Optimal exact tests for complex alternative hypotheses on cross tabulated data

Daniel Yekutieli

Statistics and OR
Tel Aviv University

CDA course
29 July 2017

Plan

1. Death penalty example
2. Methodology
3. Theoretical results
4. Relation to likelihood ratio tests
5. Job Satisfaction example and simulation
6. Discussion

Death Penalty example (Agresti 2002, Table 2.13)

326 subjects are the defendants in indictments involving cases with multiple murders in Florida

Victim's Race	Defendent's Race	Death Penalty	Count
White	White	Yes	19
		No	132
	Black	Yes	11
		No	52
Black	White	Yes	0
		No	9
	Black	Yes	6
		No	97

Research question and some notations

Does probability of receiving death sentence depend on defendant's race?

- X – Race of Victim, Y – Race of Defendant, Z – Death Penalty verdict
- π_{ijk} – the probability that X takes on its ith value and Y takes on its jth value and Z takes on its kth value
- Marginal OR between between defendant race and death penalty

$$\theta_{YZ} = (\pi_{+11} \cdot \pi_{22+})/(\pi_{+12} \cdot \pi_{+21}), \text{ for } \pi_{+jk} = \pi_{1jk} + \pi_{2jk}.$$

- Conditional OR between defendant race and death penalty

$$\theta_{YZ|X=1} = (\pi_{111} \cdot \pi_{122})/(\pi_{121} \cdot \pi_{112})$$

$$\theta_{YZ|X=2} = (\pi_{211} \cdot \pi_{222})/(\pi_{221} \cdot \pi_{212})$$

Victim's race assoc w. defendant race and death penalty

	White defendant	Black defendant
White victim	151	63
Black victim	9	103

$\hat{\theta}_{XY} = 27.1$, 0.95 CI for θ_{XY} is $[12.7, 64.8]$

	Death penalty	No Death penalty
White victim	30	184
Black victim	6	105

$\hat{\theta}_{XZ} = 2.87$, 0.95 CI for θ_{XZ} is $[1.13, 8.73]$

Victim's race is a confounder

As death penalty and white defendant are more likely for a white victim than for a black victim, white defendants have higher probability of receiving death penalty just because they are more likely to kill a white victim.

And indeed we see:

	Death penalty	No Death penalty
White defendant	19	141
Black defendant	17	149

$\hat{\theta}_{YZ} = 1.18$, 0.95 CI for θ_{XZ} is $[0.56, 2.52]$

Hypotheses

- Null hypothesis – conditional on victim's race defendant's and death penalty are independent:

$$H_0 : \ \theta_{YZ|X=1} = 1, \theta_{YZ|X=2} = 1$$

- The alternative hypothesis is Simpson's paradox – the marginal association has a different direction than the conditional associations:

$$H_1 : \ \theta_{YZ|X=1} < 1, \ \theta_{YZ|X=2} < 1, \ \theta_{YZ} > 1$$

Observed counts

White victims: $\hat{\theta}_{YZ|X=1} = 0.68$

	Death penalty	No Death penalty	
White defendant	19	132	151
Black defendant	11	52	63
	30	184	214

Black victims: $\hat{\theta}_{YZ|X=2} = 0$

	Death penalty	No Death penalty	
White defendant	0	9	9
Black defendant	6	97	159
	6	106	112

Data sample space

$$\Omega = \{(N_{111}, N_{211}) : \ N_{111} \in (0, \cdots, 6), \ N_{211} \in (0, \cdots, 30)\}$$

White victims	Death penalty	No Death penalty	
White defendant	N_{111}	$151 - N_{111}$	151
Black defendant	$30 - N_{111}$	$33 + N_{111}$	63
	30	184	214

Black victims	Death penalty	No Death penalty	
White defendant	N_{211}	$9 - N_{211}$	9
Black defendant	$6 - N_{211}$	$97 + N_{211}$	103
	6	106	112

$$\Pr_{H_0}(N_{111} = x, N_{211} = y) = dhyper(x; 151, 63, 30) \cdot dhyper(y; 9, 103, 6)$$

