Testing Independence of Exchangeable Random Variables

Marcus Hutter

DeepMind

Latest version & more @

http://www.hutter1.net/official/bib.htm#exiid

22 October 2022

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.

Contents

Keywords

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

1 Introduction

We consider the problem of testing whether a set of exchangeable random variables $X_1,...,X_n$ are independent, solely from observations $x_{1:n}:=x_1x_2...x_n$ sampled from Q. A distribution $Q(x_1,...,x_n)$ is called (finitely) exchangeable if it is invariant under all (finite) permutations of its argument. We make no structural assumptions on the underlying probability space $(\mathcal{X}^n,\Sigma,Q)$ beyond Q being exchangeable, and of course that $\mathcal{X}\supseteq\{x_1,...,x_n\}$. Less formally, assume we have observed $x_{1:n}$, which we believe to be well-shuffled, and want to know whether they originated from some iid distribution $P_{\boldsymbol{\theta}}$. The shuffling implies that the X_t are identically distributed, but it does not make them independent.

A priori one may think this is a hopeless problem. For instance, if we remove the 'identically distributed' condition, every $x_{1:n}$ is independent w.r.t. *some* non-iid distribution. One can always take $P[X_t=x_t]=1\ \forall t$, i.e. no valid test can reject the hypothesis that $x_{1:n}$ are independent, unless one makes some further assumptions on P.

The exchangeability assumption implies that the only useful information in $x_{1:n}$ is the counts $n_x:=|\{x_t:x_t=x\}|$ of each $x\in\mathcal{X}$. They form a minimal sufficient statistic. Due to the assumed lack of structure in $\mathcal{X}$, the specific label x also bears no information, so we may as well injectively map each x_t to a label from $\{1,...,d''\}$, where $d''\in\mathbb{N}$ is the number of *different* x_t in $x_{1:n}$, and can even sort them e.g. w.r.t. decreasing n_x. That means, the only useful information in $x_{1:n}$ is actually the second-order counts $m_k:=|\{x:n_x=k\}|$. All this will be made clear later. We are primarily interested in the case of low duplicity, i.e. most n_x are small, though our results are general.

Contents. We provide some motivating examples in Section 2, which also informally show that the second-order counts m_k can indeed sometimes reveal that $x_{1:n}$ did not come from an iid process. A practical example is data duplication in machine learning and the test set contamination problem it results in. In Section 3 we introduce first-order counts n_x and second-order counts m_k, exchangeable distributions, and the nature of statistical tests in this context for finite and countable $\mathcal{X}$. In particular we reduce iid distributions to multinomial distributions, and then for our tests to a mixture of Poisson distributions. We then show in Section 4 that we can reduce every $\mathcal{X}$ whatsoever to $\mathcal{X}=\mathbb{N}$ or $\mathcal{X}=\{1:d\}$ with discrete σ-algebra $\Sigma=2^{\mathcal{X}}$. After this lengthy preparation, we are finally able to develop our statistical tests in Section 5. The tests we consider are based on the observation that a mixture of Poisson distributions is "smooth", so if m_k as a function of k is not sufficiently smooth, this can be used as evidence for dependence. The tests are summarized in Theorem 11. We experimentally verify our tests in Section 6 on artificially generated data. In Section 7 we give an outlook on alternative ways of deriving iid tests for exchangeable data. Section 8 concludes.

In Appendix A, we state/derive a number of technical lemmas we require to derive our tests. For improving the power of our tests, in Appendix B we derive upper bounds on our test statistics analogous to Section 5 but without the Poisson approximation, i.e. directly for iid $\mathcal{X}$ or the multinomial distribution. Details on the multinomial and product of Poisson distributions and their relation can be found in Appendix C. They are used to derive upper bounds on the variance of our tests in the multinomial model. A list of notation can be found in Appendix D.

Unrelated work. Independence tests in the literature most often refer to testing whether a pair of random variables (X,Y) is independent, given a number of iid(!) sample pairs $\{(x_t,y_t)\}$ (mutual information and chi-square tests are popular). Our setup is totally different and much harder.

Another setup is stochastic processes. Dependence can be tested via estimating auto-correlation coefficients, but this requires ordered data and $\mathcal{X}=\mathbb{R}$. One could use some other independence test on the pairs $\{(x_{t-1},x_t)\}$ without the $\mathcal{X}=\mathbb{R}$ assumption to test a Markov vs iid hypothesis, and/or adapt auto-correlation tests to unordered data. We briefly remark on this in Section 7.

2 Examples & Potential Applications

In this section we will provide some motivating examples and potential applications. This will also provide some intuition why rejecting the hypothesis H_{iid} that data is iid is possible at all, but also the difficulty from not having any more structure available. We consider biased coin flips (binomial process), Black Jack, and data duplication. We also discuss the relevance to machine learning, whose dominant training paradigm still operates under the iid assumption.

Binomial. Consider a binary sequence $x_{1:n}=x_1x_2...x_n\in\{0,1\}$ of length $n=1000$, say. If $x_{1:n/2}=1^{n/2}$ and $x_{n/2+1:n}=0^{n/2}$ we confidently reject the hypothesis H_{iid} that $x_{1:n}$ was sampled i.i.d. But we are unlikely to observe such a sequence if $x_{1:n}$ is well-shuffled (sampled from an exchangeable process). If we shuffle 500 ones and 500 zeros, a typical $x_{1:n}=0111100101...0100100100$ looks random, or does it? There are exactly $n_0=n_1=n/2=500$ ones and zeros. While for a fair coin, we expect *about* $n/2$ ones, would or should you believe anyone telling you that this is a sequence of fair coin flips? The probability of observing *exactly* $n/2$ ones in a sequence of n fair coin flips is around $\sqrt{2/\pi n}=2.5\%$, so a test for $N_1\stackrel{?}{=}n/2$ would confidently reject the hypothesis that the $x_{1:n}$ above arose from a fair coin.

What about $n=1'000'000$ and $n_1=314'159$. Obviously this is not from a fair coin, but our aim is to test for iid, not fairness. Could such a sequence have been the result of a biased coin? Since we assume the bits to be perfectly shuffled, n_1 is a sufficient statistic, so any test plausibly should only depend on n_1, and not the sequence $x_{1:n}$ itself. The probability that $N_1=n_1$ is ≤ 0.00086 for a coin of any bias, i.e. test $N_1\stackrel{?}{=}n_1$ would reject H_{iid}. Of course, tests have to be designed before observing the data, and a-priori $n_1=314'159$ is unlikely, so such a test is unlikely to have any power. We can of course combine tests and apply a union bound, but not too many, otherwise the tests become too weak. $n/2$ seems special, so maybe we should put such a test in the mix, but what about $314'159$? It's the first 6 digits of π. Maybe this is too much numerology, but what about testing for prime n_1? The density of primes p is around $1/\ln p$, so a-priori we should expect an n_1 around $3\cdot 10^5$ to be composite (with confidence $1-1/\ln(n_1)\doteq 93\%$). Maybe this is just not enough to reject H_{iid}, but we could always up the numbers.

Imagine n so large that the binary representation of n_1 contains some encrypted message or a long segment of Chaitin's number of wisdom. In general, finding every pattern in n_1 is an AI-complete problem. There are universal tests which in principle could test for all such eventualities, and in some situations practical approximations thereof can be very powerful. We discuss them briefly in Section 7, but we were not

able to make them work as well as the specific tests develop in this paper, so will not consider universal tests any further (nor will we delve into numerology any further).

Black Jack. A standard deck of cards without Jokers consists of 52 cards of 13 ranks, each in four suits, two red and two black. If we shuffle together infinitely many such decks and then draw n cards, this equivalently to drawing cards uniformly iid from the $|\mathcal{X}|=52$ different card faces. If we have only one deck and draw all 52 cards from it, we obviously observe every card exactly once. Such an outcome would be extremely unlikely had we drawn 52 cards from an infinite set of decks (see (5)).

Assume now an unknown number of decks have been shuffled together. Assume n cards have been drawn so far from this pile and we remembered their face (called card counting strategy). An interesting question is to infer the number of decks the cards have been drawn from. Or consider the weaker question: Are $x_{1:n}$ consistent with H_{iid}? If yes, we cannot infer the next card face better than by chance (1/52), so should not waste our time trying to do so, and wait with raising the stakes for when we have seen more cards. The answer to this question is relevant even if we know the number of decks c. For Black Jack, 1-8 decks are used, in casinos often 6. If n is small, we will not be able to reject H_{iid}, but if we have seen all cards, then each card will appear exactly c times, again ruling out H_{iid}. If we are close to the end of the pile, most faces will have appeared c times, none more, and only a few significantly less. Even mid-way through the pile, each face has appeared at most c times, which is evidence against H_{iid} for small c. 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). That is, latest half-way through a single deck, this fact is revealed. If cards are drawn from 2 decks, more than 52 cards are needed to reveal that they are not iid (Figure 3 bottom right).

Data duplication. In modern Machine Learning, esp. Deep Learning, data $x_{1:n}$ is abundant (large n) and observation spaces are huge (large $\mathcal{X}$). For instance, ImageNet consists of over 14 million images, usually resized or cropped to e.g. 224×224 pixels of $256\times256\times256$ colors, i.e. $\mathcal{X}=256^{3\times224\times224}$. We can as well assume that $\mathcal{X}$ is infinite. Assume $x_{1:n}$ contains no duplicate images and is well shuffled. As we will show later, no valid test can reject H_{iid} in this case. But if $x_{1:n}$ contains duplicates one may be able to reject H_{iid}, similarly to the Black Jack example above, even without knowing anything about the observation space $\mathcal{X}$. For instance, assume every observation is duplicated, i.e. every x that appears in $x_{1:n}$ appears exactly twice. For uncountable $\mathcal{X}$, if x is sampled from a probability *density*, the probability of sampling the same x twice is 0. So duplications can only happen if $P_{\boldsymbol{\theta}}$ contains point masses, i.e. is not purely continuous, i.e. $P_{\boldsymbol{\theta}}(x)>0$ for some x. For finite or countable $\mathcal{X}$, this is necessarily true. But if $P_{\boldsymbol{\theta}}(x)>0$, then the frequency of seeing x is binomially distributed. While seeing some x twice is plausible, seeing *all* x exactly twice is very unlikely, so we can reject H_{iid}: If data is iid and some items are duplicate, we should also see triples and quadruples, etc.

Relevance for machine learning. The predominant training and evaluation protocol in Machine Learning in general and Deep Learning in particular is still to assume the data is iid, train on most of the data and evaluate on the rest. Interestingly, this is true even for models dealing with definitely non-iid text. For instance, for Transformers,

text is crudely chopped into chunks of equal length and then shuffled. The empirical test loss is an unbiased estimator of the true loss, so is a proper way of comparing the performance of different models. Data sizes in modern machine learning are huge, so that even 10% held-out data is so much that test noise is often of little concern. That's at least the general story.

But in Deep Learning, data these days is often scraped from the whole internet, and duplications abound. For instance, assume the whole data set contains 3 copies of each data item. If we randomly split off 10% as the test set, then the train set contains nearly all (99%) of the test set items. With heldout-validation a pure memorizer without any generalization capacity will perform nearly perfectly on the test set [BLH22], but will fail in practice on future data. Indeed shuffling the data makes this problem the worst [SEBF21, GB19]. The problem is known as test set contamination, and well known.

The standard solution is to decontaminate or clean the data, e.g. removing duplicates, but this does not suffice. One has to remove *approximate* duplicates too. But at what threshold for example should a document that cites a training-set document verbatim be removed from the test-set? When are two images scraped from the internet rescaled or cropped or jpeg compressed versions of the same image, and even if so, should they be regarded duplicates? While approaches exist that meliorate the problem, in theory this problem is ill-defined [Hut06, FAQ], and in practice a huge, actually AI-complete, problem [BMR+20, App.C].

Test set evaluation is empirically sound for iid data, therefore the failure of this paradigm must be attributed to the non-iid nature of the data. The strength and weakness of the iid tests developed in this paper are that they are completely model- and data-agnostic. This makes them universally applicable and valid, but also very weak. On the other hand, as discussed above, finding good/perfect model- and data-type-sensitive tests is itself a difficult/impossible research question beyond the scope of this article.

3 Problem Formalization and Preliminaries

We now introduce notation and concepts used throughout the paper: general notation, the multinomial and Binomial distributions, first-order and second-order counts, exchangeable distributions, and statistical tests. The reader familiar with these concepts could skim this section to just pick up the notational convention used in this article.

Notation. We use calligraphic upper letters such as $\mathcal{X}$ for sets and $|\mathcal{X}|$ or $\#\mathcal{X}$ for the size of $\mathcal{X}$. Probability spaces are denoted by $(\mathcal{X}^n,\Sigma,P)$ with $d=|\mathcal{X}|\in\mathbb{N}\cup\{\infty\}$. $P_{\boldsymbol{\theta}}$ denotes iid distributions. Q denotes exchangeable distributions. Capital letters X,N,M,T,E,O,D,C,U,V,Z,Y denote random variables, and corresponding lower case letters samples corresponding to them. We will use the shorthand $P(x):=P[X=x]$ and similarly for other random variables. The variance of Z is $\mathbb{V}[Z]:=\mathbb{E}[Z^2]-\mathbb{E}[Z]^2$ and the covariance of Y and Z is $\mathrm{Cov}[Y,Z]:=\mathbb{E}[YZ]-\mathbb{E}[Y]\mathbb{E}[Z]$. In addition to the classical $O()$ notation, $O_P(f(n))$ denotes all random functions $F(n)$ for which $\forall\delta>0\ \exists c>0\ \forall n$: $P[|F(n)|\geq c\cdot|f(n)|]\leq\delta$. We use $f(n)\lesssim g(n)$ to denote $f(n)\leq g(n)\cdot[1\pm O_P(1/\sqrt{n})]$, and similarly $\gtrsim$ and $\eqsim$. We use $\lesssim$ even when stronger asymptotic or non-stochastic bounds would be possible, since we mostly care about the leading-order approximation in n and

not the approximation error as long as it tends to zero almost surely for $n\to\infty$. See Appendix D for more standard notation and beyond.

Multinomial distribution. Let $X_{1:n}:=(X_1,...,X_n)$ be $\mathcal{X}$-valued random variables. Let $x_{1:n}$ be sampled from some probability distribution P, where $x_t\in\mathcal{X}$ for $t\in\{1:n\}:=\{1,...,n\}$. Though we are interested in general measurable (finite, infinite, uncountable) $\mathcal{X}$, we will show in Section 4 that without loss of generality we can and hence will assume that $\mathcal{X}=\{1:d\}$ or $\mathcal{X}=\mathbb{N}$ ($d=\infty$), and $\Sigma=2^{\mathcal{X}}$. We will use the shorthand $P(x_{1:n}):=P[X_{1:n}=x_{1:n}]$ and $P(x_t)=P[X_t=x_t]$ and similarly for other random variables. The null hypothesis $H_{\text{iid}}=\{P_{\boldsymbol\theta}\}$ is that $X_{1:n}$ are i.i.d. In this case $P=P_{\boldsymbol\theta}$ for some $\boldsymbol\theta\in[0;1]^d$ with $\sum_{x=1}^d\theta_x=1$ and $P_{\boldsymbol\theta}(x):=\theta_x$. Let $n_x:=\#\{t:x_t=x\}\in\{0:n\}$ be the number of times x appears in $x_{1:n}$, called (first-order) *item count*, and $N_x:=\#\{t:X_t=x\}$ be the corresponding random variables. Then $P_{\boldsymbol\theta}$ can be written as

$$P_{\boldsymbol\theta}(x_{1:n}) \;=\; P_{\boldsymbol\theta}(x_1)\cdot....\cdot P_{\boldsymbol\theta}(x_n) \;=\; \theta_{x_1}\cdot....\cdot\theta_{x_n} \;=\; \prod_{x=1}^d\theta_x^{n_x}$$

This expression is independent of the order of $x_{1:n}$, which leads to the multinomial distribution

$$P_{\boldsymbol\theta}[N_{1:d}=n_{1:d}] \;=\; \binom{n}{n_1,...,n_d}\prod_{x=1}^d\theta_x^{n_x} \tag{1}$$

In particular the probability of event $N_x=k$ has a binomial distribution (see Appendix C for further details):

$$\begin{aligned} P_{\boldsymbol\theta}[N_x=k] \;&=:\; P_{\theta_x}(k) \;=\; f_k^n(\theta_x) \\ f_k^n(\theta) \;&:=\; P_\theta(k) \;=\; \tbinom{n}{k}\theta^k(1-\theta)^{n-k} \end{aligned} \tag{2}$$

Poisson distribution.

Lemma 1 (Poisson distribution). *The Poisson(λ) distribution $P_\lambda(k):=\lambda^k e^{-\lambda}/k!=:g_k(\lambda)$ for $\lambda\geq 0$ and $k\in\mathbb{N}_0$ has the following properties: $\mathbb{E}[k]=\mathbb{V}[k]=\lambda$. For fixed k it is unimodal in λ with maximum at $\lambda^*=k$ and $\max_\lambda P_\lambda(k)=P_k(k)=(1-\dot\varepsilon_k)/\sqrt{2\pi k}$ and $0\leq\dot\varepsilon_k\to 0$ (16).*

With a slight overload in notation, let $P_{\boldsymbol\lambda}(\boldsymbol n):=\prod_{x=1}^d\lambda_x^{n_x}e^{-\lambda_x}/n_x!$ for $\boldsymbol n\in\mathbb{N}_0^d$ be a product of d independent but not identical Poissons, where $\boldsymbol n\equiv n_{1:d}$. It is well-known that $P_{\boldsymbol\theta}(n_{1:d})=P_{\boldsymbol\lambda}[N_{1:d}=n_{1:d}|N=n]$ for $\boldsymbol\lambda=n\boldsymbol\theta$ and $N:=N_+=N_1+...+N_d$. The mean and variance of Poisson(λ) are both λ, hence $\mathbb{E}_{\boldsymbol\lambda}[N]=\sum_x\mathbb{E}_{\boldsymbol\lambda}[N_x]=\sum_x\lambda_x=n$ and similarly $\mathbb{V}_{\boldsymbol\lambda}[N]=n$ (using independence). This means that $N=n\pm O_P(\sqrt{n})$ is close to n for large n, and therefore under certain conditions, $P_{\boldsymbol\theta}[N_x=k]\approx P_{\boldsymbol\lambda}[N_x=k]$ even without conditioning on $N=n$. For instance, $\mathbb{E}_{\boldsymbol\theta}[N_x]=n\theta_x=\lambda_x=\mathbb{E}_{\boldsymbol\lambda}[N_x]$ and $\mathbb{V}_{\boldsymbol\theta}[N_x]=n\theta_x(1-\theta_x)\leq n\theta_x=\lambda_x=\mathbb{V}_{\boldsymbol\lambda}[N_x]$. Unfortunately, for the events R we care about, it is extremely cumbersome to quantify the relation $P_{\boldsymbol\theta}(R)\approx P_{\boldsymbol\lambda}(R)$. We will do so in Appendix C. In the main Section 5 we adopt a simpler approach: Noting that N is itself Poisson(n) distributed,

$$\begin{aligned} P_{\boldsymbol\theta}(\boldsymbol n) \;&=\; P_{\boldsymbol\lambda}(\boldsymbol n)/P_{\boldsymbol\lambda}[N=n] \;=\; c_n\cdot P_{\boldsymbol\lambda}(\boldsymbol n) \\ P_{\boldsymbol\lambda}[N=n] \;&=\; P_n(n) \;=\; n^n e^{-n}/n! \;=:\; 1/c_n \;\eqsim\; 1/\sqrt{2\pi n} \end{aligned} \tag{3}$$

Hence, for any event R, we can upper $P_{\boldsymbol\theta}[R]\leq c_n\cdot P_{\boldsymbol\lambda}[R]$. For our tests, $P[R]$ is typically exponentially small in n, so the blow-up by $c_n=O(\sqrt{n})$ is insignificant in theory: Increasing sample size n to $n+O(\log n)\approx n$ cancels c_n. We therefore can and will treat the item counts $N_1,...,N_d$ as independent Poisson distributed $N_x\sim P_{\lambda_x}$, i.e. $\boldsymbol{N}\equiv N_{1:d}\sim P_{\boldsymbol\lambda}$, which greatly facilitates the developments of our tests. In Appendices B and C we show that c_n can directly be replaced by 1 in many cases of practical interest.

Second-order count multiplicity. Let $m_k:=\#\{x:n_x=k\}$ be the number of x that appear k times in $x_{1:n}$, called (second-order) *count multiplicities*, and $M_k:=\#\{x:N_x=k\}$ be the corresponding random variables. Note that $M_k=0$ for $k>N$ but $M_k=0$ also for many $k\leq N$ due to $\sum_{k=0}^\infty k\cdot M_k=N$ and $\sum_{k=0}^\infty M_k=d$. Let $M_+:=\sum_{k=1}^\infty M_k=\#\{x:N_x>0\}=\#\{X_1,...,X_n\}$ be the number of different X_t, not counting multiplicities. We are mostly interested in $d=\infty$, in which case $M_0=\infty$ is not a useful statistic. We therefore exclude M_0 in $\boldsymbol{M}:=M_{1:n}$.

Exchangeable distributions. Non-iid distributions will be denoted by Q. A distribution Q is exchangeable if it is invariant under permutations of $x_{1:n}$, i.e. $Q(x_{1:n})=Q(x_{\pi(1:n)})$, where $\pi\in S_n$ is any permutation of $1:n$. As in the iid-case, Q only depends on the counts $\boldsymbol{n}$. Let $\mathcal{Q}$ be the class of all exchangeable distributions Q.

For instance, for $d=2$, Laplace's rule $Q(x_{t+1}|x_{1:t})=(\#\{\tau\leq t:x_\tau=x_{t+1}\}+1)/(t+2)$ has exchangeable but non-iid distribution $Q(x_{1:n})=n_1!n_2!/(n+1)!$. Similarly for the Good-Turing and Ristad distributions [Hut18].

Exchangeable distributions occur naturally as follows: Assume $X'_{1:n}$ are drawn from an arbitrary distribution Q', and then perfectly shuffled such as to destroy any order information. Formally, $X_{1:n}=X'_{\Pi(1:n)}$, where Π is drawn uniformly from all permutations S_n. It is easy to see that $X_{1:n}$ are exchangeable random variables. In particular exchangeable $X_1,...,X_n$ are identically distributed, i.e. $Q[X_t=x]=Q[X_{t'}=x]$.

Invariant statistical tests. A (valid) statistical test of significance $0<\alpha<1$ is a reject region $R\subset\mathcal{X}^n$ such that $P_{\boldsymbol\theta}[R]\leq\alpha\ \forall\boldsymbol\theta$. We can reject the hypothesis $H_{\rm iid}$ that $\boldsymbol{x}\equiv x_{1:n}$ is iid with confidence $1-\alpha$ *iff* $\boldsymbol{x}\in R$, that is, $H_{\rm iid}$ is falsely rejected (Type I error) with probability at most α. Reject regions are most often defined via a test statistic $T:\mathcal{X}^n\to\mathbb{R}$ and $R=\{\boldsymbol{x}:T(\boldsymbol{x})>c\}$ for some critical value $c\in\mathbb{R}$. T at critical level $c_\alpha:=\inf\{c:\sup_{\boldsymbol\theta}P_{\boldsymbol\theta}[T(\boldsymbol{X})>c]\leq\alpha\}$ has significance α. The p-value of a test T for data $\boldsymbol{x}$ is $p:=\sup_{\boldsymbol\theta}P_{\boldsymbol\theta}[T(\boldsymbol{X})>T(\boldsymbol{x})]$ is the smallest level α at which we can reject $H_{\rm iid}$: T can reject $H_{\rm iid}$ with confidence $1-p$.

Since we assume $X_{1:n}$ are exchangeable (shuffled), it is natural to ask for a test to reject $H_{\rm iid}$ independently of the order in which $X_1,...,X_n$ are presented. That is, T should be a function of the item counts $N_{1:d}$ only.

Furthermore, we do not want to make any structural assumptions on $\mathcal{X}$. While each $Q\in\mathcal{Q}$ is *not* necessarily invariant under permutations of elements of $\mathcal{X}$, the class $\mathcal{Q}$ itself is. Since we want to test against all $\mathcal{Q}$, it is natural to consider tests that are not affected by permuting $\mathcal{X}$, that is, $T(X_{1:n})=T(\pi(X_1),...,\pi(X_n))$, where this π is any permutation of elements in $\mathcal{X}$. Combining both invariances, we must have $T(X_{1:n})=T(N_{1:d})=T(N_{\pi(1:d)})$. T is invariant under reordering of $N_{1:d}$ *iff* it only depends on $M_0,...,M_n$. It may be possible to make an argument for order-independent tests that

among the most powerful tests w.r.t. to some invariant sub-class of $\mathcal{Q}$ there is always an invariant test, i.e. they include all minimax optimal tests.

Definition 2 (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.*

The power of tests. Neyman-Pearson use alternative hypotheses to determine the power $\beta=Q[T>c]$ of a test T for Q ($1-\beta=$Type II error). In our case, the alternative hypothesis is the set of exchangeable distributions without the iid distributions $H_{\overline{\text{iid}}}:=\mathcal{Q}\setminus H_{\text{iid}}$. There are no uniformly most powerful (UMP) tests for $H_{\overline{\text{iid}}}$, not even close; $H_{\overline{\text{iid}}}$ is too broad. Each test will have high power for some subset of $\mathcal{Q}$ and low power for other $Q\in\mathcal{Q}$. We do not formally define "interesting" subsets of $\mathcal{Q}$ and derive the power of tests for them or find UMPs for these subsets. We focus on developing tests which have known small (upper bound on the) Type I error α = small size α = significance level α = probability of falsely rejecting H_{iid} when it is actually true. We therefore rarely mention $\mathcal{Q}$, so unless explicitly mentioned to the contrary, distributions and sampling refers to iid or multinomial or Binomial. Our work is closer in spirit to Fisher hypothesis testing without alternative hypothesis, but we do demonstrate the power of the tests empirically in Section 6 on some hand-selected Q.

