In the next few lectures we will address modern data analysis methods that seek to overcome shortcomings of the simple hypothesis testing framework we discussed for differential abundance estimation.

In particular, the following are three problems we are addressing (see slides):

1. Unstable variance estimates due to moderate sample sizes degrade gene ranking for differential abundance.

2. Parametric assumptions (e.g. t-test) are not appropriate.

3. Multiple testing

We will study the following approaches for each:

1. Empirical Bayes Methods
2. Permutation methods
3. Multiple testing correction
Testing revisited:

We will use this notation based on linear modeling and consider a two-group test as example:

**Goal:** Find genes that are differentially abundant in group A (e.g. dormant pathogen) vs. group B (e.g. multiplying pathogen).

We can model expression of gene $g$, sample $i$ as

$$E(y_{gi}) = \mu_g + \beta_g x_i + \epsilon_{gi}$$

where $\epsilon_{gi} \sim N(0, \sigma_g^2)$ & $x_i$ is an indicator

$$x_i = \begin{cases} 
1 & \text{if group B} \\
0 & \text{o.w.} 
\end{cases}$$

With this $\hat{\beta}_g$ is an estimate of the difference in expression of gene $g$.

Recall the standard testing formulation:

1. Compute an observed difference statistic ($\hat{\beta}_g$)
2. Scale statistic considering error (e.g. $t$-statistic)

$$t = \frac{\hat{\beta}_g}{\text{s.e.}(\hat{\beta}_g)} = \frac{1}{\sqrt{n}} \frac{\hat{\beta}_g}{\hat{\sigma}_g}$$

3. Assume "no difference" null hypothesis (e.g. $\beta_g = 0$)
4. Compute $P(\hat{\beta}_g | \beta_g = 0)$: the $p$-value

Reject "no difference" hypothesis if $p$-value $\leq \alpha$ (e.g. 0.05)
As we saw estimating $g$ is unstable for small/moderate sample sizes. However, we have $g = 1, \ldots, m = 25k$, can we borrow information across genes to improve these estimates? This is where empirical Bayes methods are used. You may see this general frequentist modeling strategy under different names:

1. Shrinkage
2. Mixed models
3. Random effect models
4. Hierarchical models

The modeling strategy is the same, but the treatment (intervene) done by these methods is different.

The core idea is to use Bayes rule as a method to incorporate prior knowledge/belief into analyses:

$$P(G | X) \propto P(X | G) P(G) \propto \frac{\text{posterior probability}}{\text{prior probability}}$$

The likelihood of data $X$ given (e.g.) some hypothesis $G$ having observed data $G$

In general, settings obtaining $P(G)$ (the prior) is hard or difficult to defend. Empirical Bayes methods use data $X$ to estimate prior parameters.
The simplest example (the normal-normal model):

\[ y_g = \theta_g + \epsilon_g \]

\[ \epsilon_g \sim N(0, \sigma^2) \]

\[ \theta_g \sim N(0, \tau^2) \]

with \( \sigma^2 \) and \( \tau^2 \) known.

We want to know what the posterior distribution of \( \theta_g | y_g \) is.

\[
\begin{align*}
    p(\theta_g | y_g) &\propto p(y_g | \theta_g) p(\theta_g) \\
    &\propto \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{1}{2} \frac{(y_g - \theta_g)^2}{\sigma^2} \right) \times \frac{1}{\sqrt{2\pi \tau^2}} \exp \left( -\frac{1}{2} \frac{\theta_g^2}{\tau^2} \right)
\end{align*}
\]

