

EVAL 6970: Meta-Analysis Vote Counting, The Sign Test, Power, Publication Bias, and Outliers

Dr. Chris L. S. Coryn

Kristin A. Hobson

Fall 2013

Agenda

- Vote counting and the sign test
 - In-class activity
- Statistical power
- Publication bias
 - In-class activity
- Outlier analysis
 - In-class activity

Vote Counting

- A method of narrative review in which the number of statistically significant studies is compared to the number of statistically nonsignificant studies using p -values
- The p -value is a function of both the observed effect size and obtained sample size and even if an effect is substantial the obtained p -value will not be statistically significant unless the sample size is sufficiently large

Vote Counting

- Essentially, a statistically nonsignificant p -value is treated as evidence that an effect is absent
- Even so, small, medium, and large effects may yield a statistically nonsignificant p -value due to inadequate statistical power
- Low statistical power is pervasive in most social inquiry

The Sign Test

- Similar to the vote counting method, the sign test is used to count the number of studies with findings in one direction compared to the number of findings in the other direction irrespective of whether the findings are statistically significant
- If a treatment were truly ineffective, it would be expected that half of the studies would lie on each side of the no-effect line

The Sign Test

- The sign test is considered a valid approach when
 - No numeric data are provided from studies, but directions of effects are provided
 - Numeric data are of such different types that they cannot be combined statistically
 - Studies are so diverse in their populations or other characteristics that a pooled effect size is meaningless, but studies are addressing a questions sufficiently similar that the direction of effect is meaningful

The Sign Test

- The sign test takes into account neither the actual effect's magnitudes observed in studies nor the amount of evidence within studies (e.g., sample sizes and precision)
- Can be tested for statistical significance

$$= 2 * \text{BINOMDIST}(n, N, 0.5, \text{TRUE})$$

- With n of the smaller two numbers

First In-Class Activity

- Using your class project data, conduct an analysis using both the vote counting and sign test methods, including the p -value for the sign test, to determine whether the intervention would be considered effective or ineffective
- Contrast the results of the vote counting and sign test methods against those from your meta-analysis

Type I and Type II Error

Type I Error
(false-positive)



Type II Error
(false-negative)



Type I and Type II Error

	H_0 true	H_0 false
Fail to Reject	Correct decision $1 - \alpha$	Type II error β
Fail to Accept	Type I error α	Correct decision $1 - \beta$

Statistical Power

- A Type I error is the conditional prior probability of rejecting H_0 when it is true, where this probability is typically expressed as alpha (α)
- Alpha is a prior probability because it is specified before data are collected, and it is a conditional prior probability, p , because H_0 is assumed to be true

$$\alpha = p(\text{Reject } H_0 | H_0 \text{ true})$$

- where $|$ means assuming or given

Statistical Power

- Both p and α are derived from the same sampling distribution and are interpreted as long-run, relative-frequency probabilities
- Unlike α , p is not the conditional prior probability of a Type I error because it is estimated for a particular sample result
- Alpha sets the risk of a Type I error rate, akin to a false-positive because the evidence is incorrectly taken to support the hypothesis

Statistical Power

- Statistical power, and the concept of Type II error, is the conditional prior probability of making the correct decision to reject H_0 when it is actually false, where

$$\text{Power} = p(\text{Reject } H_0 | H_0 \text{ false})$$

Statistical Power

- A Type II error, or false-negative, occurs when the sample result leads to the failure to reject H_0 when it is actually false
- The probability of a Type II error is usually represented by β , and it is also a conditional prior probability

