To C or not to C

Conditioning in association tests

Zheng Gao

gaozheng@umich.edu

Table of contents

Section 1

Ancillarity

Ancillarity

- Ghosh, Reid, & Fraser (2010)[1]: "... statistics with distributions not depending on the model parameters."
- Little (1989)[2]: "let X and Y be random variables with joint distribution that factorizes in the form

$$p(x, y \mid \theta, \phi) = p(x \mid y, \theta) p(y \mid \phi),$$

then Y contains no information about θ and is called an ancillary statistic[3]. "

[1] Ghosh, Malay, N. Reid, and D. A. S. Fraser. "Ancillary statistics: A review." Statistica Sinica (2010): 1309-1332.
[2] Little, Roderick JA. "Testing the equality of two independent binomial proportions." The American Statistician 43.4 (1989): 283-288.
[3] Cox, D. R., and D. Hinkley. "Chapman and Hall." Theoretical Statistics (1974).

Ancillarity

- Ghosh, Reid, & Fraser (2010)[1]: "... statistics with distributions not depending on the model parameters."
- Little (1989)[2]: "let X and Y be random variables with joint distribution that factorizes in the form

$$p(x, y \mid \theta, \phi) = p(x \mid y, \theta) p(y \mid \phi),$$

then Y contains no information about θ and is called an ancillary statistic[3]. "
- Formally, the two disagree!

[1] Ghosh, Malay, N. Reid, and D. A. S. Fraser. "Ancillary statistics: A review." Statistica Sinica (2010): 1309-1332.
[2] Little, Roderick JA. "Testing the equality of two independent binomial proportions." The American Statistician 43.4 (1989): 283-288.
[3] Cox, D. R., and D. Hinkley. "Chapman and Hall." Theoretical Statistics (1974).

Ancillarity

- Ghosh, Reid, & Fraser (2010)[1]: "... statistics with distributions not depending on the model parameters."
- Little (1989)[2]: "let X and Y be random variables with joint distribution that factorizes in the form

$$p(x, y \mid \theta, \phi) = p(x \mid y, \theta) p(y \mid \phi),$$

then Y contains no information about θ and is called an ancillary statistic[3]. "
- Formally, the two disagree!
- the latter, i.e., "statistics with distributions not depending on the model parameters of interest" is used.

[1] Ghosh, Malay, N. Reid, and D. A. S. Fraser. "Ancillary statistics: A review." Statistica Sinica (2010): 1309-1332.
[2] Little, Roderick JA. "Testing the equality of two independent binomial proportions." The American Statistician 43.4 (1989): 283-288.
[3] Cox, D. R., and D. Hinkley. "Chapman and Hall." Theoretical Statistics (1974).

Examples of ancillary statistics

- The baseball example
- The horticulturist example
- A regression example
- The 2x2 table!

Examples of ancillary statistics: baseball batting

- Observer tries to determine batter's ability by
- ... observing $N \sim \text{Poi}(\lambda)$ number of at-bats,
- ... record the number of hits $X \sim \text{Binom}(p, N)$.

Examples of ancillary statistics: baseball batting

- Observer tries to determine batter's ability by
- ... observing $N \sim \text{Poi}(\lambda)$ number of at-bats,
- ... record the number of hits $X \sim \text{Binom}(p, N)$.

In this case,

- N is the ancillary statistic since
- ... its distribution does not depend on p,
- ... although it does provide information on the accuracy of $\hat{p}$.

Examples of ancillary statistics: the horticulturist

- Observer tries to determine the probability of red flowers by
- ... observing $N \sim \text{Binom}(\phi, 4)$ plants which has flowered,
- ... record the number of red flowers $X \sim \text{Binom}(p, N)$.

Examples of ancillary statistics: the horticulturist

- Observer tries to determine the probability of red flowers by
- ... observing $N \sim \text{Binom}(\phi, 4)$ plants which has flowered,
- ... record the number of red flowers $X \sim \text{Binom}(p, N)$.

