WORKSHOP

Meta-Analysis:
A Tool for Evidence Synthesis in Plant Pathology

L. V. Madden
P. A. Paul

Ohio State University

Meta-analysis workshop : Outline

- Basic concepts (with demonstration)
  - A little history and the goals of meta-analysis
  - Effect sizes: results from multiple studies
  - Graphical appraisal of the effect sizes
  - Fixed vs. random-effect meta-analysis
  - Heterogeneity among studies, impact of heterogeneity
  - Confidence intervals, prediction intervals
- Example: use of SAS procedures and macros
- Risk prediction
- *Power* of meta-analysis
  - Fallacy of counting *P* values (avoid vote counting)
- *Bias* of meta-analysis and how to assess
- Moderator variables in a meta-analysis

Genesis of Meta-Analysis

- The psychotherapy debate (1952-1977)
- Glass (1976); Smith & Glass (1977)
  - "**META-ANALYSIS**"
- Rosenthal; Rosenthal & Rubin (1978)
- Schmidt & Hunter (1977)
- Precursors:
  - Pearson (1904): correlations
  - Fisher (1932): *P* values
  - Yates & Cochran (1938, …): "Ag" experiments
- Medical research (1980s-): heart disease, cancer, etc. – ubiquitous since the 1990s
  - *"It is obvious that the new scientific discipline of meta-analysis is here to stay"* -- Chalmers & Lau (1993)

Social sciences: psychology, education, Employment testing, personnel evaluation, etc,

Meta-Analysis

- "The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" -- Glass (1976)
- "Averaging results across studies" -- Hunter & Schmidt (2004)
- "…the combination of results from multiple independent studies" -- Sutton & Higgins (2008)
- "[combination of the] results of previous research in order to arrive at summary conclusions to resolve uncertainty about the underlying medical question"-- Mittlbock & Heinzl (2006)
- Basic concept:
  - ***"…a single study will not resolve a major issue. Indeed, a small sample study will not even resolve a minor issue. Thus, the foundation of science is the culmination of knowledge from the results of many studies."***

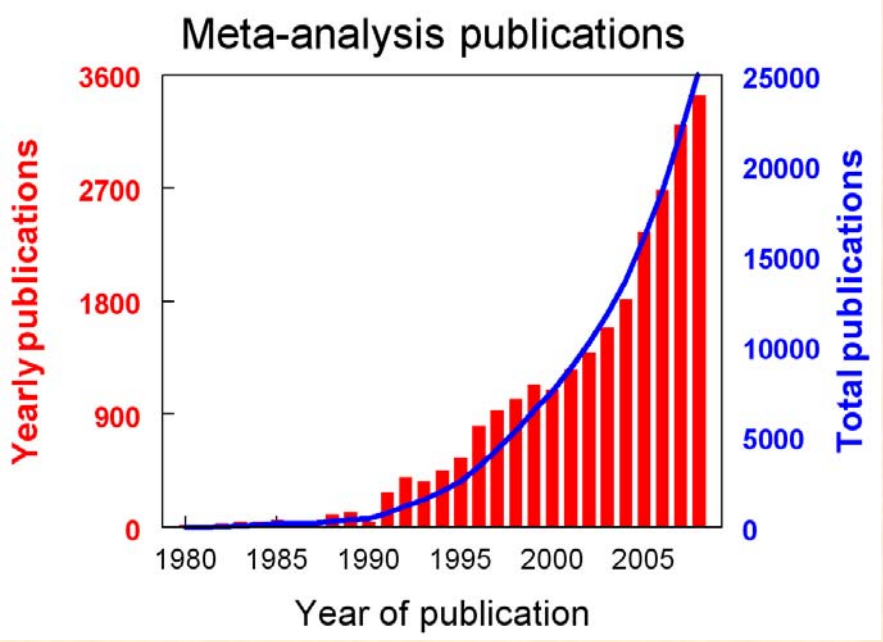
    -- Hunter & Schmidt (2004)

Meta-Analysis

- Controversy
  - "an exercise in mega-silliness" -- Eysenck (1978)
  - The problem of *"garbage-in, garbage-out"*: empirical data in certain studies may be untrustworthy
  - The problem of *"mixing apples and oranges"*: studies may differ too much from each other (methodology, treatments, measured responses, etc.), making synthesis problematic
  - Publication bias: only the 'good' results get published

•Nevertheless, **meta-analysis has become the standard for evidence synthesis in many disciplines**

–The above concerns can be (mostly) nullified with a formal selection of studies for analysis, based on clearly defined criteria



An illustration: an individual study

An investigation of the effect of treatment T, on severity of crop disease. Example:

- 2+ treatments or factor levels (T, C [=control], …)
- 4 replications
- Response: *y* (disease severity)
- Estimated **Effect Size** of interest (estimated parameter, or combination of estimated parameters, from an individual study):
  - Difference in mean disease for T and C

$$D = \hat{\mu}_C - \hat{\mu}_T = \bar{y}_C - \bar{y}_T$$

  - Or, % control, *C%* (relative reduction in disease compared to the control)

$$C\% = 100(\hat{\mu}_C - \hat{\mu}_T)/\hat{\mu}_C = 100(1 - \hat{\mu}_T/\hat{\mu}_C)$$

  - Or, transformation of the above for statistical reasons (e.g., log-response ratio): $L = \ln(\hat{\mu}_T/\hat{\mu}_C)$

L is especially useful when the mean in the control could be small or large -- e.g., D=3 is large when the control mean is 5 (**C% = 100·3/5 = 60%**), but small when the control mean is 50 (**C% = 100·3/50 = 6%**)

An illustration, *continued*

- Use **z** as a generic symbol for the estimated effect size (***D***, ***C%***, ***L***, …)
  - ***z*** is an estimate of a parameter ζ (*true* effect size)
- Record the estimated **effect size** (***z***) of interest (e.g., difference of two treatment means), and also the variance of ***z*** (label this s^2; known as the *sampling [within-study] variance*) for the study
- Note: When the effect size is the difference of means (i.e., when ***z = D***), then s^2 is the square of the *standard error of the difference* of means ($s^2 = SED^2$)
  - $s^2 = SED^2 = 2 \cdot V/n = (LSD/t^*)^2$
    - » Where ***V*** is the residual variance (mean square error), and ***n*** is the number of replicates (blocks)
    - » ***t**** is the critical value for a Student *t* distribution (~2 for a 5% significance level, with large residual *df*); approximate by standard normal if *df* is large enough
- Examples:
  - LSD = 4, t^*=2 ➔ $s^2 = (4/2)^2 = 4$
  - V = 3, n=4 ➔ $s^2 = 2 \cdot 3/4 = 1.5$

An illustration, *continued*

- When the effect size is ***L (z = L)***, then s^2 (V_L) is more complicated:

$$s^2 = V_L = \frac{V}{n}\left(\frac{1}{\bar{y}_C^2} + \frac{1}{\bar{y}_T^2}\right)$$

- Much more to say on different effect sizes (more on this later)
- A single study has the pair of required statistics, (z, s^2)

- Now suppose there are several studies, with the same treatments. Label the studies with index ***i***. If there are ***K*** studies (e.g., ***K*=10**), then ***i* =1, …, 10**

- Estimated effect size for study ***i*** is z_i, with variance s_i^2
- The pair (z_i, s_i^2) becomes a "*data point*" for a meta-analysis, and the unknown true effect size (a parameter) for study *i* is ζ_i

Fusarium head blight of wheat

- As part of the U.S. Wheat and Barley Scab Initiative, “Uniform Fungicide Trials” have been conducted for 10+ years in several states
  - Methodology has been standardized, so that all studies are conducted in a very similar manner
  - Usually, 4-7 treatments (including the control)
- These data have been used for several meta-analyses. See: *Phytopathology* 97: 211-220; *Phytopathology* 98: 999-1011.
  - Here, we use results for the effect of Folicur (tebuconazole) on DON (ppm in grain)
    - There were *K*= 101 studies in this analysis
    - *y:* DON (ppm)
    - Two treatments used
      - T: Folicur (applied at Feekes 10.5.1)
      - C: Check

Primary interest:
Percent Control (***C%***)

Effect size (z_i):
Log-response ratio

$$z_i = L_i = \ln(\hat{\mu}_{T,i} / \hat{\mu}_{C,i})$$

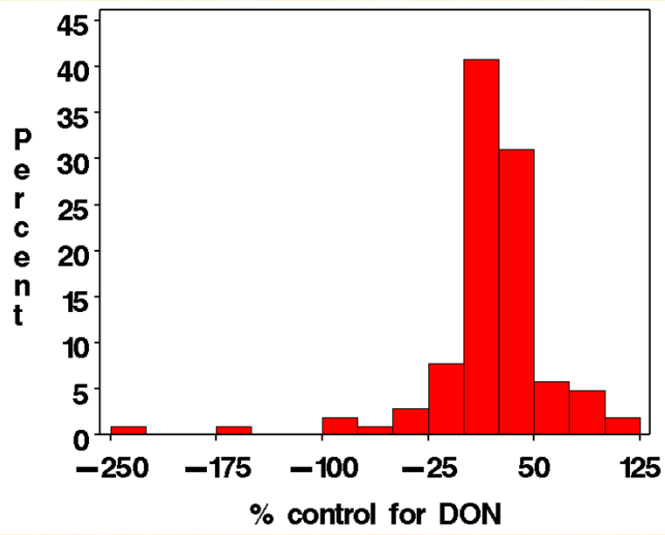
The meta-analytical data set:

i z_i s_i^2

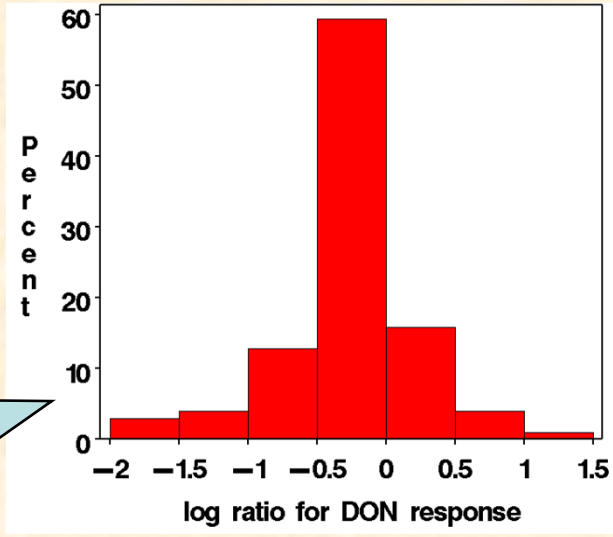
Study	$\hat{\mu}_C$	$\hat{\mu}_T$	C%	$\ln(\hat{\mu}_T / \hat{\mu}_C)$	V_L
1	10.3	4.8	53.4	-0.764	0.029
2	8.0	4.4	45.0	-0.598	0.017
3	3.9	3.8	2.8	-0.029	0.011
4	5.3	2.7	49.1	-0.674	0.036
5	8.2	7.1	13.4	-0.144	0.019
...					