$$\beta = p(\text{Fail to reject } H_0 | H_0 \text{ false})$$

- Because power and β are complimentary

$$\text{Power} + \beta = 1.00$$

Power for Fixed-Effect Model

- The test for significance for the main effect

$$Z = \frac{M}{\sqrt{V_M}}$$

- For a two-tailed test where

$$p = 2[1 - (\Phi |Z|)]$$

Power for Fixed-Effect Model

- Which is based on

$$\lambda = \frac{\delta}{\sqrt{V_\delta}}$$

- Where

$$\text{Power} = 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda)$$

- See pages 268-269 for a worked example

Power for Random-Effects Model

- The test for significance for the main effect

$$Z^* = \frac{M^*}{\sqrt{V_{M^*}}}$$

- For a two-tailed test where

$$p^* = 2[1 - (\Phi |Z^*|)]$$

Power for Random-Effects Model

- Which is based on

$$\lambda^* = \frac{\delta^*}{\sqrt{V_{\delta^*}}}$$

- With

$$V_{\delta^*} = \frac{V_Y + \tau^2}{k}$$

Power for Random-Effects Model

- Where

$$\text{Power} = 1 - \Phi(c_\alpha - \lambda^*) + \Phi(-c_\alpha - \lambda^*)$$

- See pages 271-272 for a worked example

Publication Bias

- Publication bias is concerned with biases that arise from missing studies in a meta-analysis
 - If missing studies are a random subset of all relevant studies, failure to include these studies will result in less information, wider confidence intervals, and less powerful tests
 - If missing studies are systematically different from located studies, then the sample of studies will be biased

