# Statistical significance and undersampling

Dr. Joseph Lizier

THE UNIVERSITY OF SYDNEY

# Statistical significance: session outcomes

- Understand estimation of information as a statistic, and how to test for significance of that statistic.
- Understand where and how such tests can be performed analytically.
- Apply statistical significance testing of MI, CMI etc., using JIDT

- Primary references:
  - J.T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics and AI, 1:11, 2014; appendix A.5

# Mutual information

– Recall statistical interpretation of MI:

$$\mathrm{I}(X;Y) = 0 \;\leftrightarrow X \text{ is independent of } Y$$

– In theory …
– In practice, or from empirical data:
  a. We can have X is independent of Y,
  b. But measure I(X;Y) $\neq$ 0 !

– Q1: Is a given estimate of I(X;Y) different to or consistent with 0?
– Q2: How many samples do we need to determine this, or indeed an accurate value of I(X;Y)?

# Q1: Is estimate I(X;Y) consistent with 0?

- Taking a statistical view, we form a statistical test of I(X;Y):

- Null hypothesis $H_0$: X is independent of Y
- Alternative hypothesis: X has a dependence on Y

- To test $H_0$ → Test probability of sampling the statistic I(X;Y) assuming it is distributed under $H_0$:
  1. Form surrogate distribution of $I(X;Y^s)$ where $Y^s$ are surrogates for Y generated under $H_0$
     - Which have same statistical properties as Y, but potential relationship to X is destroyed.
     - With p(x | y) distributed as p(x) (whilst p(y) retained)

$$\mathrm{I}(X;Y) = \sum_{x \in A_x, y \in A_y} p(x,y) \log_2 \frac{p(x|y)}{p(x)}$$

  2. Measure p-value of $p(I(X;Y) \leq I(X;Y^s))$ and take one-tailed test against $\alpha$

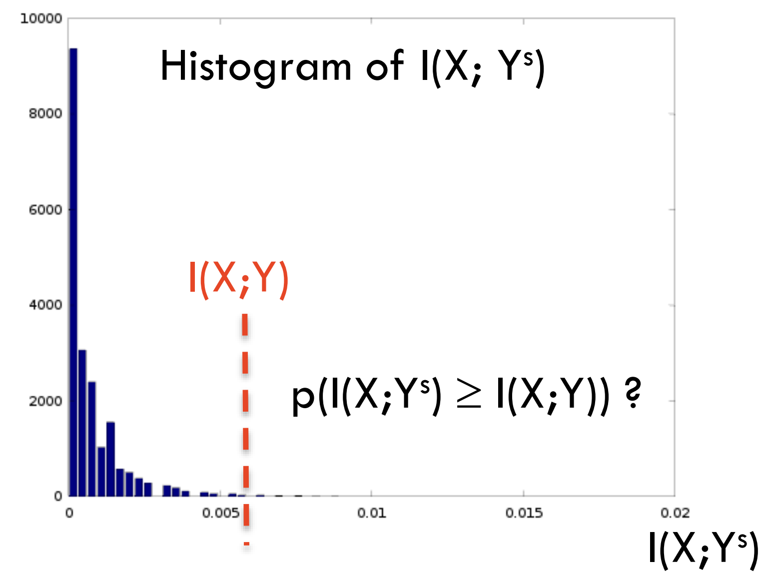
J.T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics & AI, 1:11, 2014; appendix A.5

# Q1: Is estimate I(X;Y) consistent with 0?

- In practice, we generate the surrogate distribution empirically:
  - By resampling† the N samples y of Y to create a surrogate variable $Y^{s1}$; which has no per-sample relation to X;
  - Then computing $I(X;Y^{s1})$;
  - And repeating many times (*S*) to get the distribution $I(X;Y^{s})$.

