$ $\quad Q_E = \sum_{j=1}^{m} Q_{Wj} = \sum_{j=1}^{m} \sum_{i=1}^{k_j} w_{ij} \left(E_{ij} - \overline{E}_j\right)^2$

$Q_M = \sum_{j=1}^{m} \sum_{i=1}^{k_j} w_{ij} \left(\overline{E}_j - \overline{\overline{E}}\right)^2$ test Q_M vs. X^2 with m-1 df, where m is # groups

Significant Q_M implies groups are different (significant Q_E implies there is still structure remaining)

Model: Study effect sizes covary with continuous variables

Example H_0: **Does competition intensity change with age?**

Use Weighted GLM: $E_i = b_o + b_1 X_i + \varepsilon$

$$b_1 = \frac{\sum_{i=1}^{n} w_i X_i E_i - \frac{\sum_{i=1}^{n} w_i X_i \sum_{i=1}^{n} w_i E_i}{\sum_{i=1}^{n} w_i}}{\sum_{i=1}^{n} w_i X_i^2 - \frac{\left(\sum_{i=1}^{n} w_i X_i\right)^2}{\sum_{i=1}^{n} w_i}}$$

$$b_0 = \frac{\sum_{i=1}^{n} w_i E_i - b_1 \sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

This is weighted regression

Homogeneity: $Q_M = \frac{b_1^2}{s_{b_1}^2}$ (Q_M vs. X2 with 1 df)

Significant Q_M implies X explains significant component of variation in *E*

-What are we doing? Summarizing effect sizes as if 'primary' data

-If w_i = 1.0, then we're calculating standard means & SS

$$\bar{\bar{E}} = \sum_{i=1}^{n} w_i E_i \Big/ \sum_{i=1}^{n} w_i$$

$$Q_T = \sum_{i=1}^{n} w_i \left(E_i - \bar{\bar{E}} \right)^2$$

Also note, Q_T is partitioned, just like SS

-Therefore, think of meta-analysis as ANOVA, regression, etc.

-Meta-analytic models are actually **Weighted GLM**

-Weighted GLM is a standard statistical method used to account for different weights of objects (recall PGLS for phylogeny)

Since meta-analysis is analyzed in this general framework, more complicated designs can also be tested (e.g., ANCOVA, 2-factor ANOVA, etc.)

-Represent analyses using standard matrix algebra

$$\mathbf{E} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{E} = \begin{bmatrix} E_1 \\ \vdots \\ E_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ \vdots & \vdots & & \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix}$$

(For no structure, **X** is vector of 1's)

$$\mathbf{W} = \begin{bmatrix} w_i & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_n \end{bmatrix}$$

W 'in' error term (w_i inverse of variance)

-Solve model as: $\beta = \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{E}$

-Q_T, Q_M, etc. calculated as weighted SS

-Allows for simple-complicated designs

-Can be generalized to multivariate (though multivariate effect sizes nearly impossible to obtain for a set of published studies!)

-All previous models are 'fixed effects' models

-Fixed-effects model: assume only one true effect size shared by all studies (studies therefore only differ by sampling error)

-Random-effects model: assume studies differ by sampling error and random component (pooled study variance: σ^2_{pooled})

σ^2_{pooled} found from running a fixed effects model

σ^2_{pooled} is incorporated in weights for random model: $w_{i(rand)} = \frac{1}{v_i + \sigma^2_{pooled}}$

$$\sigma^2_{pooled} = \frac{Q_T - (n-1)}{\sum_{i=1}^{n} w_i - \frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i}}$$

No Structure

$$\sigma^2_{pooled} = \frac{Q_E - (n-m)}{\sum_{j=1}^{m}\left(\sum_{i=1}^{k_j} w_{ij} - \frac{\sum_{i=1}^{k_j} w_{ij}^2}{\sum_{i=1}^{k_j} w_{ij}}\right)}$$

Categorical Model

$$\sigma^2_{pooled} = \frac{Q_E - (n-2)}{\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i^2 \left(\frac{\sum_{i=1}^{n} w_i X_i^2 - 2X_i \sum_{i=1}^{n} w_i X_i + X_i^2 \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i X_i^2 - \left(\sum_{i=1}^{n} w_i X_i\right)^2} \right)}$$

Continuous Model

-Competition in biological communities (Gurevitch et al., 1992)

-Subset of data (N=43) from 3 habitats (terrestrial, lentic, marine)

-Data: mean, std, n data from experiment/control

-H_0: Does competition differ among habitats?

