William Cipolli
55 University Ave.
Hamilton, NY 13346
☎ *+1 (203) 848 5643*
✉ *will@cipolli.com*
www.cipolli.com

Research Statement

Currently, I am an Assistant Professor of Mathematics at Colgate University. My methodological research focuses on Bayesian Nonparametric approaches to a variety of problems including multiple testing, density estimation, and supervised learning. More specifically, I aim to address key questions related to assumptions in parametric approaches by envisioning new, flexible solutions that are computationally efficient and widely available. My work has been published in journals like Statistics and Computing, and Computational Statistics and Data Analysis.

My future work in supervised learning includes a nonparametric classification schema which has been submitted to Advances in Classification and Data Analysis. I am currently working with collaborators to extend this methodology for prediction and feature selection using Householder reflections to make the approaches computationally efficient. Due to the remarkable success of these approaches in applications to real data, I'm currently working on implementing these methods into R packages which would make these theoretical approaches widely available.

I also complete many collaborations by providing simulation techniques to other statisticians and quantitative expertise to those in other sciences including Sociology, Psychology, Biology and the Medical field. These interdisciplinary collaborations involve translating qualitative hypotheses to a quantitative, actionable solutions through careful experimental design and model building. These solutions provide precise estimates and compelling evidence for such impactful theories about the world. These works have been published in top journals of Statistical Simulation, Sociology, Psychology, Biology and Medicine.

Research in Progress

The Score Test for a Zero-inflated Bivariate Poisson Distribution (Completing initial simulations) – *A collaboration with Roy Bower*

This work will extend our first collaboration *The Score Test for a Bivariate Poisson Distribution* to score tests for testing independence and zero inflation in Lakshminarayana's bivariate Poisson distribution. The parameters of the bivariate distribution are estimated using maximum likelihood methods where appropriate and we will conduct a variety of simulation studies to show that the score test maintains a significance level close to the nominal one.

Prediction and Feature Selection for High Dimensional Data through Smoothed Polya Trees(Theoretical work complete)

This work will extend the *Supervised Learning for via Smoothed Polya Trees* paper to the prediction scenario. We aim to propose a new and computationally efficient way of fully smoothing multivariate Polya trees using Householder Reflections in place of computationally expensive Givens Rotations, which require a significant MCMC effort. This will yield a flexible, nonparametric density estimate from which we can make predictions while possibly improving our original classification scheme with additional smoothing.

Computationally Tractable Approximate and Smoothed Polya Trees for Accelerated Failure Time Models with Heteroskedastic Error (To be started Summer 2018)

This work will extend a model proposed in *Computationally Tractable Approximate and Smoothed Polya Trees* to the case where errors are heteroskedastic. This makes the model more flexible as it can be applied even when variability is unequal across groups in a study.

The Decline of the Islamic State: An Empirical Test of Organizational Squeeze (Data collection and literature review in progress) – *A collaboration with Derek Silva*

This work will propose a quantitative model of "squeeze theory" on ISIS; e.g. we expect more events in US and Europe when we bomb, take more land back, etc. While sociologists and other researchers have completed qualitative research, very little has been done to quantitatively measure this theory. We've reached out to sources around the world, including researchers at the United Nations for data as we work toward building this model. Many researchers expect that the squeeze theory should be quite clearly quantifiable and are excited for the results.

The Score Test for a Bivariate Pareto Distribution (Debugging Simulation Code) – *A collaboration with Roy Bower*

Two score tests are proposed: one for testing independence based on Sankaran and Nair's bivariate Pareto distribution and one for testing whether Sankaran and Nair's parameterization reduces to the more popular bivariate Pareto distribution introduced by Lindley–Singpurwalla. The marginal distributions of both bivariate parameterizations are univariate Pareto II distributions, and the parameters of the bivariate distribution are estimated using numerical methods. The simulation studies show that both score tests maintain a significance level close to the nominal size. To check the efficiency of the derived score tests, the estimated significance level and power of the likelihood ratio and Wald tests are also compared. One real-world data set is used to illustrate the application of both score tests.

