ST 501: Fundamentals of Statistical Inference

Discrete Random Variables (part III)

Minh Tang

Fall 2024

Department of Statistics, North Carolina State University.

Estimation and Chebyshev's inequality

We have seen many examples where we first specify values for the parameters of the distribution and then compute probabilities associated with the r.v. Where (and or how) do we get these values ?

The modus operandi of many statistician is to (1) assume a particular model (that depends on certain parameters) for random variable of interest and (2) using existing data, estimate the corresponding parameters for the model.

Example A pollster polls $n = 100$ people, recording which of the two candidates A and B each person prefer. At the end, $X = 60$ people prefer candidate A and 40 people prefer candidate B. If the pollster assumes $X \sim \mathrm{Bin}(n, p)$, what is a reasonable value for p ?

Example Cook-In is planning to open a new store on Western Boulevard. Corporate management surveyed a random store on a weekday and observe that the store received 95 people during the lunch hours from 11 am to 1 pm. If the surveyor assumes $X \sim \mathrm{Pois}(\lambda)$, what is a reasonable value for λ ?

There are many possible techniques for these estimation problems. The two simplest, and most widely applicable, techniques are (1) maximum likelihood estimation and (2) method of moments.

For maximum likelihood estimation, we try to find the value of the parameters so that the probability of the observed data, given those parameters, is maximized, i.e., what parameter values make the observed data most likely ?

Example For the pollster example, assuming $X \sim \mathrm{Bin}(n, p)$, for any given p,

$$P(X = 60) = \binom{n}{60} p^{60}(1-p)^{n-60}.$$

We wish to maximize the above expression with respect to p.

Since n is known and $\binom{n}{60}$ does not depend on p, this is equivalent to maximizing $p^{60}(1-p)^{n-60}$. Since the logarithm is monotone increasing, this is equivalent to maximizing

$$\log(p^{60}(1-p)^{n-60}) = 60 \log p + (n-60)\log(1-p).$$

Setting the derivative of the above expressions to 0, we have

$$\frac{60}{p} - \frac{n-60}{1-p} = 0 \implies \frac{60(1-p)-(n-60)p}{p(1-p)} = 0 \implies \frac{60-np}{p(1-p)} = 0$$

Then $\hat{p} = 60/n$ is the maximum likelihood estimate for p.

Note In general, if $X \sim \mathrm{Bin}(n, p)$, then on observing $X = x$, the MLE of p is $\hat{p} = x/n$.

Example For the Cook-In example, assuming $X \sim \mathrm{Pois}(\lambda)$, then for any given λ

$$P(X = 95) = \frac{e^{-\lambda}\lambda^{95}}{95!}.$$

Similarly reasoning as in the previous example, we want to find λ to maximize

$$\log(e^{-\lambda}\lambda^{95}) = 95 \log \lambda - \lambda$$

Taking derivatives of the above expression and setting it to 0 yield the MLE $\hat{\lambda}$, i.e.,

$$\frac{95}{\lambda} - 1 = 0 \Longrightarrow \hat{\lambda} = 95$$

In summary, if $X \sim \mathrm{Pois}(\lambda)$ then on observing $X = x$, $\hat{\lambda} = x$ is the MLE of λ.

Once we obtain an estimate, the next natural question is "How accurate are these estimates"? As we will see later (in ST 372), this question can be answered in many different ways. The following is one simple and yet, for many purposes, reasonably powerful answer.

Proposition (Chebyshev's inequality)
Let X be a random variable with finite mean μ and finite variance σ^2. Then for any $k > 0$

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Example Suppose the pollster polls n people. Let $\hat{p} = X/n$ be the random variable denoting the MLE of p. Then

$$\mathbb{E}[\hat{p}] = \mathbb{E}[X]/n = np/n = p,$$
$$\mathrm{Var}[\hat{p}] = \mathrm{Var}[X]/n^2 = np(1-p)/n^2 = \frac{p(1-p)}{n}.$$

We then have

$$P\left(|\hat{p} - p| \geq k\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{k^2}$$

Suppose $k = 10$. Then

$$P\left(|\hat{p} - p| \geq 10\sqrt{\frac{p(1-p)}{n}}\right) \leq 0.01$$

Now, as n increases, $p(1-p)/n$ decreases. For example

$$\begin{aligned}
\sqrt{p(1-p)/n} &\leq 1/20 \quad \text{for } n = 100 \\
\sqrt{p(1-p)/n} &\leq 1/63 \quad \text{for } n = 1000 \\
\sqrt{p(1-p)/n} &\leq 1/200 \quad \text{for } n = 10000
\end{aligned}$$

In summary, for sufficiently large n, $|\hat{p} - p|$ is small with probability close to 1.