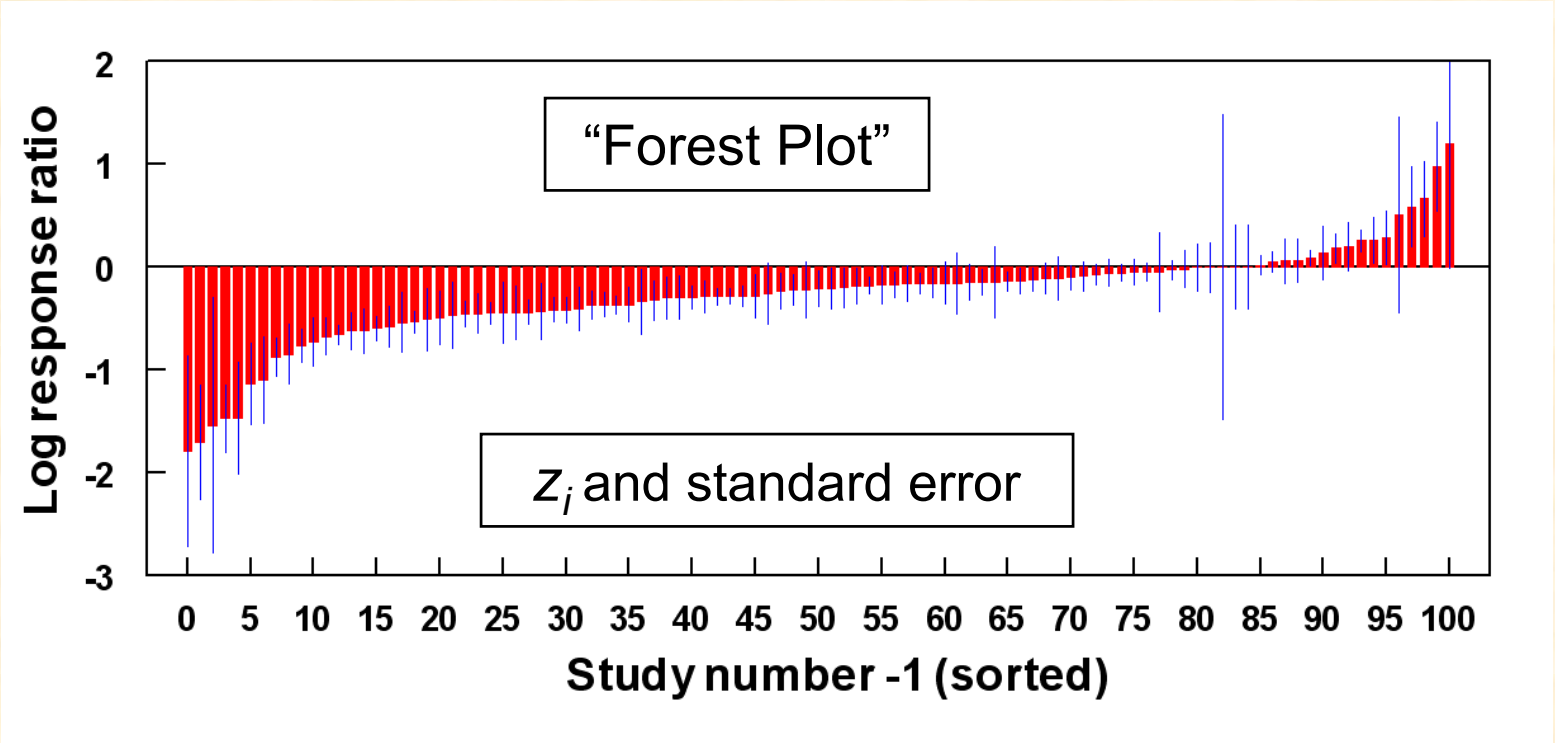
K=101 studies:
Each study becomes an "observation"

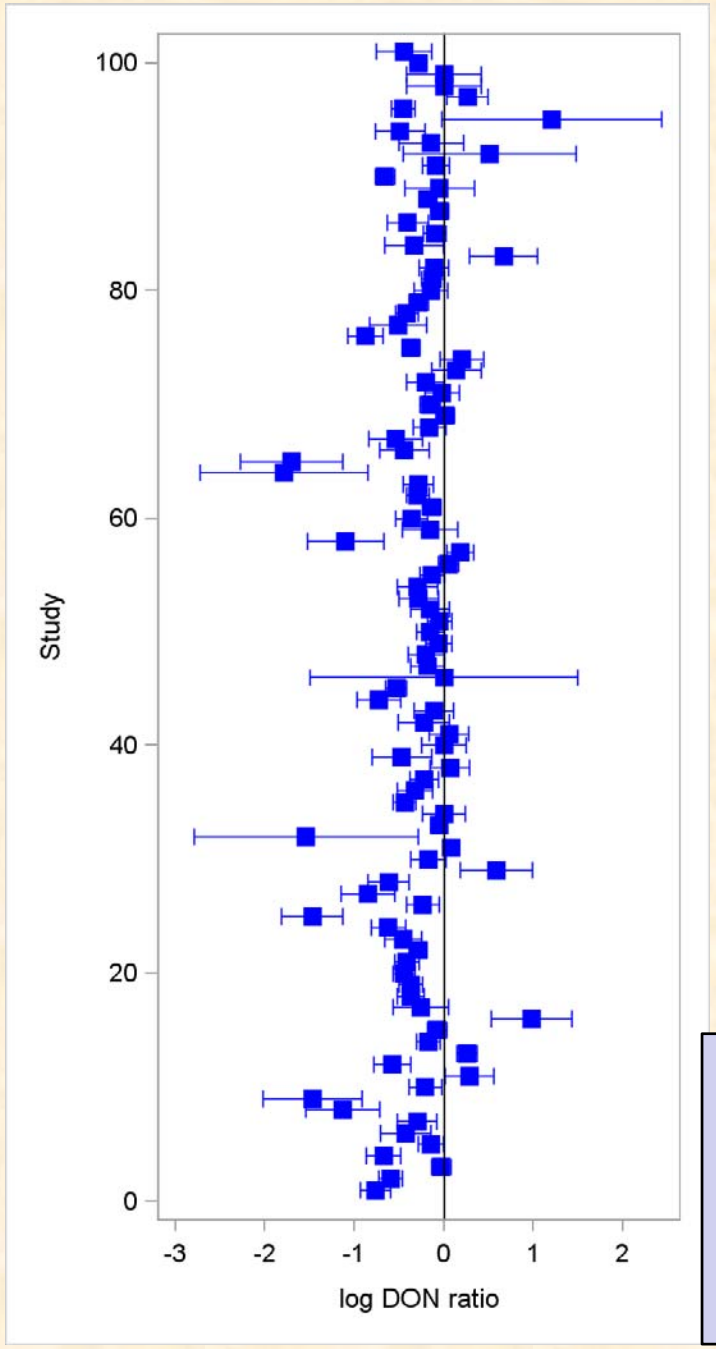
Graphs



Histogram of z_i

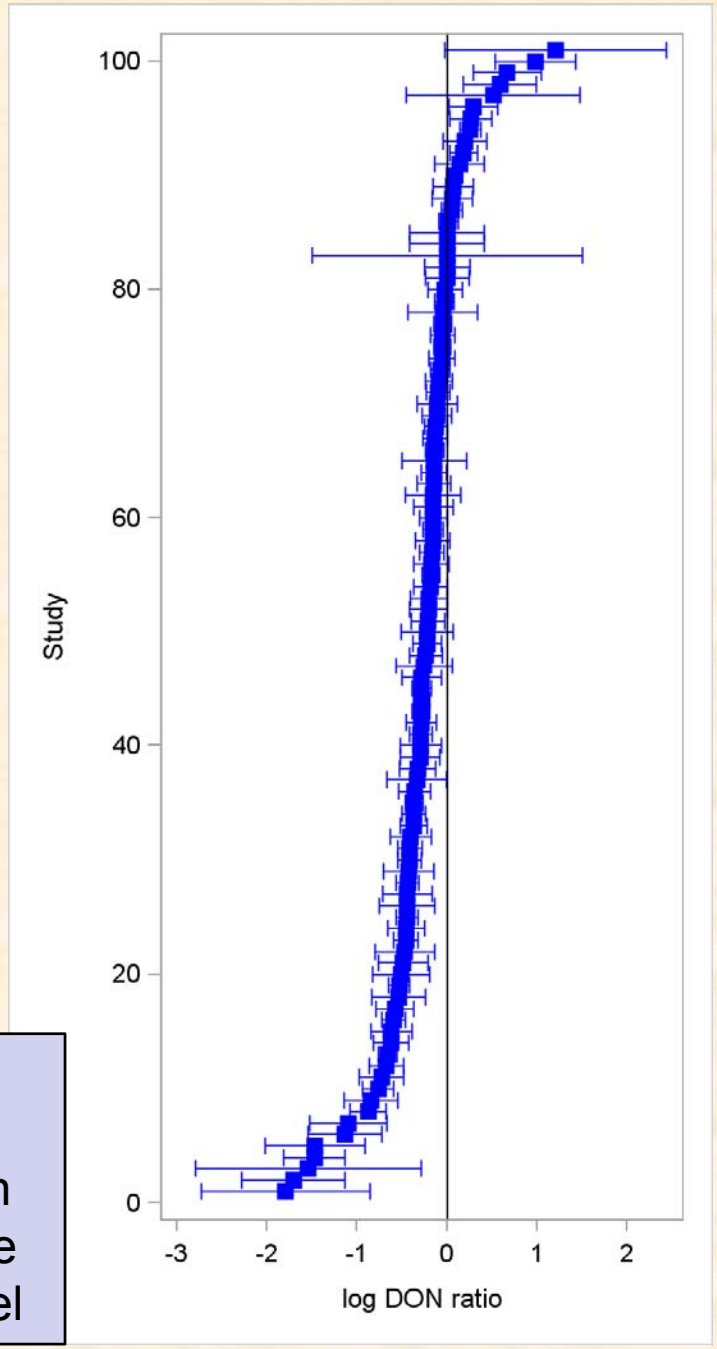






Forest Plot

Once you have the "data", then you analyze with a model



Model Fitting

Effect size for study *i* (the result from study *i* becomes a data point in the meta-analysis)

$$z_i = \zeta + u_i + \varepsilon_i$$

Expected effect size, overall

Among-study variability term. Random effect of study *i* on the effect size.

