ECMA 31000: Solutions to Problem Set 3

Joe Hardwick

Question 1 Show that if X_n is a sequence of random variables such that for some $r > 0$ and some random variable Y,

$$n^r (X_n - X) \xrightarrow{d} Y,$$

then $X_n \xrightarrow{p} X$.

ANS: Note that since $n^r (X_n - X) \xrightarrow{d} Y$ and $n^{-r} \xrightarrow{p} 0$, by Slutsky's Theorem, $n^{-r} n^r (X_n - X) \xrightarrow{d} 0 \cdot Y = 0$. Therefore $X_n - X \xrightarrow{d} 0$, which implies $X_n - X \xrightarrow{p} 0$ because 0 is a constant limit, so the result follows.

Question 2 a) We stated in class that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does not imply

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Prove that, if $\{X_n\}_{n \geq 1}$ and $\{Y_n\}_{n \geq 1}$ are independent sequences with $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, and X, Y are independent random variables, then in fact

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

ANS: Let $F_{X_n}, F_{Y_n}, F_X, F_Y$ denote the distribution functions of X_n, Y_n, X, Y respectively, while $F_{X_n Y_n}$ and F_{XY} denote the distribution functions of $\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ and $\begin{pmatrix} X \\ Y \end{pmatrix}$, respectively. Note that

$$\begin{aligned} F_{X_n}(x) &\to F_X(x), \\ F_{Y_n}(y) &\to F_Y(y), \end{aligned}$$

for all points (x, y) such that x is a continuity point of F_X and y is a continuity point of F_Y. At such points, by independence of X_n, Y_n,

$$\begin{aligned} F_{X_n Y_n}(x, y) &= F_{X_n}(x) F_{Y_n}(y) \\ &\to F_X(x) F_Y(y) \\ &= F_{XY}(x, y). \end{aligned}$$

It follows by definition of convergence in distribution that

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

b) Use this result to show that if $\{X_i\}_{i\geq 1}$ and $\{Y_i\}_{i\geq 1}$ are iid sequences that are independent of each other, $\mathrm{E}(X_i) = \mu_X$, $\mathrm{E}(Y_i) = \mu_Y$, and $Var(X_i) = Var(Y_i) = 1$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i - \mu_X \\ Y_i - \mu_Y \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

Find constants $r \geq 0$ and c such that

$$n^r \left(\frac{\bar{X}_n}{\bar{Y}_n} - c \right)$$

has a non-degenerate limiting distribution in the following two cases: (i) $\mu_X \in \mathbb{R}$, $\mu_Y \neq 0$, and (ii) $(\mu_X, \mu_Y) = (0, 0)$.

ANS: Let $X_n := \sqrt{n}\left(\bar{X}_n - \mu_X\right)$, $Y_n := \sqrt{n}\left(\bar{Y}_n - \mu_Y\right)$. Then X_n is independent of Y_n, so since $X_n \xrightarrow{d} \mathcal{N}(0,1)$ and $Y_n \xrightarrow{d} \mathcal{N}(0,1)$,

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$$

where X and Y are independent $\mathcal{N}(0,1)$ variates. The joint pdf of X, Y is given by product $f_X f_Y$ because X is independent of Y:

$$\begin{aligned} f_X(x) f_Y(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^2} \exp\left(-\frac{(x^2+y^2)}{2}\right), \end{aligned}$$

which is the density function of a bivariate normal distribution with $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

In case (i),

$$\sqrt{n} \begin{pmatrix} \bar{X}_n - \mu_X \\ \bar{Y}_n - \mu_Y \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right),$$

so by the delta method, with $g(x, y) = x/y$, we have

$$Dg(\mu_X, \mu_Y) = \left(\frac{1}{\mu_Y}, -\frac{\mu_X}{\mu_Y^2} \right)$$

$$\sqrt{n} \left(\frac{\bar{X}_n}{\bar{Y}_n} - \frac{\mu_X}{\mu_Y} \right) \xrightarrow{d} \mathcal{N}\left(0, Dg(\mu_X, \mu_Y)\, Dg(\mu_X, \mu_Y)'\right)$$

$$\stackrel{d}{=} \mathcal{N}\left(0, \frac{1}{\mu_Y^2} + \frac{\mu_X^2}{\mu_Y^4}\right),$$

provided $\mu_Y \neq 0$. If $\mu_X = \mu_Y = 0$, then

$$\begin{pmatrix} \sqrt{n}\bar{X}_n \\ \sqrt{n}\bar{Y}_n \end{pmatrix} \stackrel{d}{\to} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

so by the continuous mapping theorem

$$\frac{\bar{X}_n}{\bar{Y}_n} \stackrel{d}{\to} \frac{\mathcal{N}(0,1)}{\mathcal{N}(0,1)},$$

where the two normal distributions on the RHS are independent. Note: This distribution is a standard Cauchy distribution with pdf

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Therefore, $r = 0$ and $c \in \mathbb{R}$ leads to a non-degenerate limit.

Question 3 a) Show that if $X_n \sim Binomial(n,p)$ then

$$\frac{X_n - np}{\sqrt{np(1-p)}} \stackrel{d}{\to} \mathcal{N}(0,1).$$

b) Show that if $X_n \sim \chi_n^2$ then

$$\frac{X_n - n}{\sqrt{2n}} \stackrel{d}{\to} \mathcal{N}(0,1).$$

Hint: Find a way to use the CLT.