“It doesn’t happen here" – An inquiry into sexual assault on female only college campuses (Starting R1 preparation) – *A collaboration with Kris Macomber and Juliette Grimmett*

We plan to apply for an NIH R1 grant to conduct a nationwide survey across thirty-seven women’s colleges and universities in the United States. This inquiry is designed to illuminate the campus climate in terms of sexual violence. While many campuses say “it doesn’t happen here" many are in proximity to large state schools, have partner institutions or a reputation as a “suitcase school" some of these institutions provide little or no education about sexual assault or support for victims of sexual violence. In addition to making same-sex couples invisible, it leaves vulnerable students without important services. We hope to elucidate what happens on these campuses and start a new line of research in sexual violence as findings are likely to require further exploration including possible connections to experiences in women’s high schools.

Introductory Statistics with R (Editing first draft) – *A collaboration with Josh Tebbs and Roy Bower*

In 2014 the American Statistical Association (ASA) and the Mathematical Association of America (MAA) released a report Guidelines for Assessment and Instruction in Statistics Education, which outlined significant updates to the professional associations’ recommendations for teaching introductory statistics at the college level. This text explores these topics through published research across disciplines in an inquiry-guided manner, leading students toward critical statistical thinking and numeracy. Shifting the classroom towards statistical thinking allows students to ask questions statisticians can answer. Incorporating R in the classroom allows students to answer such questions. Allowing students to see the entire life cycle of statistical analysis gives them experience in creating knowledge and critically evaluating and consuming information. This text innovates in STEM education by pushing the boundary of an introductory class by putting faith in the technological skills of the modern student.

Authored Research

Cellular metabolism and oxidative stress as a possible determinant for longevity in small breed and large breed dogs. (In revision at PLOS One) – *A collaboration with Ana Jimenez*

Among species, larger animals tend to live longer than smaller ones, however, the opposite seems to be true for dogs - smaller dogs tend to live significantly longer than larger dogs across all breeds. We were interested in the mechanism that may allow for small breeds to age more slowly compared with large breeds in the context of cellular metabolism and oxidative stress. Primary dermal fibroblasts were grown in tissue culture from small and large breed dogs as puppies and seniors. Our data suggest that as dogs of both size classes age, basal respiration, and proton leak is significantly higher in older dogs, regardless of size class. We found that all aspects of glycolysis were significantly higher in larger breeds compared with smaller breeds. We found significant differences between age classes in GSH concentration, and a negative correlation between DNA damage and mean breed lifespan in puppies, each of these parameters showing a significant increase in older dogs, particularly in the larger breeds. Interestingly, RS production showed no differences across size and age class. Thus, large breed dogs may have higher glycolytic rates, and DNA damage, suggesting a potential mechanism for their decreased lifespan compared with small breed dogs.

Supervised Learning for via Smoothed Polya Trees (In revision at Advances in Data Analysis and Classification)

Many classification approaches make distributional assumptions about the data, most infamously the Gaussian distribution. We can create a more flexible approach by generalizing the Gaussian assumption by assuming the nonparametric multivariate Polya tree prior. The flexibility gained from relaxing the distributional assumptions from our analysis can prove paramount when even minor deviations from the distributional assumptions occur; the ability of the Polya tree approach to pick out even slight deviations could significantly improve classification over a model that is assumption-heavy. The outcomes obtained by this model can be compared to other methods including decision trees, k-nearest neighbor, naive Bayes, artificial neural networks, support vector machines, etc.

On The Distribution of Monochromatic Complete Subgraphs and Arithmetic Progressions (Submitted to Experimental Mathematics) – *A collaboration with Aaron Robertson*

We consider the distribution of the number of monochromatic complete subgraphs over edgewise 2-colorings of complete graphs as a type of statistical Ramsey theory. We also consider the analogous for monochromatic arithmetic progressions. We present convincing evidence that both distributions are very well-approximated by the family of Delaporte distributions.