Within-study variability term; residual or *"sampling variation"*. Assume known.

Distributional assumptions:

σ^2 *:* among-study variance

$$u_i \sim N(0, \sigma^2)$$

$$\varepsilon_i \sim N(0, s_i^2)$$

s_i^2*:* sampling variance (separate for each study; assume known). Assume u and ε are independent

$$z_i \sim N(\zeta, \sigma^2 + s_i^2)$$

One estimates ζ and σ^2

Meta-analysis models

- **Random-effects model** (explicit consideration of among-study variability)
  - $\sigma^2 \geq 0$

$$z_i = \zeta + u_i + \varepsilon_i$$

$$u_i \sim N(0, \sigma^2)$$

$$\varepsilon_i \sim N(0, s_i^2)$$

- **Fixed-effects model** (assume that there is no random variation in the true effect size) – the "old-fashioned" approach
  - i.e., u_i = 0 for all studies, which means that $\sigma^2 = 0$
  - In this case, think of ζ as a *common* (not expected) effect

$$z_i = \zeta + \varepsilon_i$$

$$\varepsilon_i \sim N(0, s_i^2)$$

Meta-Analysis

- Parameter estimation, for ζ and σ^2
  - **Method of moments** (the classical meta-analytical approach, but *may* not be the most general or accurate approach)
  - **Maximum likelihood** (and restricted maximum likelihood): **ML** and **REML**
    - Iterative and more computer-intensive, but is usually superior
  - **Bayesian approach** - a useful alternative
- *In general, an investigator uses <u>one</u> estimation method (but we demonstrate several here, for teaching purposes)*
- Tests: Student-*t*, Standard Normal, *F*, Chi-square, etc.
- At its core, meta-analysis is a method of obtaining weighted averages of effect sizes

$$\hat{\zeta} = \frac{\sum w_i z_i}{\sum w_i}, \quad w_i = \frac{1}{\sigma^2 + s_i^2} \qquad \mathrm{SE}(\hat{\zeta}) = \left(\sum w_i\right)^{-1/2}$$

Random-effect meta-analysis

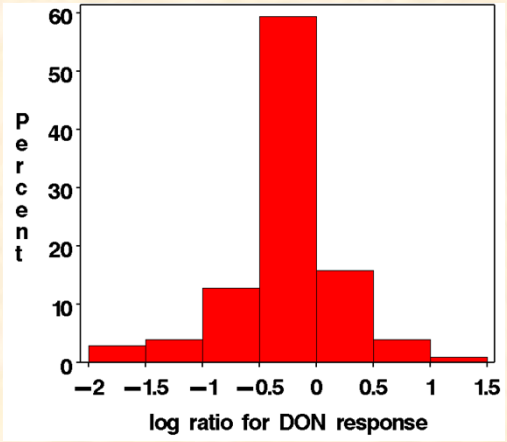
$\hat{\zeta}$	SE($\hat{\zeta}$)	95% CI for $\hat{\zeta}$	$t = \hat{\zeta} / \mathrm{SE}(\hat{\zeta})$	*P* value	Control % (*C*%)	95% CI for *C*%
–0.24	**0.028**	**–0.30 ↔ –0.19**	**–8.85**	**<0.001**	**21.6%**	**17.2% ↔ 25.8%**

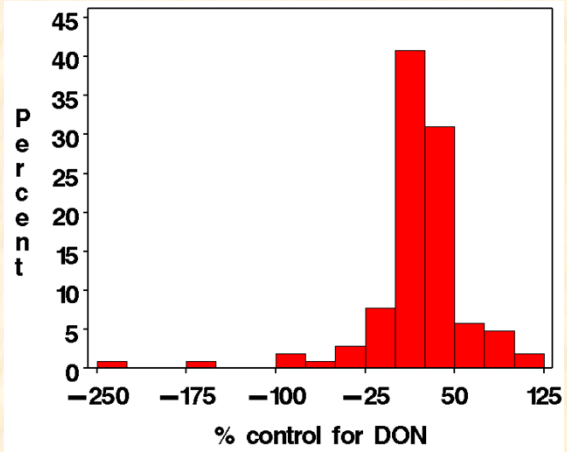
$$C\% = 100 \cdot (1 - \exp(\hat{\zeta}))$$

H_0: $\zeta = 0$ (i.e., log response ratio = 0, Same as response ratio = 1, or percent control = 0%)

H_a: $\zeta \neq 0$

Results based on ML estimation





Estimation methods

ML (or **REML**) methods are very general and useful, especially for those familiar with likelihood-based model fitting (recommended)

Several methods can be used. **Moment Method** is most commonly used in the many specialized computer programs

Method	$\hat{\zeta}$ (SE)	Confidence Interval (95%)
ML	**-0.244 (.0276)**	**-0.299 ↔ -0.189**
REML	**-0.244 (.0278)**	**-0.299 ↔ -0.189**
Moment	**-0.245 (.0285)**	**-0.301 ↔ -0.188**
Fixed	***-0.223 (.0163)***	***-0.255 ↔ -0.192***
Bayesian	**-0.242 (.0281)**	**-0.298 ↔ -0.184**

Fixed-effect estimates (**assumed $\sigma^2=0$**), also very common, should not be used, in general. *False sense of precision*. Shown for demonstration.

Bayesian approaches are becoming more and more common. Here, *noninformative priors* were used. Useful for dealing with uncertainty in variance estimate.

Heterogeneity and risk probabilities

- The among-study variance (σ^2) is of value for:
  - Properly estimating the expected effect size and its standard error
  - Assessing the magnitude of effect-size heterogeneity
    - If $\sigma^2 = 0$, then:
      - One could use fixed-effect analysis with $\sigma^2 = 0$, but there is really no reason to do so (random-effect analysis is just as easy [now], which automatically takes care of the among-study variability [if present])
  - Specialized post-model fitting analyses of value:
    - **Prediction interval** for effect size (interval in which future (other) *individual* effect sizes will fall, with specified probability [say, 95%])
    - The probability that the effect size in a randomly selected future study will be *less than* (or will be *greater than*) any constant of interest (ϑ)

Heterogeneity and risk probabilities

- One can test for significance of σ^2 in several ways, including with a likelihood ratio test (but very computationally intensive)
- Often, one simply wants to know the *impact* of heterogeneity
  - Higgins & Thompson developed three (interrelated) indices for impact, to ascertain whether among-study heterogeneity is having a substantial *effect* on the results (H^2, I^2, R^2)
    - I^2: Percentage of total variability that is due to among-study heterogeneity (defined directly in terms of moment estimates)
    - "R^2": Based on *ratio* of the width of the confidence interval for estimated effect size (ζ) for a random effect and fixed effect analysis
      - Larger then 1.5 (or 2) means that among-study variation is having a substantial effect on all the results

$$"R^2" = \left(\frac{SE(\hat{\zeta})_{random}}{SE(\hat{\zeta})_{fixed}} \right)^2$$

Meta-analysis: Among-study variability

ML estimation for DON and Folicur;
Profile likelihood CI method

Likelihood-ratio statistic (LRS) and Chi-square test (one can also use a standard normal Z test).
If interval does not include 0, then variance is significantly greater than 0.

$\hat{\sigma}^2$	95% CI for $\hat{\sigma}^2$	*P* value	"R^2"
0.036	**0.020 ↔ 0.063**	**<0.001**	**2.9**

Relative impact of heterogeneity (>1.5 is high)

H_0: $\sigma^2 = 0$ (i.e., no heterogeneity in the [true] effect size among studies: there is a common effect size across studies)

H_a: $\sigma^2 > 0$ (i.e., heterogeneity in the [true] effect size)

Higgins and Thompson metric:

$$"R^2" = \left(\frac{SE(\hat{\zeta})_{random}}{SE(\hat{\zeta})_{fixed}} \right)^2 = \left(\frac{.0276}{.0163} \right)^2 = 2.9$$

Confidence Interval (for expected value):

$$\hat{\zeta} \pm t^{*} \mathrm{SE}(\hat{\zeta})$$

$\hat{\zeta}$	SE($\hat{\zeta}$)	**95% CI**	$t = \hat{\zeta} / \mathrm{SE}(\hat{\zeta})$	*P* value	Control % (*C%*)	**95% CI for *C%***
–0.24	**0.028**	**–0.30 ↔ –0.19**	**–8.85**	**<0.001**	**21.6%**	**17.2% ↔ 25.8%**

Prediction Interval (for individual effect sizes):

$$\hat{\zeta} \pm t^{*} \cdot \left(\mathrm{SE}^{2}(\hat{\zeta}) + \hat{\sigma}^{2} \right)^{0.5}$$

