15. Stein's Method

We have mentioned previously that a shortcoming of the limit theorems presented thus far is that they do not come with rates of convergence.

A proof of the Berry-Esseen theorem for normal convergence rates in the Kolmogorov metric is given in Durrett, and a proof of Poisson convergence with rates in the total variation metric is given there as well.

Rather than reproduce these classical results, we will obtain similar bounds using Stein's method in order to give a glimpse of this relatively modern technique which can be applied to all sorts of different distributions and often allows one to weaken assumptions such as independence as well.

As our purpose is expository, we will present some of the more straightforward approaches rather than seek out the best possible constants and conditions.

Stein's method refers to a framework based on solutions of certain differential or difference equations for bounding the distance between the distribution of a random variable X and that of a random variable Z having some specified target distribution.

The metrics for which this approach is applicable are of the form

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]|$$

for some suitable class of functions $\mathcal{H}$, and include the Kolmogorov, Wasserstein, and total variation distances as special cases. These cases arise by taking $\mathcal{H}$ to be the set of indicators of the form $1_{(-\infty,a]}$, 1-Lipschitz functions, and indicators of Borel sets, respectively. Convergence in each of these three metrics is strictly stronger than weak convergence (which can be metrized by taking $\mathcal{H}$ as the set of 1-Lipschitz functions with sup norm at most 1).

The basic idea is to find an operator $\mathcal{A}$ such that $E[(\mathcal{A}f)(X)] = 0$ for all f belonging to some sufficiently large class of functions $\mathcal{F}$ if and only if $\mathscr{L}(X) = \mathscr{L}(Z)$.

For example, we will see that $Z \sim \mathcal{N}(0,1)$ if and only if $E[f'(Z) - Zf(Z)] = 0$ for all Lipschitz functions f.

If one can then show that for any $h \in \mathcal{H}$, the equation

$$(\mathcal{A}f)(x) = h(x) - E[h(Z)]$$

has solution $f_h \in \mathcal{F}$, then upon taking expectations, absolute values, and suprema, one finds that

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]| = \sup_{h \in \mathcal{H}} |E[(\mathcal{A}f_h)(X)]| .$$

Remarkably, it is often easier to work with the right-hand side of this equation and the techniques for analyzing distances between probability distributions in this manner are collectively known as Stein's method.

Stein's method is a vast field with over a thousand existing articles and books and new ones written all the time, so we will only be able to scratch the surface here. In particular, we will not prove any results for dependent random variables. (Other than supplying convergence rates, the principal advantage of Stein's method is that it often enables one to prove limit theorems when there is some weak or local dependence, whereas characteristic function approaches typically fall apart when there is dependence of any sort.)

An excellent place to learn more about Stein's method (and the primary reference for this exposition) is the survey *Fundamentals of Stein's method* by Nathan Ross.

Normal Distribution.

We begin by establishing a characterizing operator for the standard normal.

Lemma 15.1. *Define the operator* $\mathcal{A}$ *by*

$$(\mathcal{A}f)(x) = f'(x) - xf(x).$$

If $Z \sim N(0,1)$*, then* $E[(\mathcal{A}f)(Z)] = 0$ *for all absolutely continuous* f *with* $E|f'(Z)| < \infty$*.*

Proof. Let f be as in the statement of the lemma. Then Fubini's theorem gives

$$\begin{aligned}
E[f'(Z)] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f'(x)e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} f'(x)e^{-\frac{x^2}{2}}dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} f'(x)e^{-\frac{x^2}{2}}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} f'(x)\left(-\int_{-\infty}^{x} ye^{-\frac{y^2}{2}}dy\right)dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} f'(x)\left(\int_{x}^{\infty} ye^{-\frac{y^2}{2}}dy\right)dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} ye^{-\frac{y^2}{2}}\left(-\int_{y}^{0} f'(x)dx\right)dy + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} ye^{-\frac{y^2}{2}}\left(\int_{0}^{y} f'(x)dx\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} ye^{-\frac{y^2}{2}}\left(f(y)-f(0)\right)dy + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} ye^{-\frac{y^2}{2}}\left(f(y)-f(0)\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} yf(y)e^{-\frac{y^2}{2}}dy - f(0)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}}dy \\
&= E[Zf(Z)] - f(0)E[Z] = E[Zf(Z)].
\end{aligned}$$

