Interpreting evidence cheatsheet

Gavin Band, WHG GMS Programme 2021

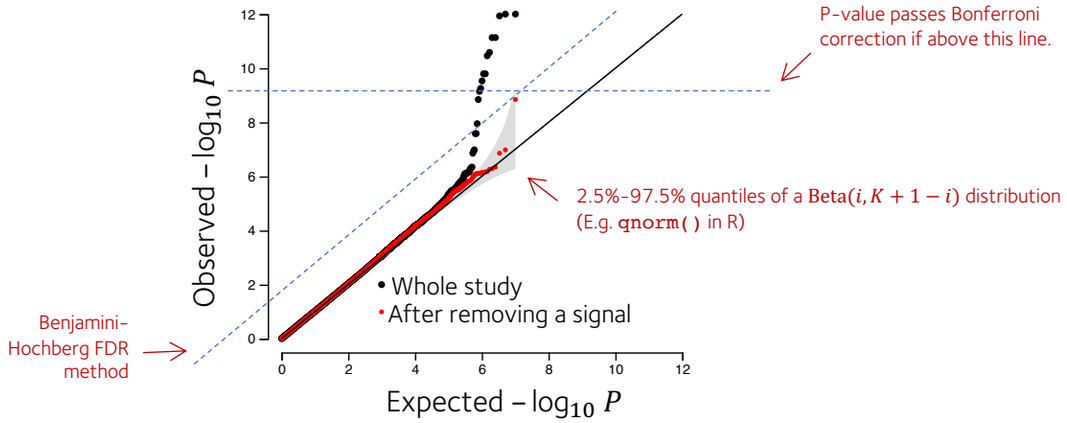
Let's suppose you have run your GWAS or other large-scale study, and now you have a bunch of effect size estimates, standard errors, and P-values or Bayes factors. How do you interpret them?

Notation. I'll write $\hat{\beta}_i$, se_i and p_i. for the estimated effect size, standard error, and P-value for the ith comparison in your study. Meanwhile β_i (with no hat) will denote the *true effect for comparison i.* And suppose there are K comparisons in total.

Model. To make sense of this, we have to have some sort of model of what's going on. We will use the following model in which we imagine that **true nonzero effects are rare**, and that they occur with a fixed probability Π. That is,

$$P(\beta_i \neq 0 | \text{our study design}) = \Pi \qquad \text{(before we see any data)}.$$

First please note that **this model is daft.** *Can* we divide nonzero from zero effects like this? For example, for a GWAS of polygenic traits it seems quite likely that much of the genome is associated, it's just that the effects are very small. This is worth thinking about in interpreting results! Thinking this way leads to the idea of trying to estimate the distribution of true effects, which is clearly what we want. Nevertheless this dichotomous model of 'most things are zero but some aren't' does often hold approximately in many studies (like the one depicted below) and we'll go with it here.



Interpretation 1 – the '**controlling for multiple comparisons**' interpretation. It typically focusses on the P-values, and works like this: we act as though the worst thing in the world is to make any false positive calls (at all) in our study, and we try to pick a P-value threshold T to avoid that. In other words we try arrange that:

A probability, or **expectation over repeated studies** →

$$P\left(\min_i p_i < T \,|\, \beta_i \equiv 0 \text{ for all } i\right) < \alpha$$

For some chosen desired false positive rate α. As expressed above this is all about the minimum P-value, whose distribution is Beta as drawn on the above qqplot. The simplest approach is to require $T = \alpha/K$ (Bonferonni), which is equivalent to drawing a line at the 97.5% quantile of that Beta distribution and declaring everything above it 'significant'.

For a human GWAS study of a single trait, it has traditionally been taken that there are about a million 'independent tests worth' of variants in the genome, so to achieve $\alpha = 0.05$ we need the threshold to be $T = 5\times10^{-8}$. (But many studies widely compute their own thresholds.)

It is true that if you were to repeat studies over and over, and use a rule like Bonferonni to declare significance at $\alpha = 0.05$, you would mistakenly declare an association < 5% of the time.

This is however unsatisfying because: 1: most people don't have the luxury of repeating their study over and over, 2: most people have already seen their data and want to know 'how likely was this signal to be a real association?'

Interpretation 2 – the **scientific approach**. How likely is it that an effect is nonzero given we see a P-value smaller than T? It is easy to work out:

$$P(\beta \neq 0 | p < T) = \frac{P(p < T | \beta \neq 0) \cdot P(\beta \neq 0)}{P(p < T)}$$

← Apply Bayes' rule to LHS

Or in other words:

$$P(\beta \neq 0 | p < T) = \frac{\text{power} \cdot \Pi}{\text{power} \cdot \Pi + T \cdot (1 - \Pi)}$$

What this makes clear is that for a given threshold T the interpretation of the P-value depends on both the association test power $P(p < T | \beta \neq 0)$ and the prior probability Π.

Power what now? The power says "how likely are we to see $p < T$ if the effect is really nonzero? A bit of thought will convince you that this depends on the true effect size. (It also depends on the other things that affect power, like the sample size and the predictor and outcome frequencies). In most experiments, we don't know the distribution of true effects up front – so we have to guess at it or try to estimate it.

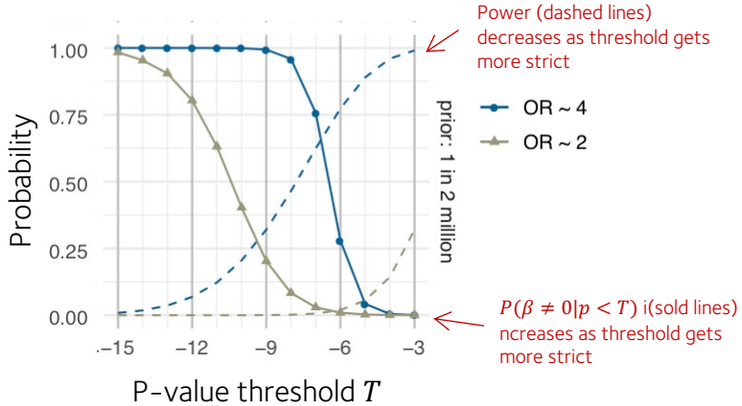
Prior what now? It is explicit in the above that the evidence for association **depends on the prior as well as the power**.

I've called this the 'scientific' approach because it relates the things we want to know – how big are the effects? How many effects are there? Now we've seen the data, how likely was this to be an effect?

For GWAS-like effects of small magnitude the power can be approximately worked out. For example, in a case-control trait with case proportion ϕ and N samples, the association test standard error is approximately

$$\text{se} \approx \frac{1}{\sqrt{2N\phi(1-\phi)f(1-f)}}$$

where f is the variant frequency. For a given effect size b, the power is then approximately computed by the $|x| > b$ tails of the Gaussian with that standard error. This leads to a power analysis like this:



Example The plot on the left is for a disease at 1% frequency and a SNP at 50% frequency, with a prior of 1 in 2 million.

For an experiment to be successful, we want to have a threshold that gives us high power (chance to detect signals) and a high chance that the detected signals are real.