Using probabilistic couplings in data analysis

Víctor H. Cervantes[1], Ehtibar N. Dzhafarov[2]

Purdue University
[1]cervantv@purdue.edu
[2]ehtibar@purdue.edu

Quantum Contextuality in Quantum Mechanics and Beyond
Prague
May 19, 2019

Stochastically unrelated random variables

- Consider a coin flipped n_1 times here and another coin flipped n_2 times in the USA.
- The number of head of these coins may be represented by random variables $X_1 \sim \text{Binomial}(n_1, p_1)$ and $X_2 \sim \text{Binomial}(n_1, p_2)$.

Stochastically unrelated random variables

- Random variables X_1 and X_2 are generally taken as independent random variables.
- There is no logical justification for this.
- We investigated a more principled approach: using all possible couplings and choosing one that is optimal in accordance with certain criteria.
- We did this for the case where both X_1 and X_2 have the same n.

Stochastically unrelated random variables

Bell inequalities

- Note that in the usual Alice-Bob setting, the situation is similar when considering each pair of measurements performed by Alice with varying choices of Bob, and vice versa.

Couplings

- A coupling of a pair of random variables $\{X, Y\}$ is a random variable $\left(\widetilde{X}, \widetilde{Y}\right)$ (with jointly distributed components), such that $\widetilde{X} \overset{d}{=} X, \quad \widetilde{Y} \overset{d}{=} Y$, where $\overset{d}{=}$ stands for "has the same distribution as."
- A coupling always exists, generally non-uniquely.

Coupling of two binomial random variables

Optimal Coupling

- We applied the maximum likelihood meaning of optimality to the task of identifying and comparing two probabilities from two stochastically unrelated sets of binary events.

Coupling of two binomial random variables

- Let $X_1 \sim \text{Binomial}(n, p_1)$ and $X_2 \sim \text{Binomial}(n, p_2)$ be two stochastically unrelated random variables for a given number of observations n.
- Let $Z = (Z_1, Z_2)$ be a coupling of X_1 and X_2

Coupling of two binomial random variables

- Z is a random 2×2 matrix whose cells follow a multinomial distribution with parameters $(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ such that $\theta_{11} + \theta_{12} = p_1$ and $\theta_{11} + \theta_{21} = p_2$.

$$\begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}$$

Coupling of two binomial random variables

- Z is a random 2×2 matrix whose cells follow a multinomial distribution with parameters $(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ such that $\theta_{11} + \theta_{12} = p_1$ and $\theta_{11} + \theta_{21} = p_2$.

$$\begin{bmatrix} \theta_{11} & p_1 - \theta_{11} \\ p_2 - \theta_{11} & 1 - p_1 - p_2 + \theta_{11} \end{bmatrix}$$

Coupling of two binomial random variables

- Z is a random 2×2 matrix whose cells follow a multinomial distribution with parameters $(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$ such that $\theta_{11} + \theta_{12} = p_1$ and $\theta_{11} + \theta_{21} = p_2$.

$$\begin{bmatrix} \theta_{11} & p_1 - \theta_{11} \\ p_2 - \theta_{11} & 1 - p_1 - p_2 + \theta_{11} \end{bmatrix}$$

- Given data for $X_1 = x_1$ and $X_2 = x_2$, we wish to explore the likelihood of the possible couplings Z.

Coupling of two binomial random variables

- Note that a realization of a coupling Z is of the following form

$$\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

  where
  $m_{11} + m_{12} + m_{21} + m_{22} = n,$
  $m_{11} \in \{\max(x_1 + x_2 - n, 0), \ldots, \min(x_1, x_2)\},$
  $m_{11} + m_{12} = x_1,$
  $m_{11} + m_{21} = x_2.$
- $\Pr(Z = \{m_{11}, m_{12}, m_{21}, m_{22}\}) = \binom{n}{m_{11}\, m_{12}\, m_{21}\, m_{22}} \prod_{i=1}^{2} \prod_{j=1}^{2} \theta_{ij}^{m_{ij}}$