Score Test for Independence Using the Normal Copula and Poisson Marginals (In Revision at Communications in Statistics: Case Studies and Data Analysis) – *A collaboration with Roy Bower*

An increasingly popular approach to model the dependence between random variables centers on the use of copula functions. A variety of bivariate copula families are explored, in particular the Guassian copula. Furthermore, a score test for testing independence of response variables is proposed for the specific case where the marginal distributions are known to be Poisson. The simulation study shows the test keeps the significance level close to the nominal one. Similarly, the estimated significance level and power of the likelihood ratio and Wald tests are also compared to show our test is numerically stable. A real-world data set is used to demonstrate the application of our test.

The Score Test for a Bivariate Poisson Distribution (Submitted to Communications in Statistics Part B: Case Studies and Data Analysis) – *A collaboration with Roy Bower*

A score test for testing independence in Lakshminarayana's bivariate Poisson distribution is proposed. The marginal distributions of the bivariate model are the univariate Poisson distributions, and the parameters of the bivariate distribution are estimated using maximum likelihood methods. The simulation study shows that the score test maintains a significance level close to the nominal one. To check the efficiency of the derived score test, the estimated significance levels and powers of the Likelihood Ratio and Wald tests are also compared. A relevant data set is used to demonstrate the application of the bivariate Poisson model and score test with success.

The paradox of group mind "People in a group" have more mind than "a group of people"(Published in the Journal of Experimental Psychology: General) – *A collaboration with Erin Cooley*

Three studies examine how subtle shifts in framing can alter the mind perception of groups. Study 1 finds that people generally perceive groups to have less mind than individuals. However, Study 2 demonstrates that changing the framing of a group from "a group of people" to "people in a group," substantially increases mind perception—leading to comparable levels of mind between groups and individuals. Study 3 reveals that this change in framing influences people's sympathy for groups, an effect mediated by mind perception. We conclude that minor linguistic shifts can have big effects on how groups are perceived—with implications for mind perception and sympathy for mass suffering.

In Search of a Five-Star: The Centrality of Body Discourses in the Scouting of High School Football Athletes(Accepted at Sociology of Sport Journal) – *A collaboration with Derek Silva*

This project explores the question of how scouting reports represent the body, skills, and physical characteristics of the subject. The hypothesis is that body discourses focus predominantly on physical characteristics of the athlete In other words; we predict that representations related to size and strength are more important than more football-specific characteristics. It is also hypothesized that scouting 'experts,' here conceptualized as analysts working for each major media outlet, are rather arbitrary with their subjected player analyses tend to be more important than more objective measurements such as listed weight, height, and calculated body mass index (BMI). To explore these questions, we have attempted to identify the most important predictors of high rated athletes – typically referred to in this context as "five-star" prospects.

Bifactor Models of the Strengths and Difficulties Questionnaire in a Large U.S. Community Sample (Presented at the annual meeting of the American Psychological Association - Denver) – *A collaboration with Kate Flory*

This study analyzed 25-item screenings of children and adolescents ages 3-16 obtained in a CDC funded study. The best-fitting model was a new bifactor model where all factors are allowed to correlate which is an improvement on the bifactor model in existing research literature.

Computationally Tractable Approximate and Smoothed Polya Trees (Published in Statistics and Computing)

This work introduces a discrete approximation to the Polya tree prior that enjoys surprisingly simple and efficient conjugate updating. This approximation is illustrated via simulation and in two applied contexts: the implementation of a nonparametric meta-analysis involving studies on the relationship between alcohol consumption and breast cancer and random intercept Poisson regression for Ache armadillo hunting times. The discrete approximation Dirac measures are then replaced with Gaussian densities to provide a smoothed mixture of Polya trees that can be used in standard contexts; the smoothed approximation is illustrated on density estimation of the eruption times of the Old Faithful geyser.

Bayesian Nonparametric Multiple Testing (Published in Computational Statistics & Data Analysis)

Multiple testing, or multiplicity, problems often require testing several means with the assumption that we will reject infrequently, as motivated by the need to analyze DNA microarray data. The goal is to keep the combined rate of false discoveries and non-discoveries as small as possible. We propose a discrete approximation to a Polya tree prior that enjoys fast, conjugate updating, centered at the usual Gaussian distribution, thus generalizing Scott and Berger (2006) to a nonparametric setting. This new technique and the advantages of our approach are demonstrated using extensive simulation and data analysis accompanied by a Java web application. The numerical studies demonstrate that our new procedure shows promising FDR and estimation of key values in the mixture model with very reasonable computational speed.

