Multiple test procedures and smile plots

Roger Newson (King's College, London, UK)

roger.newson@kcl.ac.uk

The ALSPAC Study Team

http://www.alspac.bris.ac.uk

- Dangers of multiple tests and confidence intervals.
- Corrected confidence intervals for "micro-scale data mining".
- Smile plots and multiple test procedures.
- Controlling the familywise error rate (FWER) for "medium-scale data mining".
- Controlling the false discovery rate (FDR) for "mega-scale data mining".

Dangers of multiple tests and confidence intervals

- Scientists often have good reasons for wanting to measure multiple parameters. (Especially when scanning genomes.)
- Unfortunately, 5% of *sample* differences will be significant at the 5% level (and have 95% confidence intervals excluding zero), even if all *population* differences are zero.
- Epidemiologists, including genetic epidemiologists, are commonly accused of making much of their living out of “significant differences” of this kind. (See Colhoun *et al.*, 2003.)
- A sceptical public will therefore rightly be suspicious of “significant” results published, especially if they are “highlights” from a large number of parameters measured.
- Therefore, scientists need to be able to address this scepticism. (Especially in themselves.)

Corrected confidence intervals in Stata

- Given n true null hypotheses and a threshold P-value α, the Bonferroni-corrected threshold for testing the smallest and “most significant” P-value is equal to α/n.

- If the P-values are from 2-tailed tests based on multivariate Normal test statistics, then we can use the less conservative Šidák-corrected threshold, equal to $1-(1-\alpha)^{1/n}$.

- Most scientists, most of the time, view P-values as a means to the end of defining confidence intervals (or other confidence regions).

- It is possible (using `correlate` or `parmest`) to calculate Bonferroni-corrected $100\,(1-\alpha/n)\%$ confidence intervals, or Šidák-corrected $100\,(1-\alpha)^{1/n}\%$ confidence intervals, for each of the n parameters.

- We can then be *at least* $100\,(1-\alpha)\%$ confident that *all* of the n parameters are inside their respective corrected confidence limits.

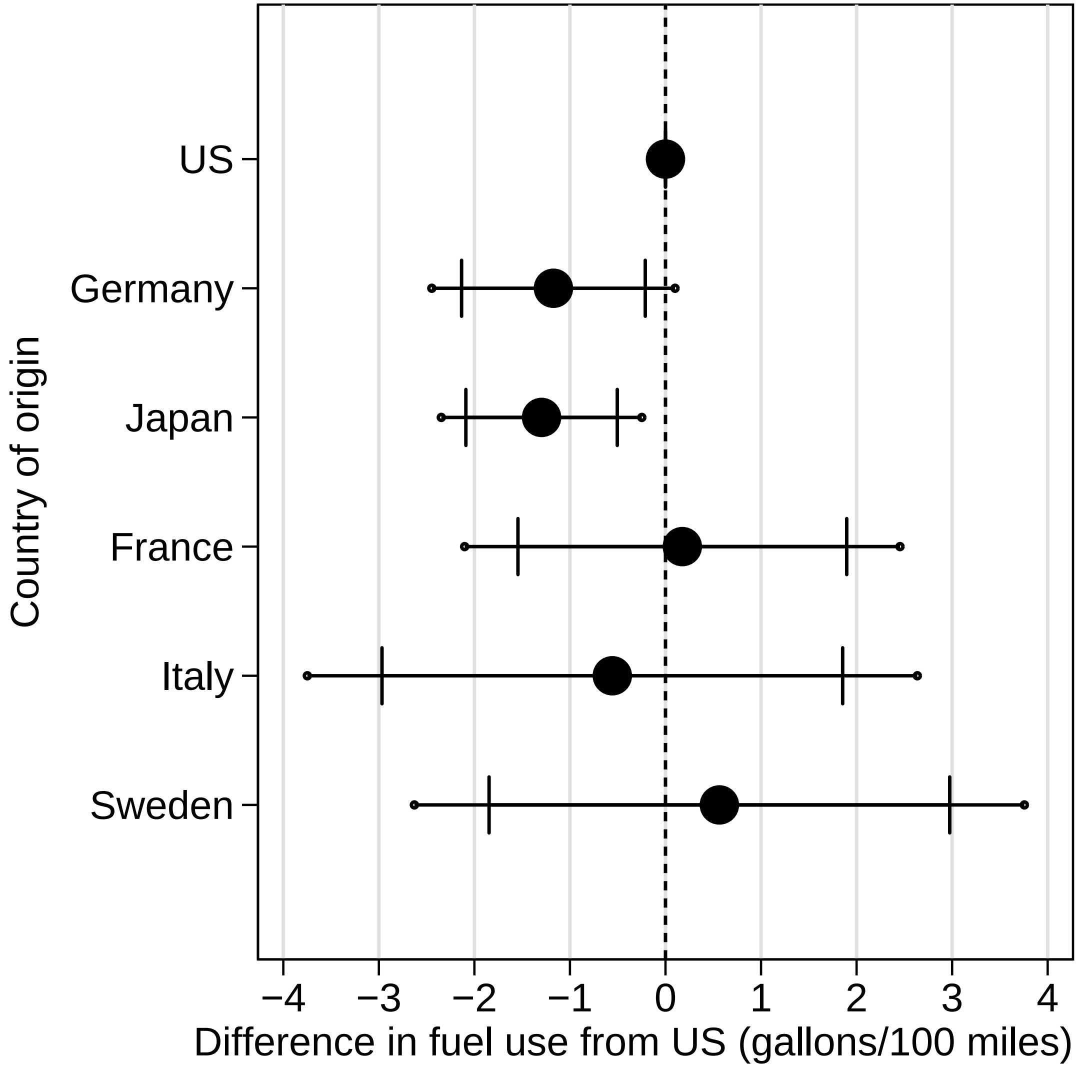
Differences in fuel consumption between non-US and US cars in the `auto` data

