Journal of Statistical Planning and Inference 236 (2025) 106243

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A family of discrete maximum-entropy distributions

David J. Hessen

Utrecht University, The Netherlands

ARTICLE INFO

Keywords:
maximum-entropy
Discrete support
Discrete normal distribution
Symmetric distribution
Exponential family

ABSTRACT

In this paper, a family of maximum-entropy distributions with general discrete support is derived. Members of the family are distinguished by the number of specified non-central moments. In addition, a subfamily of discrete symmetric distributions is defined. Attention is paid to maximum likelihood estimation of the parameters of any member of the general family. It is shown that the parameters of any special case with infinite support can be estimated using a conditional distribution given a finite subset of the total support. In an empirical data example, the procedures proposed are demonstrated.

1. Introduction

Entropy is the expected information inherent to the support of the distribution of a random variable (Shannon, 1948). Like the continuous normal distribution, the discrete normal distribution can be characterized as the distribution with maximum-entropy given specified mean and variance. Kemp (1997) derived the maximum-entropy distribution of a discrete random variable with a given mean, a given variance, and integer support. Another characterization of the discrete normal distribution with integer support was given by Roy (2003). In general, however, the sample space of a discrete random variable is either finite or countably infinite and the values it can take, need not be equally spaced nor integers. Moreover, the two parameter discrete normal distribution is quite restrictive and might not describe the true distribution of a discrete random variable well in practice. In this paper, therefore, a family of discrete maximum-entropy distributions with general discrete support is derived. Members of this family are distinguished by the number of specified non-central moments. It is shown that the discrete normal, the geometric, the discrete uniform, and the Bernoulli distributions are members of this family. Furthermore, a subfamily of discrete symmetric distributions is derived.

If the support is finite, then a member of the proposed family can easily be applied in practice because then the normalizing constant in the probability mass function involves a finite sum of terms. If the support is infinite, however, then the normalizing constant involves an infinite series for which a summation procedure may not exist. In this paper, therefore, it is shown that if the distribution of a discrete random variable is a member of the proposed family and has infinite support in the total population, then the random variable has the same distribution in any subpopulation defined by a subset of the support in the total population. As a consequence, the parameters of the discrete distribution in the total population can be estimated using finite support.

The rest of this paper has been organized as follows. In the next section, the family of discrete maximum-entropy distributions is derived and special cases are discussed. In the third section, attention is paid to maximum likelihood estimation. The fourth section deals with general goodness of fit tests and in the fifth section, an empirical data example is given.

E-mail address: d.j.hessen@uu.nl.

https://doi.org/10.1016/j.jspi.2024.106243
Received 18 June 2024; Accepted 24 September 2024
Available online 1 October 2024
0378-3758/© 2024 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2. A general probability mass function

Theorem 1. *The distribution of the discrete random variable X with support $S \subseteq \{x_1, x_2, x_3, \ldots\} \subset \mathbb{R} = (-\infty, \infty)$ and probability mass function*

$$P(X = x) = \pi_x = \delta \exp\left(\sum_{i=1}^{k} \beta_i x^i\right), \tag{1}$$

where $\delta = \left\{\sum_{x\in S} \exp\left(\sum_{i=1}^{k} \beta_i x^i\right)\right\}^{-1}$ is a normalizing constant, is characterized as the maximum-entropy distribution with specified $\mu_i = E(X^i) = \sum_{x\in S} x^i \pi_x$, for $i \in \{0, 1, \ldots, k\}$ and $k \leq |S| - 1$.

Proof. The entropy of a discrete distribution is $-\sum_{x\in S} \pi_x \ln \pi_x$. The maximum-entropy distribution with given $\mu_i = E(X^i) = \sum_{x\in S} x^i \pi_x$, for $i \in \{0, 1, \ldots, k\}$, is obtained by taking the derivative of the Lagrangian

$$\mathcal{L} = -\sum_{x\in S} \pi_x \ln \pi_x + \sum_{i=0}^{k} \beta_i \left(\sum_{x\in S} x^i \pi_x - \mu_i\right)$$

with respect to π_x, setting it equal to zero and solving for π_x. The derivative is

$$\frac{\partial \mathcal{L}}{\partial \pi_x} = -\ln \pi_x - 1 + \sum_{i=0}^{k} \beta_i x^i.$$