4 Reducing general $\mathcal{X}$ to $\mathbb{N}$

In this section we discuss general probability spaces $(\mathcal{X}^n,\Sigma,P)$, only to discover that we can without loss of generality restrict our analysis to $\mathcal{X}=\mathbb{N}$ and $\mathcal{X}=\{1:d\}$ with discrete σ-algebra $\Sigma=2^{\mathcal{X}}$. The only assumption we have to make on Σ is that it contains all singletons, $\{\boldsymbol{x}\}\in\Sigma\ \forall\boldsymbol{x}\in\mathcal{X}^n$, in order for the events $\boldsymbol{X}=\boldsymbol{x}$ to be measurable. For example, every T1 or Hausdorff space provided with the Borel sets satisfies this, in particular $\mathbb{R}^n$.

Infinite $\mathcal{X}$. So far we have considered finite and countable $\mathcal{X}$. Consider now $\mathcal{X}^n=\mathbb{R}^n$ with joint Gaussian density $\rho=\text{Gauss}(\mathbf{0},\boldsymbol{\Xi})$. The probability that $x_{1:n}$ contains repetitions is zero, hence $x_1,...,x_n$ are all different, so $M_0=\infty$, $M_1=n$, $M_k=0\ \forall k\geq 2$. Since an invariant test T is a function of $M_{0:n}$ only, there is a constant c' such that almost surely $T(M_{0:n})=c'$, hence $Q_\rho[T>c]$ is identically 0 or identically 1, i.e. the same for all $\boldsymbol{\Xi}$. It cannot be 1, since T must satisfy $Q_\rho[T>c]\leq\alpha<1$ for iid Q_ρ ($\Xi_{tt'}\propto[\![t=t']\!]$), but then T never rejects H_{iid}, even if ρ is non-iid and maximally correlated ($\Xi_{tt'}=1\ \forall tt'$). In general, any uncountable $\mathcal{X}$ can be equipped with a σ-algebra and non-atomic measure, leading to the same conclusion. Now consider $\mathcal{X}=\varepsilon\mathbb{Z}$ and discretize ρ. For $\varepsilon\to 0$ the conclusion still holds. Formally, for every $\delta>0$ there exists an $\varepsilon>0$, such that all $x_1,...,x_n$ are different with probability at least $1-\delta$. Hence the conclusion also holds for countably infinite $\mathcal{X}$.

Proposition 3 (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}. In particular, $X_{1:n}$ are almost surely all different if sampled from a non-atomic measure, e.g. if the measure has a density w.r.t. to the Lebesgue measure on $\mathcal{X}=\mathbb{R}^d$.*

Reduction of $\mathcal{X}$ to $\mathbb{R}$. Let $\mathcal{X}_{pp}:=\{x\in\mathcal{X}:P[\{x\}]>0\}$, which is countable, $\beta:=P[\mathcal{X}_{pp}]$, then $P_{pp}[A]:=P[A|\mathcal{X}_{pp}]$ is a pure point measure and $P_{\overline{pp}}[A]:=P[A|\mathcal{X}\setminus\mathcal{X}_{pp}]$ is non-atomic:

$$P[A] \;=\; \beta\cdot P_{pp}[A] \;+\; (1-\beta)\cdot P_{\overline{pp}}[A]$$

i.e. every measure P can be decomposed into a pure point measure and a non-atomic rest.

Consider now a point measure $\tilde{P}_{pp}$ on $\mathbb{R}$ with $\tilde{P}_{pp}[\{2i\}]:=P_{pp}[\{y_i\}]$ and zero elsewhere, where $\mathcal{X}_{pp}=\{y_1,y_2,...\}$ is some enumeration of elements in $\mathcal{X}_{pp}$. Now define

$$\tilde{P}[A] \;=\; \beta\cdot\tilde{P}_{pp}[A] \;+\; (1-\beta)\cdot\text{Gauss}_{0,1}[A] \tag{4}$$

As far as the second-order counts $\boldsymbol{M}:=(M_1,...,M_n)$ and $M_0=\infty$ are concerned, $\tilde{P}[\boldsymbol{M}=\boldsymbol{m}]=P[\boldsymbol{M}=\boldsymbol{m}]$, since for the discrete part we bijected $\mathcal{X}_{pp}$ to (a subset of) $2\mathbb{N}$ with same probability mass, and $\boldsymbol{M}$ is invariant under such bijection. As for the non-atomic part, in both cases, we almost surely each time sample a novel x not seen before, i.e. only M_1 is affected and increases by 1 with probability $1-\beta$. That is, we can restrict ourselves to measures on $\mathbb{R}$ of the form (4):

Proposition 4 ($\mathcal{X}=\mathbb{R}$ suffices). *For every invariant test T,*
$P[T>c]\leq\alpha$ for iid P on $\mathcal{X}$ $\Longleftrightarrow$ $\tilde{P}[T>c]\leq\alpha$ for iid $\tilde{P}$ on $\mathbb{R}$ of the form (4).

Note that T only depends on the counts $\boldsymbol{M}$ and M_0, so the same T is defined across every $\mathcal{X}$, and gives the same result independent from which infinite space $\{x_1,...,x_n\}\subseteq\mathcal{X}$ came from.

Reduction of $\mathbb{R}$ to $\mathbb{N}$. Consider $\tilde{P}_\ell[\{2i-1\}]:=1/\ell$ for $i=1,...,\ell$ and 0 on $\mathbb{R}$ else. Assume we draw n iid samples from $\tilde{P}_\ell$. The probability of sampling some x twice is

$$\tilde{P}_\ell[\exists t\neq t':X_t=X_{t'}] \;\leq\; \sum_{t\neq t'}\tilde{P}_\ell[X_t=X_{t'}] \;=\; \sum_{t\neq t'}\frac{1}{\ell} \;\leq\; \frac{n^2}{\ell} \tag{5}$$

That is, for fixed n, the probability for all x being unique ($P_\ell[\boldsymbol{M}=(n,0,0,...)]$) tends to 1 for $\ell\to\infty$. Combining this with the point measure above implies for all $\boldsymbol{m}:=m_{1:n}$,

$$\beta\cdot\tilde{P}_{pp}(\boldsymbol{m})+(1-\beta)\cdot\tilde{P}_\ell(\boldsymbol{m}) \quad\longrightarrow\quad \tilde{P}(\boldsymbol{m}) \quad\text{for}\quad \ell\to\infty \tag{6}$$

Proposition 5 ($\mathcal{X}=\mathbb{N}$ suffices). *For every invariant test T and infinite $\mathcal{X}$,*
$P[T>c]\leq\alpha$ for all iid P on $\mathcal{X}$ $\quad\Longleftrightarrow\quad$ $\tilde{P}[T>c]\leq\alpha$ for all iid $\tilde{P}$ on $\mathbb{N}$.

This justifies our restriction to finite $\mathcal{X}=\{1:d\}$ and countable $\mathcal{X}=\mathbb{N}$ ($d=\infty$).

Finite $\mathcal{X}$. *Embedding finite $\mathcal{X}$ into infinite $\mathcal{X}$:* Observing (only) $x_{1:n}$, we we do not want to make any assumption from which space $\mathcal{X}$ they have been sampled from. Obviously $|\mathcal{X}|\geq M_+$ is needed. But any P on a finite domain, say $\{1:d'\}$, can be extended to infinite $\mathcal{X}$ by setting $P[\mathcal{X}\setminus\{1:d'\}]=0$ without affecting $\boldsymbol{M}$, i.e. infinite $\mathcal{X}$ also contain all finitely supported P. So if a test has confidence $1-\alpha$ for $|\mathcal{X}|=\infty$, then it also has confidence at least $1-\alpha$ for $|\mathcal{X}|<\infty$. The converse however is not true:

Knowing $\mathcal{X}$ is finite and its size d: While $d=\infty$ includes all measures that have finite support, i.e. $\theta_x=0$ for all $x>d'$, *knowing* that d is finite provides extra information, so

the $d=\infty$ analysis does *not* automatically include the case where d is known and finite. While the $d=\infty$ tests remain valid for $d<\infty$, stronger tests are possible for $d<\infty$. The reason is that M_0 depends on $|\mathcal{X}|$.

Example: Recall that no invariant test can reject H_{iid} if all x_t are different ($M_1=n$), but this relied on $\mathcal{X}$ being infinite. On the other hand, if we know/assume $|\mathcal{X}|=n$, then the probability of seeing every $x\in\mathcal{X}$ exactly once is

$$P_{\boldsymbol{\theta}}[\forall t\neq t':X_t\neq X_{t'}] \;=\; n!\cdot\prod_{x\in\mathcal{X}}\theta_x \;\leq\; n!\cdot\max_{\boldsymbol{\theta}}\prod_{x\in\mathcal{X}}\theta_x \;=\; n!\prod_{x\in\mathcal{X}}\frac{1}{n} \;=\; \frac{n!}{n^n} \;\leq\; \sqrt{n}\cdot e^{1-n}$$

This is extremely small for large n, so H_{iid} can be rejected with very high confidence. In particular having observed $x_{1:n}$ and knowing nothing about $\mathcal{X}$, we cannot choose $|\mathcal{X}|=d=\#\{x:N_x\geq 1\}\equiv M_+=\sum_{k=1}^n M_k$, even if we were somehow able to deal with the fact that such d would itself be random.

Approximating infinite $\mathcal{X}$ by finite $\mathcal{X}$: While we made the case for countably infinite $\mathcal{X}$, for fixed n and the approximate results we aim at, we actually do not need to consider $d=\infty$, but $\ell=d=n^3$ suffices: We sort x in decreasing order of $P[\{x\}]$, truncate $\mathcal{X}_{pp}$ to $\frac{1}{2}n^3$ elements and choose $\ell=\frac{1}{2}n^3$ (see above). So any potential complications from $d=\infty$ can easily be avoided, but it turns out that mostly $d=\infty$ is more convenient.

5 I.I.D. Tests

We are finally in a position to develop some tests. We first outline the common idea behind all tests developed in this section. In Section 7 we discuss alternative approaches. We then derive a couple of tests that feel natural. Although they follow a common theme, they are quite diverse in the sense that every test highlights a new or different feature or power or technical difficulty. The most basic test uses a single M_k, all others are linear combinations thereof, except the last one, which is a logarithmic combination. The even and odd tests E and O are global sums of M_k over all even/odd k. The slope test $D_k=M_k-M_{k-1}$ is a bit more difficult to derive but also allows for a lower bound test. The curvature test $C_k=2M_k-M_{k-1}-M_{k+1}$ and its logarithmic version $\bar{U}_k=\ln(M_k^2/M_{k-1}M_{k+1})$ can be very strong. Also, while some tests require to use the empirical variance (E,O,$\bar{U}_k$), for others a theoretical upper bound is possible (D_k,C_k,M_k) and better (M_k). P, $\mathbb{E}$, $\mathbb{V}$, $\mathbb{W}$ are w.r.t. the mixture of Poisson distributions $P_{\boldsymbol{\lambda}}$ (3), which approximates $P_{\boldsymbol{\theta}}$.

The general idea behind the tests. First note that the Poisson distribution $P_\lambda(k)=\lambda^k e^{-\lambda}/\Gamma(k+1)$ is smooth if we take the liberty of plugging in $k\in\mathbb{R}$ (see e.g. $\mathbb{E}[M_k]$ curve in Figures 2 left for "uniform"). It has a unique maximum at $k=\lambda$ and is log-concave, so a rather benign function. For large λ, $P_\lambda(k)$ is also "smooth" in $k\in\mathbb{N}_0$ in the sense that its finite-difference approximations of slope and curvature (and higher) are small. It still has a unique maximum at $k\approx\lambda$ and is log-concave, and indeed approximately Gaussian with mean and variance λ. $P_\lambda(k)$ is also differentiable in λ, which we will also exploit. Now consider

$$\begin{aligned}\mathbb{E}[M_k] \;&=\; \mathbb{E}[\#\{x:N_x=k\}] \;=\; \mathbb{E}\textstyle\sum_x[\![N_x=k]\!] \\ &=\; \textstyle\sum_x P_{\boldsymbol{\lambda}}[N_x=k] \;=\; \sum_x P_{\lambda_x}(k) \;=\; \sum_x g_k(\lambda_x)\end{aligned} \tag{7}$$

That is, $\mathbb{E}[M_k]$ is a sum of Poisson(λ_x) distributions. Depending on the distribution of λ_x, $\mathbb{E}[M_k]$ as a function of k may have multiple extrema, but as a mixture of Poissons it cannot be less smooth and typically is even more smooth (see e.g. $\mathbb{E}[M_k]$ curve in Figures 2 left for "linear" mixture). 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. Since invariant tests can only depend on $\boldsymbol{M}$, they must test for some such property. For instance, no Poisson and hence no mixture of Poissons can have $\mathbb{E}[M_k] = 0$ for all odd k (see e.g. Figures 2 left for "even-n"), so $M_k = 0$ for all odd k is strong evidence against $\boldsymbol{X}$ being iid.

Linear tests. Most of our tests are (signed) linear combinations of a subset of the M_k. The general template for upper bounds on the mean and variance is as follows:

Proposition 6 (Poisson upper bounds for linear tests). *Let $T = \sum_k \alpha_k M_k$ for $\alpha_k \in \mathbb{R}$. Provided all involved sums and integrals are absolutely convergent, we have $\tau := \mathbb{E}[T] \le n \cdot \sup_{\lambda>0} g(\lambda)/\lambda =: \tau^{ub}$, where $g(\lambda) := \sum_k \alpha_k P_\lambda(k) = \sum_k \alpha_k \lambda^k e^{-\lambda}/k!$, and $\mathbb{V}[T] \le \sum_k \alpha_k^2 \mathbb{E}[M_k] \le V^{ub}$, where $V^{ub} := \sum_k \alpha_k^2 \mu_k^{ub}$ with $\mu_k^{ub} \ge \mathbb{E}[M_k]$ upper bounding the expectations of M_k.*

Instead of upper bound V^{ub} as defined above, for some of our tests we use random $V^{ub} = \sum_k \alpha_k^2 M_k$, which is an upper bound in expectation justified by Lemma 17.

Proof. Using (7) and Lemma 12 we can upper bound

$$\mathbb{E}[T] \;=\; \textstyle\sum_k \alpha_k \sum_x P_{\lambda_x}(k) \;=\; \sum_x g(\lambda_x) \;\le\; n \cdot \sup_{\lambda>0} g(\lambda)/\lambda$$

For the variance, for $Z_k^x := [\![N_x = k]\!]$, we have $Z_k^+ = M_k$ and $Z_k^x Z_{k'}^x = 0$. Furthermore, $\mathrm{Cov}[Z_k^x, Z_{k'}^{x'}] = 0$ for $x \ne x'$ since N_x and $N_{x'}$, hence Z_k^x and $Z_{k'}^{x'}$ are independent. Hence Lemma 19 implies

$$\mathbb{V}[T] \;=\; \mathbb{V}[\sum_k \alpha_k Z_k^+] \;\le\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^+] \;=\; \sum_k \alpha_k^2 \mathbb{E}[M_k] \;\le\; \sum_k \alpha_k^2 \mu_k^{ub} \qquad \blacksquare$$

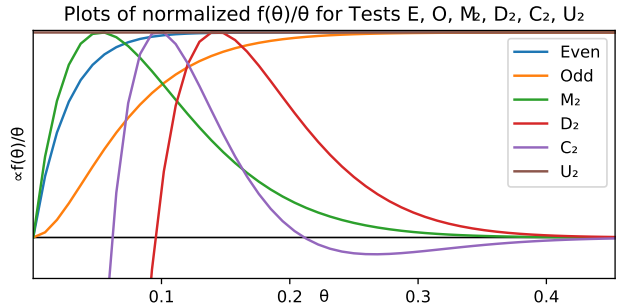

Figure 1: Normalized $g(\lambda)/\lambda \approx f(\theta)/\theta$ for Tests $E, O, M_k, D_k, C_k, \bar{U}_k$ for $k = 2$

Second-order count tests $\boldsymbol{M_k}$. We first determine upper bounds for the second-order counts M_k for each k separately. We can apply Proposition 6 with $\alpha_{k'} = [\![k' = k]\!]$ or just directly apply Lemma 12:

$$\begin{aligned} \mu_k \;:=\; \mathbb{E}[M_k] \;&=\; \sum_x g_k(\lambda_x) \;\le\; n \cdot \sup_{\lambda>0} \frac{\lambda^{k-1} e^{-\lambda}}{k!} \\ &=\; n \frac{(k-1)^{k-1} e^{-(k-1)}}{k!} \;=\; \frac{n}{k} \frac{1 - \dot{\varepsilon}_{k-1}}{\sqrt{2\pi(k-1)}} \;=:\; \mu_k^{ub} \end{aligned} \tag{8}$$

The bound follows from $g_k(\lambda)/\lambda = \lambda^{k-1}e^{-\lambda}/k!$ being maximal for $\lambda^* = k-1$ (cf. Figure 1). The approximate expression follows from Lemma 21 with $1/(12k+1) \le \dot{\varepsilon}_k \le 1/12k$. Note that for $k=1$ the bound is valid but vacuous. We see that the relative frequency of k-multiplicities $\bar{\mu}_k = \mu_k/n$ is upper bounded by $O(k^{-3/2})$, i.e. the expected *number* of such items is $k \cdot \mathbb{E}[M_k] = O(1/\sqrt{k})$. From Lemma 1 we know that for a single Poisson $\max_\lambda P_\lambda(k) \approx 1/\sqrt{2\pi k}$, hence a mixture of Poisson(λ)'s cannot be larger, which is consistent with the result above.

By Proposition 6 we (also) have $\mathbb{V}[M_k] \le \mathbb{E}[M_k] \le \mu_k^{ub}$, so by Lemma 17 with $Z_x = [\![N_x = k]\!] \in [0;1]$, the p-value for rejecting H_{iid} is

$$p \lesssim \Phi_n((\mu_k^{ub} - M_k)/\sqrt{\mu_k^{ub}}) \le \exp(-\tfrac{1}{2}n(\bar{M}_k - \bar{\mu}_k^{ub})^2/\bar{\mu}_k^{ub} + O(1)) = e^{-O(n/k^{3/2})}$$

We need to use $\Phi_n(y) := c_n \cdot \Phi(y)$ (see (3) and Definition 14) rather than Φ, since in reality $M_k \sim P_{\boldsymbol{\theta}}$ while $\mathbb{E}$ and $\mathbb{V}$ were w.r.t. $P_{\boldsymbol{\lambda}}$, so $p = P_{\boldsymbol{\theta}}[T(\boldsymbol{M}) > c] \le c_n \cdot P_{\boldsymbol{\lambda}}[T(\boldsymbol{M}) > c] \approx c_n \cdot \Phi(\cdots) =: \Phi_n(\cdots)$ for all of our tests T. The Φ_n-bound only holds if $M_k > \mu_k^{ub}$ and the exponential bound if furthermore n is sufficiently large. We could also have chosen $V^{ub} = M_k$, a random upper bound on $\mathbb{V}[M_k]$, but if $M_k < \mu_k^{ub}$, then $p \ge \frac{1}{2}$, so the test has no power, and if $M_k > \mu_k^{ub}$, then using μ_k^{ub} leads to a stronger test than using M_k.

Example 7. *Our running example for all tests will be a data set where 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 $p \lesssim \exp(-\frac{1}{2}n(\frac{1}{2} - \frac{1}{2e})^2/\frac{1}{2e}) \doteq e^{-0.271n}$. i.e. H_{iid} can be extremely confidently rejected for moderately large n. For $k \ne 2$, the tests have no power ($\bar{M}_k = 0 < \bar{\mu}_k^{ub}$).* ♢

Even and odd tests E and O. The above example suggests non-trivial upper bounds on the even and odd second-order counts M_k, and a test based on that, but we have to be a bit careful. For $d = \infty$, there are $d - n = \infty$ many unobserved $x \in \mathcal{X}$, hence $M_0 = \infty$. Similarly, for $\lambda_x = n/d$ and $d \to \infty$, every x is observed exactly once, hence $M_1 = n$ and all other $M_k = 0$, again not leading to a useful test. The general solution is to exclude M_0 and M_1. First, for $\alpha_k^{\text{even}} := k \cdot [\![k \ne 0 \text{ even}]\!]$ and $\alpha_k^{\text{odd}} := k \cdot [\![k \ne 1 \text{ odd}]\!]$ we define (see also Figure 1)

$$g_{\text{even}}(\lambda) := \sum_k \alpha_k^{\text{even}} P_\lambda(k) = \sum_{k \ne 0 \text{ even}} k \cdot P_\lambda(k) = \sum_{k \ne 0 \text{ even}} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \tfrac{\lambda}{2}[1 - e^{-2\lambda}]$$

$$g_{\text{odd}}(\lambda) = \sum_k \alpha_k^{\text{odd}} P_\lambda(k) = \sum_{k \ne 1 \text{ odd}} k \cdot P_\lambda(k) = \sum_{k \ne 1 \text{ odd}} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \tfrac{\lambda}{2}[1 - e^{-\lambda}]^2$$

The last expressions follow from pulling out a $\lambda e^{-\lambda}$ from the sum and recognizing the Taylor series expansion of $\sinh(\lambda)$ and $\cosh(\lambda)$. The even and odd test statistics

$$E := \sum_k \alpha_k^{\text{even}} M_k = \sum_{k \ne 0 \text{ even}} k \cdot M_k \quad \text{and} \quad O := \sum_k \alpha_k^{\text{odd}} M_k = \sum_{k \ne 1 \text{ odd}} k \cdot M_k.$$

Using (7), the expectation can be upper bounded by Proposition 6 as

$$\mathbb{E}[E] = \sum_{k \ne 0 \text{ even}} k \sum_x P_{\lambda_x}(k) = \sum_x g_{\text{even}}(\lambda_x) \le n \cdot \sup_{\lambda > 0} \tfrac{1}{2}[1 - e^{-2\lambda}] = \frac{n}{2} =: \varepsilon^{ub} = \bar{\varepsilon}^{ub} n$$

$$\mathbb{E}[O] = \sum_{k \ne 0 \text{ odd}} k \sum_x P_{\lambda_x}(k) = \sum_x g_{\text{odd}}(\lambda_x) \le n \cdot \sup_{\lambda > 0} \tfrac{1}{2}[1 - e^{-\lambda}]^2 = \frac{n}{2} =: o^{ub} = \bar{o}^{ub} n$$

That is, excluding singletons, we should expect at most half of the data items to appear evenly often, and at most half oddly often. The even/odd upper bounds are "attained" for $\lambda^* \to \infty$. Using Proposition 6 again, we can also upper bound the variances of E and O:

$$\mathbb{V}[E] \;\le\; \sum_k (\alpha_k^{\text{even}})^2\; \mathbb{E}[M_k] \;=\; \sum_{k\neq 0 \text{ even}} k^2 \mathbb{E}[M_k] \quad \text{and} \quad \mathbb{V}[O] \;\le \sum_{k\neq 1 \text{ odd}} k^2 \mathbb{E}[M_k]$$

Unfortunately no meaningful finite $\boldsymbol{\lambda}$-independent upper bounds on them are possible. Trying to use the same method as for upper bounding $\mathbb{E}[E]$ leads again to $\lambda^* \to \infty$, but this time the expression diverges. Note that for fixed $\boldsymbol{\lambda}$, the variance is finite; it just does not have a (non-vacuous) uniform upper bound. So in this case we have to resort to using the empirical upper bound $V^{ub}_{\text{even}} := \sum_{k\neq 0 \text{ even}} k^2 M_k$ for the variance, and similarly for $V^{ub}_{\text{odd}} := \sum_{k\neq 1 \text{ odd}} k^2 M_k$. By Lemma 17 and Lemma 20, the p-values for rejecting H_{iid} are

$$\begin{aligned} p \;&\lesssim\; \Phi_n((\varepsilon^{ub} - E)/\sqrt{V^{ub}_{\text{even}}}) \;\le\; \exp(-\tfrac{1}{2} n (\bar{E} - \tfrac{1}{2})^2 / \sqrt{\bar{V}^{ub}_{\text{even}}} + O(1)) \\ p \;&\lesssim\; \Phi_n((o^{ub} - O)/\sqrt{V^{ub}_{\text{odd}}}) \;\le\; \exp(-\tfrac{1}{2} n (\bar{O} - \tfrac{1}{2})^2 / \sqrt{\bar{V}^{ub}_{\text{odd}}} + O(1)) \end{aligned}$$

As before, the Φ_n-bound only holds if $E > \varepsilon^{ub}$ and $O > o^{ub}$; the exponential bound only holds if $\bar{E} > \frac{1}{2}$ and $\bar{O} > \frac{1}{2}$ and sufficiently large n.