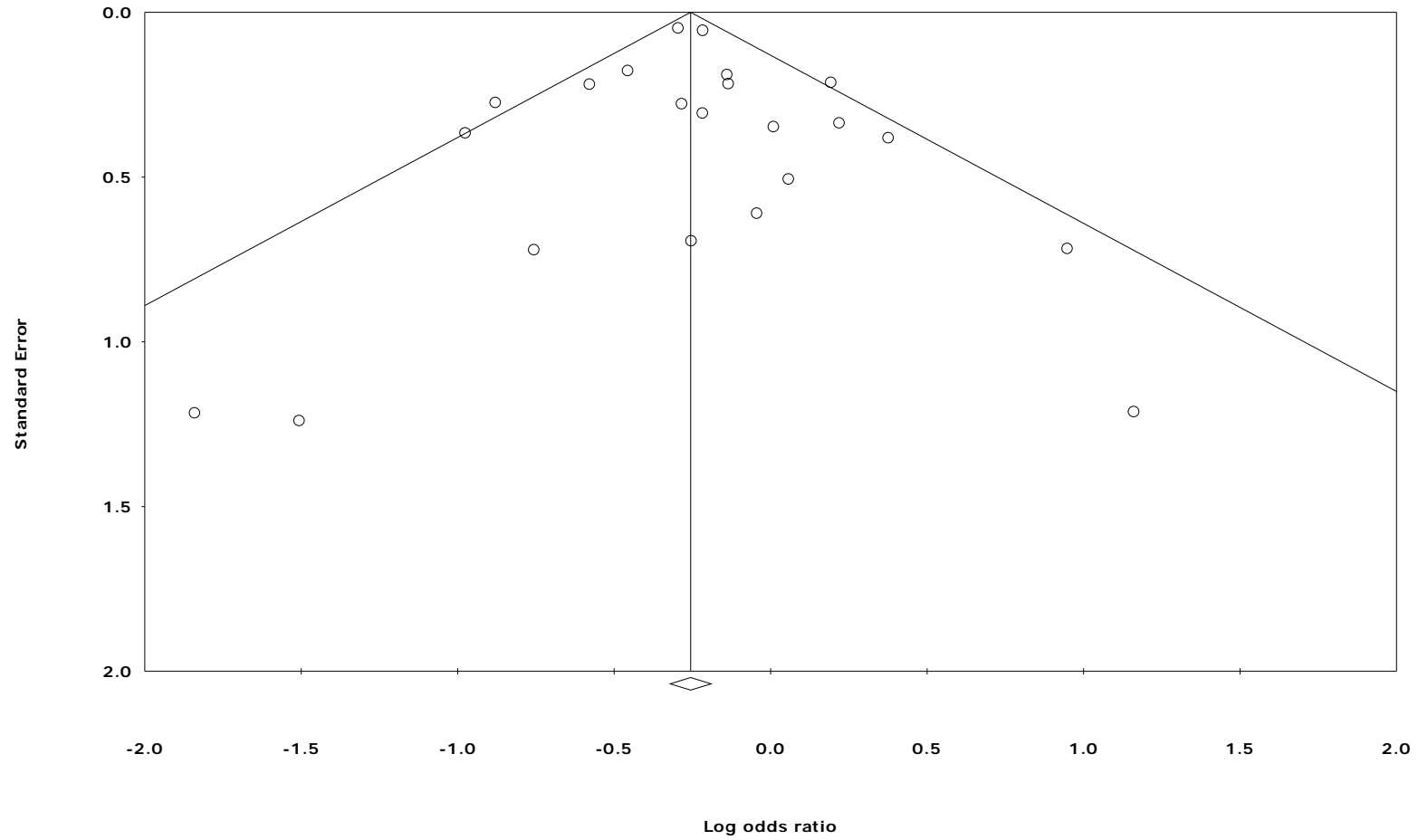
Publication Bias

- Publication bias methods are used to determine if bias is likely, the impact of bias, and to make adjustments
- If these methods are used to estimate an adjusted effect to remove bias, one of three situations potentially arise
 - The resulting effect is essentially unchanged
 - The effect changes, but the basic conclusion remains unaffected
 - The basic conclusion is called into question