In this case,

- N is, again, the ancillary statistic since
- ... its distribution does not depend on p,
- ... although it, again, provides information on the accuracy of $\hat{p}$.

© Zheng Gao

Examples of ancillary statistics: regression

- Determine β with n observations from the model

$$Y \sim F, \quad (X|Y,\beta) \sim Y\beta + \epsilon.$$

  (reversed X and Y to match notations from before)
- ... OLS estimate

$$\begin{aligned}\widehat{\beta} &= (Y'Y)^{-1}Y'X = (Y'Y)^{-1}Y'(X\beta + \epsilon) \\ &= \beta + \frac{\sum_i y_i \epsilon_i}{\sum_i y_i^2} \stackrel{\mathrm{d}}{=} \mathrm{N}\left(\beta, \frac{1}{\sum_i y_i^2}\right).\end{aligned}$$

- ... How do you perform inference on β?

Examples of ancillary statistics: regression

- Determine β with n observations from the model

$$Y \sim F, \quad (X|Y,\beta) \sim Y\beta + \epsilon.$$

(reversed X and Y to match notations from before)

- ... OLS estimate

$$\begin{aligned}\widehat{\beta} &= (Y'Y)^{-1}Y'X = (Y'Y)^{-1}Y'(X\beta + \epsilon) \\ &= \beta + \frac{\sum_i y_i \epsilon_i}{\sum_i y_i^2} \overset{\mathrm{d}}{=} \mathrm{N}\left(\beta, \frac{1}{\sum_i y_i^2}\right).\end{aligned}$$

- ... How do you perform inference on β?

Most of us (I think!) would perform conditional inference, i.e., width of CI depends on Y.

- Y is ancillary since/if
- ... its distribution does not depend on β,
- ... Y provides information only on the accuracy of $\hat{\beta}$.

Although there was an argument for unconditional inference, if we interpret the relationship as only a linear approximation to the conditional expectations.

Examples of ancillary statistics: 2×2 tables

- Determine if there is an association ($\mathrm{OR} = \frac{\mu_{11}\mu_{22}}{\mu_{21}\mu_{12}} = 1$) using N (constant) observations from a multinomial model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \mathrm{Multinomial}(N, (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})).$$

$$\begin{array}{cc|c} n_{11} & n_{12} & n_1 \\ n_{21} & n_{22} & n_2 \\ \hline m_1 & m_2 & N \end{array}$$

Examples of ancillary statistics: 2×2 tables

- Determine if there is an association (OR $= \frac{\mu_{11}\mu_{22}}{\mu_{21}\mu_{12}} = 1$) using N (constant) observations from a multinomial model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})).$$

$$\begin{array}{cc|c} n_{11} & n_{12} & n_1 \\ n_{21} & n_{22} & n_2 \\ \hline m_1 & m_2 & N \end{array}$$

We now show that one of the marginals, say, (n_1, n_2), is ancillary.

- Re-parameterize

$$\phi = \mu_{11} + \mu_{12}, \quad p_1 = \frac{\mu_{11}}{\mu_{11} + \mu_{12}}, \quad p_2 = \frac{\mu_{21}}{\mu_{21} + \mu_{22}}.$$

- so that

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\phi p_1, \phi(1-p_1), (1-\phi)p_2, (1-\phi)(1-p_2))).$$

Examples of ancillary statistics: 2×2 tables

Denote the re-parameterized model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\phi, p_1, p_2)).$$

The likelihood function is

$$\begin{aligned} & p((n_{11}, n_{12}, n_{21}, n_{22})|(\phi, p_1, p_2)) \\ = & \binom{N}{n_{11}, n_{12}, n_{21}, n_{22}} (\phi p_1)^{n_{11}} (\phi(1-p_1))^{n_{12}} ((1-\phi)p_1)^{n_{21}} ((1-\phi)(1-p_1))^{n_{22}} \end{aligned}$$

