Missing Data

Patrick Kearns
Tenzing Ingty
3/5/2014

What is missing data?

- What do we mean by missing data?
  - Not in the right form or easily accessed
  - Has data, but is missing key things (e.g. sample size)
  - No useful data (e.g. on p-values)
  - Significance of p-values?
  - All the above

What are we missing?

- Correlations
- Variance
- Sample sizes
- Means

Usefulness for meta-analysis	Study statistics	What is available		Addressing what's missing
high	Completely reported	Has all the data for inclusion	→	Nothing missing!
	Selectively reported	All the data are available but not in forms that are easily integrated into meta-analysis (e.g., data in figures, sample sizes need to be determined from table, *t*-tests and means are not reported, etc.)	→	Extract data from figure or tables (see Chapter 5), convert available statistics (e.g., *t*-test into effect size)
	Partially reported	Has some data (e.g., sample sizes) but is missing information that cannot be estimated directly from what is available (e.g., variance estimates)	→	Recalculation or conversion of available statistics (back calculation from *P*-values), or within-study imputation methods
	Qualitatively reported	No useful data except for *P*-values or discussion regarding the significance or non-significance of analysis	→	Recalculation of statistics, or use within-study imputation methods or multiple-imputation methods
low	Unreported	No statistics or data are available, although may have specified a protocol for the analysis in the Methods section	→	Exclude from meta-analysis or use an alternative approach to meta-analysis (e.g., vote-count methods)

Why are we missing data?

- Publication page/letter limits
- Lack of publication of null results
- Perceived lack of importance
  - P-value

How does it affect us?

- Exclude studies
  - Too much exclusion can induce type II error
  - Study quality?
- Too small a sample size can under/overestimate effect size

How to handle it

- Talk with other researchers
- Algebraic recalculations
  - Able to convert p-values and other stats metrics to usable data
  - Rely on the authors to have data/stats that don't violate assumptions of normality

Hedge's D from T-Test

d	effect size
$\bar{X}$	sample mean
T and C	treatment and control groups
SD	standard deviation
n	sample size
J	bias correction factor
s	pooled SD

$$d = \frac{(\bar{X}_T - \bar{X}_C)}{s} J \qquad J = 1 - \frac{3}{4(n_T + n_C) - 9}$$

$$s = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_C - 1)SD_C^2}{n_T + n_C - 2}}$$

$$d = t\sqrt{\frac{n_T + n_C}{n_T n_C}} \qquad d = \frac{2t}{\sqrt{n_{total}}} \qquad d = t_R\sqrt{\frac{2(1 - r_R)}{n_{total}}}$$

Hedge's D from ANOVA

d	effect size
$\bar{X}$	sample mean
T and C	treatment and control groups
SD	standard deviation
n	sample size
J	bias correction factor
s	pooled SD

$$d = \frac{(\bar{X}_T - \bar{X}_C)}{s} J \qquad J = 1 - \frac{3}{4(n_T + n_C) - 9}$$

$$s = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_C - 1)SD_C^2}{n_T + n_C - 2}}$$

$$|d| = \sqrt{\frac{F(n_T + n_C)}{n_T n_C}} \qquad |d| = 2\sqrt{\frac{F}{n_{total}}}$$

Correlation coefficient (r)

key terms	
r	Pearson product-moment correlation
x and y	variables under analysis
n	total sample size

definition

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}$$

linear regression

$$r = \beta\left(\frac{SD_x}{SD_y}\right) \quad \text{if } y = \alpha + \beta x$$

SD = standard deviation, α = intercept, β = slope

biserial r (r_b)

$$r \approx r_b$$

point-biserial r (r_{pb})

$$r_b = \frac{r_{pb}\sqrt{n_T n_C}}{u(n_T + n_C)}$$

u = ordinate of unit normal distribution (see Terrell 1982)

independent t-test

$$r_{pb} = \sqrt{\frac{t^2}{t^2 + n_T + n_C - 2}} \qquad r_{pb} = \sqrt{\frac{t^2}{t^2 + df}} \qquad |r_{pb}| = \frac{P}{\sqrt{P^2 + 4}}$$

df = degrees of freedom, P = P-value

Hedges' d

$$r = \sqrt{\frac{d^2 n_T n_C}{d^2 n_T n_C + n(n-1)}}$$

$n = n_T + n_C$