

Unsupervised Learning: High-Dimensional Hypothesis Testing











Hypothesis Testing

- A different topic from previous lectures on **unsupervised** learning (clustering, dimension reduction, graphical models)
- Useful for many statistical problems for **rigorous** scientific discoveries:
How can I know if my discoveries arise from random noise?
- Main example in this lecture: feature importance testing; useful for feature screening
- It is a seemingly much simpler topic
- But hypothesis testing in **high dimensions** has its challenges!

Hypothesis Testing

- High-dimensional hypothesis testing
 - ▶ challenges of high-dimensional hypothesis testing
 - ▶ controlling family-wise error rate (FWER)
 - ▶ controlling false discovery rate (FDR)

Hypothesis Testing

<p>Good or Bad Metabolizer?</p> <p>Level of Protein 1?</p> <p>Level of Protein 2?</p> <p>Level of Protein M?</p>										
	Good	Good	Bad	Bad	Good	Bad	Good	Bad	Good	Bad
	2	1	5	11	17	8	2	0	56	1
	4	8	10	34	3	1	0	1	76	1

	5	6	2	15	23	11	3	6	8	9

Hypothesis Testing

- We have M features, each of which is measured in n observations.
- (In this lecture only, M is number of features, rather than p .)
- We can also have multiple responses, e.g.,
 - ▶ blood pressure
 - ▶ tumor size
 - ▶ survival time
 - ▶ cancer subtype
- We wish to test the null hypothesis

H_{0j} : j th feature is not associated with the response

for $j = 1, \dots, M$.

Testing One Hypothesis

Suppose we had only one hypothesis to test:

H_0 : feature is not associated with the response.

Type of test depends on types of responses:

- binary response (e.g. cancer versus normal): **two-sample t-test**
- categorical response (e.g. cancer type 1 versus cancer type 2 versus cancer type 3): **F-statistic for one-way ANOVA**
- survival response (e.g. time to recurrence): **score statistic for Cox proportional hazards model**
- quantitative response (e.g. blood pressure): **t-statistic for regression coefficient**

We always get a **test statistic** and a **p-value** for the association between the feature and the response.

Quick Review: Test Statistics & P-Values

- Suppose we have measured the expression level of gene A in n patients, and want to test its association with the response, cancer versus normal.
- We compute a two-sample t-statistic quantifying its association with the response.
- Large (absolute) t-statistic: strong association with response.
- How big is big?

Quick Review: Test Statistics & P-Values

- The **p-value** is the probability of observing a t-statistic at least this large, **under the null hypothesis of no association between the feature and the response**.
- A p-value of 0.02 means that the probability of seeing such a strong association between the response and the feature by chance, under the null hypothesis, is 0.02.
- **A small p-value** indicates strong evidence for association: i.e. **reject the null hypothesis**.

Type I Error

- Type I error occurs when we reject the null hypothesis when the null hypothesis actually holds.
- When we reject the null hypothesis if a p-value is below 0.05, we are **controlling Type I Error at level $\alpha = 0.05$** .
- We are not required to control Type I error at level $\alpha = 0.05$. A smaller α leads to a more conservative approach.
- Physicists control Type I error at around $\alpha = 5 \times 10^{-7}$. (Then again, atoms are cheaper than patients and mice.)

Multiple Testing

- Now, we test the association of each of $M > 1$ features with the response
- Suppose we compute a p-value for the association between each feature and the response, and reject each null at level α .
- The probability of falsely rejecting the j th null hypothesis is α .
- The probability of falsely rejecting at least one of M null hypotheses can be much greater than α !

Multiple Testing

- Let $\alpha = 0.05$, and assume that the tests are all independent.
- The probability of falsely rejecting at least one null hypothesis is $1 - (1 - \alpha)^M$:
 - ▶ 0.05, when $M = 1$;
 - ▶ 0.0975, when $M = 2$;
 - ▶ 0.40, when $M = 10$;
 - ▶ 0.9999649, when $M = 200$.
- In other words, we will erroneously reject the null hypothesis with higher probability if we test **a lot of hypotheses** and reject the null hypothesis with the usual threshold, like 0.05.