Example 8. *In our running example in which each data item is doubled ($M_2 = n/2$), we have $E = n\bar{E} = n$ and $V^{ub}_{even} = 2n$, hence the even test has p-value $p \lesssim e^{-n/8\sqrt{2}}$, i.e. H_{iid} can be extremely confidently rejected for moderately large n. The odd test has no power ($O = 0$). If we triple each item, then $O = n/3$ and $E = 0$ and $V^{ub}_{odd} = 3n$, and the odd test has $p \lesssim e^{-n/8\sqrt{3}}$.* $\diamondsuit$

Slope tests $\boldsymbol{D_k := M_k - M_{k-1}}$. As mentioned at the beginning of the section, $\mathbb{E}[M_k]$ as a function of k is "smooth", so it is natural to test for a small difference=slope $D_k := M_k - M_{k-1}$. Let $D^x_k = [\![N_x = k]\!] - [\![N_x = k-1]\!]$, hence $D_k = \sum_x D^x_k$. Then similar to before

$$\begin{aligned} \delta_k \;&:=\; \mathbb{E}[D_k] \;=\; \sum_x P_{\lambda_x}[N_x = k] - P_{\lambda_x}[N_x = k-1] \;=\; \sum_x g_\delta(\lambda_x) \;\le\; n \cdot \sup_{\lambda > 0} \frac{g_\delta(\lambda)}{\lambda} =: n\bar{\delta}^{ub}_k \\ &\text{where} \quad g_\delta(\lambda) := \frac{\lambda^k e^{-\lambda}}{k!}\left[1 - \frac{k}{\lambda}\right] \end{aligned} \tag{9}$$

The maximum of $P_\lambda(k)$ is at $\lambda = k$ but the bracket $[1 - \frac{k}{\lambda}]$ kills this maximum, moving it to $\approx k + \sqrt{k}$ (Figure 1). Since $g_\delta(\lambda) \le 0$ for $\lambda \le k$, we can assume $\lambda > k$, hence

$$\frac{d}{d\lambda} \ln \frac{g_\delta(\lambda)}{\lambda} \;=\; \frac{d}{d\lambda}[(k-1)\ln\lambda - \lambda - \ln k! + \ln(1 - k/\lambda)] \;=\; \frac{k-1}{\lambda} - 1 + \frac{k/\lambda^2}{1 - k/\lambda} \;\stackrel{!}{=}\; 0$$

Multiplying with $\lambda^2(1 - k/\lambda)$ leads to a quadratic equation in λ which has two solutions $\lambda^*_\pm = k - \frac{1}{2} \pm \sqrt{k + \frac{1}{4}}$, only $\lambda^*_+ > k$ is valid, and is indeed the global maximum. Note that $g_\delta(\lambda) = 0$ for $k \ge 2$ but not for $k = 1$, so Lemma 12 and hence the following bound only

applies for $k \geq 2$. The reason is that D_1 involves M_0, but $M_0 = \infty$ for infinite $\mathcal{X}$, hence $D_1 = -\infty$ and as a test is vacuous. A tedious calculation shows that

$$\bar{\delta}_k^{ub} \;=\; \frac{g_\delta(\lambda_+^*)}{\lambda_+^*} \;=\; \frac{1-\ddot{\varepsilon}_k}{k^2\sqrt{2\pi e}}, \quad \text{where} \quad \ddot{\varepsilon}_k = O(1/\sqrt{k}) \tag{10}$$

That is, the slope of (a mixture of) Poissons is upper bounded by $\bar{\delta}_k^{ub} = O(1/k^2)$. This is smaller than μ_k^{ub} by a factor of $1/\sqrt{ke}$, so can lead to a stronger test than test M_k, provided that indeed M_k deviates from M_{k-1} sufficiently.

By Proposition 6, the variance of D_k can be upper bounded by $\mathbb{V}[D_k] \leq \mathbb{E}[M_k] + \mathbb{E}[M_{k-1}] = \mu_k + \mu_{k-1}$. We can theoretically upper bound this by $V_k^{ub} = \mu_k^{ub} + \mu_{k-1}^{ub}$ or empirically estimate it by $V_k^{ub} = M_k + M_{k-1}$. So by Lemma 17 with $Z_x = D_k^x \in [-1;1]$ the p-value for rejecting H_{iid} is

$$p \;\lesssim\; \Phi_n((\delta_k^{ub} - D_k)/\sqrt{V_k^{ub}}) \;\leq\; \exp(-\tfrac{1}{2}n(\bar{D}_k - \bar{\delta}_k^{ub})^2/\bar{V}_k^{ub} + O(1)) \;=\; e^{-O(n/k^{5/2})}$$

The empirical choice for V_k^{ub} can be smaller=better than the theoretical upper bound if the bound is loose, but can also be larger=worse, since $M_k + M_{k-1} \to \mathbb{E}[M_k + M_{k-1}] \leq \mu_k^{ub} + \mu_{k-1}^{ub}$ is only guaranteed in the iid case, but we precisely want to test for non-iid, in which case $M_k + M_{k-1}$ may be larger than $\mu_k^{ub} + \mu_{k-1}^{ub}$ even asymptotically. In all of our experiments, the empirical choice $V_k^{ub} = M_k + M_{k-1}$ performed better.

Example 9. *Testing our previous example where each data item is doubled ($D_2 = M_2 = n/2$), for $k = 2$ we have $\lambda_+^* = 3$, hence $\bar{\delta}_2^{ub} = f_2(\lambda_+^*)/\lambda_+^* = 1/2e^3 \doteq 0.0249$ and $\bar{\mu}_2^{ub} + \bar{\mu}_1^{ub} = 1/2e + 1 \doteq 1.184 > \frac{1}{2} = \bar{M}_2 + \bar{M}_1$, so in this case, using $\bar{M}_2 + \bar{M}_1$ as V_k^{ub} is indeed better, and $p \lesssim \exp(-\frac{1}{2}n(\frac{1}{2} - \frac{1}{2e})^2/\frac{1}{2}) = e^{-0.2257n}$. For $k \neq 2$, the D_k tests have no power ($\bar{M}_k = 0 < \bar{\mu}_k^{ub}$).* ♢

We can also lower bound D_k by upper bounding $-D_k$. The maximizing λ^* is then λ_-^* and $g_\delta(\lambda_-^*)$ is the same (apart from a minus sign) to leading order in k, and $p \lesssim \Phi_n((|\delta_k^{ub}| + D_k)/\sqrt{V_k^{ub}})$. In the example above, $D_3 = -M_2$ would have (the same) power as D_2.

Linear curvature tests $C_k := 2M_k - M_{k-1} - M_{k+1}$. As apparent from the graphs, the curvature of a (mixture of) Poisson as a function of $k \geq 2$ is also bounded, which gives us another test. Let

$$g_\gamma(\lambda) \;:=\; 2P_\lambda(k) - P_\lambda(k-1) - P_\lambda(k+1) \;=\; \frac{\lambda^k e^{-\lambda}}{k!}\left[2 - \frac{k}{\lambda} - \frac{\lambda}{k+1}\right]$$

be the negative curvature of P_λ. As before, we need to maximize $g_\gamma(\lambda)/\lambda$ for $k \geq 2$. Another tedious calculation shows

$$\frac{d}{d\lambda}\frac{g_\gamma(\lambda)}{\lambda} \;=\; ... \;=\; \frac{\lambda^{k-3}e^{-\lambda}}{k!} \cdot \left(1 - \frac{k}{\lambda}\right)\left[\left(\frac{\lambda}{k+1} - 1\right)^2 - 3\right]$$

This has 3 zeros with $\lambda^* = k$ corresponding to the unique maximum of $g_\gamma(\lambda)/\lambda$ (cf. Figure 1). The other two are minima. With $n \cdot \bar{C} := C_k := 2M_k - M_{k-1} - M_{k+1}$ being the

empirical negative curvature, we can upper bound its expectation as

$$n\bar{\gamma} \;:=\; \mathbb{E}[C_k] \;=\; \sum_x g_\gamma(\lambda_x) \;\le\; n\cdot\max_{\lambda>0}\frac{g_\gamma(\lambda)}{\lambda} \;=\; n\frac{g_\gamma(k)}{k}$$
$$=\; \frac{n}{k(k+1)}\frac{k^k e^{-k}}{k!} = \frac{n(1-\dot{\varepsilon}_k)}{k(k+1)\sqrt{2\pi k}} \;=:\; n\bar{\gamma}_k^{ub} \tag{11}$$

By Proposition 6 we have

$$\mathbb{V}[C_k] \;\le\; \mathbb{E}[4M_k+M_{k-1}+M_{k+1}] \;=\; 4\mu_k+\mu_{k-1}+\mu_{k+1}$$
$$\le\; n\left[4\frac{1-\dot{\varepsilon}_k}{\sqrt{2\pi k}}+\frac{1-\dot{\varepsilon}_{k-1}}{\sqrt{2\pi(k-1)}}+\frac{1-\dot{\varepsilon}_{k+1}}{\sqrt{2\pi(k+1)}}\right] \;=:\; 6n\frac{1+\tilde{\varepsilon}_k}{\sqrt{2\pi k}}$$

where $\tilde{\varepsilon}_k = O(1/\sqrt{k})$. Together by Lemma 17 with $Z_x = 2[\![N_x=k]\!]-[\![N_x=k-1]\!]-[\![N_x=k+1]\!] \in [-2;2]$,

$$p \;\lesssim\; \Phi_n(\sqrt{n}(\bar{\gamma}_k^{ub}-\bar{C}_k)/\sqrt{\bar{V}_k^{ub}}) \;\le\; \exp(-\tfrac{1}{2}n(\bar{C}_k-\bar{\gamma}_k^{ub})^2/\bar{V}_k^{ub}+O(1)) \;=\; e^{-O(n/k^{7/2})}$$

Similarly to the slope case, we can choose $\bar{V}_k^{ub}$ as $4\bar{M}_k+\bar{M}_{k-1}+\bar{M}_{k+1}$ or $4\bar{\mu}_k^{ub}+\bar{\mu}_{k-1}^{ub}+\bar{\mu}_{k+1}^{ub}$. In all of our experiments, the empirical choice performed better, but our running example below shows that the theoretical upper bounds can be better in certain circumstances.

Example 10. *Continuing our previous example where each data item is doubled ($C_2 = 2M_2 = n$), for $k=2$ we have $\bar{\gamma}_2^{ub} = 1/3e^2 = 0.0451$ and $4\bar{\mu}_2^{ub}+\bar{\mu}_1^{ub}+\bar{\mu}_3^{ub} \doteq 1.8260 < 2 = 4\bar{M}_2+\bar{M}_1+\bar{M}_3$, so in this case, using the former as V_k^{ub} is slightly better, and $p \lesssim \exp(-\frac{1}{2}n(1-0.0451)^2/1.8260) \doteq e^{-0.2497n}$. For $k \neq 2$, the C_k tests have no power.* ♢

Logarithmic curvature tests $\bar{U}_k := 2\ln M_k - \ln M_{k-1} - \ln M_{k+1}$. One weakness of the tests so far is that they rely on absolute moment bounds. If all x are equally likely, this is ok, but if λ_x are diverse, the Poisson mixture becomes wider and hence lower and hence $\mathbb{E}[M_k]$ and its slope and curvature become smaller. Since the upper bounds must include the worst-case when all λ_x are the same, the bounds become quite loose. We can fix this by normalizing the curvature by $\mathbb{E}[M_k]$. The mathematics becomes somewhat tedious, but there is a more elegant alternative with a very similar effect. We consider the negative curvature of $\ln M_k$:

$$\bar{U}_k \;:=\; 2\ln\bar{M}_k - \ln\bar{M}_{k-1} - \ln\bar{M}_{k+1} \;=\; 2\ln M_k - \ln M_{k-1} - \ln M_{k+1}$$

This is scale invariant, i.e. if all $M_\cdot$ in the vicinity of k are scaled down by some factor α, $\bar{U}_k$ stays unaffected, i.e. does not become smaller=weaker. Of course this is only useful if we can derive a good upper bound on its expectation. Since $\bar{U}_k$ is non-linear we need some new approach:

Consider the function $g:\mathbb{R}^3 \to \mathbb{R}$ with $g(\bar{\boldsymbol{Z}}) := \bar{U}_k$, where $\bar{\boldsymbol{Z}} := (\bar{M}_k, \bar{M}_{k-1}, \bar{M}_{k+1})^\top$. Noting that $\bar{M}_k$ concentrates around $\bar{\mu}_k$ for large n, we perform a second-order Taylor-series expansion of g around $\bar{\boldsymbol{\zeta}} := (\bar{\mu}_k, \bar{\mu}_{k-1}, \bar{\mu}_{k+1})$. The CLT then implies $g(\bar{\boldsymbol{Z}}) \approx \text{Gauss}(g(\bar{\boldsymbol{\zeta}}), \mathbb{V}[\bar{\boldsymbol{Z}}^\top \nabla g(\bar{\boldsymbol{\zeta}})])$. Formally we use the multivariate delta method, Lemma 16.

$$\bar{\upsilon}_k \;:=\; g(\bar{\boldsymbol{\zeta}}) \;=\; 2\ln\bar{\mu}_k - \ln\bar{\mu}_{k-1} - \ln\bar{\mu}_{k+1} \;=\; 2\ln\mu_k - \ln\mu_{k-1} - \ln\mu_{k+1}$$

Let us define some auxiliary probability distribution over x solely for technical purposes without ascribing any meaning to it.

$$\begin{aligned}
\tilde{P}_{\boldsymbol{\lambda},k}[X{=}x] \;&:=\; \frac{\lambda_x^k e^{-\lambda_x}/k!}{\sum_x \lambda_x^k e^{-\lambda_x}/k!}, \quad \text{then} \\
\frac{\mu_{k+1}}{\mu_k} \;&=\; \frac{\sum_x \frac{\lambda_x}{k+1}\lambda_x^k e^{-\lambda_x}/k!}{\sum_x \lambda_x^k e^{-\lambda_x}/k!} \;=\; \frac{\tilde{\mathbb{E}}_{\boldsymbol{\lambda},k}[\lambda_X]}{k+1} \\
\frac{\mu_{k-1}}{\mu_k} \;&=\; \frac{\sum_x \frac{k}{\lambda_x}\lambda_x^k e^{-\lambda_x}/k!}{\sum_x \lambda_x^k e^{-\lambda_x}/k!} \;=\; k\cdot\tilde{\mathbb{E}}_{\boldsymbol{\lambda},k}\Big[\frac{1}{\lambda_X}\Big] \;\geq\; \frac{k}{\tilde{\mathbb{E}}_{\boldsymbol{\lambda},k}[\lambda_X]}
\end{aligned}$$

where we applied Jensen's inequality in the last step to convex function $1/\lambda$. Taking the product, the dependence on unknown $\boldsymbol{\lambda}$ cancels out:

$$\bar{v}_k \;=\; \ln\frac{\mu_k}{\mu_{k-1}}\frac{\mu_k}{\mu_{k+1}} \;\leq\; \ln\frac{k+1}{k} =: \bar{v}_k^{ub} \;\leq\; \frac{1}{k} \tag{12}$$

Jensen's inequality for $1/\lambda$ is sharp *iff* $\tilde{P}_{\boldsymbol{\lambda},k}$ is a Dirac measure. That is, as before, the bound is attained when all $\lambda^* = \lambda_x$ are the same, but unlike before, the maximum is flat and attained for *any* λ^*, even far away from k. Indeed, the log-curvature $\bar{v}_k$ is the same, namely $\bar{v}_k^{ub} = \ln\frac{k+1}{k}$, for all λ^*. As for the variance,

$$\mathbb{V}[\bar{\boldsymbol{Z}}^\top\nabla g(\bar{\boldsymbol{\zeta}})] \;=\; \mathbb{V}\left[\begin{pmatrix}\bar{M}_k\\ \bar{M}_{k-1}\\ \bar{M}_{k+1}\end{pmatrix}^\top\begin{pmatrix}2/\bar{\mu}_k\\ -1/\bar{\mu}_{k-1}\\ -1/\bar{\mu}_{k+1}\end{pmatrix}\right] \;=\; \mathbb{V}\left[2\frac{M_k}{\mu_k} - \frac{M_{k-1}}{\mu_{k-1}} - \frac{M_{k+1}}{\mu_{k+1}}\right]$$

With $Z_k^x := [\![N_x = k]\!]$ and $\alpha_k := 2/\mu_k$ and $\alpha_{k\pm1} := -\mu_{k\pm1}$ and all other $\alpha_{k'} := 0$, we have $Z_k^+ = M_k$, hence

$$\mathbb{V}[\bar{\boldsymbol{Z}}^\top\nabla g(\bar{\boldsymbol{\zeta}})] \;=\; \mathbb{V}[\sum_x \alpha_k Z_k^+] \;\leq\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^+] \;=\; \frac{1}{\mu_{k-1}} + \frac{4}{\mu_k} + \frac{1}{\mu_{k+1}}$$

where we used Lemma 19a in the inequality and $\mathbb{E}[Z_k^+] = \mathbb{E}[M_k] = \mu_k$ in the last equality.

We need an approximation or upper bound on this, but we only have lower bounds on $1/\mu_k$. We solve this problem by replacing μ_k with their empirical estimates $M_k = \mu_k \pm O_P(\sqrt{n})$. This is the second time we are forced to use the empirical estimate to upper bound the variance, but for a slightly different reason than for E and O. Using Lemma 16 with $\boldsymbol{Z}_x := (Z_k^x, Z_{k-1}^x, Z_{k+1}^x) \in \{0,1\}^3$, the p-value for the logarithmic curvature test is

$$p \;\lesssim\; \Phi_n\left(\sqrt{n}\frac{\bar{v}_k^{ub} - \bar{U}_k}{\sqrt{\bar{M}_{k-1}^{-1} + 4\bar{M}_k^{-1} + \bar{M}_{k+1}^{-1}}}\right) \;=\; e^{-O(n/k^{7/2})}$$

Summary. In the table below we summarize the most important quantities for the tests $T : \mathcal{X}^n \to \mathbb{R}$ derived in this section. For $\tau := \mathbb{E}[T]$ we derived tight upper bounds even for small k. We only show the $n \gg k \gg 1$ approximations in the table and refer to the exact expressions. For the variance V_k^{ub} we only show the better upper bound (empirical except for M_k).

Test Name	$T:=n\bar{T}:=$	$\bar{\tau}:=\mathbb{E}[\bar{T}]\leq$	$\mathbb{V}[\bar{T}]\lesssim\bar{V}^{ub}=$	λ^*	$O(\ln\frac{1}{p})$
Even$\neq 0$	$E:=\sum_x N_x[\![N_x\neq 0 \text{ even}]\!]$	$\bar{\varepsilon}^{ub}=1/2$	$\frac{1}{n}\sum_{k\neq 0 \text{ even}}k^2M_k$	∞	n
Odd$\neq 1$	$O:=\sum_x N_x[\![N_x\neq 1 \text{ odd}]\!]$	$\bar{o}^{ub}=1/2$	$\frac{1}{n}\sum_{k\neq 1 \text{ odd}}k^2M_k$	∞	n
2nd-Count	$M_k:=\sum_x[\![N_x=k]\!]$	$\bar{\mu}_k^{ub}\overset{(8)}{=}\frac{1-\dot{\varepsilon}_{k-1}}{k\sqrt{2\pi(k-1)}}$	$\bar{\mu}_k^{ub}$	$k-1$	$\frac{n}{k^{3/2}}$
Slope	$D_k:=M_k-M_{k-1}$	$\bar{\delta}_k^{ub}\overset{(10)}{=}\frac{1-\ddot{\varepsilon}_k}{k^2\sqrt{2\pi e}}$	$\bar{M}_k+\bar{M}_{k-1}$	$\sqrt{k+1/4}^{\,k-1/2+}$	$\frac{n}{k^{5/2}}$
Lin.Curv.	$C_k:=2M_k-M_{k-1}-M_{k+1}$	$\bar{\gamma}_k^{ub}\overset{(11)}{=}\frac{1+\ddot{\varepsilon}_k}{k(k+1)\sqrt{2\pi k}}$	$4\bar{M}_k+\bar{M}_{k-1}+\bar{M}_{k+1}$	k	$\frac{n}{k^{7/2}}$
Log.Curv.	$\bar{U}_k:=\ln(M_k^2/M_{k-1}M_{k+1})$	$\bar{\upsilon}_k^{ub}\overset{(12)}{=}\ln\frac{k+1}{k}\leq\frac{1}{k}$	$\bar{M}_{k-1}^{-1}+4\bar{M}_k^{-1}+\bar{M}_{k+1}^{-1}$	any	$\frac{n}{k^{7/2}}$

Theorem 11 (IID tests). *Consider the test statistics $\bar{T}$ and associated upper bounds on their mean $\bar{\tau}$ and variance $\mathbb{V}[\bar{T}]$ from the above table. Then test $\bar{T}(x_{1:n})\geq\sqrt{n}[\bar{\tau}^{ub}+z_\alpha\sqrt{\bar{V}^{ub}}]$ rejects that $x_{1:n}$ is iid with confidence $\gtrsim 1-c_n\alpha$, i.e. at significance level $\lesssim c_n\alpha$, where $z_\alpha:=\Phi^{-1}(1-\alpha)$. The accuracy of $\lesssim$ is $O_P(1/n\sqrt{\mathbb{V}[\bar{T}]})$, except for E and O for which it is $O((\sum_x\theta_x^3)/(\sum_x\theta_x^2)^{3/2})$. See Lemma 17 (with $Z_+=T$ and $Z_x=T_x$) for p-values and further details.*

In our experiments, $\theta_x=1/d$ (uniform) and $\theta_x=2x/d/(d+1)$ (linear), so $(\sum_x\theta_x^3)/(\sum_x\theta_x^2)^{3/2}\leq 1.3/\sqrt{d}$ with $d=30...100$, so the Gaussian approximation is not great but ok for $\alpha=0.05$. For the other tests, the Gaussian approximation is quite good.

Proof. Follows directly from the derivations in this section and Lemma 17 and the fact that all T have the required decomposition $T=\sum_x Z_x$ ($\boldsymbol{Z}_x$ for $\bar{U}_k$ within $O_P(1/n)$). Z_x is bounded, so Lemma 17(iii) applies, except for E and O:

Let $E^x:=\sum_{0\neq k\neq n \text{ even}}k\cdot[\![N_x=k]\!]$, then $E=\sum_x E^x$. Similar to the derivation of $\mathbb{E}[E]$ one can show for $\alpha>0$ that

$$\begin{aligned}
\mathbb{E}[|E^x-\mathbb{E}[E^x]|^\alpha] &\eqsim (\tfrac{1}{2}\lambda_x)^\alpha \\
\text{This implies}\quad \sigma_+^2 &\equiv \textstyle\sum_x\mathbb{V}[E^x] \eqsim (\tfrac{1}{2}\lambda_x)^2 \\
\text{and}\qquad\quad \rho_+ &\equiv \textstyle\sum_x\mathbb{E}[|E^x-\mathbb{E}[E^x]|^3] \eqsim (\tfrac{1}{2}\lambda_x)^3 \\
\text{hence}\quad \rho_+/\sigma_+^{3/2} &\eqsim (\textstyle\sum_x\theta_x^3)/(\sum_x\theta_x^2)^{3/2}.
\end{aligned}$$

The condition in Lemma 17(iii) applies if $\rho_+/\sigma_+^{3/2}=O(1/\sqrt{n})$. For the CLT to hold asymptotically $\rho_+/\sigma_+^{3/2}\to 0$ suffices; see Theorem 15. The expressions for O are the same. ■

6 Toy/Control Experiments

We verify the tests developed in Section 5 on artificially generated data. We generate iid data for the extremes of all θ_x being the same, and $\boldsymbol{\theta}$ being maximally diverse. This is used for testing the validity of our tests (correct low Type I error). We then "corrupt" the samples in various ways to create non-iid data to determine the power of the tests in rejecting H_{iid} (low Type II error). For instance, 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. This is not meant to be a comprehensive evaluation of the tests, but a sanity check that the tests work as intended.

Remark: As discussed in Section 4 we can restrict our attention to $\mathcal{X}=\mathbb{N}$. On the other hand, it makes no difference whether we sample from $\mathcal{X}$ or $\mathcal{X}':=\{x:\theta_x>0\}$, so experimentally we can as well assume that $d=d'=|\mathcal{X}'|$, but given the caveat described in Section 4, we still should imagine $\mathcal{X}=\mathbb{N}$ and the parameter d in the experiments below now decoupled from $\mathcal{X}$. Alternatively, mentally replace every d in this section by d'.

Data generation. *Iid sampling* (iid) For the iid distributions $P_{\boldsymbol{\theta}}$ we tested two "extreme" choices. (uniform) One in which all $\boldsymbol{\theta}$ are the same for d categories, and 0 for all others, i.e. w.l.g. $P_{\boldsymbol{\theta}}[x]=\theta_x=1/d\ \forall x\in\{1:d\}$, and $P_{\boldsymbol{\theta}}[\mathcal{X}\setminus\{1:d\}]=0$. This should be the hardest case to not accidentally reject H_{iid}, since $\mathbb{E}[M_k]$ (7) is maximally peaked out (Figure 2 top left). (linear) The other extreme is for which the probabilities θ_x of categories x are equally/uniformly "distributed", i.e. $\theta_x\propto x$ for $x\in\{1:d\}$, i.e. $\theta_x=2x/d(d+1)$. $\mathbb{E}[M_k]$ is maximally washed out in this case (Figure 3 top left).

Exact data duplication (even-n) We created non-iid distributions Q out of the iid ones as follows: To mimic the data duplication problem, we sampled $x_{1:n/2}$ iid and then duplicated each item to $x_{1:n/2}x_{1:n/2}$, and then shuffled (though the tests only depend on the counts n_x, so shuffling is not necessary). This makes all n_x even, hence all $m_k=0$ for odd k.