- `eclplot` plots mean fuel use differences in the `auto` data between cars from 5 non-US countries and US cars.

- For each non-US country, the difference is displayed with uncorrected 95% and Šidák-corrected 98.98% confidence limits. (Note that `parmest` allows non-integer confidence levels.)

- Japanese cars consume fewer gallons of fuel per 100 miles than US cars, even considering that there are 5 comparisons.

- We are less sure about German cars.



Multiple test procedures and smile plots

- Most scientists, most of the time, do not use corrected confidence intervals.
- They may be unreliable at confidence levels far above 99%, and are certainly conservative.
- Multiple test procedures *may* be more reliable and less conservative. They define confidence regions for a non-numeric parameter, *usually* "the set of null hypotheses that are true".
- Typically, they take, as input, a set of observed P-values and an uncorrected P-value threshold, calculate a corrected P-value threshold, and define a subset of "rejected null hypotheses" (or "discoveries"), whose P-values are at or below the corrected threshold.
- The `smileplot` package, downloadable from SSC, contains the programs `multproc`, `smileplot` and `smileplot7`. It takes, as input, a data set with one observation per measured parameter and data on P-values (eg a `parmest`, `statsby` or `postfile` output).
- The program `multproc` carries out a range of multiple test procedures. The programs `smileplot` and `smileplot7` express a multiple test procedure graphically by plotting the P-values on a reverse log scale on the Y-axis against another variable (*usually* the corresponding estimates, but *possibly* the positions of the corresponding genes on a chromosome) on the X-axis.

