UNIVERSITY OF LONDON

Goldsmiths
UNIVERSITY OF LONDON

Sept. 21st, 2021

A Weak Proof of the Central Limit Theorem with Moment Generating Function

Study Notes | Written by Larry Cui

Central Limit Theorem (CLT) is one of the two most important theorems in statistics (the other is the Large Number Theorem). In this note, the theorem of moment generating function (*mgf*) is used to prove the CLT without proving the mgf theorem itself. So we call this a weak proof. Notice that mgf may not always exist for all functions, so sometimes people use characteristic function instead to prove CLT.

1 Moment Generating Function

Definition

Let W be a random variable. The moment generating function (mgf) for W is denoted $M_W(t)$ and given by

$$M_W(t) = E(e^{tW}) = \begin{cases} \sum_{\text{all } k} e^{tW} p_W(k) & \text{if } W \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tW} f_W(w)\, dw & \text{if } W \text{ is continuous} \end{cases}$$

at all values of t for which the expected value exists.

In short, mgf is the expected value of e^{tW} for the generating function that has the pdf of $p_W(k)$ or $f_W(w)$, in discrete or continuous way, respectively. Sometimes, due to the property of the function, the expected value may go to infinity, and we say under such situation the mgf does not exist.

An illustration of mgf for **binomial random variable** X with pdf:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

is

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (1-p+pe^t)^n \end{aligned}$$

Another example is for **Poisson random variable**:

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

and

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \sum_{k=0}^{\infty} e^{tk}\frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!} \\ &= e^{-\lambda}e^{\lambda e^t} \\ &= e^{-\lambda+\lambda e^t} \end{aligned}$$

2 Some Important Features of mgf

mgf Theorem

Suppose that $W1$ and $W2$ are random variables for which $M_{W_1}(t) = M_{W_2}(t)$ for some interval of t's containing 0. Then $f_{W_1}(w) = f_{W_2}(w)$.

We will leave the proof of this theorem to further notes, but use this result directly to prove CLT.

That being said, we can still get some feeling about this theorem by intuition, for we know that

$$M_W^{(0)}(0) = \int e^{0w} f_W(w)\,dw = \int f_W(w)\,dw = 1$$

and

$$M_W^{(1)}(0) = \int we^{0w} f_W(w)\,dw = \int w f_W(w)\,dw = E(W)$$

and

$$M_W^{(2)}(0) = \int w^2 e^{0w} f_W(w)\,dw = \int w^2 f_W(w)\,dw = E(W^2)$$

$$\cdots$$

$$M_W^{(n)}(0) = \int w^n e^{0w} f_W(w)\,dw = \int w^n f_W(w)\,dw = E(W^n)$$

So we know that if two function's mgf are the same, then their expected values for random variables itself and higher orders are also the same. And since expected value, variance and other properties of a function can be represented by its mgf, we can almost tell that the two functions will behave same. The missing piece of this intuitive understanding is obvious, however: it does not rule out the possibility that why two different functions cannot generate the same mgf.

We move on to the next two important properties of mgf.

lemma a. Let W be a random variable with mgf $M_W(t)$. Let $V = aW + b$. Then

$$M_V(t) = e^{bt}M_W(at)$$

Proof: we prove the lemma for continuous function, the discrete function is the same.

$$\begin{aligned} M_V(t) &= \int e^{t(aW+b)} f_V(v)\, dv \\ &= \int e^{t(aW+b)} \frac{1}{|a|} f_W(w) a\, dw \\ &= \int e^{t(aW+b)} f_W(w)\, dw \\ &= e^{bt} \int e^{atW} f_W(w)\, dw \\ &= e^{bt} M_W(at) \end{aligned}$$

lemma b. Let $W_1, W_2, \ldots, W_n$ be independent random variables with mgfs $M_{W_1}(t), M_{W_2}(t), \ldots, M_{W_n}(t)$, respectively. Let $W = W_1 + W_2 + \cdots + W_n$. Then

$$M_W(t) = M_{W_1}(t) \cdot M_{W_2}(t) \cdots M_{W_n}(t)$$

Proof: In order to prove lemma b., we only need to prove that $M_W(t) = M_{W_1}(t) \cdot M_{W_2}(t)$ if $W = W_1 + W_2$. Because W_1 and W_2 are independent variables, we have pdf and cdf for W:

$$f_W(w) = f_{W_1}(w_1) f_{W_2}(w_2) \qquad \text{and} \qquad F_W(w) = \iint f_{W_1}(w_1) f_{W_2}(w_2)\, dw_1\, dw_2$$

and mgf for W:

$$\begin{aligned} M_W(t) = E(e^{t(W_1+W_2)}) &= \iint e^{t(W_1+W_2)} f_{W_1}(w_1) f_{W_2}(w_2)\, dw_1\, dw_2 \\ &= \int e^{tW_1} f_{W_1}(w_1)\, dw_1 \cdot \int e^{tW_2} f_{W_2}(w_2)\, dw_2 \\ &= M_{W_1}(t) \cdot M_{W_2}(t) \end{aligned}$$

3 Normal Distribution Function

pdf of a normal distribution

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right] \quad , \quad -\infty < y < \infty$$

Depending on the values of μ and σ, the pdf curve can take different shapes. When $\mu = 0$ and $\sigma = 1$, the normal distribution becomes $\frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}$ and is called **standard**.