Approximate data duplication (even-m) We also tested duplication and then injectively corrupt the data to $x'_{1:n/2}$, so that all $x'_{1:n/2}$ differ from all $x_{1:n/2}$. This mimics approximate duplicates. The effect is that for each count n_x there is a deterministically corrupted x' with same count $n_{x'}$, which in effect means that all m_k are even. This is a hard signal to detect without explicitly searching for it, and indeed our tests don't. See Section 2 for more discussion.

No empty categories (no-empty) We also sampled, $\boldsymbol{x}$ iid and then increased the count n_x for each $x\in\{1:d\}$ by 1. This eliminates all empty categories for $x\leq d$, but note that $\mathcal{X}$ itself is intended to be infinite, so does not really remove *all* empty categories. Technically we sample $x_{1:n-d}$ iid from $\{1:d\}$ and then add $x_{n-d+x}=x$ for $x\in\{1:d\}$ and then shuffle. Such $x_{1:n}$ is *not* iid w.r.t. any $P_{\boldsymbol{\theta}}$, but the signal is in general very weak, and none of our tests were able to pick it up.

No unique&empty categories (no-unique) Finally we increased each count by two so that every data item appears at least twice, but unlike even-n can also appear an odd number of times. Technically we sample $x_{1:n-2d}$ iid from $\{1:d\}$ and then add $x_{n-2d+x}=x_{n-d+x}=x$ for $x\in\{1:d\}$ and then shuffle. As long as the original iid sample has enough items of low multiplicities, there is a clear signal that the data is non-iid as explained in Section 2.

Choice of "hyper"-parameters. We also have to choose d, n, and k. Since this is not a systematic empirical study, not even on toy data, but only to illustrate the tests and corroborate the theoretical arguments, we chose some arbitrary and some interesting values without any claim of coverage. k is typically chosen where the tests are strongest, essentially where μ_k is large.

Test setup and graphs. We tested our tests ($T\in\{E,O,M_k,D_k,C_k,\bar{U}_k\}$) for over/under-confidence on the artificial iid data and their power on the artificial non-iid data above. We sampled data $x_{1:n}$ a 10'000 times from P or Q, and computed 10'000

p-values for each test. Let $\tilde{T} := \Phi((\tau^{ub}-T)/\sqrt{V^{ub}})$, where τ^{ub} and V^{ub} are the upper bounds for mean and variance of T we derived for our tests (see test table before Theorem 11). We report $\hat{P}[\tilde{T} \leq \alpha]$, the fraction of p-values below α, as a function of α (Figures 2 left). For an ideal uniformized test (see Section 7), $P_{\boldsymbol{\theta}}[\tilde{T} \leq \alpha] = \alpha$, but this rarely possibly to achieve simultaneously for *all* $\boldsymbol{\theta}$. If data is sampled iid from $P_{\boldsymbol{\theta}}$, a valid test should be below this diagonal line, or at least not much above, or at the very least (approximately) below α for the α we care about, typically $\alpha = 0.1\%...5\%$. For non-iid data we want the power $\beta(\alpha) := Q[\tilde{T} \leq \alpha]$ to be as large as possible, ideally close to 1 for the α we care about, since this is the probability our test correctly rejects H_{iid} if data is sampled from Q.

We also plot the empirical second-order counts M_k as a function of k for one sample $x_{1:n}$, together with their true *uncorrupted* expectation $\mathbb{E}_{\boldsymbol{\theta}}[M_k]$, which in some informal sense is the iid distribution closest to the corrupted non-iid distribution. We also plot the empirical average $\langle M_k \rangle$ over the 10'000 runs, which in case of iid data is very close to the true mean $\mathbb{E}_{\boldsymbol{\theta}}[M_k]$ (Figures 2&3 left).

Explanation of the figures. The legends in Figures 2&3 also display the hyper-parameters d, n (and k for the right graphs), together with the sampling procedure (uniform|linear) and corruption model (iid|even-n|no-unique). Test M uses the theoretical upper bound μ^{ub} for its variance, all other tests use the empirical upper bounds for their variance. u-test is a uniformly at random sampled "test" $u \sim$Uniform[0;1] for control purposes, and should be very close to the diagonal (exact); deviations are due to the finite sample approximation. The % behind the test is the fraction of times $\hat{\beta} = \hat{Q}[p \leq 0.05]$ (or $P_{\boldsymbol{\theta}}$ if iid), the test rejected H_{iid} at 5% significance level. The true reject probability is $\hat{\beta} \pm O(\hat{\beta}(1-\hat{\beta})/10'000)^{1/2}$. In each graph, we also plotted the 10'000 p-values for the (at $\alpha = 5\%$) most powerful test (with uniformly at random y-component). For a perfect test on iid data, these points would be uniformly distributed in $[0;1] \times [0;1]$.

Experimental results per data type. *Tests are not over-confident* (uniform-iid): Figure 2 top row is an example that shows that no test is over-confident (right). For uniform, all θ are them same, hence $\mathbb{E}[\boldsymbol{M}]$ is proportional to a Binomial(θ^*) with maximum around $n\theta^* = n/d \approx 3 = k$ (left), for which most tests are most sensitive. We tested the tests on a variety of further combinations, but without any surprises. All tests are valid in the sense that they reject H_{iid} on iid data at level α with probability less then α, or at least not much more, so we refrained from showing the iid control plots for most of the non-iid experiments. The occasional slight over-confidence could either be due to finite sample size n and our asymptotic approximations (multinomial to Gaussian approximation) of our tests, or variance from finite (10'000) experiment repetitions.

Tests are under-confident (linear-iid): The second row shows that tests can be quite under-confident (right). For linear, the θ values are spread out, hence $\mathbb{E}[\boldsymbol{M}]$ is a broad mixture over Binomials of different θ (left). The broader the distribution of $\mathbb{E}[\boldsymbol{M}]$, the less confident the tests are.

Tests can be powerful (uniform-even-n): The third row shows that many tests effectively detect if every data item is duplicated, as is also obvious from the spikes in $\boldsymbol{M}$ (left). Unsurprisingly the most powerful one is the even-test E which is tailored for this kind of data, closely followed by the logarithmic curvature test U_2 with 99.6% rejection rate.

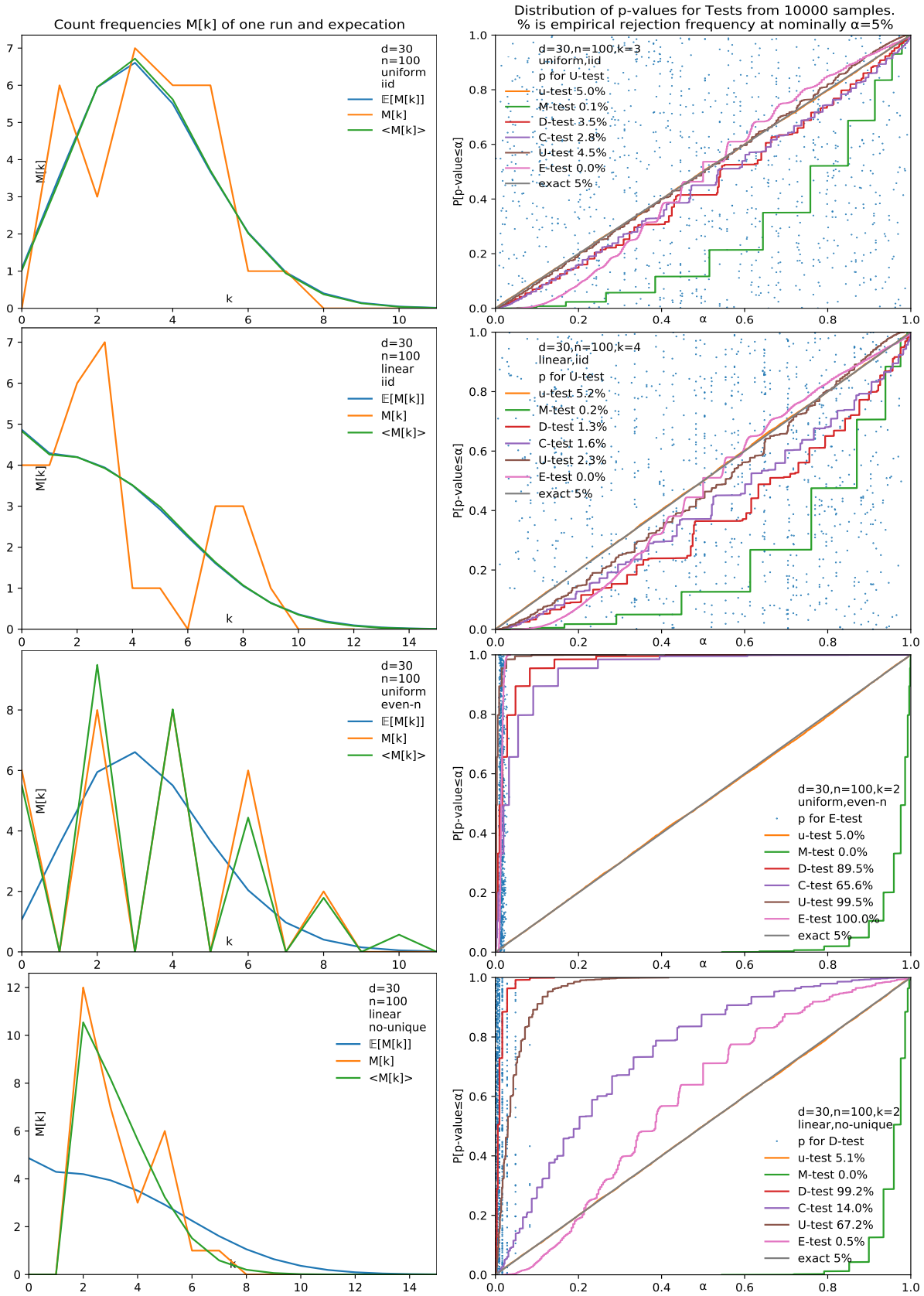

Figure 2: (**Testing the tests**) 10'000 data sets $x_{1:n}$ are sampled iid from $P_{\boldsymbol{\theta}}$ or non-iid Q. *(left)* One sample, average, and expected second-order counts M_k as a function of k. *(right)* p-value distribution of various tests. For iid $P_{\boldsymbol{\theta}}$, a curve above/below the diagonal means over/under-confidence, i.e. we want on or below. For non-iid we want far above. % is the empirical rejection frequency at nominally $\alpha = 5\%$.

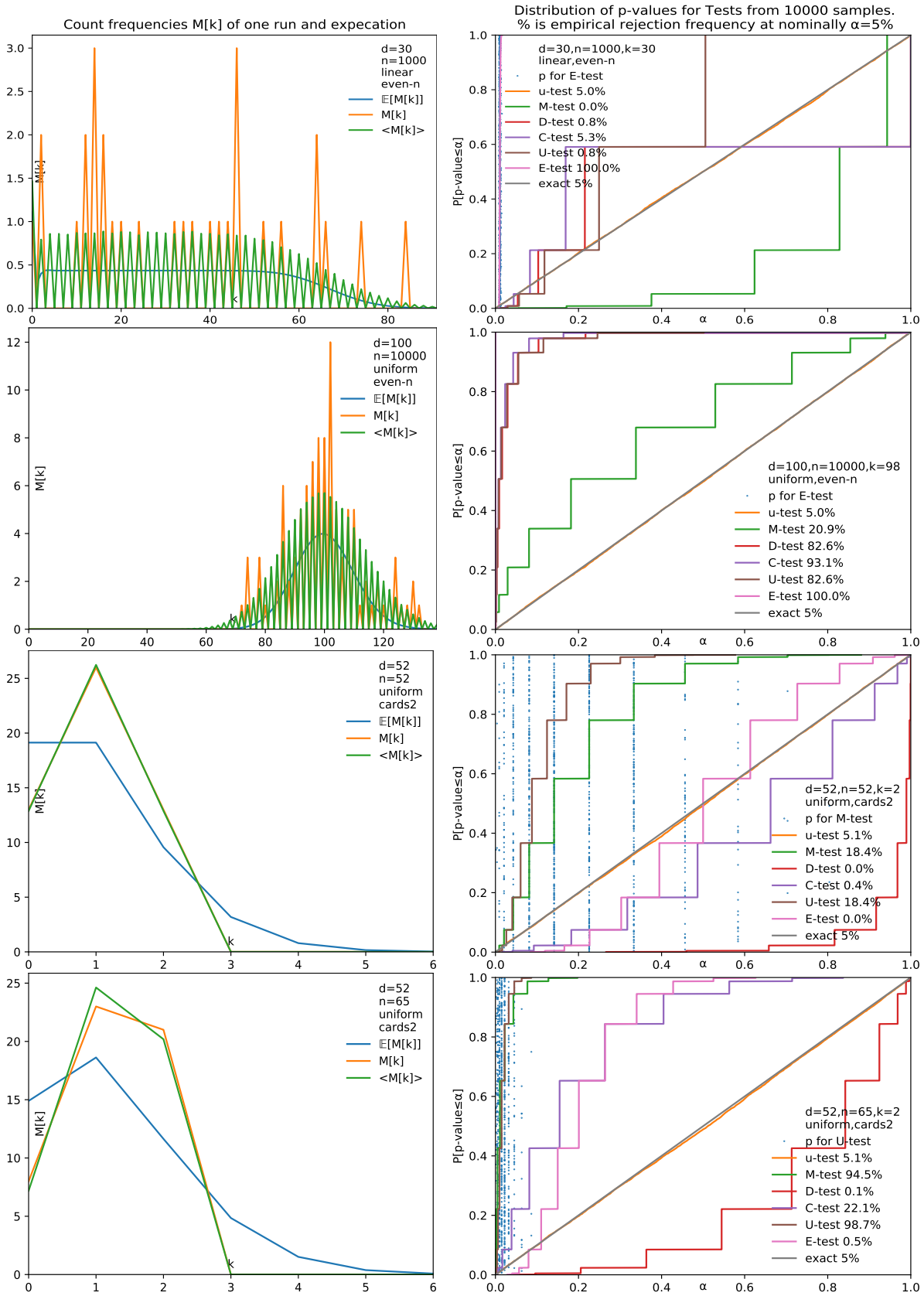

Figure 3: (**Testing the tests**) 10'000 data sets $x_{1:n}$ are sampled iid from $P_{\boldsymbol{\theta}}$ or non-iid Q. *(left)* One sample, average, and expected second-order counts M_k as a function of k. *(right)* p-value distribution of various tests. For iid $P_{\boldsymbol{\theta}}$, a curve above/below the diagonal means over/under-confidence, i.e. we want on or below. For non-iid we want far above. % is the empirical rejection frequency at nominally $\alpha = 5\%$.

Having no singletons can sometimes be significant (linear-no-unique): Adding two copies to each observed x roughly shifts $\boldsymbol{M}$ two to the right ($M_0 = M_1 = 0$), so there is little non-iid signal in M_k for higher k (bottom left). If most counts are very low, large M_2 with zero M_1 is suspicious. Indeed the curvature test M_2 rejects H_{iid} 99.2% of the time (bottom right).

Tests are often weak for larger k (uniform-even-n): Figure 3 top row shows a case with large (first-order) multiplicities (large N_x), i.e. most data items appear between 10 and 70 times. linear washes out any peak in $\boldsymbol{M}$ which is indeed uniform in even k from 0 to 60. This makes all M_k very small, around 1. While some tests get a very weak signal, it is not strong enough to reject H_{iid}. The only effective test is even, since combining all even M_k amplifies the signal. The second row is similar: here most N_x are around 100 with M_k around 4, and besides E (only) the linear curvature test is able to reject H_{iid} with 92.9% confidence.

Cards from 2 decks (uniform-cards2): The bottom two rows of Figure 3 show the result of drawing n cards from a pair of shuffled 52-card decks (104 cards in total). Drawing half of the cards ($n = 52$) is not enough to reveal that $x_{1:n}$ is not iid (third row), but drawing $n \geq 65$ suffices (last row). 240 cards from 6 decks with 312 cards in total are required to reject H_{iid}. M_5 rejects with 96% confidence (not shown).

Experimental results per test. Comparing the various tests across the shown data examples and beyond, they largely perform as expected. They are valid in the sense that the Type I error is $\lesssim \alpha$, as we would expect, since they have been derived to satisfy this. We also see that the Type I error is often significantly smaller than α. Every test has its own strengths and weaknesses, and its power depends on the type of non-iid data they face. All tests are sensitive to some non-iid signal, and oblivious to (many) others. Among the developed tests there was no uniformly best test, but they could be combined to an approximately most powerful test as described in Section 7. For all k-dependent tests (i.e. all but E and O), the more uniform $\boldsymbol{\theta}$, i.e. the more concentrated the μ_k are around some k, and hence the larger μ_k, the more powerful the test in general.

Experimental results not shown. We confirmed that the alternative choice for the variance bounds, namely empirical for M, and theoretical for all others when available, are worse than our (opposite) choice. The Odd test was never effective on any of the created data, but unsurprisingly is very effective if data is dominated by odd multiplicities. As discussed above, our tests were not strong enough to detect the non-iid nature of 'approximately duplicated data' (even-m) and 'no empty categories' (no-empty).

7 More/Alternative Tests

Here we present some alternative ways of deriving tests. Either we have not worked out the details or specifics, or we could not make them work, or they were inferior to the ones derived in Section 5. They are nevertheless interesting and may be made to work with more effort. Also disseminating failed attempts can be useful in itself. We discuss the standard ways of combining test to broaden their power by summation or maximization with Bonferroni correction. Pushed to the extreme we get a universal Martin-Löf randomness test, which in principle can detect all effective non-iid signals including the "exotic" ones from the coin flip example in Section 2. We then derive a

general template for invariant tests from a generalized likelihood ratio test. This can be used to develop combinatorial tests and compression tests. Finally we reduce the problem of testing independence of exchangeable data to the empirical moment problem, which unfortunately is hard.

Summing tests. Assume we have tests $T_1,T_2,\ldots$. We could combine them by summing them up, $T_+=\sum_k T_k$. We could either sum their upper bounds τ_k^{ub} to get an upper bound on $\mathbb{E}[T_+]$, or derive an improved joint upper bound τ^{ub} by maximizing $\mathbb{E}[T_+]$ directly. For instance, $E=\sum_{k\ldots} k\cdot M_k$ could be interpreted as a sum of tests $T_k=k\cdot M_k$, but summing up the individual bounds $\mathbb{E}[E]\leq\sum_{k\ldots} k\cdot\mu_k^{ub}\leq n\cdot\sum_k k/k^{3/2}=\infty$ is vacuous, while the joint bound $\mathbb{E}[T_+]\leq n/2$ was non-vacuous. Also, summing reduces the relative variance and may lead to stronger tests. On the other hand, if some T_k are highly negative, they could ruin the sum and make the combined test weaker. Using a weighted (e.g. by $1/\sqrt{V_k^{ub}}$) sum may be better than a plain sum, but makes it harder to get improved upper bounds on its expectation.

Bonferroni. Let c_k^α be such that $\sup_{\boldsymbol\theta} P_{\boldsymbol\theta}[T_k>c_k^\alpha]\leq\alpha$, e.g. $c_k^\alpha\approx\mu_k^{ub}+z_\alpha\sqrt{\mu_k^{ub}}$ for $T_k=M_k$, and similarly for other tests, or hybrid combinations. For finite $K\subset\mathbb{N}$, combined test $T_K:=\max\{T_k-c_k^{\alpha/|K|}:k\in K\}$ with $c_K=0$ has significance α ($\alpha/|K|$ is the Bonferroni correction), i.e. $P_{\boldsymbol\theta}[T_K>0]\leq\alpha$. Combining tests in this way, if some T_k are highly negative, they do not impact the joint test T_K. They are just ineffective and make T_K a bit weaker, since they reduce $\alpha/|K|$.

Uniformizing tests. We can always uniformize tests T with $\sup_{\boldsymbol\theta} P_{\boldsymbol\theta}[T(\boldsymbol X)>c_\alpha]\leq\alpha$ to $\sup_{\boldsymbol\theta} P_{\boldsymbol\theta}[\tilde T\leq\delta]\leq\delta$ as follows: Since $c_\alpha:U\subseteq[0,1]\to\mathbb{R}$ is monotone decreasing, it has a left-continuos monotone decreasing inverse $\bar F:\mathbb{R}\to[0;1]$, $\bar F(c_\alpha)=\alpha$, hence $\sup_{\boldsymbol\theta} P_{\boldsymbol\theta}[\bar F(T(\boldsymbol X))\leq\bar F(c_\alpha)]\leq\alpha$, hence $\tilde T(\boldsymbol X):=\bar F(T(\boldsymbol X))$ does as claimed. For instance, if T is standard Normal, then $c_\alpha=z_\alpha=\Phi^{-1}(1-\alpha)$, hence $\bar F(z)=1-\Phi(z)$. Alternatively, we can define decreasing $\bar F(t):=\sup_{\boldsymbol\theta}\bar F_{\boldsymbol\theta}(t)$ via decreasing survival function $\bar F_{\boldsymbol\theta}(t):=P_{\boldsymbol\theta}[T(\boldsymbol X)>t]$, then again for all $\boldsymbol\theta$,

$$P_{\boldsymbol\theta}[\tilde T\leq\delta]\;=\;P_{\boldsymbol\theta}[\bar F(T)\leq\delta]\;\leq\;P_{\boldsymbol\theta}[\bar F_{\boldsymbol\theta}(T)\leq\delta]\;\leq\;P_{\boldsymbol\theta}[T\geq\bar F_{\boldsymbol\theta}^{-1}(\delta)]\;=\;\bar F_{\boldsymbol\theta}(\bar F_{\boldsymbol\theta}^{-1}(\delta))\;=\;\delta$$

Universal tests. In uniform form, the Bonferroni correction is simply $\tilde T(\boldsymbol X)=|K|\cdot\min\{\tilde T_k:k\in K\}$. If $\tilde T$, i.e. just one of the tests $\tilde T_k$ is small, we can reject H_{iid}, but we pay a price of $|K|$ in what small means. This generalizes to infinitely many tests, say $K=\mathbb{N}$, by defining

$$\begin{aligned}\tilde T\;&:=\;\min\{k(k+1)\tilde T_k:k\in\mathbb{N}\}:\qquad(13)\\ P_{\boldsymbol\theta}[\tilde T\leq\delta]\;&=\;P_{\boldsymbol\theta}[\exists k:k(k+1)\tilde T_k\leq\delta]\;\leq\;\sum_k P_{\boldsymbol\theta}[k(k+1)\tilde T_k\leq\delta]\;\leq\;\sum_k\frac{\delta}{k(k+1)}\;=\;\delta\end{aligned}$$

where we applied the union bound in the first inequality. If $\tilde T_1,\tilde T_2,\tilde T_3,\ldots$ is an effective enumeration of *all* upper semi-computable tests [LV08], then $U:=\tilde T$ is a so-called universal test. No other effective test can be significantly stronger than U. One can show that $U(\boldsymbol X)=\sup_{\boldsymbol\theta} P_{\boldsymbol\theta}(\boldsymbol X)/M(\boldsymbol X)$, where $M(\boldsymbol x)$ is Solomonoff's universal a-priori distribution [Sol64, HLV07, Hut07, Hut17] Another interpretation is that U is a Martin-Löf

randomness test [LV08] but w.r.t. the *class* of iid distributions. The iid randomness deficiency of $\boldsymbol{x}$ can be defined as $d(\boldsymbol{x}) := \inf_{\boldsymbol{\theta}}\{\log_2 M(\boldsymbol{x}) - \log_2 P_{\boldsymbol{\theta}}(\boldsymbol{x})\}$, which implies $P_{\boldsymbol{\theta}}[d(\boldsymbol{X}) \geq \log_2(1/\delta)] \leq \delta \ \forall \boldsymbol{\theta}$.

Likelihood Ratio (LR) tests. Any probability distribution Q on $\mathcal{X}^n$ can be converted into a uniformized iid test

$$\tilde{T}(\boldsymbol{X}) := \sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}(\boldsymbol{X})/Q(\boldsymbol{X}): \tag{14}$$

$$P_{\boldsymbol{\theta}}[\tilde{T} \leq \delta] \ \leq \ P_{\boldsymbol{\theta}}[\tfrac{P_{\boldsymbol{\theta}}(\boldsymbol{X})}{Q(\boldsymbol{X})} \leq \delta] \ = \ P_{\boldsymbol{\theta}}[\delta \tfrac{Q(\boldsymbol{X})}{P_{\boldsymbol{\theta}}(\boldsymbol{X})} \geq 1] \ \leq \ \mathbb{E}_{\boldsymbol{\theta}}[\delta \tfrac{Q(\boldsymbol{X})}{P_{\boldsymbol{\theta}}(\boldsymbol{X})}] \ = \ \delta \textstyle\sum_{\boldsymbol{x}} P_{\boldsymbol{\theta}}(\boldsymbol{x}) \tfrac{Q(\boldsymbol{x})}{P_{\boldsymbol{\theta}}(\boldsymbol{x})} \ = \ \delta$$

We can even choose Q to depend on $\boldsymbol{\theta}$ or allow semi-probabilities $\sum_{\boldsymbol{x}\in\mathcal{X}^n} Q(\boldsymbol{x}) \leq 1$. For $Q(x) := M(x)$ we recover the universal test U above.