| X |
|---|
| $x_1$ |
| $x_2$ |
| $x_3$ |
| $x_4$ |
| $x_5$ |
| $x_6$ |
| … |

| Y | $Y^{s1}$ | $Y^{s2}$ | $Y^{s3}$ | $Y^{s4}$ | … | $Y^{sS}$ |
|---|---|---|---|---|---|---|
| $y_1$ | $y_{10}$ | $y_8$ | $y_{27}$ | $y_{45}$ | … | $y_{94}$ |
| $y_2$ | $y_4$ | $y_{37}$ | $y_{58}$ | $y_{73}$ | … | $y_{29}$ |
| $y_3$ | $y_{23}$ | $y_{88}$ | $y_{38}$ | $y_{55}$ | … | $y_{13}$ |
| $y_4$ | $y_5$ | $y_{12}$ | $y_{44}$ | $y_{76}$ | … | $y_{89}$ |
| $y_5$ | $y_{72}$ | $y_{51}$ | $y_{22}$ | $y_{11}$ | … | $y_3$ |
| $y_6$ | $y_{16}$ | $y_{99}$ | $y_{81}$ | $y_{21}$ | … | $y_{65}$ |
| … | … | … | … | … | … | … |
| $I(X;Y)$ | $I(X;Y^{s1})$ | $I(X;Y^{s2})$ | $I(X;Y^{s3})$ | $I(X;Y^{s4})$ | … | $I(X;Y^{sS})$ |




† Via permutation or bootstrap sampling.

J.T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics & AI, 1:11, 2014; appendix A.5

# Q1: Is estimate $I(X;Y)$ consistent with 0?

- Same principle holds for conditional MI $I(X;Y|Z)$, but
  - We generate surrogate distribution $p(x|y,z)$ as $p(x|z)$.

- Notice the difference to $I(X;Y)$?
  - Becomes a directional test here.
  - *Asymptotically* it doesn't matter if we resample x or y
  - Often we're interested in a directional test anyway (e.g. with transfer entropy – see Information Transfer session).

# Aside: normalising measurements

- Sometimes we normalise information estimates by removing the component due to finite sample size:

$$I^n(X;Y) = I(X;Y) - \langle I(X;Y^s) \rangle$$

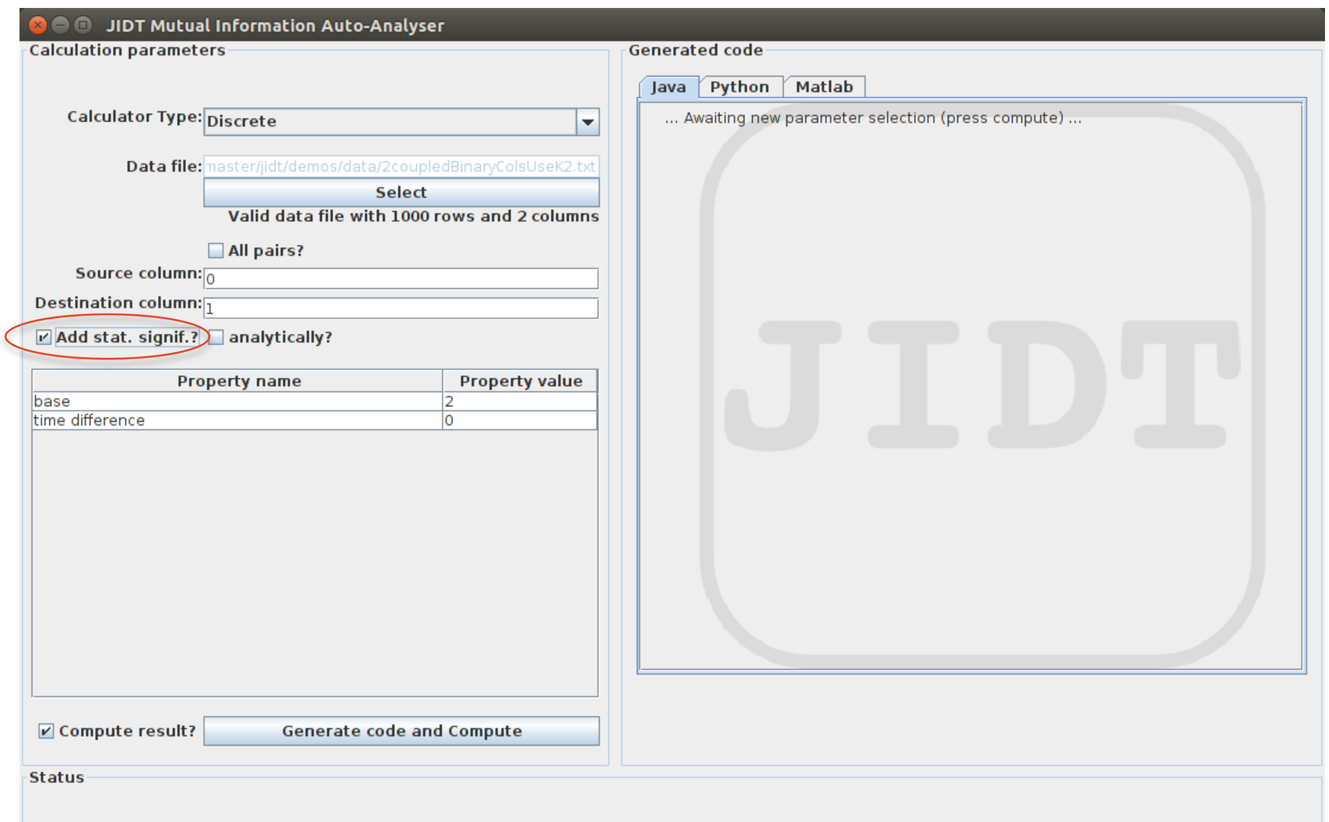
- This is equivalent to bias correction (so not necessary if bias correction works well).

Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 4.5.2.

# Statistical significance test in JIDT

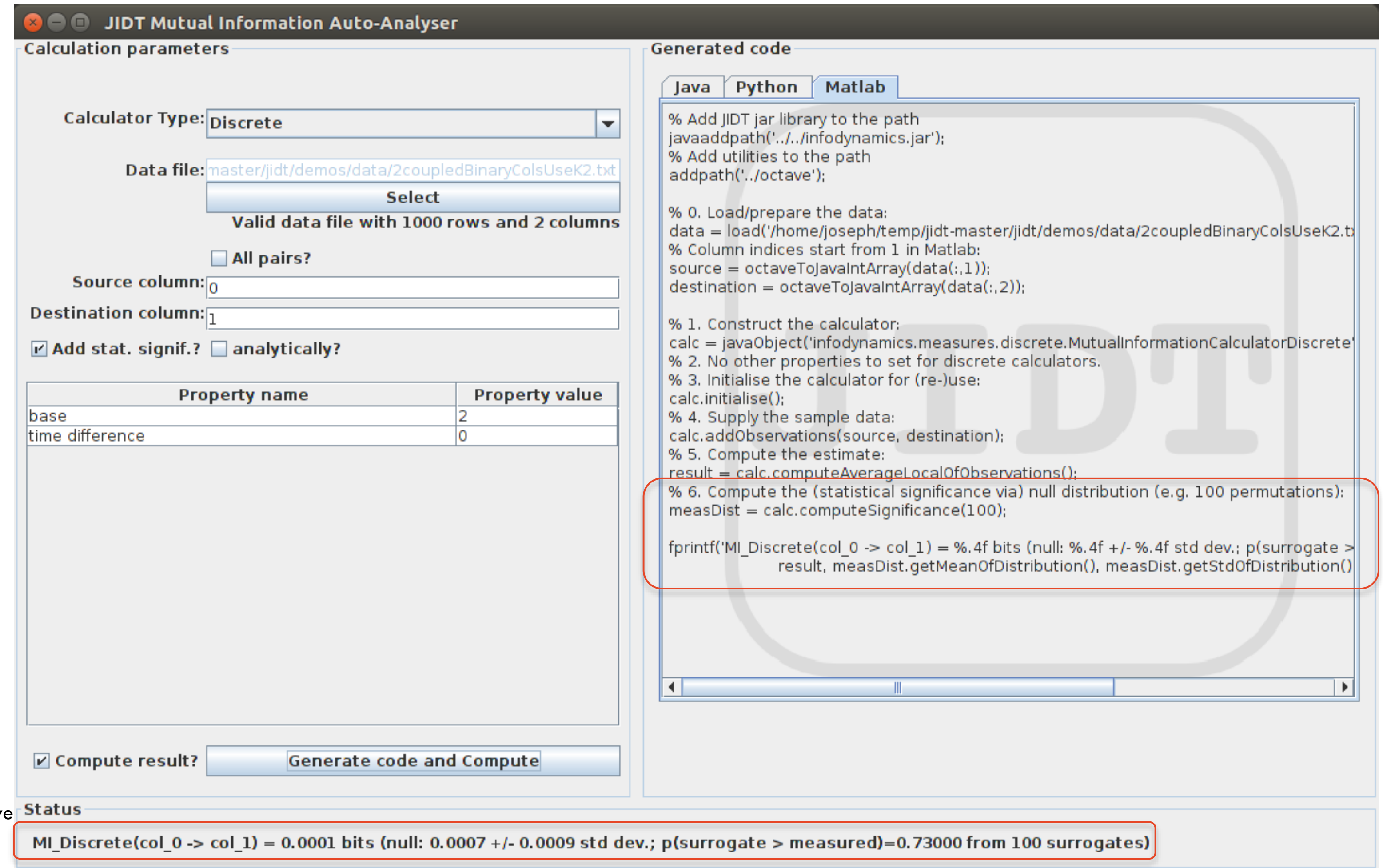
– Generate from AutoAnalyser by clicking the checkbox next to "Add stat. signif.?"

