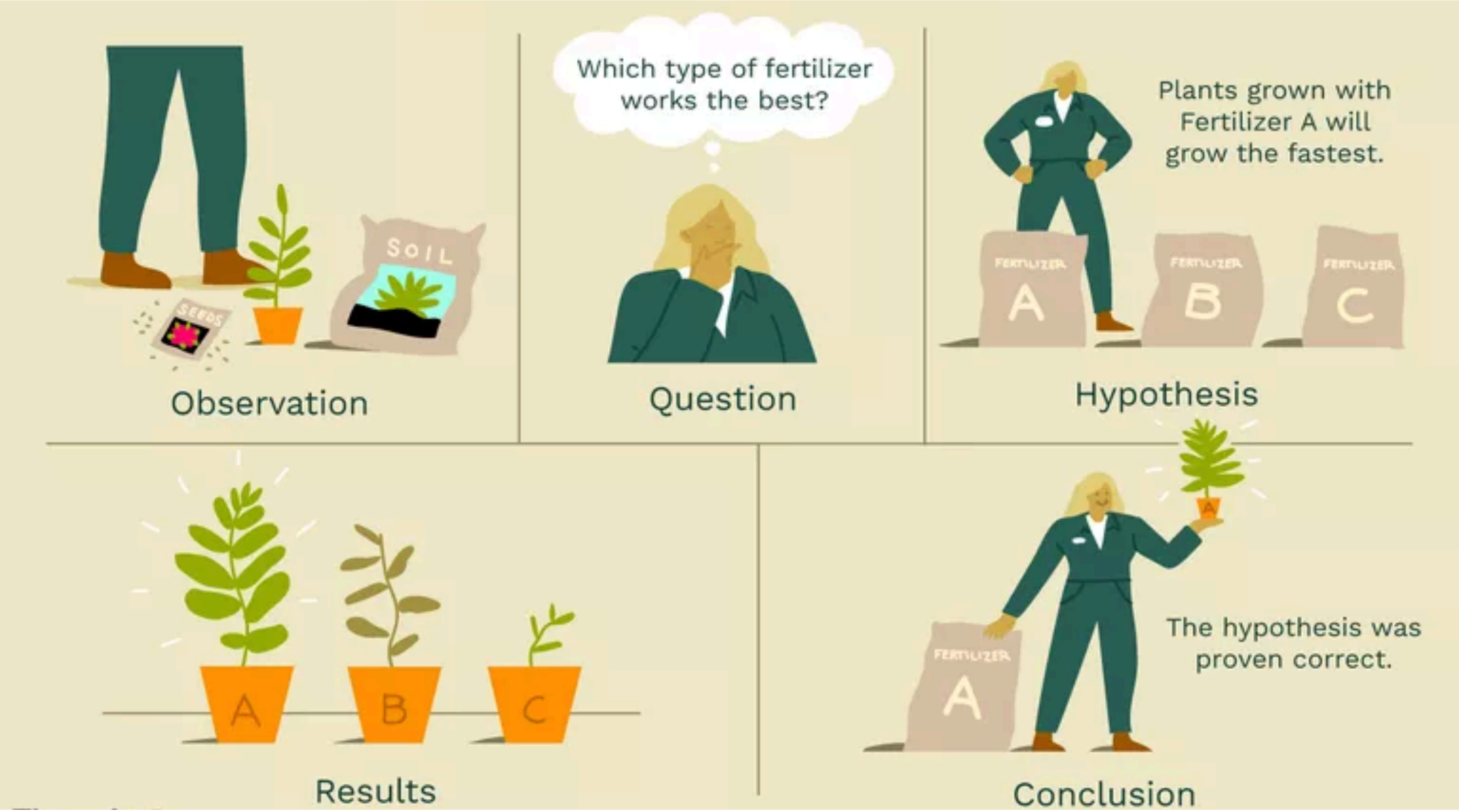
Testing Hypotheses

The role of Hypotheses in the Scientific Method



saverio.giallorenzo@gmail.com

The role of Hypotheses in Quantitative Studies

In quantitative studies, hypotheses are **statistical hypotheses**, i.e., they are **testable by considering the observed data as values taken from a collection of random variables**.

The difference between the two models (the actual and random one) is deemed statistically significant if, according to a threshold probability, the actual data is unlikely to have occurred randomly.

The hypothesis that the two models are comparable is called **null hypothesis**.

If we can reject (through testing) the null hypothesis, we have ground for believing that there is a (dependence) relationship between two or more phenomena represented by the data. This second hypothesis (or class of hypotheses) is called **alternative**.

Towards testing for correlation in network studies

The term “**correlation**” refers to the degree to which a pair of variables are linearly related. Correlations are useful because they can indicate a predictive relationship - not necessarily a causal one.