A formal statement of the above result is that

Proposition
Let $X \sim \mathrm{Bin}(n, p)$. Then $\hat{p} - p \to 0$ as $n \to \infty$. More specifically, for any $\epsilon > 0$

$$P(|\hat{p} - p| \leq \epsilon) \to 1$$

as $n \to \infty$.

Proposition
Let $X \sim \mathrm{Pois}(n\lambda)$. Then $\hat{\lambda} - \lambda \to 0$ as $n \to \infty$. More specifically, for any $\epsilon > 0$,

$$P(|\hat{\lambda} - \lambda| \leq \epsilon) \to 1$$

as $n \to \infty$.

Note The Poisson result can be interpreted as follows. Suppose λ is rate parameter for the average number of arrivals in one unit of time (or the number of events in one unit of area/volume). Then if we observe the random variable X for the number of arrivals in n units of time (for some sufficiently large n), then the MLE estimate $\hat{\lambda}$ will converge to λ.

Proof of Chebyshev's inequality

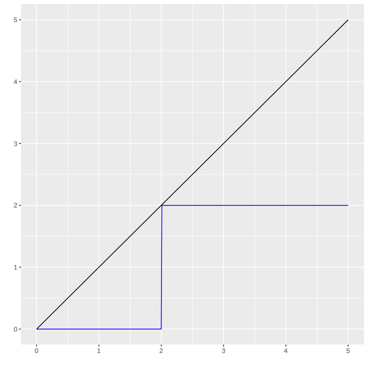
We first present the simple Markov inequality.

Proposition
Let X be a random variable. Then, for any $t > 0$,

$$P(|X| \geq t) \leq \frac{\mathbb{E}[|X|]}{t}$$

Proof of Markov's inequality Fixed a given t. Let Z be the random variable

$$Z = tI_{\{|X| \geq t\}} = \begin{cases} t & \text{if } |X| \geq t \\ 0 & \text{if } |X| < t \end{cases}$$



We note that $|X| \geq Z$ always. We therefore have

$$\mathbb{E}[|X|] \geq \mathbb{E}[Z] = \mathbb{E}[tI_{\{|X| \geq t\}}] = t\mathbb{E}[I_{\{|X| \geq t\}}] = tP(|X| \geq t)$$

and hence

$$P(|X| \geq t) \leq \frac{\mathbb{E}[|X|]}{t}.$$

Proof of Chebyshev's inequality For any random variable X with finite mean μ and finite variance σ^2

$$P(|X - \mu| \geq k\sigma) = P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}.$$

Moment generating functions

His reputation is expanding, faster than the universe. He once have an awkward moment, just to see how it feels.

(Dos Equis): The most interesting man in the world

Definition
Let X be a random variable. The moment generating function (mgf) for X is a real-valued function defined as

$$m_X(t) = \mathbb{E}[e^{tX}], \qquad t \in \mathbb{R}$$

The mgf is said to be well-defined if there exists a $b > 0$ such that $m_X(t)$ exists (is finite) for all t with $|t| < b$.