Examples of ancillary statistics: 2×2 tables

Denote the re-parameterized model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\phi, p_1, p_2)).$$

The likelihood function is

$$\begin{aligned}
& p((n_{11}, n_{12}, n_{21}, n_{22}) | (\phi, p_1, p_2)) \\
= & \binom{N}{n_{11}, n_{12}, n_{21}, n_{22}} (\phi p_1)^{n_{11}} (\phi(1 - p_1))^{n_{12}} ((1 - \phi)p_1)^{n_{21}} ((1 - \phi)(1 - p_1))^{n_{22}} \\
= & \binom{N}{n_{11}, n_{12}, n_{21}, n_{22}} \times (p_1)^{n_{11}} (1 - p_1)^{n_1 - n_{11}} p_1^{n_{21}} (1 - p_1)^{n_2 - n_{21}} \times \phi^{n_1} (1 - \phi)^{n_2}
\end{aligned}$$

Examples of ancillary statistics: 2×2 tables

Denote the re-parameterized model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\phi, p_1, p_2)).$$

The likelihood function is

$$\begin{aligned}
&p((n_{11}, n_{12}, n_{21}, n_{22})|(\phi, p_1, p_2)) \\
=&\binom{N}{n_{11}, n_{12}, n_{21}, n_{22}}(\phi p_1)^{n_{11}}(\phi(1-p_1))^{n_{12}}((1-\phi)p_1)^{n_{21}}((1-\phi)(1-p_1))^{n_{22}} \\
=&\binom{N}{n_{11}, n_{12}, n_{21}, n_{22}} \times (p_1)^{n_{11}}(1-p_1)^{n_1-n_{11}}p_1^{n_{21}}(1-p_1)^{n_2-n_{21}} \times \phi^{n_1}(1-\phi)^{n_2} \\
=&C(x,y) \times p(\underbrace{(n_{11}, n_{21})}_{x} \,|\, \underbrace{(n_1, n_2)}_{y}, \underbrace{(p_1, p_2)}_{\theta}) \times p(\underbrace{(n_1, n_2)}_{y} \,|\phi)
\end{aligned}$$

Examples of ancillary statistics: 2×2 tables

Denote the re-parameterized model

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{Multinomial}(N, (\phi, p_1, p_2)).$$

The likelihood function is

$$\begin{aligned}
&p((n_{11}, n_{12}, n_{21}, n_{22})|(\phi, p_1, p_2)) \\
=&\binom{N}{n_{11}, n_{12}, n_{21}, n_{22}}(\phi p_1)^{n_{11}}(\phi(1-p_1))^{n_{12}}((1-\phi)p_1)^{n_{21}}((1-\phi)(1-p_1))^{n_{22}} \\
=&\binom{N}{n_{11}, n_{12}, n_{21}, n_{22}} \times (p_1)^{n_{11}}(1-p_1)^{n_1-n_{11}} p_1^{n_{21}}(1-p_1)^{n_2-n_{21}} \times \phi^{n_1}(1-\phi)^{n_2} \\
=&C(x,y) \times p(\underbrace{(n_{11}, n_{21})}_{x} \,|\, \underbrace{(n_1, n_2)}_{y}, \underbrace{(p_1, p_2)}_{\theta}) \times p(\underbrace{(n_1, n_2)}_{y} \,|\phi)
\end{aligned}$$

Recall the definition of ancillarity...

- "let X and Y be random variables with joint distribution that factorizes in the form
$$p(x, y \mid \theta, \phi) = p(x \mid y, \theta)p(y \mid \phi),$$
then Y contains no information about θ and is called an ancillary statistic. "
- ... therefore, $\mathbf{n_1}$ **and** $\mathbf{n_2}$ **are ancillary** (for any functionals of (p_1, p_2))!

© Zheng Gao

Section 2

Why ancillarity? (spoiler: conditionality principle)

Conditional inference: What is conditional inference?