Part of data

Study	Habitat	Nc	Ne	Xc	Xe	Sc	Se	d	var(d)
1	Terrestrial	7	7	78.14	79.71	40.65	40.65	0.0362	0.2858
2	Terrestrial	7	7	18.86	26	9.17	9.17	0.7289	0.3047
3	Terrestrial	6	6	-1.8	-2.1	0.49	0.49	0.5651	0.3466
4	Terrestrial	5	5	-2.2	-2.8	0.224	0.447	1.5329	0.5175
5	Terrestrial	7	7	-2.1	-3	0.265	0.529	2.0139	0.4306
6	Terrestrial	6	6	-2.3	-4.2	0.49	1.225	1.8799	0.4806
7	Terrestrial	3	3	85.3	285.7	115.008	153.806	1.1806	0.7828
8	Terrestrial	3	3	0	3	0	2.425	1.3996	0.8299
9	Terrestrial	3	3	0	2	0	2.078	1.0889	0.7655
10	Terrestrial	3	3	0	1.67	0	1.732	1.0909	0.7658
11	Terrestrial	5	5	17	17	7.603	5.367	0	0.4
12	Terrestrial	5	5	47	37	10.286	9.391	-0.9171	0.4421
13	Terrestrial	4	4	87	272	37.712	183.532	1.2142	0.5921
14	Terrestrial	18	20	-0.113	0.294	0.255	0.215	1.6975	0.1435
15	Terrestrial	20	20	-0.163	0.412	0.588	0.218	1.2709	0.1202
16	Terrestrial	18	20	0.14	0.632	0.38	0.359	1.3051	0.128
17	Terrestrial	20	20	-0.184	0.259	0.326	0.238	1.5213	0.1289
18	Terrestrial	20	20	-0.075	0.354	0.487	0.182	1.1438	0.1164
19	Terrestrial	20	20	0.147	0.541	0.34	0.299	1.2062	0.1182
20	Lentic	4	4	281.11	-201.03	158.038	27.52	3.6961	1.3538

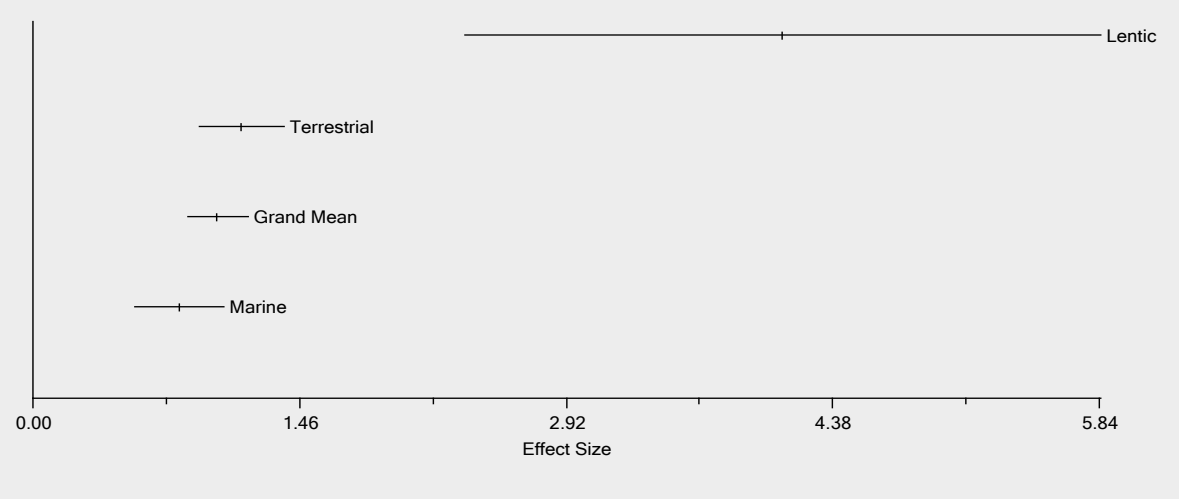
-*E*: Group effect sizes differed from zero (except lentic)

Q_M: Effect sizes differed among groups

Conclusion: competition occurs and differs among habitats

Group	#Studies	E+	df	95% CI
Terrestrial	19	1.1417	18	0.8999 to 1.3
Lentic	2	4.1072	1	-7.1465 to 15.3609
Marine	22	0.7985	21	0.5419 to 1.0550
E++	43	1.0099		0.8408 to 1.1789