Invariant LR tests for finite $\mathcal{X}$. We want invariant tests, so Q should only depend on $\boldsymbol{m}$. Let $Q(\boldsymbol{m}) := Q[\{\boldsymbol{x} : \boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{m}\}]$, where with slight overload of notation $\boldsymbol{m}(\boldsymbol{x})$ denotes the second-order counts of $\boldsymbol{x}$. If $\mathcal{X}$ is finite, we can symmetrize any Q via

$$\bar{Q}_d(\boldsymbol{x}) \ := \ \frac{Q(\boldsymbol{m})}{\#\{\boldsymbol{x} : \boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{m}\}}, \quad \text{where} \quad \#\{\boldsymbol{x} : \boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{m}\} \ = \ \tbinom{d}{m_+} \cdot \tbinom{m_+}{\boldsymbol{m}} \cdot \tbinom{n}{\boldsymbol{n}}$$

Note that $\binom{n}{\boldsymbol{n}} := n!/\prod_{x=1}^d n_x! = n!/\prod_{k=1}^n k!^{m_k}$ also depends on $\boldsymbol{m}$ only. The same is true for

$$\sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}(\boldsymbol{x}) \ = \ \prod_{x=1}^d \Big(\frac{n_x}{n}\Big)^{n_x} \ = \ \prod_{k=1}^n \Big(\frac{k}{n}\Big)^{k \cdot m_k}$$

$$\text{Thus} \quad \tilde{T}_d(\boldsymbol{x}) \ := \ \frac{\sup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}(\boldsymbol{x})}{\bar{Q}_d(\boldsymbol{x})} \ = \ \frac{n!}{Q(\boldsymbol{m})} \cdot \binom{d}{m_+} \cdot \binom{m_+}{\boldsymbol{m}} \cdot \prod_{k=1}^n \Big(\frac{k^k}{n^k k!}\Big)^{m_k} \tag{15}$$

depends on $\boldsymbol{m}$ only. By construction, $\tilde{T}_d$ is a valid uniformized test (14). We are primarily interested in $d = \infty$, but unfortunately the test becomes vacuous for $d \to \infty$. By the argument in Section 4, tests to leading order in n remain valid if we replace infinite $\mathcal{X}$ by finite $\mathcal{X}$ of size n^3, i.e. we can use $\tilde{T}_{n^3}$ but this is very crude. We can avoid the dependence on d by using $\boldsymbol{\theta}$-dependent Q:

Invariant LR tests for infinite $\boldsymbol{\mathcal{X}}$. Let $\mathcal{X}'' = \{x : n_x > 0\}$ and allow Q to depend on $\boldsymbol{\theta}$ and decompose it as

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\boldsymbol{x}) \ &= \ Q(\boldsymbol{x}|\boldsymbol{n}, \mathcal{X}'', \boldsymbol{m}) \cdot Q(\boldsymbol{n}|\mathcal{X}'', \boldsymbol{m}) \cdot Q(\mathcal{X}''|\boldsymbol{m}) \cdot Q(\boldsymbol{m}) \\ &= \ Q(\boldsymbol{x}|\boldsymbol{n}) \qquad\quad \cdot Q(\boldsymbol{n}|\boldsymbol{m}) \qquad \cdot Q(\mathcal{X}''|m_+) \cdot Q(\boldsymbol{m}) \\ &= \ \binom{n}{\boldsymbol{n}}^{-1} \ \cdot \ \binom{m_+}{\boldsymbol{m}}^{-1} \cdot \Big(m_+! \prod_{x \in \mathcal{X}''} \theta_x\Big) \cdot Q(\boldsymbol{m}) \end{aligned}$$

Note that $\boldsymbol{n}$ implies $\mathcal{X}''$ and $\boldsymbol{m}$, the other choices in the second line are motivated by invariance. In the last line we chose uniform probabilities for the first two factors. For the third factor we chose the true sampling probabilities θ_x of the (unique) symbols in $\mathcal{X}''$. The $m_+!$ is because order plays no role in set $\mathcal{X}''$. Note that

$$\sum_{\mathcal{X}'' \subset \mathcal{X} : |\mathcal{X}''| = m_+} \Big(m_+! \prod_{x \in \mathcal{X}''} \theta_x\Big) \ \leq \ \sum_{x_{1:m_+} \in \mathcal{X}^{m_+}} \theta_{x_1} \cdot \ldots \cdot \theta_{x_{m_+}} \ = \ \Big(\sum_{x \in \mathcal{X}} \theta_x\Big)^{m_+} \ = \ 1$$

hence $Q(\mathcal{X}''|m_+)$ is indeed a valid semi-probability. Combining this with $P_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{x\in\mathcal{X}''}\theta_x^{n_x}$ we get

$$\tilde{T}(\boldsymbol{x}) \;:=\; \sup_{\boldsymbol{\theta}}\frac{P_{\boldsymbol{\theta}}(\boldsymbol{x})}{Q_{\boldsymbol{\theta}}(\boldsymbol{x})} \;=\; \sup_{\boldsymbol{\theta}}\frac{\prod_{x\in\mathcal{X}''}\theta_x^{n_x-1}}{m_+!\cdot Q(\boldsymbol{m})}\binom{n}{\boldsymbol{n}}\binom{m_+}{\boldsymbol{m}}$$
$$= \;\frac{n!\binom{m_+}{\boldsymbol{m}}}{m_+!Q(\boldsymbol{m})}\cdot\prod_{k=1|2}^{n}\left[\frac{1}{k!}\left(\frac{k-1}{n-m_+}\right)^{k-1}\right]^{m_k}$$

where we used that the maximum is attained at $\theta_x = (n_x-1)/(n-m_+)$ and a similar rearranging of terms as in the previous paragraph. This expression is similar to (15) but independent from d as desired.

Combinatorial tests. We have reduced the choice of T to a choice of Q, but what have we gained? Note that we need to ensure $\sum_{\boldsymbol{m}}Q(\boldsymbol{m})\leq 1$, where the sum is over all valid second-order counts $\mathcal{M}:=\{\boldsymbol{m}\geq\boldsymbol{0}:\sum_{k=1}^{n}k\cdot m_k=n\}$. Any Q satisfying this constraint results in a valid invariant uniformized test T. $\text{Part}(n):=|\mathcal{M}|$ is the number of partitions of n into a sum of natural numbers without regard to order [AS84]. We could choose $Q(\boldsymbol{m})=\text{Part}(n)^{-1}$ uniformly. This choice is closely related to the Good-Turing estimator [Hut18]. Another improved choice would be Ristad's estimator $Q(\boldsymbol{m})=\frac{1}{m_+}\binom{m_+}{\boldsymbol{m}}/\binom{n}{m_+}$ [Ris95]. The former led to essentially vacuous tests, the latter to very weak tests.

Compression tests. Similar to Solomonoff's M, $Q(\boldsymbol{m}):=2^{-K(\boldsymbol{m}|n)}$, where $K(\boldsymbol{m}|n)$ is the prefix Kolmogorov complexity of $\boldsymbol{m}$ given n, is a universal distribution leading to a universal and in this case invariant test $U(\boldsymbol{m})=2^{K(\boldsymbol{m})}\sup_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}(\boldsymbol{x})$. In theory this is (at least asymptotically) the strongest test possible. In practice, since K is not computable, it needs to be approximated by feasible codes. The task here is to find short prefix-free codes for $\boldsymbol{m}$ of length $\text{CL}(\boldsymbol{m}|n)$, and use test $T_{\text{CL}}(\boldsymbol{m}) := 2^{\text{CL}(\boldsymbol{m})}\sup_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}(\boldsymbol{x})$. We tried a couple of codes such as naive $\text{CL}(m_k)\approx\log_2(m_k+1)$, or coding the differences $\text{CL}(m_k-m_{k-1})\approx\log_2(2|m_k-m_{k-1}|+1)$, and variations thereof. The resulting tests were weak to vacuous.

Moment method. Let $\vartheta := \theta/(1-\theta)\in\mathbb{R}_0^+$ and $\rho[\vartheta\leq b] := \sum_x[\![\vartheta_x\leq b]\!]$ be a non-negative measure on $\mathbb{R}_0^+$. Note that $\rho[\mathbb{R}_0^+]=\sum_{x=1}^{d}[\![\vartheta_x\leq\infty]\!]=d\neq 1$, i.e. ρ is *not* a probability measure. Now let P_ν be a measure on $\mathbb{R}_0^+$ that has density $dP_\nu(\vartheta)/d\rho(\vartheta) := n\vartheta/[(1+\vartheta)^n\mu_1]$ w.r.t. ρ. Then

$$\mu_k \;=\; \mathbb{E}[M_k] \;=\; \sum_{x=1}^{d}\binom{n}{k}\theta_x^k(1-\theta_x)^{n-k} \;=\; \sum_{x=1}^{d}\binom{n}{k}\frac{\vartheta_x^k}{(1+\vartheta_x)^n}$$
$$=\; \binom{n}{k}\int_0^\infty\frac{\vartheta^k}{(1+\vartheta)^n}d\rho(\vartheta) \;=\; \frac{\mu_1}{n}\binom{n}{k}\int_0^\infty\vartheta^{k-1}dP_\nu(\vartheta) \;=\; \frac{\mu_1}{n}\binom{n}{k}\nu_{k-1}$$

where $\nu_{k-1}:=\mathbb{E}_\nu[\vartheta^{k-1}]$ is the $k-1$st moment of P_ν. Note that $\mathbb{E}_\nu[1]=\nu_0=n\mu_1/n\mu_1=1$, hence P_ν is indeed a probability measure. Now given μ_k for $k\geq 1$ is equivalent to being given ν_{k-1} for $k\geq 1$ and μ_1. We can therefore ask whether some given sequence of real numbers $(\nu_{k-1}:=n\binom{n}{k}\mu_k/\mu_1 : k\geq 1)$ are moments of some probability measure, say, P_ν. If so, using also μ_1 we can (re)construct ρ from it, which in turn defines a mixture of Binomial distributions with ϑ-distribution ρ, which in turn defines an iid distribution

with θ-distribution $\rho[\theta/(1-\theta) \leq b]$. This iid distribution has expected second-order counts $\mathbb{E}[M_k]$ we started with. ($\mu_0 = \infty$ is always true for $d = |\mathcal{X}| = \infty$, while $\mu_0 < \infty$ requires $d < \infty$ and poses strong extra conditions on ρ hence P_ν hence (ν_k).) If (ν_k) are *not* moments of a probability distribution, then (μ_k) are *not* expectations from a mixture of Binomial distributions.

In principle we can exploit the above correspondence to develop iid tests. Unfortunately, the moment problem, inferring a probability measure from moments, is hard. Furthermore, we do not actually have the moments μ_k but only empirical estimates M_k thereof. That is, we would need to determine whether $\hat{\nu}_{k-1} := nM_k/\binom{n}{k}M_1$ are approximately moments. More precisely, we need tests which tell us whether there exist $\nu_k \in [\hat{\nu}_k \pm O(\sqrt{\mathbb{V}[\hat{\nu}_k]})]$ for all $k \geq 1$ that are moments of some probability distribution. If not, we can reject H_{iid}.

8 Conclusion

Summary. We developed various tests for the (in)dependence of exchangeable data without exploiting or being given any structure in the observation space $\mathcal{X}$. We reduced the problem to $\mathcal{X} = \mathbb{N}$ which greatly simplified the analysis. A necessary condition for any invariant test to have power is to observe duplicate items in $x_{1:n}$. We derived a number of tests based on the observation that the second-order counts m_k are "smooth" in k if data are iid, and demonstrated their (lack of) power empirically. We also presented some alternative ideas for developing tests.

Outlook. While we have experimentally verified that our tests have power for some non-iid distributions, it could be interesting to identify sub-classes of exchangeable distributions and compute the theoretical power of our tests. Some of the alternative approaches to developing tests from Section 7 could be worked out, esp. universal compression-based tests and moment-based tests. It would be interesting to apply our tests to some real data, but our invariant/agnostic tests only have power if $x_{1:n}$ contains exact duplicates, and even then the non-iid signal may be too weak to detect. In ML practice, we likely need to exploit some structure in the data. The simplest solution would be to aggregate *similar* x into the same category ($\mathcal{X}_{\text{orig}} \to \mathcal{X}_{\text{agg}}$). The tests become more powerful to the extent that this increases the number of duplicates. To guide the aggregation we need some metric or at least topology on $\mathcal{X}$, and to be effective not just any but "good" ones. Alternatively one could develop tests directly for $\mathcal{X}_{\text{orig}}$ exploiting its structure. For instance, for $\mathcal{X} = \mathbb{R}$, one could check whether x_t and $x_{t'}$ are correlated, e.g. for zero-mean x_t whether $\frac{1}{n^2}\sum_{t \neq t'} x_t x_{t'} \not\approx 0$, or higher-order (central) moments $\frac{1}{n^3}\sum_{t,t',t''} x_t x_{t'} x_{t''} \not\approx 0$, etc. In any case, as discussed in Section 2, this is an AI-complete problem already for unstructured $\mathcal{X}$ and even more so for structured $\mathcal{X}$, without a general clean solution except impractical universal tests.

We have only considered invariant tests. We argued that this is a natural choice, but more convincing arguments would be good. How limiting is it to only consider invariant tests? Are there non-invariant tests that have more power on some invariant sub-class of $\mathcal{Q}$?

Acknowledgements. I thank *Tor Lattimore* and *Bryn Elezedy* for great feedback on earlier drafts.

References

[AS84] Milton Abramowitz and Irene A. Stegun, editors. *Pocketbook of Mathematical Functions*. H. Deutsch, Thun [Switzerland], 1984.

[Bee72] Paul Beek. An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 23(3):187–196, 1972.

[BLH22] Jörg Bornschein, Yazhe Li, and Marcus Hutter. Sequential Learning Of Neural Networks for Prequential MDL. (arXiv:2210.07931), October 2022.

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020.

[Ess56] C. G. Esseen. A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956(2):160–170, July 1956.

[GB19] Kyle Gorman and Steven Bedrick. We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy, 2019. Association for Computational Linguistics.

[HLV07] Marcus Hutter, Shane Legg, and Paul M. B. Vitányi. Algorithmic probability. *Scholarpedia*, 2(8):2572, 2007.

[Hut06] M. Hutter. Human knowledge compression prize. open ended, http://prize.hutter1.net/, 2006.

[Hut07] Marcus Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.

[Hut17] Marcus Hutter. Universal learning theory. In C. Sammut and G. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*, pages 1295–1304. Springer, second edition, 2017.

[Hut18] Marcus Hutter. Tractability of batch to sequential conversion. *Theoretical Computer Science*, 733:71–82, 2018.

[LV08] Ming Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, New York, 3rd ed edition, 2008.

[Rai19] Martin Raic. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25(4A), November 2019.

[Ris95] E. S. Ristad. A natural law of succession. Technical Report CS-TR-495-95, Princeton University, 1995.

[SEBF21] Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We Need To Talk About Random Splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, 2021. Association for Computational Linguistics.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Was10] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, New York Berlin Heidelberg, corr. 2. print., [repr.] edition, 2010.

A Technical Lemmas

To derive our tests we require a couple of technical lemmas we derive in this section. Lemma 12 is heavily used to derive upper bounds on the expected value of our tests. Lemma 17 is a rather standard way of creating a test of significance $\alpha \in [0;1]$ from upper bounds on the mean and variance of an asymptotically Gaussian test statistic. Lemmas 18&19 are bounds on the variance of linear combinations of negatively correlated random variables. Lemma 16 is a version of the delta-method in statistics. In this section, P, $\mathbb{E}$, $\mathbb{V}$, $\mathbb{W}$ are generic, and properties of random variables explicitly stated.

A general upper bound for expectations. The following Lemma is the work horse for deriving upper bounds on the expected value of our tests.

Lemma 12 (Upper bound for expectations). *Consider $g : \mathbb{R}^+ \to \mathbb{R}$ and extend its definition to $g : \mathbb{R}_0^+ \to \mathbb{R} \cup \{\pm\infty\}$ via $g(0) = \limsup_{\lambda \to 0} g(\lambda)$ and $\lambda_x \geq 0$ and $\sum_x \lambda_x = n$. Then $\sum_{x=1}^d g(\lambda_x) \leq n \cdot \sup_{\lambda > 0} g(\lambda)/\lambda$. Assuming the maximum is attained and unique, set $\lambda^* := \operatorname{argmax}_{\lambda > 0} g(\lambda)/\lambda$. The l.h.s. is approximately maximized by setting $\lambda_x \approx \lambda^*$ for $d' \approx n/\lambda^*$ of the x and the remaining $\lambda_x = 0$. The maximum is exactly attained if $d \geq d' \in \mathbb{N}$. For Lipschitz g with $g(0) = 0$ and $d' \leq d$, the gap in the bound is $O(1)$ (relative gap is $O(1/n)$). Otherwise the gap can be unbounded.*

That is, the "only" effect of the constraint $\sum_x \lambda_x = n$ is that we maximize $g(\lambda)/\lambda$ rather than $g(\lambda)$ and multiply the result with n rather than d. For our core case $d = \infty$, $d' \leq d$ is satisfied. For $d < d' < \infty$, the bound can be very loose. We will comment on that when it is due.

Proof. Consider $L(\boldsymbol{\lambda}) = \sum_{x=1}^d g(\lambda_x)$ with $\lambda_x \geq 0$ and constraint $\sum_x \lambda_x = n$ and $g : \mathbb{R}_0^+ \to \mathbb{R} \cup \{\pm\infty\}$ and $g(0) = g(0^+) = 0$ (we lift the last assumption at the end of the proof). We can bound $L(\boldsymbol{\lambda})$ as follows: First, all x for which $\lambda_x = 0$ neither contribute to L, nor to the constraint, so we can simply ignore such x, and by symmetry replace d by $d' := \#\{x : \lambda_x > 0\} \leq d$, and henceforth assume $\lambda_x > 0$.

$$L(\boldsymbol{\lambda}) \;=\; \sum_{x=1}^{d'} g(\lambda_x) \;=\; \sum_{x=1}^{d'} \frac{g(\lambda_x)}{\lambda_x} \lambda_x \;\leq\; \sum_{x=1}^{d'} \sup_{\lambda > 0} \frac{g(\lambda)}{\lambda} \lambda_x \;=\; n \cdot \sup_{\lambda} \frac{g(\lambda)}{\lambda}.$$

If the global maximum is attained, choose any maximizer $\lambda^* = \operatorname{argmax}_\lambda g(\lambda)/\lambda$. If $d' \leq d$ and n/λ^* is an integer, the upper bound is exact, i.e. $\max_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}) = n \cdot g(\lambda^*)$ for $d' = n/\lambda^*$. If n/λ^* is not an integer, we can set $d' = \lfloor n/\lambda^* \rfloor$ or $d' = \lceil n/\lambda^* \rceil$ assuming it does not exceed d. Since we still need to respect $\sum_{x=1}^{d'} \lambda_x = n$, we either have to adjust all λ^* by at most $\pm 1/n$, or adjust one λ_x by at most ± 1. In either case, for Lipschitz g, the relative slack will be $O(1/n)$ (which we generally ignore).

We now lift the assumption that $g(0)$ is defined and 0. Let $c := \limsup_{\lambda \to 0} g(\lambda) \in \mathbb{R} \cup \{\pm\infty\}$. If $c = 0$, we can simply extend g to $g(0) = 0$ and the above proof applies. If $c > 0$, then $\sup_\lambda f(\lambda)/\lambda = \infty$, and the bound vacuously holds. If $c < 0$, $\limsup_{\lambda \to 0} \sum_{x=d'+1}^{d} g(\lambda) = (d - d')c < 0$, and hence can be replaced by 0 for an upper bound. ■

On sums and averages and functions of random variables. Let Y and Z be generic real-valued random variables. For a random variable, the corresponding lower greek denotes its expectation, e.g. $\zeta = \mathbb{E}[Z]$. Variance is denoted by $\sigma^2 = \mathbb{V}[Z]$. and third absolute central moment by $\rho = \mathbb{W}[Z] := \mathbb{E}[|Z-\zeta|^3]$.

Let Z_x with $x \in \mathcal{X}$ be a collection of independent but typically not identically distributed random variables. We define $Z_+ := \sum_{x\in\mathcal{X}} Z_x$, sometimes over other, even uncountable, domains as long as only a countable number of Z_x are non-zero. Furthermore $\bar{Z} := Z_+/n$, i.e. *the bar always denotes division by* n, and is *not* the average of all Z_x over x. Mostly $|\mathcal{X}| = \infty$ anyway, but Z_x and Z_+ and $\bar{Z}$ implicitly depend on the sample size n. We use the $_x$, $+$, and $^-$ convention also for other variables like expectations. We sometimes drop the $+$ if it does not cause any confusion. The reason for this convention is that typically Z_+ and others scale linearly with $n \to \infty$, and $\bar{Z}$ and others are $\Theta(1)$ or even converge to a finite non-zero value for $n \to \infty$.

Definition 13 (Sums and "averages" of independent random variables)**.** *Let Z_x with $x \in \mathcal{X}$ be a collection of independent random variables with sum $Z_+ := \sum_x Z_x$ and $\bar{Z} := Z_+/n$. Provided all involved sums and integrals are absolutely convergent,*

$$\begin{array}{llll} \zeta_x := \mathbb{E}[Z_x], & \sigma_x^2 := \mathbb{V}[Z_x], & \rho_x := \mathbb{W}[Z_x] & = \Theta(1) \\ \zeta_+ := \sum_x \zeta_x = \mathbb{E}[Z_+], & \sigma_+^2 := \sum_x \sigma_x^2 = \mathbb{V}[Z_+], & \rho_+ := \sum_x \rho_x & = \Theta(n) \\ \bar{\zeta} := \zeta_+/n = \mathbb{E}[\bar{Z}], & \bar{\sigma}^2 = \sigma_+^2/n := n \cdot \mathbb{V}[\bar{Z}], & \bar{\rho} = \rho_+/n & = \Theta(1) \end{array}$$

The last $\Theta()$-column is indicative only. Note that $\bar{\sigma}^2$ is *not* the variance of $\bar{Z}$, and $\sigma_+ := \sqrt{\sigma_+^2}$ (*not* $\sum_x \sigma_x$).

We make repeated use of the Esseen version of the Berry-Esseen theorem [Ess56] with improved constant [Bee72], which is a strengthening of the Central Limit Theorem.

Definition 14 (Standard Normal distribution)**.** *Let $\Phi(y) = \int_{-\infty}^{y} e^{-x^2/2} dx/\sqrt{2\pi}$ be the Cumulative Distribution Function (CDF) of the standard Normal. Define $z_\alpha := \Phi^{-1}(1-\alpha)$ for significance α (typically $\alpha = 0.05$ and $z_{0.05} \doteq 1.64$). The p-value $p := \Phi(-y) \leq e^{-y^2/2}/y\sqrt{2\pi}$, which is sharp (in ratio) for $y \to \infty$.*

Theorem 15 (**Berry-Esseen: $Z_+ \approx \text{Gauss}(\zeta_+, \sigma_+^2)$ and $\bar{Z} \approx \text{Gauss}(\bar{\zeta}, \bar{\sigma}^2/n)$**)**. *With the notation above, let F be the distribution function of $Y := (Z_+ - \zeta_+)/\sigma_+ = \sqrt{n}(\bar{Z}-\bar{\zeta})/\bar{\sigma}$, then $\sup_y |F(y) - \Phi(y)| \leq \rho_+/\sigma_+^3 = \bar{\rho}/\sqrt{n}\bar{\sigma}^3$, i.e. if $\bar{\rho}/\sqrt{n}\bar{\sigma}^3 \to 0$, then Y converges in distribution to a standard Normal.*

The following lemma is a version of the multivariate delta-method in statistics [Was10]. The statements follows directly from a second-order Taylor series expansion of $g(\bar{Z})$ around $\bar{\zeta}$ and a multivariate version of the theorem above [Rai19]. The result implies that Gaussian confidence intervals or p-values for $g(\bar{Z})$ are asymptotically correct.

Lemma 16 (**Multivariate delta method: $g(\bar{\boldsymbol{Z}}) \approx \text{Gauss}(g(\bar{\boldsymbol{\zeta}}), \mathbb{V}[\bar{\boldsymbol{Z}}^\top \nabla g(\bar{\boldsymbol{\zeta}})])$**)**. *With the notation above, except that $\bar{\boldsymbol{Z}} = \frac{1}{n}\sum_x \boldsymbol{Z}_x \in \mathbb{R}^d$ can be vector-valued, with $\bar{\boldsymbol{\zeta}} = \mathbb{E}[\bar{\boldsymbol{Z}}]$ and $g : \mathbb{R}^d \to \mathbb{R}$ a twice differentiable function with $\nabla g(\bar{\boldsymbol{\zeta}}) \neq 0$ and $\sigma_g^2 := \nabla g(\bar{\boldsymbol{\zeta}})^\top \mathbb{V}[\bar{\boldsymbol{Z}}] \nabla g(\bar{\boldsymbol{\zeta}}) = \mathbb{V}[\bar{\boldsymbol{Z}}^\top \nabla g(\bar{\boldsymbol{\zeta}})]$ and F the distribution function of $Y := \sqrt{n}(g(\bar{\boldsymbol{Z}}) - g(\bar{\boldsymbol{\zeta}}))/\sigma_g$. Then $\sup_y |F(y) - \Phi(y)| = O(\bar{\rho}/\sqrt{n}\bar{\sigma}^3)$, where $\bar{\sigma}^2 := n \cdot \mathbb{E}[||\bar{\boldsymbol{Z}} - \bar{\boldsymbol{\zeta}}||_2^2]$ and $\bar{\rho} = \frac{1}{n}\sum_x \mathbb{E}[||\boldsymbol{Z}_x - \boldsymbol{\zeta}_x||_2^3]$.*

Upper-bound test. We will develop various independence tests for $\boldsymbol{X}=(X_1,...,X_n)$ based on applications of the basic test below. It is a rather standard way of creating a test of significance $\alpha\in[0;1]$ from upper bounds on the mean and variance of an approximately Gaussian test statistic, except possibly that it is in terms of infinite sums of random variables with finite total mean and variance.