□

Of course, if $\|f'\|_\infty < \infty$, then $E|f'(Z)| < \infty$. It turns out that the condition $E[(\mathcal{A}f)(W)] = 0$ for all absolutely continuous f with $\|f'\|_\infty < \infty$ is also sufficient for $W \sim N(0,1)$.

To see that this is the case, we prove

Lemma 15.2. *If* Φ *is the distribution function for the standard normal, then the unique bounded solution to the differential equation*

$$f'(w) - wf(w) = 1_{(-\infty,x]}(w) - \Phi(x)$$

is given by

$$f_x(w) = \begin{cases} \sqrt{2\pi}e^{\frac{w^2}{2}}\left(1-\Phi(x)\right)\Phi(w), & w \le x \\ \sqrt{2\pi}e^{\frac{w^2}{2}}\Phi(x)(1-\Phi(w)), & w > x \end{cases}.$$

Moreover, f_x *is absolutely continuous with* $\|f_x\|_\infty \le \sqrt{\frac{\pi}{2}}$ *and* $\|f_x'\|_\infty \le 2$*.*

Proof. Multiplying both sides of the equation $f'(t) - tf(t) = 1_{(-\infty,x]}(t) - \Phi(x)$ by the integrating factor $e^{-\frac{t^2}{2}}$ shows that a bounded solution f_x must satisfy

$$\frac{d}{dt}\left(e^{-\frac{t^2}{2}}f_x(t)\right) = e^{-\frac{t^2}{2}}\left[f_x'(t) - tf_x(t)\right] = e^{-\frac{t^2}{2}}\left[1_{(-\infty,x]}(t) - \Phi(x)\right],$$

and integration gives

$$\begin{aligned}
f_x(w) &= e^{\frac{w^2}{2}}\int_{-\infty}^{w} e^{-\frac{t^2}{2}}\left(1_{(-\infty,x]}(t) - \Phi(x)\right)dt \\
&= -e^{\frac{w^2}{2}}\int_{w}^{\infty} e^{-\frac{t^2}{2}}\left(1_{(-\infty,x]}(t) - \Phi(x)\right)dt.
\end{aligned}$$

When $w \le x$, we have

$$f_x(w) = e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1 - \Phi(x)\right) dt$$
$$= \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w),$$

and when $w > x$, we have

$$f_x(w) = -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(0 - \Phi(x)\right) dt$$
$$= \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x) \frac{1}{\sqrt{2\pi}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)).$$

To check boundedness, we first observe that for any $z \ge 0$,

$$1 - \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-\frac{t^2}{2}} dt \le \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{(s+z)^2}{2}} ds$$
$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} e^{-sz} ds \le e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} ds = \frac{1}{2} e^{-\frac{z^2}{2}},$$

and, by symmetry, for any $z \le 0$,

$$\Phi(z) = 1 - \Phi\left(|z|\right) \le \frac{1}{2} e^{-\frac{z^2}{2}}.$$

Since f_x is nonnegative and $f_x(w) = f_{-x}(-w)$, it suffices to show that f_x is bounded above for $x \ge 0$.

If $w > x \ge 0$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)) \le \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}};$$

If $0 < w \le x$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w)$$
$$\le \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{x^2}{2}} \cdot 1 \le \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}};$$

and if $w \le 0 \le x$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w) \le \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}}.$$

The claim that f_x is the only bounded solution follows by observing that the homogeneous equation $f'(w) - wf(w) = 0$ has solution $f_h(w) = Ce^{\frac{w^2}{2}}$ for $C \in \mathbb{R}$, so the general solution is given by $f_x(w) + Cf_h(w)$, which is bounded if and only if $C = 0$.