Conditionality principle (Birnbaum 1962): When the experiment E can be described as a mixture of several component experiments E_y where y is an ancillary statistic, inference (about the parameter) in the following two situations should be the same:

- Observing (x, y).
- Observing x from the component experiment E_y.

Conditional inference: What is conditional inference?

Conditionality principle (Birnbaum 1962): When the experiment E can be described as a mixture of several component experiments E_y where y is an ancillary statistic, inference (about the parameter) in the following two situations should be the same:

- Observing (x, y).
- Observing x from the component experiment E_y.

In other words:

- Whatever experiment that didn't happen doesn't count.
- We only care about the conditional distribution $p(x|y)$.

© Zheng Gao

Conditional inference: What is conditional inference?

Conditionality principle (Birnbaum 1962): When the experiment E can be described as a mixture of several component experiments E_y where y is an ancillary statistic, inference (about the parameter) in the following two situations should be the same:

- Observing (x, y).
- Observing x from the component experiment E_y.

In other words:

- Whatever experiment that didn't happen doesn't count.
- We only care about the conditional distribution $p(x|y)$.

Example:

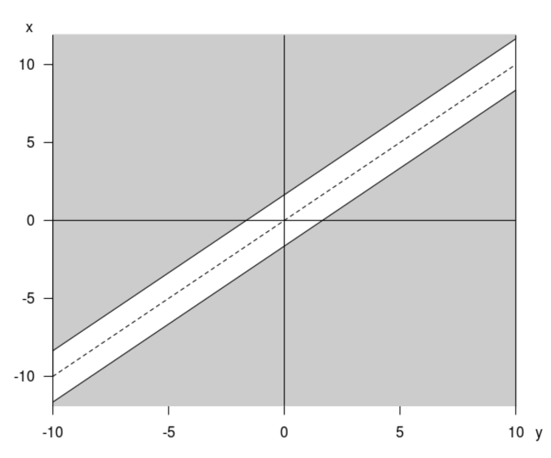
- Testing for $\beta = 1$ in the regression example with just 1 sample

$$Y \sim \mathrm{N}(0, 3), \quad (X|Y, \beta) \sim \mathrm{N}(Y\beta, 1).$$

- Marginally,

$$X \sim \mathrm{N}(0, 4).$$

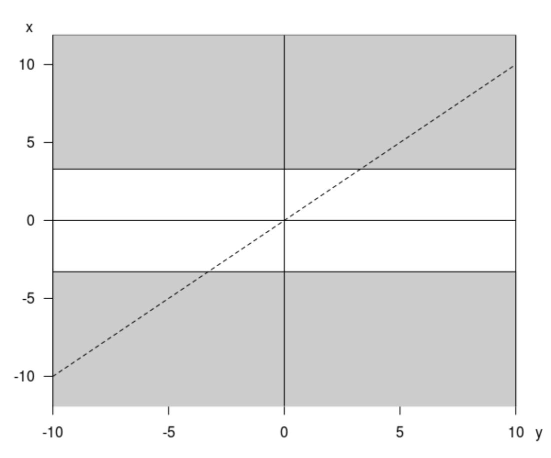
Conditional inference: example



Rejection region based on

- Cond. dist. $p(x|y, \beta)$.

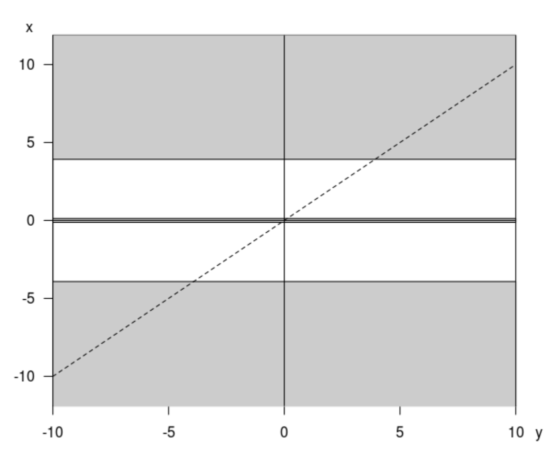
Conditional inference: example



Rejection region based on

- Cond. dist. $p(x|y, \beta)$.
- Marginal dist. $p(x|\beta)$.