$\hat{\zeta}$	SE($\hat{\zeta}$)	**95% Pred. Int.**	$t = \hat{\zeta} / \mathrm{SE}(\hat{\zeta})$	*P* value	Control % (*C%*)	**95% Pred. Int. for *C%***
–0.24	**0.028**	**–0.62 ↔ 0.13**	**–8.85**	**<0.001**	**21.6%**	**-14.3% ↔46.3%**

$$SE(\hat{\zeta}) = 0.0276$$

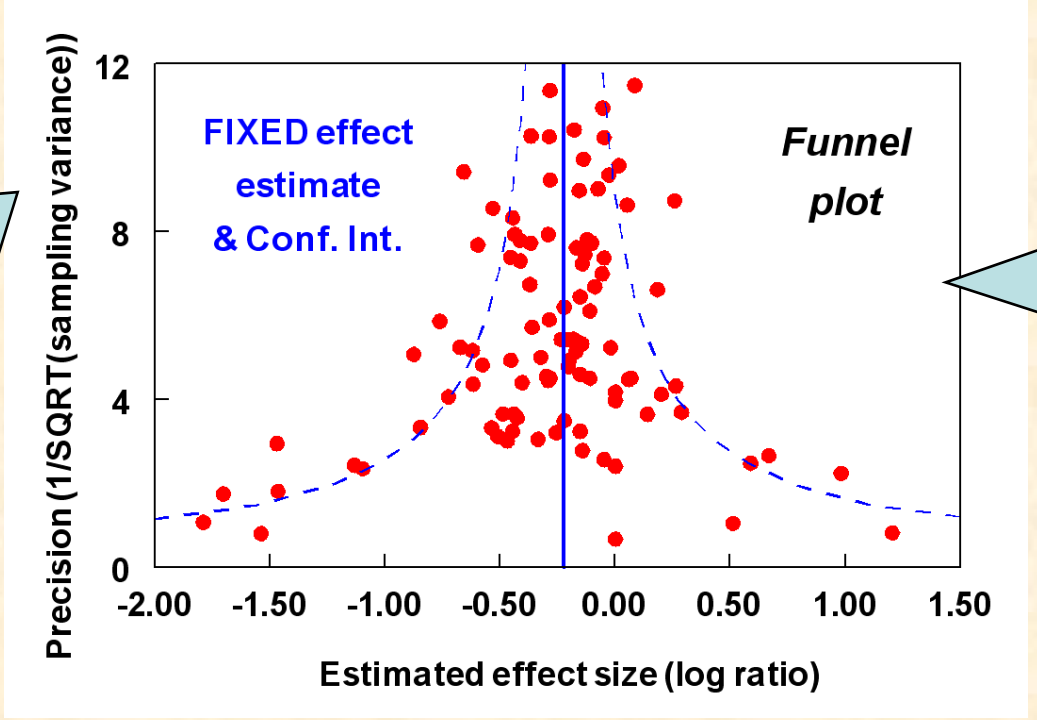
$$\hat{\sigma}^{2} = 0.0365, \ \hat{\sigma} = 0.191$$

$$C\% = 100 \cdot (1 - \exp(\hat{\zeta}))$$

Other graphical views of effect sizes

- With a large *K*, some *specialized* graphs *may* be useful
  - These graphs *can* simultaneously be used to determine if a fixed-effects (common-effects) analysis is warranted, and if there is bias due to missing studies (*publication bias*, discussed later)
  - **Funnel plot:** Graph of "precision" (= 1/[sampling variance]$^{1/2}$) vs. estimated effect size for all the studies);
    - $1/[s_i^2]^{1/2}$ vs. z_i

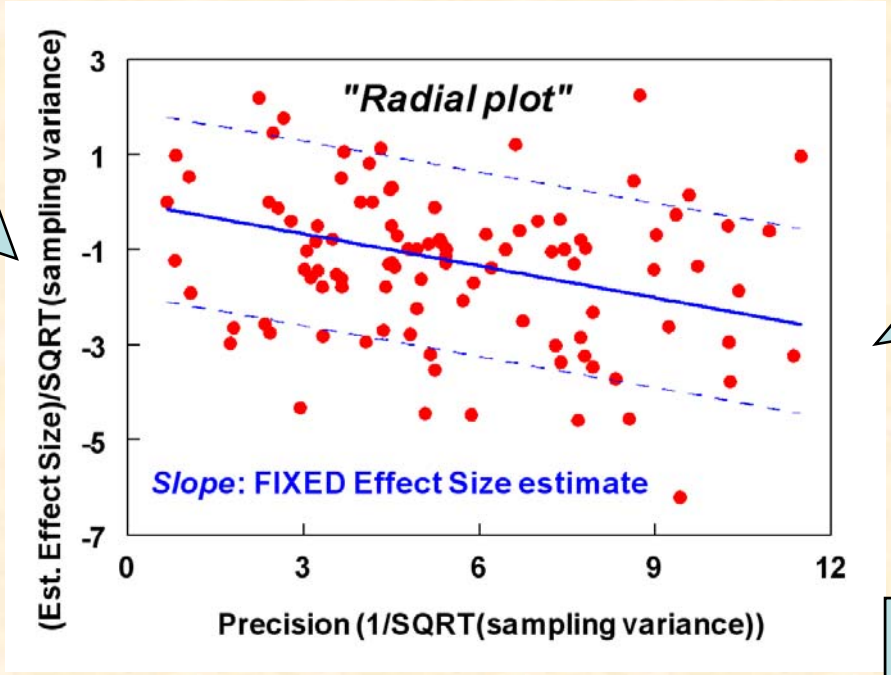
If among-study variance is 0 (justifying fixed-effects), almost all points should be inside the dashed lines). Evidence here is for random effects



If not upside down funnel, and not symmetrical, then selective reporting of results may be occurring. No bias here.

- With a large *K*, some *specialized* graphs *may* be useful
  - In addition to **funnel plot**, a so-called "Radial plot" or "Galbraith plot") may be useful
  - **Radial plot:** Graph of *standardized* estimated effect size versus "precision" (1/[sampling variance]$^{1/2}$)
    - $z_i/[s_i^2]^{1/2}$ vs. $1/[s_i^2]^{1/2}$

If among-study variance is 0 (justifying fixed-effects), almost all points should be inside the dashed lines). Evidence here is for random effects



If no bias, there should be a random scatter around the line (no gaps at certain precisions or at certain effect sizes)

These graphs are **guides** only, and may not be useful with much smaller *K*

Other effect sizes: Example analysis

- Before considering applications of meta-analysis and more advanced topics, an example analysis will be carried out in the workshop
- So far, we have used the log-response ratio (L_i) as the effect size (z_i), and use the estimated expected value to determine percent control (through a back transformation)
- There are numerous possible effect sizes, depending on the objectives of the investigator and the nature of the studies being assembled and analyzed in the meta-analysis
- Many meta-analytical textbooks give details on effect sizes and their sampling variances for individual studies

Effect Sizes: Treatment effects

- **Continuous data**
  - Difference in means (D_i)
  - Log ratio (L_i), or percent control…
    - Valuable when *relative* changes matter
    - *May* be useful when the response variable is not (quite) the same for all studies (different scales)
  - Standardized mean difference (d_i)
    - *Very* common in social sciences
    - Advocated when the response variable differs among studies (different scales)
      - (Often overlooked: assumes a linear relation among scales)
- **Discrete data (*not covered*)**
  - Difference of proportions (risk difference), relative risk (ratio of proportions), odds ratio, and their transformations (*log-odds*)

$$D_i = \hat{\mu}_{C,i} - \hat{\mu}_{T,i}$$

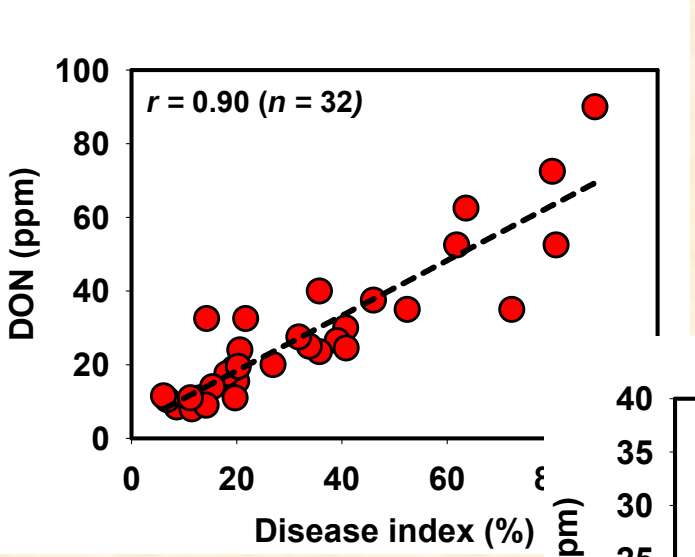
$$L_i = \ln(\hat{\mu}_{T,i} / \hat{\mu}_{C,i})$$

$$d_i = \frac{(\hat{\mu}_{C,i} - \hat{\mu}_{T,i})}{S_{pooled_i}}$$

$S_{pooled,i}$:
Within-study standard deviation

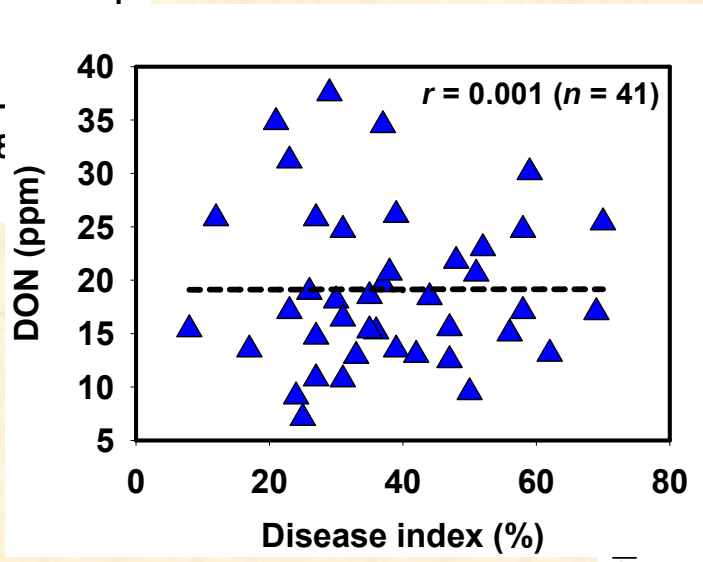
Sampling variance formula depends on the effect size and response variable

