ALLOY INFERENCE: TESTS OF A SINGLE NULL

A PREPRINT

Gary J. Cornwall*

Office of the Chief Economist

U.S. Bureau of Economic Analysis

Gary.Cornwall@bea.gov

August 18, 2025

ABSTRACT

This paper presents a new joint testing framework that fuses multiple test statistics into a single, more powerful inference tool. Using the probability integral transform to confine the support to the unit-hypercube, I use simulated null cases and Archimedean copulas to approximate the underlying joint null distribution of two or more statistics. Analogous to an alloy in metallurgy, where the final product has [typically] stronger properties than its constituent parts, I show how two or more tests can be combined to outperform a single test statistic in finite samples. To illustrate the performance of this approach, I provide a stylized example using the game of craps such that trade-offs can be be assessed in economic terms. Under potential uncertainty in the fairness of game dice, the proposed method–a combination of the Student-t and χ^2 statistic–provides increased power, producing a revenue distribution which second-order stochastically dominates its constituent parts.

1 Introduction

Empirical research relies on the results of hypothesis tests to evaluate falsifiable statements about the world through observable data. After formulating a research design and a testable hypothesis, the dominant paradigm in statistical analysis is to employ a single test statistic (e.g., a t-statistic) to draw inferences. This has two well-known forms of error Neyman and Pearson (1933): a failure to reject the null when it is false (Type II) and a rejection of the null when it is true (Type I). If one were to choose a different test statistic, one that tests the same null hypothesis directly or a

*The results and opinions are those of the author(s) and do not reflect those of the U.S. Department of Commerce or the U.S. Bureau of Economic Analysis. I would like to thank Scott Wentland, Jeremy Moulton, Marina Gindelsky, Beau Sauley, and Ben Bridgman for helpful comments and suggestions on prior drafts. I would also like to thank participants at the 2024 meeting of the Midwest Econometrics Group for their invaluable feedback.

variation thereof, not only may the probability of Type II errors change–convention fixes the probability of a Type I– the resulting inferences may disagree with respect to evidence against the null. Hence, the convention in applied statistical work is for the researcher to proceed with a single test statistic, usually determined by the prevailing norms in their field or subfield. This raises a key question: rather than making a choice to test a null with a single statistic, can we combine multiple test statistics such that we see improved performance?

In this paper, I answer this question by explicating a new framework for testing a null hypothesis using a collection of test statistics, comparing the performance of this "alloy test statistic" against its individual parts and other composite inference approaches.[2] I find that this new approach does, in fact, substantially outperform traditional individual test statistics in finite sample settings. It combines information from multiple testable statements, each reflective in their own way of the research question, to maximize, from the observed data, information upon which inferences are drawn and reduce error. The proposed alloy method shares a common antecedent with advances in the forecast combination literature (see Wang et al. (2023) for a recent review), which leverage non-perfectly overlapping information from different models to enhance prediction accuracy.[3] What is less clear from recent literature is precisely *how* improvements in prediction and error-reduction are achieved in so-called 'black box' methods that combine non-overlapping information sets. Thus, a parallel objective of this paper is to shed new light on how methods like this achieve better outcomes that traditional statistical tools.

I begin with a common pedological example in statistics–the fairness of a six-sided die–and show how joint evaluation of the Student's t-statistic and a χ^2 statistic can outperform either one individually. Beyond its simplicity, the example also highlights a relevant aspect of applied statistical research, where a researcher is interested in a broader, more fundamental question (e.g., is a die fair? or, does a time series have a unit root?), and individual test statistics are used to evaluate variations of this 'fundamental null' differently.[4] Consistent with probability theory, I argue that the null distributions of these test statistics are marginal representations of some larger, but unknown, probability space. Like forecast combinations, each statistic can provide imperfectly overlapping information about a fundamental null hypothesis. Thus, inferences from the joint distribution of two (or more) test statistics provide more information than

[2]Contextually it is important to note that approaches like the harmonic p-value Wilson (2019) are designed not for multiple tests addressing the same null statement, but rather are used in the case where many different null statements are being tested simultaneously.

[3]For example, Chen and Cornwall (2021) and Chen et al. (2021) explored this question in a high dimensional setting, using supervised machine learning techniques to differentiate null and alternative cases. Using an agnostic definition of seasonality – a concept for which there is no uniform, agreed upon definition – Chen and Cornwall (2021) used a number of test statistics as features in a Random Forest framework to differentiate seasonal from non-seasonal time series finding that – when combined – the collection of test statistics outperformed any individual statistic over the parameter space. In similar fashion, Chen et al. (2021) examined the use of multiple statistics in the case of unit root testing, a more well-defined problem than seasonality, finding an increase in power over a single test alternative. In an earlier version of this paper, the authors also showed that, given statistics calculated from a simulated data set a boosting algorithm (Schapire and Freund, 2013) recovers the full Receiver Operating Characteristic Curve of the test statistic itself. This implies that these algorithms are approximating the null distribution and the joint distribution of any alternatives they are exposed to.

[4]More generally, this approach makes explicit what the typical applied statistical analysis entails: a statement of interest, a measurement or set of measurements that reflect that statement of interest, and testable statements based on the behavior of those measurements. Although individual test statistics are, strictly speaking, testing different null hypotheses (i.e., t-statistic and χ^2 statistic in the dice example below are different), I describe in more detail in the next section how these can collectively evaluate a fundamental null hypothesis in practice.

using either one individually.[5] The results below demonstrate improved ability of joint testing to detect deviations from the fundamental null hypothesis while maintaining a fixed probability of a Type I error.

Finally, if we can improve hypothesis testing for a simple example like determining the fairness of a die, is the improvement economically meaningful? To explore this, I turn to an application of the fair die example, where probabilities are well-established and payoffs are known, such that we can assess trade-offs in terms of dollar values. More concretely, in Section 3 I provide a stylized example based the generation of economic activity through casino craps, where the existence of undetected unfair dice reduces a casino's revenue from the game. Generating roughly 450 million dollars in revenue in Las Vegas alone (Yakowicz, 2023), craps is not only a popular game, but also one which contains even odds, or near even odds bets.[6] In a hypothetical scenario where casinos face uncertainty about the fairness of their dice, I simulate a 'horse race' of casinos using different test statistics to detect the fairness of dice in finite samples. Casinos that leverage joint or alloyed testing will better detect, and close significantly more, unfair tables–conditional upon a fixed testing scheme–than either of the individual inputs. As a result, the distribution of revenue from the remaining tables second order stochastically dominates those produced by either the Student-t or χ^2 test approach. Moreover, an alloyed test's improvement over a single hypothesis test is greater when samples are relatively small, which is common in many sub-fields of economics and applied statistical analysis.

2 Framework

Let us take a step back and be more concrete in the in the problem setup and establish parameters of a simple example and along with some notation. Suppose we have some population $X \sim F(x)$ from which we obtain a sample $\mathcal{D} = v_1, \ldots, v_n$. For now it is assumed that $\mathcal{D}$ is a purely random sample with no measurement issues, and that the population, X, is arbitrarily large or infinite in size such that $n/N \to 0$ as $N \to \infty$. The goal is to draw inference about some aspect of X from $\mathcal{D}$.