To test the hypothesis of correlation, we measure a set of variables (e.g., two) on a set of cases drawn via a probability sample from a population and then calculate a **correlation coefficient** between the variables.

If the correlation measure found in the sample appears in rare cases with respect to all possible random configurations - usual measures are 5%, 1% or 0,1% - we have a high degree of confidence in rejecting the null hypothesis (that the two variables are independent) and accept the alternative (that they are correlated).

The most common of correlation coefficients is **Pearson’s coefficient.**

Testing Hypotheses in Network Analysis

Applying correlation tests over network data has some peculiarities, mainly because standard (statistical) tests make assumptions about the data which are violated by network data.

E.g., standard tests assume that the observations are statistically independent, which, in the case of adjacency matrices, they are not. To visualise this, consider that all the values along one row of an adjacency matrix pertain to a single node. If that node has a special quality, such as being very antisocial, it will affect all of their relations with others, introducing a lack of independence among all those cells in the row.

Another typical assumption of standard tests is that the variables are drawn from a population with a particular distribution, such as a normal distribution. However, many network population distributions are not normal.

Testing Hypotheses in Network Analysis

The main solutions to solve the problem of statistical testing of network hypotheses are two.

The usage (and development) of statistical models specifically designed for studying the distribution of ties in a network (e.g., exponential random graph models and actor-oriented longitudinal models).

The usage of a generic methodology of randomisation/permutation tests and use them with (modified) standard methods, like regression. These methods are easy to use and interpret, and can be customised for different research questions - this is the path followed by UCINET.

Permutation Tests

Suppose we believe that a teacher favours tall kids. Hence, we hypothesise that height and exam scores in this teacher's classes are (positively) correlated.

To test our hypothesis, we let the teacher give all the students a test, measure their height, and then correlate the two variables. Let us say we get a moderate (Pearson's) correlation coefficient of 0.384.
Is that number significant to asses our hypothesis or have we been just lucky and obtained a rare random configuration?

To remove the uncertainty, we might let the teacher give and grade a statistically-relevant number of tests (e.g., 100) and then perform our assessment with much more certainty.

Permutation Tests

Unfortunately, that procedure takes time.

E.g., if the teacher had three different classes (a reasonable amount of classes to teach to) and tested the students once per month (a reasonable amount of time for them to learn some new concepts to test), our study would take ~3 years (on a 10-months/year schedule).

Still, we have concerns due to path dependence (knowledge debt of students) and we need to wait a lot of time for the study to be finished.

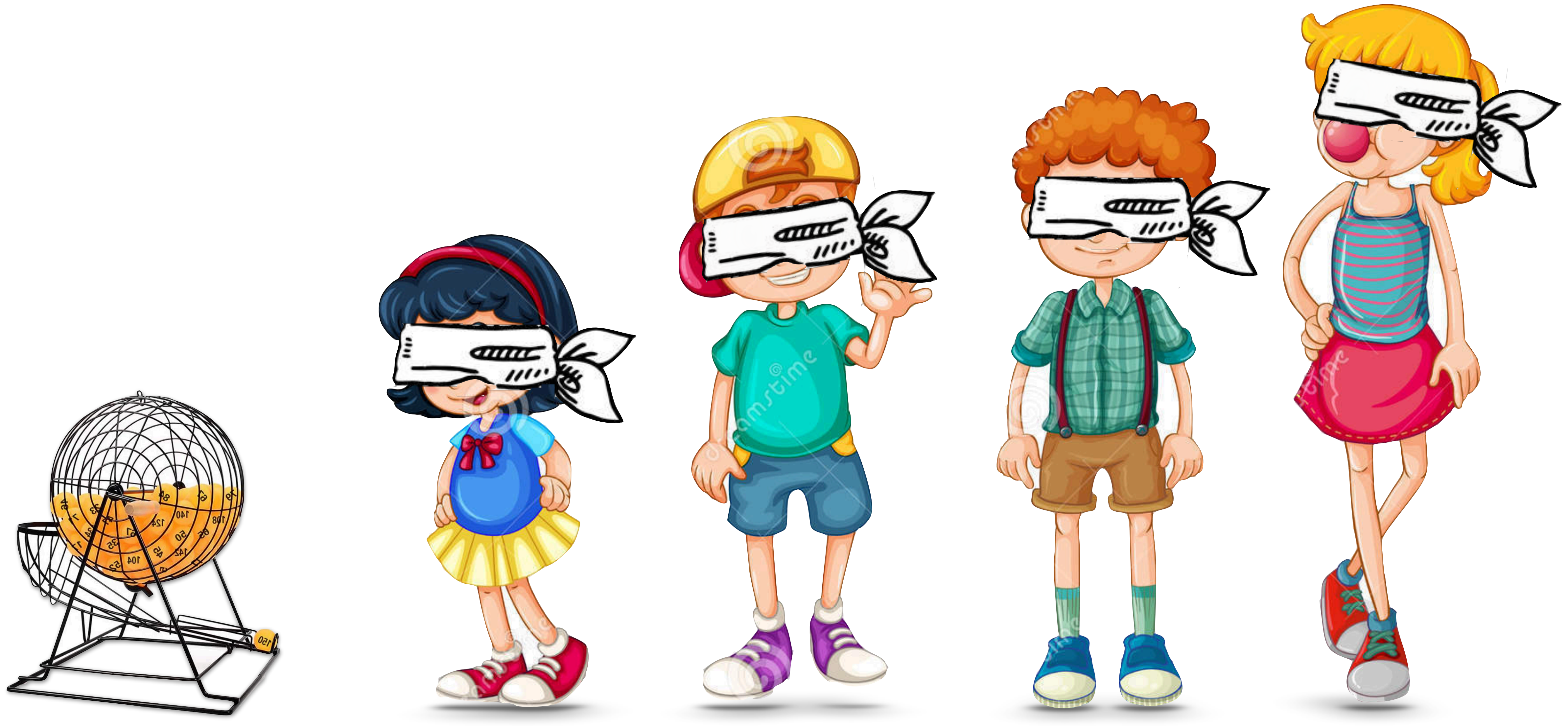
Is there a cheaper (time-, resource-wise) solution?