Setting the derivative equal to zero and solving for π_x yields Eq. (1), where $\delta = \exp(-1 + \beta_0)$. Taking into account that $\sum_{x\in S} \pi_x = 1$ and solving for δ gives $\delta = \left\{\sum_{x\in S} \exp\left(\sum_{i=1}^{k} \beta_i x^i\right)\right\}^{-1}$. Since $k + 1$ points can perfectly be connected by a polynomial of order k, $k \leq |S| - 1$. □

The family of discrete maximum-entropy distributions with general probability mass function in Eq. (1) is recognized as a subfamily of the exponential family, where $\beta_1, \ldots, \beta_k$ are the canonical parameters, $x, x^2, \ldots, x^k$ are the sufficient statistics, and $-\ln \delta$ is the cumulant function (Efron, 2022). The following four well-known discrete distributions are special cases.

Remark 1. If $k = 2$, $S = \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, $\beta_1 = \ln(\lambda q^{-1/2})$, and $\beta_2 = \ln q^{1/2}$, then X is a discrete normal random variable with probability mass function

$$P(X = x) = A\lambda^x q^{x(x-1)/2}, \tag{2}$$

where $\lambda > 0$, $0 < q < 1$, and $A = \left\{\sum_{x=-\infty}^{\infty} \lambda^x q^{x(x-1)/2}\right\}^{-1}$ is a normalizing constant (Kemp, 1997).

Remark 2. If $k = 1$, $S = \{1, 2, 3, \ldots\}$, and $\beta_1 \leq 0$, then X is a geometric random variable with $P(X = x) = \theta(1 - \theta)^{x-1}$, where $\theta = 1 - \exp(\beta_1)$.

Remark 3. If $S = \{x_1, \ldots, x_m\}$ and $\beta_i = 0$, for all i, then X is uniformly distributed with $P(X = x) = m^{-1}$, for all $x \in S$.

Remark 4. If $S = \{0, 1\}$, then X has a Bernoulli distribution with $P(X = x) = \varepsilon^x/(1 + \varepsilon)$, where $\varepsilon = \exp\left(\sum_{i=1}^{k} \beta_i\right)$, for $x \in \{0, 1\}$.

Remark 5. If S is finite and $k = |S| - 1$, then the discrete distribution is saturated.

Replacing $\sum_{i=1}^{k} \beta_i x^i$ in Eq. (1) by its kth order Taylor polynomial at ν yields

$$P(X = x) = \gamma \exp\left\{\sum_{r=1}^{k} \lambda_r (x - \nu)^r\right\}, \tag{3}$$

where $\gamma = \delta \exp(\lambda_0) = \left[\sum_{x\in S} \exp\left\{\sum_{r=1}^{k} \lambda_r (x - \nu)^r\right\}\right]^{-1}$, $\lambda_0 = \sum_{i=1}^{k} \beta_i \nu^i$, and $\lambda_r = \sum_{i=r}^{k} \binom{i}{r} \beta_i \nu^{i-r}$, for $r \in \{1, \ldots, k\}$. The number of parameters is now $k + 1$, whereas only k parameters are identifiable. This indeterminacy can be solved by arbitrarily fixing one of $\nu, \lambda_1, \ldots, \lambda_{k-1}$ to an arbitrary constant. The constraint should not be on λ_k because $\lambda_k = \beta_k$. Multiplying out the terms of the exponent in the right-hand side of Eq. (3) and collecting powers of x yields $\beta_i = \sum_{r=i}^{k} \binom{r}{i} \lambda_r (-\nu)^{r-i}$, for $i \in \{1, \ldots, k\}$.

Remark 6. If $k = 2$, $\lambda_1 = 0$, and $\lambda_2 = -\frac{1}{2\phi^2}$, then X is a discrete normal random variable with probability mass function

$$P(X = x) = \gamma \exp\left\{-\frac{1}{2}\left(\frac{x - \nu}{\phi}\right)^2\right\}, \tag{4}$$

where $\gamma = \left[\sum_{x\in S} \exp\left\{-\frac{1}{2}\left(\frac{x-\nu}{\phi}\right)^2\right\}\right]^{-1}$, $\phi^2 = -\frac{1}{2\beta_2}$, and $\nu = -\frac{\beta_1}{2\beta_2}$.