For concreteness, suppose $X \sim \text{Multinomial}(\text{n}; \pi_1, \ldots, \pi_\text{K})$ where K is the total number of categories and the probabilities satisfy $\sum_{i=1}^{K} \pi_i = 1$. We might be interested in testing whether $\mathcal{D}$ is consistent with the hypothesis that $\pi_i = 1/K$ for all $i \in \{1, \ldots, K\}$, which represents an assumed structure of X. How we decide to test this statement depends on a number of factors including, but not limited to, the sample size of $\mathcal{D}$, which measures of $\mathcal{D}$ are relevant for assessing the hypothesis, and our knowledge of available statistical methods leveraging these measures.[7]

To be even more specific, consider the following. Suppose $k \in \{1, \ldots, 6\}$ and that $X \sim \text{Multinomial}(\text{n}; \pi_1, \ldots, \pi_6)$, where in truth, $\sum_{i=1}^{6} \pi_i = 1$ with $\pi_i = 1/6$ for all i. An astute reader will recognize this data-generating process

[5] It is well-known that entropy in joint distributions is bounded from above by the sum of the marginal entropies with equality holding only when the margins are independent (MacKay, 2003). This implies that observing a set of outcomes provides more information in a joint setting where the margins are not deterministic functions of one another.

[6] For example, in a pass-the-line bet the win probability is 0.4929 with a house take of approximately 1.4%. Other forms of betting at the table such as "Odds Bets" have even odds

[7] This list is by no means comprehensive and merely is illustrative of some of the factors that serve as inputs to drawing statistical inference through hypothesis testing. Other factors to consider include the specific alternative of interest, additional sampling issues such as sample bias or measurement error, and possible clustering or dependence among sample observations, to name a few.

as that of a fair, six-sided die – one of the most commonly used pedagogical examples in probability and statistics. Further, suppose two students are given the same sample, $\mathcal{D}$, drawn from X. Since the interest lies not in the specific finite-sample behavior but rather the choice of statistical test, it is assumed $n = 100$ for concreteness.

Each student considers how to test whether the die is fair based on the tools they are familiar with. Student A recognizes that one approach is to count the occurrences of each outcome and compare them to the expected frequencies using a chi-squared statistic. On the other hand, Student B is unfamiliar with the chi-squared statistic, but recalls that the expected outcome of a fair die is 3.5 and identifies the Student-t test as an appropriate evaluation method.

Student A obtains the vector $(17, 14, 13, 11, 29, 16)$ corresponding to the counts of each outcome and uses this to calculate a chi-squared statistic of 12.320 with corresponding p-value of 0.031 and rejects the null hypothesis that each outcome occurs with equal probability at the 5% significance level. On the other hand, using that same sample, Student B calculates the sample mean of 3.690 and sample standard deviation of 1.756, leading to a Student-t statistic of 1.082 with corresponding p-value of 0.282. Student B thus fails to reject the null statement at the 5% significance level.This scenario illustrates how different statistical methods, despite operating on the same data and attempting to address the same fundamental question, can produce different conclusions. This motivates the need for a joint evaluation framework that accounts for multiple statistical perspectives simultaneously.

In this, admittedly trivial, example there is a right answer. The sample was in fact generated from a distribution with uniform probabilities over the six outcomes. It should be relatively clear that the chi-squared test performed by Student A is both necessary and sufficient to determine the uniformity of the associated probability vector. However, it should also be clear that in (smaller) finite samples there is a great deal of variability introduced. Conversely, the Student-t test conducted by Student B is only a necessary condition to determine the uniformity of the associated probability vector. It is relatively trivial to think of a non-uniform probability vector that would produce the same mean as a uniform one. Ultimately, the nuance of this example is that the amount of evidence, supplied by the sample, is sufficient to reject the null hypothesis of only one of these tests.

This raises an important question: if a researcher had access to both tests, how should they decide between them? On the one hand, the chi-squared test is evaluating a necessary and sufficient condition of uniformity but is plagued by slow convergence and [potentially] poor finite sample properties. On the other hand, the Student-t test is not actually testing the statement of interest, rather it is evaluating a derivative of it, the expected outcome. A purely dichotomous approach – where a single statistic or p-value dictates the conclusion – overlooks the fact that each test provides complementary information about the underlying hypothesis.

In order to consider joint inference between the two tests, let H_b^0 be a general statement about the world under investigation. From the example above this would be: "This six-sided die is fair". Following convention the alternative will be defined as H_b^1; "This six-sided die is unfair". Note that H_b^0 and H_b^1 on their own do not imply a specific measurement; rather, they serve as foundational statements one wishes to consider. For now, focus will be on null hypothesis significance testing, so a specific alternative will generally not be specified going forward.

With this fundamental statement in mind, testable statements are constructed conditional upon specific assumptions ($\mathcal{A}$) and available measurements ($\mathcal{M}$). I will denote this testable statement as H_j^0 with $j = 1, \ldots, m$ indexing the testable statements derivable from the fundamental statement. Given the assumptions, measures, and any additional constraints facing the researcher (*i.e.,* sample size, time-to-decision, etc.) a test statistic is constructed. This test statistic, $x_j = f(\mathcal{M}_j|\mathcal{D})$, is a function projecting measures of observed data identified as pertinent to H_j^0 onto $\mathbb{R}$, or some subset thereof. For now, I will refer to the support of the statistic as I_j^0. The null distribution with density $p_j(x_j|H_j^0 = \text{True})$ is a probabilistic representation of the testable null statement over this support, and together with H_j^0 forms the tuple $\langle\, H_j^0,\, I_j^0,\, p_j(x_j)\,\rangle$. Broadly speaking x_j is evaluated under the null to calculate the probability one would observe a statistic at least as extreme as what was observed, conditional upon the null being true, commonly known as a p-value.

Returning to the case of our two students, both received a sample $\mathcal{D}$ representing n rolls of the die. Student A is presented with the tuple,

$$\langle\, H_2^0,\, I_2^0,\, p_2(x_1)\,\rangle \begin{cases} p(v_1 = 1) = \ldots = p(v_6 = 6) = 1/6 \\ \mathbb{R}^+ \\ \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{\frac{-x}{2}} \end{cases} \tag{1}$$

where

$$x_1 = \sum_{j=1}^{6} \left(\frac{(\sum_{i=1}^{n} 1(v_i = j) - n/6)^2}{n/6} \right) \tag{2}$$

is the well-known chi-squared statistic. While on the other hand, Student B is presented with,

$$\langle\, H_1^0,\, I_1^0,\, p_1(x_1)\,\rangle \begin{cases} \bar{v} = \mu_0 \\ \mathbb{R} \\ \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}} \end{cases} \tag{3}$$

where

$$x_2 = \frac{n^{-1}\sum_{i=1}^{n} v_i - \mu_0}{\sqrt{\frac{(n-1)^{-1}\sum_{i=1}^{n}(v_i - \mu_0)^2}{n}}} \tag{4}$$

is the well-known Student-t statistic.

Recall that the statement being addressed is the "fairness" of the six-sided die from which the sample was derived. The measures, outcome frequencies and expectation, are both margins upon which fairness is judged. This is concept is important as it leads directly to Assumption 1.