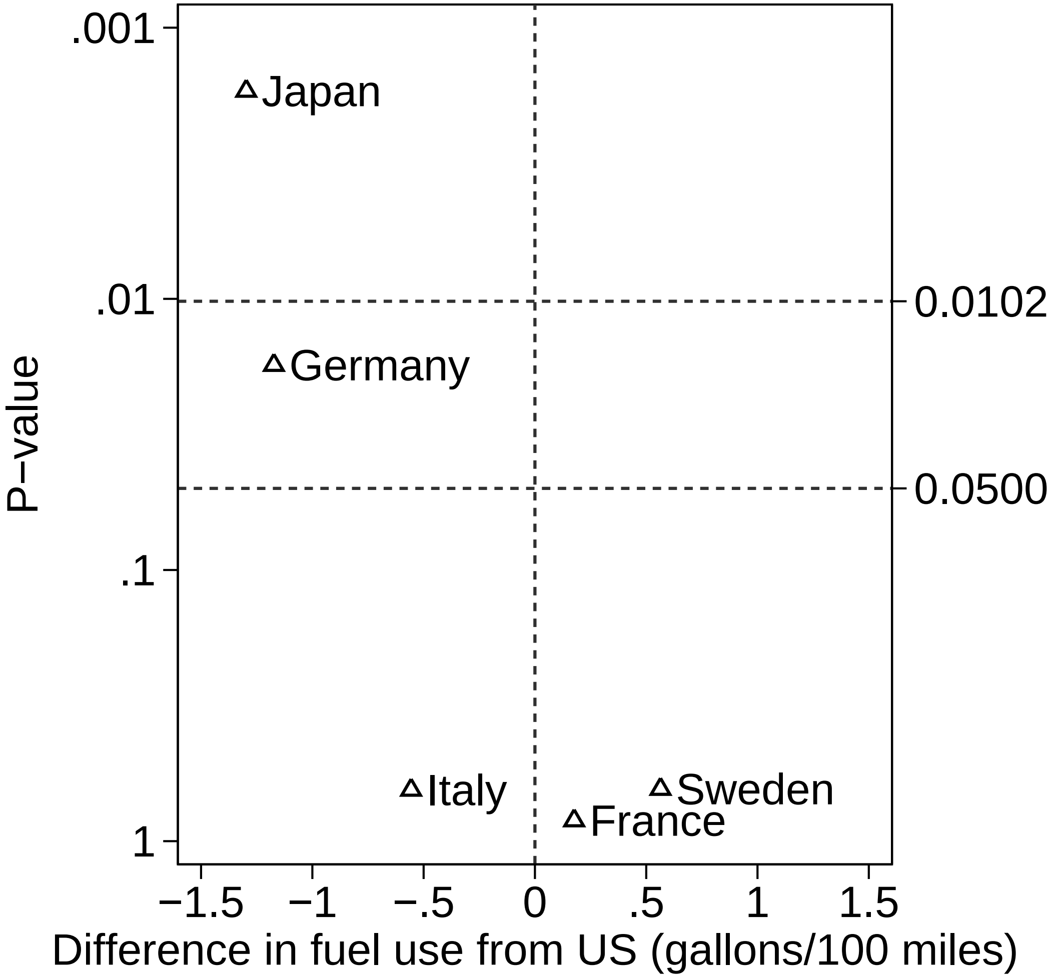
Smile plot for differences in fuel consumption between non-US and US cars

- The data points are mean differences from US cars (labelled by country).

- The X-axis measures *practical* significance. The reference line indicates the null hypothesis.

- The Y-axis measures *statistical* significance. The reference lines indicate uncorrected and Šidák-corrected threshold P-values.

- A doubling (halving) of the number of measured parameters shifts the corrected threshold up (down) by *approximately* 0.3 $\log_{10}$-units.



Example: Mother's diet in pregnancy and child's history of eczema and wheezing (ALSPAC study, Bristol University)

- Mothers of 12028 children completed a food frequency questionnaire (FFQ) at 32 weeks gestation.

- They also completed questionnaires when the child was 6, 30 and 42 months old on wheezing and eczema history.

- 33 FFQ-derived diet exposures were calculated. The 5 outcomes were 18-30 month eczema, 30-42 month wheezing, "early transient" wheezing, "late-onset" wheezing, and "persistent" wheezing.

- Associations were measured by logistic regression, using per-category odds ratios for categorical exposures and per-doubling odds ratios for continuous exposures.

- There were therefore $33 \times 5 = 165$ unadjusted odds ratios, with confidence limits and P-values. (And 165 corresponding confounder-adjusted odds ratios.)

Controlling the familywise error rate for medium-scale data mining

- The uncorrected threshold P-value of a multiple test procedure may be the **familywise error rate (FWER)**, which is the probability that *at least one* true null hypothesis is rejected.

- Procedures controlling the FWER include the Bonferroni, Šidák, Holm and Holland-Copenhaver procedures.