Coupling of two binomial random variables

Likelihood

Thus, the likelihood is defined by

$$\mathcal{L}(\theta_{11}, p_1, p_2 | n, x_1, x_2) =$$

$$\Pr(Z_1 = x_1, Z_2 = x_2) =$$

$$\sum_{m_{11}=a}^{b} \Pr(Z = \{m_{11}, x_1 - m_{11}, x_2 - m_{11}, n - x_1 - x_2 + m_{11}\})$$

where $a = \max(x_1 + x_2 - n, 0)$ and $b = \min(x_1, x_2)$

Coupling of two binomial random variables

Maximizing the likelihood

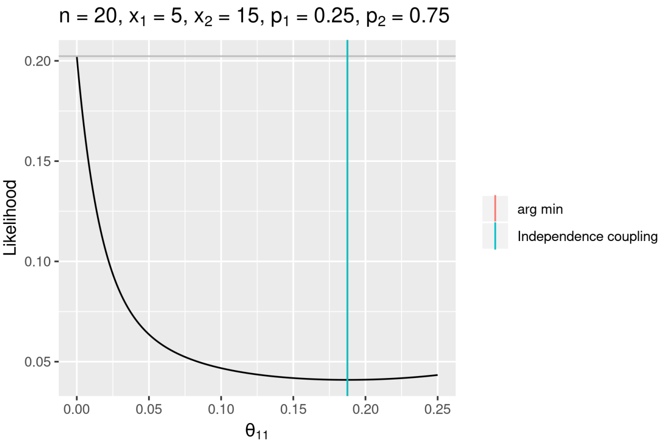
- Given data, the likelihood can easily be maximized numerically.
- Also, by functional invariance of likelihood estimators,

$$\hat{p}_i = {}^{x_i}/_{n}, i = 1, 2$$

.

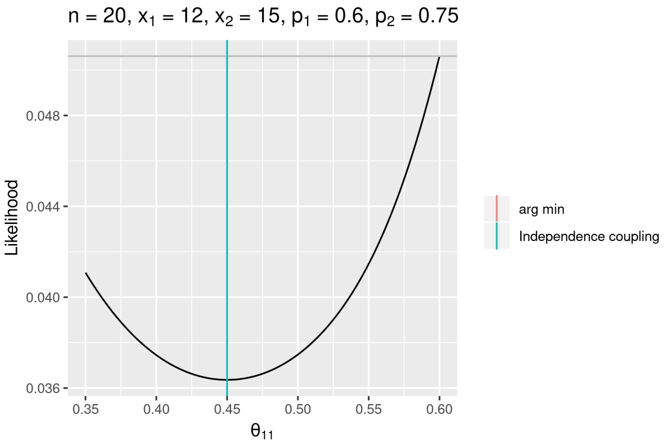
Coupling of two binomial random variables

Maximizing the likelihood (examples)



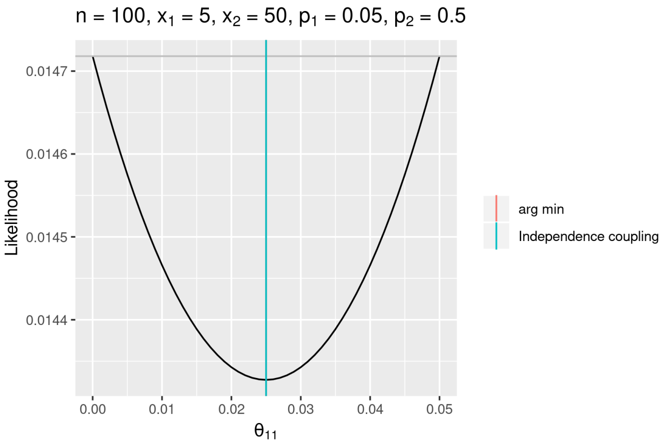
Coupling of two binomial random variables

Maximizing the likelihood (examples)



Coupling of two binomial random variables

Maximizing the likelihood (examples)