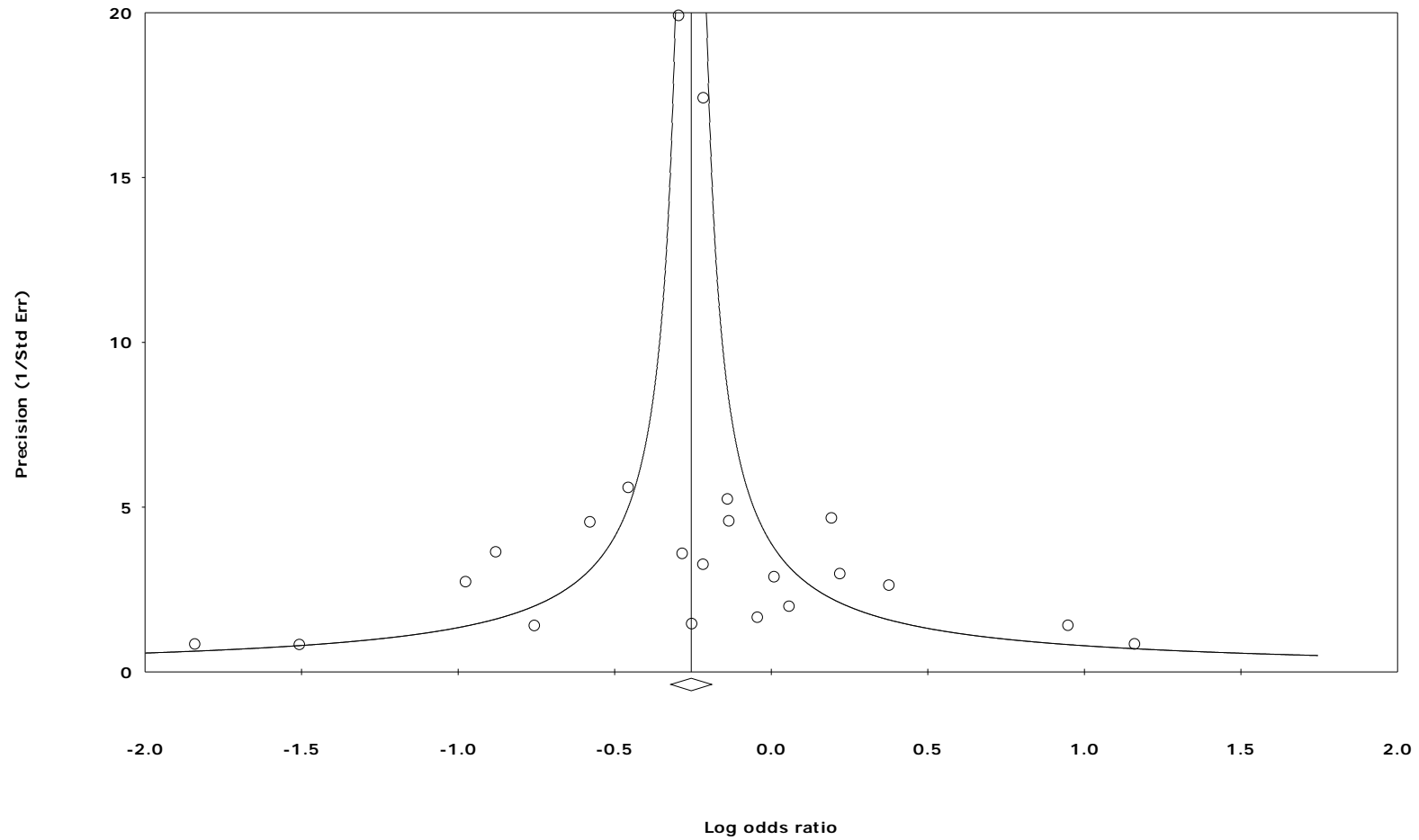
Publication Bias

- The models used to assess publication bias assume
 - Large studies are likely to be published regardless of statistical significance
 - Moderately-sized studies are at risk for being lost, but with moderate sample sizes even modest effects will be statistically significant, and so only some of these studies will be lost
 - Small studies are at the greatest risk of being lost
 - Because of small sample sizes only very large effects are likely to be significant and those with small and moderate effects are likely to be unpublished

Funnel Plot of Standard Error



Funnel Plot of Precision



Evidence of Bias

- From the funnel plots
 - In the absence of publication bias the studies will be distributed symmetrically about the combined effect size
 - In the presence of bias, the bottom of the plot would tend to show a higher concentration of studies on one side of the mean than the other
 - This would reflect the fact that smaller studies (which appear toward the bottom) are more likely to be published if they have larger than average effects, which makes them more likely to meet the criterion for statistical significance

Estimate of the Unbiased Effect Size: Trim and Fill

- Duval and Tweedie's Trim and Fill builds on the idea behind the funnel plot; that in the absence of bias the plot would be symmetric about the summary effect
- If there are more small studies on the right than on the left, the concern is that studies may be missing from the left
- The Trim and Fill procedure imputes these missing studies, adds them to the analysis, and then re-computes the summary effect size

Estimate of the Unbiased Effect Size: Trim and Fill

Comprehensive meta analysis - [Publication bias]

File Edit Format View Computational options Analyses Color Help

← Core analysis ↕ Next table ▾ Funnel plot

Duval and Tweedie's trim and fill

		Fixed Effects			Random Effects			Q Value
	Studies Trimmed	Point Estimate	Lower Limit	Upper Limit	Point Estimate	Lower Limit	Upper Limit	
Observed values		0.77448	0.72553	0.82674	0.78260	0.69266	0.88422	31.51262
Adjusted values	1	0.77365	0.72477	0.82583	0.77979	0.68963	0.88173	32.88028

Look for missing studies where?

Not specified

To left of mean

To right of mean

Look for missing studies using which model?