Conditional inference: example



Rejection region based on

- Cond. dist. $p(x|y,\beta)$.
- Marginal dist. $p(x|\beta)$.
- Cond. dist. $p(x|z,\beta)$, where $Z = \mathbb{1}[X < 0]$.

© Zheng Gao

Conditional inference

- All procedures have calibrated levels, marginally. That is,

$$\mathbb{P}[\text{rejection} \mid \beta = 1] = \alpha.$$

However,

[4] Fraser, Donald AS. "Ancillaries and conditional inference." Statistical Science 19.2 (2004): 333-369.

© Zheng Gao

Conditional inference

- All procedures have calibrated levels, marginally. That is,

$$\mathbb{P}[\text{rejection} \mid \beta = 1] = \alpha.$$

  However,

- Cond. on ancillary statistics seem to yield more "reasonable" procedures.
- Not all conditioning is good, as the third example clearly demonstrates.

[4] Fraser, Donald AS. "Ancillaries and conditional inference." Statistical Science 19.2 (2004): 333-369.

Conditional inference

- All procedures have calibrated levels, marginally. That is,

$$\mathbb{P}[\text{rejection} \mid \beta = 1] = \alpha.$$

  However,
- Cond. on ancillary statistics seem to yield more "reasonable" procedures.
- Not all conditioning is good, as the third example clearly demonstrates.

Question

- What does it mean to be more "reasonable"??

[4] Fraser, Donald AS. "Ancillaries and conditional inference." Statistical Science 19.2 (2004): 333-369.

Conditional inference

- All procedures have calibrated levels, marginally. That is,

$$\mathbb{P}[\text{rejection} \mid \beta = 1] = \alpha.$$

  However,
- Cond. on ancillary statistics seem to yield more "reasonable" procedures.
- Not all conditioning is good, as the third example clearly demonstrates.

Question

- What does it mean to be more "reasonable"??

Answer

- I don't know...

[4] Fraser, Donald AS. "Ancillaries and conditional inference." Statistical Science 19.2 (2004): 333-369.

Conditional inference

- All procedures have calibrated levels, marginally. That is,

$$\mathbb{P}[\text{rejection} \mid \beta = 1] = \alpha.$$

However,

- Cond. on ancillary statistics seem to yield more “reasonable” procedures.
- Not all conditioning is good, as the third example clearly demonstrates.

Question

- What does it mean to be more “reasonable”??

Answer

- I don’t know...

One possible explanation (see also, Fraser (2004)[4]):

- Robustness against model misspecification: even when we get the distribution of y wrong, the test can still be used.

Still, conditionality principle is a principle, not an explanation, not a theorem.

[4] Fraser, Donald AS. "Ancillaries and conditional inference." Statistical Science 19.2 (2004): 333-369.

Section 3

Association tests in 2x2 tables

Examples of ancillary statistics: 2×2 tables

n_1 and n_2 are ancillary (in the multinomial model).

- The same is (trivially) true for product binomial model.
- ... and the hypergeometric model.

The C principle — should you choose to accept it — says that we should condition on one of the marginals.

[5] Little, Roderick JA. "Testing the equality of two independent binomial proportions." The American Statistician 43.4 (1989): 283-288.

Examples of ancillary statistics: 2×2 tables

n_1 and n_2 are ancillary (in the multinomial model).

- The same is (trivially) true for product binomial model.
- ... and the hypergeometric model.

The C principle — should you choose to accept it — says that we should condition on one of the marginals.

However, conditioning a both margins may still be controversial, since

- (n_1, n_2) and (m_1, m_2) are not jointly ancillary!
- (m_1, m_2) is only approximately ancillary,
- ... i.e., "carries little information about the OR"[5] (whatever that means, statements are vague, though quantifiable.).

