Testing Independence of Exchangeable Random Variables

Marcus Hutter

DeepMind, London, UK
http://www.hutter1.net/

Abstract

Given well-shuffled data, can we determine whether the data items are statistically (in)dependent? Formally, we consider the problem of testing whether a set of exchangeable random variables are independent. We will show that this is possible and develop tests that can confidently reject the null hypothesis that data is independent and identically distributed and have high power for (some) exchangeable distributions. We will make no structural assumptions on the underlying sample space. One potential application is in Deep Learning, where data is often scraped from the whole internet, with duplications abound, which can render data non-iid and test-set evaluation prone to give wrong answers.

Keywords: independent; identically distributed; exchangeable random variables; statistical tests; unstructured data.

Table of Contents

Table of Contents

IID and Exchangeable Distributions

Definition (Exchangeable distributions)

- Probability space $(\mathcal{X}^n, \Sigma, Q)$
- Probability Q is (finitely) exchangeable $\iff$ $Q(x_1, ..., x_n)$ is invariant under all (finite) permutations of its argument.
- $\mathcal{Q} := \{\text{exchangeable } Q\}$

- In particular $x_1, ..., x_n$ are equally distributed: $Q[X_t = x] = Q[X_{t'} = x]$

Definition (IID Distributions)

- Probability space $(\mathcal{X}^n, \Sigma, P_{\boldsymbol{\theta}})$
- Q is independent and identically distributed (iid) $\iff$ $Q(x_1, ..., x_n) = P_{\boldsymbol{\theta}}(x_1, ..., x_n) := \theta_{x_1} \cdot ...\theta_{x_n}$ for some $\boldsymbol{\theta} \in [0;1]^{\mathcal{X}}$ with $\sum_{x\in\mathcal{X}} \theta_x = 1$
- $H_{\text{iid}} := \{\text{iid } Q\} \equiv \{P_{\boldsymbol{\theta}}\}$

- In particular iid $P_{\boldsymbol{\theta}}$ are exchangeable: $H_{\text{iid}} \subset \mathcal{Q}$

Problem Setup

Main Question Considered in this Talk

How to test whether exchangeable random variables $X_1, ..., X_n$ are independent, solely from observations $x_{1:n} := x_1 x_2 ... x_n$ sampled from some exchangeable Q.

- *(Only) assumptions:* $\mathcal{X} \supseteq \{x_1, ..., x_n\}$ and Q is exchangeable.
- *Less formally:* Assume $x_{1:n}$ is well-shuffled. Did it originate from some iid distribution P_θ?
- The *only useful information* in $x_{1:n}$ is the counts $n_x := |\{x_t : x_t = x\}|$ of each $x \in \mathcal{X}$, and indeed actually only the *second-order multiplicities* $m_k := |\{x : n_x = k\}|$.
- So we may as well assume $\mathcal{X} \subseteq \mathbb{N}$ (will be proven).
- We are primarily interested in low multiplicities n_x.

Binomial Example

- Shuffle $n = 1000$ coins with 500 heads up without turning them.
- Looks random?
  Probability of 500 heads from flipping 1000 coins iid is only 2.5%.
- $\Longrightarrow$ Test "$N_{\text{heads}} \stackrel{?}{=} n/2$" rejects H_{iid}.
- What about $n = 1'000'000$ and $n_1 = 314'159$.
- Obviously not fair, but maybe from coin with bias around n_1/n?
- n_1 is Prime and $P[\text{prime}] \approx 1/\ln(n_1) \doteq 7\%$ is small.
- n_1 is also first 6 digits of π. Again n_1 is suspicious.
- How to avoid numerology: Universal tests [Hut22]

Black Jack Example

- Cards are drawn from $c \in \mathbb{N}$ card decks of 52 cards each deck.
- The first few draws look uniformly iid.
- Closer to the end of the pile, the non-iid nature is revealed (exploited in card-counting)
- For instance: the chance of seeing no face twice when drawing 26 cards iid from 52 faces is less than 0.2% (cf. the birthday paradox), thus is strong evidence for $c = 1$.
- Our tests are **not** tailored to this setting, but our most advanced test is sensitive to this signal.

Data Duplication in Machine Learning

- Data is scraped from the whole internet and *duplications* abound.
- If, say, a photo appears more than once, the chance that it originated from independent shoots is close to zero.
- This is *evidence* that the scraped data is **not** *iid*.
- *Why is this relevant?* ML still mostly assumes iid and train/test split
- *Problem:* If, for instance, the whole data set contains 3 copies of each data item, then 99% of the items in the 10% hold-out set appear as well in the train set.
- A *pure memorizer* without any generalization capacity will perform nearly perfectly on the hold-out set,
  but will fail in practice on any newly taken photo.
- Removing *approximate duplicates* is a huge ill-defined AI-complete problem.
- *Conclusion:* Detecting that unordered/shuffled/exchangeable data is non-iid can prevent falling prey to **bad** *overfitting* due to misleading low test error.