Exact test for death penalty data

- To construct an exact test we need to order the 217 sample points according their strength of evidence in favor of Simpson's paradox
- The exact significance level of the observed data point is the sum of the probabilities of the data points with greater or equal strength of evidence than that of the observed data point.
- However, as Simpson's paradox involves effects having conflicting signs, determining strength of evidence in favor of Simpson's paradox is difficult

For example:

does data point $(20, 0)$ with $\hat{\theta}_{YZ|X=1} = 0$, $\hat{\theta}_{YZ|X=2} = 0.810$, and $\hat{\theta}_{YZ} = 1.34$ offer more evidence in favor of Simpson's paradox than the observed data point $(19, 0)$?

Our proposed test

The posterior probability of the event corresponding to Simpson's paradox

$$\mathcal{P}_1 = \{(\pi_{111} \cdots \pi_{222}) : \ \theta_{YZ|X=1} < 1, \ \theta_{YZ|X=2} < 1, \ 1 < \theta_{YZ} \}.$$

Computing the posterior distributions

- We use a Dirichlet prior with concentration parameters $(0.5 \cdots\ 0.5)$ for $(\pi_{111} \cdots\ \pi_{222})$

- For $(N_{111} \cdots\ N_{222})$, the posterior distribution of $(\pi_{111} \cdots\ \pi_{222})$ is Dirichlet with concentration parameters $(N_{111} + 0.5 \cdots\ N_{222} + 0.5)$

- To compute the posterior probabilities needed to compute our statistics, we sample $(\pi_{111}, \cdots \pi_{222})$ from the posterior and count the proportion of samples that $(\pi_{111} \cdots\ \pi_{222})$ is in $\mathcal{P}_1^{Smpsn}$ or in $\mathcal{P}_0(\epsilon)$.

Exact p-value

- Data point $(20, 0)$ with $\Pr_{H_0}(20, 0) = 0.087$ has the largest posterior probability of $\mathcal{P}_1$: 0.085954 $(s.e. < 0.0001)$.
- The observed table with $\Pr_{H_0}(19, 0) = 0.064$ has the second largest posterior probability of $\mathcal{P}_1$, 0.07983 $(s.e. < 0.0001)$.
- Data point $(21, 0)$ with $\Pr_{H_0}(21, 0) = 0.101$ has the third largest posterior probability of of $\mathcal{P}_1$, 0.07955 $(s.e. < 0.0001)$.

Thus the significance level of the observed table is:

$$0.151 = 0.087 + 0.064$$

Setup

- The parameter is $\boldsymbol{p} \in \mathcal{P}$ and $\pi(\boldsymbol{p})$ is the prior distribution.
- the data is $\boldsymbol{N} \in \Omega$; $\Pr(\boldsymbol{n}|\,\boldsymbol{p})$ is the likelihood.
- The alternative hypothesis is $H_1 : \boldsymbol{p} \in \mathcal{P}_1$, where $\mathcal{P}_1 \subseteq \mathcal{P}$ is the discovery event and $\mathcal{P}_0 \subseteq \mathcal{P} - \mathcal{P}_1$ the non-discovery event.
- The null hypothesis H_0 does not have to correspond to an explicit subset or point in $\mathcal{P}_0$, all we will need is that H_0 specifies a null distribution $\Pr_{H_0}(\boldsymbol{N} = \boldsymbol{n})$ on Ω.
- Tests are mappings $\mathcal{T} : \ \Omega \to \{0, 1\}$, where $\mathcal{T} = 1$ corresponds to rejecting H_0, and for $S \subseteq \Omega$, let $\mathcal{T}(S) := I(\boldsymbol{n} \in S)$.
- The significance level of $\mathcal{T}(S)$ is $Pr_{H_0}(\boldsymbol{N} \in S)$.