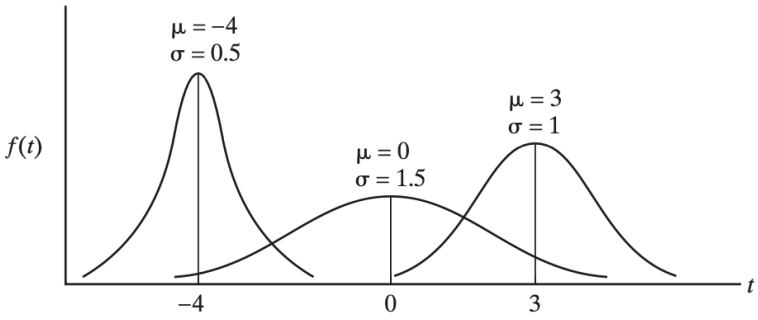

Figure 1: normal curves with different μ and σ

The adding up of pdf $f_Y(y)$ to 1 needs some polar coordinates integral techniques. But first of all, let's see if we can simplify this function for the integration purpose. We can tell from Figure 1 above that μ determine the location of the curve, but has nothing to do with the shape itself. So let $\mu = 0$ won't lose generosity to the calculation while gives us a huge convenience.

Proof: the basic idea is that we square the cdf of $f_Y(y)$, if the result equals 1, then cdf must be 1. We re-write the pdf as follows (deleting the μ part):

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2\right]$$

If we square the cdf of a function, the mathematical meaning is to find the joint density in an xy-plane, where x and y axes have the same pdf.

$$\text{cdf}^2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}\,dx \cdot \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2\sigma^2}}\,dy = \frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x^2+y^2)}\,dx\,dy$$

Let $x = r\cos\theta$ and $y = r\sin\theta$, then $dx\,dy = r\,dr\,d\theta$, and

$$\begin{aligned}
\frac{1}{2\pi\sigma^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x^2+y^2)}\,dx\,dy &= \frac{1}{2\pi\sigma^2}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} r\,dr\,d\theta \\
&= \frac{1}{2\pi\sigma^2}\int_0^{\infty} e^{-\frac{r^2}{2\sigma^2}} r\,dr \int_0^{2\pi} d\theta \\
&= \frac{1}{2\pi\sigma^2}\cdot\sigma^2\cdot 2\pi \\
&= 1
\end{aligned}$$

mgf of a normal distribution

$$M_Y(t) = \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$$

and

$$M_Y(t) = e^{\frac{t^2}{2}} \quad \text{when } \mu = 0 \text{ and } \sigma = 1$$

Calculating process of $M_Y(t)$:

$$M_Y(t) = E(e^{ty}) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{ty}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]dy$$
$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{y^2-2\mu y+\mu^2-2\sigma^2 ty}{2\sigma^2}\right]dy$$

Using a little technique, we can re-arrange the terms in the brackets as

$$y^2-(2\mu+2\sigma^2 t)y+(\mu+\sigma^2 t)^2-(\mu+\sigma^2 t)^2+\mu^2$$
$$=[y-(\mu+\sigma^2 t)]^2-\sigma^4 t^2+2\mu t\sigma^2$$

The last two terms involves no y, so can be moved out of the integral

$$M_Y(t) = \exp\left(\mu t+\frac{\sigma^2 t^2}{2}\right)\underbrace{\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left[\frac{y-(\mu+t\sigma^2)}{\sigma}\right]^2\right]dy}_{\text{equals 1}}$$

The latter part is still the cdf of a normal curve with $\mu' = (\mu + t\sigma^2)$, so the mgf is shortened to the first part of the equation.

4 Central Limit Theorem

CLT: Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ (both $< \infty$). Define

$$U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \qquad \text{where } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Then the distribution function of U_n converges to the standard normal distribution function as $n \to \infty$. That is, for all u (u is the value that variable U_n takes)

$$\lim_{n\to\infty} P(U_n \leqslant u) = \int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\,dt$$

The ordinary language interpretation of the CLT is that, no matter what the pdf is for the original function $f_Y(y)$, the sum or arithmetic average of Y_i, when i is big enough, [1] tends to match the normal distribution.

Why U_n?