Example Let X be a Bernoulli random variable with $P(X = 1) = p$. Then

$$\begin{aligned}\mathbb{E}[e^{tX}] &= e^{t\times 0} \times P(X = 0) + e^{t\times 1} \times P(X = 1) \\ &= e^0 \times (1 - p) + e^t \times p \\ &= (1 - p) + pe^t\end{aligned}$$

which exists for all $t \in \mathbb{R}$. We emphasize that a mgf is a function that depends only on t.

mgf and sums

Theorem
Let X and Y be independent discrete random variables. Then for $Z = X + Y$,

$$\begin{aligned} m_Z(t) &= \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{tX+tY}] = \mathbb{E}[e^{tX}e^{tY}] \\ &= \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = m_X(t) \times m_Y(t). \end{aligned}$$

Proof Let S and T be discrete sets with $X \in S$ and $Y \in T$. Then

$$\begin{aligned}
\mathbb{E}[e^{tX+tY}] &= \sum_{(x,y)\in S\times T} e^{tx+ty} P(X=x, Y=y) \\
&= \sum_{(x,y)\in S\times T} e^{tx} e^{ty} P(X=x) P(Y=y) \\
&= \sum_{x\in S} \sum_{y\in T} e^{tx} P(X=x) e^{ty} P(Y=y) \\
&= \sum_{x\in S} e^{tx} P(X=x) \sum_{y\in T} e^{ty} P(Y=y) = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]
\end{aligned}$$

Example

Let $Z \sim \mathrm{Bin}(n, p)$. Then $Z = X_1 + X_2 + \cdots + X_n$ where the X_i are mutually independent Bernoulli random variables with a common success probability p. We have

$$\begin{aligned} m_Z(t) &= \mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t(X_1 + X_2 + \cdots + X_n)}] \\ &= \mathbb{E}[e^{tX_1}] \times \mathbb{E}[e^{tX_2}] \times \cdots \times \mathbb{E}[e^{tX_n}] \\ &= (1 - p + pe^t)^n \end{aligned}$$

Example Let $X \sim \mathrm{Geom}(p)$ be a geometric r.v. with success probability p. Then

$$
\begin{aligned}
\mathbb{E}[e^{tX}] &= \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1}p \\
&= pe^t \sum_{x=1}^{\infty} e^{t(x-1)}(1-p)^{x-1} \\
&= pe^t \sum_{\ell=0}^{\infty} ((1-p)e^t)^{\ell}
\end{aligned}
$$

If we now assume that $(1-p)e^t < 1$, i.e., $t < -\log(1-p)$, then

$$
\mathbb{E}[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}, \qquad t < -\log(1-p).
$$

Example Let $Z \sim \mathrm{NB}(r, p)$ be a negative binomial rv for the number of trials to get r successes with success probability p. Then $Z = X_1 + X_2 + \cdots + X_r$ where the X_i are mutually independent geometric rv with success probability p. Hence

$$\mathbb{E}[e^{tZ}] = \mathbb{E}[e^{tX_1}] \times \cdots \times \mathbb{E}[e^{tX_r}] = (\mathbb{E}[e^{tX_1}])^r = \left(\frac{pe^t}{1 - (1 - p)e^t}\right)^r.$$

which is well-defined for all $t < -\log(1 - p)$.

Example Let $X \sim \mathrm{Pois}(\lambda)$. Then

$$\begin{aligned}
\mathbb{E}[e^{tX}] &= \sum_{x=0}^{\infty} \frac{e^{tx} e^{-\lambda} \lambda^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} \\
&= e^{-\lambda} \exp(\lambda e^t) = \exp(\lambda(e^t - 1))
\end{aligned}$$

which is well-defined for all $t \in \mathbb{R}$.

Uniqueness of mgf

Theorem
Let X and Y be random variables. Suppose there exists a $b > 0$ such that

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{tY}] \qquad \textit{for all } |t| < b$$

Then X and Y have the same distribution.

The proof of the above theorem is way outside the scope of the class. In simple words, two distributions with the same mgf are necessarily the same distribution.

Example Let $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ with X and Y being independent random variables. Let $Z = X + Y$. Then

$$m_Z(t) = m_X(t) \times m_Y(t) = (1 - p + pe^t)^n \times (1 - p + pe^t)^m = (1 - p + pe^t)^{n+m}$$

which is the momement generating function for a $\mathrm{Bin}(n + m, p)$ random variable. Thus $Z \sim \mathrm{Bin}(n + m, p)$.

Suppose, however, that $X \sim \mathrm{Bin}(n, p_1)$ and $Y \sim \mathrm{Bin}(m, p_2)$ with $p_1 \neq p_2$. Then, for $Z = X + Y$,

$$m_Z(t) = m_X(t) \times m_Y(t) = (1 - p_1 + p_1 e^t)^n (1 - p_2 + p_2 e^t)^m$$

which cannot be written as $(1 - p + pe^t)^s$ for any p and any s. Thus $Z = X + Y$ cannot be a binomial rv.