Let us make a thought experiment!

Permutation Tests

Instead of giving them tests, we let the students draw blindly (i.e., randomly) their scores from a box.

Then we ask "could it happen, by chance alone, that most of the high scores happened to go to the tall people? Unlikely, but it - that *random 0.384 correlation coefficient - could happen.*



How much unlikely? We can calculate that by just repeating e.g., 100 times the blind-drawing experiment and calculate how many times we got that 0.384 correlation (and above).

That would let us resolve a problem that would have taken 3 years into a couple of days (without considering path-dependency issues).

Permutation Tests

To find our solution, we essentially reversed the problem.

Instead of looking for certainty of correlation (the ~3-year study), we looked for certainty of non-randomness of the correlation.

E.g., if the percentile of cases in which we find those correlation coefficients is high (e.g., 20%) we cannot reasonably conclude that the teacher is biased.

Permutation Tests

In permutation tests computers simulate millions of blind-drawings quickly and give us enough data to sufficiently test our hypotheses (of non-randomness).

A permutation test essentially calculates many (if not all) the ways that the experiment could have come out, given that scores were actually independent of height, and **counts the proportion** of random assignments yielding a correlation as large as the one actually observed.

That proportion is called the **p-value** or significance of the test and estimates the degree of confidence in rejecting the null hypothesis - usual measures are 5%, 1% or 0,1%.

Permutation Tests, a simple(r) example

E.g., we want to study whether lung diseases have a correlation with smoking.

Null hypothesis: there is no correlation between smoking and lung diseases, i.e., they are independent

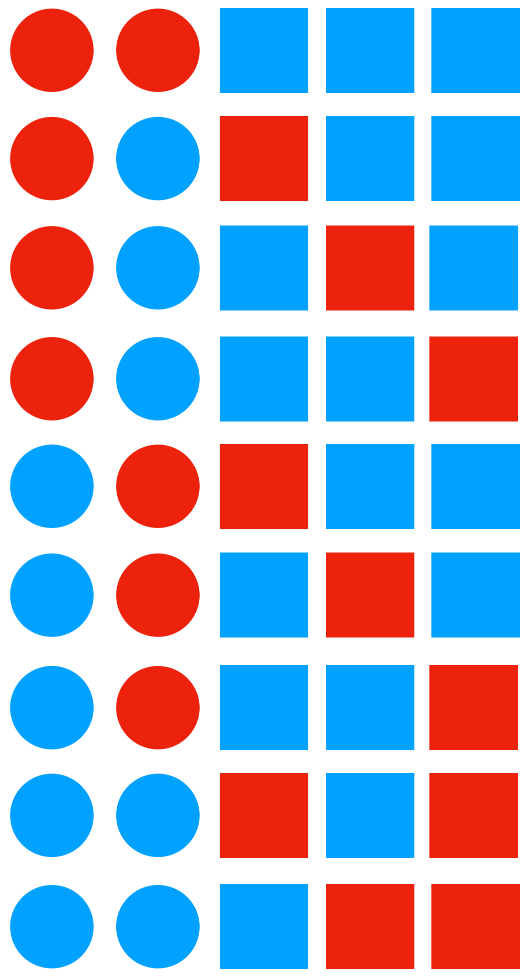
Alt. hypothesis: there is a correlation between smoking and lung diseases