[5] Little, Roderick JA. "Testing the equality of two independent binomial proportions." The American Statistician 43.4 (1989): 283-288.

Two (three, four) schools of thought

1. **Condition on one margin, or none!** — Pearson's chi-square, Barnard's CSM, Yule's, Student, Welch's t-tests, etc.
2. **Condition on two margins** — Fisher's exact test (approx. by Yates)
3. ... the dark side (topic for another day: likelihood principle, Bayesianism).

Section 4

Conservativeness of Fisher's Exact test?

Why is Fisher's exact test "conservative" then?

Discreteness.

- The data was discrete to start with.
- Exacerbated by conditioning.

Why is Fisher's exact test "conservative" then?

TABLE 2

Relative frequencies of the 36 2 × 2 tables generated by samples from two binomial distributions, $n_1 = n_2 = 5$, $\mathbf{p} = 1/2$, classified by values of the m_1, m_2 margin

	m_1, m_2 margin												
$p_1 - p_2$	10, 0	9, 1	8, 2	7, 3	6, 4	5, 5	4, 6	3, 7	2, 8	1, 9	0, 10	*Total*	*Overall probability*
−1.0						1 (0.004)						1	0.001
−0.8					5 (0.024)		5 (0.024)					10	0.010
−0.6				10 (0.083)		25 (0.099)		10 (0.083)				45	0.044
−0.4			10 (0.222)		50 (0.238)		50 (0.238)		10 (0.222)			120	0.117
−0.2		5 (0.5)		50 (0.417)		100 (0.397)		50 (0.417)		5 (0.5)		210	0.205
0.0	1 (1.0)		25 (0.556)		100 (0.476)		100 (0.476)		25 (0.556)		1 (1.0)	252	0.246
+0.2		5 (0.5)		50 (0.417)		100 (0.397)		50 (0.417)		5 (0.5)		210	0.205
+0.4			10 (0.222)		50 (0.238)		50 (0.238)		10 (0.222)			120	0.117
+0.6				10 (0.083)		25 (0.099)		10 (0.083)				45	0.044
+0.8					5 (0.024)		5 (0.024)					10	0.010
+1.0						1 (0.004)						1	0.001
Total	1	10	45	120	210	252	210	120	45	10	1	1024	1.000

The first column contains the single table (5, 0; 5, 0), the second the two tables (4, 1; 5, 0), (5, 0; 4, 1), etc. The figures in parentheses are the elements of the hypergeometric distribution for given values of the m_1, m_2 margin.

Section 5

Should we care?

Asymptotic equivalence

"Well-known" asymptotic "equivalence" of these tests

Asymptotic equivalence

"Well-known" asymptotic "equivalence" of these tests

- in what sense?
- Level, i.e., $\mathbb{P}$[type I error]?
- Power, i.e., $1 - \mathbb{P}$[type II error]?

Answers more scarce than I believed.

Asymptotic equivalence

"Well-known" asymptotic "equivalence" of these tests

- in what sense?
- Level, i.e., $\mathbb{P}[\text{type I error}]$?
- Power, i.e., $1 - \mathbb{P}[\text{type II error}]$?

Answers more scarce than I believed.

What is large sample, anyway? Are sample sizes in modern application large enough?

A modern genetic application

Genetic compositions (at p genomic locations) are compared between n_1 cases and n_2 controls, using association tests on 2×2 tables.

	Variant A	Variant B	
Cases	n_{11}	n_{12}	n_1
Controls	n_{21}	n_{22}	n_2
	m_1	m_2	N

- N from $1,000$s to $500,000$. Imbalance in n_1, n_2 is typically not that bad.

A modern genetic application

Genetic compositions (at p genomic locations) are compared between n_1 cases and n_2 controls, using association tests on 2×2 tables.

	Variant A	Variant B	
Cases	n_{11}	n_{12}	n_1
Controls	n_{21}	n_{22}	n_2
	m_1	m_2	N

- N from $1,000$s to $500,000$. Imbalance in n_1, n_2 is typically not that bad.