Parameter ν is not in general the mode nor the mean but can generally be interpreted as a location parameter. Parameter ϕ in Eq. (4) is not the standard deviation but can be interpreted as a scaling or dispersion parameter.

Theorem 2. *Parameter ν is a local extreme point of the continuous function $f(x) = \gamma \exp\left\{\sum_{r=1}^{k} \lambda_r (x-\nu)^r\right\}$ that passes through (x, π_x), for all $x \in S$.*

Proof. The derivative of $f(x)$ with respect to x is $f(x)\sum_{r=1}^{k} \lambda_r r(x-\nu)^{r-1}$ and equals zero if $x = \nu$. □

Theorem 2 implies that ν is the mode of X if continuous $f(x)$ is unimodal and $\nu \in S$. Next, it is well-known that if $P(X = \nu - \alpha_x) = P(X = \nu + \alpha_x)$, for all $x \in S$, then $P(X = x)$ is symmetric and ν is the mean. In the following theorem, conditions are given under which the probability mass function in Eq. (3) is symmetric and $\nu = \mu_1 = E(X)$.

Theorem 3. *If $\nu - \alpha_x$ and $\nu + \alpha_x$ are in S, for all x, and $\lambda_r = 0$, for odd r, then $P(X = x)$ is symmetric and $\nu = \mu_1 = |S|^{-1}\sum_{x\in S} x$. If in addition S is finite, then $\nu = (x_1 + x_{|S|})/2$.*

Proof. Since $P(X = \nu - \alpha_x) = P(X = \nu + \alpha_x) = \delta \exp\left(\sum_r \lambda_r \alpha_x^r\right)$, for even r and all $x \in S$, $P(X = x)$ is symmetric and $\nu = \mu_1$. Let $S_1 = \{x \mid x = \nu - \alpha_x\}$ and $S_2 = \{x \mid x = \nu + \alpha_x\}$. Then, $|S|^{-1}\sum_{x\in S} x = |S|^{-1}\left\{\sum_{x\in S_1}(\nu - \alpha_x) + \sum_{x\in S_2}(\nu + \alpha_x)\right\} = \nu$. If S is finite, then $x_1 = \nu - \alpha_{x_1}$ and $x_{|S|} = \nu + \alpha_{x_1}$. Solving the second equation for α_{x_1} and substitution into the first equation yields $x_1 = 2\nu - x_{|S|}$. Solving this equation for ν yields $\nu = (x_1 + x_{|S|})/2$. □

Remark 7. ν is indeterminate if S is infinite.

In the case S is finite, the condition given by $|S|^{-1}\sum_{x\in S} x = (x_1 + x_{|S|})/2$ is a necessary (but not sufficient) condition for $P(X = x)$ to be symmetric. So, if S is finite and $|S|^{-1}\sum_{x\in S} x \neq (x_1 + x_{|S|})/2$, then $P(X = x)$ is not symmetric. On the other hand, if $S = \{x_1, \ldots, x_m\}$ and $x_j = x_1 + (j-1)\alpha$, for $j \in \{1, \ldots, m\}$ and $\alpha > 0$ (equally spaced), then $\nu = |S|^{-1}\sum_{x\in S} x = (x_1 + x_{|S|})/2 = x_1 + \alpha(m-1)/2$, but this does not imply that $P(X = x)$ is symmetric.

Theorem 3 defines a subfamily of discrete symmetric distributions. However, if $P(X = x)$ is not symmetric but $\lambda_r = 0$, for odd r, then the continuous function $f(x)$ (in the proof of Theorem 3) that passes through (x, π_x), for all $x \in S$, is symmetric about ν on $\mathbb{R}$. So, a more general subfamily of discrete distributions is defined by the general probability mass function in Eq. (3) and the single condition that $\lambda_r = 0$, for odd r.

Remark 8. If $\lambda_r = 0$, for odd r, then X has probability mass function

$$P(X = x) = \gamma \exp\left\{\sum_{r=1}^{t} \lambda_{2r}(x-\nu)^{2r}\right\}, \tag{5}$$

where $\gamma = \left[\sum_{x\in S} \exp\left\{\sum_{r=1}^{t} \lambda_{2r}(x-\nu)^{2r}\right\}\right]^{-1}$, $t = \frac{k}{2}$ if k is even, and $t = \frac{k-1}{2}$ if k is odd.