Alternative Observation:

0.6

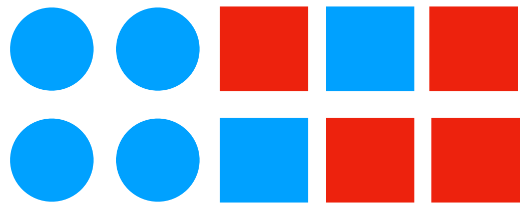
Legend: Red, has disease. Square: smoker

(Random) permutations:



-1
-0.1
-0.1
-0.1
-0.1
-0.1
-0.1
0.6
0.6

Permutations that show the same ratio as the sample:



The likelihood of observing a correlation coefficient equal to the observed one over all (random) permutations is 3/10 = 30%.

We cannot reject the null hypothesis (sample size problem).

Network Hypotheses

There are four main levels of analysis for network hypotheses:

- **monadic** hypotheses, e.g, “central people/nodes tend to be happier”, are the closest to non-network ones and the correlation stands between a characteristic of the node (e.g., centrality) and another (e.g., happiness);
- **dyadic** hypotheses consider relations among pairs (dyads) of nodes, e.g., “the shorter the distance between two people’s offices, the more they communicate over time”, and are normally organised as $N \times N$ matrices we correlate;
- **mixed** hypotheses consider the relation between monadic and dyadic properties, e.g., “does the gender (monadic property) of a person affect whom that person is friends (dyadic) with?”. Oftentimes these kinds of hypotheses are tested as dyadic by rephrasing the monadic property;
- **network-level** hypotheses consider the structure of the network with respect to some variable, e.g., “does the structure of the network determine the time a group finds a solution?”, with the cases being whole networks instead of nodes.

Monadic Hypotheses

Regression testing do not make assumptions on the independence of variables and can be used to test monadic hypotheses of network data.

Regression coefficients.

		1	2	3	4	5
		Coef	Beta	SE	T	c.Sig
		------	------	------	------	------
1	Intercept	26.986	0.000	12.449		
2	#PRIORS	0.601	0.422	0.345	1.740	0.104

The algorithm:

1. performs a standard multiple regression across the corresponding cells of the vectors;
2. randomly permutes the elements of the dependent vector and recomputes the regression, storing the resulting values;
3. Repeat 2. until it can properly estimate the standard errors for the statistics of interest. For each coefficient, we count the proportion of random permutations that yielded a coefficient as the one computed in 1.

Dyadic Hypotheses

The Quadratic Assignment Procedure (QAP) is a technique designed to correlate whole matrices and it is frequently used to test dyadic hypotheses.

Briefly, the technique tests the correlation between two matrices by reshaping them into thousands of random-yet-similar pairs.

To do that it takes the data matrices and randomly rearranges their rows (and matching columns). Because this is done randomly, the resulting matrices are independent from the data matrix they came from but they preserve the properties of the original (e.g., mean, standard deviation, number of edges, number of cliques, etc).

To obtain the p-value, QAP counts the proportion of the random configurations that have a correlation coefficient as large as the one observed.

```
Matrix: QAP Correlations

                1     2
            PADGB PADGM
            ----- -----
  1 PADGB   1.000 0.372
  2 PADGM   0.372 1.000

Matrix: QAP P-Values

              1  2
             PA PA
             DG DG
              B  M
             -- --
  1 PADGB   0.000 0.001
  2 PADGM   0.001 0.000
```

Monadic-to-dyadic Hypotheses

The standard approach to testing the association between a node attribute and a dyadic relation is to convert the problem into a purely dyadic hypothesis by constructing a dyadic variable from the node attribute.

There are two main techniques to do this conversion, depending on whether the monadic attribute is scalar or nominal.

- scalar attributes, e.g., age, wealth, height, can be converted into dyadic attributes as a comparison between the values of the two nodes. E.g, A is 24 years old and B is 43, then the value of the “age” relation between them is 19.

- nominal attributes, e.g., role, gender, political affiliation, can be converted into dyadic attributes by exact match, e.g., A and B are “clerks” while C is a “manager”, then A and B have a relation (the match) while C has no relations with neither A nor B.

Whole-network Hypotheses

We can use normal correlation tests where we can make the (statistics) classical assumption that the networks are obtained via a random sample.

If that assumption falls, we can resort to permutation tests and contrast the correlation coefficient between the groups/networks and the dependent variable.

While permutation tests can help in assessing the significance of observations in a given (non-random) sample, the results derived from the latter are not general as the one of the random one.

The permutation results obtained from a non-random sample are significant for that sample, however, since we do not represent the whole population we lost generality—e.g., that teacher favours tall students, but we cannot extend that observation to other teachers.