Finally, we observe that, by construction, f_x is differentiable at all points $w \ne x$ with $f_x'(w) = wf_x(w) + 1_{(-\infty,x]}(w) - \Phi(x)$, so that

$$|f_x'(w)| \le |wf_x(w)| + \left|1_{(-\infty,x]}(w) - \Phi(x)\right| \le |wf_x(w)| + 1.$$

For $w > 0$,

$$|wf_x(w)| = \left| -we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt \right| \le we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left|1_{(-\infty,x]}(t) - \Phi(x)\right| dt$$
$$\le we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} dt \le we^{\frac{w^2}{2}} \int_{w}^{\infty} \frac{t}{w} e^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} \int_{w}^{\infty} te^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} e^{-\frac{w^2}{2}} = 1,$$

and for $w < 0$,

$$|wf_x(w)| = |-wf_{-x}(-w)| \leq 1,$$

hence $|f_x'(w)| \leq |wf_x(w)| + 1 \leq 2$.

Since f_x is continuous and differentiable at all points $w \neq x$ with uniformly bounded derivative, it is Lipschitz and thus absolutely continuous. $\square$

An immediate consequence of the preceding lemma is

Theorem 15.1. *A random variable W has the standard normal distribution if and only if*

$$E[f'(W) - Wf(W)] = 0$$

for all Lipschitz f.

Proof. Lemma 15.1 establishes necessity.

For sufficiency, observe that for any $x \in \mathbb{R}$, taking f_x as in Lemma 15.2 implies

$$|P(W \leq x) - \Phi(x)| = \left|E\left[1_{(-\infty,x]}(W) - \Phi(x)\right]\right| = |E\left[f_x'(W) - Wf_x(W)\right]| = 0. \qquad \square$$

The methodology of Lemma 15.2 can be extended to cover more general test functions than indicators of half-lines.

Indeed, the argument given there shows that for any function $h : \mathbb{R} \to \mathbb{R}$ such that

$$Nh := E[h(Z)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(z) e^{-\frac{z^2}{2}} dz$$

exists in $\mathbb{R}$, the differential equation

$$f'(w) - wf(w) = h(w) - Nh$$

has solution

$$(*) \qquad f_h(w) = e^{\frac{w^2}{2}} \int_{-\infty}^{w} (h(t) - Nh)\, e^{-\frac{t^2}{2}} dt.$$

Some fairly tedious computations which we will not undertake here show that

Lemma 15.3. *For any $h : \mathbb{R} \to \mathbb{R}$ such that Nh exists, let f_h be given by $(*)$.*

If h is bounded, then

$$\|f_h\|_\infty \leq \sqrt{\frac{\pi}{2}} \|h - Nh\|_\infty, \quad \|f_h'\|_\infty \leq 2\|h - Nh\|_\infty.$$

If h is absolutely continuous, then

$$\|f_h\|_\infty \leq 2\|h'\|_\infty, \quad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}} \|h'\|_\infty, \quad \|f_h''\|_\infty \leq 2\|h'\|_\infty.$$

(That the relevant derivatives are defined almost everywhere is part of the statement of Lemma 15.3.)

We can now give bounds on the error in normal approximation for sums of i.i.d. random variables.

We will work in the Wasserstein metric

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[h(W)] - E[h(Z)]|$$

where

$$\mathcal{H}_W = \{h : \mathbb{R} \to \mathbb{R} \text{ such that } |f(x) - f(y)| \leq |x - y| \text{ for all } x, y \in \mathbb{R}\}.$$

If $Z \sim N(0,1)$, then the preceding analysis shows that

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - W f_h(W)]|$$

where f_h is given by $(*)$.