Lemma 17 (Basic upper-bound test for independent random variables). *Consider the hypothesis H_{ind} that Z_x for $x\in\mathcal{X}$ are independent random variables with known upper bounds $\mathbb{E}[Z_+]\leq\zeta_+^{ub}$ and $\sigma_+^2:=\mathbb{V}[Z_+]\leq\mathbb{E}[V^{ub}]$, where $Z_+:=\sum_x Z_x$, and $\mathbb{V}[V^{ub}]\leq O(\rho_+^2/\sigma_+^2)$ for some $V^{ub}\equiv n\bar{V}^{ub}$, and $\rho_+:=\mathbb{W}[Z_+]$.*

- *(i) At significance level α, if $Z_+>\zeta_+^{ub}+z_\alpha\sqrt{V^{ub}}$ we can reject H_{ind} with confidence $\approx 1-\alpha$ (typically $\alpha=0.05$ and $z_{0.05}\doteq 1.64$), where $z_\alpha:=\Phi^{-1}(1-\alpha)$.*
- *(ii) The p-value is upper bounded by $p\lesssim\tilde{T}:=\Phi((\zeta_+^{ub}-Z_+)/\sqrt{V^{ub}})$, provided $Z_+>\zeta_+^{ub}$. That is, we can reject H_{ind} if $p\leq\alpha$.*
- *(iii) The above test and p-value are approximate with accuracy $O_P(\rho_+/\sigma_+^3)$ ($O_P((b-a)/\sigma_+)$ if $Z_x\in[a;b]$), conservative for $\sigma_+^3/\rho_+\to\infty$, and asymptotically exact only if also $\mathbb{E}[Z_+]\to\zeta_+^{ub}$ and $\mathbb{E}[V^{ub}]\to\sigma_+^2$.*
- *(iv) If $n\bar{Z}:=Z_+>\zeta_+^{ub}=:n\bar{\zeta}^{ub}$, then $p\lesssim\Phi(\sqrt{n}(\bar{\zeta}^{ub}-\bar{Z})/\sqrt{\bar{V}^{ub}})\leq\exp(-\frac{1}{2}n(\bar{Z}-\bar{\zeta}^{ub})^2/\bar{V}^{ub})$ for sufficiently large n (and $\bar{Z}$ and $\bar{\zeta}^{ub}$ and $\bar{V}^{ub}$ fixed or $\Theta(1)$).*
- *(v) If $Z_x\leq b$ and $\mathbb{V}[V^{ub}]=0$, then $p\leq\exp\{-\frac{1}{2}n(\bar{Z}-\bar{\zeta}^{ub})^2/[\bar{V}^{ub}+b(\bar{Z}-\bar{\zeta}^{ub})/3n]\}$ provided $\bar{Z}-\bar{\zeta}^{ub}$.*

The bound in terms of Φ is sharper than the exponential bounds in (iv) and (v). Φ is used in our experiments; (iv) is more convenient for analytical/asymptotic analysis in our toy examples; (v) theoretically pleasing since it is non-asymptotic, i.e. exact if using theoretical (non-stochastic, non-empirical) upper bounds on $\mathbb{V}[Z_+]$.

Proof. Let $\zeta_x:=\mathbb{E}[Z_x]$ and $\sigma_x^2:=\mathbb{V}[Z_x]$. Let $Z_+:=\sum_x Z_x$ be their sum with $\mathbb{E}[Z_+]=\zeta_+$ and $\mathbb{V}[Z_+]=\sigma_+^2$. Let $Y:=(Z_+-\zeta_+)/\sigma_+$ with CDF F, where $\sigma_+:=\sqrt{\sigma_+^2}$.

(i) $P[Z_+>\zeta_+^{ub}+z_\alpha\sqrt{V^{ub}}]\lesssim P[Z_+>\zeta_++z_\alpha\sigma_+]=P[Y>z_\alpha]=1-F(z_\alpha)\approx 1-\Phi(z_\alpha)=\alpha$. The first $\lesssim$ follows from $\zeta_+^{ub}\geq\zeta_+$ and is an exact inequality if $V^{ub}\geq\mathbb{V}[Z_+]$. For random V^{ub}, we have

$$V^{ub} \;=\; \mathbb{E}[V^{ub}]-O_P(\sqrt{\mathbb{V}[V^{ub}]}) \;\geq\; \sigma_+^2-O_P(\rho_+/\sigma_+) \;=\; \sigma_+^2(1-O_P(\rho_+/\sigma_+^3))$$

i.e. $\lesssim$ holds to relative accuracy $O(\rho_+/\sigma_+^3)$. The approximate equality $\approx$ holds, since Y is approximately Normal: $\sup_y|F(y)-\Phi(y)|=O(\rho_+/\sigma_+^3)$ by the (Berry-)Esseen theorem.

(ii&iv) For $Z_+>\zeta_+$ we have

$$\begin{aligned} p \;&=\; 1-F(Y) \;\approx\; \Phi(-Y) \;=\; \Phi\left(\frac{\zeta_+-Z_+}{\sigma_+}\right) \;\lesssim\; \Phi\left(\frac{\zeta_+^{ub}-Z_+}{\sqrt{V^{ub}}}\right) \\ &=\; \Phi\left(\sqrt{n}\frac{\bar{\zeta}^{ub}-\bar{Z}}{\sqrt{\bar{V}^{ub}}}\right) \;\leq\; \exp\left(-n\frac{(\bar{Z}-\bar{\zeta}^{ub})^2}{2\bar{V}^{ub}}\right) \end{aligned}$$

(iii) The general accuracy has already been shown in (i). For bounded Z_x we have

$$\rho_+ \;=\; \textstyle\sum_x\mathbb{E}[|Z_x-\zeta_x|^3] \;\leq\; (b-a)\sum_x\mathbb{E}[(Z_x-\zeta_x)^2] \;=\; (b-a)\sum_x\sigma_x^2 \;=\; (b-a)\sigma_+^2$$

If the upper bounds are exact ($\zeta_+=\zeta_+^{ub}$ and $\sigma_+^2=\mathbb{E}[V^{ub}]$) and the relative variance of V^{ub} tends to zero ($V^{ub}/\sigma_+^2\to 1$), all approximations in (i) and (ii) become (asymptotically)

exact.
(v) Since $Z_x \le \zeta_x + b$, Bernstein's (one-sided) inequality applied to $Z_x - \zeta_x$ gives

$$P[Z_+ - \zeta_+ \ge t] \;=\; P\Big[\sum_x (Z_x - \zeta_x) \ge t\Big] \;\le\; \exp\!\left(-\frac{t^2/2}{\sum_x \sigma_x^2 + bt/3}\right)$$

For bounding the p-value, we replace t by the observed $Z_+ - \zeta_+$:

$$p \;\le\; \exp\!\left(-\frac{(Z_+ - \zeta_+)^2/2}{\sigma_+^2 + b(Z_+ - \zeta_+)/3}\right) \;\le\; \exp\!\left(-\frac{(Z_+ - \zeta_+^{ub})^2/2}{V^{ub} + b(Z_+ - \zeta_+^{ub})/3}\right)$$

where the last inequality is true since the expression can be shown to be monotone increasing in ζ_+ and σ_+^2 provided $Z_+ \ge \zeta_+^{ub}$. ■

Negatively correlated random variables. We need bounds on the variance of linear combinations of negatively correlated random variables. For positive combinations, the non-diagonal covariance terms can simply be dropped, but we need bounds for mixed signed combinations. Luckily the correlations are sufficiently weak to get weaker but still useable upper bounds.

Lemma 18 (**Expectation and variance of correlated random variables.**). *Let Z_k be correlated random variables. Provided all involved sums and integrals are absolutely convergent,*

$$\begin{array}{lll}
\mathbb{E}[(\sum_k \alpha_k Z_k)^2] \;=\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^2] & \text{if} & \mathbb{E}[Z_k Z_{k'}] = 0 \;\forall k \ne k'. \\
\mathbb{E}[(\sum_k \alpha_k Z_k)^3] \;=\; \sum_k \alpha_k^3 \mathbb{E}[Z_k^3] & \text{if} & \mathbb{E}[Z_k Z_{k'} Z_{k''}] = 0 \;\forall k \ne k' \ne k'' \ne k. \\
\mathbb{V}[\sum_k \alpha_k Z_k] \;\le\; \sum_k \alpha_k^2 \mathbb{V}[Z_k] & \text{if} & Cov[Z_k, Z_{k'}] \le 0 \text{ and } \alpha_k \ge 0 \;\forall k \ne k'.
\end{array}$$

Proof.
$$\begin{aligned}
\mathbb{E}[(\textstyle\sum_k \alpha_k Z_k)^2] \;&=\; \textstyle\sum_{k,k'} \alpha_k \alpha_{k'} \mathbb{E}[Z_k Z_{k'}] \;=\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^2] \\
\mathbb{E}[(\textstyle\sum_k \alpha_k Z_k)^3] \;&=\; \textstyle\sum_{k,k',k''} \alpha_k \alpha_{k'} \alpha_{k''} \mathbb{E}[Z_k Z_{k'} Z_{k''}] \;=\; \sum_k \alpha_k^3 \mathbb{E}[Z_k^3] \\
\mathbb{V}[\textstyle\sum_k \alpha_k Z_k] \;&=\; \textstyle\sum_{k,k'} \alpha_k \alpha_{k'} \mathrm{Cov}[Z_k, Z_{k'}] \;\le \sum_k \alpha_k^2 \mathbb{V}[Z_k]
\end{aligned}$$
■

Lemma 19 (**Double collection of (un)correlated random variables.**). *Let $\alpha_k \in \mathbb{R}$ and Z_k^x be a double collection of random variables and $Z_k^+ := \sum_x Z_k^x$. Assume all involved sums and integrals below are absolutely convergent. Then*
(a) *If $Z_k^x \in [0;1]$ and Z_k^x are uncorrelated in x, i.e. $\mathrm{Cov}[Z_k^x, Z_{k'}^{x'}] = 0$ for all $x \ne x'$ and k,k', and additionally $\mathbb{E}[Z_k^x Z_{k'}^x] = 0 \;\forall k \ne k'$. Then $\mathbb{V}[\sum_k \alpha_k Z_k^+] \le \sum_k \alpha_k^2 \mathbb{E}[Z_k^+]$.*
(b) *If $Z_k^x \in [0;1]$ and $\mathbb{E}[Z_k^x Z_{k'}^x Z_{k''}^x] = 0 \;\forall k \ne k' \ne k'' \ne k$. Then $\rho_+ := \sum_x \mathbb{W}[\sum_k \alpha_k Z_k^x] \le 8 \sum_k |\alpha_k|^3 \mathbb{E}[Z_k^+]$.*
(c) *If $Z_k^x \in \mathbb{R}$ and $\alpha_k \ge 0 \;\forall k$ and $\mathrm{Cov}[Z_k^x, Z_{k'}^{x'}] \le 0$ if $k \ne k'$,*
then $\mathbb{V}[\sum_k \alpha_k Z_k^+] \le \sum_k \alpha_k^2 \mathbb{V}[Z_k^+]$

Proof.(a)
$$\mathbb{V}[\textstyle\sum_k \alpha_k Z_k^x] \;\le\; \mathbb{E}[(\sum_k \alpha_k Z_k^x)^2] \;=\; \sum_k \alpha_k^2 \mathbb{E}[(Z_k^x)^2] \;\le\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^x]$$

where the equality follows from Lemma 18. This and independence of Z_k^x and $Z_{k'}^{x'}$ for $x \ne x'$ implies

$$\begin{aligned}
\mathbb{V}[\textstyle\sum_k \alpha_k Z_k^+] \;&=\; \mathbb{V}[\sum_k \alpha_k \sum_x Z_k^x] \;=\; \textstyle\sum_x \mathbb{V}[\sum_k \alpha_k Z_k^x] \\
&\le\; \textstyle\sum_x \sum_k \alpha_k^2 \mathbb{E}[Z_k^x] \;=\; \sum_k \alpha_k^2 \mathbb{E}[Z_k^+]
\end{aligned}$$

(b) The proof follows a similar structure as (a) but with $\mathbb{V}$ replaced by $\mathbb{W}$ and α_k^2 by $|\alpha_k|^3$ and using $\mathbb{W}[Z] \leq 8\mathbb{E}[|Z^3|]$:

$$\begin{aligned}\tfrac{1}{8}\rho_x \;&:=\; \tfrac{1}{8}\mathbb{W}[\sum_k \alpha_k Z_k^x] \;\leq\; \mathbb{E}[|\sum_k \alpha_k Z_k^x|^3] \\ &\leq\; \mathbb{E}[(\sum_k |\alpha_k||Z_k^x|)^3] \;=\; \sum_k |\alpha_k|^3 \mathbb{E}[|Z_k^x|^3] \;\leq\; \sum_k |\alpha_k|^3 \mathbb{E}[Z_k^x]\end{aligned}$$

where the equality follows from Lemma 18. This implies

$$\tfrac{1}{8}\rho_+ \;\equiv\; \tfrac{1}{8}\textstyle\sum_x \rho_x \;\leq\; \sum_x \sum_k |\alpha_k|^3 \mathbb{E}[Z_k^x] \;=\; \sum_k |\alpha_k|^3 \mathbb{E}[Z_k^+]$$

(c)
$$\begin{aligned}\mathbb{V}[\textstyle\sum_k \alpha_k Z_k^+] \;&=\; \sum_{k,k'} \alpha_k \alpha_{k'} \sum_{x,x'} \mathrm{Cov}[Z_k^x, Z_{k'}^{x'}] \\ &\leq\; \textstyle\sum_k \alpha_k^2 \sum_{x,x'} \mathrm{Cov}[Z_k^x, Z_k^{x'}] \;=\; \sum_k \alpha_k^2 \mathrm{Cov}[Z_k^+, Z_k^+]\end{aligned}$$ ■

While the following lemma is not needed to show that any of our current tests are asymptotically Normal, we state it here for future use. It shows that certain unbounded $Z_x \widehat{=} T_x$ also satisfy the condition in Lemma 17(iii).

Lemma 20 (Upper bounds on moments). *For $Z_k^x := M_k^x := [\![N_x = k]\!]$, hence $Z_k^+ = M_k$, and $T_x := \sum_k \alpha_k M_k^x$ and $T_+ = \sum_k \alpha_k M_k$, $\tau_x := \mathbb{E}[T_x]$, $\sigma_x^2 := \mathbb{V}[T_x]$, $\rho_x := \mathbb{W}[T_x]$, following the notational convention of Definition 13, we have*
$\tau_+ := \mathbb{E}[T_+] \leq c \cdot n$ *and* $\mathbb{E}[\bar{T}] = \bar{\tau} \leq c$ *if* $\alpha_k \leq c \cdot k$
$\sigma_+^2 := \mathbb{V}[T_+] \leq c \cdot n$ *and* $\mathbb{V}[\bar{T}] = \bar{\sigma}^2/n \leq c/n$ *if* $\alpha_k^2 \leq c \cdot k$
$\rho_+ := \sum_x \mathbb{W}[T_x] \leq 8c \cdot n$ *and* $\sum_x \mathbb{W}[\bar{T}_x] = \bar{\rho}/n^2 \leq 8c/n^2$ *if* $|\alpha_k|^3 \leq c \cdot k$

Proof. $\mathbb{E}[T_+] = \mathbb{E}[\sum_k \alpha_k M_k] \leq c \cdot \mathbb{E}[\sum_k k M_k] = c \cdot \mathbb{E}[N] = c \cdot n$
Lemma 18a implies $\mathbb{V}[T_+] = \mathbb{V}[\sum_k \alpha_k Z_k^+] \leq \sum_k \alpha_k^2 \mathbb{E}[Z_k^+] \leq c \cdot \sum_k k \cdot \mathbb{E}[M_k] = c \cdot n$
Lemma 18b implies $\rho_+ = \sum_x \mathbb{W}[\sum_k \alpha_k Z_k^x] \leq 8 \sum_k |\alpha_k|^3 \mathbb{E}[Z_k^+] \leq 8c \cdot \sum_k k \cdot \mathbb{E}[M_k] = 8cn$ ■

Finally, we need the standard Stirling approximation:

Lemma 21 (Stirling approximation). $\ln n! = n \ln \frac{n}{e} + \ln \sqrt{2\pi n} + O(\frac{1}{n})$ *or more precisely*

$$\frac{1 - \dot{\varepsilon}_k}{\sqrt{2\pi k}} \;:=\; \frac{k^k e^{-k}}{k!} \quad \textit{with} \quad e^{-1/12k} \leq 1 - \dot{\varepsilon}_k \leq e^{-1/(12k+1)} \tag{16}$$

We make frequent use of this representation/approximation. Rapidly $0 \leq \dot{\varepsilon}_k \leq 1/12k \to 0$, but for most practical purposes simply setting $\dot{\varepsilon}_k = 0$ or its lower bound even for $k = 1$ should be fine. We used the exact expression in the experiments, but the asymptotic one provides more insight.

B I.I.D. Tests for Multinomial Distribution

Here we develop tests analogs to those in Section 5 but without the Poisson approximation, i.e. directly for iid $\boldsymbol{X}$ i.e. multinomial $\boldsymbol{N}$. The derivations for the upper bounds on the expectations of the test statistics T are structurally very similar. Since this section closely mirrors Section 5, we only point out the differences, and refer to Section 5 for

explanation of various steps and detailed explanations and discussion. Upper bounding the variances is significantly more complicated, and is deferred to Appendix C. P, $\mathbb{E}$, $\mathbb{V}$, $\mathbb{W}$ are w.r.t. the multinomial distribution $P_{\boldsymbol{\theta}}$ (1).

For large n, the basic functions $g(\theta) \approx f^n(\lambda)$ and Figure 1 looks virtually unchanged, just with λ replaced by $n\theta$. The p-value expressions are also unchanged, except now in terms of Φ instead of Φ_n, i.e. without fudge factor c_n, and the upper bounds for $\mathbb{E}[T]$ derived here are slightly different. Appendix C also shows that under certain conditions, $\mathbb{V}_{\boldsymbol{\theta}}[T] \approx \mathbb{V}_{\boldsymbol{\lambda}}[T]$. The running example of duplicate data items is also unchanged, even the specific constants remain the same for $n \to \infty$.

The general idea behind the tests. Like the Poisson, the binomial (2) is also "smooth" in k and θ, has a unique maximum at $k \approx n\theta$, is log-concave with small slope and curvature, so is also a rather benign function. Indeed it is (also) approximately Gaussian with mean $n\theta$ and variance $n\theta(1-\theta)$. Analogous to (7), consider

$$\mathbb{E}[M_k] \;=\; \sum_x P_{\boldsymbol{\theta}}[N_x = k] \;=\; \sum_x P_{\theta_x}(k) \;=\; \sum_x f_k^n(\theta_x) \;=\; \sum_x \binom{n}{k} \theta_x^k (1-\theta_x)^{n-k} \tag{17}$$

Proposition 6 stays nearly the same. Since we use it repeatedly we (re)state it here in terms of $P_{\boldsymbol{\lambda}}$. The proof is the same with the obvious substitutions of $P_{\boldsymbol{\theta}}$ instead of $P_{\boldsymbol{\lambda}}$ and $\sum_x \theta_x = 1$ instead of $\sum_x \lambda_x = n$, which explains the "missing" factor n.

Proposition 22 (Multinomial upper bounds for linear tests). *Let $T = \sum_k \alpha_k M_k$ for $\alpha_k \in \mathbb{R}$. Provided all involved sums and integrals are absolutely convergent, we have $\tau := \mathbb{E}[T] \le \sup_{\theta>0} f(\theta)/\theta =: \tau^{ub}$, where $f(\theta) := \sum_k \alpha_k P_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$, and $\mathbb{V}[T] \le \sum_k \alpha_k^2 \mathbb{E}[M_k] \le V^{ub}$, where $V^{ub} := \sum_k \alpha_k^2 \mu_k^{ub}$ with $\mu_k^{ub} \ge \mathbb{E}[M_k]$ upper bounding the expectations of M_k.*

Second-order count tests M_k.

$$f_k^n(\theta) \;\le\; f_k^n\big(\tfrac{k}{n}\big) \;\le\; \sqrt{\tfrac{n}{2\pi k(n-k)}} \;=\; \tfrac{1}{\sqrt{2\pi k}}\big[1+O(\tfrac{k}{n})\big], \quad \text{hence} \tag{18}$$

$$\mu_k \;:=\; \mathbb{E}[M_k] \;\le\; \tfrac{n}{k} f_{k-1}^{n-1}(\tfrac{k-1}{n-1}) =: \mu_k^{ub} \;\le\; \tfrac{n}{k}\sqrt{\tfrac{n-1}{2\pi(k-1)(n-k)}} \;=\; \tfrac{n}{k\sqrt{2\pi(k-1)}}[1+O(\tfrac{k}{n})] \tag{19}$$

These expressions very similar to (8) just for $f_{k-1}^{n-1}(\theta)$ and maximizer $\theta^* = \frac{k-1}{n-1}$ (cf. Figure 1). In Appendix C we show that under certain conditions, $\mathbb{V}[M_k] \le \mathbb{E}[M_k]$ remains approximately valid also for P_θ, hence

$$p \;\lesssim\; \Phi((\mu_k^{ub} - M_k)/\sqrt{\mu_k^{ub}}) \;\le\; \exp(-\tfrac{1}{2} n(\bar{M}_k - \bar{\mu}_k^{ub})^2/\bar{\mu}_k^{ub}) \;=\; e^{-O(n/k^{3/2})}$$

now without the fudge factor c_n. The same is true for the other tests.

Even and odd tests E and O. In Section 5 we explained why we need to exclude M_0 and M_1 from E and O. In addition, for $\theta_x = 1/d$ and $d \to \infty$, every x is observed exactly once, hence $M_1 = n$ and all other $M_k = 0$, also not leading to a useful test, so we also need to exclude M_n, i.e. $\alpha_k^{\text{even}} := k \cdot [\![0 \ne k \ne n \text{ even}]\!]$ and $\alpha_k^{\text{odd}} := k \cdot [\![1 \ne k \ne n \text{ odd}]\!]$. Let

$$f_{\text{even}}(\theta) \;:=\; \sum_k \alpha_k^{\text{even}} P_\theta(k) \;=\; \sum_{0 \ne k \ne n \text{ even}} k \cdot f_k^n(\theta) \;=\; \tfrac{1}{2} n\theta[1 - (1-2\theta)^{n-1} - 2\theta^{n-1}[\![n \text{ even}]\!]] \;\le\; \tfrac{1}{2} n\theta$$

$$f_{\text{odd}}(\theta) \;:=\; \sum_{1 \ne k \ne n \text{ odd}} k \cdot f_k^n(\theta) \;=\; \tfrac{1}{2} n\theta[1 - 2(1-\theta)^{n-1} + (1-2\theta)^{n-1} - 2\theta^{n-1}[\![n \text{ odd}]\!]] \;\le\; \tfrac{1}{2} n\theta$$

The equalities follow from binomial identities. The expressions in the brackets have the form $[1-h(\theta)]$ for 2×2 different functions $h(\theta)$: For f_{even} and f_{odd} and for even and odd n. In one case $h(\theta)\geq 0$ is trivial. In the other 3 cases this follows by finding the minimum $\theta^* = \frac{1}{3}|\frac{1}{2}|\frac{2}{3}$ resp. via $dh(\theta)/d\theta = 0$ and showing $h(\theta^*)\geq 0$. This establishes the upper bounds. The remaining definitions, derivations, and arguments are the same as in Section 5.