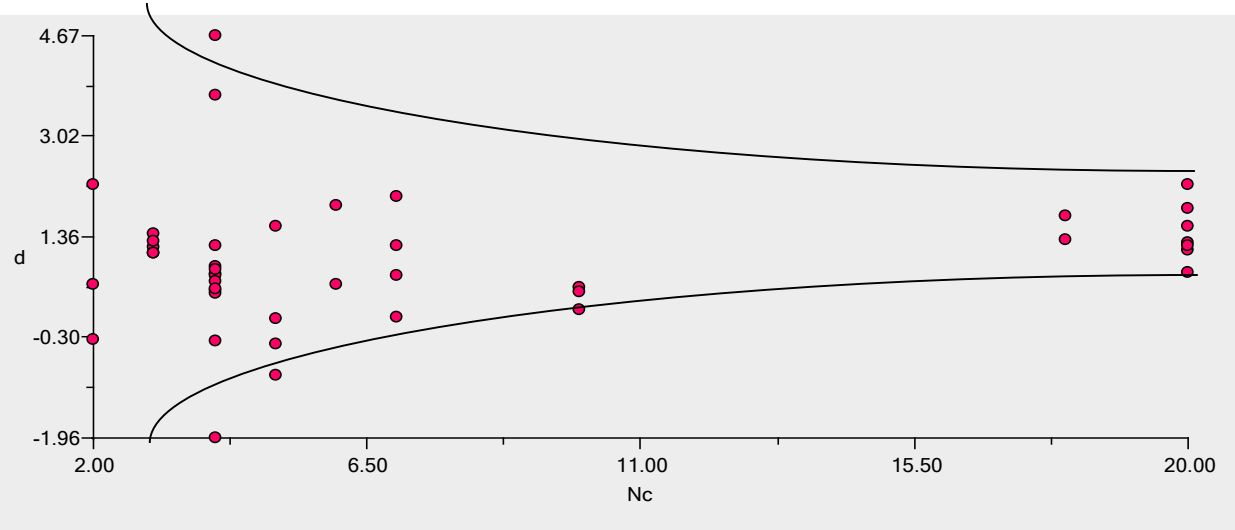
Model	df	Q	Prob(Chi-Square)
Between	2	16.4798	0.00026
Within	40	69.5016	0.00262
Total	42	85.9814	0.00007



-Common concern is that only studies with significant results get published, resulting in bias

-Can be assessed in a number of ways:

-**Funnel Plot**: plot effect size vs. sample size: should be funnel shaped (larger variance with smaller n). If overabundance of extreme values (for given n) with lack of data 'in' funnel, might be publication bias



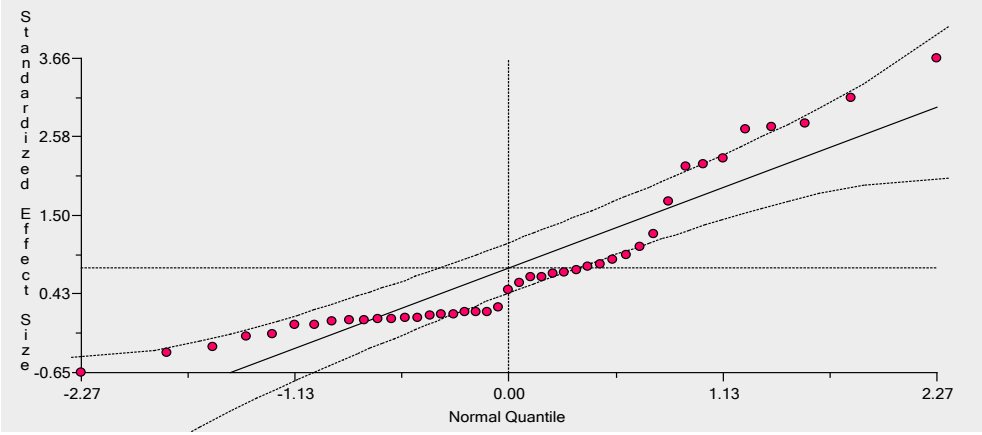
-Rank-Correlation Tests: Look at rank-correlation of standardized effect size vs. sample size $E_i^* = \frac{\left(E_i - \overline{\overline{E}}\right)}{\sqrt{v_i^*}}$ where $v_i^* = v_i - \left(\sum 1/v_j\right)^{-1}$

-**Fail-Safe Numbers**: For the 'file drawer problem'. How many non-significant studies must be added to change result to non-significant (if large #, then result is robust)

$$N_R = \frac{\left(\sum_1^n Z(p_i)\right)^2}{Z_\alpha^2} - n$$

N: # studies, Z(pi): Z-scores for study significance values, Za: 1-tail probability

-**Normal Quantile Plot**: Standardized effect size vs. normal quantile (gaps or strange nonlinearities may indicate publication bias)