U_n has a strange looking compared to sum or average of Y_i. But if the sum or average of Y_i has already followed the normal distribution, why should we construct U_n in such an awkward

[1] By convention, $i > 30$ will give us a pretty decent approximation.

looking way to bring out the CLT? The reason is for **standardization**. [2] Let's presume Y_i follows normal distribution for the time being, then the sum or average would have an expected value and variance as

$$E\left(\sum_{i=1}^{n} Y_i\right) = E(Y_1) + E(Y_2) + \cdots + E(Y_n) = n\mu$$

and [**larsen2012introduction**]

$$V\left(\sum_{i=1}^{n} Y_i\right) = V(Y_1) + V(Y_2) + \cdots + V(Y_n) = n\sigma^2$$

or

$$E(\overline{Y}) = \frac{1}{n}[E(Y_1) + E(Y_2) + \cdots + E(Y_n)] = \mu$$

and

$$V(\overline{Y}) = \frac{1}{n^2}[V(Y_1) + V(Y_2) + \cdots + V(Y_n)] = \frac{\sigma^2}{n}$$

Conclusion: variable U_n is constructed in such a way that it always has $\mu = 0$ and $\sigma = 1$, a standard form easy for analysis and calculation, where simplified normal distribution probability $e^{-t^2/2}$ are available everywhere.

Before proving the theorem, we also need a lemma about the mgf:

lemma c.

mgf can by approximated by polynomial. If we use Taylor's Theorem (here Maclaurin series, specifically), that

$$M(t) = M(0) + M'(0)t + M''(0)\frac{t^2}{2!} + \cdots + M^{(n)}(0)\frac{t^n}{n!} + M^{(n+1)}(\xi)\frac{t^{n+1}}{(n+1)!}$$

where $0 < \xi < t$.

Lemma c gives us an alternative to find a variable's mgf without knowing its pdf.

Proof of CLT:

Because $U_n = \dfrac{1}{\sqrt{n}} \dfrac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma}$, if let $Z_i = \dfrac{Y_i - \mu}{\sigma}$, we can write U_n as:

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$$

Since we know from the condition of the theorem that $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$, then

$$E(Z_i) = 0 \quad \text{and} \quad V(Z_i) = 1$$

Z_i is based on Y_i, which in turn comes from the same population, so each and every Z_i must

[2] Actually, we can never know the exact values of μ and σ of the total population. But how to find approximated values are another topic.

have the same pdf and mgf. Now we can re-write $M_{U_n}(t)$ as

$$\begin{aligned} M_{U_n}(t) &= M_{\frac{1}{\sqrt{n}}\sum Z_i}(t) \\ &= M_{\sum Z_i}\left(\frac{t}{\sqrt{n}}\right) \\ &= \left[M_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \qquad \text{by lemma b.} \end{aligned}$$

but $M_{Z_1}(t)$ can be written as a polynomial series by lemma c,

$$M_{Z_1}(t) = M_{Z_1}(0) + M'_{Z_1}(0)t + M''_{Z_1}(\xi)\tfrac{t^2}{2}, \qquad \text{where } 0 < \xi < t$$

and because $M_{Z_1}(0) = E(e^{0Z_1}) = E(1) = 1$, and $M'_{Z_1}(0) = E(Z_1) = 0$, the above series reduces to

$$M_{Z_1}(t) = 1 + \frac{M''_{Z_1}(\xi)}{2}t^2, \qquad \text{where} \quad 0 < \xi < t$$

as t is dummy variable here, it can be replaced by any other variables. If we replace it with $t/\sqrt{n}$, then

$$M_{Z_1}\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{M''_{Z_1}(\xi')}{2}\left(\frac{t}{\sqrt{n}}\right)^2 \qquad \text{where } 0 < \xi' < \frac{t}{\sqrt{n}}$$

Now we have an updated form for $M_{U_n}(t)$:

$$\begin{aligned} M_{U_n}(t) &= \left[1 + \frac{M''_{Z_1}(\xi')}{2}\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\ &= \left[1 + \frac{M''_{Z_1}(\xi')t^2/2}{n}\right]^n \end{aligned}$$

Notice that as $n \to \infty$, $\xi' \to 0$, so $M''_{Z_1}(\xi') \to M''_{Z_1}(0)$, and

$$M''_{Z_1}(0) = E(Z_1^2) = V(Z_1) + E(Z_1)^2 = 1$$

further reduce $M_{U_n}(t)$ to

$$M_{U_n}(t) = \left[1 + \frac{t^2/2}{n}\right]^n \qquad \text{where } \; n \to \infty$$

which is a familiar form for e to the power of anything above n, and here comes our conclusion:

$$M_{U_n}(t) = e^{\frac{t^2}{2}}$$

This is exactly the mgf of a normal distribution function with $\mu = 0$ and $\sigma = 1$. Also be noted that we didn't use pdf of the generating function to reach the mgf, which is in line with the CLT assertion that it applies to all functions no matter what pdf it is. So the proof.