– Available on all MI, CMI and calculators based on these.



# Statistical significance test in JIDT

- The `getSignificance(numPermutations)` method returns an `EmpiricalMeasurementDistribution` object (see Javadocs) from which you can retrieve full surrogate distribution, mean, std dev, and p-value of statistic.



# Analytic surrogate distributions

- For some estimators, we have an analytic representation of the surrogate distribution.
- Specifically, $2N \times I(X;Y^s)$ or $2N \times I(X;Y^s|Z)$ in ***nats*** follow $\chi^2$ distributions with the following degrees of freedom:

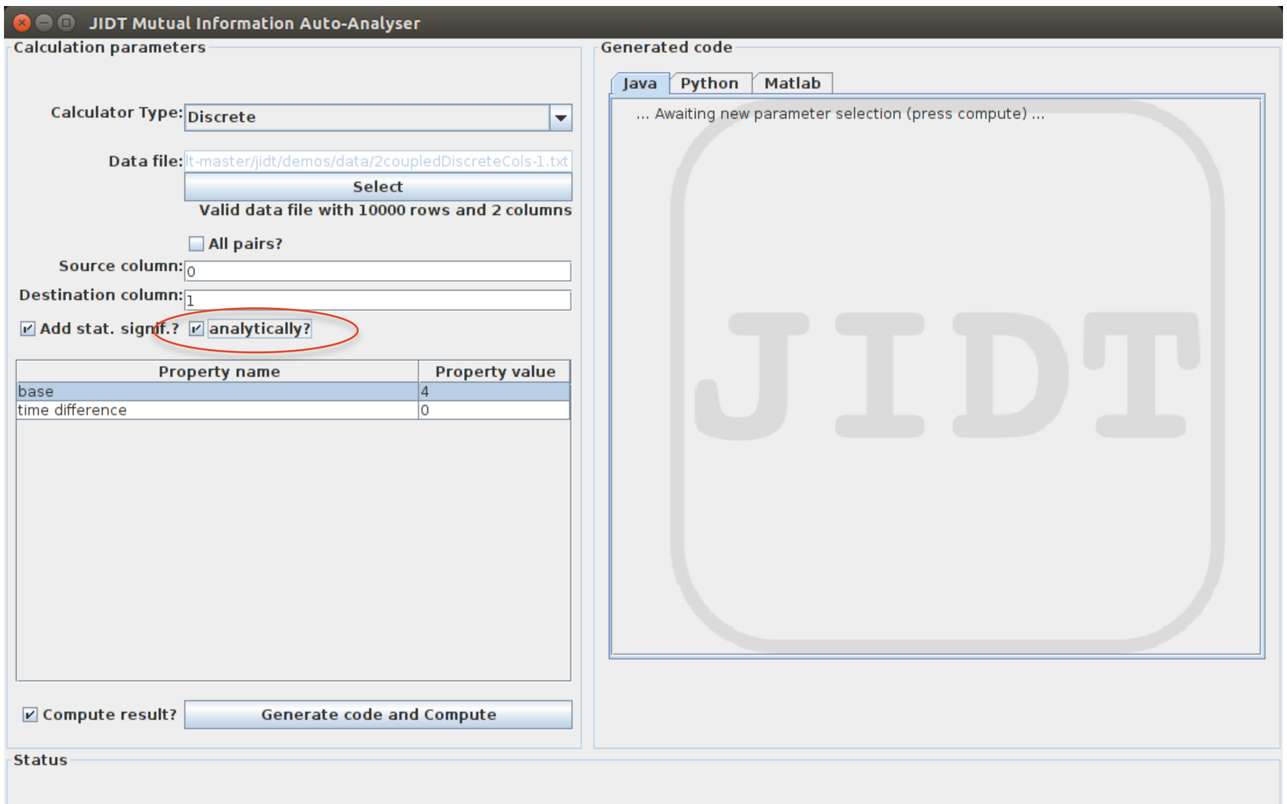
| Estimator | Mutual info $I(X;Y^s)$ | Conditional Mutual Info $I(X;Y^s\|Z)$ |
|---|---|---|
| Linear-Gaussian | $\dim(\boldsymbol{X})\dim(\boldsymbol{Y})$ | $\dim(\boldsymbol{X})\dim(\boldsymbol{Y})$ |
| Discrete (plug-in) | $(\|A_X\|-1)(\|A_Y\|-1)$ | $(\|A_X\|-1)(\|A_Y\|-1)\|A_Z\|$ |