Since Lipschitz functions are absolutely continuous, the second part of Lemma 15.3 applies with $\|h'\|_\infty = 1$.

From these observations and some elementary manipulations we have

Theorem 15.2. *Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with $E[X_i] = 0$ and $E[X_i^2] = 1$ for all $i = 1, \ldots, n$. If $W = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ and $Z \sim N(0,1)$, then*

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) \leq \frac{3}{n^{\frac{3}{2}}} \sum_{i=1}^n E\left[|X_i|^3\right].$$

Proof. Let f be any differentiable function with f' absolutely continuous, $\|f\|_\infty, \|f'\|_\infty, \|f''\|_\infty < \infty$.

For each $i = 1, \ldots, n$, set

$$W_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j = W - \frac{1}{\sqrt{n}} X_i.$$

Then X_i and W_i are independent, so $E[X_i f(W_i)] = E[X_i]E[f(W_i)] = 0$.

It follows that

$$E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i f(W)\right] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i)\right)\right].$$

Adding and subtracting $E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right]$ yields

$$E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] + E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right].$$

The independence and unit variance assumptions show that

$$E[X_i (W - W_i) f'(W_i)] = E\left[\frac{1}{\sqrt{n}} X_i^2 f'(W_i)\right] = \frac{1}{\sqrt{n}} E[X_i^2] E[f'(W_i)] = \frac{1}{\sqrt{n}} E[f'(W_i)],$$

so

$$E[W f(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] + E\left[\frac{1}{n} \sum_{i=1}^n f'(W_i)\right],$$

and thus

$$
\begin{aligned}
&|E[f'(W) - Wf(W)]| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} f'(W_i)\right] - E[f'(W)]\right| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} \left(f'(W_i) - f'(W)\right)\right]\right| \\
&\quad \le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} \left|f'(W_i) - f'(W)\right|\right]
\end{aligned}
$$

The Taylor expansion (with Lagrange remainder)

$$f(w) = f(z) + f'(z)(w - z) + \frac{f''(\zeta)}{2}(w - z)^2$$

for some ζ between w and z gives the bound

$$|f(w) - f(z) - (w - z)f'(z)| \le \frac{\|f''\|_\infty}{2}(w - z)^2,$$

so

$$
\begin{aligned}
\frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] &\le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \frac{\|f''\|_\infty}{2}(W - W_i)^2\right|\right] \\
&= \frac{\|f''\|_\infty}{2\sqrt{n}} \sum_{i=1}^{n} E\left|X_i \left(\frac{X_i}{\sqrt{n}}\right)^2\right| = \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right].
\end{aligned}
$$

Also, the mean value theorem shows that

$$\frac{1}{n} E\left[\sum_{i=1}^{n} \left|f'(W_i) - f'(W)\right|\right] \le \frac{1}{n} E\left[\sum_{i=1}^{n} \left(\|f''\|_\infty |W_i - W|\right)\right] = \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i|.$$

Since $1 = E[X_i^2] = E\left[\left(|X_i|^3\right)^{\frac{2}{3}}\right] \le E\left[|X_i|^3\right]^{\frac{2}{3}}$, we have $E\left[|X_i|^3\right] \ge 1$, hence $E|X_i| \le E\left[|X_i|^3\right]^{\frac{1}{3}} \le E\left[|X_i|^3\right]$. (The conclusion is trivial if $E\left[|X_i|^3\right] = \infty$.)

Putting all of this together gives

$$
\begin{aligned}
|E[f'(W) - Wf(W)]| &\le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} \left|f'(W_i) - f'(W)\right|\right] \\
&\le \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right] + \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i| \le \frac{3\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right],
\end{aligned}
$$

and the result follows since

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - Wf_h(W)]|$$

and $\|f_h''\|_\infty \le 2\|h'\|_\infty = 2$ for all $h \in \mathcal{H}_W$. □

Of course the mean zero variance one condition is just the usual normalization in the CLT and so imposes no real loss of generality. If the random variables have uniformly bounded third moments, then Theorem 15.2 gives a rate of order $n^{-\frac{1}{2}}$ which is the best possible.