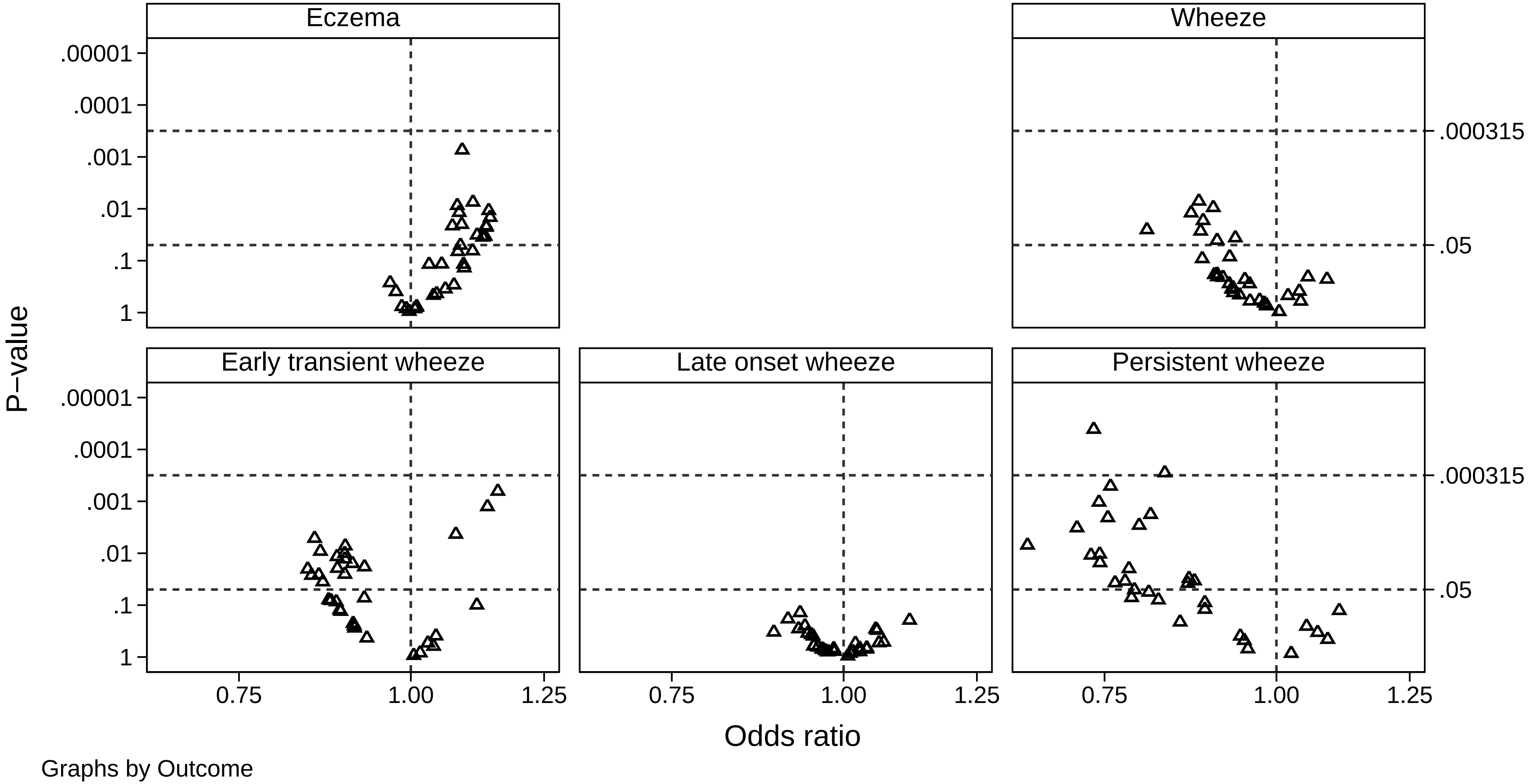
- They define a power-set-valued confidence region for "the set of null hypotheses that are true", namely the power set of the set of non-rejected null hypotheses.

- If the FWER is controlled at α, then we are $100(1-\alpha)\%$ confident that *all* rejected null hypotheses will be false.

- The price of this confidence is that the corrected critical P-value (and therefore the power to detect a difference of a given size) tends to zero as the number of estimated parameters becomes large.