Assumption 1 *There exists some joint probability distribution,*

$$p(x_1, x_2|H_b^0 = \textit{True})$$

, which represents our probabilistic assessment of the null, H_b^0. The null distribution of each testable statement is the marginalization of this joint distribution with respect to all other statements. That is:

$$p(x_1|H_1^0 = \textit{True}) = \int_{\mathbb{S}} p(x_1, x_2|H_b^0 = \textit{True})dx_2, \tag{5}$$

where $\mathbb{S}$ indicates the support, I_2^0.

This assumption, though seemingly non-trivial in this context, is actually guaranteed by the rules of probability. Even independent random variables can be expressed as a joint distribution through the product of their marginals.

This is not to say we know what $p(x_1, \ldots, x_s|H_b^0 = \text{True})$ is, or even that it is analytically tractable. For context, suppose the die that produced $\mathcal{D}$ had been rolled $n = 3$ times instead of the 100 mentioned earlier. If one were to enumerate all possible sample vectors, 216 in this case, one would find that there are 50 unique Student-t statistics and only 3 unique chi-squared statistics. From there, it is trivial to show that the resulting frequency table, when divided by the number of possible vectors (216) forms a valid joint probability table for the Student-t and chi-squared statistics. One can then quickly verify that the product of the marginal distributions does not equal the joint and thus these test statistics are not independent. The joint distribution conceptually captures the full extent of information available about the fundamental null. In other words, each testable null hypothesis is informed by one or more measures of the data, and these measures are derived from a margin of the larger data generating process. As a result, the test statistic represents a probabilistic distribution over that margin, conditional on the testable null and the assumptions about the measurement.

In small sample, discrete outcome settings it may be possible to fully enumerate the outcome space and thus directly identify the joint distribution of two or more test statistics. However, this complete enumeration quickly becomes infeasible as the number of possible outcomes and/or the sample size increases. Moreover, even in discrete problems we often compare observed, small sample test statistics to the corresponding asymptotic continuous distribution rather than its discrete counterpart (*e.g.,* any Student-t statistic in the $n = 3$ case would be compared to the [continuous] Student-t distribution with three degrees of freedom). Instead, the joint distribution can be approximated using work by Sklar (1959) which states that, for any joint probability distribution $p(x_1, x_2)$ there exists a unique copula, $c(u_1, u_2)$ such that:

$$p(x_1, x_2) = c(u_1 = P_1(x_1), u_2 = P_2(x_2))p_1(x_1)p_2(x_2), \tag{6}$$

subject to some regularity conditions on the marginal distributions (*e.g.,* they are continuous, Nelsen (2007)). We observe both $p(x_1|H_1^0 = \text{True})$ and $p(x_2|H_2^0 = \text{True})$ and since Assumption 1 says that $p(x_1, x_2)$ must exist, then all that is needed is to identify an appropriate copula, $c(u_1, u_2)$.

Since conditions under which $H_b^0 = \text{True}$ are known, I propose we identify the appropriate copula through simulation. That is, for a stated fundamental null, a large set of synthetic data is constructed from which the appropriate test statistics are calculated. It is well-known that numerical approximations of the joint distribution with $\hat{f}(u_1, \ldots, u_m) \rightarrow$

$f(u_1, \ldots, u_m)$ as the number of simulations approaches infinity.[8] This approach is often used in cases where a continuous null distribution is not analytically tractable (*e.g.,* deriving the critical values for tests of unit roots in time series, see MacKinnon (2010) for one such example).

Returning to the example of our two students, in Figure 1a I provide a simulated representation of the chi-squared and Student-t statistics conditional upon H_b^0 : "This six-sided die is fair". Each point is a pair, (t, χ), representing the calculated test-statistic from a sample of $N = 100$ rolls from one fair six-sided die. A total of one million six-sided die were rolled to create the collection of points you see and the corresponding contours are estimates from these points.[9] The dashed lines correspond to marginal critical values used to evaluate the statistic at $\alpha = 0.05$.

Before going further and attempting to identify the appropriate copula, Figure 1a highlights two problems with evaluating a joint space. First, consider the support of the joint distribution which in this case is $\mathbb{R} \times \mathbb{R}^+$. It is nearly impossible to saturate this space in a simulated environment since we are unlikely to visit extreme outcomes in a feasible number of iterations. Second, the form of the test on the margin – one-tailed versus two-tailed – creates multiple, and possibly disjoint, rejection regions. In the example provided above, the rejection regions (where both tests reject their corresponding null statement) is the rectangles in the upper right and left corners respectively. However, given a decision boundary along the margin, we also have three disjoint "disagreement zones", areas where one test would indicate the sample is consistent with that of a fair die and the other inconsistent.

I contend that bounding the space in the unit hypercube by first transforming each marginal distribution into its corresponding p-value solves both of these issues. This transformation not only limits the support to $U(0, 1)^d$ but also standardizes the decision boundaries for each margin to some threshold value, chosen by the researcher, near the origin. Figure 1b represents this transformation and is a numeric approximation to the joint distribution of p-values. Note that in this case the margins are known to be uniformly distributed since the testable statement is true in both cases and these statistics are well-behaved.

Table 1: Correlation of Simulated Statistics

ρ / τ	t	χ^2	p_t	p_{χ^2}
t		**-0.001**	**0.000**	**0.001**
χ^2	-0.003		**-0.257**	**-1.000**
p_t	0.000	-0.349		**0.257**
p_{χ^2}	0.001	-0.930	0.371	

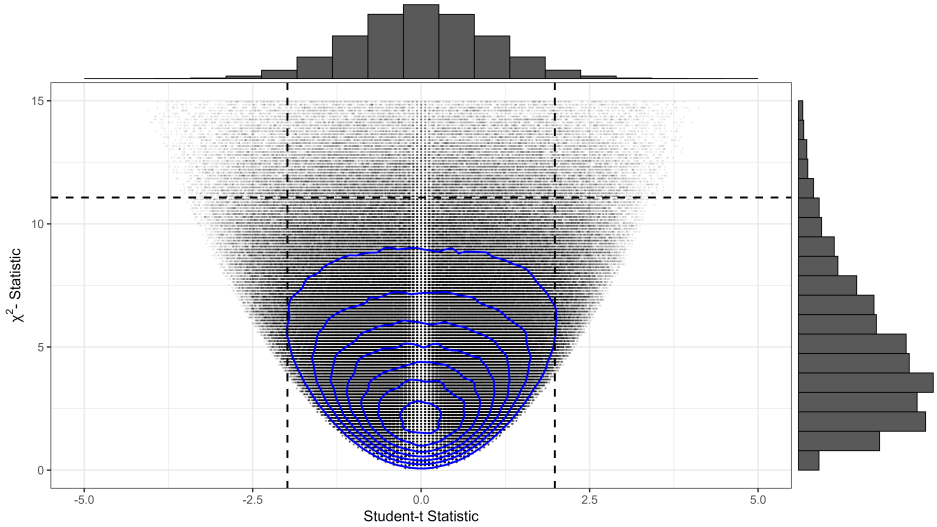
Note:

Evident from this transformation is the shape and dependence of the joint distribution have changed. In Table 1 the Pearson's ρ and Kendall's τ correlation coefficients are provided in the lower and upper triangular elements respectively.