We conclude with an example of a CLT with local dependence which can be proved using very similar (albeit more computationally intensive) methods.

Definition. A collection of random variables $\{X_1, ..., X_n\}$ is said to have *dependency neighborhoods* $N_i \subseteq \{1, 2, ..., n\}$, $i = 1, ..., n$, if $i \in N_i$ and X_i is independent of $\{X_j\}_{j \notin N_i}$.

Theorem 15.3. *Let $X_1, ..., X_n$ be mean zero random variables with finite fourth moments. Set $\sigma^2 = \mathrm{Var}\left(\sum_{i=1}^n X_i\right)$ and define $W = \sigma^{-1}\sum_{i=1}^n X_i$. Let $N_1, ..., N_n$ denote the dependency neighborhoods of $\{X_1, ..., X_n\}$ and let $D = \max_{i\in[n]} |N_i|$. Then for $Z \sim N(0,1)$,*

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) \leq \frac{D^2}{\sigma^3}\sum_{i=1}^n E\left[|X_i|^3\right] + \frac{D^{\frac{3}{2}}}{\sigma^2}\sqrt{\frac{28}{\pi}\sum_{i=1}^n E\left[X_i^4\right]}.$$

Poisson Distribution.

To illustrate some of the diversity in Stein's method techniques, we now look at size-biased couplings in Poisson approximation.

Definition. For a random variable $X \geq 0$ with $\mu = E[X] \in (0, \infty)$, we say that X^s has the *size-biased distribution* with respect to X if $E\left[Xf(X)\right] = \mu E\left[f(X^s)\right]$ for all f such that $E\left|Xf(X)\right| < \infty$.

To see that X^s exists, note that our assumptions imply that $Qf := \frac{1}{\mu}E\left[Xf(X)\right]$ is a well-defined linear functional on the space of continuous functions with compact support. Since X is nonnegative, we have that $Qf \geq 0$ for $f \geq 0$. Therefore, the Riesz representation theorem implies that there is a unique positive measure ν with $Qf = \int f d\nu$. Since $Q1 = \frac{1}{\mu}E[X] = 1$, ν is a probability measure. Thus $X^s \sim \nu$ satisfies

$$\frac{1}{\mu}E\left[Xf(X)\right] = Qf = \int f d\nu = E\left[f(X^s)\right].$$

Alternatively, one can adapt the argument from the following lemma to construct the distribution function of X^s in terms of that of X.

Lemma 15.4. *Let X be a nondegenerate $\mathbb{N}_0$-valued random variable with finite mean μ. Then X^s has mass function*

$$P\left(X^s = k\right) = \frac{kP\left(X = k\right)}{\mu}.$$

Proof.

$$\mu E\left[f\left(X^s\right)\right] = \sum_{k=0}^{\infty}\mu f(k)P\left(X^s = k\right) = \sum_{k=0}^{\infty}kf(k)P\left(X = k\right) = E\left[Xf(X)\right].$$

□

Size-biasing is an important consideration in statistical sampling.

For example, suppose that a school has $N(k)$ classes with k students.

Then the total number of classes is $n = \sum_{k=1}^{\infty} N(k)$ and the total number of students is $N = \sum_{k=1}^{\infty} kN(k)$.

If an outside observer were interested in estimating class-size statistics, they might ask a random teacher how large their class is.

Letting X denote the teacher's response, we have $P(X = k) = \dfrac{N(k)}{n}$ since $N(k)$ of the n classes have k students.

On the other hand, they might ask a random student how large their class is.

The student's response, Y, would have $P(Y = k) = \dfrac{kN(k)}{N}$ because $kN(k)$ of the N students are in a class of k students.