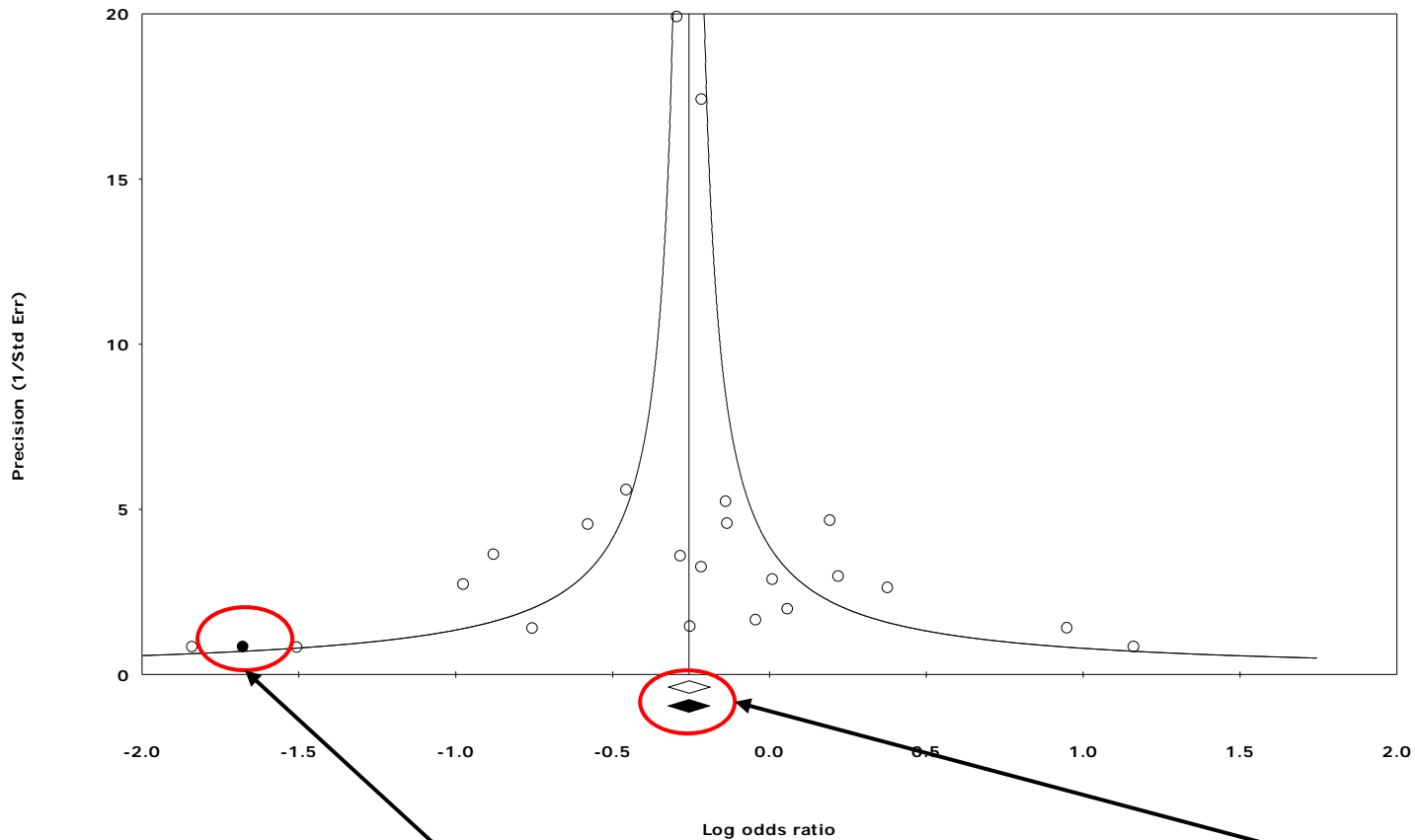
Not specified

Fixed effect model

Random effects model

Here, there is one imputed missing study

Estimate of the Unbiased Effect Size: Trim and Fill



Imputed study

Imputed summary effect

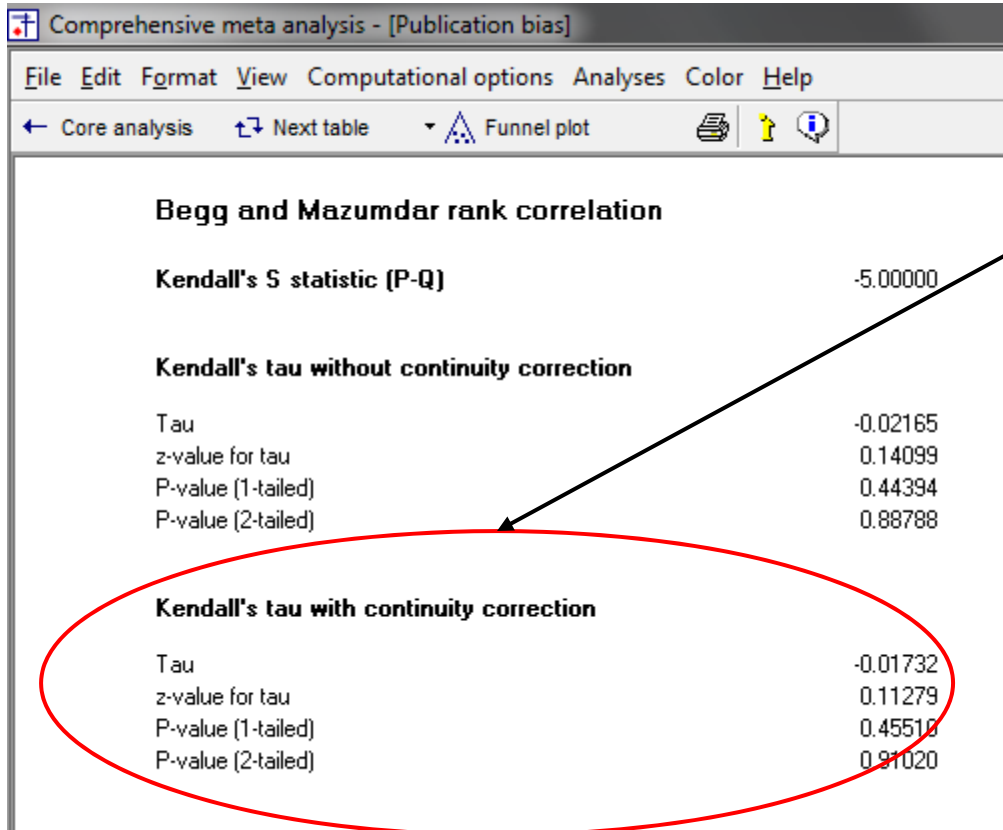
Correlation and Regression Methods

- Begg and Mazumdar's rank correlation test reports the rank correlation (Kendall's τ) between the standardized effect size and the variances (or standard errors) of these effects
 - Kendall's τ is interpreted much the same way as any correlation, with a value of zero indicating no relationship between effect size and precision, and deviations from zero indicating the presence of a relationship

Correlation and Regression Methods

- If asymmetry is caused by publication bias large standard errors (small studies) will be associated with larger effect sizes
 - If larger effects are represented by low values, τ would be positive, while if larger effects are represented by larger values, τ would be negative
 - Since asymmetry could appear in the reverse direction, the significance test is two-sided

Correlation and Regression Methods



Comprehensive meta analysis - [Publication bias]

File Edit Format View Computational options Analyses Color Help

← Core analysis ↗ Next table ▾ Funnel plot

Begg and Mazumdar rank correlation

Kendall's S statistic (P-Q) -5.00000

Kendall's tau without continuity correction

Tau	-0.02165
z-value for tau	0.14099
P-value (1-tailed)	0.44394
P-value (2-tailed)	0.88788

Kendall's tau with continuity correction

Tau	-0.01732
z-value for tau	0.11279
P-value (1-tailed)	0.45516
P-value (2-tailed)	0.91020

Kendall's τ -b
(corrected for ties, if
any) is -0.017, with
a 1-tailed p -value
(recommended) of
0.455