- A FWER-correcting procedure is therefore typically not much less conservative than the Bonferroni procedure, and is *not* a good way to assess a "mega-scale" data mining expedition, such as a productive scientific career.

Holland-Copenhaver smile plots (FWER=0.05) for 5 outcomes and 33 diet exposures



Of 165 associations measured, only 2 (involving persistent wheeze) are "discovered" by the procedure. *However*, we may be 95% confident that *both* discoveries are true.

Controlling the false discovery rate for “mega-scale” data mining

- Let R denote the number of discoveries made by a multiple test procedure, and let V denote the number of such discoveries that are false. The **false discovery rate (FDR)** of a multiple test procedure is defined as $E[Q]$, where

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

- The `smileplot` package offers a range of FDR-controlling procedures, including the Simes procedure and the Yekutieli-Benjamini procedure.

- The FDR is a hybrid quantity. If all null hypotheses are true, then it is equal to the FWER.

- *On the other hand*, if we are “data mining in a fairly rich seam”, then R will almost never be zero, and the FDR will then approximate the proportion of discoveries that are false.

- *Therefore*, instead of aiming to control the number of false discoveries at zero, a FDR-controlling procedure aims to control the number of false discoveries as a proportion of the number of *true* discoveries, from which a productive scientist makes his/her reputation.

False discovery rates: the story so far

- FDR-controlling procedures (eg the Simes procedure) were first proposed in 1995. However, the seminal paper justifying them for most practical uses was published in 2001.

- FDR-controlling procedures remain controversial, but look very promising. (I currently use the Simes procedure to provide "footnote analyses", rather than "bottom line analyses".)

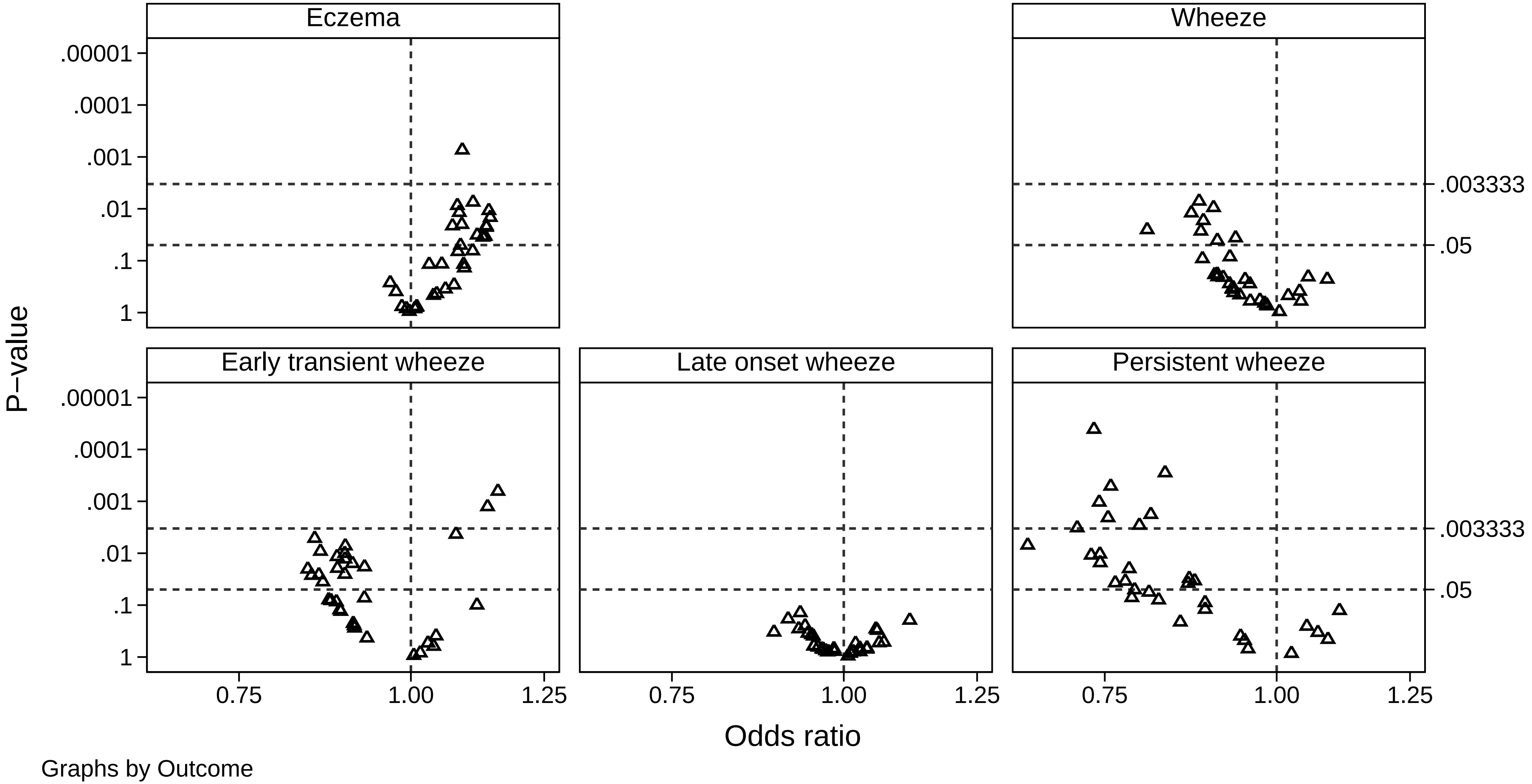
- Genovese and Wasserman (2002) showed that, *if* all the P-values are independent, *and* there is a non-zero probability that a null hypothesis is false, *then* the Simes-corrected P-value (and therefore power) converges to a non-zero minimum as the size of the data-mining expedition becomes large. (Unlike the old FWER-corrected P-values, which converge to zero.)