Unrelated Work

- Testing *independence of a pair of random variables* (X, Y), given a number of *iid* sample pairs $\{(x_t, y_t)\}$
(e.g. mutual information and chi-square tests).

- *Stochastic processes:* Dependence can be tested via estimating auto-correlation coefficients. Requires ordered data and $\mathcal{X} = \mathbb{R}$.
Might be extendible beyond linear order and beyond $\mathcal{X} = \mathbb{R}$.

- Our setup is totally different and much harder.

Table of Contents

List of Notation

Symbol	Type	Explanation
$\mathcal{X}$		sample space of size $d = \lvert\mathcal{X}\rvert$, mostly $d = \infty$ and $\mathcal{X}$ cour
n		number of samples, sample size
X		$\mathcal{X}$-valued random variable
$\boldsymbol{X}$	$\equiv X_{1:n}$	n iid or exchangeable random variables
$\boldsymbol{x} \equiv x_{1:n}$	$\in \mathcal{X}^n$	sample of size n
N_x	$= \#\{X_t : X_t = x\}$	(first-order) count=multiplicity of x in $\boldsymbol{X}$
M_k	$= \#\{x : N_x = k\}$	(second-order) count=multiplicity of k in $\boldsymbol{N}$
$\boldsymbol{M}$	$= (M_1, M_2, ...)$	vector of M_k excluding M_0
$x, n_x, m_k, \boldsymbol{m}, ...$		realization of random variable $X, N_x, M_k, \boldsymbol{M}, ...$
$P(x)$	$:= P[X = x]$	probability that X is x
$P_\theta(k)$	$\equiv f_k^n(\theta)$	$:= \binom{n}{k}\theta^k(1-\theta)^{n-k}$ binomial distribution over $\mathbb{N}_0$
$P_{\boldsymbol{\theta}}$	$\in H_{\text{iid}}$	iid (multinomial) distribution over $\mathcal{X}^n$ $(\mathbb{N}_0^{\mathcal{X}})$
$P_\lambda(k)$	$\equiv g_k(\lambda)$	$:= \lambda^k e^{-\lambda}/k!$ Poisson distribution over $\mathbb{N}_0$
$P_{\boldsymbol{\lambda}}$		product of Poisson(λ_x) distributions over $\boldsymbol{n} \in \mathbb{N}_0^{\mathcal{X}}$
Q	$\in \mathcal{Q}$	exchangeable distribution

List of Notation

Symbol	Type	Explanation
Z		generic random variable
$\mathbb{E}$		expectation w.r.t. $P_{\boldsymbol{\theta}}$ or $P_{\boldsymbol{\lambda}}$ unless otherwise noted
$\sigma^2 = \mathbb{V}[Z] :=$	$\mathbb{E}[Z^2] - \mathbb{E}[Z]^2$	variance of Z and other random variables
$\mathrm{Cov}[Y, Z] :=$	$\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]$	covariance of Y and Z
$\bar{Z}$	$:= Z/n$	**not** an average of random variables
ζ	$:= \mathbb{E}[Z]$	corresponding lower-case greek letters denote expectation
ζ^{ub}	$\in \mathbb{R}$	upper bound on expectation
V^{ub}		deterministic or stochastic upper bound on variance
T	$: \mathcal{X}^n \to \mathbb{R}$	generic test statistic
$E, O, M_k, D_k, C_k, \bar{U}_k$		specific test statistics
α	$= P_{\boldsymbol{\theta}}[T > c_\alpha]$	Type I error, prob. of falsely rejecting H_{iid}
$\beta(\alpha) =$	$Q[T > c_\alpha]$	power of test T at level α for Q

IID $\to$ Multinomial $\rightsquigarrow$ Poisson

- *Binomial distribution:* $P_\theta(k) \equiv f_k^n(\theta) := \binom{n}{k}\theta^k(1-\theta)^{n-k}$ $(0 \leq \theta \leq 1,\ k \in \mathbb{N}_0)$
- *Poisson distribution:* $P_\lambda(k) \equiv g_k(\theta) := \dfrac{\lambda^k e^{-\lambda}}{k!}$ $\quad (\lambda \geq 0,\ k \in \mathbb{N}_0)$
- *IID=True Distribution:* $P_{\boldsymbol{\theta}}(x_{1:n}) \ = \ P_{\boldsymbol{\theta}}(x_1)\cdot ... \cdot P_{\boldsymbol{\theta}}(x_n) \ = \ \theta_{x_1}\cdot ... \cdot \theta_{x_n} \ = \ \prod_{x\in\mathcal{X}} \theta_x^{n_x}$
- *Multinomial distribution:* $P_{\boldsymbol{\theta}}(n_{1:d}) \ = \ \binom{n}{n_1,...,n_d}\prod_{x\in\mathcal{X}} \theta_x^{n_x}$
- *Product of Poissons:* $P_{\boldsymbol{\theta}}(n_{1:d}) \ = \ \displaystyle\prod_{x\in\mathcal{X}} P_{\lambda_x}(n_x) \ = \ \prod_{x\in\mathcal{X}} \frac{\lambda_x^{n_x} e^{-\lambda_x}}{n_x!}$