Slope tests $D_k := M_k - M_{k-1}$. For the slope test we have

$$\delta_k \;:=\; \mathbb{E}[D_k] \;=\; \sum_x P_{\theta_x}[N_x=k] - P_{\theta_x}[N_x=k-1] \;=\; \sum_x f_\delta(\theta_x) \;\leq\; \sup_{\theta>0}\frac{f_\delta(\theta)}{\theta} \;=:\; n\bar{\delta}_k^{ub}$$
$$\text{where}\quad f_\delta(\theta) := f_k^n(\theta) - f_{k-1}^n(\theta) \;=\; \tbinom{n+1}{k}\theta^{k-1}(1-\theta)^{n-k}[\theta - \tfrac{k}{n+1}]$$

The last expression follows from inserting (2) and elementary algebra. The maximum of $P_\theta(k)$ is at $\theta = k/n$ but the bracket $[\theta - \frac{k}{n+1}]$ kills this maximum, moving it to $\approx k+\sqrt{k}$ (Figure 1). As in Section 5 we can find it exactly by differentiating

$$\ln(f_\delta(\theta)/\theta) \;=\; \ln\tbinom{n+1}{k} + (k-2)\ln\theta + (n-k)\ln(1-\theta) + \ln(\theta - \tfrac{k}{n+1})$$
$$\tfrac{d}{d\theta}\ln(f_\delta(\theta)/\theta) \;=\; \tfrac{k-2}{\theta} - \tfrac{n-k}{1-\theta} + \tfrac{1}{\theta - k/(n+1)} \;\propto\; -[(n^2-1)\theta^2 + (k-2kn+n+1)\theta + k^2 - 2k] \;\stackrel{!}{=}\; 0$$

The last expression follows from multiplication with $\theta(1-\theta)[(n+1)\theta - k]$ and rearranging terms. This is a quadratic equation in θ with solution

$$\theta^* \;=\; \frac{2kn-k-n-1+\sqrt{k^2(5-4n)+k(4n^2-2n-6)+(n+1)^2}}{2(n^2-1)} \;=\; \frac{k-\frac{1}{2}}{n} + \frac{\sqrt{k+\frac{1}{4}}}{n} + O\Big(\frac{k^{3/2}}{n^2}\Big)$$

which is indeed the global maximum.

$$\bar{\delta}_k^{ub} \;=\; \frac{f_\delta(\theta^*)}{n\theta^*} \;=\; \frac{1-O(1/\sqrt{k})}{k^2\sqrt{2\pi e}} \tag{20}$$

The remainder is the same as in Section 5 with $\Phi_n \rightsquigarrow \Phi$ and $\mathbb{V}_{\boldsymbol{\theta}}[D_k] \approx \mathbb{V}_{\boldsymbol{\lambda}}[D_k] \leq ....$

Linear curvature tests $C_k := 2M_k - M_{k-1} - M_{k+1}$. Let

$$f_\gamma(\theta) \;:=\; 2P_\theta(k) - P_\theta(k-1) - P_\theta(k+1) \;=\; f_k^n(\theta)\Big[2 - \frac{k}{n-k+1}\frac{1-\theta}{\theta} - \frac{n-k}{k+1}\frac{\theta}{1-\theta}\Big]$$

be the negative curvature of P_θ. Unlike in Section 5 we cannot maximize $f_\gamma(\theta)/\theta$ exactly anymore, but the following upper bound is quite tight (cf. Figure 1)

$$\sup_{\theta>0}[f_\gamma(\theta)/\theta] \;\leq\; \sup_{\theta>0}\{f_k^n(\theta)/\theta\}\cdot\max_{z>0}[2-\alpha/\vartheta-\beta\vartheta] \;=\; \mu_k^{ub}[2-2\sqrt{\alpha\beta}]$$

where $\vartheta := \frac{\theta}{1-\theta}$, $\alpha = \frac{k}{n-k+1}$, $\beta = \frac{n-k}{k+1}$, and $\max_z$ is attained at $\vartheta^2 = \alpha/\beta$. Hence

$$n\bar{\gamma} \;:=\; \mathbb{E}[C_k] \;=\; \sum_x f_\gamma(\theta_x) \;\leq\; \max_{\theta>0}\frac{f_\gamma(\theta)}{\theta}$$
$$\leq\; n\bar{\gamma}_k^{ub} \;:=\; \mu_k^{ub}\Big[2-2\sqrt{\tfrac{k}{k+1}\tfrac{n-k}{n-k+1}}\Big] \;\leq\; \frac{\mu_k^{ub}}{(k+\frac{1}{2})(1-\frac{k}{n+1/2})} \;\approx\; \frac{n}{k^2\sqrt{2\pi k}} \tag{21}$$

where $\approx$ holds for $n \gg k \gg 1$. The remainder is the same as in Section 5.

Logarithmic curvature tests $\bar{U}_k := 2\ln M_k - \ln M_{k-1} - \ln M_{k+1}$. The derivation is the same as in Section 5 with minimal changes: With $\tilde{P}_{\boldsymbol{\theta},k}[X=x] := f_k^n(\theta_x)/\mu_k$ and $\vartheta_x := \theta_x/(1-\theta_x)$ we get

$$\frac{\mu_{k+1}}{\mu_k} = \frac{1}{\mu_k}\sum_x f_{k+1}^n(\theta_x) = \frac{1}{\mu_k}\sum_x \frac{n-k}{k+1}\vartheta_x f_k^n(\theta_x) = \frac{n-k}{k+1}\cdot\tilde{\mathbb{E}}_{\boldsymbol{\theta},k}[\vartheta_X]$$

$$\frac{\mu_{k-1}}{\mu_k} = \frac{1}{\mu_k}\sum_x f_{k-1}^n(\theta_x) = \frac{1}{\mu_k}\sum_x \frac{k}{n-k+1}\frac{1}{\vartheta_x} f_k^n(\theta_x) = \frac{k}{n-k+1}\cdot\tilde{\mathbb{E}}_{\boldsymbol{\theta},k}\left[\frac{1}{\vartheta_X}\right] \geq \frac{\frac{k}{n-k+1}}{\tilde{\mathbb{E}}_{\boldsymbol{\theta},k}[\vartheta_X]}$$

where we applied Jensen's inequality in the last step to convex function $1/\vartheta$. Taking the product, the dependence on unknown $\boldsymbol{\theta}$ cancels out:

$$\bar{v}_k := \ln\frac{\mu_k}{\mu_{k-1}}\frac{\mu_k}{\mu_{k+1}} \leq \ln\frac{n-k+1}{n-k}\frac{k+1}{k} =: \bar{v}_k^{ub} \leq \frac{1}{k}+\frac{1}{n-k} \tag{22}$$

The remainder is the same as in Section 5.

Summary. In the table below we summarize the most important quantities for the tests $T:\mathcal{X}^n\to\mathbb{R}$ derived in this section for comparison to the ones derived in Section 5 based on the Poisson approximation. For $\tau := \mathbb{E}[T]$ we derived tight upper bounds for all, even small k. The $n \gg k \gg 1$ approximations in the table are the same as for the multinomial model, but the referred to exact expressions differ.

Test Name	$T := n\bar{T} :=$	$\bar{\tau} := \mathbb{E}[\bar{T}] \leq$	$\mathbb{V}[\bar{T}] \lesssim \bar{V}^{ub} =$	θ^*	$O(\ln\frac{1}{p})$
Even$\neq 0$	$E := \sum_x N_x[\![N_x \neq 0 \text{ even}]\!]$	$\bar{\varepsilon}^{ub} = 1/2$	$\frac{1}{n}\sum_{k\neq 0 \& n \text{ even}} k^2 M_k$	$1/3 \mid 1/2$	n
Odd$\neq 1$	$O := \sum_x N_x[\![N_x \neq 1 \text{ odd}]\!]$	$\bar{o}^{ub} = 1/2$	$\frac{1}{n}\sum_{k\neq 1 \& n \text{ odd}} k^2 M_k$	$2/3 \mid 1/2$	n
2nd-Count	$M_k := \sum_x [\![N_x = k]\!]$	$\bar{\mu}_k^{ub} \overset{(19)}{\approx} \frac{1}{k\sqrt{2\pi k}}$	$\bar{\mu}_k^{ub}$	$\frac{k-1}{n-1}$	$\frac{n}{k^{3/2}}$
Slope	$D_k := M_k - M_{k-1}$	$\bar{\delta}_k^{ub} \overset{(20)}{\approx} \frac{1}{k^2\sqrt{2\pi e}}$	$\bar{M}_k + \bar{M}_{k-1}$	$\approx \frac{1}{n}[k+\sqrt{k}]$	$\frac{n}{k^{5/2}}$
Lin.Curv.	$C_k := 2M_k - M_{k-1} - M_{k+1}$	$\bar{\gamma}_k^{ub} \overset{(21)}{\approx} \frac{1}{k^2\sqrt{2\pi k}}$	$4\bar{M}_k + \bar{M}_{k-1} + \bar{M}_{k+1}$	$\approx k/n$	$\frac{n}{k^{7/2}}$
Log.Curv.	$\bar{U}_k := \ln(M_k^2/M_{k-1}M_{k+1})$	$\bar{v}_k^{ub} \overset{(22)}{\approx} \frac{1}{k}$	$\bar{M}_{k-1}^{-1} + 4\bar{M}_k^{-1} + \bar{M}_{k+1}^{-1}$	any	$\frac{n}{k^{7/2}}$

Claim 23 (IID tests). *Consider the test statistics $\bar{T}$ and associated upper bounds on their mean $\bar{\tau}$ and variance $\mathbb{V}[\bar{T}]$ from the above table. Then test $\bar{T}(x_{1:n}) \geq \sqrt{n}[\bar{\tau}^{ub} + z_\alpha\sqrt{\bar{V}^{ub}}]$ rejects that $x_{1:n}$ is iid with confidence $\gtrsim 1-\alpha$, i.e. at significance level $\lesssim \alpha$, where $z_\alpha := \Phi^{-1}(1-\alpha)$ (typically $\alpha = 0.05$ and $z_{0.05} \doteq 1.64$). The conditions under which $\lesssim$ is reasonably accurate are discussed in Appendix C. See Lemma 17 (with $Z_+ = T$ and $Z_x = T_x$) for p-values and further details.*

C Technical Lemmas for Multinomial vs Poisson

In this section we introduce the multinomial distribution $P_{\boldsymbol{\theta}}$ and Poisson process $P_{\boldsymbol{\lambda}}$ more carefully with the aim to find useful relations between their means and variances for our tests. We confirm that $\mathbb{E}_{\boldsymbol{\lambda}}[T]$ derived in Section 5 is close to $\mathbb{E}_{\boldsymbol{\theta}}[T]$ derived directly in Appendix B, not just for or specific tests but more generally. For the variance, we have only derived expressions for $\mathbb{V}_{\boldsymbol{\lambda}}$ (Section 5 and appendix A). The results in this Section show that under certain conditions $\mathbb{V}_{\boldsymbol{\theta}} \lesssim \mathbb{V}_{\boldsymbol{\lambda}}$, which in turn allows to avoid the fudge factor $c_n \approx \sqrt{2\pi n}$ used in Section 5. The precise conditions under which this is possible have yet to be worked out. Since we are comparing the multinomial with the Poisson distribution, P, $\mathbb{E}$, $\mathbb{V}$ are appropriately indexed with $\boldsymbol{\theta}$ or $\boldsymbol{\lambda}$.

Poisson distribution/process. The Poisson(λ) distribution $P_\lambda(k) := \lambda^k e^{-\lambda}/k! =: g_k(\lambda)$ for $\lambda \geq 0$ and $k \in \mathbb{N}_0$ has $\mathbb{E}_\lambda[k] = \mathbb{V}_\lambda[k] = \lambda$. For finite $\mathcal{X}$, let $(\Omega := \mathbb{N}_0^{\mathcal{X}}, 2^\Omega, P_{\boldsymbol{\lambda}})$ be the probability space of independent Poisson(λ_x) for $x \in \mathcal{X}$. For random variables $\boldsymbol{N}$ and atomic events $\{\boldsymbol{n}\}$ with $\boldsymbol{n} \in \mathbb{N}_0^{\mathcal{X}}$, with slight overload in notation we have

$$P_{\boldsymbol{\lambda}}(\boldsymbol{n}) \;:=\; P_{\boldsymbol{\lambda}}[\boldsymbol{N} = \boldsymbol{n}] \;:=\; \prod_{x \in \mathcal{X}} P_{\lambda_x}(n_x) \;=\; \prod_{x \in \mathcal{X}} \frac{\lambda_x^{n_x} e^{-\lambda_x}}{n_x!}$$

where $\lambda_x \geq 0$. We will assume $\sum_x \lambda_x = n$. For infinite $\mathcal{X}$, one defines $P_{\boldsymbol{\lambda}}$ on all finite partitions of $\mathcal{X}$, with λ of a partition being the sum (or measure in general) of λ_x over the x in the partition. These probabilities so defined on the "cylinder" set are indeed "self-consistent" in the sense that they can uniquely be extended in a standard way to a measure on $\mathbb{N}_0^{\mathcal{X}}$ with σ-algebra generated by the cylinders (a Poisson process). The details are of no concern to us.

What is important is that partitions of $\mathcal{X}$ are also products of Poissons. In particular, for the cylinders,

$$\begin{aligned}
E_k^x \;&:=\; \{\boldsymbol{n} \in \mathbb{N}_0^{\mathcal{X}} : n_x = k\} \\
E_{\bar{k}}^{\bar{x}} \;&:=\; \{\boldsymbol{n} \in \mathbb{N}_0^{\mathcal{X}} : n_{\bar{x}} = \bar{k}\}, \quad \text{where} \quad n_{\bar{x}} := \textstyle\sum_{x'' \neq x} n_{x''} \\
E_{\overline{kk'}}^{\overline{xx'}} \;&:=\; \{\boldsymbol{n} \in \mathbb{N}_0^{\mathcal{X}} : n_{\overline{xx'}} = \overline{kk'}\}, \quad \text{where} \quad n_{\overline{xx'}} := \textstyle\sum_{x'' \in \mathcal{X} \setminus \{x,x'\}} n_{x''} \\
P_{\boldsymbol{\lambda}}[E_k^x] \;&=\; P_{\boldsymbol{\lambda}}[N_x = k] \;=\; P_{\lambda_x}(k) \;\equiv\; g_k(\lambda_x) \;\equiv\; \lambda_x^k e^{-\lambda_x}/k! \\
P_{\boldsymbol{\lambda}}[E_k^x \cap E_{k'}^{x'} \cap E_{\overline{kk'}}^{\overline{xx'}}] \;&=\; P_{\boldsymbol{\lambda}}[E_k^x] \cdot P_{\boldsymbol{\lambda}}[E_{k'}^{x'}] \cdot P_{\boldsymbol{\lambda}}[E_{\overline{kk'}}^{\overline{xx'}}] \;=\; P_{\lambda_x}(k) \cdot P_{\lambda_{x'}}(k') \cdot P_{\lambda_{\overline{xx'}}}(\overline{kk'}) \\
\text{where} \quad \lambda_{\overline{xx'}} \;&:=\; \textstyle\sum_{x'' \in \mathcal{X} \setminus \{x,x'\}} \lambda_{x''} \;=\; n - \lambda_x - \lambda_{x'}
\end{aligned}$$

and similar for other combination of events. For the total sample size $N := N_+ = \sum_{x \in \mathcal{X}} N_x$, and associated event $E_n^{\mathcal{X}} := \{\boldsymbol{n} : \sum_x n_x = n\}$ we have

$$P_{\boldsymbol{\lambda}}[E_n^{\mathcal{X}}] \;\equiv\; P_{\boldsymbol{\lambda}}[N = n] \;=\; P_n(n) \;=\; n^n e^{-n}/n! \;\approx\; 1/\sqrt{2\pi n} \;=:\; 1/c_n$$

Independence also implies the elementary identities

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\lambda}}[N_x] \;&=\; \mathbb{V}_{\boldsymbol{\lambda}}[N_x] \;=\; \lambda_x \quad \text{and} \quad \mathrm{Cov}_{\boldsymbol{\lambda}}[N_x, N_{x'}] \;=\; 0 \quad \text{for} \quad x \neq x', \quad \text{hence} \\
\mathbb{E}_{\boldsymbol{\lambda}}[N] \;&=\; \sum_x \mathbb{E}_{\boldsymbol{\lambda}}[N_x] \;=\; \sum_x \lambda_x \;=\; n \quad \text{and} \quad \mathbb{V}_{\boldsymbol{\lambda}}[N] \;=\; \sum_x \mathbb{V}_{\boldsymbol{\lambda}}[N_x] \;=\; \sum_x \lambda_x \;=\; n
\end{aligned}$$

This means that $N = n \pm O_P(\sqrt{n})$ is close to n for large n.

Multinomial distribution. The multinomial distribution respects similar identities as the product of Poissons, except independence. Under certain conditions it is close to Poisson and close to independent. Traditionally, the multinomial is defined on sample space $\Omega_n := \{\boldsymbol{n} : \sum_x n_x = n\}$. For comparison to the Poisson this is inconvenient. We enlarge the support from Ω_n to $\Omega = \mathbb{N}_0^{\mathcal{X}}$ and define the multinomial zero on $\Omega \setminus \Omega_n$. For $\mathcal{X} = \{1:d\}$,

$$P_{\boldsymbol{\theta}}(\boldsymbol{n}) \;:=\; P_{\boldsymbol{\theta}}[\boldsymbol{N} = \boldsymbol{n}] \;:=\; \binom{n}{n_1, \ldots, n_d} \prod_{x=1}^d \theta_x^{n_x} \quad \text{for} \quad \sum_x n_x = n \quad \text{and 0 else}$$

where $\theta_x \geq 0$ and $\sum_x \theta_x = 1$. For countable $\mathcal{X}$ the above formula still applies, since only finitely many n_x can be non-zero, and $n_x = 0$ gives no contribution. For uncountable

$\mathcal{X}$ it can be extended in the same way as $P_{\boldsymbol{\lambda}}$ by partitioning $\mathcal{X}$. A special case is the binomial distribution

$$P_\theta(k) \;=\; f_k^n(\theta) \;:=\; \tbinom{n}{k}\theta^k(1-\theta)^{n-k} \quad \text{for } 0\leq\theta\leq 1 \text{ and } k\in\{0:n\} \text{ and } 0 \text{ else}$$

Some identities analogous to $P_{\boldsymbol{\lambda}}$ are

$$\begin{aligned} P_{\boldsymbol{\theta}}[E_k^x] \;&=\; P_{\boldsymbol{\theta}}[N_x{=}k] \;=\; P_{\theta_x}(k) \;\equiv\; f_k^n(\theta_x) \\ P_{\boldsymbol{\theta}}[E_k^x{\cap}E_{k'}^{x'}] \;&=\; P_{\boldsymbol{\theta}}[N_x{=}k{\wedge}N_{x'}{=}k'] \;=:\; f_{kk'}(\theta_x,\theta_{x'}) \;=\; \tbinom{n}{k\;k'}\theta_x^k\theta_{x'}^{k'}(1{-}\theta_x{-}\theta_{x'})^{n-k-k'} \end{aligned}$$

and similar for other combination of events. By definition, $P_{\boldsymbol{\theta}}[N{=}n]{=}P_{\boldsymbol{\theta}}[\Omega_n]{=}1$. We also have the elementary identities

$$\mathbb{E}_{\boldsymbol{\theta}}[N_x] \;=\; n\theta_x, \quad \mathbb{V}_{\boldsymbol{\theta}}[N_x] \;=\; n\theta_x(1{-}\theta_x) \quad \text{and} \quad \mathrm{Cov}_{\boldsymbol{\theta}}[N_x,N_{x'}] \;=\; -n\theta_x\theta_{x'} \quad \text{for} \quad x{\neq}x'$$

Note that N_x and $N_{x'}$ are negatively correlated, and E_k^x and $E_k^{x'}$ are *not* independent under $P_{\boldsymbol{\theta}}$ unlike $P_{\boldsymbol{\lambda}}$.

Exact relations between multinomial and Poisson distribution. For $\boldsymbol{\lambda}{=}n\boldsymbol{\theta}$, which we henceforth assume, the multinomial and product of Poissons are closely related. A straightforward bound for any *linear* combination $S{:=}\sum_x\beta_x N_x$ of N_x is

$$\begin{aligned} \mathbb{V}_{\boldsymbol{\theta}}[S] \;&=\; \sum_x\beta_x^2\mathbb{V}_{\boldsymbol{\theta}}[N_x]{+}\sum_{x\neq x'}\beta_x\beta_{x'}\mathrm{Cov}_{\boldsymbol{\theta}}[N_x,N_{x'}] \;=\; \sum_x\beta_x^2 n\theta_x(1{-}\theta_x){-}\sum_{x\neq x'}\beta_x\beta_{x'}n\theta_x\theta_{x'} \\ &=\; \sum_x\beta_x^2\lambda_x{-}n(\sum_x\beta_x\theta_x)^2 \;\leq\; \sum_x\beta_x^2\mathbb{V}_{\boldsymbol{\lambda}}[N_x] \;=\; \mathbb{V}_{\boldsymbol{\lambda}}[S] \end{aligned}$$

Unfortunately we need bounds for *non*-linear functions of N_x such as $M_k{=}\sum_x[\![N_x{=}k]\!]$, which are much harder to come by. First note that $P_{\boldsymbol{\lambda}}$ conditioned on $N{=}n$ exactly equals $P_{\boldsymbol{\theta}}$:

$$P_{\boldsymbol{\lambda}}[\boldsymbol{N}{=}\boldsymbol{n}|N{=}n] \;=\; \frac{P_{\boldsymbol{\lambda}}[\boldsymbol{N}{=}\boldsymbol{n}]}{P_{\boldsymbol{\lambda}}[N{=}n]} \;=\; \frac{\prod_x\lambda_x^{n_x}e^{-\lambda_x}/n_x!}{n^ne^{-n}/n!} \;=\; \binom{n}{\boldsymbol{n}}\prod_x\left(\frac{\lambda_x}{n}\right)^{n_x} \;=\; P_{\boldsymbol{\theta}}[\boldsymbol{N}{=}\boldsymbol{n}] \tag{23}$$

$$\text{hence} \quad \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{n})}{P_{\boldsymbol{\theta}}(\boldsymbol{n})} \;=\; \frac{P_{\boldsymbol{\lambda}}(\boldsymbol{n}|n)P_{\boldsymbol{\lambda}}(n)}{P_{\boldsymbol{\lambda}}(\boldsymbol{n}|n)} \;=\; P_{\boldsymbol{\lambda}}[N{=}n] \;=:\; \frac{1}{c_n} \;\eqsim\; \frac{1}{\sqrt{2\pi n}} \quad \text{provided} \quad \sum_x n_x{=}n$$

This implies that for *all* events $E{\subseteq}\Omega_n$, $P_{\boldsymbol{\lambda}}[E|n]{=}P_{\boldsymbol{\theta}}[E]$ but $P_{\boldsymbol{\lambda}}[E]{=}c_n{\cdot}P_{\boldsymbol{\theta}}[E]$, i.e. they differ by a factor of $c_n{=}O(\sqrt{n})$. The intuition is that the $P_{\boldsymbol{\theta}}$ probability mass of E is spread out in $P_{\boldsymbol{\lambda}}$ over $\sum_x n_x{=}n{\pm}O(\sqrt{n})$. For events E of probability exponentially small in n, the slack of a sub-polynomial c_n is benign in theory, but unfortunately not in practice. Another useful exact relation is

$$P_{\boldsymbol{\theta}}[E_k^x] \;=\; P_{\boldsymbol{\lambda}}[E_k^x|N{=}n] \;=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x{\cap}\{N{=}n\}]}{P_{\boldsymbol{\lambda}}[N{=}n]} \;=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x{\cap}E_{\bar{k}}^{\bar{x}}]}{P_{\boldsymbol{\lambda}}[N{=}n]} \;=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x]{\cdot}P_{\boldsymbol{\lambda}}[E_{\bar{k}}^{\bar{x}}]}{P_{\boldsymbol{\lambda}}[N{=}n]} \tag{24}$$

where $\bar{k}{=}n{-}k$ and $\lambda_{\bar{x}}{=}\sum_{x''\neq x}\lambda_{x''}$. This can also be verified by explicit calculation similar to (23).

On the other hand we will show that under certain conditions, $P_{\boldsymbol{\theta}}[E_k^x]{\approx}P_{\boldsymbol{\lambda}}[E_k^x]$ even without conditioning on $N{=}n$. For instance, we already know thar $\mathbb{E}_{\boldsymbol{\theta}}[N_x]{=}n\theta_x{=}$

$\lambda_x = \mathbb{E}_{\boldsymbol{\lambda}}[N_x]$ and $\mathbb{V}_{\boldsymbol{\theta}}[N_x] = n\theta_x(1-\theta_x) \approx n\theta_x = \lambda_x = \mathbb{V}_{\boldsymbol{\lambda}}[N_x]$ for small θ_x. There is no contradiction to $P_{\boldsymbol{\lambda}}[E] = c_n \cdot P_{\boldsymbol{\theta}}[E]$, since $E_k^x \not\subseteq \Omega_n$. Here the intuition is that E_k^x itself is spread out over a wide range of n', and $P_{\boldsymbol{\lambda}}[E_k^x|n']$ is approximately independent of n' at least over the range $n' \in [n \pm O(\sqrt{n})]$. If this is satisfied, then

$$P_{\boldsymbol{\lambda}}[E_k^x] \;=\; \sum_{n'} P_{\boldsymbol{\lambda}}[E_k^x|N=n']P_{\boldsymbol{\lambda}}(n') \;\approx\; P_{\boldsymbol{\lambda}}[E_k^x|N=n]\sum_{n'} P_{\boldsymbol{\lambda}}(n') \;=\; P_{\boldsymbol{\theta}}[E_k^x]$$

Approximate relations between multinomial and Poisson. We now derive our fundamental relation between a single Poisson $P_\lambda(k)$ and binomial $P_\theta(k)$ for $\lambda = n\theta$, which is the basis for all other approximations.

Lemma 24 (Expansion of log(Poisson/binomial)). *For $\kappa := k/n =: 1-\gamma$ and $\theta = \lambda/n$ fixed and $n \to \infty$,*

$$\ln\frac{P_\lambda(k)}{P_\theta(k)} \;\equiv\; \ln\frac{g_k(\lambda)}{f_k^n(\lambda)} \;=\; n(\kappa-\theta) - n\gamma\ln\frac{1-\theta}{\gamma} + \ln\sqrt{\gamma} + O(\frac{\kappa}{\gamma n})$$

For κ close to θ we can further approximate this by

$$\ln\frac{P_\lambda(k)}{P_\theta(k)} \;=\; \frac{n}{2\gamma}(\kappa-\theta)^2[1+O(\tfrac{1}{\gamma}|\kappa-\theta|)] + \ln\sqrt{\gamma} + O(\frac{\kappa}{\gamma n})$$

For $\kappa, \theta = o(n^{-1/2})$ this implies

$$\begin{aligned} P_\lambda(k) \;&=\; P_\theta(k)\cdot[1+(\tfrac{n}{2}(\kappa-\theta)^2(1+O(n(\kappa-\theta)^2+\kappa)) - \tfrac{1}{2}\kappa - O(\kappa^2+\kappa/n))] \\ &=\; P_\theta(k)\cdot[1\pm o(1)] \end{aligned}$$

For $\kappa, \theta \le cn^{-1/2-\delta}$ with $c, \delta > 0$ it implies $P_\theta(k) \lesseqgtr P_\lambda(k)\cdot[1 \pm c'n^{-2\delta}]$ for some $c' < \infty$.