Testing equality of two probabilities

Likelihood

If we assume equality of proportions, the restricted likelihood becomes

$$\mathcal{L}(\theta_{11}, p|n, x_1, x_2) = \sum_{m_{11}=a}^{b} \frac{1}{2^{-(n-x_1-x_2+m_{11})}} \times$$

$$\frac{n!}{m_{11}!(x_1 - m_{11})!(x_2 - m_{11})!(n - x_1 - x_2 + m_{11})!} \times$$

$$\theta_{11}^{m_{11}} (2p - 2\theta_{11})^{x_1+x_2-2m_{11}} (1 - 2p + \theta_{11})^{n-x_1-x_2+m_{11}}$$

Testing equality of two proportions

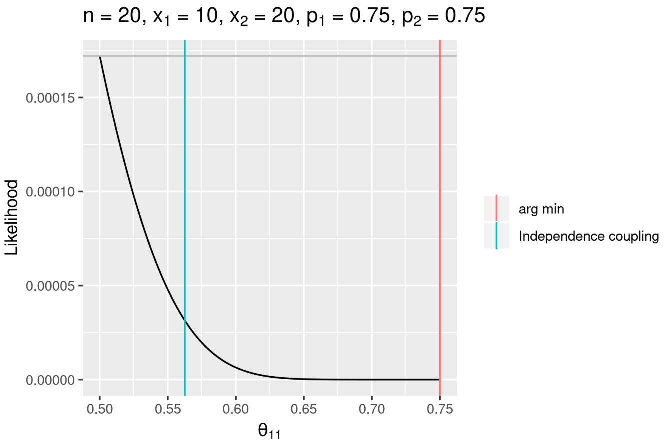
Maximizing the likelihood (examples)

For all cases we have explored, the optimal coupling maximizing (p, θ_{11}), is given by

- $\hat{p} = \frac{x_1+x_2}{2n} = \frac{1}{2}(\frac{x_1}{n} + \frac{x_2}{n})$
- $\hat{\theta}_{11} = \begin{cases} \max(0, {}^{(x_1+x_2)}/_{n} - 1) & \text{(minimal coupling)} \\ \min({}^{x_1}/_{n}, {}^{x_2}/_{n}) & \text{(maximal coupling)} \end{cases}$

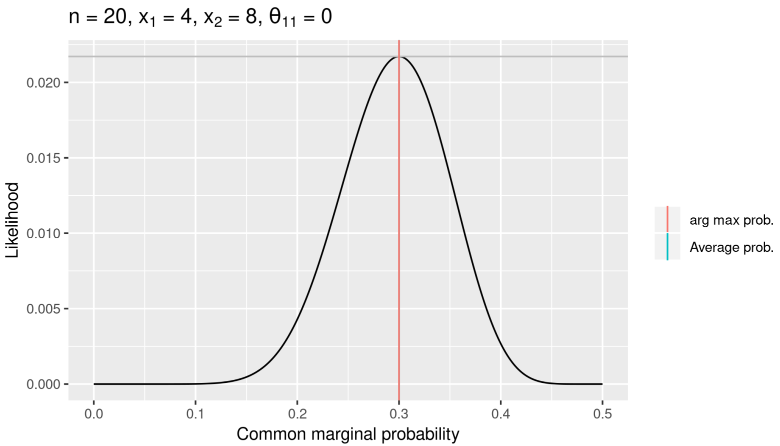
Testing equality of two proportions

Maximizing the likelihood (examples)



Testing equality of two proportions

Maximizing the likelihood (examples)



Testing equality of two proportions

Testing equality

$$H_o : p_1 = p_2 = p$$

vs.

$$H_a : p_1 \neq p_2$$

$$\hat{\lambda} = \frac{\max\{\mathcal{L}(\theta_{11}, p_1, p_2 | n, x_1, x_2)\}}{\max\{\mathcal{L}(\theta_{11}, p | n, x_1, x_2)\}}.$$