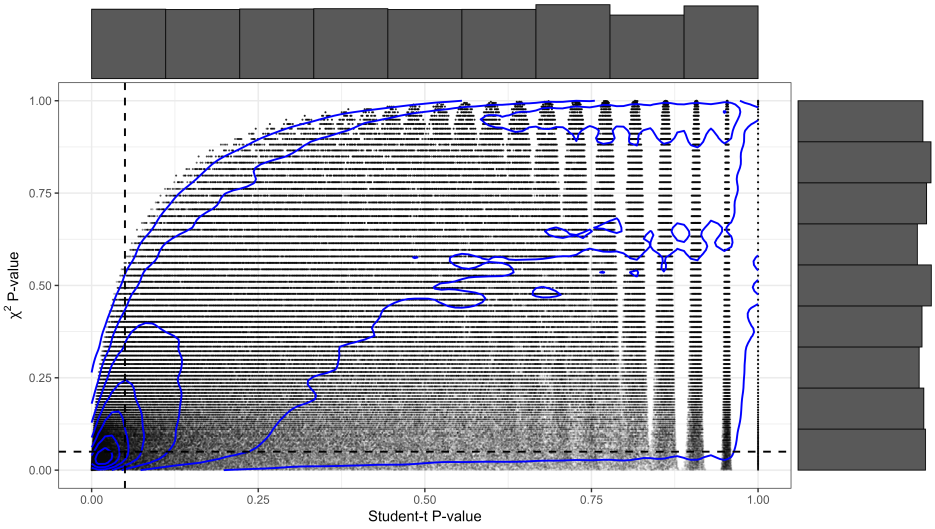
[8] It is important to note that in a discrete setting like the die example, the joint distribution is also discrete. There is a limited number of possible means for samples, of size n, obtained from a single, m-sided die, $m + (n-1)(m-1)$ and a corresponding limited number of variances. By definition this means that there are a fixed number of t-statistics, and by extension chi-squared statistics. To reiterate the points made earlier, my use of a copula in the context of a discrete finite example is ultimately an estimate of the true underlying probability distribution represented in continuous fashion.

[9] The striations in the figure result from the fact that, in finite samples, there are a finite number of means and variances possible when sampling from a fair six-sided die and thus a finite number of both t-statistics and χ^2 statistics.

Figure 1: Numeric Approximation of the Joint Distribution



(a) Each point is a pair of statistics calculated from a sample of 100 rolls from a fair, six-sided die. The dashed lines indicate marginal thresholds at the $\alpha = 0.05$ level. Contour lines are estimated from the collection of statistics and the marginal histograms are provided.



(b) Each point is a pair of p-values calculated the statistics in Figure 1a. The dashed lines indicate marginal thresholds at the $\alpha = 0.05$ level. Contour lines are estimated from the collection of statistics and the marginal histograms are provided.

The intuition here is that two or more testable statements under the fundamental null will produce p-values with higher tail-dependence in the joint space. This tail-dependence will vary in location based on the decision rule imposed in the p-value calculation. For example, two positively dependent test statistics each with a left-tailed decision would produce additional density near the $(0, 0)$ and $(1, 1)$ vertices of the unit square.[10] Transforming the marginal distributions into their corresponding p-values also has the added benefit of fitting the properties of Archimedean copulas and their generalizations (e.g., a Tawn copula (Tawn, 1988)) which combine uniform marginals with a strictly decreasing and convex generator function, $\psi(t; \theta)$ (Nelsen, 2007). This leads to Assumption 2.

Assumption 2 *Assume that for a given fundamental null hypothesis H_b^0, each test statistic x_j is transformed via its corresponding continuous and strictly increasing cumulative distribution function, F_j, into a p-value $u_j = F_j(x_j)$, which is uniformly distributed on the interval $[0, 1]$ under H_b^0. Then, there exists a copula $c : [0, 1]^m \to [0, \infty)$ from the Archimedean family, or a suitable generalization thereof, with generator function $\psi : [0, \infty) \to [0, 1]$ and dependence parameter $\theta \in \Theta \subseteq \mathbb{R}$, such that the joint distribution of the p-values can be expressed as:*

$$p(u_1, u_2, \ldots, u_m | H_b^0) = c(u_1, u_2, \ldots, u_m | \theta). \tag{7}$$

The generator, $\psi(t; \theta)$ is assumed to be strictly decreasing and convex, satisfying $\psi(0) = 1$ and $\lim_{t\to\infty} \psi(t) = 0$.

Leveraging the p-value transformation, this assumption serves two purposes. First, it states that under these conditions the joint distribution of the transformed test statistics is characterized by the copula alone. Second, the p-value transformation standardizes the rejection area based on the decision structure. Given that the Archimedean family is well-known to provide flexible representation of tail dependence behavior, it follows that such a copula is well-suited to model the joint distribution under H_b^0. Moreover, given an observed sample, the estimation of Archimedean copula in both the bivariate and higher dimensional cases through parametric estimators is well-developed (see Genest and Rivest (1993); Nelsen (2007); Genest et al. (2009); Joe (2014) for example).

Once again let us return to the single die example provided to our two students. Assuming the die is fair, it is trivial to simulate data from any arbitrary sample size and estimate the appropriate Archimedean copula. Using the simulations illustrated in Figures 1a and 1b, Table 2 outlines measures of fit for different Archimedean copulas and their rotations.[11] Made clear from this table is the relative fit of a Tawn Copula (denoted as Type II, more on this in a moment) rotated $180°$. Figure 2 provides a visual comparison of the fit with respect to contours of the simulated rolls of one million fair, six-sided die, and the sampling of one million draws from the fitted copula. I would like to take a quick moment and point out that despite the large sample of simulated outcomes, copula are continuous over the unit-hypercube and

[10] For illustration, consider two positively correlated test statistics, x and y, both governed by left-tailed decision rules. In this setting, departures from the null hypothesis manifest as both statistics taking on values closer to the origin. Consequently, when the null is false, the joint density of x and y concentrates in the lower-left region of their support, reflecting enhanced left-tail dependence. Alternatively, imagine that x and y are associated with two-tailed tests. In that case, deviations from the null in either direction lead to reduced p-values; following the transformation to p-values—where rejection is signaled by small values—the joint density again accumulates near the origin. In both scenarios, the dependence structure induced by the decision mechanism results in a copula that predominantly captures left-tail dependence in the transformed space.

[11] Rotations of the copula can be used to control which corner(s) of the unit square experience the tail dependence. Rotations can be made in increments of $90°$ clockwise or counter-clockwise, for example, to rotate an Archimedean copula $180°$ as in the table one would replace each pair, (x, y) with $(x' = 1 - x, y' = 1 - y)$ and simplify accordingly.