Optimal tests are Bayes classifiers

Our tests are Bayes rules for the following loss function:

$$L(S; \lambda_1, \lambda_2) = \lambda_1 \cdot I(\boldsymbol{N} \in S,\ \boldsymbol{P} \in \mathcal{P}_0) + \lambda_2 \cdot I(\boldsymbol{N} \notin S,\ \boldsymbol{P} \in \mathcal{P}_1).$$

To derive the Bayes rules note that the marginal distribution of $\boldsymbol{N}$ is

$$\Pr(\boldsymbol{N} = \boldsymbol{n}) = \int_{\boldsymbol{p}} \pi(\boldsymbol{p}) \cdot \Pr(\boldsymbol{N} = \boldsymbol{n} | \, \boldsymbol{p}) \, d\boldsymbol{p},$$

and the conditional distribution of $\boldsymbol{p}$ given $\boldsymbol{n}$ is

$$\pi(\boldsymbol{p} | \, \boldsymbol{n}) = \Pr(\boldsymbol{N} = \boldsymbol{n} | \, \boldsymbol{p}) \cdot \pi(\boldsymbol{p}) / \Pr(\boldsymbol{N} = \boldsymbol{n}).$$

Thus the average risk can be expressed

$$\begin{aligned} &\sum_{\boldsymbol{n}} \Pr(\boldsymbol{n}) \cdot \int_{\boldsymbol{p}} \pi(\boldsymbol{p} | \, \boldsymbol{n}) \cdot [\lambda_1 \cdot I(\boldsymbol{n} \in S,\ \boldsymbol{P} \in \mathcal{P}_0) + \lambda_2 \cdot I(\boldsymbol{n} \notin S,\ \boldsymbol{P} \in \mathcal{P}_1)] \, d\boldsymbol{p} \\ &= \sum_{\boldsymbol{n} \in S} \Pr(\boldsymbol{n}) \cdot \lambda_1 \cdot \Pr(\boldsymbol{P} \in \mathcal{P}_0 | \, \boldsymbol{n}) + \sum_{\boldsymbol{n} \notin S} \Pr(\boldsymbol{n}) \cdot \lambda_2 \cdot \Pr(\boldsymbol{P} \in \mathcal{P}_1 | \, \boldsymbol{n}) \end{aligned}$$

Specifying the Bayes classifier

- S that minimizes the average risk is

$$S^{Bayes}(\lambda_1,\lambda_2) = \{\boldsymbol{n} :\ \frac{\lambda_1}{\lambda_2} \leq \frac{\Pr(\boldsymbol{P} \in \mathcal{P}_1 |\ \boldsymbol{n})}{\Pr(\boldsymbol{P} \in \mathcal{P}_0 |\ \boldsymbol{n})}\}$$

- To derive level α tests we specify the Bayes classifiers according to the significance level (instead of λ_1 and λ_2).
- Thus, for

$$S^{Bayes}(\delta) = \{\boldsymbol{n} :\ \delta \leq \frac{\Pr(\boldsymbol{P} \in \mathcal{P}_1 |\ \boldsymbol{n})}{\Pr(\boldsymbol{P} \in \mathcal{P}_0 |\ \boldsymbol{n})}\},$$

  We define $S^{Bayes}(\alpha) := S^{Bayes}(\delta_\alpha)$ with

$$\delta_\alpha = \max\{\delta :\ Pr_{H_0}(\boldsymbol{N} \in S^{Bayes}(\delta)) \leq \alpha\ \}$$

Mean most powerful tests

Definition 1.

1. The *mean significance level* of $\mathcal{T}(S)$ is $Pr(\boldsymbol{N} \in S | \boldsymbol{p} \in \mathcal{P}_0)$.
2. The *mean power* of $\mathcal{T}(S)$ is $Pr(\boldsymbol{N} \in S | \boldsymbol{p} \in \mathcal{P}_1)$.
3. $\mathcal{T}(S)$ is a *mean most powerful* test if all tests with less or equal mean significance level have less or equal mean power.