A Comparison of Three Diagnostic Tests for Diagnosis of Carpal Tunnel Syndrome Using Latent Class Analysis (Published in The Journal of Bone and Surgery) – *A collaboration with John Fowler*

The current reference standard for carpal tunnel syndrome is under debate. Recent studies have demonstrated similar diagnostic accuracy between ultrasound and nerve conduction studies. Given the lack of a universally accepted reference standard for carpal tunnel syndrome, latent class analysis is an established method to determine the true diagnostic accuracy of these tests. The purpose of this study is to determine sensitivity and specificity of ultrasound (US), nerve conduction studies (NCS), and CTS-6 for diagnosis of carpal tunnel syndrome (CTS) using latent class analysis. We used latent class models to estimate the sensitivity and specificity of these three diagnostic tests in the absence of a gold standard. All models were fitted in WinBUGS, including identifiable models regression carpal tunnel syndrome status on age and gender.

Publications

A. Jimenez, J. Winward, U. Beattle, and W. Cipolli, “Cellular metabolism and oxidative stress as a possible determinant for longevity in small breed and large breed dogs.” In revision at PLOS One., 2017.

W. Cipolli and T. Hanson, “Supervised learning via smoothed Polya trees.” In revison at Advances in Data Analysis and Classification, 2017.

A. Robertson, W. Cipolli, and M. Dascalu, “On the distribution of monochromatic complete subgraphs and arithmetic progressions.” Submitted to Advances in Experimental Mathematics., 2017.

R. Bower, J. Hussey, J. Zhang, J. Quattro, M. Muhling, W. Cipolli, and J. Hardin, “The score test for independence of two marginal Poisson variables.” Submitted to Communications in Statistics - Case Studies and Data Analysis., 2017.

R. Bower, J. Hussey, J. Zhang, J. Quattro, W. Cipolli, and J. Hardin, “A copula approach for testing independence using Poisson cumulative distribution functions.” In revision at to Communications in Statistics - Case Studies and Data Analysis, 2017.

E. Cooley, B. Payne, W. Cipolli, C. Cameron, A. Berger, and K. Gray, “The paradox of group mind: “people in a group" have more mind than “a group of people",” *Journal of Experimental Psychology: General*, vol. 146, pp. 1691–699, May 2017.

D. M. Silva, W. Cipolli, and R. Bower, “In search of a five-star: The centrality of body discourses in the scouting of high school football athletes.” Accepted at Journal of Sport and Social Issues, 2017.

K. Flory, B. A. Bell, K. Burgess, E. R. Siceloff, W. Cipolli, and R. Bower, “Bifactor models of the strengths and difficulties questionnaire in a large U.S. community sample,” Presented at the annual meeting of the American Psychological Association, Denver, CO, 2016.

W. Cipolli and T. Hanson, “Computationally tractable approximate and smoothed Polya trees,” *Statistics and Computing*, pp. 1–13, April 2016.

W. Cipolli, T. Hanson, and A. McLain, “Bayesian nonparametric multiple testing,” *Computational Statistics and Data Analysis*, vol. 101, pp. 64–79, September 2016.

J. Fowler, W. Cipolli, and T. Hanson, “A comparison of three diagnostic tests for diagnosis of carpal tunnel syndrome using latent class analysis,” *Journal of Bone & Joint Surgery*, vol. 97, pp. 1958–1961, December 2015.

Professional Consulting

1/2018 – Scoring Internet of Things (IoT) Solutions

This International Data Corporation (IDC) study explores creating an IDC score for IoT solutions based on various attributes of the solution. The study describes each of the criteria used in the scoring, as well as a mathematical formula for scoring, that creates an automated way of choosing between IoT solutions based on utility.