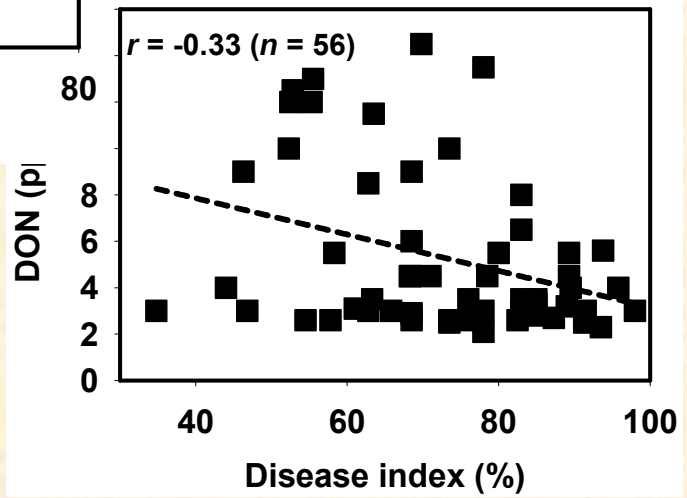
Effect sizes for relationships or associations



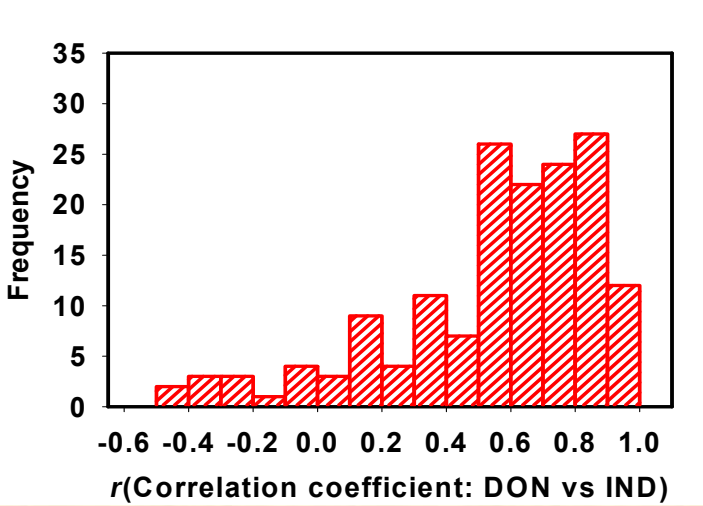
Correlation coefficient (r_i) or the Fisher transformation (Z_{ri}).

Sampling variance of Z_{ri} is $s_i^2 = 1/(n_i\text{-}3)$

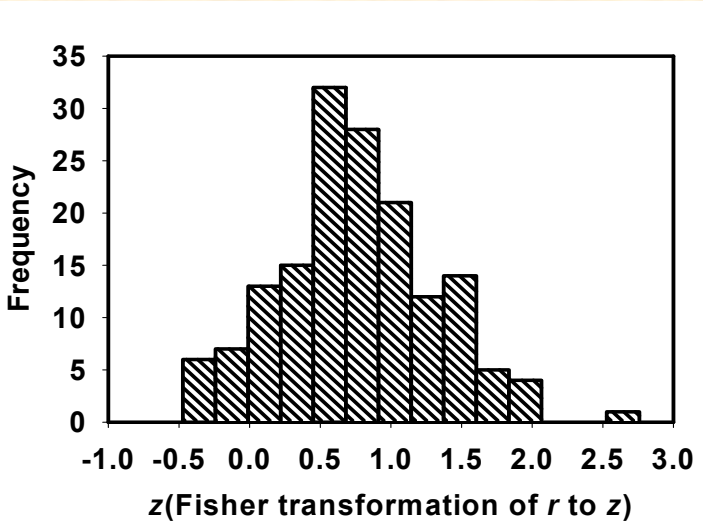


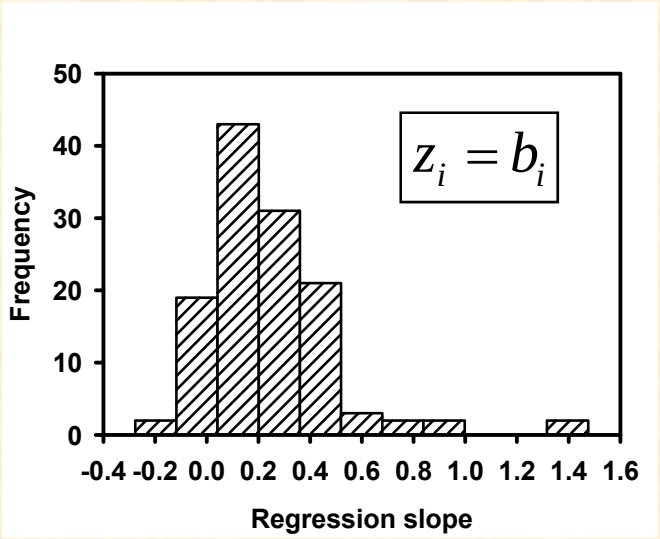


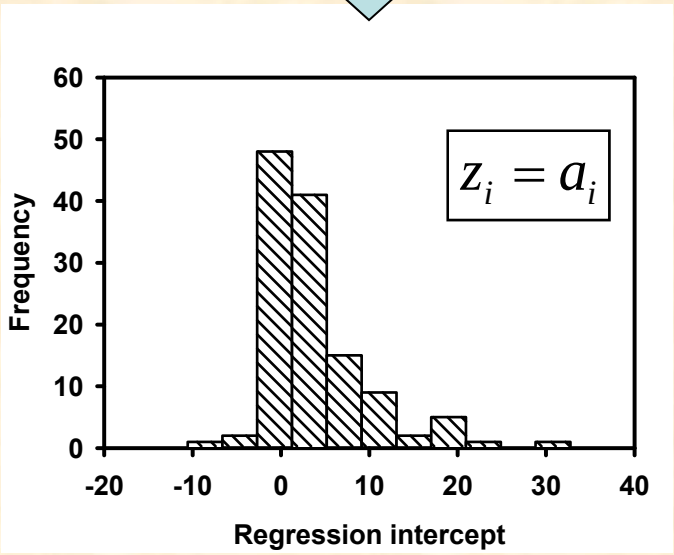
Slope (b_i) and/or intercept (a_i) of model fitted to the data for each study (i = 1, …, K). Sampling variance: square of estimated standard error of slope or intercept



$$z_i = Z_{r_i} = \frac{1}{2}\log\left(\frac{1+r_i}{1-r_i}\right)$$







Example:

Yield Loss in Sweet Corn Caused by *Puccinia sorghi*: A Meta-Analysis

Denis A. Shah and **Helene R. Dillard,** Department of Plant Pathology, New York State Agricultural Experiment Station, 630 W. North St., Geneva 14456

- Effect size: slope of the regression model for crop loss as a function of disease severity (i.e., $z_i = b_i$, where b_i is the slope for the i-th study (i = 1, …, 20)

http://oardc.osu.edu/APS-statsworkshop/downloads/downloads.htm

Heterogeneity and risk probabilities

- The *mean* effect size and its standard error (or confidence interval) are of interest for determining the *expected* outcome in the *long run* (over many studies [or over many fields]), but these statistics cannot be used directly to determine how likely a *given* effect size will be in a *single* (future) study or in a field treated in the same manner, or in a collection of *individual* fields
  - *Prediction intervals* are useful for dealing with single studies
- More directly, one can estimate the probability that the effect size in a randomly selected future study will be *less than* (or will be *greater than*) any constant of interest (ϑ)
  - For instance, with DON control for Fusarium head blight, a grower might be most interested in knowing the probability (p_0) that *C*% > 0% (i.e., *L* < 0) or maybe that *C*% > 50% (i.e., *L* < -0.69)

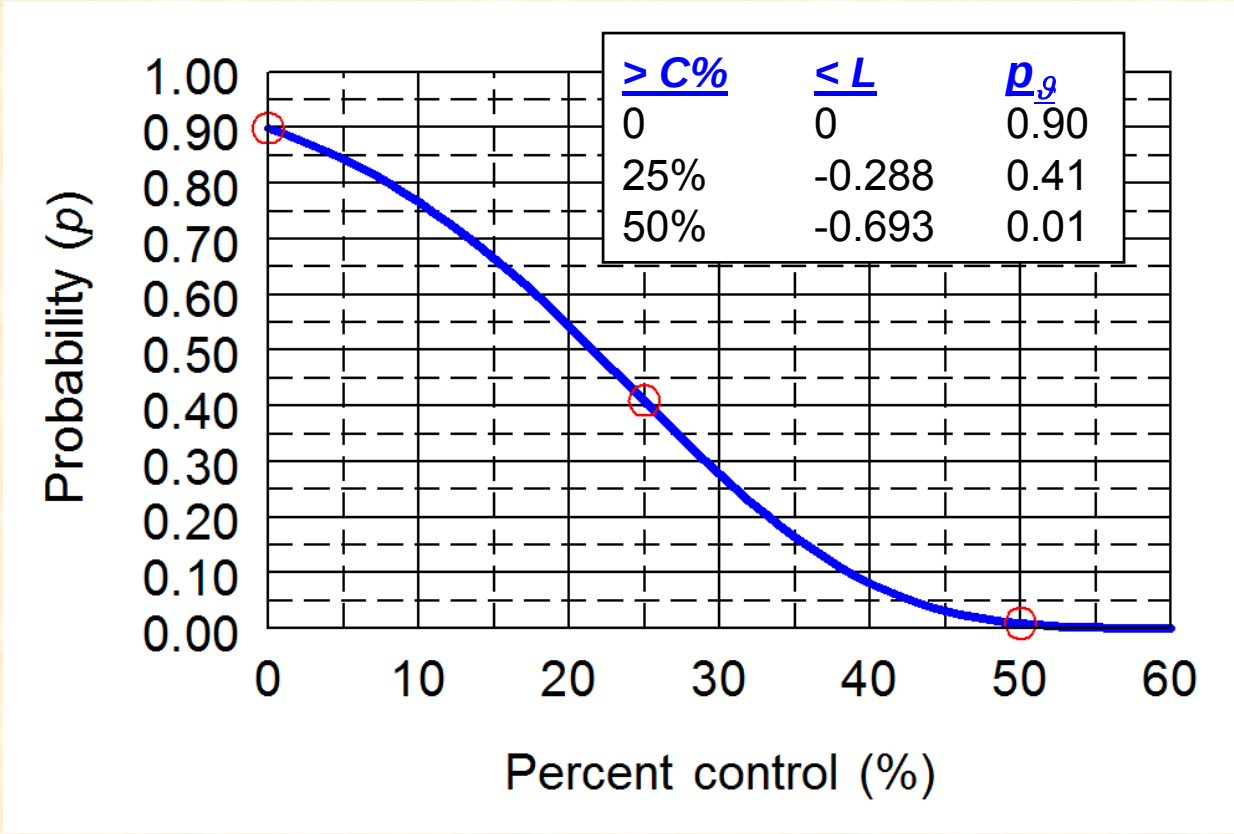
$$p_{\vartheta} = \Phi((\vartheta - \hat{\zeta}) / \hat{\sigma})$$