Hypothesis Testing

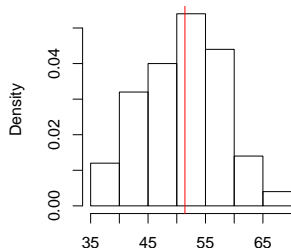
We will get a lot of false positives if we use 0.05 as a cut-off for rejecting the null hypothesis.

Protein	T-stat	P-Value
1	3.2	0.00137
2	-1.8	0.0718
3	5.8	6e-9
4	13.2	0
5	1.4	0.1615
6	-0.2	0.8414
.	.	.
.	.	.
.	.	.
M	4	6e-5

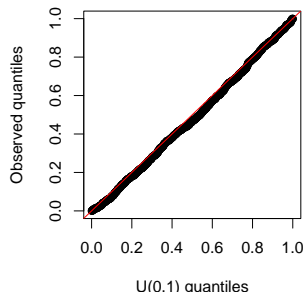
A Simple Illustration

- Consider 1000 hypotheses & 50 samples in two groups (25 treatment, 25 control): e.g. mRNA expression differences between treatment and control for 1000 genes
- Suppose all features from both groups are sampled from $N(0, 1)$:
nothing significant!
- 10 smallest p-values:**
0.0005, 0.0069, 0.0076, 0.0087, 0.0123, 0.0124, 0.0132, 0.0160, 0.0168, 0.0176
- All of these will be rejected **by chance** at the level of 5%!!

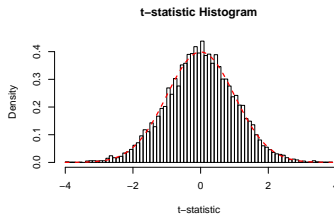
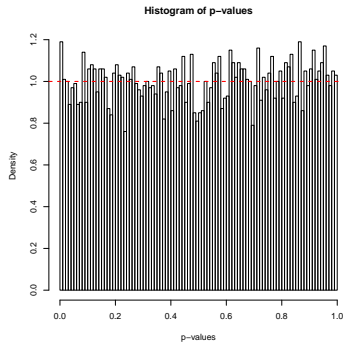
no of rejected hypotheses



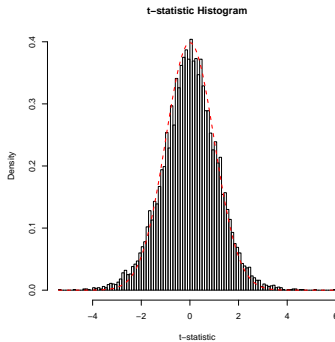
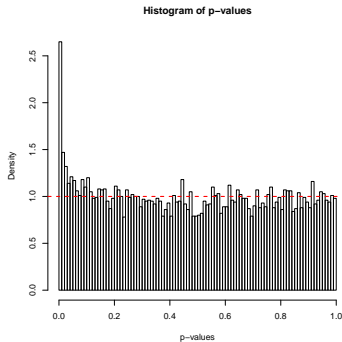
QQ plot of observed p-value



Example 1: All Null Hypotheses Hold



Example 2: Not All Null Hypotheses Hold



Bonferroni Correction for Family-wise Error Rate Control

- Family-wise error rate (overall type I error) control: want the probability of falsely rejecting any hypothesis to be controlled
- Simple solution: A Bonferroni correction.
- Rather than rejecting the j th null hypothesis if its p-value is less than 0.05, we use the threshold $0.05/M$.
- More generally, to control the overall Type I error (probability of falsely rejecting the null hypothesis) at level α , we must reject the j th null hypothesis if its p-value is below α/M .

Why does the Bonferroni correction control Type I Error?

- Let A denote the event that at least one null hypothesis is falsely rejected, and let A_j be the event that the j th null hypothesis is falsely rejected. So

$$P(A) = P\left(\bigcup_{j=1}^M A_j\right) \leq \sum_{j=1}^M P(A_j).$$

This means that if we keep $P(A_j)$ below α/M , then $P(A)$ will be below α .

- And keeping $P(A_j)$ below α/M means that we reject the j th null hypothesis if the p-value is below α/M .

Bonferroni Bottom Line

- To summarize, to control the overall Type I error at level α , we reject the j th null hypothesis if its p-value is below α/M .
- So if we are testing each SNP for association with a phenotype in a GWAS with $M = 10^6$ SNPs, then we reject the j th null hypothesis if its p-value is below 0.05×10^{-6} .
- This is **very conservative**! We will hardly ever reject the null hypothesis.
- Slightly less conservative alternatives to Bonferroni, that also control overall Type I error, are also available (Holm (1979), Hochberg (1988), Hommel (1988)).
- Holm's method dominates Bonferroni under arbitrary assumptions.