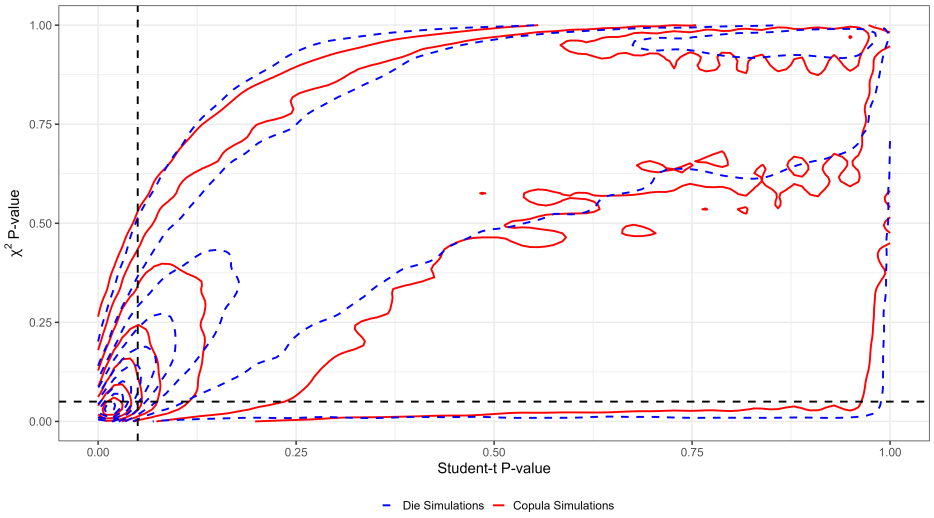
Table 2: A Single Die: Archimedean Copula Choice

Copula	$\hat{\theta}$	Log-Lik	AIC	BIC
Clayton	0.55	82,755	-165,509	-165,497
Gumbel	1.27	57,338	-114,675	-114,663
Frank	2.38	73,207	-146,413	-146,401
Joe	1.29	31,509	-63,017	-63,005
Tawn (Type I)	1.72	73,882	-147,760	-147,736
Tawn (Type II)	1.28	36,253	-72,503	-72,479
Rotations				
Clayton	0.40	43,888	-87,774	-87,762
Gumbel	1.31	84,300	-168,599	-168,588
Joe	1.43	73,959	-147,916	-147,904
Tawn (Type I)	1.31	52,992	-105,981	-105,958
Tawn (Type II)	2.26	119,487	-238,970	-238,947

Note: Estimates are based on simulations of one million fair, six-sided die. Note that the number of simulations can be increased to arbitrarily increase precision of the relevant parameters.

merely approximate the finite sample from a discrete event like the roll of a six-sided die. As mentioned earlier, this is no different from comparing one of the finite test statistics obtained from n rolls of the die to its continuous null distribution.

Figure 2: Contour Comparison: Die Rolls and Tawn (Type II) Copula



Note: Using the one million dice rolled to create Figures 1a and 1b a series of copulas from the Archimedean family were fit. From Table 2 a rotated Tawn copula best describes the observed statistics. To compare, a sample of one million draws was constructed from the continuous Tawn Copula and the resulting contours plotted along those previously observed.

The Tawn copula (Tawn, 1988) is a modification of the well-known Gumbel copula, an Archimedean copula known for right-tail dependence, to allow for asymmetric dependence over the support. It is a three parameter copula, (θ, α, β), such that $\theta \in [1, \infty)$ controls the level of overall dependence, and $\alpha, \beta \in [0, 1]$ are asymmetry parameters and can be

written as,

$$P(u_1, u_2|H_b^0) = u_1^{1-\alpha} u_2^{1-\beta} \mathrm{e}^{\left(\left(\beta \log\left[\frac{1}{u_1}\right]\right)^\theta + \left(\alpha \log\left[\frac{1}{u_2}\right]\right)^\theta \right)^{1/\theta}}. \tag{8}$$

When $\alpha = \beta = 1$, the Tawn copula reduces to the standard Gumbel copula. For the case of the fair, six-sided die simulated in Figure 1b the estimated parameters, assuming we rotate the distribution $180°$, are $\theta = 2.261,\ \alpha = 1.000,\ \beta = 0.284$.

Returning to the example of the two students, suppose now that Student C utilizes this joint inference approach to assess whether the sample $\mathcal{D}$ (the same one observed by Students A and B) is consistent with a fair die. Recall that Student A obtained a p-value of 0.031 using the chi-squared test and Student B obtained a p-value of 0.282 using the Student-t test. (For clarity, assume that u_1 corresponds to the p-value from the Student-t test and u_2 corresponds to that from the chi-squared test.) Student C then evaluates the joint probability

$$P\big(u_1 \leq 1 - 0.282,\ u_2 \leq 1 - 0.031\big) \tag{9}$$

using the fitted copula. Based on the estimated copula parameters and the observed p-values, Student C finds that this joint probability is 0.015, meaning that he or she would reject the null of a fair, six-sided die at a 5% significance level. By incorporating the joint distribution of the p-values, this procedure resolves the seemingly paradoxical situation of conflicting test results. The joint inference procedure achieves this by leveraging more information than is available in the marginal representations.

Finally, if a single p-value represents the probability of observing data as extreme as that actually observed – conditional upon the null being true – then the joint inference procedure extends this concept to the multivariate setting. In this framework, the combined p-value corresponds to the probability of observing a set of p-values at least as extreme as those produced by the data, given the fundamental null H_b^0. This approach, therefore, allows for a more nuanced evaluation of the evidence, harnessing the full dependence structure among the testable statements. While the example contained herein is relatively straightforward, the pseudo-code outlined in Algorithm 1 can be used more generally, for example in the case of testing for unit roots or seasonality in time series.

While conducting joint inference via Algorithm 1 allows one to combine information from a set of test statistics rather than relying on a single one, the question of comparative power has yet to be answered. Suppose instead that a six-sided die has the following probability vector governing its outcomes: $(p(1) = 70/600,\ p(2) = 75/600,\ p(3) = 100/600,\ p(4) = 100/600,\ p(5) = 115/600,\ p(6) = 140/600)$; a vector we will return to in Section 3. The fundamental null of H_b^0 : "This six-sided die is fair" still applies. How does the power of the joint inference processed proposed herein compare to that of its constituent parts? Moreover, how does this joint testing procedure compare to alternative "composite" p-value methods such as Fisher's combined p-value Fisher (1970) and the harmonic p-value Wilson (2019).

Algorithm 1 Joint Inference via Archimedean Copula

Require: Observed data $\mathcal{D}$, significance level α, number of simulations N, fundamental null H_b^0, test statistic functions $\{f_j\}_{j=1}^m$, and corresponding null CDFs $\{F_j\}_{j=1}^m$.
Ensure: Decision on H_b^0 (reject or fail to reject)
Simulate p-values:
for $s = 1$ to N **do**
  Simulate dataset $\mathcal{D}^{(s)}$ under H_b^0
  for $j = 1$ to m **do**
    $x_j^{(s)} \leftarrow f_j\big(\mathcal{D}^{(s)}\big)$
    $u_j^{(s)} \leftarrow F_j\big(x_j^{(s)}\big)$
  end for
end for
Let $\mathcal{U} = \{(u_1^{(s)}, \ldots, u_m^{(s)}) : s = 1, \ldots, N\}$.
Fit Copula:
Estimate Archimedean copula parameters $\hat{\theta}$ from $\mathcal{U}$ to obtain fitted copula $C(u_1, \ldots, u_m; \hat{\theta})$.
Compute Observed p-values:
for $j = 1$ to m **do**
  $x_j \leftarrow f_j\big(\mathcal{D}\big), u_j \leftarrow F_j\big(x_j\big)$
end for
Evaluate Joint p-value:
$P_{\text{joint}} \leftarrow C(\tilde{u}_1, \ldots, \tilde{u}_m; \hat{\theta})$
if $P_{\text{joint}} \leq \alpha$ **then**
  return Reject H_b^0
else
  return Fail to reject H_b^0
end if