Geneticists are worried about “rare variants”, or low variant counts (small m_1).

- When m_1 is small, asymptotics doesn’t apply
- Pearson’s chi-square, etc. fail to control for type I error.
- Barnard’s CSM Test (1945) may work! (idk if anyone uses it...)
- Most people run logistic regressions, afaik.

© Zheng Gao

"Rare-variants"

... is typically defined as a fraction of the number of subjects *N*, say $\epsilon = 0.5\%$.

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)
[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

“Rare-variants”

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

"Rare-variants"

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn't make much sense — one could always apply Fisher's exact test, because it is exact.

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)
[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

"Rare-variants"

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn't make much sense — one could always apply Fisher's exact test, because it is exact.
- The threshold for "rare-variant" is better defined as the **"minimum calibration number[6]"** — the smallest m_1 (and m_2) such that rejection region is non-empty at the specified level α (so that it is meaningful to perform tests).

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

"Rare-variants"

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn't make much sense — one could always apply Fisher's exact test, because it is exact.
- The threshold for "rare-variant" is better defined as the **"minimum calibration number[6]"** — the smallest m_1 (and m_2) such that rejection region is non-empty at the specified level α (so that it is meaningful to perform tests).
- Idea appeared in Sec 10 of Yates (1984)[7]. I wasn't aware...

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

"Rare-variants"

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn't make much sense — one could always apply Fisher's exact test, because it is exact.
- The threshold for "rare-variant" is better defined as the **"minimum calibration number**[6]**"** — the smallest m_1 (and m_2) such that rejection region is non-empty at the specified level α (so that it is meaningful to perform tests).
- Idea appeared in Sec 10 of Yates (1984)[7]. I wasn't aware...

Some properties of this **"minimum calibration number"**:

- The MCN depends on the ancillary marginal (n_1, n_2), a lot. And therefore...

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

“Rare-variants”

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn’t make much sense — one could always apply Fisher’s exact test, because it is exact.
- The threshold for “rare-variant” is better defined as the **“minimum calibration number**[6]**”** — the smallest m_1 (and m_2) such that rejection region is non-empty at the specified level α (so that it is meaningful to perform tests).
- Idea appeared in Sec 10 of Yates (1984)[7]. I wasn’t aware...

Some properties of this **“minimum calibration number”**:

- The MCN depends on the ancillary marginal (n_1, n_2), a lot. And therefore...
- One could overcome the curse of rare variants by choosing appropriate designs!
- See a demo here: https://power.stat.lsa.umich.edu/u-pass/

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

“Rare-variants”

... is typically defined as a fraction of the number of subjects N, say $\epsilon = 0.5\%$.

(Single SNP-based) association tests are not performed if $m_1 < \epsilon N$.

- Doesn’t make much sense — one could always apply Fisher’s exact test, because it is exact.
- The threshold for “rare-variant” is better defined as the **“minimum calibration number**[6]**”** — the smallest m_1 (and m_2) such that rejection region is non-empty at the specified level α (so that it is meaningful to perform tests).
- Idea appeared in Sec 10 of Yates (1984)[7]. I wasn’t aware...

Some properties of this **“minimum calibration number”**:

- The MCN depends on the ancillary marginal (n_1, n_2), a lot. And therefore...
- One could overcome the curse of rare variants by choosing appropriate designs!
- See a demo here: https://power.stat.lsa.umich.edu/u-pass/

My point: finite-sample applicability of the tests is still very much a problem!

[6] Z Gao, J Terhorst, C Van Hout, S Stoev, U-PASS: unified power analysis and forensics for qualitative traits in genetic association studies, Bioinformatics (2019)

[7] Yates, Frank. "Tests of significance for 2× 2 contingency tables." Journal of the Royal Statistical Society: Series A (General) 147.3 (1984): 426-449.

Thank you!

Questions and Comments