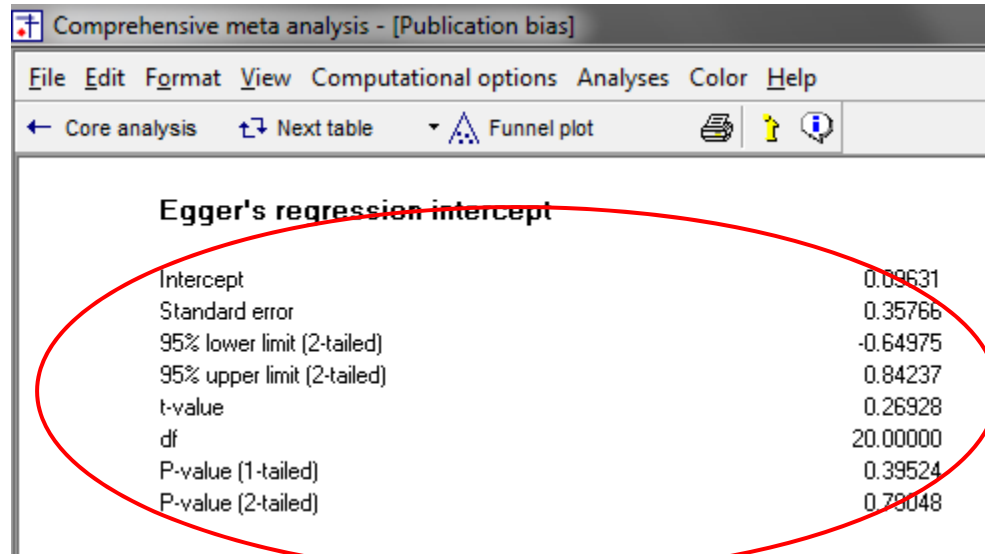
Correlation and Regression Methods

- Egger's linear regression method, like the rank correlation test, quantifies the bias captured by the funnel plot using the actual values of the effect sizes and their precision
- In the Egger test, the standardized effect (effect size divided by standard error) is regressed on precision (inverse of standard error)
- Small studies generally have a precision close to zero, due to their large standard error

Correlation and Regression Methods

- In the absence of bias such studies would be associated with small standardized effects and large studies associated with large standardized effects
 - This would create a regression line whose intercept approaches the origin
 - If the intercept deviates from this expectation, publication bias may be the cause
 - This would occur when small studies are disproportionately associated with larger effect sizes

Correlation and Regression Methods



Comprehensive meta analysis - [Publication bias]

File Edit Format View Computational options Analyses Color Help

← Core analysis ↕ Next table ▾ Funnel plot

Egger's regression intercept

Intercept	0.09631
Standard error	0.35766
95% lower limit (2-tailed)	-0.64975
95% upper limit (2-tailed)	0.84237
t-value	0.26928
df	20.00000
P-value (1-tailed)	0.39524
P-value (2-tailed)	0.79048

Egger's Test of the Intercept indicates an intercept of 0.096, with $t = 0.269$, $df = 20$, and a two-tailed p -value of 0.790