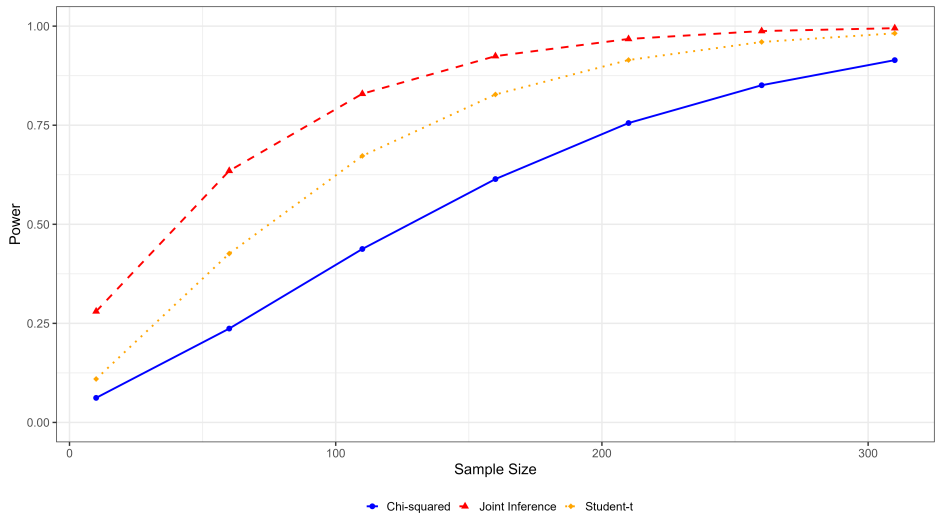
In Figure 3 I plot power curves, assuming an unfair die with the aforementioned probabilities, for an increasing sample size and compare the proposed joint testing method and its constituent parts. At each sample size a simulation of $100,000$ of these unfair die was used and the power shown is the proportion of those simulations where the corresponding null distribution was rejected. Since inferences from the joint distribution has access to more information than either of the marginal distributions its power curve strictly dominates.

Fisher's combined p-value (Fisher, 1970) is an early attempt at drawing inference simultaneously from multiple p-values. The statistic which I will denote as F, asymptotically a chi-squared statistic, can be expressed as,

$$F = -2\sum_{i=1}^{k} \log[p_i], \tag{10}$$

where $i = 1, \ldots, k$ is the number of tests under consideration and p_i is the p-value of the i^{th} statistic. This statistic is then evaluated under a $\chi^2(2k)$ distribution in a right-tailed sense. More plainly, the collection of p-values turns into a statistic, which when compared to the appropriate null distribution turns into a p-value, a concept not to dissimilar to what is outlined herein. A major weakness of this approach is that it necessarily assumes that $p_i \perp p_j$ for all $i \neq j$.

Figure 3: Joint Inference Is More Powerful: Evidence from a Six-Sided Die



Note: Using the probability vector $(p(1) = 70/600,\ p(2) = 75/600,\ p(3) = 100/600,\ p(4) = 100/600,\ p(5) = 115/600,\ p(6) = 140/600)$ and following Algorithm 1, the joint inference method outperforms its constituent parts in detecting deviations from H_b^0. Values shown are the proportion of rejections from $100,000$ samples generated from the unfair die at each sample size.

While there have been many attempts at extending the Fisher type method to dependent statistics, the most recent is the harmonic p-value (Wilson, 2019).[12] The object of interest, which I will denote as W, can be expressed as,

$$W = \frac{\sum_{i=1}^{k} \omega_i}{\sum_{i=1}^{k} \frac{\omega_i}{p_i}}, \tag{11}$$

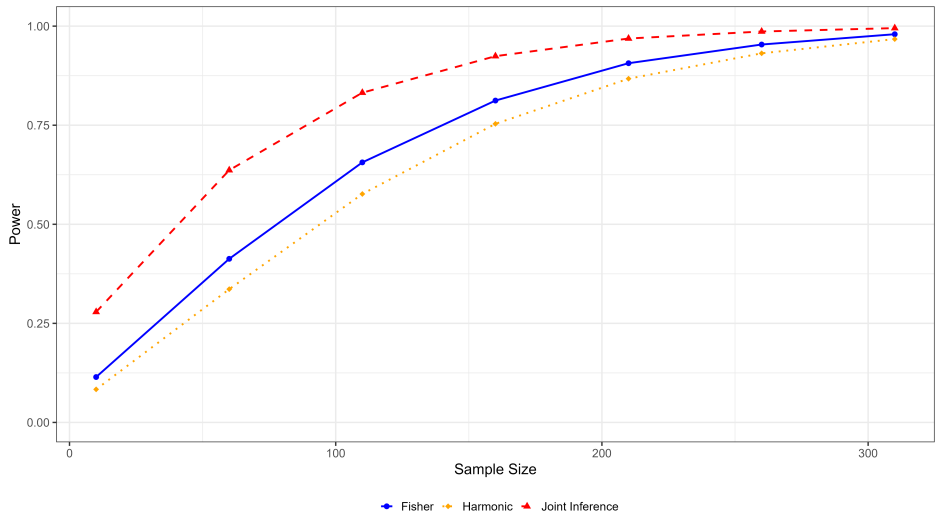
where $i = 1, \ldots, k$ is the number of tests under consideration and $\Omega = (\omega_i, \ldots, \omega_k)$ is a set of user specified weights such that $\sum_{i=1}^{k} \omega_i = 1$. The asymptotic distribution, in the number of tests k, is based off of the Landau distribution and is outlined in the author's original work.[13]

In Figure 4 I plot results in similar fashion to Figure 3 with the comparison group being the aforementioned composite p-value approaches instead of the marginal distributions. Relative to the marginal distribution, the Fisher combined p-values behaves in terms of power nearly identically to the Student-t statistic with a negligible loss in power comparatively. The harmonic p-value falls between the the Student-t and chi-squared test given $\omega_1 = \omega_2 = 0.50$. This is consistent with the idea of the harmonic p-value being equivalent–in some sense–to Bayesian model averaging. The take away is that these composite p-value methods are in some way failing to maximally extract the additional, non-overlapping information provided by the two statistics as compared to the copula based method I propose.

To summarize, often we have interest in testing a statement about the world that can be expressed through numerous possible measures of observed data. These are typically measures of convenience, and are (often though not always) marginally representative of the statement under consideration. Conditional on the null being true, I argue that we can take these marginally representative statements and consider their corresponding null distributions as marginal

[12] This is a version of composite p-value construction analogous to Bayesian model averaging.
[13] Practically, I have implemented this method via the authors own R package, harmonicmeanp.

Figure 4: Comparing Against Other Composite P-Value Approaches



Note: Using the same probability vector from Figure 3 and following Algorithm 1, the joint inference method outperforms two other composite p-value approaches, the Fisher combined p-value and harmonic p-value respectively, in detecting deviations from H_b^0. Values shown are the proportion of rejections from $100,000$ samples generated from the unfair die at each sample size.

representations of a larger joint space. Though it is trivial to show this joint distribution not only exists but is unique in small sample, discrete settings, it is non-trivial in larger sample and/or continuous settings. Using Sklar's theorem, I show how one can approximate the joint distribution from test statistics derived from a simulated environment where the fundamental null is known to be true and subsequently draw inference about the observed data. I show that, in the case of a six-sided die, this method is more powerful at detecting deviations from the stated null than either of its constituent parts. The increase in power I find directly corresponds to that found in Chen et al. (2021) and Chen and Cornwall (2021) under a higher dimensional setting.