- Benjamini and Yekutieli (2001) showed that, if the P-values are *not* independent, then the Simes procedure still works for P-values from two-tailed multivariate normal test statistics, and the Yekutieli-Benjamini procedure works in the more general case.

- FDR-controlling procedures still probably err on the side of conservatism. (*However*, this seems to be less of a problem if we are "mining lean paydirt".)

- If *all* null hypotheses are true, then FDR-controlling procedures are FWER-controlling. (So they can only help scientists who help themselves by finding "worthwhile paydirt".)

Simes smile plots (FDR=0.05) for 5 outcomes and 33 diet exposures



This time, 11 associations (involving eczema, early transient wheeze and persistent wheeze) are "discovered". *However*, we may *not* be 95% confident that *all* of them are true.

So what *can* we be confident about?

- When I first read about FDR-controlling procedures, they seemed to me to have wonderful properties in the limiting case, where a scientist carries out an infinitely large number of infinite-sized data mining expeditions.

- *If* the scientist publishes a new, independent data mining expedition every week, controlling the FDR at 5%, *then*, by consistency laws, s/he *may* end his/her career 100% confident that 95% of his/her discoveries were true.

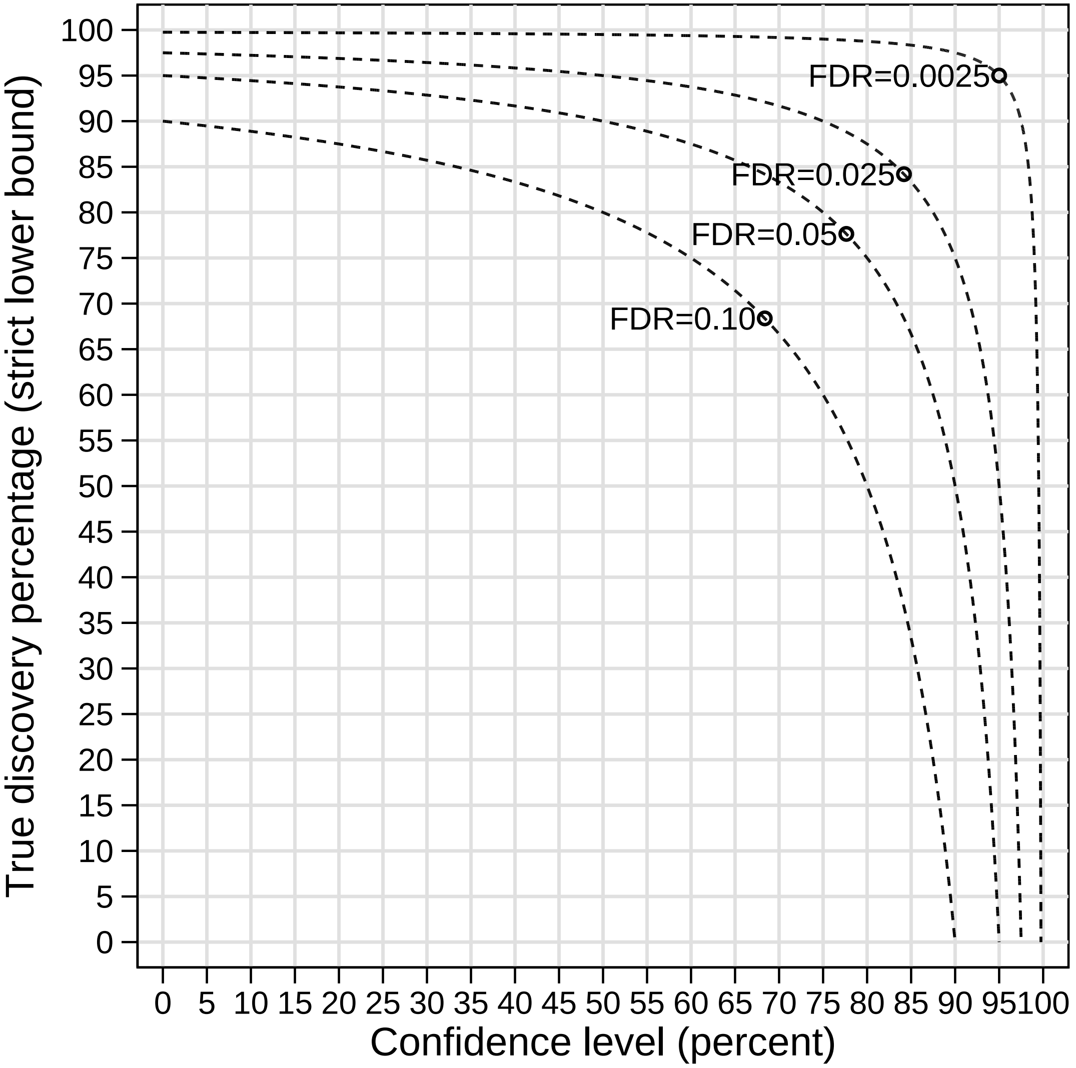
- *However*, as lower-ranking scientists, we might want to be able to make confidence statements about the list of discoveries in our own single and finite data mining expedition. (As we could do with the old FWER-controlling procedures.)

- Fortunately, it is easily shown (Newson, 2003) that, if we control the FDR at a level $\alpha = \beta \times \gamma$, then we can be $100(1 - \beta)\%$ confident that, *if* any discoveries are made, *then* strictly more than $100(1 - \gamma)\%$ of them will be true.