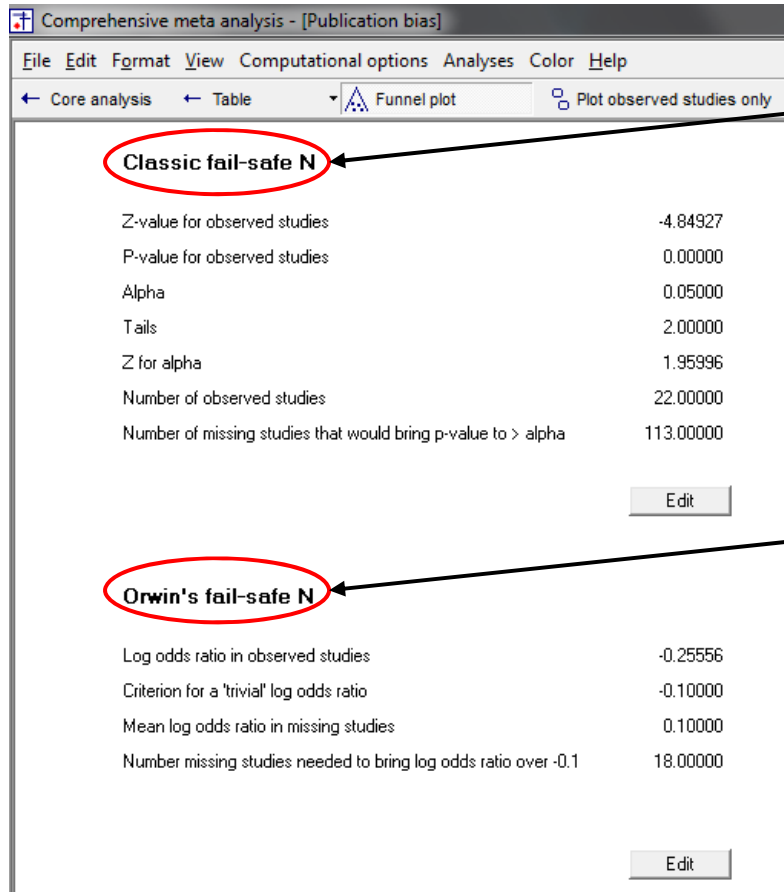
Fail-Safe N

- Rosenthal's Fail-Safe N test computes the number of missing studies (with mean effect of zero) that would need to be added to the analysis to yield a statistically nonsignificant overall effect
- The Orwin variant of this test addresses two problems with Rosenthal's method; that it focuses on statistical rather than clinical significance, and that it assumes a nil overall effect in the missing studies

Fail-Safe N

- Orwin's test allows for selecting both the smallest effect value deemed to be clinically important and a value other than nil for the mean effect in the missing studies
- This method can be used to model a series of other distributions for missing studies
- All Fail-Safe N methods lead to widely varying estimates

Fail-Safe N



Rosenthal's
Fail-Safe N

Orwin's Fail-
Safe N

Second In-Class Activity

- Using your class project data, produce a funnel plot of precision (including any imputed studies using Duval and Tweedie's Trim and Fill method), Kendall's τ , Eggers Test of the Intercept, and Rosenthal's Fail-Safe N
- Interpret all of the publication bias results

Outlier Analysis

- An outlier is an observation that is numerically distant from the rest of the data; that is, a value that appears to deviate markedly from other members of the sample in which it occurs
- Identification of outliers serves multiple purposes and in the context of meta-analysis can most usefully be used as part of a sensitivity analysis that includes meta-analyses with and without outliers (i.e., removed from the analysis)

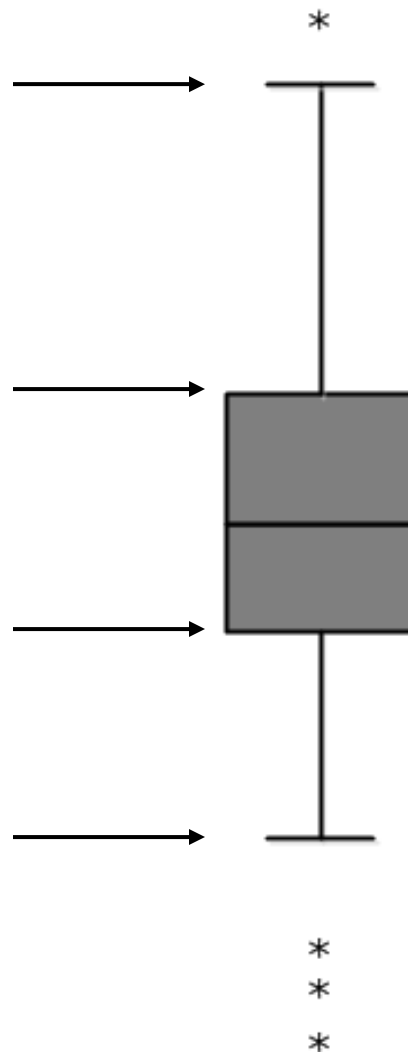
Outlier Analysis

Upper fence/whisker:
Greatest value excluding
outliers

Q3 (Upper quartile, 0.75):
25% of data greater than
this value

Q1 (Lower quartile, 0.25):
25% of data less than this
value

Lower fence/whisker:
Least value excluding
outliers



Outlier: $1.5 \times \text{IQR}$
(The IQR is the inter-
quartile range = the
distance between
Q1 and Q3)

Median (0.50)

Outliers: $1.5 \times \text{IQR}$
(The IQR is the inter-
quartile range = the
distance between
Q1 and Q3)

Third In-Class Activity

- Using your class project data, conduct a statistical outlier analysis on the study effect sizes
- If there are outliers, estimate the summary effect including all studies as well as excluding statistical outliers
 - How similar or dissimilar are the summary effects?