Assignment 05

PUBH 8878

## Setup

set.seed(8878)  
  
library(ggplot2)  
library(dplyr)  
library(tibble)

## Data for the assignment

We simulate test statistics using the two-groups model discussed in lecture, then convert to two-sided -values.

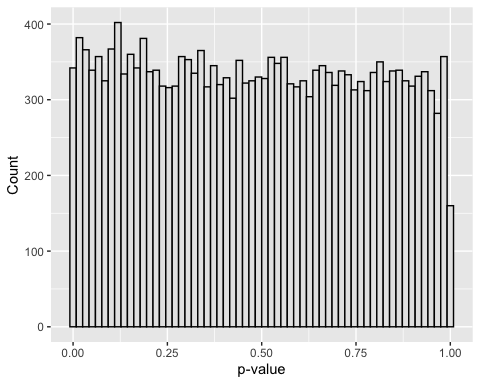
m <- 20000  
pi0\_true <- 0.95 # fraction of nulls  
sigma\_true <- 2.0 # sd under alternative  
tau\_true <- 1.0 # sd under null (standard normal null)  
  
H <- rbinom(m, 1, 1 - pi0\_true) # 1 = non-null, 0 = null  
z <- numeric(m)  
z[H == 0] <- rnorm(sum(H == 0), 0, tau\_true) # null z  
z[H == 1] <- rnorm(sum(H == 1), 0, sigma\_true) # alt z  
p <- 2 \* pnorm(-abs(z)) # two-sided p-values  
  
dat <- tibble(i = seq\_len(m), z = z, p = p, H = H)

A quick look:

dat |> summarize(  
 m = n(),  
 m0 = sum(H == 0),  
 m1 = sum(H == 1),  
 pi0 = mean(H == 0),  
 min\_p = min(p),  
 med\_p = median(p)  
) |>  
 kableExtra::kable(digits = 3)

| m | m0 | m1 | pi0 | min\_p | med\_p |
| --- | --- | --- | --- | --- | --- |
| 20000 | 19010 | 990 | 0.951 | 0 | 0.486 |

ggplot(dat, aes(p)) +  
 geom\_histogram(bins = 60, color = "black", fill = "grey90") +  
 labs(x = "p-value", y = "Count")



## Problem 1: Some theory (20 pts)

**(a)** (4 pts) Prove that if a continuous test statistic has null cdf , then the one-sided -value is **Uniform(0,1)** under the null. (Hint: use the probability integral transform.)

**(b)** Show that . Then, let . Prove the averaging identity . (Hint: write both numerator and denominator as integrals over A.)

**(c)** (3 pts) Briefly describe the key difference between *FDR control at level q* (e.g., BH) and reporting local false discovery rates (lfdr) for individual hypotheses.

## Problem 2: Implement BH step‑up from scratch (20 pts)

We will implement the BH decision rule and compare to built‑ins.

**(a)** (10 pts) Write an R function bh\_from\_scratch(p, q) that:

1. orders the input vector of p-values p
2. finds (take if the set is empty)
3. returns a list with $k, the BH threshold , and a logical vector reject of length m marking rejections.

Then run it at q = 0.10 on the vector dat$p. Produce a plot overlaying the ordered p\_(i) and the BH line , and mark the chosen cutoff.

**(b)** (5 pts) Compare your rejections to p.adjust(dat$p, method="BH") <= 0.10. They should match exactly. Report the number of discoveries.

**(c)** (5 pts) Report the *empirical FDP* on this simulated data, , using the latent truth H (remember: H==0 means null). Comment briefly.

## Problem 3: Simulation study of BH FDR control (20 pts)

Design a small simulation to assess how the BH FDR behaves as a function of and the alternative strength.

* Fix m = 5000, q = 0.10. For each and alternative sd :
  + simulate 200 independent datasets via the two-groups model with tau = 1,
  + apply BH at level q,
  + record the FDP for each replication using the latent truth.
* Plot the average FDP and its simulation SE versus for each . Does BH control FDR near under independence?

## Problem 4: Empirical-Bayes BFDR from p-values (20 pts)

We will estimate the Bayesian FDR at a threshold using

**(a)** (8 pts) Implement the Storey (2002) estimator

and report the smoothed estimate obtained by fitting a cubic spline or loess of versus and evaluating at

**(b)** (6 pts) For a grid of thresholds , compute and plot it as a function of .

**(c)** (6 pts) Pick the smallest whose estimated BFDR is and report how many discoveries you would make at that threshold. Compare to the BH discoveries at from Problem 2.

Compare your to the estimate from the qvalue package and report both.

## Problem 5: q-values and discovery sets (10 pts)

Compute qvalues <- qvalue::qvalue(dat$p) and:

**(a)** (4 pts) Report how many features have qvalues$qvalues <= 0.10. Compare to BH at and to your BFDR-based threshold in Problem 4.

**(b)** (6 pts) Sort features by their q-values (ascending). Let be the running mean of the first q-values. Plot versus and mark the largest with . Explain why selecting the first features is a reasonable discovery rule.

## Problem 6: Brief discussion (20 pts)

Write a concise paragraph (6–10 sentences) addressing the following:

* How do BH, BFDR thresholding, and q-value selection compare on these data in terms of number of discoveries and estimated error rates?
* What assumptions underlie BH control, and how might LD (dependence among tests) in GWAS affect it?
* What are potential pitfalls of estimating from the empirical distribution of -values?