STATS 413 PROBLEM SET 5

This problem set is due at **noon ET** on **Oct 28, 2021**. Please upload your solutions to Canvas in a PDF file. You are encouraged to collaborate on problem sets with your classmates, but the final write-up (including any code) **must be your own**.

1. Convergence in distribution vs convergence in probability.

(a) Let $\mathbf{x}$ and $\mathbf{y}$ be (scalar) random variables. Consider the sequence of random variables $\mathbf{z}_n \triangleq \mathbf{x} + \frac{1}{n}\mathbf{y}$. What is the limit of $(\mathbf{z}_n)$, and in what sense does $(\mathbf{z}_n)$ converge to this limit? (Does it converge in probability? Does in converge in distribution?)

Solution: (1) $\lim_{n\to\infty} P(|\mathbf{x}-\mathbf{x}| > \epsilon) = \lim_{n\to\infty} P(0 > \epsilon) = 0 \Rightarrow \mathbf{x} \xrightarrow{p} \mathbf{x}$

(2) $\lim_{n\to\infty} P(|\dfrac{1}{n}\mathbf{y} - 0| > \epsilon) = \lim_{n\to\infty} P(\dfrac{1}{n}\mathbf{y} > \epsilon) + P(\dfrac{1}{n}\mathbf{y} < -\epsilon)$

$= \lim_{n\to\infty} 1 - P(\mathbf{y} \le n\epsilon) + P(\mathbf{y} < -n\epsilon) = 1 - 1 + 0 = 0 \Rightarrow \frac{1}{n}\mathbf{y} \xrightarrow{p} 0$

(2) alternative: $\lim_{n\to\infty} P(|\dfrac{1}{n} - 0| > \epsilon) = 0 \Rightarrow \dfrac{1}{n} \xrightarrow{p} 0$ and $\mathbf{y} \xrightarrow{p} \mathbf{y}$, so by continuous mapping $\frac{1}{n}\mathbf{y} \xrightarrow{p} 0 * \mathbf{y} = 0$

Using continuous mapping with (1) and (2): $\mathbf{x} + \frac{1}{n}\mathbf{y} = \mathbf{z}_n \xrightarrow{p} \mathbf{x}$

(b) Let $(\mathbf{x}_n)$ and $(\mathbf{y}_n)$ be IID sequences of (scalar) random variables. Consider the sequence of random variables $\mathbf{z}_n \triangleq \mathbf{x}_n + \frac{1}{n}\mathbf{y}_n$. What is the limit of $(\mathbf{z}_n)$, and in what sense does $(\mathbf{z}_n)$ converge to this limit?

Solution: Given $\mathbf{x}_n$ is IID: $\mathbf{x}_n$ has $F_n(x) = F(x)$. Thus, $\lim_{n\to\infty} F_n(x) = F(x) \Rightarrow \mathbf{x}_n \xrightarrow{d} \mathbf{x}$.

Take the case where $\mathbf{x}_n \overset{\text{i.i.d}}{\sim} Ber(\frac{1}{2})$ and $\mathbf{x} \sim Ber(\frac{1}{2})$, then $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$. However, $\mathbf{x}_n \overset{p}{\nrightarrow} \mathbf{x}$. $P(|\mathbf{x}_n - \mathbf{x}| > \epsilon) = \frac{1}{2}$, because $\mathbf{x}_n, \mathbf{x} \in \{0,1\}$, $p = \frac{1}{2}$. Thus $P(\mathbf{x}_n = 1, \mathbf{x} = 0) + P(\mathbf{x}_n = 0, \mathbf{x} = 1)$ $= \frac{1}{2}$.

Similarly to (2) above, we can see $\frac{1}{n}\mathbf{y}_n \xrightarrow{p} 0$. This implies $\frac{1}{n}\mathbf{y}_n \xrightarrow{d} 0$.

Using continuous mapping again: $\mathbf{x} + \frac{1}{n}\mathbf{y} = \mathbf{z}_n \xrightarrow{d} \mathbf{x}$