Theorem (IID = Multinomial $\approx$ Poisson Product)

For many events $E \subseteq \mathcal{X}^n$ *and random variables* $Z : \mathcal{X}^n \to \mathbb{R}$ *of interest, for large* n, $P_{\boldsymbol{\theta}}[E] \approx P_{\boldsymbol{\lambda}}[E]$, $\mathbb{E}_{\boldsymbol{\theta}}[Z] \approx \mathbb{E}_{\boldsymbol{\lambda}}[Z]$, $\mathbb{V}_{\boldsymbol{\theta}}[Z] \lesssim \mathbb{V}_{\boldsymbol{\lambda}}[Z]$. *This holds in particular for (restricted linear combinations of) basics events* $E_k^x := \{\boldsymbol{n} : n_x = k\}$ *and* $M_k^x = [\![N_x = k]\!]$.

Second-Order Count Multiplicities

- *First-order counts:* $n_x \;:=\; \#\{x_t : X_t = x\}$ = multiplicity of x in $x_{1:n}$.
- *Second-order count multiplicities:*
  $m_k := \#\{x : n_x = k\}$ = number of x that appear k times in $x_{1:n}$.

Basic properties of m_k

- $m_k = 0$ for $k > n$ but $m_k = 0$ also for many $k \leq n$ due to
- $\sum_{k=0}^{\infty} k \cdot m_k = m$ and $\sum_{k=0}^{\infty} m_k = d = |\mathcal{X}|$.
- $m_+ := \sum_{k=1}^{\infty} m_k = \#\{x : n_x > 0\} = \#\{x_1, ..., x_n\} = d - m_0$
  is the number of different x_t in $\boldsymbol{x}$, not counting multiplicities.
- We are mostly interested in $d = \infty$, in which case $m_0 = \infty$
- We therefore exclude m_0 in $\boldsymbol{m} := m_{1:n}$.

Invariant Statistical Tests

Definition (Statistical tests)

- $T : \mathcal{X}^n \to \mathbb{R}$ is a (valid) test statistic with critical value c_α for Type I error α *iff* $P_{\boldsymbol{\theta}}[T(\boldsymbol{X}) > c] \leq \alpha\ \forall \boldsymbol{\theta}$.
- T rejects H_{iid} that $\boldsymbol{x}$ is iid with confidence $1 - \alpha$ *iff* $T(\boldsymbol{x}) > c$.
- The p-value of T for data $\boldsymbol{x}$ is $p := \sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}[T(\boldsymbol{X}) > T(\boldsymbol{x})]$.
- T can reject H_{iid} with confidence $1 - p$.

- Since we assume $X_{1:n}$ are exchangeable (shuffled), it is natural to ask for T to be independent of the order in which $X_1, ..., X_n$ are presented.
- Since the **class** $\mathcal{Q}$ of exchangeable Q is invariant under permutations of elements of $\mathcal{X}$, it is natural to ask T to be as well.

Definition (Invariant tests T)

We call tests $T : \mathcal{X}^n \to \mathbb{R}$ that are invariant under permutations of the argument $x_1, ..., x_n$ as well as invariant under permutations of the elements in $\mathcal{X}$, invariant tests. Invariant tests are functions of $M_0, ..., M_n$ only.

Exchangeable Distr. and Power of Tests

- Q is *exchangeable* :iff $Q(x_{1:n}) = Q(x_{\pi(1:n)})$, where $\pi \in S_n$ is any permutation of $1:n$.
- $\implies Q$ only depends on the counts $\boldsymbol{n}$. $\mathcal{Q} := \{\text{exchangeable Q}\}$
- *Examples:* Laplace's rule $Q(x_{1:n}) = n_1!n_2!/(n+1)!$ is exchangeable. Others: KT, Good-Turing, Ristad.
- All *shuffle*d data ($\pi(1:n) \sim \text{Uniform}(S_n)$) have $Q \in \mathcal{Q}$
- *Power* $\beta = Q[T > c]$ of test T ($1 - \beta =$Type II error).
- There are *no uniformly most powerful (UMP) tests* for $\mathcal{Q} \setminus H_{\text{iid}}$.
- Different tests will have high power for some subset of $\mathcal{Q}$ and low power for other $Q \in \mathcal{Q}$.
- We focus on developing tests with correct (small) *Type I error* $\alpha =$ small size $\alpha =$ *significance level* α
- We demonstrate the (lack of) *power empirically*.