Testing equality of two proportions

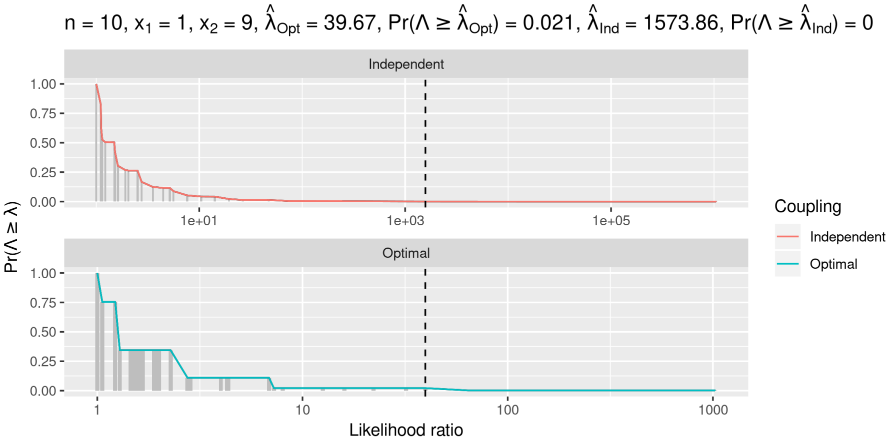
Testing equality

We approximate the distribution of $\hat{\lambda}$ via parametric bootstrap:

1. Given n, x_1, x_2 find $\hat{\lambda}$, and $\hat{\theta}_{11}, \hat{p}$ such that $\mathcal{L}(\hat{\theta}_{11}, \hat{p}|n, x_1, x_2) = \max\{\mathcal{L}(\theta_{11}, p|n, x_1, x_2)\}$
2. For each possible sample of Z distributed with $\hat{\theta}_{11}, \hat{p}$ find $\lambda(n, z_1, z_2)$.

Testing equality of two proportions

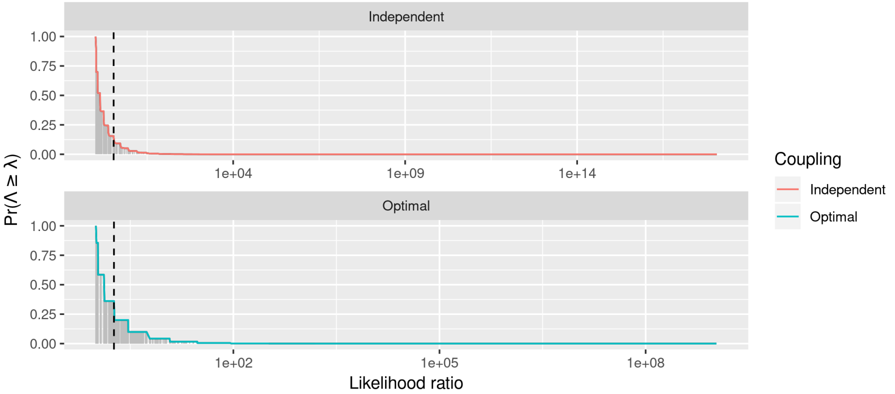
Testing equality (examples)



Testing equality of two proportions

Testing equality (examples)

$n = 30,\ x_1 = 12,\ x_2 = 18,\ \hat{\lambda}_{Opt} = 1.83,\ \Pr(\Lambda \geq \hat{\lambda}_{Opt}) = 0.362,\ \hat{\lambda}_{Ind} = 3.35,\ \Pr(\Lambda \geq \hat{\lambda}_{Ind}) = 0.155$



Closing Remarks

- Optimal couplings are readily identifiable, and the independent coupling is rarely optimal.
- Considerations of stochastical unrelatedness and couplings lead to rethink the basic assumptions of statistical analysis.
- Some conclusions may coincide between optimal and independence couplings (e.g., some point estimates).
- Decisions may not necessarily be the same: given the same data and choice of significance level, the optimal coupling leads to a more conservative test.

Thank you!