3. Maximum likelihood estimation

Theorem 4. *If $B \subseteq S$ and the marginal distribution of discrete X is a member of the family of maximum-entropy distributions with general probability mass function in Eq. (3), then the conditional probability mass function of X given B is*

$$P(X = x \mid B) = \pi_{x|B} = \mathbf{1}_B(x)\gamma_B \exp\left\{\sum_{r=1}^{k} \lambda_r (x-\nu)^r\right\}, \quad \text{for all } x \in S, \tag{6}$$

where $\mathbf{1}_B(x)$ is an indicator function and $\gamma_B = \left[\sum_{x\in B} \exp\left\{\sum_{r=1}^{k} \lambda_r (x-\nu)^r\right\}\right]^{-1}$.

Proof. The conditional probability distribution of X given B is

$$P(X = x \mid B) = \frac{P(X = x, B)}{P(B)} = \frac{\mathbf{1}_B(x)P(X = x)}{\sum_{x\in B} P(X = x)}, \quad \text{for all } x \in S. \tag{7}$$

Substitution from Eq. (3) into Eq. (7) gives Eq. (6). □

Now, let n_x be the observed frequency of x in a sample of size n and let $O = \{x \mid n_x > 0\}$. Let B be a finite subset of S and a superset of O if S is infinite and let $B = S$ if S is finite. Assuming independence of observations, the likelihood function given support B equals

$$L(\lambda_1, \ldots, \lambda_k, \nu) = \gamma_B^n \exp\left\{\sum_{x\in O} n_x \sum_{r=1}^{k} \lambda_r (x-\nu)^r\right\}. \tag{8}$$

The log-likelihood function given support B and its first partial derivatives with respect to the parameters can be used to find the maximum likelihood estimates. The first partial derivatives of the log-likelihood function given support B with respect to λ_r and ν are

$$\frac{\partial l(\lambda_1,\ldots,\lambda_k,\nu)}{\partial \lambda_r} = \sum_{x\in O} n_x(x-\nu)^r - n\sum_{x\in B}(x-\nu)^r\pi_{x|B}, \tag{9}$$

$$\frac{\partial l(\lambda_1,\ldots,\lambda_k,\nu)}{\partial \nu} = n\sum_{x\in B}\sum_{r=1}^{k} r\lambda_r(x-\nu)^{r-1}\pi_{x|B} - \sum_{x\in O} n_x\sum_{r=1}^{k} r\lambda_r(x-\nu)^{r-1}, \tag{10}$$

from which it follows that the maximum likelihood estimate of $E\{(X-\nu)^r \mid B\}$ equals $n^{-1}\sum_{x\in O} n_x(x-\nu)^r$ and the maximum likelihood estimate of $E\left\{\sum_{r=1}^{k} r\lambda_r(X-\nu)^{r-1} \mid B\right\}$ equals $n^{-1}\sum_{x\in O} n_x \sum_{r=1}^{k} r\lambda_r(x-\nu)^{r-1}$. Note that these maximum likelihood estimates are realizations of unbiased estimators of the unconditional expected values $E\{(X-\nu)^r\}$ and $E\left\{\sum_{r=1}^{k} r\lambda_r(X-\nu)^{r-1}\right\}$ if the observations have been randomly drawn from the total population distribution.

Remark 9. If $B = S$ and $\nu = 0$, then $\lambda_r = \beta_r$, for $r \in \{1,\ldots,k\}$, and the maximum likelihood estimate of non-central moment $\mu_r = E(X^r)$ is given by sample moment $n^{-1}\sum_{x\in O} n_x x^r$, for $r \in \{1,\ldots,k\}$. Consequently, the maximum likelihood estimate of β_i coincides with its moment estimate.