All Tests are Powerless Against Densities

- Consider (non-iid and iid) densities ρ on $\mathcal{X}^n = \mathbb{R}^n$, e.g. Gaussian.
- Then all $x_1, ..., x_n$ are different (almost surely).
- Hence for any test statistic T, $T(\boldsymbol{M}) = T(n, 0, 0, ...)$ is the same for all $x_{1:n}$ and all densities ρ, whether iid or not.
- Hence no test can discern iid from non-iid **densities**.
- Same conclusion for any $\mathcal{X}$ and non-atomic measure
- Same conclusion for countably infinite $\mathcal{X}$ by discretizing ρ on $\varepsilon\mathbb{Z}$ and $\varepsilon \to 0$.

Proposition (All tests are powerless against densities)

If $\mathcal{X}$ is infinite and all $x_1, ..., x_n$ are different, no valid invariant test can reject H_{iid}. This is **not** *true for finite $\mathcal{X}$. See $c = 1$ card counting example.*

Reducing General $\mathcal{X}$ to $\mathbb{R}$ to $\mathbb{N}$

Proposition ($\mathcal{X} = \mathbb{R}$ suffices)

For every invariant test T, $P[T > c] \leq \alpha$ *for iid P on $\mathcal{X}$*
$\iff \tilde{P}[T > c] \leq \alpha$ *for iid $\tilde{P}$ on $\mathbb{R}$ constructed below.*

Proof: Decompose P into pure point measure and atom-free rest. Construct measure $\tilde{P}$ on $\mathbb{R}$ with same point measure on $\mathbb{N}$ and any density on $\mathbb{R} \setminus \mathbb{N}$. Then $\tilde{P}(\boldsymbol{m}) = P(\boldsymbol{m})$. ∎

Proposition ($\mathcal{X} = \mathbb{N}$ suffices)

For every invariant test T and infinite $\mathcal{X}$,
$P[T > c] \leq \alpha$ *for* **all** *iid P on $\mathcal{X}$* $\iff$ $\tilde{P}[T > c] \leq \alpha$ *for* **all** *iid $\tilde{P}$ on $\mathbb{N}$.*

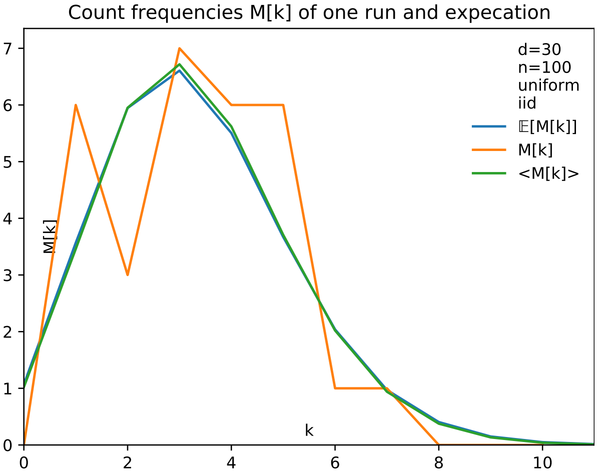
Proof: Approximate $\mathbb{R}$ by $\varepsilon\mathbb{Z}$ and let $\varepsilon \to 0$ and $\mathbb{N} \simeq \varepsilon\mathbb{Z}$. ∎

Proposition ($|\mathcal{X}| = n^3$ suffices to leadning order in n)

Table of Contents

The Poisson Distribution is "Smooth"

- $P_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{\Gamma(k+1)}$ is "smooth" in k $\approx$ *blue curve*
- Unique maximum at $k = \lambda = n/d$
- log-concave
- $\Rightarrow$ benign function



Mixtures of Poissons are (More) Smooth

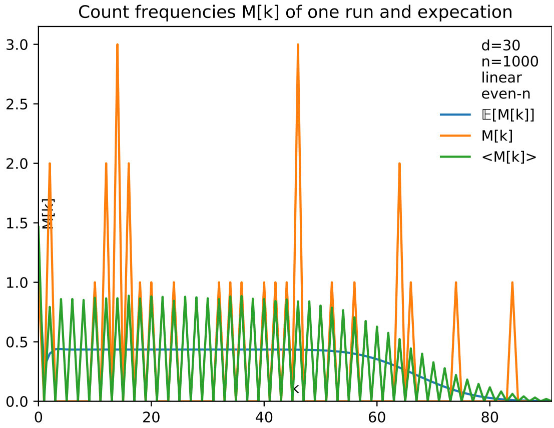
- $\mathbb{E}[M_k] = \mathbb{E}[\#\{x : N_x = k\}] = \mathbb{E}\sum_x [\![N_x = k]\!]$
  $= \sum_x P_{\boldsymbol{\lambda}}[N_x = k] = \sum_x P_{\lambda_x}(k) = \sum_x g_k(\lambda_x) = \sum_x \lambda_x^k e^{-\lambda_x}/k!$

- That is, $\mathbb{E}[M_k]$ is a sum of Poisson(λ_x) distributions.