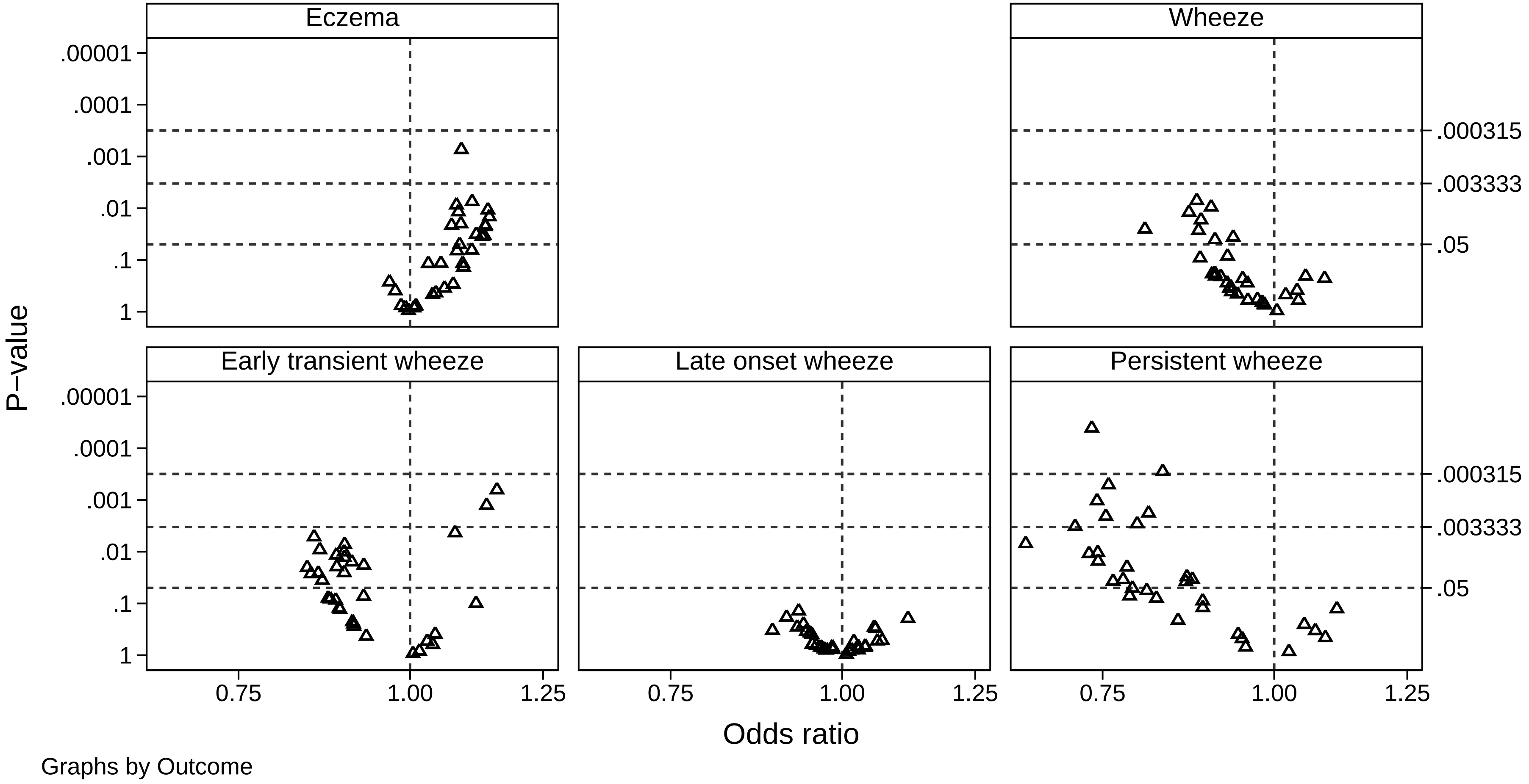
- For instance, if we control the FDR at 0.05, then we can be 95% confident that *some* of our discoveries are true, or, alternatively, 90% confident that *most* of our discoveries are true.

The trade-off between confidence level and true discovery percentage

- Each curve corresponds to a level of FDR.
- The bottom of each curve gives the level of confidence $100(1-\mathrm{FDR})$ that *some* of the discoveries are true.
- With a small sacrifice of confidence level, we can be confident about much more.
- If the FDR is 0.0025, then we can be 95% confident that over 95% of any discoveries are true.
- (Note that a confidence level is *not* a Bayesian posterior probability.)



Simes-Holland-Copenhaver smile plots for 5 outcomes and 33 diet exposures



11 associations were “discovered”. We may be 95% confident that *some* of them are true, or 90% confident that *most* of them are true. Or 95% confident that *the top 2* are true.

Further reading

Benjamini, Y. and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29: 1165–1188.

Colhoun, H. M., P. M. McKeigue and G. Davey Smith. 2003. Problems of reporting genetic associations with complex outcomes. *The Lancet* 361: 865-872.

Genovese, C. and L. Wasserman. 2002. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* 64: 499-517.

Newson, R. 2003. Multiple test procedures and smile plots. *The Stata Journal*, in press.

The `parmest`, `eclplot` and `smileplot` packages, used in this presentation, can all be downloaded from SSC. (In Stata, type `help ssc` to find how to do this.)