A Less Conservative Notion of Error Control

- The Bonferroni correction allows us to be confident that we will **almost never** falsely reject the null hypothesis.
- But in the analysis of omics data, this approach can be a little too conservative.
- If an investigator spends a lot of money measuring DNA methylation levels at millions of sites, he or she might be willing to sometimes falsely reject the null hypothesis, in exchange for correctly rejecting the null hypothesis more frequently.
- In other words: willing to have some false positives, in exchange for more true positives.
- A new notion of error control: **false discovery rate**.
- Proposed in 1995, but still a very active area of research!

Possible Outcomes from M Hypothesis Tests

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- V is number of false positives.
- T is number of false negatives.
- The Bonferroni correction guarantees that $P(V \geq 1) \leq \alpha$.

False Discovery Rate

$$FDR = E \left(\frac{\text{number of null hypotheses falsely rejected}}{\text{number of null hypotheses rejected}} \right).$$

For instance, in a gene expression experiment,

$$FDR = E \left(\frac{\text{number of genes incorrectly declared significant}}{\text{number of genes declared significant}} \right).$$

In other words, this is the **fraction of discoveries that are false positives**.

False Discovery Rate

	Called Not Significant	Called Significant	Total
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M - R$	R	M

- FDR: V/R , which is random... we often **control $E(V/R)$** .
- By controlling the FDR, we are guaranteeing that **we don't have too many false discoveries**.
- If we use an FDR threshold of 0.2, then no more than 20% of the null hypotheses that we reject were actually true. (Unfortunately, we don't know which ones!)
- In a GWAS, if we reject the null hypothesis of no association for SNPs with $FDR \leq 0.2$, then we expect no more than 20% of those SNPs to be false positives.

Benjamini-Hochberg Algorithm for FDR Control

- 1 Fix the false discovery rate, α .
- 2 Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(M)}$ denote the ordered p-values for the M hypothesis tests.

- 3 Define

$$L = \max \left\{ k : p_{(k)} < \alpha \frac{k}{M} \right\}.$$

- 4 Then, reject the j th null hypothesis if $p_j \leq p_{(L)}$, the Benjamini-Hochberg rejection threshold.
- 5 In other words, find the maximum order statistic (k) such that

$$\frac{M \times p_{(k)}}{k} \leq \alpha$$

Then reject all tests j for which $p_j \leq p_{(k)}$

FDR vs Bonferroni Control

- Benjamini & Hochberg:

- ▶ Find the maximum order statistic (k) such that

$$p(k) \leq \frac{\alpha k}{M}$$

- ▶ Reject all tests j with $p_j < p(k)$.
- ▶ We can also get a **q-value**, which is like a p-value for that test, but quantifies the **FDR** if any tests with smaller p-values are rejected.
- ▶ `p.adjust(ps, method="BH")`

- Bonferroni:

- ▶ Reject test j if

$$p_j \leq \frac{\alpha}{M}$$

- ▶ `p.adjust(ps, method="bonferroni")`

Example in R: All Null Hypotheses Hold

```
x <- matrix(rnorm(1000*50),ncol=50)
y <- sample(c(0,1),50,rep=TRUE)
ps <- NULL
for(i in 1:1000) ps <- c(ps,
  t.test(x[i,y==0],x[i,y==1])$p.value)
cat("Around 5% of p-values are below 0.05:",
mean(ps<.05),fill=TRUE)
fdrs.bonf <- p.adjust(ps, method="bonferroni")
plot(fdrs.bonf)
fdrs.bh <- p.adjust(ps, method="BH")
plot(ps,fdrs.bh)
plot(fdrs.bh)
```

Example in R: Not All Null Hypotheses Hold

```
x <- matrix(rnorm(1000*50),ncol=50)
y <- sample(c(0,1),50,rep=TRUE)
x[1:100,y==0] <- x[1:100,y==0] + 1
ps <- NULL
for(i in 1:1000) ps <- c(ps,
  t.test(x[i,y==0],x[i,y==1])$p.value)
cat("Way more than 5% of p-values are below 0.05:",
  mean(ps<.05),fill=TRUE)
```