Since the family of discrete maximum-entropy distributions is a subfamily of the exponential family, any special case for which ν is a fixed constant, can be fitted to data as a generalized linear model using iterative weighted least squares to find the maximum likelihood estimates of $\lambda_1,\ldots,\lambda_k$ (Nelder and Wedderburn, 1972; Charnes et al., 1976; McCullagh and Nelder, 1989; Dobson and Barnett, 2008). The expected frequency of x can be modeled with a Poisson distribution and a log link function. If ν is estimated, then $\lambda_1,\ldots,\lambda_{k-1}$ should be subjected to at least one constraint and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher, 1987) can be used to obtain the maximum likelihood estimates. The BFGS algorithm is a quasi-Newton method in which the Hessian matrix of second derivatives is approximated using updates specified by (approximate) evaluations of the first partial derivatives. Under regularity conditions (Agresti, 2013, p.529; Lehmann, 1999, pp. 499–501), maximum likelihood estimators under a subpopulation distribution have been shown to be asymptotically unbiased, consistent, and asymptotically efficient estimators of the parameters of the total population distribution (Hessen, 2023).

4. Goodness of fit

Let $B = \{x_1,\ldots,x_m\} \subseteq S$. The observed sample frequencies of the elements of B can be viewed as realizations of random frequencies that jointly follow a conditional multinomial distribution given B with parameters n and $\pi_{x|B}$, for all $x \in B$. Therefore, to assess the goodness of fit of a member of the family of discrete maximum-entropy distributions in the subpopulation defined by B, where $O \subseteq B \subseteq S$, Pearson's chi-square test might be appropriate. The observed value of Pearson's statistic is given by

$$\sum_{x\in B}\frac{(n_x - n\hat{\pi}_{x|B})^2}{n\hat{\pi}_{x|B}}, \tag{11}$$

where $\hat{\pi}_x$, for all x, is the estimate $\pi_{x|B}$ under the fitted member. If the assumed distribution holds in the subpopulation defined by B, then the Pearson statistic asymptotically follows a chi-square distribution (Mukhopadhyay, 2016). If the discrete maximum-entropy distribution is unconstrained, then this chi-square distribution has $|B| - 1 - k$ degrees of freedom. If the distribution is constrained to be symmetric, then the degrees of freedom are $|B| - k/2$, for even k, and $|B| - (k-1)/2$, for odd k.

An alternative goodness-of-fit test is the generalized likelihood ratio test of the assumed conditional distribution given B against the conditional saturated distribution given B. The likelihood function for the conditional saturated distribution given B is

$$L_B^* = \prod_{x\in B}\pi_{x|B}^{n_x}. \tag{12}$$

It is well-known that the value that maximizes the likelihood function with respect to $\pi_{x|B}$, for all $x \in B$, subject to $\sum_{x\in B}\pi_{x|B} = 1$ and $\sum_{x\in B} n_x = n$, is n_x/n, for all $x \in B$. So, the maximum of the log-likelihood function under the conditional saturated distribution given B is

$$\hat{l}_B^* = \sum_{x\in B} n_x \ln n_x - n\ln n. \tag{13}$$

The observed sample value of the likelihood ratio statistic is then given by

$$2\left(\hat{l}_B^* - \ln \hat{L}_B\right), \tag{14}$$

where $\hat{L}_B$ is the maximum of the likelihood function under the assumed conditional probability distribution given B. Under regularity conditions and the assumed conditional probability distribution, the likelihood ratio statistic has the same asymptotic chi-square distribution as the Pearson statistic (Wilks, 1938).

If observed sample frequencies are too low, then Pearson's chi-square test and the generalized likelihood ratio test of the assumed distribution against the saturated distribution are not appropriate because then both statistics are far from chi-square. Under these circumstances, a generalized likelihood ratio test of the assumed distribution against a less general alternative might be more appropriate. In addition, to select a discrete maximum-entropy distribution from among a number of likely candidates while taking the number of parameters into account, the AIC (Akaike, 1974) can be used.