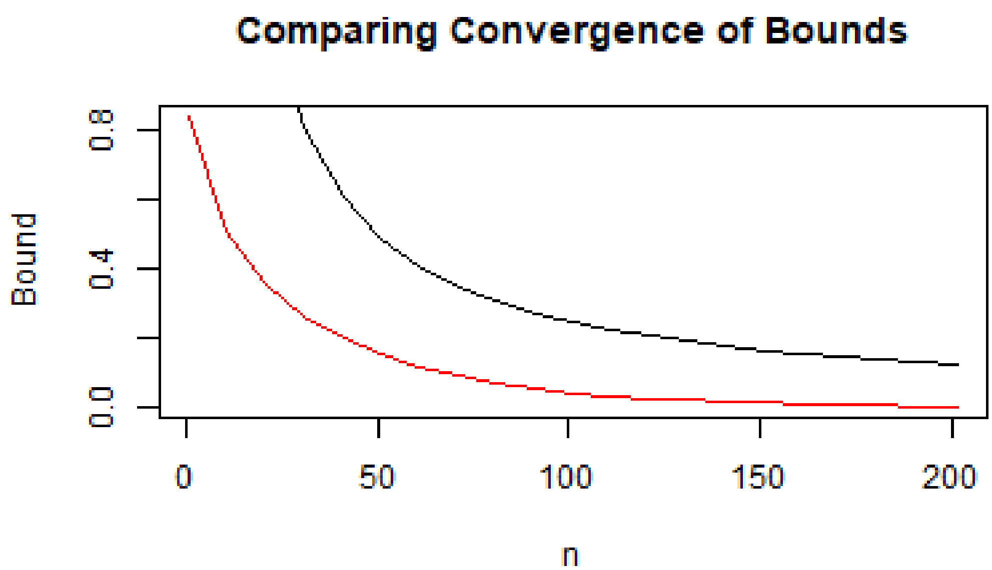
2. Chebyshev's inequality vs central limit theorem. Consider a sequence of n independent fair coin tosses ($p = \frac{1}{2}$) $\mathbf{x}_1, \dots, \mathbf{x}_n \in \{0,1\}$ ($\mathbf{x}_i = 1$ if the i-th toss comes up heads). Intuitively, the fraction of heads ($\bar{\mathbf{x}}_n \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$) is tightly concentrated around $\frac{1}{2}$. In this problem, we consider two ways of approximating $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$.

(a) Use Chebyshev's inequality to obtain a (upper) bound on $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$.
Hint: Recall Chebyshev's inequality is $\mathbf{P}(|\mathbf{x} - \mu| > \epsilon) \le \frac{\mathsf{var}[\mathbf{x}]}{\epsilon^2}$.

Solution: $\sum \mathbf{x}_i \sim Bin(n, \frac{1}{2})$, by Chebyshev's inequality we get $\mathbf{P}(|\bar{\mathbf{x}} - \frac{1}{2}| > .1) \le \frac{\mathsf{var}[\bar{\mathbf{x}}]}{0.1^2}$ $= \frac{\frac{1}{n^2} n \frac{1}{4}}{0.1^2} = \frac{1}{n * 0.04}$

(b) Use the central limit theorem to approximate $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$. Compare this approximation with bound from (a).

Solution: $(\bar{\mathbf{x}}_n - \frac{1}{2})$ is approx. $N(0, \sigma^2/n)$, where $\sigma^2 = \frac{1}{2}(1-\frac{1}{2})$, so $(\bar{\mathbf{x}}_n - \frac{1}{2})$ is approx. $N(0, \frac{1}{n*4})$. Also, $\sqrt{n}\frac{\bar{\mathbf{x}}_n - \frac{1}{2}}{.5}$ is approx. $N(0,1)$. This gives us: $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1) = 2(\mathbf{P}((\bar{\mathbf{x}}_n - \frac{1}{2})) > 0.1))$
$= 2(1 - \mathbf{P}((\bar{\mathbf{x}}_n - \frac{1}{2}) \leq 0.1)) = 2(1 - \mathbf{P}(\sqrt{n}\frac{\bar{\mathbf{x}}_n - \frac{1}{2}}{.5} \leq \sqrt{n}\frac{0.1}{.5})) \approx 2(1 - \Phi(\frac{\sqrt{n}}{5}))$
Below we can see a graph where the red line is the CLT bound and the black is the Chebyshev's inequality bound. We can see the CLT converges at a faster rate.



(c) **(1 extra pt)** The main benefit of using Chebyshev's inequality is the resulting bound is valid *non-asymptotically*; *i.e.* you know that $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$ is at most something, and this bound is valid for any n. The central limit theorem provides a good approximation of $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$ that becomes more accurate as n grows, but you don't know whether the approximation is an over or under-estimate of $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$. Is it possible to get the best of both worlds; *i.e.* is it possible to obtaining a tighter bound on $\mathbf{P}(|\bar{\mathbf{x}}_n - \frac{1}{2}| > 0.1)$ that is non-asymptotically valid?
Solution: Look into Hoeffding's inequality