- $\mathbb{E}[M_k]$ may have multiple extrema in k

- but as a mixture of Poissons it cannot be less smooth

- and typically is even more smooth, see plot for $\lambda_x \propto x$



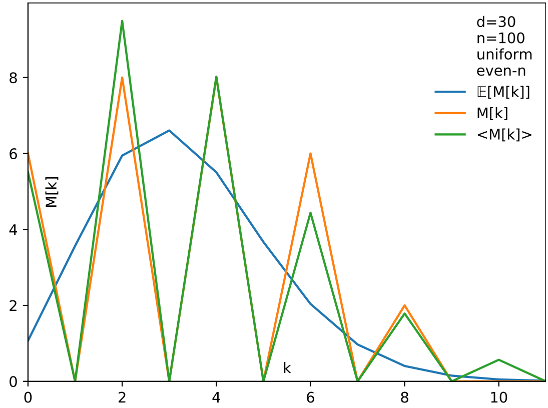
The General Idea Behind the Tests

- Since $\bar{M}_k \to \mathbb{E}[\bar{M}_k]$ for $n \to \infty$, M_k as a function of k will inherit any (lack of) structure in $\mathbb{E}[M_k]$, just with noise added (see first plot).

- Since invariant tests can only depend on $\boldsymbol{M}$, they must test for some such structure of $\mathbb{E}[M_k]$.

- Example: No Poisson (mixture) can have $\mathbb{E}[M_k] = 0$ for all odd k

- $\Rightarrow$ Such M_k is strong evidence against $\boldsymbol{X}$ being iid.



The Specific Idea Behind the Tests

For any test $T = T_n$ with critical values c_p, which is asymptotically Gaussian, the p-value can be approximately upper bounded by

$$p \overset{(a)}{=} \sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}[T > c_p] \overset{(b)}{\approx} \sup_{\boldsymbol{\theta}} \Phi\left(\frac{\mathbb{E}_{\boldsymbol{\theta}}[T] - T}{\sqrt{\mathbb{V}_{\boldsymbol{\theta}}[T]}}\right)$$

$$\overset{(c)}{\leq} \Phi\left(\frac{\tau^{ub} - T}{\sqrt{V^{ub}}}\right) \overset{(d)}{\leq} \exp\left(-\frac{n(\bar{\tau}^{ub} - \bar{T})^2}{2\bar{V}^{ub}}\right)$$

(a) by definition of c_p

(b) by T being asymptotically Gaussian

(c) by $\tau^{ub} :\geq \mathbb{E}_{\boldsymbol{\theta}}[T]\ \forall\boldsymbol{\theta}$ and $V^{ub} :\geq \mathbb{V}_{\boldsymbol{\theta}}[T]\ \forall\boldsymbol{\theta}$ and if $\tau^{ub} \geq T$

(d) by $\bar{T} := T/n$ and $\bar{V}^{ub} := V^{ub}/n$ and large n and $\Phi(y) := \int_{-\infty}^{y} e^{-x^2/2}dx/\sqrt{2\pi} \leq e^{-y^2/2}/y\sqrt{2\pi}$

Random V^{ub} is also ok provided $\mathbb{E}[V^{ub}] \geq \mathbb{V}[T]$ and $\dfrac{\sqrt{\mathbb{V}[V^{ub}]}}{\mathbb{E}[V^{ub}]} \to 0$

Upper Bounds for $\mathbb{E}$ and $\mathbb{V}$ of Tests

Proposition (Upper bounds for linear tests)

- *Let* $T = \sum_k \alpha_k M_k$ *for* $\alpha_k \in \mathbb{R}$. *Then*
- $\tau := \mathbb{E}[T] \leq n \cdot \sup_{\lambda>0} g(\lambda)/\lambda =: \tau^{ub}$,
- *where* $g(\lambda) := \sum_k \alpha_k P_\lambda(k) = \sum_k \alpha_k \lambda^k e^{-\lambda}/k!$, *and*
- $\mathbb{V}[T] \leq \sum_k \alpha_k^2 \mathbb{E}[M_k] \lesssim V^{ub}$,
- *where* $V^{ub} := \sum_k \alpha_k^2 \mu_k^{ub}$ *or* $V^{ub} := \sum_k \alpha_k^2 M_k$,
- *with* $\mu_k^{ub} \geq \mathbb{E}[M_k] =: \mu_k$ *upper bounding the expectations of* M_k.

