

The “Multiple Testing Problem” – What Is It and What Can We Do About It?

Susan E. Hodge
Lisa J. Strug

1

You’re studying large families (say, 6 children), and you get a family with 6 girls.

What would you think?

- Unusual family – father produces only X-bearing sperm? Rare event?
- Out of 64 families, *expect* one to have all girls.

Same evidence → different conclusions?

2

This illustrates essence of “multiple testing problem”: Situations in which –

Observations are same, but interpretations differ.

Depends on context.

3

Genetic Examples

- *Linkage analysis* – whether a disease locus is linked to a known marker locus.
- *Association analysis* – whether a disease phenotype is associated with alleles at a known marker locus.

Testing one marker vs. testing many.

4

Say we get a “significant” result (say, $p < .05$).

- If studied only one marker, that’s viewed as truly “significant.”
- If part of genome scan, viewed as “not significant.”

5

Again:

Observations are same, but interpretations differ.

Paradoxical.

6

Thinking about statistical evidence in a
different way

– the “evidential” paradigm –
can help resolve these paradoxes.

(But note: No magic bullet; no “free lunch.”)

7

Overview of Talk

- I. What is the “multiple testing problem”?
(We’ve seen some examples.)
- II. Evidential paradigm.
- III. Back to the original problem, and a
different way of thinking about it.

8

II. Evidential Paradigm

Decouples “measures of evidence” from “error probabilities.”

Define –

- A. ***Measure of evidence***
- B. Error probabilities
- C. Decoupling

9

A. Measure of Evidence

Define the “likelihood ratio” (LR) as our measure of evidence (as opposed to the p-value).

- What is the LR?
- Why does it make sense as a measure of evidence?

10

The likelihood ratio between two hypotheses is

$$LR \equiv \frac{L(\text{Hypothesis 1} | \text{data})}{L(\text{Hypothesis 0} | \text{data})} = \frac{L(H_1)}{L(H_0)}$$

“Likelihood” is proportional to “probability,” indicates which hypothesis is better supported by the observed data (i.e., under which hypothesis are observed data “less surprising”).

11

An Example

I have two grown daughters, April & Myriam. April loves classical music; Myriam prefers rock.

I come home from work one day, & through open window I hear classical music on kitchen radio.

Is April there (H_0)? or Myriam (H_1)?

12

Quantify this question:

Hypothesis	P(Class. hyp.)	P(Rock hyp.)
April (H_0)	0.90	0.10
Myriam (H_1)	0.05	0.95

Look at these probabilities... Classical music
not very surprising (not rare) if April is there...
quite surprising (quite rare) if Myriam is there...

In fact, it's 18[✎] less surprising or rare for April.

13

Given that I hear classical, we say the April
hypothesis is $.90/.05 = 18$ times “more likely”
than the Myriam one.

$$LR = \frac{L(\text{April})}{L(\text{Myriam})} = \frac{0.90}{0.05} = 18$$

14

Genetic Applications

Linkage analysis

(θ = recomb. frac. = measure of distance between loci)

H_0 : Disease is not linked to marker locus ($\theta = 0.5$)

H_1 : Disease is linked to marker locus (e.g., $\theta = 0.05$)

Association analysis

(OR = odds ratio)

H_0 : Disease not associated w/ marker allele (OR = 1.0)

H_1 : Disease associated w/ marker allele (e.g., OR = 2.0)

15

Criterion: We will choose a value of " k " ($k > 1$) and will agree that a $LR \geq k$ represents "strong" evidence in favor of H_1 , and that $LR \leq 1/k$ represents "strong" evidence in favor of H_0 .

Possible values of k could be 32, 100, 1,000...

(E.g., if we choose $k=32$, then $1/k \approx .03$.)

16

LR between those two values –
e.g., $.03 < k < 32$
represents “weak” evidence.

An advantage of Evidential paradigm.