Φ(•) is the cumulative normal distribution, use to obtain probability that effect size is less than ϑ

$$p_{\vartheta} = \Phi((\vartheta - \hat{\zeta}) / \hat{\sigma})$$

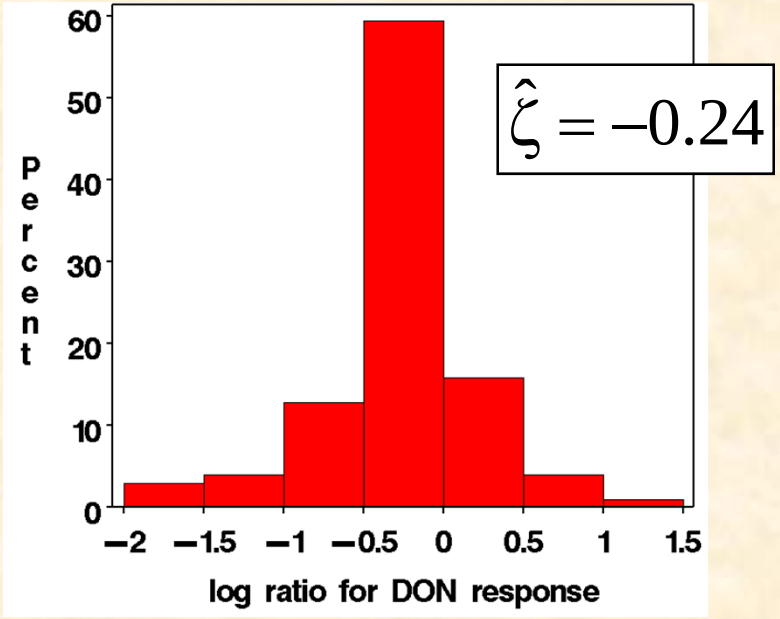
Risk probability for DON control:
Determined for log ratio and than converted back to percent control

$$\hat{\zeta} = -0.244, \hat{C}\% = 21.6\%, \hat{\sigma} = \sqrt{0.036} = 0.19$$



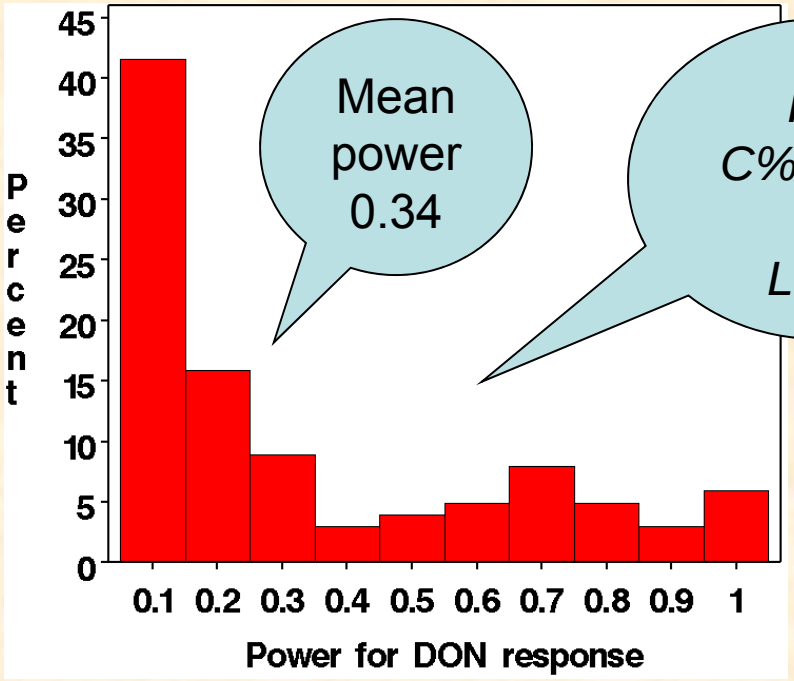
Statistical power: a major benefit of meta-analysis

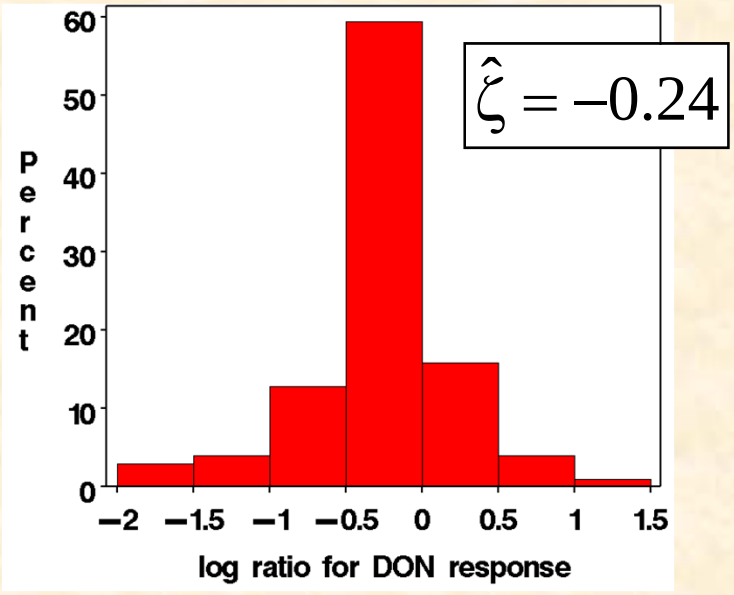
- Assume H_a ($\zeta \neq 0$) is true (treatment is truly effective)
- Consider statistical **power**
  - **Prob. of rejecting H_0 when H_0 is false (i.e., probability of making correct decision here)**
  - To justify the use of meta-analysis, we can first estimate the power for *each* study (assuming H_a is true for *every* study ($\zeta_i \neq 0$)
    - Hypothetical and unrealistic here, but useful for demonstration purposes



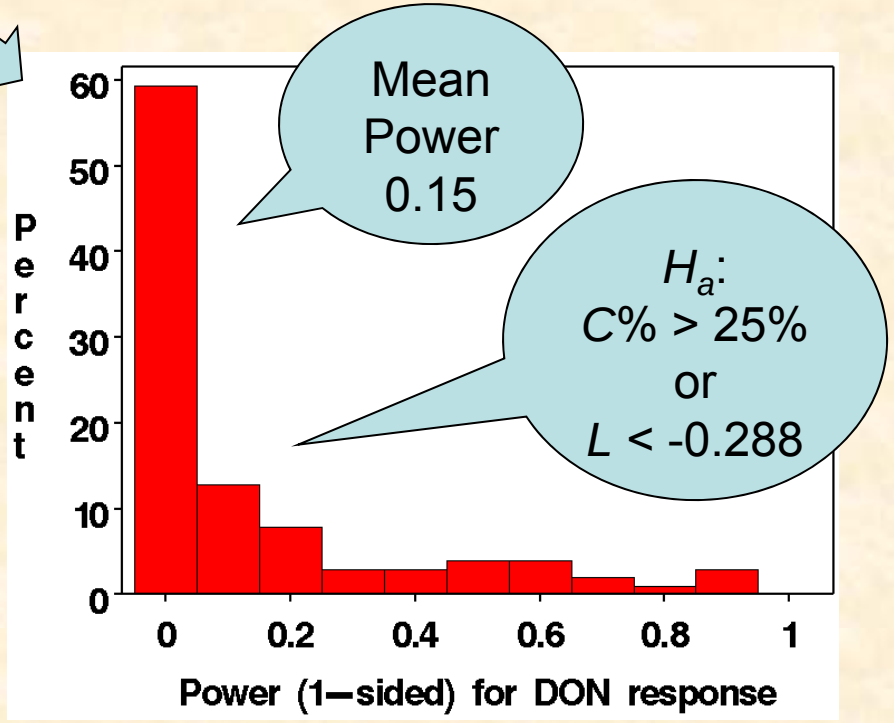
Calculations based on assumed normality for z_i (L_i here), and (shifted) Student t distribution for estimated ζ_i

One can consider two-sided ("not equal") or one-sided ("less than 0", "greater than ϑ percent") alternatives





By any definition, power is low for the *individual* studies (in terms of treatment effect on DON)