Table 1
ACT mathematics sum score x, observed frequency n_x, and estimated percentile rank pr_x, for $x \in \{1, \ldots, 40\}$, in a sample of 4329 students.

x	n_x	pr_x	x	n_x	pr_x	x	n_x	pr_x	x	n_x	pr_x
1	1	0.00	11	192	12.66	21	147	56.93	31	91	87.06
2	1	0.00	12	192	16.78	22	163	60.75	32	83	89.19
3	3	0.03	13	192	21.22	23	147	64.38	33	73	91.16
4	9	0.11	14	201	25.85	24	140	67.82	34	72	92.98
5	18	0.35	15	204	30.56	25	147	71.08	35	75	94.62
6	59	0.90	16	217	35.26	26	126	74.15	36	50	96.07
7	67	1.93	17	181	39.88	27	113	77.06	37	37	97.31
8	91	3.59	18	184	44.38	28	100	79.79	38	38	98.33
9	144	5.96	19	170	48.74	29	106	82.37	39	23	99.10
10	149	9.01	20	201	52.93	30	107	84.79	40	15	99.65

Table 2
Goodness of fit results for different discrete distributions fitted to the sum scores of 4329 students on 40 ACT mathematics items.

Form	k	df	Pearson χ^2	p-value	LR χ^2	p-value	AIC
Canonical	2	38	380.92	0.00	445.19	0.00	694.09
	3	37	190.38	0.00	212.82	0.00	463.73
	4	36	40.85	0.27	40.80	0.27	293.70
	5	35	36.80	0.39	35.78	0.43	290.69
	6	34	33.12	0.51	30.61	0.64	287.51
	7	33	32.14	0.51	30.41	0.60	289.31
Symmetric	2	39	387.80	0.00	446.59	0.00	693.49
	4	38	341.35	0.00	365.00	0.00	613.91
	6	37	327.21	0.00	357.57	0.00	608.47
	8	36	327.46	0.00	357.57	0.00	610.47

Table 3
Estimation results for the maximum-entropy distribution with $k = 6$ that fits best to the sum scores of 4329 students on 40 ACT mathematics items.

Parameter	$m.l.e.$	std.error	z-value	p-value
β_1	$2.33e{+}00$	3.74e−01	6.23	4.79e−10
β_2	$-2.44e{-}01$	5.67e−02	−4.30	1.72e−05
β_3	$1.37e{-}02$	4.23e−03	3.24	1.21e−03
β_4	$-4.40e{-}04$	1.65e−04	−2.67	7.62e−03
β_5	$7.62e{-}06$	3.22e−06	2.36	1.81e−02
β_6	$-5.51e{-}08$	2.49e−08	−2.22	2.65e−02

5. Empirical data application

The data in Table 1 are the sum scores of 4329 students on 40 ACT mathematics items (Kolen and Brennan, 2004). The data are freely available in the R package `equate` (Albano, 2016; R Core Team, 2020).

The data are viewed as the observations of a random sample. The random variables that constitute the random sample are discrete random variables with common finite support $S = \{0, 1, 2, \ldots, 40\}$. If the students are randomly sampled from a target population, then the data can be used to create norms such as percentile ranks. For creating norms, the distribution of the discrete sum score in the target population has to be estimated. To avoid imprecise estimation and over-fitting as much as possible the selected parametric form for the sum score distribution should not be too simple nor too complex. A number of discrete maximum-entropy distributions in the canonical parameterization of Eq. (1) has been fitted to the data using the R statement

```
glm(xcount~poly(scale,degree=k,raw=T),data=ACTmath,family=Poisson)
```

where `xcount` is n_x and `scale` is x in the data file `ACTmath` of the `equate` package. The number of non-central moments k has been varied and it turns out to be sufficient to only report the results for $k \in \{2, \ldots, 7\}$. In addition, a number of discrete symmetric distributions has been fitted to the data using the R statement

```
glm(xcount~poly((scale-nu)^2,degree=t,raw=T),data=ACTmath,family=Poisson)
```

where $t = \frac{k}{2} \in \{1, 2, 3, 4\}$ and `nu` is $\nu = 40/2 = 20$ because the values in the support are equally spaced integers. The goodness of fit results for the selected distributions are given in Table 2.

The results in Table 2 show that out of the discrete distributions for which the goodness of fit results are presented, the maximum-entropy distribution in canonical form with $k = 6$ fits best to the data. Parameter estimation results for this best fitting distribution are given in Table 3.

Estimates of the discrete normal distribution (canonical form with $k = 2$), the best fitting distribution (canonical form with $k = 6$), the symmetric distribution with the lowest AIC ($t = 3$ or $k = 6$), and the saturated distribution, are plotted in Fig. 1.