17

II. Evidential Paradigm

Decouples “measures of evidence” from
“error probabilities.”

- A. Measure of evidence
- B. Error probabilities**
- C. Decoupling

18

B. Error Probabilities

How do we justify using the LR as our measure of evidence?

(Anybody could walk in off the street and propose an evidence measure – e.g., astrological signs!)

Justification: How does this measure of evidence *behave*?

A reliable measure of evidence has *low error probabilities*.

19

Say H_0 is true (e.g., no linkage, or no association).
Three possible outcomes –

- If $LR \geq k$, that's "misleading":
So $P[LR \geq k | H_0]$ = an error prob. – minimize this.
- If LR between $1/k$ and k , that's "weak" (inconclusive):
 $P[1/k < LR < k | H_0]$ – minimize this also.
- If $LR \leq 1/k$, that's leading to correct conclusion:
 $P[LR \leq 1/k | H_0]$ – maximize this.

20

Similarly, if H_1 is true (i.e., linkage or association, etc.), we consider same three outcomes, but interpretations are reversed:

- $P[LR \leq 1/k | H_1]$ is error probability;
- Also minimize $P[1/k < LR < k | H_1]$;
- Maximize $P[LR \geq k | H_1]$.

21

Analogous to classical paradigm

$P[LR \geq k | H_0]$ analogous to type I error;
 $P[LR \geq k | H_1]$ analogous to power.

But – we specify and use these error probabilities differently:

22

Classical Approach

Collect data, then interpret p -value as error probability *and* as evidence.

Example:

- Collect some data.
- Calculate $p < 0.0001$.
- Interpret this “probability” as a measure of evidence.

23

Evidential Approach

Don't discuss error probabilities *after* data collection (relevant *only* for planning).

Once the data are collected, the error probabilities become irrelevant --
(e.g., if weather report predicts rain with 50% probability...)
-- look only at evidence

24

How do error probabilities behave?

- Work by Royall and others analyzes magnitudes of error prob's when we set k at different values – also as function of sample size.
- Brief summary: Values of k as low as 32, can result in very low error probabilities for reasonable sample sizes.

25

A few examples (with $k=32$)

- *Absolute upper bound:*
Both error prob's $\leq 1/32 \approx .03$.
- *Asymptotic upper bound:*
Both $\leq \Phi[-\sqrt{(2\ln 32)}] \approx .0043$.

(These are *upper bounds*, not actual achieved values.)

26

An Example

Some linkage results, testing
 $H_0: \theta=0.5$ vs. $H_1: \theta = 0.05$, with $k=32$.

Error Probabilities

Sample	H_0 true	H_1 true
20 FIGs	.0013	.0003
20 SPs	.0013	.0018
20 fams.	.0000	--

FIG = fully informative gamete; SP = sib pair

27

Focus on Weak Evidence

Since it's "easy" to keep the probability of *misleading* evidence low,
focus on lowering the probability of *weak* evidence:

- Choose alternative hypothesis (H_1) carefully,
- Don't set k too high
- Increase sample size.

28

II. Evidential Paradigm

Decouples “measures of evidence” from
“error probabilities.”

- A. Measure of evidence
- B. Error probabilities
- C. *Decoupling***

29

C. Decoupling

We keep the “error probabilities” *separate*
from the “measure of evidence”:

- Design our experiment to have acceptable error probabilities.
- After collecting data, look *only* at measure of evidence.
- Do not try to make one quantity serve both functions.

30

Contrast with classical hypothesis testing, where p -value is used as both error prob. and measure of evidence:

- P -value is technically defined as, “prob. of observing result this deviant or more from H_0 , if H_0 is true.”
- But then also used as measure of evidence: p -value of .04 → “acceptable but not very strong evidence”; p -value of .0001 → “strong evidence”; etc.

31

There are many reasons why p -value isn't a good measure of evidence.

A whole separate talk.