- For non-linear tests $f(T)$, we linearize by Taylor expansion $f(T) = f(\tau) + (T-\tau)f'(\tau) + O(T-\tau)^2$.

- More precisely, we apply the delta-method in statistics.

Basics M_k Test

Basic test: $T = M_k$, i.e. $\alpha_{k'} = [\![k' = k]\!]$. Then

$$\mu_k := \mathbb{E}[M_k] \leq n \cdot \sup_{\lambda > 0} \frac{\lambda^{k-1} e^{-\lambda}}{k!} = n \frac{(k-1)^{k-1} e^{-(k-1)}}{k!} =: \mu^{ub} \leq \frac{n}{k\sqrt{2\pi(k-1)}}$$

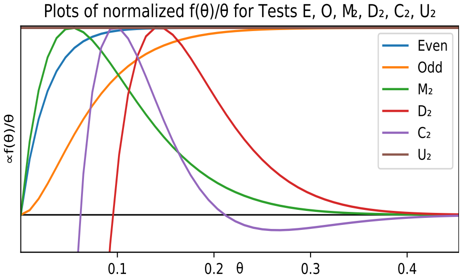
We (also) have $\mathbb{V}[M_k] \leq \mathbb{E}[M_k] \leq \mu_k^{ub}$ and also $\mathbb{V}[M_k] \lesssim M_k$

Example

- Assume each data item is duplicated and appears exactly twice.
- In this case, $M_2 = n/2$ and all other $M_k = 0$.
- For $k = 2$ we have $\bar{\mu}_2^{ub} = 1/2e \doteq 0.184$ and hence
- *p-Value:* $p \lesssim \exp(-\frac{1}{2} n (\frac{1}{2} - \frac{1}{2e})^2 / \frac{1}{2e}) \doteq e^{-0.271n}$.
- I.e. H_{iid} can be extremely confidently rejected for moderately large n.
- For $k \neq 2$, the tests have no power ($\bar{M}_k = 0 < \bar{\mu}_k^{ub}$).

Other/More Powerful Tests

- Table only shows $n \gg k \gg 1$ approximation
- For V_k^{ub} we only show the better upper bound (empirical except for M_k)



Test statistics $T : \mathcal{X}^n \to \mathbb{R}$ with upper bounds on their mean and variance:

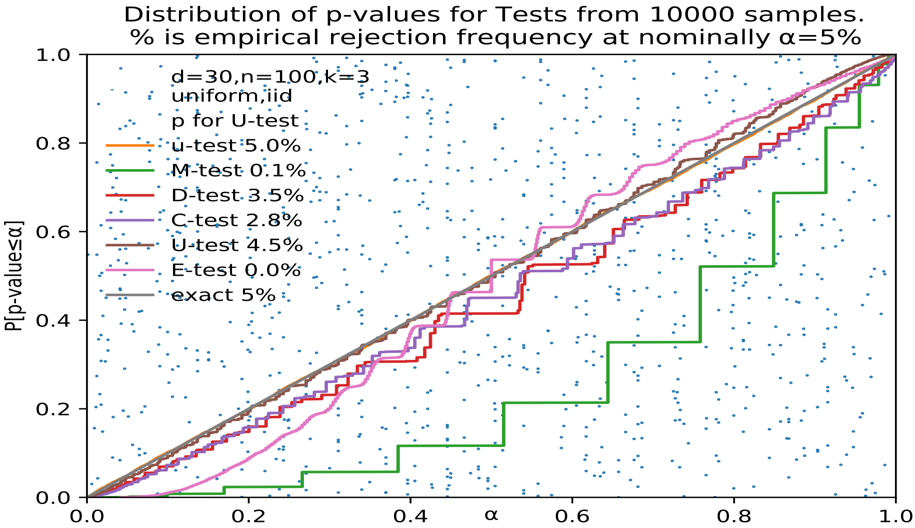
Test Name	$T := n\bar{T} :=$	$\bar{\tau} := \mathbb{E}[\bar{T}] \leq$	$\mathbb{V}[\bar{T}] \lesssim \bar{V}^{ub} =$	λ^*
Even$\neq 0$	$E := \sum_x N_x [\![N_x \neq 0 \text{ even}]\!]$	$\bar{\varepsilon}^{ub} = 1/2$	$\frac{1}{n}\sum_{k\neq 0 \text{ even}} k^2 M_k$	∞
Odd$\neq 1$	$O := \sum_x N_x [\![N_x \neq 1 \text{ odd}]\!]$	$\bar{o}^{ub} = 1/2$	$\frac{1}{n}\sum_{k\neq 1 \text{ odd}} k^2 M_k$	∞
2nd-Count	$M_k := \sum_x [\![N_x = k]\!]$	$\bar{\mu}_k^{ub} \leq \frac{1}{k\sqrt{2\pi(k-1)}}$	$\bar{\mu}_k^{ub}$	$k-1$
Slope	$D_k := M_k - M_{k-1}$	$\bar{\delta}_k^{ub} \leq \frac{1}{k^2\sqrt{2\pi e}}$	$\bar{M}_k + \bar{M}_{k-1}$	$k-1/2+\sqrt{k+1/4}$
Lin.Curv.	$C_k := 2M_k - M_{k-1} - M_{k+1}$	$\bar{\gamma}_k^{ub} \approx \frac{1}{k(k+1)\sqrt{2\pi k}}$	$4\bar{M}_k + \bar{M}_{k-1} + \bar{M}_{k+1}$	k
Log.Curv.	$\bar{U}_k := \ln(M_k^2 / M_{k-1} M_{k+1})$	$\bar{\upsilon}_k^{ub} = \ln\frac{k+1}{k} \leq \frac{1}{k}$	$\bar{M}_{k-1}^{-1} + 4\bar{M}_k^{-1} + \bar{M}_{k+1}^{-1}$	any