We only care about the exponent and terms that depend on \( \theta_g \):

\[
\begin{align*}
    \left[ -\frac{1}{2} \frac{(y_g - \theta_g)^2}{\sigma^2} + \frac{\theta_g^2}{\tau^2} \right] &= -\frac{1}{2} \left[ \frac{y_g^2 - 2y_g \theta_g + \theta_g^2}{\sigma^2} + \frac{\theta_g^2}{\tau^2} \right] \\
    &= -\frac{1}{2} \left[ \frac{y_g^2}{\sigma^2} - 2y_g \frac{\theta_g}{\sigma^2} + \frac{\theta_g^2}{\tau^2} \right] + \frac{\theta_g^2}{\sigma^2 + \tau^2}
\end{align*}
\]

\[
\begin{align*}
    \frac{y_g^2 \tau^2 - 2y_g \theta_g \tau^2 + \tau^2 \sigma^2 + \sigma^2 \theta_g^2}{\sigma^2 + \tau^2} &= \frac{1}{\sigma^2 + \tau^2} \\
    \theta_g \left( \frac{\tau^2 + \sigma^2}{\sigma^2 + \tau^2} \right) - 2y_g \frac{\theta_g}{\sigma^2 + \tau^2} &= \left( \theta_g - y_g \frac{\tau^2}{\sigma^2 + \tau^2} \right)^2 \\
    &= \frac{\theta_g^2}{\sigma^2 + \tau^2} \\
    \theta_g^2 &= \frac{\tau^2}{\sigma^2 + \tau^2} \\
    \theta_g &= \frac{\tau^2}{\sigma^2 + \tau^2}
\end{align*}
\]
This implies that
\[ \Theta_3 | y_3 \sim N(\lambda y_3, \sigma_3^2) \] where \( \lambda = \frac{\sigma_3^2}{\sigma_y^2 + \sigma^2} \)

Consider the case where \( \sigma^2 = 0 \) (there is no variance in the prior:

\[ p(\theta) \sim N(0, 0) \]

In this case \( \Theta_3 | y_3 \sim N(0, 0) \) since \( \lambda = 0 \). Regardless of what data is \( (y_3) \) the posterior will equal the prior.

Now consider the case where \( \sigma^2 \to \infty \) (there is no precision in the prior:

\[ p(\theta) \sim N(0, \infty) \]

In this case \( \Theta_3 | y_3 \sim N(y_3, \sigma_3^2) \) since \( \lambda = 1 \). In this case the posterior equals the likelihood. So, in summary depending what the relationship between prior variance and data variance is \( (\lambda = \frac{\sigma_3^2}{\sigma_y^2 + \sigma^2}) \), the posterior estimate is “shrunken” towards the prior.

In HWII, you will consider more interesting models:

\[ y_3 = \Theta_3 + \varepsilon_3 \]

\[ \varepsilon_3 \sim N(0, \sigma_3^2) \]

\[ \varepsilon_3 \sim N(0, \Omega_{33}) \]

\[ y_3 \sim N(y_3, \sigma_y^2) \]

\[ \Theta_3 \sim N(y_3, \sigma_3^2) \]

\[ E\Theta_3 | y_3 = y_3 + (1 - \lambda) \mu \] (the estimate is “shrunk” towards \( \mu \).
The paper I linked (Smyth et al.) considers the more complicated case where $\beta$ is not known. We'll explore it more in HW II.

Final Note: Under the normal-normal Model, the marginal distribution of $Y_0$ is also normal:

$$p(Y_0) = \int p(Y_0 | o) p(o) do \sim N(\mu, \sigma^2 + \tau^2)$$

This idea of introducing overdispersion via hierarchical models was also used in DEseq where a Gamma-Poisson model is used (as negative binomial).

Another reason Bayes methods are attractive is that it provides a more satisfying view of testing where we can reason about

$$\frac{p(\beta > 0 | Y_0)}{p(\beta = 0 | Y_0)}$$

they comparing posterior probability distributions.

These are referred to as Bayes Factors, which correspond to the ratio of posterior to prior odds:

$$\frac{P(\beta > 0 | Y_0)}{P(\beta = 0 | Y_0)} = \frac{P(Y_0 | \beta > 0)}{P(Y_0 | \beta = 0)}$$

This is the Bayes Factor.