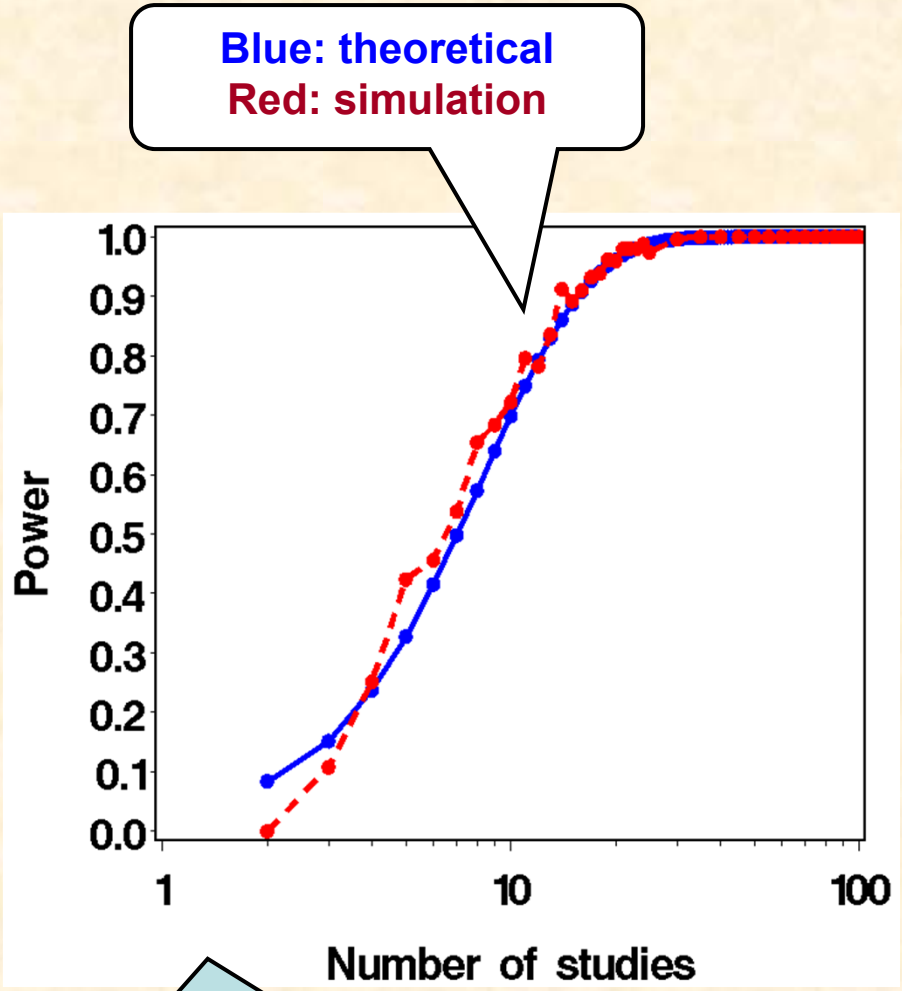


We can estimate the power for the meta-analysis of the 101 studies (we do not need to assume that H_a is true for every study, just that $\zeta \neq 0$)

***Power* > 0.999**

Power in meta-analysis

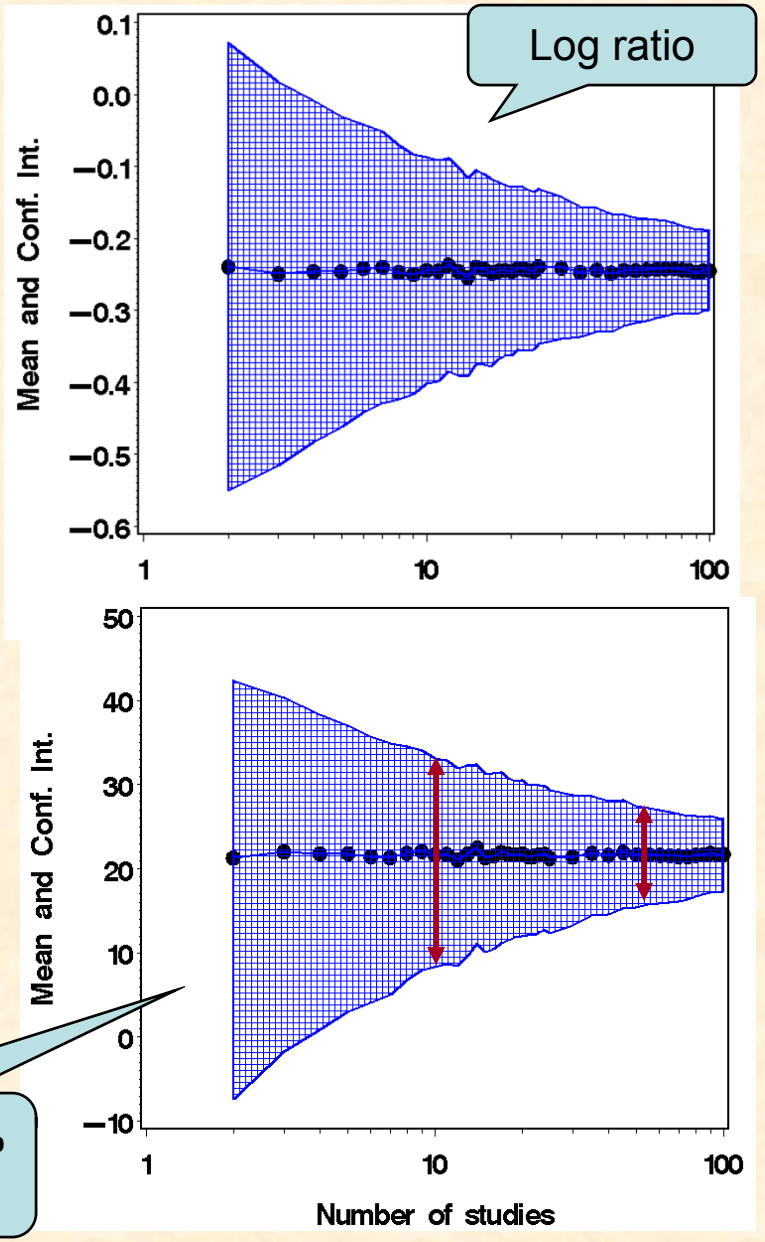
- One can determine power for any number ($2 \rightarrow K$) of randomly-selected studies:
  - **Theoretically**, assuming a normal distribution for z_i (with known variance terms)
  - With **simulation**, which is a better approach to deal with unequal sampling variances, but computationally demanding
- Reminder:
  - Mean individual power (assuming that Folicur always has an effect) was 0.34
  - Meta-analysis ***Power* > 0.999** with 101 studies (no assumption about individual studies)
    - Even a *Power* of 0.8 could be reached with < 20 studies



See Littell et al. (2006) for details

Precision in meta-analysis

- Often, the null hypothesis will not be true, and this can be shown with a reasonable (or small) number of studies.
- It may be more informative to consider the estimated mean and its 95% confidence interval for different number of (randomly-selected) studies. Easily obtained using simulation
- Choose a study number that gives a desired width of the confidence interval



The fallacy of counting *P* values (instead of doing a meta-analysis)

- *Suppose K* independent studies were conducted, and that there is *truly* a significant treatment effect (say, $\zeta_i < 0$) in every study (i.e., H_a is always true) **-- returning to our hypothetical scenario**
- But also *suppose* that individual-study power is 0.40 (not a very high chance of detecting the true effect)
- A typical "qualitative" ("narrative") summary is to count the number of significant results (studies where $P \leq 0.05$): vote counting
  - **Conclude that the treatment is effective if *at least half* the studies are significant**
- With a large number of studies (say, $K = 150$), 40% will have significant results (on average) with this power
  - **Thus, one would falsely conclude here that treatment was not effective, even though it was (truly) effective in every study.**

Fallacy of counting *P* values

- As the number of studies *increases*, it becomes *less and less* likely to every find 50+% of the studies with significant results (when individual power < 0.5).
  - **In fact, there is a higher chance of finding 50+% of the studies with significant results if fewer studies are considered (a major violation of good statistical practice)**
- Demonstration:
  - Chance of at least half the studies being significant ($P \leq 0.05$) when H_a is always true and individual-study power is **0.40** (low, but higher than in example)

Studies	*Prob*
10	0.17
20	0.13
30	0.10
50	0.06
100	0.02

With a small number of studies, one actually has a better chance of finding half (or more) of the studies being significant

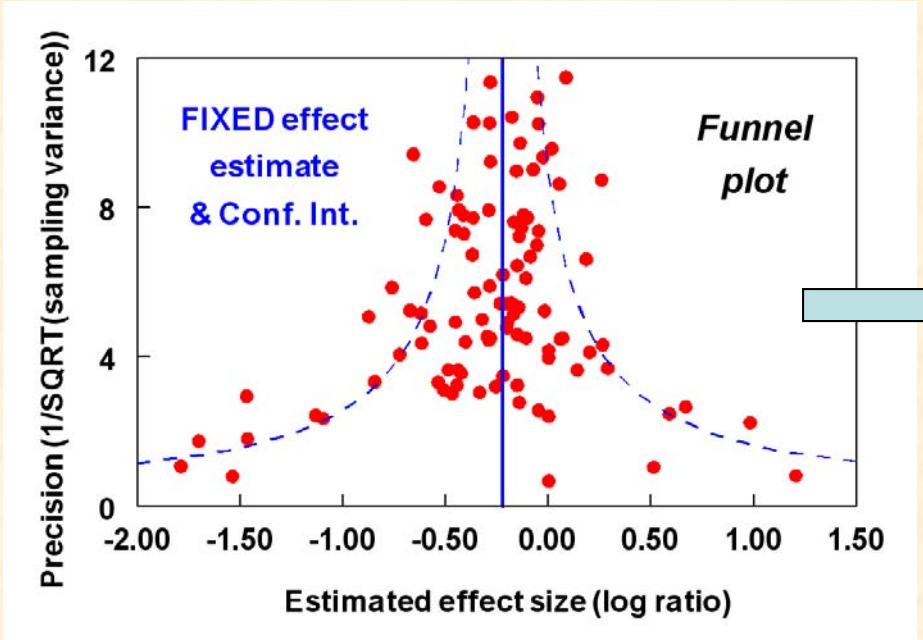
There are valid ways to combine *P* values to determine overall significance (going back to work by Fisher), but these are not discussed here.