Example, Continued

```
fdrs.bonf <- p.adjust(ps, method="bonferroni")
plot(fdrs.bonf)
cat("Significant features after Bonferroni correction:",
    which(fdrs.bonf < 0.05))
fdrs.bh <- p.adjust(ps, method="BH")
plot(ps,fdrs.bh)
plot(fdrs.bh)
cat("Significant features after BH-correction:",
    which(fdrs.bh < 0.05))
```

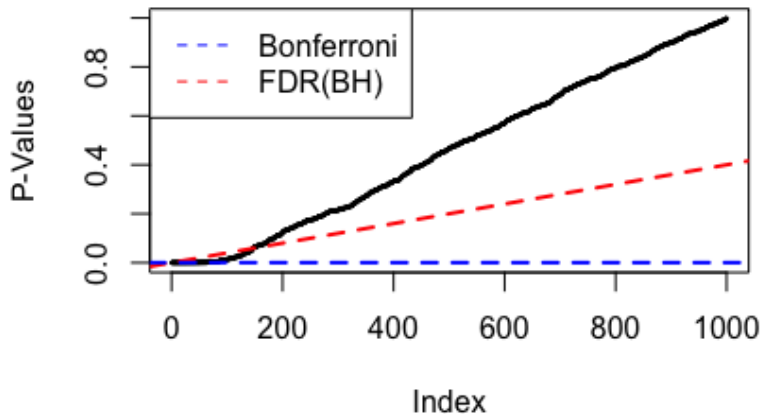
Example, Continued

Visualize the two procedures with $\alpha = 0.4$:

```
plot(sort(ps,decreasing=FALSE), ylab="P-Values",cex=0.2)
abline(a=0, b=0.4/1000, col="red",lty="dashed",lwd=2)
abline(a=0.4/1000, b=0, col="blue",lty="dashed",lwd=2)
legend("topleft", legend = c("Bonferroni", "FDR(BH)"),
      lty = c("dashed", "dashed"), col = c("blue", "red"))
```

Simulated Example

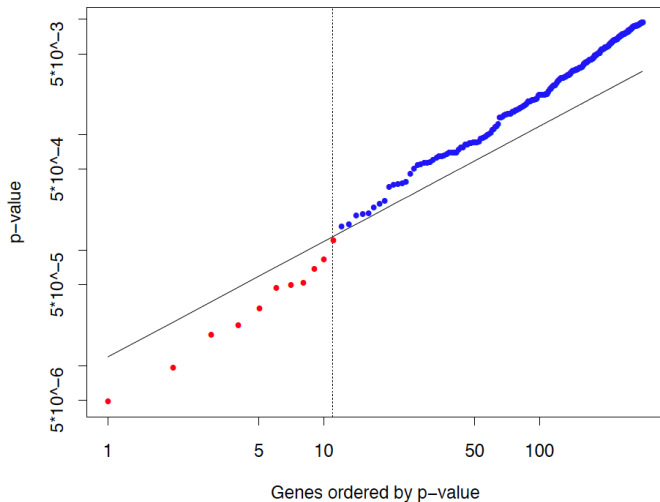
Suppose there are 1000 features, 50 samples (half control, half treatment).
100 features are associated with the response.



Example: Radiation Treatment Sensitivity

- Microarray data set on sensitivity of cancer patients to ionizing radiation treatment.
- Citation: Rieger, Hong et al, PNAS, 2004
- $M = 12,625$ genes, $n = 58$ samples: 44 samples with a normal reaction, 14 patients with severe reaction to radiation.
- Compute two-sample t-statistic for each gene's association with response to radiation.
- NO genes are declared significant after Bonferroni correction at level $\alpha = 0.05$.
- 11 genes had FDR below $\alpha = 0.15$.
- For each gene we can get a **q-value**, which is like a p-value for that gene, but quantifies the **FDR associated with that gene**.

Example: Radiation Treatment Sensitivity



Correlated Hypotheses

- So far, we have not discussed the effect of **correlation** in the data
- In real data examples, genes, proteins etc work together, and are correlated
- How would methods of multiple testing adjustment work if tests are correlated?