Table of Contents

Data Generation and Test Evaluation

- We verify the validity of our tests on *artificial*ly generated *data* (correct low Type I error).
- We generate *iid* data for all θ_x being the same and $\boldsymbol{\theta}$ being maximally diverse.
- We then "corrupt" the samples in various ways to create *non-iid* data.
- We also sample from *finite population* w/o replacement (Black Jack) to determine the tests's *power* in rejecting $H_{\rm iid}$ (*low Type II error*).
- We estimate the *p-value distribution* and *rejection frequency* at nominally $\alpha = 0.05$ from $10'000$ sampled data sets $\boldsymbol{x}$.

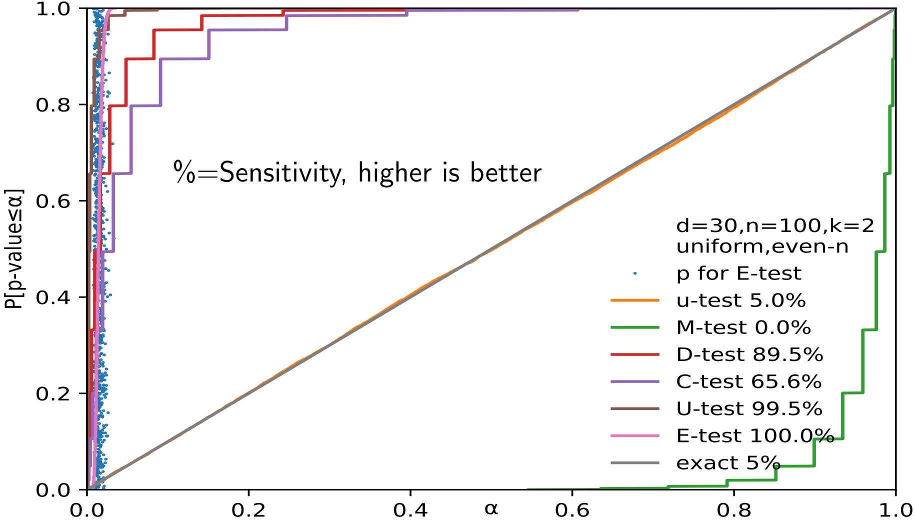
Testing the Tests on IID Data



For iid P_θ we want the curve to be on or below the diagonal
A curve above/below means over/under-confidence

Testing the Tests on Non-IID Data

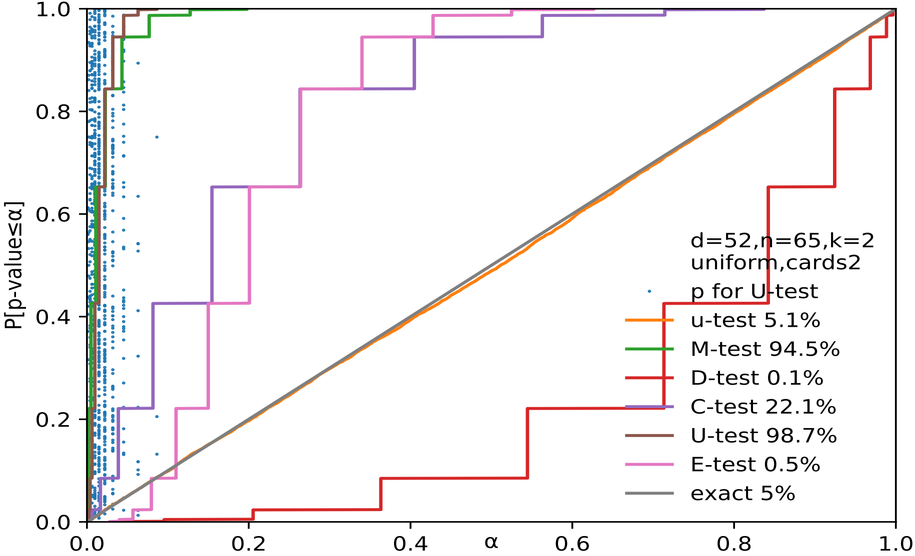
Every data item is duplicated, i.e. $M_k = 0$ *for all odd* k



For non-iid we want far above the diagonal, esp. for small α.