- Where:
  - dim(**X**) means the number of dimensions in a multivariate **X**
  - $|A_X|$ means the alphabet size of discrete variable X.
  - Discrete estimate is converted to nats in the distribution! (but back to bits in JIDT)

L. Barnett and T. Bossomaier, Physical Review Letters 109, 138105+ (2012)
J.T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics & AI, 1:11, 2014; appendix A.5

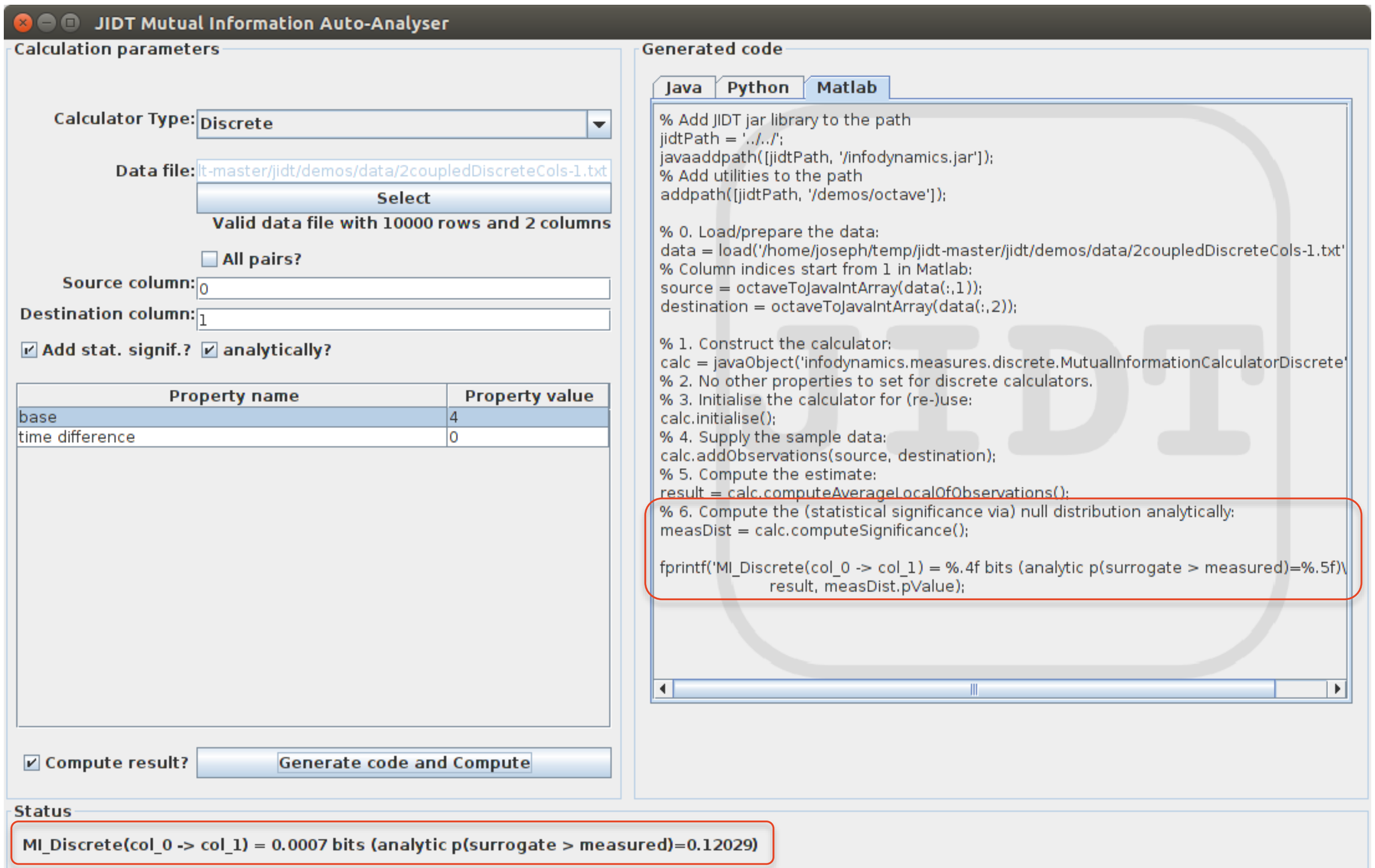
# Analytic statistical significance test in JIDT