Proposition 2. $\forall \delta$, $\mathcal{T}(S^{Bayes}(\delta))$ *is a mean most powerful test.*

The proof is very similar to the proofs i haven't given in the two previous lectures

A few remarks

- Determining $\mathcal{P}_1$, $\mathcal{P}_0$, and $\pi(\boldsymbol{p})$, produces a family of mean most powerful tests.
- By construction, $\mathcal{T}(S^{Bayes}(\alpha))$ has significance level α and has more mean power than all mean most powerful tests with significance level $< \alpha$.
- According to Proposition 2, $\mathcal{T}(S^{Bayes}(\alpha))$ also has more mean power than *all* tests with smaller or equal mean significance levels.
- Ideally, the prior distribution captures the knowledge regarding the parameters. In our examples in we use conjugate non-informative priors that provide easy test statistic computation and yield general optimal tests for each alternative null hypothesis.

A few more remarks

$\mathcal{P}_1$ is dictated by application, but $\mathcal{P}_0$ can be any subset of $\mathcal{P} - \mathcal{P}_1$.

- We suggest either setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$, or setting $\mathcal{P}_0 = \mathcal{P}_0(\epsilon)$ to be a "small" ball around the null parameter value $\boldsymbol{p}_0$.
- For $\mathcal{P}_0 = \{\boldsymbol{p}_0\}$, the mean significance level equals the significance level, thus $\mathcal{T}(S^{Bayes}(\alpha))$ would have more mean power then all level α tests. Setting $\mathcal{P}_0 = \mathcal{P}_0(\epsilon)$ is a numeric solution for producing a very similar tests.
- Setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ for which (1) holds, has the great technical advantage that to construct the test, for each data point, we only need to assess the posterior probability of $\mathcal{P}_1$. I think that in most cases the choice of $\mathcal{P}_0$ has little effect (!!?)

$$\frac{\Pr(\boldsymbol{P} \in \mathcal{P}_1 |\, \boldsymbol{n})}{\Pr(\boldsymbol{P} \in \mathcal{P}_0 |\, \boldsymbol{n})} = \frac{\Pr(\boldsymbol{P} \in \mathcal{P}_1 |\, \boldsymbol{n})}{1 - \Pr(\boldsymbol{P} \in \mathcal{P}_1 |\, \boldsymbol{n})}, \tag{1}$$

Relation btwn our tests and likelihood ratio tests

- For simple hypotheses, $H_0 : \boldsymbol{p} = \boldsymbol{p}_0 \in \mathcal{P}_0$ vs. $H_1 : \boldsymbol{p} = \boldsymbol{p}_1 \in \mathcal{P}_1$, our test reduces to the likelihood ratio test if $\mathcal{P}_0 = \{\boldsymbol{p}_0\}$ and $\mathcal{P}_1 = \{\boldsymbol{p}_1\}$ or if the prior distribution assigns all its probability to $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$.
- The likelihood ratio test for composite hypotheses tests $H_0 : \boldsymbol{p} \in \mathcal{P}_{null}$ vs. $H_1 : \boldsymbol{p} \notin \mathcal{P}_{null}$ using the statistic

  $$\Lambda(\boldsymbol{n}) = \frac{\sup_{\boldsymbol{p} \in \mathcal{P}_{null}} \Pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{p})}{\sup_{\boldsymbol{p} \in \mathcal{P}} \Pr(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{p})}.$$

  If $\mathcal{P}_1 = \mathcal{P} - \mathcal{P}_{null}$ and setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$, $\Lambda(\boldsymbol{n})$ is similar to one minus our statistic, except that we consider the average rather than the supremum of the likelihood. HOWEVER if $\mathcal{P}_1$ is a “small” subset of $\mathcal{P} - \mathcal{P}_{null}$ our test that sorts the sample space according to $\mathcal{P}_1$ can be considerably more powerful.
- This is shown in the next example and in our contingency table examples in which $\Lambda(\boldsymbol{n})$ is the X^2 statistic.