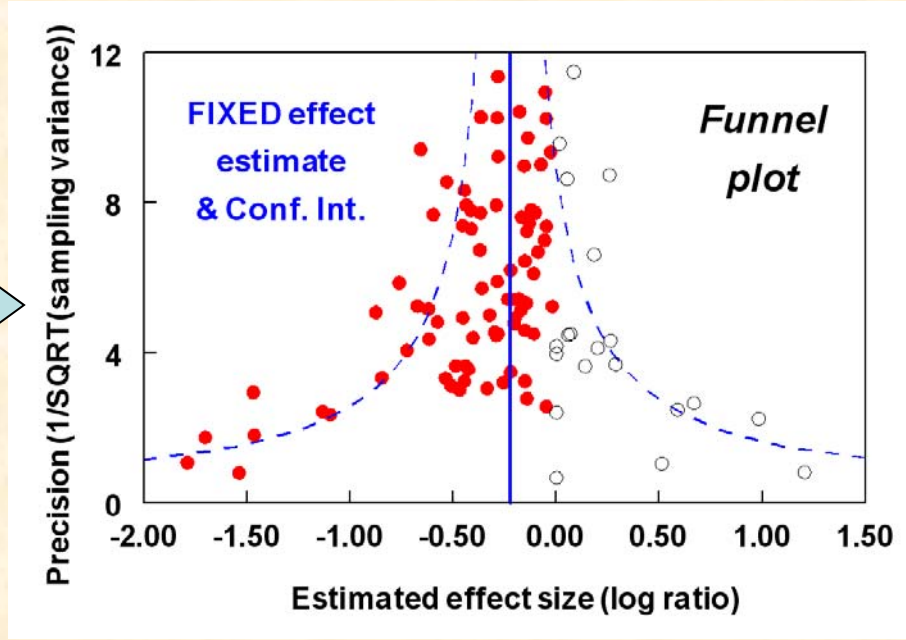
Publication Bias

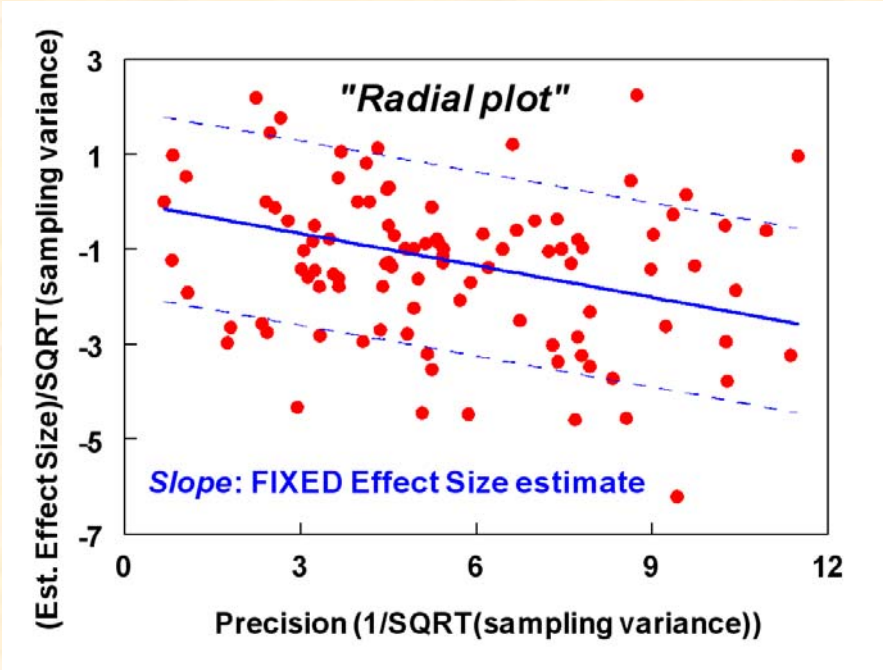
- Most meta-analyses make the tacit assumption that the studies under review are a random sample from a hypothetical population of possible studies (or the *effects* in each study comprise a random sample of possible effects)
  - Unlikely to be true, of course, in a technical sense.
  - It is likely that larger studies, or studies with significant results, will be published or made available for review
  - The "*nightmare*" of meta-analysis (van Houwelingen, 1997).
- If inclusion of a study in the dataset depends on the realized effect size, then the meta-analytical results will be biased:

$$E(\hat{\zeta}) \neq \zeta$$

- Not of concern, for the most part, with Fusarium head blight example. The national initiative encouraged the 'publication' of all studies in proceedings and reports







If no bias, there should be a random scatter around the line (no gaps at certain precisions or at certain effect sizes)

These graphs are **guides** only, and may not be useful with much smaller *K*

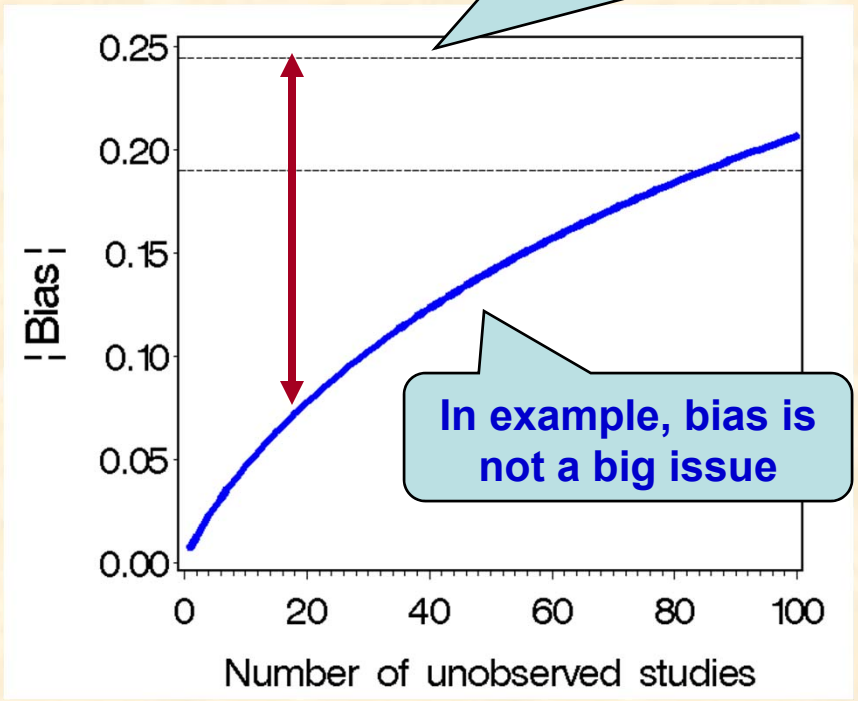
Publication Bias: Solutions

- Ignore the "selection bias" of studies (usual "solution")
  - The population being analyzed consists of the available (i.e., published, found, selected) studies, not an undefined larger population **-- Limits scope of inference**
- Use various analytical methods (including complex weighting of effect sizes), based on various assumptions regarding the study selection process
  - Conduct sensitivity analysis to see consequences of different selection choices, which can lead to a bias adjustment, or can show how many unused studies would invalidate the results (**file drawer problem**)
- However, it is impossible to determine the study-selection mechanism from the available studies
- A very interesting new alternative is to determine the **upper bound on the bias** for any number of unpublished studies
  - Copas & Jackson (2004) show that the worst-case bias (for any selection mechanism) is straight-forward to calculate

Publication Bias: an upper bound

- Copas & Jackson approach:
  - Fusarium head blight example
  - *K* = 101 studies
  - Effect size: log ratio
    - **mean = -0.24**
- Equations not shown

Compare **|Bias|** to **|mean|** from the published studies, or **mean-(2·SE)**



Calculate the absolute value of the (worst-possible) bias for different numbers of unobserved or unpublished studies

Example, if there are **20** unpublished studies, the total number of studies is 121 (not 101), and the mean effect size could be as large as -0.24**+0.077** (-0.163) or as small as -0.24**-0.077** (-0.317)

Study heterogeneity ($\sigma^2 > 0$), *continued*

- Causes include:
  - Differences in study conditions (experimental methods, data collection approaches, etc.)
  - Environment (broad sense)
- Study conditions/environment can be accounted for in the meta-analysis through the incorporation of **moderator variables** in the model
  - Moderator variable: *study-level characteristic (continuous or categorical) that can affect the magnitude of the effect size*
    - Examples: cultivar, temperature (degree days), baseline disease intensity, etc.
  - The approach is analogous to performing a mixed-model analysis of covariance
  - Accounting for significant moderator variables can lower the estimated among-study variance and possibly the standard error of the estimated effect sizes

Meta-Analysis

Effect size for study *i*

$$z_i = \zeta + u_i + \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$$

Within-study variability term; residual or *"sampling variation"*. Assume known.

Expected effect size, overall

Among-study variability term. Random effect of study *i* on the effect size.

Effect of moderator variable(s) for the *i*-th study.

$\mathbf{X}_i$: a row vector of *p* different continuous moderator variables (e.g., mean temperature), or "dummy variables" to indicate categories or class levels (e.g., wheat type)

$\boldsymbol{\beta}$: vector of effects of the moderator variables on the effect size (vector product is a scalar (single value) for each study).

$$u_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \sim N(0, s_i^2)$$

s_i^2: sampling variance (separate for each study; assume known)

σ^2: among-study variance

$$z_i \sim N(\zeta + \mathbf{X}_i\boldsymbol{\beta}, \sigma^2 + s_i^2)$$

Moderator Variable Example:
Wheat Type (Winter [W] or Spring [S])

Wheat type	$\hat{\zeta}$	SE($\hat{\zeta}$)	95% CI for $\hat{\zeta}$	$t = \hat{\zeta} / SE(\hat{\zeta})$	*P* value	Control % (*C*%)	95% CI for *C*%
W	**–0.17**	**0.035**	**–0.24 ↔ –0.11**	**–4.9**	**<0.001**	**16%**	**10% ↔ 21%**
S	**-0.33**	**0.041**	**–0.42 ↔ –0.25**	**–8.2**	**<0.001**	**28%**	**22% ↔ 34%**

Chi-square test indicated a highly significant effect of wheat type. The estimated among-study variance, however, was only slightly decreased (from 0.036 to 0.032)

Concluding comments

- With over 25,000 journal articles in print, and numerous textbooks, meta-analysis is here to stay
  - In many disciplines, it is the standard approach to research synthesis
  - For some regulatory government agencies, meta-analysis is virtually mandatory (e.g., approval of new drugs or treatments)
- Many issues not covered here, including:
  - Effects of multiple moderator variables
  - Multivariate meta-analytical methods
    - **Multiple response variables ("endpoints")**
    - **Multiple treatments**
    - Complex evidence synthesis
  - Missing values
    - Imputation (single or multiple)
  - Cost-benefits