– This is far faster than empirical surrogate generation.
– Generate from AutoAnalyser by clicking the checkbox next to “analytic?” near “Add stat. signif.?” (when available)



# Analytic statistical significance test in JIDT

- The `getSignificance()` method returns an `AnalyticMeasurementDistribution` object (see Javadocs) from which you can retrieve p-value of statistic, and convert between estimates ↔ p-values.



# Analytic statistical significance

- Pros:
  - far faster than empirical
- Cons:
  - Is only completely correct asymptotically as $N \rightarrow \infty$. (But we only care about it when N is finite!)
  - Can be significantly away from empirical values when:
    - Distributions under analysis are highly multivariate (increasing undersampling effects), or
    - (for discrete estimator) where the distributions on the variables are heavily skewed.

- More details in `demos/octave/NullDistributions` (see on wiki)

# Q2 – how many samples do we need?

- Well, that depends on the question you want to answer … ☺
- To detect statistically significant relationship?
  - Depends on strength. Less samples required for stronger relationship
- To avoid undersampling? Depends on estimator
  - Heuristic : have $\geq 3\times$ as many samples as possible state configurations

| | |
|---|---|
| x=0<br>y=0 | x=1<br>y=0 |
| x=0<br>y=1 | x=1<br>y=1 |

  - For I(X;Y) there are $|A_X| \times |A_Y|$ total state configurations
    - E.g. for binary variables, there are $2\times2$ state configurations.
  - The number of state configurations increases as:
    - The variables have more discrete levels / larger alphabet size, or
    - The variables become multivariate. (Which is equivalent)
  - This assumes all possible state configurations are equally likely to be visited…

M. Lungarella, T. Pegors, D. Bulwinkle, and O. Sporns, "Methods for quantifying the informational structure of sensory and motor data," Neuroinformatics, vol. 3, no. 3, pp. 243–262, 2005.
J.T. Lizier, "*The local information dynamics of distributed computation in complex systems*", Springer: Berlin/Heidelberg, 2013. Section 3.3.1

# Q2 – how many samples do we need?

- Well, that depends on the question you want to answer … ☺

- But for large multivariate spaces, only a subset of the state configuration space may be explored:
  - The “typical set” of state configurations is where the “sample entropy is close to the true entropy” of that joint state.
  - Think of as set of state configurations likely to be encountered frequently enough to contribute to that entropy.
  - This is the set we need to sample well enough, with the number of samples $N \geq 3\times$ (or in general $\geq M\times$) the size of the typical set.
  - Good expressions for size of typical set for block entropy / entropy rate (see Information storage session)

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Chapter 3
K. Marton and P. C. Shields, “Entropy and the consistent estimation of joint distributions,” The Annals of Probability, vol. 22, no. 2, pp. 960–977, 1994
J.T. Lizier, “*The local information dynamics of distributed computation in complex systems*”, Springer: Berlin/Heidelberg, 2013. Section 3.3.1

# Q2 – how many samples do we need?

– Well, that depends on the question you want to answer … ☺

– That's easy enough to work with for plug-in discrete estimator.
– You can adapt it for box-kernel (heuristically).
– KSG adapts the bin width to avoid undersampling in general, but may miss subtleties in relationship.

J.T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics and AI, 1:11, 2014; appendix B.2.b
J.T. Lizier, "*The local information dynamics of distributed computation in complex systems*", Springer: Berlin/Heidelberg, 2013. Section 3.3.1

# Statistical significance : summary

- We've reviewed estimation of information as a statistic, and how to test for significance of that statistic:
  - Empirically, and
  - Analytically where possible.
- You know how to apply statistical significance testing of MI, CMI etc., using JIDT

- *Coming up*: Information processing in complex systems.

# Questions