Hypothesis Testing

Jeremy Straughter

CASOS Summer Institute
June 2020

Carnegie Mellon
Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Learning Outcomes

- Understand and formulate testable network hypotheses
- Gain an intuition for permutation testing and its use
- Test hypotheses using the ORA software

What do we mean by testing hypotheses?

- Padgett and Ansell (1993). Robust Action and the Rise of the Medici, 1400 – 1434
- Collected data on the relationships between Florentine families during the Renaissance
  - Marriage Ties
  - Business Ties
- The testable hypothesis might be that economic transactions are embedded in social relations
  - How would we test this?

Marriage

1	ACCIAIUOL	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	ALBIZZI	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0
3	BARBADORI	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
4	BISCHERI	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
5	CASTELLAN	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0
6	GINORI	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	GUADAGNI	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1
8	LAMBERTES	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9	MEDICI	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1
10	PAZZI	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
11	PERUZZI	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
12	PUCCI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	RIDOLFI	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
14	SALVIATI	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
15	STROZZI	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0
16	TORNABUON	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0

Business

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0
4	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
5	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0
6	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
7	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0
9	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1
10	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Carnegie Mellon

We may be interested in testing hypotheses at various levels of analysis

- Node-level hypotheses
- Dyadic hypotheses
- Mixed dyadic-monadic hypotheses
- Whole-network hypotheses

Carnegie Mellon

Could we use a standard statistical package?

- Unable to test this hypothesis using standard statistical packages
- Most packages are set up to correlate vectors and not matrices
- The significance tests in most packages make assumptions which are violated when using network data
  - independence among variables
  - variables drawn from a particular distribution
  - random variables

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Special Methods for Testing Hypotheses

- Develop statistical models specifically designed for studying the distribution of ties in a network
  - Exponential Random Graph Models
  - Stochastic Actor-Oriented Longitudinal Models
  - Complex models beyond the scope of this presentation
- Permutation Tests
  - Easy to use and interpret
  - Customizable for different research questions

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Let's briefly review classical significance testing…

- Based on sampling theory
- Measures a set of variables (e.g., two variables)
  - we're interested in the relationship between the variables
- Significance tells us the probability of obtaining a result that large given given that in the population the variables are independent
  - when this probability is low (less than .05) we call it statistically significant (i.e., we claim that the variables are related in the population)
  - when this probability is higher, we fail to reject the null hypothesis

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

The logic of permutation tests differ from standard statistical tests

- For example, suppose you believe we favor tall people and scores in this course are correlated with height
  - variables of height and score (correlation is .384)
- Now suppose we write down a set of math scores and have each student draw a score blindly from a hat
- What proportion of all the ways scores could be pulled would result in a correlation as large as our observation
- Compare the observed correlation against the distribution of correlations

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Permutation Tests

- The permutation test calculates all the ways that an experiment could have come out given the variables were in fact independent
- Counts the proportion of all assignments yielding a correlation as large as the one observed
  - this proportion indicates the 'p-value' or significance
- The number of permutations of N objects grows very quickly with N
- We sample uniformly from the space of all possible permutations (~20,000 permutations)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Permutation Tests

Row/ Col	1	2	3	4
1	Y1,1	Y1,2	Y1,3	Y1,4
2	Y2,1	Y2,2	Y2,3	Y2,4
3	Y3,1	Y3,2	Y3,3	Y3,4
4	Y4,1	Y4,2	Y4,3	Y4,4

Row/ Col	1	2	3	4
1	Y3,3	Y3,2	Y3,4	Y3,1
2	Y2,3	Y2,2	Y2,4	Y2,1
3	Y4,3	Y4,2	Y4,4	Y4,1
4	Y1,3	Y1,2	Y1,4	Y1,1

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Quadratic Assignment Procedure (QAP)

- QAP correlation is designed to correlate entire matrices
- To calculate the significance, the method compares the observed correlation to a reference set of thousands of correlations
- To construct a p-value, it counts the proportion of the correlations that were as large as the observed correlation
- Compare the observed correlation against the distribution of correlations

Quadratic Assignment Procedure (QAP)

- QAP regression allows us to model the values of a dyadic dependent variable using multiple independent variables
  - multiple regression (MR-QAP)
  - logistic regression (LR-QAP)
- Practical Examples
  - Florentine Families
  - Congressional Voting

QAP/MRQAP ANALYSIS REPORT

Input data: padgett

Start time: Tue May 19 16:35:43 2020

Data Description

Parameters

Dependent meta-network	padgett
Dependent data	Network: PADGB
Number of independent networks	1
Random seed	0
Number of permutations	2000
Diagonal values used	false

The table below describes how the dependent and indepedent variables were constructed. The variable labels Y, X1, X2, etc. are used consistently throughout the report.

Variable	Variable Meta-Network	Variable Description
Y	padgett	Network: PADGB
X1	padgett	Network: PADGM

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Correlation (Dependent to Independent)

This shows the correlation and related statistics between the dependent network variable and each independent network variable.