From mgf to moments

Given a discrete rv X with mgf $m_X(t) = \mathbb{E}[e^{tX}]$, consider taking derivatives of $m_X(t)$ with respect to t. For example

$$\begin{aligned} m'_X(t) &= \frac{d}{dt}\mathbb{E}[e^{tX}] \\ &= \frac{d}{dt}\sum_x e^{tx}P(X=x) \\ &= \sum_x \frac{d}{dt}e^{tx}P(X=x) = \sum_x xe^{tx}P(X=x) \end{aligned}$$

Evaluating this derivative at $t = 0$ yield

$$m'_X(0) = \sum_x xe^{0\times x}P(X=x) = \sum_x xP(X=x) = \mathbb{E}[X]$$

As another example,

$$
\begin{aligned}
m_X''(t) &= \frac{d^2}{dt^2}\sum_x e^{tx}P(X=x) \\
&= \sum_x \frac{d^2}{dt^2} e^{tx}P(X=x) = \sum_x x^2 e^{tx} P(X=x)
\end{aligned}
$$

and hence $m_X''(0) = \mathbb{E}[X^2]$.

Proposition
Let X be a random variable with mgf $m_X(t)$. Then for any integer $k \geq 1$,

$$\mathbb{E}[X^k] = \frac{d^k}{dt^k} m_X(t)\Big|_{t=0} = m_X^{(k)}(0).$$

where $m_X^{(k)}(t)$ is the kth derivative (with respects to t) of $m_X(t)$.

Note In other words, $\mathbb{E}[X^k]$ (the moments of X) are given by derivatives of the mgf of X.

Example Consider the binomial rv $X \sim \mathrm{Bin}(n, p)$ with mgf

$$m_X(t) = (q + pe^t)^n$$

Example Consider the binomial rv $X \sim \mathrm{Bin}(n, p)$ with mgf

$$m_X(t) = (q + pe^t)^n$$

Straightforward manipulations yield

$$m'_X(t) = n(q + pe^t)^{n-1} \times pe^t$$
$$\mathbb{E}[X] = m'_X(0) = np(q + p)^{n-1} = np$$
$$m''_X(t) = n(n-1)(q + pe^t)^{n-2} \times p^2 e^{2t} + npe^t(q + pe^t)^{n-1}$$
$$\mathbb{E}[X^2] = m''_X(0) = n(n-1)p^2 + np$$

A bit more tedious algebraic manipulations yield,

$$
\begin{aligned}
m_X^{(3)}(t) &= n(n-1)(n-2)(q+pe^t)^{n-3} \times p^3 e^{3t} \\
&\quad + 2n(n-1)p^2 e^{2t}(q+pe^t)^{n-2} \\
&\quad + n(n-1)(q+pe^t)^{n-2} \times p^2 e^{2t} + npe^t(q+pe^t)^{n-1}
\end{aligned}
$$

from which we obtains

$$
\mathbb{E}[X^3] = m_X^{(3)}(0) = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np
$$

Calculation of $m_X^{(4)}(t)$ by hand will completely cure any symptoms of imsonia.

31

Example Let $X \sim \mathrm{Pois}(\lambda)$ with $m_X(t) = \exp(\lambda(e^t - 1))$. Then

$$m'_X(t) = \exp(\lambda(e^t - 1)) \times \lambda e^t$$
$$m''_X(t) = \exp(\lambda(e^t - 1)) \times \lambda^2 e^{2t} + \lambda e^t \exp(\lambda(e^t - 1))$$
$$\mathbb{E}[X] = \lambda, \quad \mathbb{E}[X^2] = \lambda^2 + \lambda$$

Summary: Uses of mgf

- To prove that a random variable has a certain distribution (using the uniqueness theorem)
- To derive higher order moments of a random variable by taking derivatives of the mgf. (This could be quite tedious to do by hand, but luckily one can use e.g., Maple or Mathematica).
- To find the distribution of a sum of independent random variables (using the multiplicative theorem for mgf)