3 Joint Testing and a Game of Chance

Gambling as a whole in the United States is an industry that generated revenue in excess of 50 billion dollars in 2021 (Yakowicz, 2022). In the state of Nevada alone, home to Las Vegas and its multitude of casinos, a total of 14.8 billion dollars of revenue was generated in 2022 (Yakowicz, 2023). As of this writing a cursory examination shows that gambling is currently legal and operational or legal and pending operation in thirty-eight states plus the District of Columbia with notable exceptions being California and Texas (Association, 2023).[14] The gambling landscape continues to change at a rapid pace since the U.S. Supreme court overturned *Murphy v. National Collegiate Athletic Association* in 2018 with numerous states quickly moving to legalize retail and/or online gambling in a variety of forms.[15] Given the size and scope of the gambling industry, the fairness of games is of general concern and serves as a backdrop for illustrating the proposed testing approach.

[14]The legal landscape being framed here is current as of June 22, 2023 and is subject to change going forward.
[15]See the full decision of Murphy v. National Collegiate Athletic Association, 16 U.S. 476 (2018) for more information.

Consider the game of craps, a game of chance involving two [fair] six-sided die, available at many retail casinos. For those that may be unfamiliar with the game there are many ways to bet on craps including but not limited to: pass-the-line, place win (lose), hardways, and field; each catering to the bettors' risk preferences and producing different takes for the house. In this case I will exclusively focus on the pass-the-line bet, the most even bet at a craps table, with the expected value of a roll being $\approx \$0.986$.[16] That is, for every dollar bet, in expectation, the bettor can expect to receive 98.6 cents in return, the missing 1.4 cents represents the house "take" on the game. A cursory search at the time of writing indicates that, in Las Vegas, there are approximately 250 craps tables in operation across a number of casinos which generated roughly 450 million dollars in 2022 revenue according to Forbes (Yakowicz, 2023) making this a very profitable game overall.

To illustrate the advantages of the proposed testing approach, I consider the case of a single casino with one-hundred craps tables. This casino is provided new dice for their tables daily by a single provider and only those dice are used throughout the day, being immediately discarded at the end of the twenty-four hour period. For simplicity, the casino's tables are full at all times and produce exactly one-hundred rolls of the dice each hour. Moreover, the tables only allow pass-the-line bets in the amount of one-hundred dollars and there is no "let it ride". A bettor either wins or loses based on the rules of the pass the line bet, and then maintains the same wager value at the table.

The casino is informed by their provider that the latest batch of dice are contaminated and that 35% of the batch has been incorrectly weighted. The provider further informs the casino that even a single, unfair die at the table flips the odds in a pass-the-line wager to favor the bettor rather than the casino, leading to an expected loss. The provider is unable to provide a specific weighting of the unfair die and as a result no known alternative to a fair table is available. For simplicity, in calculations I will assume the dice are packaged in pairs and that the casino has a fixed distribution of tables. In Table 3 I have outlined the outcome probabilities, house take, and distribution of tables.

The risk-less option is to close the tables until new dice can be procured, however this provides no revenue to the casino.[17] Rather than closing the table, the casino decides it is in their best interest to station table observers at each table and record the outcome from each throw of the dice. If, in the allocated time, the data is inconsistent with H_b^0 : the table is fair, then the table is shut down, otherwise it is allowed to remain open for the remainder of the day.

Before continuing, allow me to take a moment to bound this hypothetical a bit more so as to ease both the conceptual and computational burden. Rather than doing a sequential test, the casino will only perform their statistical analysis at the end of a two hour sampling period. I will present three possible outcomes based on three different testing regimes. Under the first option, the casino will use a Student-t test to determine if a table is fair, relying upon the higher convergence rate of the test statistic over the sufficiency of the test. In the second option, the casino will rely on the chi-squared test, relying on a test that is both necessary and sufficient but slower to converge over the Student-t. The

[16] A pass the line bet is the basic bet a bettor can make at a craps table. It consists of a "come-out roll" which wins if a 7 or 11 is rolled and loses if a 2,3, or 12 are rolled; any other roll establishes a point. With a point established a bettor wins if the point is rolled prior to the appearance of a seven, and loses if a seven is rolled. There are many other kinds of bets at a craps table but this is the most basic and balanced of the bets with a high expected value for the bettor.

[17] One could imagine in a real scenario that closing down an entire swath of games may also impact the traffic of other games and/or the perception of the casino more broadly.

Table 3: Outcome Probabilities

	Number of Fair Dice		
Outcome	Two	One	Zero
2	0.028	0.019	0.014
3	0.056	0.040	0.029
4	0.083	0.068	0.055
5	0.111	0.096	0.081
6	0.139	0.128	0.114
7	0.167	0.167	0.158
8	0.139	0.147	0.150
9	0.111	0.126	0.142
10	0.083	0.099	0.115
11	0.056	0.071	0.089
12	0.028	0.039	0.054
House Take	1.414%	-1.577%	-4.933%
P(House Win)	0.5071	0.4921	0.4753
Tables	50	30	20

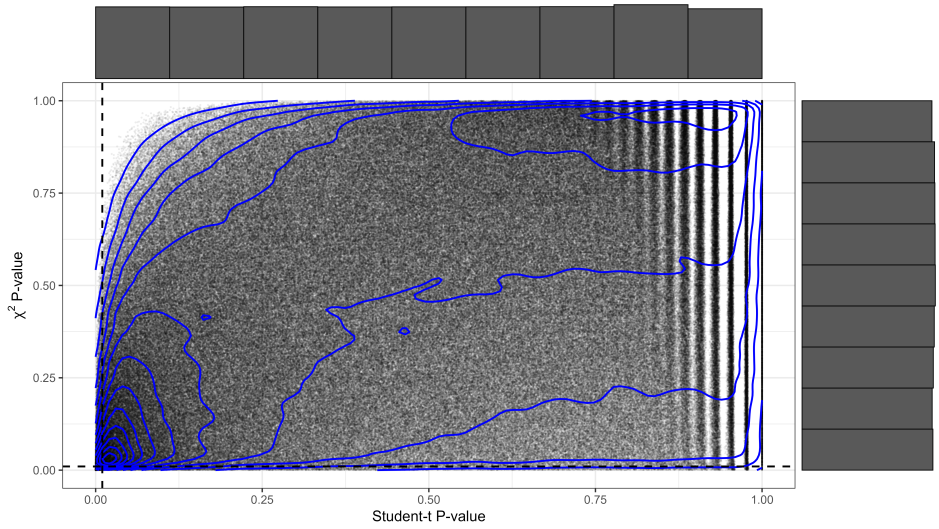
Note: The probabilities for each outcome in a game of craps, assuming both die are fair, is well known and is outlined in the first column. If a table has one fair and one unfair die the outcome probabilities are listed in column two. Finally, if both dice at the table are unfair the outcomes are listed in the third column. Note that a positive take value indicates the amount a casino makes in expectation on each dollar wagered. Negative takes lead to losses on each dollar wagered.

third option will consist of the proposed test paradigm where the Student-t and the chi-squared tests are evaluated jointly. Under all three testing paradigms the casino will require a high bar of evidence to determine if a table is unfair, fixing the probability of making a Type I error – closing a table when it is in fact a fair table – at $\alpha = 0.01$.

