Differential Abundance Testing w/ Count Data

Our goal is to use count data obtained by RNA-seq to find genes where expression (or abundance) differ between two groups of interest (e.g., normal breast vs. breast tumor samples, or between samples at different stages in development).

The Data:
- \( n_A \) samples from group A (e.g., normal breast tissue)
- \( n_B \) samples from group B (e.g., breast tumor)
- \( K_{ij} \): # of reads aligned to gene \( i \) in sample \( j \)
- \( \beta(j) \): class indicator for sample \( j \) (equal to "A" or "B")

The Model:
We formulate a model for the observed counts \( K_{ij} \)

\[
\begin{align*}
\beta(j) & \xrightarrow{\text{which group?}} K_{ij} \\
& \xrightarrow{\text{The counts}} E K_{ij} = \mu_i \beta(j)
\end{align*}
\]

Here, we run into an issue since samples are not sequenced uniformly well. Some samples have more reads than others, so we need to normalize.
\[ E_{kij} = \hat{q}_{i \hat{p}(s_j)} s_j \]

\( \hat{q}_{i \hat{p}(s_j)} \): "concentration of fragments for gene \( i \), group \( \hat{p}(s_j) \)"

\( s_j \): sampling weight for sample \( j \)

Initially, that was the extent of modeling used for RNA-seq data: \( K_{ij} = \text{Poisson}(\hat{q}_{i \hat{p}(s_j)} s_j) \). This proved to be insufficient to capture variation in biological replicates (e.g., across normal breast tissue samples). Biological variation is essential to capture.

An alternative model is a negative binomial distribution:

\[ K_{ij} \sim \text{NB}(M_{ij}, \sigma_{ij}^2) \]

\( M_{ij} \) as above (mean)

\( \sigma_{ij}^2 = M_{ij} + s_j^2 \hat{p}(s_j) \) (variance)

The Poisson model has \( \sigma_{ij}^2 = M_{ij} \), we can think of NB as an overdispersed poisson. We can also obtain the NB distribution from a conditional (hierarchical model):

\[ M_{ij} \sim \text{random}! \]

See reference in paper for details but this is a useful trick to know to model data that varies more for \( \hat{p}(s_j) \) then a simple model can capture.
The estimates

$\hat{S}(j)$: sampling (size) rate factor:

$$\hat{S}(j) = \text{median} \left( \frac{K_{ij}}{\prod_{j=1}^{m} K_{ij}} \right)^{1/m} = \text{"pseudo-reference sample"}$$

$\hat{q}_{i:p}$: mean fraction for gene $i$ group $p$ ($j$)

$$\hat{q}_{i:p} = \frac{1}{m_p} \sum_{j: p(j) = p} \frac{K_{ij}}{S(j)}$$

$\hat{v}_p$: "raw variance"

$$\hat{v}_p = \omega_p (\hat{q}_{i:p}) - z_p$$

$$\omega_p = \frac{1}{m_p - 1} \sum_{j: p(j) = p} \left( \frac{K_{ij}}{S(j)} - \hat{q}_{i:p} \right)^2$$

$$z_p = \hat{q}_{i:p} \sum_{j: p(j) = p} \frac{1}{S(j)} \text{ (correction factor)}$$

$\omega_p (\hat{q}_{i:p})$: smooth function of mean used to "borrow" information across genes to estimate variance. Useful when sample sizes are small (another useful trick).

With estimates in hand we can calculate probability of observed data $K_{ij}$. 
Recall we want to find genes where counts differ between groups A & B. With a probability model in hand we can compare the data probability under two situations: (H₀) counts do not differ across groups, and (H₁) counts do differ. We'll see the framework postulated by Ronald Fisher to do this. This is likely the most commonly used procedure in science. Ever.

The recipe:

1. Compute a statistic $X$ that measures difference between groups ($\hat{\theta}$)
2. Assuming $H₀$ is true (counts do not differ) estimate the probability of observing a value $\hat{X}$ greater than or equal to the statistic ($\hat{\theta}$) (more extreme). This is the infamous $P$-value
3. Reject $H₀$ if the $p$-value is smaller than some pre-determined level $\alpha$

Some observations:

1. This rejects the null ($H₀$). It does not prove the alternative ($H₁$).
2. Requires a probability model under $H₀$. 

[Diagram: A bell curve with shaded area under $\hat{X}$]
Back to RNA-seq.

$H_0: \hat{\theta}_1 = \hat{\theta}_2$. So, compute probability of data with equal means.

P-value: Sum of the probabilities for all possible observations

$K_{iA}$: sum of counts for group A
$K_{iB}$: sum of counts for group B

with more extreme (smaller) probability:

$$\sum_{a,b} p(a, b)$$

$P(a, b) \leq p(K_{iA}, K_{iB})$

But, that isn’t quite right, we need to constrain $a$ and $b$ to be consistent with the data: $a + b = K_{iA} + K_{iB}$

The P-value:

$$P_i = \frac{\sum_{a,b} p(a, b)}{a + b} = \frac{p(K_{iA}, K_{iB})}{K_{iA} + K_{iB}}$$

Fisher’s answer: “Genes where $P_i \leq \alpha$ (e.g. 0.05 for example) are differentially abundant”. Not quite:

“I can’t say that genes $P_i \leq \alpha$ are not differentially abundant.”