Difference btwn our tests and likelihood ratio test

$\boldsymbol{\mu} = (\mu_1 \cdots \mu_K)$, $\boldsymbol{Y} = (Y_1 \cdots Y_K)$ with $Y_k \sim N(\mu_k, 1)$.

$H_0 : \mu \equiv 0, \quad H_1 : \boldsymbol{\mu} \in \{\boldsymbol{\mu} : \ 3 \leq \mu_1\}$.

- In the likelihood ratio test the data points are ordered according to $\|\boldsymbol{y}\|$. As $\chi^2_{100,0.95} = 124.34$, the rejection region for the $\alpha = 0.05$ likelihood ratio test is $\mathcal{S} = \{\boldsymbol{y} : \ 124.34 \leq \|\boldsymbol{y}\|^2\}$
- $\mathcal{P}_1 = \{\boldsymbol{\mu} : \ 3 \leq \mu_1\}$. Setting $\mathcal{P}_0 = \mathcal{P} - \mathcal{P}_1$ and using a flat prior for $\boldsymbol{\mu}$, in our test the data points are ordered according to y_1 and the rejection for our $\alpha = 0.05$ test is $\mathcal{S}^{Bayes} = \{\boldsymbol{y} : \ 1.64 \leq y_1\}$
- Thus, for $K = 100$ and $\boldsymbol{\mu} = (3.2, 0 \cdots 0)$, the power of the likelihood ratio test is 0.179, while the power of our test is 0.940.

Job Satisfaction Example (Agresti 2002, Table 2.8)

Income	Job Satisfaction			
	Very	Little	Moderately	Very
(Dollars)	Dissatisfied	Dissatisfied	Satisfied	Satisfied
<15000	1	3	10	6
15000-25000	2	3	10	7
25000-40000	1	6	14	12
>40000	0	1	9	11

Testing independence between income and job satisfaction

- Pearson's Chi-squared test (R *chisq.test* function), corresponding to a general alternative hypothesis of dependance between of income and job satisfaction: $X^2 = 5.97$ with 9 degrees of freedom and p-value 0.743.

- Spearman's rank correlation coefficient (R *cor.test* function), alternative hypothesis of positive correlation between income and job satisfaction: $\rho = 0.177$ with p-value 0.042.

- Kendall's rank correlation coeficient (R *cor.test* function), corresponding to alternative hypothesis of concordance between of income and job satisfaction: $\tau = 0.152$ with p-value 0.043.

All significance levels are based on parametric approximation of the test statistics' distribution under the null hypothesis

Concordance

- π_{ij} is probability of respondent having income level i and job satisfaction level j
- A pair of respndents is concordant if they have different income and job satisfaction and the respondent with higher income has higher job satisfaction, its probability:

$$\Pi_C = 2\sum_i \sum_j \pi_{ij}(\sum_{i<h} \sum_{j<k} \pi_{hk})$$

- A pair of respondents is discordant if they have different income and job satisfaction and the respondent with higher income has lower job satisfaction, its probability:

$$\Pi_D = 2\sum_i \sum_j \pi_{ij}(\sum_{i<h} \sum_{k<j} \pi_{hk})$$

- Concordance is measured by Kendall's *gamma*:

$$\gamma = (\Pi_C - \Pi_D)/(\Pi_C + \Pi_D)$$

Exact test for independence vs concordance alternative

- We assume $(N_{11} \cdots\ N_{44}) \sim multinom(\pi_{11} \cdots\ \pi_{44})$,
- $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$.
- To construct the exact tests note that under H_0 conditioning on $N_{1+} = n_{1+}, \cdots, N_{+4} = n_{+4}$:

$$N_{ij} \sim MVhypergeometric(n_{1+}, \cdots\ n_{+4})$$

- There are $90, 208, 550$ possible 4-by-4 tables with the same row and columns sums as Table 2
- Setting $\hat{\pi}_{ij} = n_{ij}/n_{++}$, yields $\hat{\gamma} = 0.221$.
- The exact significance level for the test for concordance based on the $\hat{\gamma}$ statistic is $p - value = 0.0415$, computed by summing the probabilities under the null of observing the $21, 101, 151$ tables with $0.221 \leq \hat{\gamma}$