Black Jack

Drawing 65 cards from two 52-card decks, (only) U_2-Test reveals non-iid



Summary of Experimental Results

- Tests are not over-confident
- Tests can be under-confident (no problem)
- Tests can be powerful, weak, or vacuous
- Having no singletons can sometimes be significant
- Tests are often weak for larger k
- Some tests are able to detect data duplication and draws from finite card decks.
- Every test displayed its own strengths and weaknesses. There was no uniformly best test among them.
- Tests largely performed as expected.

Table of Contents

More/Alternative Tests

- *Summing tests:* $T_+ = \sum_k T_k$ to broaden power (fragile)
- *Bonferroni:* $T_K := \max_k\{T_k - c_k^{\alpha/|K|}\}$ to robustly broaden power
- *Uniformizing tests:* $T \rightsquigarrow \tilde{T}$ such that $P_{\boldsymbol\theta}[\tilde{T} \leq \delta] \leq \delta\ \forall\boldsymbol\theta$
- *Universal tests:* $\tilde{T} := \min\{k(k+1)\tilde{T}_k : k \in \mathbb{N}\}$
- *Likelihood Ratio (LR) tests:* $\tilde{T}(\boldsymbol{x}) := \sup_{\boldsymbol\theta} P_{\boldsymbol\theta}(\boldsymbol{x})/Q(\boldsymbol{x})$, any Q
- *Martin-Loef Test:* $Q(\boldsymbol{x}) := M(\boldsymbol{x}) \approx 2^{-\text{Km}(\boldsymbol{x})}=$ Solomonoff prob.
- *Generalized LR tests:* $\tilde{T}(\boldsymbol{x}) := \sup_{\boldsymbol\theta} P_{\boldsymbol\theta}(\boldsymbol{x})/Q_{\boldsymbol\theta}(\boldsymbol{x})$ w. $\sum_{\boldsymbol{x}} Q_{\boldsymbol\theta}(\boldsymbol{x}) \leq 1$
- *Invariant LR tests:* $\tilde{T}(\boldsymbol{m}) \;=\; \frac{n!\binom{m_+}{\boldsymbol{m}}}{m_+!Q(\boldsymbol{m})} \cdot \prod_{k=1|2}^{n} \left[\frac{1}{k!}\left(\frac{k-1}{n-m_+}\right)^{k-1}\right]^{m_k}$
- *Combinatorial tests:* E.g. $Q(\boldsymbol{m}) = \frac{1}{m_+}\binom{m_+}{\boldsymbol{m}}/\binom{n}{m_+}$ (Ristad)
- *Compression tests:* $Q(\boldsymbol{m}) := 2^{-\text{CodeLength}(\boldsymbol{m}|n)}$
- *Moment method:* H_{iid} iff $\forall k \geq 1$: $nM_k/\binom{n}{k+1}M_1 \approx$ a kth moment

Outlook

- Develop stronger tests that *exploit structural information* in $\mathcal{X}$ if available (topology, metric, ...).
- The simplest approach would be to *aggregate similar* x into the same category ($\mathcal{X}_{\text{orig}} \to \mathcal{X}_{\text{agg}}$).
- Derive *theoretical power of tests* for "interesting" subclasses of exchangeable distributions.
- Work out the *alternative* ideas for developing *tests*.
- Apply our tests to some *real data*.
- Could there be *stronger non-invariant tests?*

Summary

- We developed various tests for the (in)dependence of exchangeable data for unstructured observation spaces $\mathcal{X}$.
- We reduced the problem to $\mathcal{X} = \mathbb{N}$ which simplified the analysis.
- Data duplication is necessary for any invariant test to have power.
- The tests exploit that counts m_k are smooth if data are iid.
- Testing for non-iid w/o structure in $\mathcal{X}$ is hard but not impossible.
- Some tests detect data duplication and draws from finite card decks.
- Every test displayed its own strengths and weaknesses.
- There was no uniformly best test among them.
- Tests largely performed as expected.
- Plenty of work left (better/alternative tests, power of tests, better approximations, aggregation, exploit structure, real data, ...)

Thanks!
Questions?
Comments
References

[Hut22] Marcus Hutter.
Testing independence of exchangeable random variables.
Technical report, 2022.