01/2017 – Black Groups Accentuate Hypodescent by Activating Threats to the Racial Hierarchy

One reason White people categorize Black-White Biracial people as Black (called hypodescent) is to maintain the existing racial hierarchy. By creating a strict definition of who can be White, the selectivity, and thus status, of White people increases. Given that racial hierarchies are about the relative status of groups, we test whether perceiving Black groups increases hypodescent by activating fears about shifts in the racial hierarchy (i.e., a majority/minority shift). Indeed, these studies showed that White people rated and stereotyped Black-White Biracial people as more Black in Black groups (but not White groups; than when alone. Critically, this pattern was driven by White people relatively high in fear of a majority/minority shift or those experimentally led to feel this threat. Researchers were able to conclude that Black groups increase hypodescent by activating fears about shifts in the racial hierarchy, posing consequences for racial stereotyping.

10/2016 – The Othering of Muslims: Discourses of Radicalization in the New York Times, 1969–2014

Through the lens Edward Said's Orientalism and various perspectives within the othering paradigm the emergence and transformation of radicalization discourses in the news media is analyzed. Employing discourse analysis of 607 New York Times articles from 1969 to 2014, this article demonstrates that radicalization discourses are not new but are the result of complex sociolinguistic and historical developments that cannot be reduced to dominant contemporary understandings of the concept or to singular events or crises. The news articles were then compared to 850 government documents, speeches, and other official communications. The analysis of the data indicates that media conceptualizations of radicalization, which once denoted political and economic differences, have now shifted to overwhelmingly focus on Islam. As such, radicalization discourse now evokes the construct radicalization as symbolic marker of conflict between the West and the East. This work advances the established notion that the news media employ strategic discursive strategies that contribute to conceptual distinctions that are used to construct Muslims as an “alien other" to the West.

03/2014 – Judging Appropriateness of SUP Using Patient Demographic Data

This methodology explains how to record the dataset, fit a model and interpret the effects on the inappropriateness of Stress Ulcer Prophylaxis (SUP) using patient demographic data and other pertinent information. The analysis provided uses an imitated dataset - the SAS code and analysis produced are easy to read and to extend beyond the following variables: age, weight, gender, race, whether they have medical history of specialty care, prior use of concomitant antibiotics, prior use of proton pump inhibitor or H2 antagonist, severity of illness . When or if other variables are added, the statistical approach could stay the same and the added variables would be interpreted similarly.

02/2014 – Veteran Affairs Resource Utilization by Patients with HFpEF

Patients with heart failure with a preserved ejection fraction (HFpEF) represent a large portion of the heart failure population. The goal of the research is to identify differences in healthcare resource utilization by patients with HFpEF who are managed in the primary care setting compared to those receiving care in cardiology clinics. We document a statistical approach for a mock dataset created to closely match the description provided. The results shown are from the mock dataset and therefore aren't meaningful – they are only shown to describe the methodology that can be followed when the data becomes available to the client. The methodology explains how to fit a model and interpret the effects on the duration of stay using their age, gender, race, type of clinic, other conditions and smoking status. When other variables are added, the statistical approach could stay the same and the added variables would be interpreted similarly.

02/2014 – Attitude towards Homelessness Questionnaire

A statistical approach for modeling the data from surveys, which intimate sentiments towards homeless people, was provided. It is shown that the visual and textual treatments in the study lead to a slightly positive change in score, however, this difference was not significant across visual and textual treatments. The statistical insignificance of controls previously shown to be relevant in past studies is another result of interest – this sample did not hold the same biases as the past experiment allowing for further exploration on the causes for such a difference.

06/2013 – The G20 through the Internet of Things (IoT) Lens

This International Data Corporation (IDC) study outlines the methodology and results for ranking the G20 countries according to their readiness for the Internet of Things (IoT) opportunities. The study describes each of the criteria used in the ranking, as well as the weighting scheme applied, that creates the ultimate ranking of the more ready versus less ready countries for IoT opportunities. The rank-ordered G20 countries are presented as the study result. IDC's G20 Index presented through the lens of the IoT reveals just how far and above the top-tier countries are as opportunities for vendors in the space. The top-tier countries collectively astound with their share of investment in key technologies and energy consumption. These countries are at the intersection of the need for the efficiencies that IoT solutions bring and are more likely to have a mindset for and the availability of technologies that align and intersect with IoT use cases.