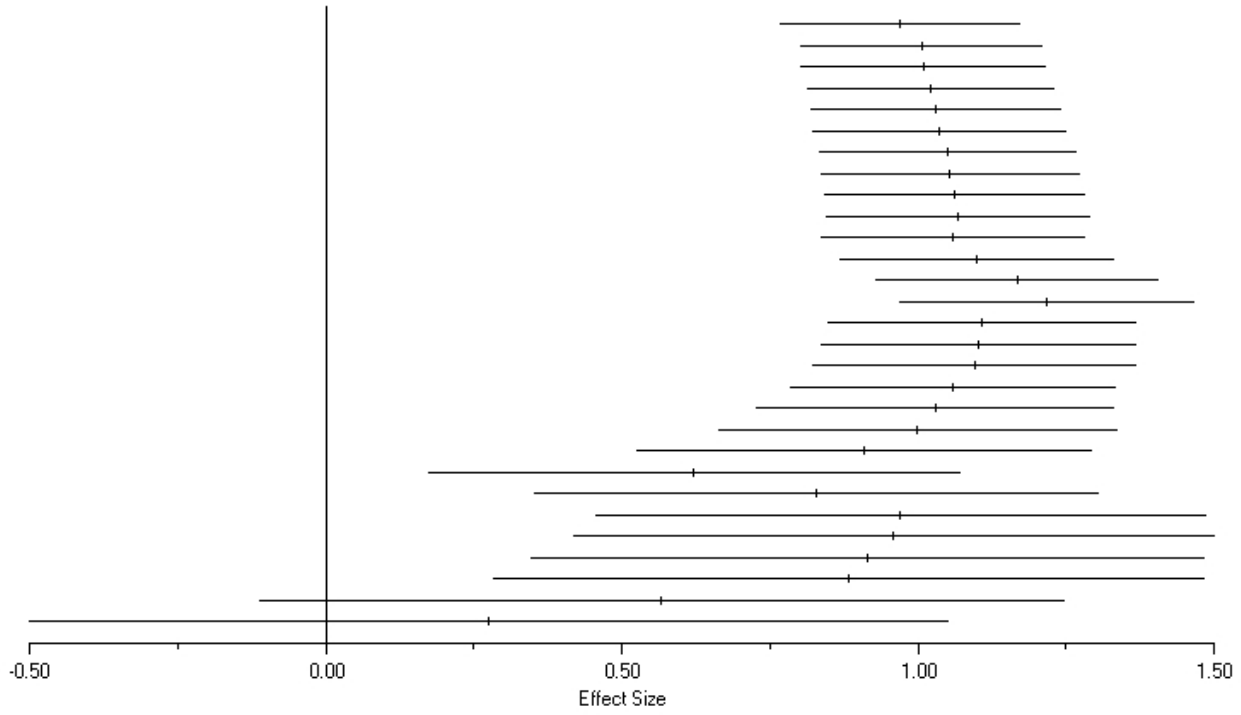


-Rank studies by some criterion (e.g., year of publication)

-Perform meta-analysis on 1st 2 studies, then 1st 3, 1st 4, etc.

-Plot cumulative effect sizes (with CI)

-Addresses *when* a synthesized result could be determined



Adams et al., 1997 (*Ecology*) proposed some resampling methods

Randomization for assessing significance of Q-statistics

Bootstrapping for assessing CI of cumulative effect sizes

Removes assumptions of testing vs. X^2 distribution

Adams et al. 1997. *Ecology.* 78:1277-1283.
See also: Rosenberg, Adams, Gurevitch. 2000. *MetaWin.* Sinauer Assoc.

-When studies come from a set of related taxa, phylogenetic non-independence is an issue

-Phylogenetic meta-analysis recently developed (Adams, 2008)

-Both PGLS and meta-analysis are GLS models, so can be combined

M-A: $\beta = \left(\mathbf{X}^t \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{W} \mathbf{E}$

PGLS: $\beta = \left(\mathbf{X}^t \mathbf{\Sigma}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{\Sigma}^{-1} \mathbf{Y}$

Steps

-SVD of S: obtain transformation matrix (D) [see Garland and Ives, 2000. Am. Nat.]

-Transform X and E as: $\mathbf{E}_{new} = \mathbf{DE}$ $\mathbf{X}_{new} = \mathbf{DX}$

-Solve meta-analysis with transformed data

$$\beta_{p-m-a} = \left(\mathbf{X}_{new}^t \mathbf{W} \mathbf{X}_{new}\right)^{-1} \mathbf{X}_{new}^t \mathbf{W} \mathbf{E}_{new}$$

NOTE: this is a fixed effects, Brownian motion model

(method generalized by Lajeunesse, 2009)

Adams. 2008. *Evolution.* 62:567-572.

Also: Lajeunesse. 2009. *Am. Nat.* 174:369-381.