Correlated Hypotheses

- Bonferroni correction works **regardless of the correlation structure** (however, we know that in general, Bonferroni is not very appealing)
- Benjamini & Hochberg can handle **positive dependence**: think of this as all genes being positively correlated with each other
- However, in practice genes may have both positive (inducers) and negative (inhibitors) correlations with each other, so this assumption may not hold

Correlated Hypotheses

- When hypotheses have general dependence structures, controlling FDR is still an active area of research
 - ▶ For some specific models, BH procedure works asymptotically for *weakly correlated* tests
 - ▶ E.g., if testing for feature importance in high-dimensional linear regression, a variant of Benjamini & Hochberg procedure is valid as long as the *features follow a sparse graphical model* (Javanmard and Hamid, EJS, 2019)

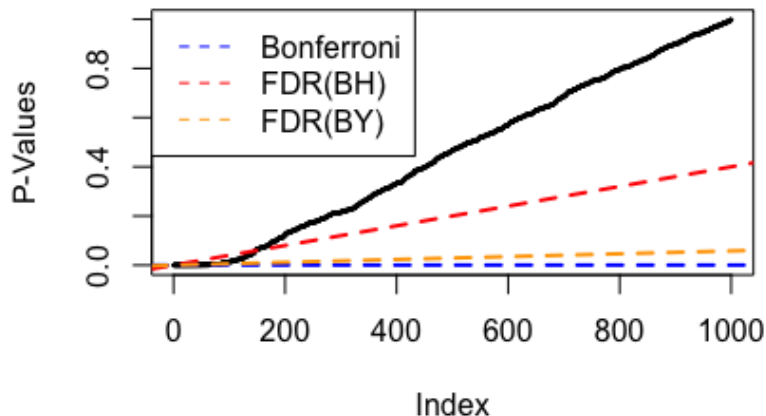
FDR Correction Under Dependence

- A control under dependence due to Benjamini & Yakutieli
 - ▶ Find the maximum order statistic (k) such that

$$p(k) \leq \frac{\alpha k}{M \times \left(\sum_{j=1}^M 1/j \right)}$$

- ▶ Reject all j with $p_j \leq p(k)$.
- $\sum_{j=1}^M 1/j \approx \log(M)$, so we pay an additional price of $\log(M)$, which makes this procedure **more conservative** than Benjamini & Hochberg (but then again, there is no free lunch!)
- In R, `p.adjust(ps, method='BY')`

Previous Simulated Example



Permutation Approach to FDR Estimation

- Another way to estimate FDRs: **by permutation**.
- Key idea: permute the responses across samples, thus breaking any relationship between the features and responses (**approximate the null distribution**)
- Easy and natural – unlike Benjamini-Hochberg, can explain it pretty simply to a non-statistician.
- **Does not require computing p-values**, which can be helpful in settings where p-values are (1) difficult to compute, or (2) unreliable.
- Also known as “plug-in estimate for FDR”.

Permutation Approach

- 1 Compute t_1, \dots, t_M , the test statistic for each of the M features.
- 2 Create K permutations of the responses, and for each permutation from $k = 1, \dots, K$, and for each feature $j = 1, \dots, M$, compute t_j^1, \dots, t_j^K .
- 3 For a range of values of the cut-point C , let

$$R_{obs} = \sum_{j=1}^M 1_{(|t_j| > C)},$$

and

$$\widehat{E(V)} = \frac{1}{K} \sum_{j=1}^M \sum_{k=1}^K 1_{(|t_j^k| > C)}.$$

- 4 Estimate the FDR by $\widehat{FDR} = \widehat{E(V)} / R_{obs}$.

Permutation Approach

- Based upon the approximation

$$E(V/R) \approx E(V)/E(R).$$

- We estimate $E(R)$ using R_{obs} – the actual number of test statistics that exceed the threshold.
- We estimate $E(V)$ by permuting the response and calculating the number of test statistics that exceed the threshold, **in the absence of any real relationship between the features and the response.**
- Actually, $\widehat{E(V)}$ estimates the number of false discoveries if all M null hypotheses hold, an **over-estimate** for $E(V)$ if not all null hold. Hence the FDR estimate is **conservative**.

At What Level to Control FDR?

