Local False Discovery Rates

Multiple testing

- **Motivation**: High-throughput studies (e.g. genomics, fMRI) routinely test *thousands* of genes or features at once.
- **Per-test vs. global error**: If each hypothesis is tested at level α and the m tests are (approximately) independent,

$$\Pr(\text{at least one false positive}) = 1-(1-\alpha)^m \approx m\alpha \quad (\alpha \ll 1).$$

- **Need for global error measures**:
  - *Family-Wise Error Rate (FWER)*: controls the probability of *any* false positive but is often overly conservative in large-scale settings.
  - We seek an error metric that:
    1. maintains meaningful error control;
    2. does *not* scale with m;
    3. preserves higher power in large-scale settings.

False discovery rates (FDR)

- **Problem setting**
  - We simultaneously test m null hypotheses

$$H_{0,i}: \ \theta_i = 0 \quad \text{vs.} \quad H_{1,i}: \ \theta_i \neq 0, \qquad i = 1, \ldots, m,$$

producing test statistics (or p-values) $p_1, \ldots, p_m$.

  - Let

$$V = \#\{\text{false positives}\}, \qquad R = \#\{\text{total rejections}\}.$$

- **Definition**

$$\mathrm{FDR} \ = \ \mathbb{E}[\frac{V}{R}]$$

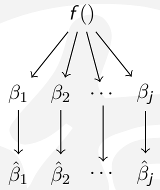
(set $V/R = 0$ when $R = 0$.)

- **Interpretation**: FDR is the expected fraction of reported discoveries that are actually null.

Empirical Bayes: Two Group Model

- **Two group model**:
  $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$
  - $f_0(z)$:density under the null.
  - $f_1(z)$ density under the alternative.
  - π_0: prior probability that an effect is null.
  - π_1: prior probability that an effect is non-null .($\pi_1 = 1 - \pi_0$).
- **Empirical Bayes idea:** Estimate the prior parameters (π_0, f_0, f_1) directly from the observed data, then use those point estimates to compute posterior quantities such as the local false discovery rates.

Effects Distribution



Deconvolution view

Local False Discovery Rate (lfdr)

- **Local false discover rates**

$$\mathrm{lfdr}(z) = P(\text{null} \mid Z = z) = \frac{P(Z = z \mid \text{null})\, P(\text{null})}{P(Z = z)} = \frac{\pi_0 f_0(z)}{f(z)}.$$

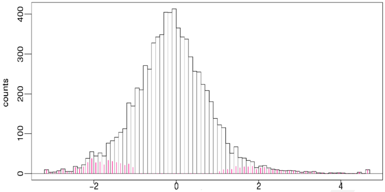
- **Empirical Bayes plug-in estimate**

$$\widehat{\mathrm{lfdr}}(z) = \frac{\widehat{\pi}_0 \widehat{f}_0(z)}{\widehat{f}(z)},$$

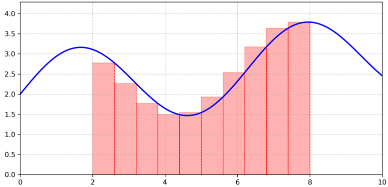
- **Controlling FDR with $\widehat{\mathrm{lfdr}}$**

  1. Sort $\widehat{\mathrm{lfdr}}_{(1)} \le \cdots \le \widehat{\mathrm{lfdr}}_{(m)}$;
  2. Find $k^* = \max\left\{k : \frac{1}{k}\sum_{i=1}^{k} \widehat{\mathrm{lfdr}}_{(i)} \le \alpha\right\}$;
  3. Reject k^* nulls.

Estimating $f(z)$:overview



$f(z)$



Riemann integral

$$\boxed{f(z) \xrightarrow{\text{binning}} \lambda_k = N \Delta f\big(z_{(k)}\big) \xrightarrow{\text{counts}} y_k \sim \mathrm{Poisson}(\lambda_k) \xrightarrow{\text{Poisson GLM}} \widehat{f}(z)}$$