Maximum likelihood estimates of the first 6 non-central moments under the best fitting discrete maximum-entropy distribution in canonical form with $k = 6$ are $\hat{\mu}_1 = 19.85$, $\hat{\mu}_2 = 461.55$, $\hat{\mu}_3 = 1.20e{+}04$, $\hat{\mu}_4 = 3.42e{+}05$, $\hat{\mu}_5 = 1.03e{+}07$, and $\hat{\mu}_6 = 3.24e{+}08$. In

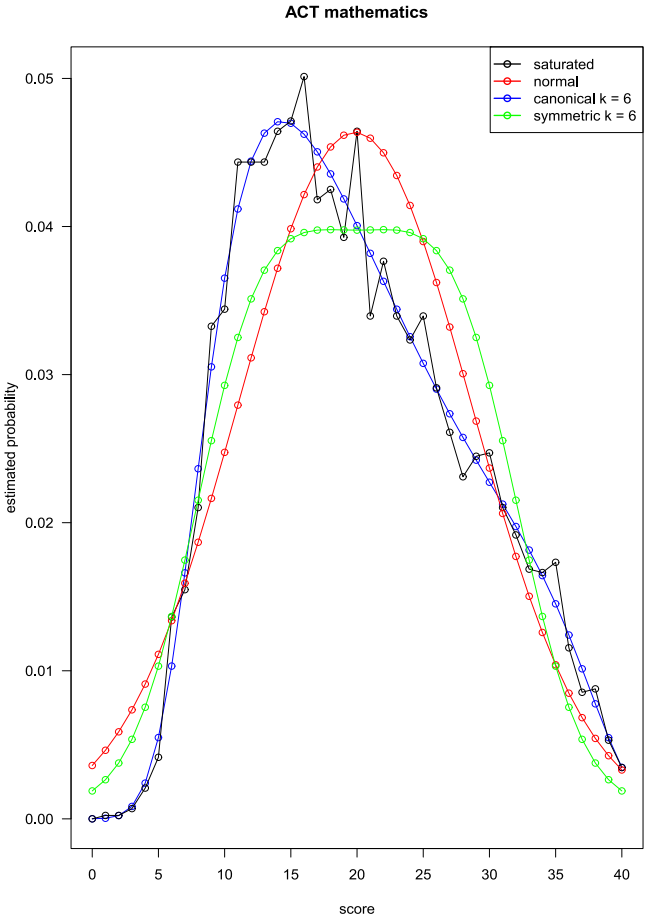

Fig. 1. Estimated discrete distributions for the sum scores of 4329 students on 40 ACT mathematics items.

Table 1, the estimated percentile rank $pr_x = 100 \sum_{u=0}^{x-1} \hat{\pi}_u$ under the best fitting discrete maximum-entropy distribution is given, for $x \in \{1, \ldots, 40\}$.

References

Agresti, A., 2013. Categorical Data Analysis. Wiley, New York.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.
Albano, A.D., 2016. Equate: An R package for observed-score linking and equating. J. Stat. Softw. 74, 1–36.
Charnes, A., Frome, E.L., Yu, P.L., 1976. The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. J. Amer. Statist. Assoc. 71, 169–171.
Dobson, A.J., Barnett, A.G., 2008. Introduction to Generalized Linear Models, third ed. Chapman and Hall/CRC.
Efron, B., 2022. Exponential Families in Theory and Practice. Cambridge University Press.
Fletcher, R., 1987. Practical Methods of Optimization, second ed. Wiley.
Hessen, D.J., 2023. Fitting and testing log-linear subpopulation models with known support. Psychometrika 88, 917–939.
Kemp, A.W., 1997. Characterizations of a discrete normal distribution. J. Statist. Plann. Inference 63, 223–229.
Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking, second ed. Springer, New York.
Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer.
McCullagh, P., Nelder, J., 1989. Generalized Linear Models, second ed. Chapman and Hall/CRC.
Mukhopadhyay, P., 2016. Complex Surveys: Analysis of Categorical Data. Springer.
Nelder, J., Wedderburn, R., 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 135, 370–384.
R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, https://www.R-project.org/.
Roy, D., 2003. The discrete normal distribution. Comm. Statist. Theory Methods 32, 1871–1883.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423, 623-656.
Wilks, S.S., 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9, 60–62.