Noting that the expected number of students in a random class is

$$E[X] = \sum_{k=1}^{\infty} k\frac{N(k)}{n} = \frac{1}{n}\sum_{k=1}^{\infty} kN(k) = \frac{N}{n},$$

we see that

$$P(Y = k) = \frac{kN(k)}{N} = \frac{k\frac{N(k)}{n}}{\frac{N}{n}} = \frac{kP(X = k)}{E[X]},$$

so $Y = X^s$.

Observe that the average number of classmates of a random student (their self included) is

$$E[Y] = \sum_{k=1}^{\infty} k\frac{kN(k)}{N} = \frac{n}{N}\sum_{k=1}^{\infty} k^2\frac{N(k)}{n} = \frac{E\left[X^2\right]}{E[X]} \geq E[X].$$

The inequality is strict unless all classes have the same number of students.

Lemma 15.5. *Let $X_1, ..., X_n \geq 0$ be independent random variables with $E[X_i] = \mu_i$, and let X_i^s have the size-bias distribution w.r.t. X_i. Let I be a random variable, independent of all else, with $P(I = i) = \frac{\mu_i}{\mu}$, $i = 1, ..., n$, $\mu = \sum_{i=1}^{n} \mu_i$. If $W = \sum_{i=1}^{n} X_i$ and $W_i = W - X_i$, then $W^s = W_I + X_I^s$ has the W size-bias distribution.*

Proof.

$$\begin{aligned}\mu E[g(W^s)] &= \mu\sum_{i=1}^{n}\frac{\mu_i}{\mu}E[g(W_i + X_i^s)] \\ &= \sum_{i=1}^{n} E[X_i g(W_i + X_i)] \\ &= E\left[\sum_{i=1}^{n} X_i g(W)\right] = E[Wg(W)].\end{aligned}$$ □

Lemma 15.6. *If $P(X = 1) = 1 - P(X = 0) = p$, then $Y \equiv 1$ has the X size-bias distribution.*

Proof.

$$P(X^s = 1) = \frac{1 \cdot P(X = 1)}{E[X]} = \frac{p}{p} = 1.$$ □

To connect size-biasing with Poisson approximation, we need the following facts, which are proved in much the same fashion as the analogous results for the normal distribution.

Theorem 15.4. *Let $\mathcal{P}_\lambda$ denote the* Poisson(λ) *distribution. An $\mathbb{N}_0$-valued random variable X has law $\mathcal{P}_\lambda$ if and only if*

$$E\left[\lambda f\left(X+1\right)-Xf(X)\right]=0$$

for all bounded f.

Also, for each $A\subseteq\mathbb{N}_0$, the unique solution of the difference equation

$$\lambda f(k+1)-kf(k)=1_A(k)-\mathcal{P}_\lambda(A),\ f_A(0)=0$$

is given by

$$f_A(k)=\lambda^{-k}e^{\lambda}(k-1)!\left[\mathcal{P}_\lambda\left(A\cap U_k\right)-\mathcal{P}_\lambda\left(A\right)\mathcal{P}_\lambda\left(U_k\right)\right]\ \text{ where } U_k=\{0,1,...,k-1\}.$$

Finally, writing the forward difference as $\Delta g(k):=g(k+1)-g(k)$, we have

$$\|f_A\|_\infty\leq\min\left\{1,\lambda^{-\frac{1}{2}}\right\}\ \text{ and }\ \|\Delta f_A\|_\infty\leq\frac{1-e^{-\lambda}}{\lambda}.$$