Step 1: Thin-binned histogram of z-scores

- **Input data** : $z_1, \ldots, z_N$ —one z-score per test, assumed i.i.d. from the unknown marginal density $f(z)$.
- **Choose a small bin width** Δ (Efron uses $\Delta \approx 0.2$). Create contiguous, non-overlapping bins

$$B_k = [a_k,\ a_k + \Delta), \qquad k = 1, \ldots, K.$$

- **Bin centres** :

$$z_{(k)} = a_k + \frac{\Delta}{2}, \qquad k = 1, \ldots, K.$$

- **Count within each bin** :

$$y_k \ = \ \#\{\, i : z_i \in B_k \}, \qquad \sum_{k=1}^{K} y_k = N.$$

- **Dataset**: $\{(y_k, z_{(k)})\}$

Step 2: Binomial → Poisson Approximation

- **Binomial sampling model** Each z_i falls into exactly one thin bin B_k, so the count

$$y_k \sim \mathrm{Bin}(N,\ \pi_k), \qquad \pi_k = \Pr\{Z \in B_k\}.$$

- **Link to the unknown density**

$$\pi_k = \int_{B_k} f(z)\,dz \approx f(z_{(k)})\,\Delta \quad \text{for thin bins } (\Delta \text{ small}).$$

- **Poisson Approximation** When N is large and π_k is small (typical in high-throughput settings),

$$\mathrm{Bin}(N, \pi_k) \implies \mathrm{Pois}(\lambda_k), \qquad \lambda_k = N\,\pi_k \approx N\,\Delta\,f(z_{(k)}).$$

- **Implication for estimation** The $(z_{(k)}, y_k)$ pairs can now be viewed as Poisson responses with mean λ_k—exactly the structure needed to fit a Poisson generalized linear model in Step 3 and obtain the smooth estimator $\widehat{f}(z)$.

Step 3: Poisson GLM to obtain $\widehat{f}(z)$

- **Model for binned counts** From Step 2 we treat the histogram counts as

$$y_k \sim \mathrm{Pois}(\lambda_k), \qquad \lambda_k = N\,\Delta\, f\big(z_{(k)}\big).$$

- **Log–linear specification**

$$\log \lambda_k = \log(N\Delta) + \beta_1\, g_1\big(z_{(k)}\big) + \cdots + \beta_p\, g_p\big(z_{(k)}\big),$$

where $\log(N\Delta)$ is an *offset* and $g_j(\cdot)$ are user-chosen smooth functions.

- **Recovering the density**

$$\widehat{f}(z) = \exp\Big(\sum_{j=1}^{p} \widehat{\beta}_j\, g_j(z)\Big), \qquad \widehat{\beta}_0 = -\log\Big[\int \exp\big(\textstyle\sum_j \widehat{\beta}_j\, g_j(u)\big)\, du\Big]$$

Estimating the Empirical Null $f_0()$

- **Theoretical null (ideal case)**

$$f_0(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

- **Real-data departures** Correlation, hidden covariates or scale shifts can produce $Z \sim \mathcal{N}(\mu_0, \sigma_0^2)$,
- **Empirical null: two estimation strategies**
  1. *Central matching* —fit a quadratic to $\log \widehat{f}(z)$ over a central window (e.g. $|z| \leq 2$), then solve for $\hat{\mu}_0, \hat{\sigma}_0$.
  2. *MLE zero-assumption* —label all observations in $|z| \leq z_0$ (typ. $z_0 = 1$) as null and maximise $\prod \phi_{\mu_0,\sigma_0}(z_i)$ to obtain $\hat{\mu}_0, \hat{\sigma}_0$.

Step 4: Estimating π_0 (Central Matching)

- **Mostly–null window** : pick a central band (e.g. $|z| \le 2$) where $f(z) \approx \pi_0 f_0(z)$.
- **Quadratic fit** :

$$\log \widehat{f}(z) \;\approx\; a_0 + a_1 z + a_2 z^2, \qquad |z| \le 2.$$

- **Coefficient comparison** : match this expansion to $\log\big[\pi_0 f_0(z)\big] = \log \pi_0 + \log f_0(z)$ where $\log f_0(z)$ itself is quadratic.
  $\log f_0(z) = -\frac{1}{2\sigma_0^2} z^2 + \frac{\mu_0}{\sigma_0^2} z - \left(\frac{\mu_0^2}{2\sigma_0^2} + \frac{1}{2}\log 2\pi\sigma_0^2\right).$
  - $a_2 \to \sigma_0$
  - $a_1 \to \mu_0$
  - $a_0 \to \pi_0$ via $a_0 = \log\big[\pi_0 f_0(0)\big]$

  Thus all empirical-null parameters are identified by the fitted coefficients.