The fixed nature of the example makes many of the calculations straight forward. For example, since the game under these conditions collapses to a binary outcome with known probabilities, the revenue distribution for a twenty-four hour period is [approximately] Gaussian with mean $\$340,800$ and standard deviation $\$48,984$ assuming all one-hundred tables are fair. Under the distribution provided in Table 3 the revenue distribution if a casino does nothing is [approximately] Gaussian with mean $-\$179,040$ and standard deviation $\$48,973$ over that same twenty-four hours. The goal is thus straight forward, collect data over two hours and then shut down tables inconsistent with the null of a fair table. The optimal outcome is that the casino shuts down zero fair tables and all unfair tables at the end of the two hour testing period; leading to a revenue distribution that is Gaussian with mean $\$141,160$ and standard deviation $\$36,050$.

The simulation of fair tables, the calculation of the appropriate test statistic(s), and their corresponding p-values, is relatively straight forward. Figure 5 plots one million p-value pairs and corresponding density produced from a fair craps table after 200 rolls. As expected there is increased density at the $(0,0)$ and $(1,1)$ vertices of the unit square. The joint density is asymmetric along the diagonal and has low density near the $(0,1)$ vertex. Similar to the single, six-sided die example these features are consistent with a dependence structure governed by a Tawn Copula, rotated 180 degrees, with parameters $\hat{\theta} = 1.630$, $\hat{\alpha} = 1.000$, and $\hat{\beta} = 0.240$.

Figure 5: Joint Distribution of a Fair Craps Table



Note: Each point is a pair of p-values produced from observed outcomes of two-hundred rolls at a single craps table. The dashed lines indicate marginal thresholds at the $\alpha = 0.01$. Contour lines are estimated from the collection of statistics and the marginal histograms are provided.

In Table 4 I provide testing outcomes for each of the three testing paradigms over $10,000$ replications. From this we can see that the joint inference method proposed herein produces higher expected revenue ($4,244) per post-testing hour with lower standard deviation ($7,999) than those produced by its constituent parts. From those same replications, the table also identifies the both the single worst and best outcome in terms of expected post-testing revenue. In both cases the joint inference approach outperforms the Student-t and chi-squared test respectively.

To summarize, rather than choosing one testing paradigm, the dominant choice of the casino under the conditions outlined is to combine them using the process outlined in Algorithm 1. This leads to higher expected revenue with less uncertainty relative to the alternatives. This implies that over the long run, in cases where a casino may be uncertain about the fairness of the dice it deploys, using the proposed method will allow them to outperform competitors using a single test paradigm from which to draw inferences.

4 Conclusion

Scientific discovery predicated upon inferences from observed data is a difficult task, and comes with a number of complications and pitfalls for researchers. One such issue is the selection and use of a single test statistic from which to draw those inferences. Recognizing that often researchers are interested not in the specific measure from which the test statistic is derived, but rather a more fundamental question, I demonstrate how one can draw upon multiple such measures and identify deviations from the null with a higher degree of accuracy. In the case of the six-sided die I show how this alloy testing method, based in an Archimedean copula, out performs its constituent parts in finite samples. Then, to illustrate the potential economic impact of this method, I show how a casino might leverage this method to

Table 4: Better Outcomes from Joint Testing

	Measure	Student-t	χ^2	Joint
Two - Fair	Mean	49.5	48.5	48.2
	Std. Dev.	(0.71)	(0.72)	(1.34)
	Minimum	45	45	41
	Maximum	50	50	50
One - Fair	Mean	18.4	26.4	15.6
	Std. Dev.	(2.68)	(1.78)	(2.76)
	Minimum	8	17	6
	Maximum	28	30	26
Zero - Fair	Mean	0.43	4.33	0.27
	Std. Dev.	(0.65)	(1.82)	(0.52)
	Minimum	0	0	0
	Maximum	5	14	4
Financial/Hour	$\mathbb{E}(R)$	\$3,912	\$711	\$4,244
	$\sigma(R)$	\$8,263	\$8,956	\$7,999
Worst Outcome	Two-Fair	49	50	46
	One-Fair	20	26	20
	Zero-Fair	5	14	3
	$\mathbb{E}(R)$	\$1,328	-\$3,924	\$1,890
Best Outcome	Two-Fair	50	50	50
	One-Fair	9	21	7
	Zero-Fair	0	0	0
	$\mathbb{E}(R)$	\$5,678	\$3,782	\$5,994

Note: The results in this table are from $10,000$ replications of the two hour testing period. All table closure decisions are done fixing $p(\text{reject}|H_b^0 : \text{Table is fair}) = 0.01$.

obtain revenue distributions which second order stochastically dominate the alternatives when facing uncertainty about the fairness of their game.

Throughout this paper I argued that we can leverage non-overlapping information contained in multiple test statistics to better identify deviations from the null statement of interest. However, as economists are fond of repeating "there is no free lunch", this comes at the cost of increased computational burden since, as explicated, it requires one to write down a null data generating process matching the statement of interest and numerically approximate the joint distribution of two or more test statistics through simulation. Depending on the cost of Type II errors this may or may not be, in its current form, worth that additional computational burden. However, this form of inference does open up new avenues of research with respect to more computational efficient representations and possible asymptotic derivations of joint distributions for common problems (e.g., the six-sided die). Perhaps the use of machine-learning techniques, where approximation of complex, high dimensional probability spaces is far more common, may lend insight into reducing the computational burden via class discrimination techniques rather than strict null hypothesis testing.

References

Association, A. G. (2023). Interactive u.s. map: Sports betting.

Chen, J. and G. Cornwall (2021). The darkside of the moon: Searching for the other half of seasonality. Technical report, Working Paper.

Chen, J., G. Cornwall, and B. Sauley (2021). Roots from trees: A machine learning approach to unit root detection. Technical report, Working Paper.

Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer.

Genest, C., B. Rémillard, and D. Beaudoin (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and economics 44*(2), 199–213.

Genest, C. and L.-P. Rivest (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association 88*(423), 1034–1043.

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

MacKinnon, J. G. (2010). Critical values for cointegration tests. Technical report, Queen's Economics Department Working Paper.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer science & business media.

Neyman, J. and E. S. Pearson (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 231*(694-706), 289–337.

Schapire, R. E. and Y. Freund (2013). Boosting: Foundations and algorithms. *Kybernetes 42*(1), 164–166.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, Volume 8, pp. 229–231.

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika 75*(3), 397–415.

Wang, X., R. J. Hyndman, F. Li, and Y. Kang (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting 39*(4), 1518–1547.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences 116*(4), 1195–1200.

Yakowicz, F. S. (2022). U.s. gambling revenue hit record $53 billion in 2021.

Yakowicz, F. S. (2023). Nevada set a gaming revenue record in 2022 with $14.8 billion.