32

An Example

H_0 : Coin is fair

H_1 : Coin has a head on both sides.

Toss 10 times, observe “heads” 9 times,
“tails” once.

How to analyze?

33

Under H_0 : Rare event, p -value is .011.

Under H_1 : *Impossible* event.

So do we reject H_0 ?

- Classical approach says Yes, at 5% level.
- Evidential approach says No, because no matter how rare under H_0 , it's much *more* rare under H_1 .

34

Summarize Evidential Paradigm:

Use the LR as a measure of evidence.

- Know that it leads to drawing wrong conclusions with low probability, i.e., error probabilities can be kept small.
- Separate (“decouple”) measure of evidence from error probabilities.

35

Overview

- I. What is the “multiple testing problem”? (We’ve seen some examples.)
- II. Evidential Paradigm.
 - A. Measure of evidence
 - B. Error probabilities
 - C. Decoupling
- III. *Back to the original problem, and a different way of thinking about it.***

36

III. Back to the Original Problem

Root of problem: Confounding “error probability” with “measure of evidence.”

Our claim: Advantage of “decoupling” – lets us consider multiple testing problem more logically and consistently.

37

Paradox of Classical Approach

“In a linkage analysis, a LR of, say, 1000:1 at locus *X* is taken to represent *different evidence*, depending on whether –

- we analyzed only locus *X*; or
- we analyzed 10 ‘candidate loci,’ including *X*; or
- we analyzed *X* as part of a genome scan...”

Violates common sense.

Classical approach conflates *evidence* with *error probability*.

38

Again:

Observations are same, but interpretations differ.

Depends on context.

39

Expressed in statistical terms:

Even though error in any one test is small, when we perform multiple tests, the probability that *at least one* of these tests has made an error may be large.

Define Family-Wise Error Rate:

$$\text{FWER} \equiv P[\geq 1 \text{ test yields } LR \geq k \mid H_0]$$

40

Classical: Argue about whether and how to adjust p -value to allow for multiple tests.

Evidential: Decouple!

- The *evidence* is what it is.
- If we have to adjust the *error probability*, we will.

41

“If nothing else” – improved clarity and precision of thought and of language.

But – we have gone further, broke down multiple testing issue into two separate situations and can say something quantitative about both.

42

Situation 1

Multiple Tests of Single Hypothesis

Examine data as we go along, then potentially collect more data, depending on what we've seen so far.

Sequential.

What researchers *do* in linkage analysis.

Conclusion: It is OK to keep collecting data until LR gives a clear result. There is still a reasonable upper bound on error prob's.

43

Situation 2

Single Test of Multiple Hypotheses

Genome scans.

Testing lots of candidate genes.

Conclusions:

- Use sample size to control error probabilities, as much as feasible.
- Use “replication” appropriately.

44

Replication is nothing new; many investigators have argued:

“Don’t get too hung up on the p -values; focus more on replication.”

This work puts that intuitive reaction on a sound logical footing, using the Evidential paradigm.

45

Summary

- Use the LR as a measure of evidence; has property that it leads a *low probability* of wrong conclusions.
- Separate (“decouple”) measure of evidence from error probabilities.
- This lets us deal with multiple testing problem more logically and consistently, by separating what we do with the evidence from what we do with the error probabilities.

46

Selected References

- Royall, R. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London 1994.
- Royall RM: On the probability of observing misleading statistical evidence (with discussion). *J Am Statist Assoc* 95: 760–780, 2000.
- Vieland, V.J. & Hodge, S.E. Review of *Statistical Evidence: A Likelihood Paradigm*, by R. Royall. *Am J Hum Genet* 63: 283–289, 1998.
- Strug, L.J & Hodge, S.E. An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling “error probabilities” from “measures of evidence.” *Hum Hered* 61: 166–188, 2006.
- Strug, L.J. & Hodge, S.E. An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. *Hum Hered* 61: 200–209, 2006.