Proof.

$$\begin{aligned}
\ln P_\lambda(k) \;&\equiv\; k\ln\lambda - \lambda - \ln k! \;=\; n(\kappa\ln\theta + \kappa\ln n - \theta) - \ln k! && \text{(by definition)} \\
\ln P_\theta(k) \;&\equiv\; \ln n! - \ln k! - \ln(\gamma n)! + n\kappa\ln\theta + n\gamma\ln(1-\theta) && \text{(by definition)} \\
\ln n!/(\gamma n)! \;&=\; n\kappa\ln\tfrac{n}{e} - \gamma n\ln\gamma - \ln\sqrt{\gamma} - O(\kappa/\gamma n) && \text{(by 2}\times\text{Stirling)} \\
\ln[P_\lambda(k)/P_\theta(k)] \;&=\; n\kappa\ln n - n\theta - \ln[n!/(\gamma n)!] - n\gamma\ln(1-\theta) && \text{(by lines 1\&2)} \\
&=\; n(\kappa-\theta) - n\gamma\ln\tfrac{1-\theta}{\gamma} + \ln\sqrt{\gamma} + O(\kappa/\gamma n) && \text{(by line 3)} \\
&=\; \tfrac{n}{2\gamma}(\kappa-\theta)^2 + \ln\sqrt{\gamma} + O(n(\kappa-\theta)^3/\gamma^2) + O(\kappa/\gamma n) && \text{(by next line)} \\
\ln\tfrac{1-\theta}{\gamma} \;&=\; \ln[1+\tfrac{\kappa-\theta}{\gamma}] \;=\; \tfrac{\kappa-\theta}{\gamma} - \tfrac{(\kappa-\theta)^2}{2\gamma^2} + O(\tfrac{\kappa-\theta}{\gamma})^3 && \text{(by Taylor)}
\end{aligned}$$

The last bound in the Lemma follows from exponentiating the previous bound, Taylor expanding the exponential, and noting that even the largest term $n(\kappa-\theta)^2 \to 0$ for $\kappa, \theta = o(n^{-1/2})$ and $1/\gamma = 1 + O(\kappa)$. ■

Assuming $k \le c \cdot n^{1/2-\delta}$ and $\theta_x \le c \cdot n^{-1/2-\delta}$ $\forall x$, the lemma implies $P_{\boldsymbol{\theta}}[E_k^x] \lesseqgtr P_{\boldsymbol{\lambda}}[E_k^x] \cdot [1 \pm c'n^{-2\delta}]$ uniformly for all x. Taking the sum over $x \in \mathcal{X}$, noting that $M_k^x = [\![N_x = k]\!] = [\![\boldsymbol{N} \in E_k^x]\!]$ and $M_k = \sum_x M_k^x$, we also have $\mathbb{E}_{\boldsymbol{\theta}}[M_k] \lesseqgtr \mathbb{E}_{\boldsymbol{\lambda}}[M_k] \cdot [1 \pm c'n^{-2\delta}]$. This extends to (positive) linear combinations of M_k:

Proposition 25 ($\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}] \approx \mathbb{E}_{\boldsymbol{\lambda}}[\boldsymbol{T}]$ and $\sum_{\boldsymbol{x}} \mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{M_k^x}] \approx \mathbb{V}_{\boldsymbol{\lambda}}[\boldsymbol{M_k}]$). *For $k_{max} \le c \cdot n^{1/2-\delta}$ and $\theta_x \le c \cdot n^{-1/2-\delta}$ $\forall x$, and random variable $T := \sum_{k \le k_{max}} \alpha_k M_k$ with $\alpha_k \ge 0$, we have $\mathbb{E}_{\boldsymbol{\theta}}[T] \lesseqgtr \mathbb{E}_{\boldsymbol{\lambda}}[T] \cdot [1 \pm c'n^{-2\delta}]$. In particular, $\mathbb{E}_{\boldsymbol{\theta}}[M_k] \lesseqgtr \mathbb{E}_{\boldsymbol{\lambda}}[M_k] \cdot [1 \pm c'n^{-2\delta}]$ and $\sum_x \mathbb{V}_{\boldsymbol{\theta}}[M_k^x] \lesseqgtr \mathbb{V}_{\boldsymbol{\lambda}}[M_k] \cdot [1 \pm c'n^{-2\delta}]$ for $k \le k_{max}$. For general $\alpha_k \lesseqgtr 0$, we have $|\mathbb{E}_{\boldsymbol{\theta}}[T] - \mathbb{E}_{\boldsymbol{\lambda}}[T]| \le \mathbb{E}_{\boldsymbol{\lambda}}[S] \cdot c'n^{-2\delta}$, where $S := \sum_k |\alpha_k| M_k$.*

For instance, for the slope test $D_k = M_k - M_{k-1}$, the correction is small *relative* to $\mathbb{E}_{\boldsymbol{\theta}}[D_k]$ iff $\mathbb{E}[M_k + M_{k-1}] \ll n^{\delta}\mathbb{E}[D_k]$. If M_k and M_{k-1} and D_k scale linearly in n, then this is the case. We derived upper bounds for $\mathbb{E}_{\theta}[T]$ directly in Appendix B, so we actually don't need to be concerned about approximation error for expectations. The stated simple bound on the variance unfortunately does not generalize to $\mathbb{V}_{\boldsymbol{\theta}}[T]$, not even $\mathbb{V}_{\boldsymbol{\theta}}[M_k]$, nor $\mathbb{V}_{\boldsymbol{\theta}}[\sum_k \alpha_k M_k^x]$.

Proof. The statement for M_k has been derived above. For T it follows from linearity of the expectation and $|P_{\boldsymbol{\theta}}[E_k^x] - P_{\boldsymbol{\lambda}}[E_k^x]| \le P_{\boldsymbol{\lambda}}[E_k^x]\varepsilon'$ derived in Lemma 24, where $\varepsilon' := c' n^{-2\delta}$:

$$|\mathbb{E}_{\boldsymbol{\theta}}[T] - \mathbb{E}_{\boldsymbol{\lambda}}[T]| \;\le\; \sum_x |\alpha_k| \cdot |P_{\boldsymbol{\theta}}[E_k^x] - P_{\boldsymbol{\lambda}}[E_k^x]| \;\le\; \sum_x |\alpha_k| P_{\boldsymbol{\lambda}}[E_k^x]\varepsilon' \;=\; \varepsilon' \mathbb{E}_{\boldsymbol{\lambda}}[S]$$

If $\alpha_k \ge 0 \;\forall k$, then $S = T$. The variance bound can be derived as follows: Lemma 24 implies

$$|(1 - P_{\boldsymbol{\theta}}[E_k^x]) - (1 - P_{\boldsymbol{\lambda}}[E_k^x])| \;=\; |P_{\boldsymbol{\lambda}}[E_k^x] - P_{\boldsymbol{\theta}}[E_k^x]| \;\le\; \varepsilon' P_{\boldsymbol{\lambda}}[E_k^x] \;\le\; \varepsilon'(1 - P_{\boldsymbol{\lambda}}[E_k^x])$$

In the last inequality we exploited $P_{\boldsymbol{\lambda}}[E_k^x] = P_{\lambda_x}(k) \le P_k(k) \le \frac{1}{\sqrt{2\pi k}} \le \frac{1}{2}$. Hence

$$1 - P_{\boldsymbol{\theta}}[E_k^x] \;\lessapprox\; (1 \pm \varepsilon')(1 - P_{\boldsymbol{\lambda}}[E_k^x])$$

For any function f this implies

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[f(M_k^x)] \;&=\; f(1) P_{\boldsymbol{\theta}}[E_k^x] + f(0)(1 - P_{\boldsymbol{\theta}}[E_k^x]) \\ &\lessapprox\; |f(1)| P_{\boldsymbol{\lambda}}[E_k^x](1 \pm \varepsilon') + |f(0)|(1 - P_{\boldsymbol{\lambda}}[E_k^x])(1 \pm \varepsilon') \;=\; (1 \pm \varepsilon')\mathbb{E}_{\boldsymbol{\lambda}}[|f(M_k^x)|]\end{aligned}$$

Specifically for the function $f_\lambda(M_k^x) := (M_k^x - \mu_\lambda)^2 \ge 0$, where $\mu_\lambda := \mathbb{E}_{\boldsymbol{\lambda}}[M_k^x]$, with $\mu_\theta := \mathbb{E}_{\boldsymbol{\theta}}[M_k^x]$, we get

$$\begin{aligned}\mathbb{V}_{\boldsymbol{\theta}}[M_k^x] \;&=\; \mathbb{E}_{\boldsymbol{\theta}}[(M_k^x - \mu_\theta)^2] \;\le\; \mathbb{E}_{\boldsymbol{\theta}}[(M_k^x - \mu_\lambda)^2] \;=\; \mathbb{E}_{\boldsymbol{\theta}}[f_\lambda(M_k^x)] \\ &\le\; (1 + \varepsilon')\mathbb{E}_{\boldsymbol{\lambda}}[f_\lambda(M_k^x)] \;=\; (1 + \varepsilon')\mathbb{V}_{\boldsymbol{\lambda}}[M_k^x]\end{aligned}$$

Reversing the role of λ and θ we get $\mathbb{V}_{\boldsymbol{\lambda}}[M_k^x] \le (1+\varepsilon')\mathbb{V}_{\boldsymbol{\theta}}[M_k^x]$ in the same way. Summing both bounds over x we get $\sum_x \mathbb{V}_{\boldsymbol{\theta}}[M_k^x] \lessapprox (1 \pm \varepsilon')\sum_x \mathbb{V}_{\boldsymbol{\lambda}}[M_k^x]$. The variance bound in the proposition now follows from independence of M_k^x w.r.t. $P_{\boldsymbol{\lambda}}$. ■

Upper bounding multinomial variances. We were able to derive upper bounds for $\mathbb{V}_{\boldsymbol{\lambda}}[T]$ (in Section 5) by exploiting independence $\mathrm{Cov}_{\boldsymbol{\lambda}}[M_k^x, M_{k'}^{x'}] = 0 \;\forall x \ne x'$, but not for $\mathbb{V}_{\boldsymbol{\theta}}[T]$, since $\mathrm{Cov}_{\boldsymbol{\theta}} \ne 0$. We need to show that $\mathrm{Cov}_{\boldsymbol{\theta}}$ is small. We do this by approximating $\mathrm{Cov}_{\boldsymbol{\theta}} \ne 0$ by $\mathrm{Cov}_{\boldsymbol{\lambda}} = 0$. The approximation error can be determined similarly as we did for the expectation.

Indeed, it can even be reduced to an application of Lemma 24: In addition to the Poisson notation above, let $E_{kk'}^{xx'} = \{\boldsymbol{n} : n_x + n_{x'} = k + k'\}$. Then for $x \ne x'$

$$\begin{aligned}P_{\boldsymbol{\theta}}[E_k^x \cap E_{k'}^{x'}] \;&=\; P_{\boldsymbol{\lambda}}[E_k^x \cap E_{k'}^{x'} | N = n] \;=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x \cap E_{k'}^{x'} \cap \{N = n\}]}{P_{\boldsymbol{\lambda}}[N = n]} \;=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x \cap E_{k'}^{x'} \cap \overline{E_{kk'}^{xx'}}]}{P_{\boldsymbol{\lambda}}[N = n]} \\ &=\; \frac{P_{\boldsymbol{\lambda}}[E_k^x] \cdot P_{\boldsymbol{\lambda}}[E_{k'}^{x'}] \cdot P_{\boldsymbol{\lambda}}[\overline{E_{kk'}^{xx'}}]}{P_{\boldsymbol{\lambda}}[N = n]} \overset{(24)}{=} P_{\boldsymbol{\lambda}}[E_k^x] P_{\boldsymbol{\lambda}}[E_{k'}^{x'}] \frac{P_{\boldsymbol{\theta}}[E_{kk'}^{xx'}]}{P_{\boldsymbol{\lambda}}[E_{kk'}^{xx'}]} \;\equiv\; g_k(\lambda_x) g_{k'}(\lambda_{x'}) \frac{f_{k+k'}(\theta_x + \theta_{x'})}{g_{k+k'}(\lambda_x + \lambda_{x'})}\end{aligned}$$

Again, one could verify this also by inserting the explicit expressions. That is, even though E_k^x and $E_{k'}^{x'}$ are not independent under $P_{\boldsymbol{\theta}}$, the probability is a "product" of 3 Poissons and 1 binomial. The above identity implies

$$\begin{aligned}
&\mathrm{Cov}_{\boldsymbol{\theta}}[M_k^x,M_{k'}^{x'}] \;\equiv\; P_{\boldsymbol{\theta}}[E_k^x\cap E_{k'}^{x'}]-P_{\boldsymbol{\theta}}[E_k^x]P_{\boldsymbol{\theta}}[E_{k'}^{x'}] \\
&= r_{kk'}(\lambda_x,\lambda_{x'})g_k(\lambda_x)g_{k'}(\lambda_{x'}) \quad \text{with} \quad r_{kk'}(\lambda_x,\lambda_{x'}) \;:=\; \left[\frac{f^n_{k+k'}(\theta_x+\theta_{x'})}{g_{k+k'}(\lambda_x+\lambda_{x'})}-\frac{f^n_k(\theta_x)}{g_k(\lambda_x)}\cdot\frac{f^n_{k'}(\theta_{x'})}{g_{k'}(\lambda_{x'})}\right] \\
&= r_{kk'}(\theta_x,\theta_{x'})f^n_k(\theta_x)f^n_{k'}(\theta_{x'}) \quad \text{with} \quad r_{kk'}(\theta_x,\theta_{x'}) \;:=\; \left[\frac{g_k(\lambda_x)}{f^n_k(\theta_x)}\frac{g_{k'}(\lambda_{x'})}{f^n_{k'}(\theta_{x'})}\frac{f^n_{k+k'}(\theta_x+\theta_{x'})}{g^n_{k+k'}(\lambda_x+\lambda_{x'})}\;-\;1\right]
\end{aligned}$$

The first/second expression is more convenient for theoretical/empirical upper bounds. Lemma 24 shows that $g_k/f_k^n\to 1$ for $\kappa,\theta=O(n^{-1/2-\delta})$ with $\delta>0$, hence $r_{kk'}$ tends to 0. More precisely

$$\begin{aligned}
r_{kk'}(\theta,\theta') \;&=\; \tfrac{n}{2}(\kappa-\theta)^2\cdot(1+O(n^{-\delta'}))-\tfrac{1}{2}\kappa \;+\; \tfrac{n}{2}(\kappa'-\theta')^2\cdot(1+O(n^{-\delta'}))-\tfrac{1}{2}\kappa' \\
&\quad -\; \tfrac{n}{2}(\kappa+\kappa'-\theta-\theta')^2\cdot(1+O(n^{-\delta'}))+\tfrac{1}{2}(\kappa+\kappa') \;+\; O(n^{-\delta''}) \\
&=\; -n(\theta-\kappa)(\theta'-\kappa')\cdot(1+O(n^{-\delta'}))+O(n^{-\delta''}) \;=\ldots=\; r_{kk'}(\lambda,\lambda')
\end{aligned}$$

where $\delta':=\min\{\frac{1}{2}+\delta,2\delta\}$ and $\delta'':=\min\{1+2\delta,\frac{3}{2}+\delta\}$. If we drop all $O()$-terms we get

$$\begin{aligned}
\mathbb{V}_{\boldsymbol{\theta}}[M_k] \;&=\; \textstyle\sum_x\mathbb{V}_{\boldsymbol{\theta}}[M_k^x]+\sum_{x\neq x'}\mathrm{Cov}_{\boldsymbol{\theta}}[M_k^x,M_{k'}^{x'}] \\
&\approx\; \textstyle\sum_x\mathbb{V}_{\boldsymbol{\theta}}[M_k^x]-\sum_{x\neq x'}n(\theta_x-\kappa)(\theta_{x'}-\kappa')f_k^n(\theta_x)f_{k'}^n(\theta_{x'}) \\
&=\; \textstyle\sum_x\mathbb{V}_{\boldsymbol{\theta}}[M_k^x]-n\sum_{x,x'}f_k^n(\theta_x)(\theta_x-\kappa)f_{k'}^n(\theta_{x'})(\theta_{x'}-\kappa')+n\sum_x[f_k^n(\theta_x)(\theta-\kappa)]^2 \\
&=\; \textstyle\sum_x\mathbb{V}_{\boldsymbol{\theta}}[M_k^x]-n[\sum_x f_k^n(\theta_x)(\theta_x-\kappa)]^2+n\sum_x[f_k^n(\theta_x)(\theta_x-\kappa)]^2
\end{aligned}$$

The middle term is negative. Adapting Lemma 12, or a bit more convenient using $f_k^n(\theta)\approx g_k(\lambda)$ by Lemma 24 and Lemma 12 directly, the last term can be upper bounded as

$$\begin{aligned}
&n\sum_x[f_k^n(\theta_x)(\theta_x-\kappa)]^2 \;\approx\; \frac{1}{n}\sum_x[g_k(\lambda_x)(\lambda_x-k)]^2 \;\leq\; \sup_{\lambda>0}\frac{[g_k(\lambda)(\lambda-k)^2]}{\lambda} \\
&\overset{(9)}{=} \left[\sup_{\lambda>0}\lambda^{3/2}\frac{g_\delta(\lambda)}{\lambda}\right]^2 \;\approx\; \left[(\lambda_+^*)^{3/2}\sup_{\lambda>0}\frac{g_\delta(\lambda)}{\lambda}\right]^2 \;=\; \lambda_+^*g_\delta(\lambda_+^*)^2 \;\approx\; kg_\delta(k)^2 \;\overset{(10)}{\approx}\; \frac{1}{2\pi ek}
\end{aligned}$$

which is very small compared to the typically linearly in n scaling $\mathbb{V}_{\boldsymbol{\theta}}[M_k]$. Ultimately we need an upper bound in terms of $\mathbb{E}[M_k]$, so

$$\begin{aligned}
&\textstyle\sum_x\mathbb{V}_{\boldsymbol{\theta}}[M_k^x] \;=\; \sum_x\mathbb{E}_{\boldsymbol{\theta}}[(M_k^x)^2]-\sum_x\mathbb{E}_{\boldsymbol{\theta}}[M_k^x]^2 \;=\; \mathbb{E}_{\boldsymbol{\theta}}[M_k]-\sum_x f_k^n(\theta_x)^2, \quad \text{hence} \\
&\mathbb{V}_{\boldsymbol{\theta}}[M_k] \;\lesssim\; \mathbb{E}_{\boldsymbol{\theta}}[M_k]+\textstyle\sum_x[f_k^n(\theta_x)^2[n(\theta_x-\kappa)]^2-1]
\end{aligned}$$

The same line of reasoning as above shows that

$$\textstyle\sum_x[f_k^n(\theta_x)^2[n(\theta_x-\kappa)]^2-1] \;\approx\; kg_\delta(k)^2-\frac{n}{k}g_k(k)^2 \;\approx\; \frac{1}{2\pi k}[\frac{1}{e}-\frac{n}{k}] \;\leq\; 0$$

The arguments above readily extend to linear combinations of M_k (cf. Lemma 18):

Claim 26 ($\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{T}]\lesssim\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$)**.** *For $k_{max}=o(n^{1/2})$ and $\sup_x\theta_x=o(n^{-1/2})$ and $T:=\sum_{k\leq k_{max}}\alpha_kM_k$ with $\alpha_k\in\mathbb{R}$, we have $\mathbb{V}_{\boldsymbol{\theta}}[T]\lesssim\mathbb{E}_{\boldsymbol{\theta}}[T]$. The smaller k_{max} and θ_x, the better the accuracy, The relative error is small under suitable further conditions.*

Law of total variation. Let T be a random variable of interest, e.g. one of our test statistics M_k or U_k, etc. Another potential approach towards proving Claim 26 is using the law of total variation:

$$\mathbb{V}_{\boldsymbol{\lambda}}[T] \;=\; \mathbb{E}_{\boldsymbol{\lambda}}[\mathbb{V}_{\boldsymbol{\lambda}}[T|N]]+\mathbb{V}_{\boldsymbol{\lambda}}[\mathbb{E}_{\boldsymbol{\lambda}}[T|N]] \;=\; \sum_{n'}\mathbb{V}_{\boldsymbol{\lambda}}[T|N=n']P_{\boldsymbol{\lambda}}[N=n']+\mathbb{V}_{\boldsymbol{\lambda}}[\mathbb{E}_{\boldsymbol{\lambda}}[T|N]]$$

Note that $\boldsymbol{\lambda}=n\boldsymbol{\theta}$ are fixed as before ($\boldsymbol{\lambda}\neq n'\boldsymbol{\theta}$ unless $n'=n$). We know that $P_{\boldsymbol{\lambda}}[N=n']$ has mean and variance n with light tails, so concentrates around $n'\in[n\pm(O\sqrt{n})]$. If $\mathbb{V}_{\boldsymbol{\lambda}}[T|N=n']$ does not change much in this interval, then the sum can be approximated by $\mathbb{V}_{\boldsymbol{\lambda}}[T|N=n]\equiv\mathbb{V}_{\boldsymbol{\theta}}[T]$. This implies

Observation 27 ($\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{T}]\lesssim\mathbb{V}_{\boldsymbol{\lambda}}[\boldsymbol{T}]$ **or even** $\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{T}]\approx\mathbb{V}_{\boldsymbol{\lambda}}[\boldsymbol{T}]$**).** *If* $\mathbb{V}_{\boldsymbol{\theta}}[T|N=n']$ *does not change much for* $n'\in[n\pm(O\sqrt{n})]$*, then* $\mathbb{V}_{\boldsymbol{\theta}}[T]\lesssim\mathbb{V}_{\boldsymbol{\lambda}}[T]$*. If in addition* $\mathbb{E}_{\boldsymbol{\lambda}}[T|N=n']$ *does not change much for* $n'\in[n\pm O(\sqrt{n})]$*, then* $\mathbb{V}_{\boldsymbol{\theta}}[T]\approx\mathbb{V}_{\boldsymbol{\lambda}}[T]$.

D List of Notation

Symbol	Type	Explanation
$a/bc=a/(bc)$		while $a/b\cdot c=(a/b)\cdot c$ though we actually always bracket the latter
$[\![\text{bool}]\!]$	$\in\{0,1\}$	=1 if bool=True, =0 if bool=False
i,j	$\in\mathbb{N}$	generic indices
$\{i:j\}$	$\subset\mathbb{Z}$	set of integers from i to j (empty if $j<i$)
$\mathbb{R},\mathbb{R}^+,\mathbb{R}_0^+$		reals, strictly positive reals, non-negative reals
$\lvert\mathcal{X}\rvert$	$\equiv\#\mathcal{X}$	size of set $\mathcal{X}$.
x	$\in\mathcal{X}$	single sample
$\mathcal{X}$		sample space of size $d=\lvert\mathcal{X}\rvert$, mostly $d=\infty$ and $\mathcal{X}$ countable.
$\mathcal{X}'$	$=\{x:\theta_x>0\}$	all x potentially observable $d'=\lvert\mathcal{X}'\rvert$.
$\mathcal{X}''$	$=\{x:n_x>0\}$	all x actually observed. $d''=\lvert\mathcal{X}''\rvert$
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
t	$\in\{1:n\}$	sample index
k	$\in\mathbb{N}_0$	second-order multiplicity index
$N_x \;=\; \#\{X_t:X_t=x\}$		(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, also $M_+:=M_1+M_2+...$
$x,n_x,m_k,\boldsymbol{m},...$		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
Y,Z		generic random variables
$\mathbb{E}$		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
$\rho = \mathbb{W}[Z]$	$:= \mathbb{E}[\lvert Z-\zeta\rvert^3]$	third absolute central moment
$O(),\Theta()$		classical $O()$ notation
$O_P()$		stochastic O-notation
$\approx$		approximately equal, informal
$f(n) \lesssim g(n)$		means $f(n) \leq g(n)\cdot[1+O_P(1/\sqrt{n})]$, and similarly $\gtrsim$ and $\eqsim$
$f(n) \lessapprox g(n)\cdot[1\pm\varepsilon]$		means $f(n) \leq g(n)\cdot[1+\varepsilon]$ and $f(n) \geq g(n)\cdot[1-\varepsilon]$
$\doteq$ e.g.	$z \doteq 1.64$	equal to within the number of displayed digits
$:=$, $\equiv$, $\overset{!}{=}$		definition, equal by earlier definition, want it to be equal
Z_x		collection of random variables with $x \in \mathcal{X}$
Z_+	$:= \sum_x Z_x$	sum of random variables
$\bar{Z}$	$:= Z_+/n$	*not* an average of random variables; also $\bar{Y} = Y/n$
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
$\dot{\varepsilon}_k, \ddot{\varepsilon}_k, ...$	$\in \mathbb{R}_0^+$	small corrections ≥ 0 tending to 0 for $k \to \infty$
$\tilde{\varepsilon}_k$	$\in \mathbb{R}$	small correction tending to 0 for $k \to \infty$
T	$: \mathcal{X}^n \to \mathbb{R}$	generic test statistic
$\tilde{T}$		uniformized test statistic ($P[\tilde{T} \leq \alpha] = \alpha$)
$E, O, M_k, D_k, C_k, \bar{U}_k$		specific test statistics
$\alpha = P_{\boldsymbol{\theta}}[T > c_\alpha]$		Type I error, prob. of falsely rejecting H_{iid}, significance level
$\beta(\alpha) = Q[T > c_\alpha]$		power of test T at level α for Q
$\diamondsuit$		end of example & end of notation & end of paper