We can now prove

Theorem 15.5. *Let X be an $\mathbb{N}_0$-valued random variable with $E[X]=\lambda$, and let $Z\sim$* Poisson(λ)*. Then*

$$d_{TV}(X,Z)\leq\left(1-e^{-\lambda}\right)E\left|X+1-X^s\right|.$$

Proof. Letting f_A be as in Theorem 15.4, the definitions of total variation and size-biasing imply

$$\begin{aligned}
d_{TV}(X,Z)&=\sup_A\left|P\left(X\in A\right)-P\left(Z\in A\right)\right|\\
&=\sup_A\left|\lambda E\left[f_A(X+1)\right]-E\left[Xf_A(X)\right]\right|\\
&=\sup_A\left|\lambda E\left[f_A(X+1)\right]-\lambda E\left[f_A(X^s)\right]\right|\\
&\leq\lambda\sup_A E\left|f_A(X+1)-f_A(X^s)\right|\\
&\leq\lambda\sup_A\|\Delta f_A\|_\infty E\left|X+1-X^s\right|\\
&\leq\left(1-e^{-\lambda}\right)E\left|X+1-X^s\right|.
\end{aligned}$$

The penultimate inequality follows by writing $f_A(X+1)-f_A(X^s)$ as a telescoping sum of $|X+1-X^s|$ first order differences. ☐

We conclude with a simple proof of Theorem 14.1 complete with a total variation bound.

Theorem 15.6. *Let $X_1,...,X_n$ be independent random variables with $P\left(X_i=1\right)=1-P\left(X_i=0\right)=p_i$, and set $W=\sum_{i=1}^n X_i$, $\lambda=E[W]=\sum_{i=1}^n p_i$. Let $Z\sim$* Poisson(λ)*. Then*

$$d_{TV}\left(W,Z\right)\leq\frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^n p_i^2.$$

Proof. Lemmas 15.5 and 15.6 show that $W^s=W_I+X_I^s=W-X_I+1$ where I is a random variable, independent of the X_i's, with $P\left(I=i\right)=\frac{p_i}{\lambda}$.

Thus, by Theorem 15.5,

$$\begin{aligned}
d_{TV}(W,Z) &\leq \left(1-e^{-\lambda}\right) E\left|W+1-W^s\right| = \left(1-e^{-\lambda}\right) E\left|X_I\right| \\
&= \left(1-e^{-\lambda}\right) \sum_{i=1}^{n} E\left|X_i\right| P(I=i) \\
&= \left(1-e^{-\lambda}\right) \sum_{i=1}^{n} p_i \frac{p_i}{\lambda} = \frac{1-e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i^2.
\end{aligned}$$

□

One can also prove Theorem 15.6 without taking a detour through size-biasing.

Indeed, suppose that f satisfies $\|f\|_\infty, \|\Delta f\|_\infty < \infty$. Then

$$\begin{aligned}
E\left[W f(W)\right] &= E\left[\sum_{i=1}^{n} X_i f(W)\right] \\
&= \sum_{i=1}^{n} p_i E\left[f(W) \,|X_i = 1\right] = \sum_{i=1}^{n} p_i E\left[f(W_i+1)\right].
\end{aligned}$$

Since $\lambda = \sum_{i=1}^{n} p_i$, we have

$$\begin{aligned}
\left|\lambda E\left[f(W+1)\right] - E[W f(W)]\right| &= \left|\sum_{i=1}^{n} p_i E\left[f(W+1)\right] - \sum_{i=1}^{n} p_i f(W_i+1)\right| \\
&\leq \sum_{i=1}^{n} p_i E\left|f(W+1) - f(W_i+1)\right| \\
&\leq \sum_{i=1}^{n} p_i \|\Delta f\|_\infty E\left|(W+1)-(W_i+1)\right| \\
&= \|\Delta f\|_\infty \sum_{i=1}^{n} p_i E\left|X_i\right| = \|\Delta f\|_\infty \sum_{i=1}^{n} p_i^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
d_{TV}(W,Z) &= \sup_A \left|P(W\in A) - P(Z\in A)\right| \\
&= \sup_A \left|\lambda E\left[f_A(W+1)\right] - E[W f_A(W)]\right| \\
&\leq \frac{1-e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i^2.
\end{aligned}$$