Significance for Pearson Correlation is the fraction of trial bootstrap values that are higher than the actual.

Significance for Hamming and Euclidean Distance is the fraction of trial bootstrap values that are lower than the actual.

The input networks are all binary valued, and therefore the Hamming distance was computed.

Variable	Variable Meta-Network	Variable Description	Correlation	Significance	Hamming Distance	Significance
X1	padgett	Network: PADGM	0.372	5.000e-04	19	5.000e-04

The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 2000

			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
X1	Correlation	0.372	-0.169	0.372	-0.002	0.091	5.000e-04	1
X1	Hamming Distance	19	19	35	30.068	2.701	1	5.000e-04

Hamming distance: treat networks as data strings and calculate the difference between the networks. ORA uses Hamming Distance for binary matrices and Euclidean Distance for matrices with continuous values.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Regression Results

Reports the results from the regression. There are three computations for standard errors: the classical formula is reported in column Std.Errors; heteroskedasticity robust standard errors are reported in column Robust Std.Errors; finally, bootstrapped standard errors are reported in column Bootstrapped Std.Errors.

Model: b0 + b1*X1

Model Fit	
Observations	120
R-Squared (R2)	0.138
Residual Sum Of Squares	11.310
Total Sum Of Squares	13.125
Standard Error	0.310

Variable	Coef	Std. Coef	Std. Errors	Robust Std.Errors	Bootstrapped Std.Errors	Sig.Y-Perm
Intercept	0.070	0	0.031	0	0.014	1
X1	0.330	0.372	0.076	0	0.082	0

R-Squared and Standard Error are both goodness-of-fit measures for linear regression models.

R-Squared indicates the percentage of the variance in the dependent variable that the Independent variables explain collectively (~13.8%)

Standard error measures the precision of the model's prediction – the standard distance between the observations and the regression line (~.31%)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Correlation (Dependent to Independent)

This shows the correlation and related statistics between the dependent network variable and each independent network variable.

Significance for Pearson Correlation is the fraction of trial bootstrap values that are higher than the actual.

Significance for Hamming and Euclidean Distance is the fraction of trial bootstrap values that are lower than the actual.

At least one input network has non-binary link values, and therefore the Euclidean distance was computed.

Variable	Variable Meta-Network	Variable Description	Correlation	Significance	Euclidean Distance	Significance
X1	Congress	Network: Age Difference	7.002e-04	0.560	188.475	0.560
X2	Congress	Network: Same Committee	0.406	0.026	6.782	0.026
X3	Congress	Network: Same Gender	-0.094	1	6.928	1
X4	Congress	Network: Same Party	0.316	0.045	5.477	0.045

The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 2000

			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
X1	Correlation	7.002e-04	-0.682	0.412	8.368e-04	0.217	0.560	0.455
X1	Euclidean Distance	188.475	187.433	190.192	188.474	0.548	0.455	0.560
X2	Correlation	0.406	-0.279	0.482	0.004	0.144	0.026	0.998
X2	Euclidean Distance	6.782	6.481	9.055	8.179	0.470	0.998	0.026
X3	Correlation	-0.094	-0.094	0.756	-0.004	0.115	1	0.389
X3	Euclidean Distance	6.928	3.464	6.928	6.634	0.395	0.389	1
X4	Correlation	0.316	-0.063	0.316	-0.002	0.083	0.045	1
X4	Euclidean Distance	5.477	5.477	6.782	6.582	0.283	1	0.045

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Correlation Matrix

This shows the correlation between all network variables.

	Y	X1	X2	X3	X4
Y	1	7.002e-04	0.406	-0.094	0.316
X1		1	-0.215	0.114	0.136
X2			1	0.221	0.056
X3				1	-0.120
X4					1

Regression Results

Reports the results from the regression. There are three computations for standard errors: the classical formula is reported in column Std.Errors; heteroskedasticity robust standard errors are reported in column Robust Std.Errors; finally, bootstrapped standard errors are reported in column Bootstrapped Std.Errors.

Model: b0 + b1*X1 + b2*X2 + b3*X3 + b4*X4

Model Fit	
Observations	90
R-Squared (R2)	0.280
Residual Sum Of Squares	14.404
Total Sum Of Squares	20
Standard Error	0.412

Variable	Coef	Std. Coef	Std. Errors	Robust Std.Errors	Bootstrapped Std.Errors	Sig.Y-Perm
Intercept	5.747e-04	0	0.107	0	0.169	0.977
X1	0.003	0.081	0.004	0	0.010	0.408
X2	0.340	0.447	0.075	0	0.123	0.005
X3	-0.162	-0.171	0.092	0	0.137	0.977
X4	0.246	0.260	0.090	0	0.091	0.044

Carnegie Mellon

isr institute for SOFTWARE RESEARCH

What you should know…

- Understand the difference between traditional data and network data
- Understand the different hypotheses that you can formulate from network data
- Understand the logic behind permutation tests and have an intuition for how ORA performs them
- Perform a QAP/MRQAP analysis in ORA
- Interpret the results of the QAP/MRQAP Analysis Report