Our exact test for concordance alternative

- Our statistic is the pstrior probability of the concordance event,

$$\Pr(0 \leq \gamma | \, N_{11} \cdots \, N_{44}) \tag{2}$$

- We use a Dirichlet prior for which posterior distribution is $Dirichlet(N_{11} + 0.5 \cdots \, N_{44} + 0.5)$
- To assess (2) we sample $(\pi_{11}, \cdots \pi_{44})$ from the posterior and record the proportion of times the concordance event occurs.
- The probability of concordance for $N_{ij} = n_{ij}$, based on a sample of 10^7 draws from the posterior, was 0.9564 ($s.e. < 0.0001$).
- To compute the significance of this statistic, we sample of $50,000$ 4-by-4 tables from the null, for each table we assess the probability of concordance, and record the proportion of tables with probability of concordance ≥ 0.9564.
- The estimated significance level was $p - value = 0.036$ ($s.e. < 0.001$).

Exact test for the positive dependence alternative

- Our statistic is the posterior probability of the event:

$$\mathcal{P}_1^{Pos} = \{(\pi_{11}, \cdots, \pi_{44}) : \ \Pr(\pi_{j|i} \le t) \ge \Pr(\pi_{j|i+1} \le t) \ \forall i,j\} \qquad (3)$$

  with $\pi_{j|i} = \pi_{ij}/\pi_{i+}$
- Using the same prior as before, we assess the statistic's value by sampling $(\pi_{11}, \cdots \pi_{44})$ from the posterior and record the of times (3) occurred.
- The observed statistic value is 0.0118 ($s.e. < 0.0001$), and its estimated significance level is $p - value = 0.0093$ ($s.e. < 0.001$)

Job Satisfaction Simulation

The simulation compares the power of the conditional exact test whose test statistic is $\hat{\gamma}$ with the conditional exact test whose test statistic is $\Pr(0 \leq \gamma | \, N_{11} \cdots \, N_{44})$

- The null distribution of $(N_{11} \cdots \, N_{44})$ is the conditional multivariate hypergeometric considered before
- The alternative distribution is $multinomial(\hat{\pi}_{ij} = n_{ij}/96)$ truncated to have $N_{1+} = n_{1+}, \cdots, N_{+4} = n_{+4}$.
- Simulation: We generate 10^5 realizations of $(N_{11} \cdots \, N_{44})$ from the alternative distribution and then use the process described before to compute the two kinds of p-values for each realized $(N_{11} \cdots \, N_{44})$
- For the $\hat{\gamma}$ statistic the mean p-value was 0.0988 and 0.537 $(s.e. < 0.005)$ of the p-values were smaller than 0.05
- For the p-values computed based on the probability of concordance statistics, the mean p-value was 0.0947 and 0.550 $(s.e. < 0.005)$ of the p-values were smaller than 0.05.

Discussion

- We presented methodology for the analysis of contingency tables in which use of exact tests is well established. However our conditional tests can easily be extended to non-parametric tests in which the null hypothesis can be generated with permutations or bootstrap samples, and also to numeric parametric tests!
- Our tests are computationally intensive. We therefore suggest using them in (1) “difficult” cases where the parameter space is high dimensional and we know how to express the alternative hypothesis as a subset of the parameter space however it is not clear how to construct a test statistic for this hypothesis; (2) in cases where there is prior information on the parameter; (3) for very high dimensional and very sparse tables in which the asymptotic results for the test statistic distribution fail.