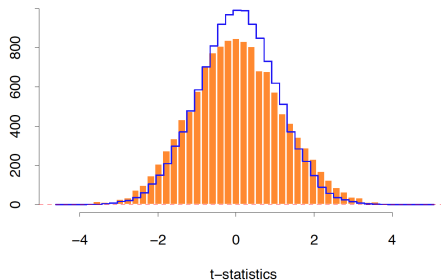
- In biology, we typically control Type I Error at level $\alpha = 0.05$ or $\alpha = 0.01$ (reject the null hypothesis if the p-value is less than 0.05 or 0.01).
- No analogous default cut-off for rejecting the null hypothesis using FDR control.
- We might want to follow up on genes whose FDR is below 10% or 20% in a gene expression experiment.
- The FDR threshold should depend on the data set, the number of genes with small FDRs, and the **number of rejected null hypotheses that we can afford to follow up on in the lab!**

P-Values Using Permutations

- We could also compute p-values using a permutation-type approach: the p-value for the j th feature is given by

$$p_j = \frac{1}{K} \sum_{k=1}^K 1_{(|t_j^k| > |t_j|)}.$$

- Example on radiation sensitivity data: orange shows true t-statistics, and blue shows t-statistics for permuted data.



Other Types of Hypothesis Testing

- We have focused on hypothesis testing on **feature-response associations**
- Can also test for
 - ▶ feature-feature associations (edges in graphical models)
 - ▶ clustering (whether two clusters are truly different from each other)
- Also faces similar challenges due to multiple testing when the number of tests, e.g., feature pairs, increases
- **General recipe:** generate p -values for each test and apply Bonferroni correction or Benjamini-Hochberg procedure

Multiple Testing & Selective Inference

Reminder: Valid inference requires pre-specifying hypotheses to test & only using data for this purpose. **Modern Data Analysis:** Explore data to determine which hypotheses to test.

- Ex: Perform feature selection (e.g. Lasso) then test selected features.
- Ex: Perform clustering then test which features are different across clusters.

Problem: Selective Inference!

Inference is incorrect (too liberal) yielding many false positives!

Selective Inference Approaches

① Data Splitting.

- ▶ Use training data for selection / data exploration & test data for inference. (Wasserman & Roeder, 2009)

② Post-Selection Inference.

- ▶ Conditions on selective model and computes proper null distributions. (Lee et al., 2016)

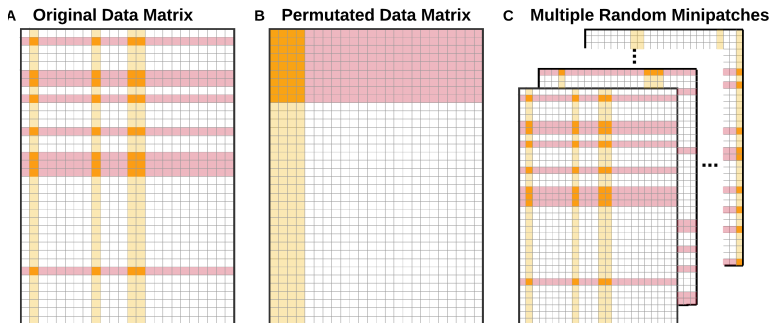
③ Knock-Off Inference.

- ▶ Creates fake copies of features (knock-offs!) and compares selection of real features to their fakes to determine FDR control. (Barber & Candès, 2015)

Research Highlight

Model-Agnostic Inference for Feature Importance

Problem: How do you test the importance of features for any machine learning model?



Solution: Fast Minipatch Ensembles yield valid model-agnostic feature importance inference. (Gan, Zheng, and Allen, 2022)

Pairwise Variable Screening

Problem: How do we screen the importance of features by incorporating the dependence among features? How can one identify spurious versus true correlation among features?

Solution: Study the **limiting behavior of the maximal absolute pairwise sample correlation** among covariates when they are independent Gaussian random variables, and use the extreme value results to identify covariates pairs that are potentially dependent and associated with a given response (Gong, Zhang, and Liu, 2021).

Research Highlight

Test for Edges in Graphical Models with Uneven Measurements

Problem: In real applications, measurements can be **uneven and irregular** across graphs, and hence **different parts of the graph can be estimated with different confidence**. Can we perform testing for all edges that account for different sample sizes over the graph? Can we ensure that most of the selected edges are true edges?

Solution: For each edge, compute its variance based on its **neighbors'** sample sizes; Compute a p -value for each edge and apply a variant of Benjamini-Hochberg procedure for FDR control (Zheng and Allen, 2022).