ANS: a) We must assume $p \in (0,1)$, else the sequence $X_n = n$ with probability 1 or $X_n = 0$ with probability 1. X_n has the same distribution as a sum of n iid $Bernoulli(p)$ random variables. That is, X_n has the same distribution as

$$\sum_{i=1}^{n} Y_i,$$

where $Y_i \sim Bernoulli(p)$ are iid. To establish convergence in distribution, it therefore suffices to examine the limit distribution of the sequence

$$\frac{\sum_{i=1}^n Y_i - np}{\sqrt{np(1-p)}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(Y_i - p)}{\sqrt{p(1-p)}}.$$

Since $\mathrm{E}(Y_i) = p$ and $Var(Y_i) = p(1-p)$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - p) \stackrel{d}{\to} \mathcal{N}(0, p(1-p))$$

by the CLT. Next, by the CMT with $g(x) = \frac{x}{\sqrt{p(1-p)}}$, we obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(Y_i - p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1).$$

For part b, note that X_n has the same distribution as a sum of n iid variables each with distribution equal to the square of a standard normal variable. That is, X_n has the same distribution as

$$\sum_{i=1}^{n} Y_i^2,$$

where $Y_i \sim \mathcal{N}(0,1)$ are iid. In this case,

$$\frac{\sum_{i=1}^{n} Y_i^2 - n}{\sqrt{2n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(Y_i^2 - 1)}{\sqrt{2}}.$$

Since $\mathrm{E}(Y_i^2) = 1$ and $Var(Y_i^2) = 2$, we obtain

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(Y_i^2 - 1)}{\sqrt{2}} \xrightarrow{d} \mathcal{N}(0,1).$$

These problems demonstrate that convergence in distribution does not require the sequence of random variables to exist on the same probability space. The essential feature is the distribution of the random variables. So, since it is possible to represent the distribution equivalently as a sum of iid draws from another distribution, we can use this representation to apply the CLT, with which we show that the sequence of distributions converges to that of a standard normal random variable.

Question 4 Let X_n be a $(K \times 1)$ sequence of random vectors converging in distribution to a random vector X.

a) Show that for any component $X_{n,i}$ of X_n, $X_{n,i} \xrightarrow{d} X_i$, where X_i is the $i-$th component of X.

ANS: By the continuous mapping theorem, with $g_i(x_1, \ldots, x_K) = x_i$, for any $i \in \{1, \ldots, K\}$,

$$g_i(X_n) = X_{n,i} \xrightarrow{d} X_i = g(X).$$

b) Suppose

$$X_n \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

for some non-singular variance matrix Σ. Show that if A_n is a sequence of $(K \times K)$ matrices such that $A_n \xrightarrow{p} A$ and $A\Sigma A' = I_K$, where I_K is the $(K \times K)$ identity matrix, then

$$\|A_n X_n\|^2 \xrightarrow{d} \chi_K^2,$$

where χ_K^2 denotes a chi-square distribution with K degrees of freedom, and $\|\cdot\|$ denotes the euclidean norm on $\mathbb{R}^K$.

ANS: Note that, as discussed in class, $A_n \xrightarrow{p} A$ and $X_n \xrightarrow{d} \mathcal{N}(0, \Sigma)$ implies

$$\begin{pmatrix} X_n \\ A_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{N}(0, \Sigma) \\ A \end{pmatrix}$$

so by the continuous mapping theorem:

$$A_n X_n \xrightarrow{d} A\mathcal{N}(0, \Sigma),$$

where $A\mathcal{N}(0, \Sigma)$ has the same distribution as $\mathcal{N}(0, A\Sigma A')$, or $\mathcal{N}(0, I_K)$. Applying the continuous mapping theorem again yields

$$\|A_n X_n\|^2 \xrightarrow{d} \|\mathcal{N}(0, I_K)\|^2$$

A special property of the multivariate normal distribution is that if its components are uncorrelated, they are all independent of one another. Finally, the sum of squares of independent standard normal variables has the χ^2_K distribution.

Question 5 (Stochastic Order Relations) a) Prove that $X_n = o_p(1) \implies X_n = O_p(1)$.

ANS: If $|X_n| \xrightarrow{p} 0$, then for any $\epsilon > 0$ and $\delta > 0$, $\exists N_{\epsilon,\delta}$ such that for all $n \geq N_{\epsilon,\delta}$,

$$\mathrm{P}(|X_n| \leq \epsilon) \geq 1 - \delta.$$

Let $\delta > 0$ be given. Set $\epsilon = 1$, and define $N_\delta := N_{1,\delta}$. For all $k < N_\delta$, define $B_{k,\delta}$ as a constant such that

$$\mathrm{P}(|X_k| \leq B_{k,\delta}) \geq 1 - \delta.$$

Such a constant can always be found because X_k has a distribution function F_{X_k} such that

$$\lim_{t \to +\infty} F_{X_k}(t) = 1,$$
$$\lim_{t \to -\infty} F_{X_k}(t) = 0.$$

It follows that for all n, and $M_\delta = \max\left\{\epsilon, \max_{k \leq N_{\epsilon,\delta}} B_{k,\delta}\right\}$, that

$$\mathrm{P}(|X_n| \leq M_\delta) \geq 1 - \delta.$$

b) Prove that $o_p(1) + o_p(1) = o_p(1)$.

ANS: Let $X_n \xrightarrow{p} 0$ and $Y_n \xrightarrow{p} 0$. Then for any $\epsilon > 0$,

$$\begin{aligned} \mathrm{P}(|X_n + Y_n| > \epsilon) &\leq \mathrm{P}(|X_n| + |Y_n| > \epsilon) \\ &\leq \mathrm{P}(\{|X_n| > \epsilon/2\} \cup \{|Y_n| > \epsilon/2\}) \\ &\leq \mathrm{P}(|X_n| > \epsilon/2) + \mathrm{P}(|Y_n| > \epsilon/2) \\ &\to 0, \end{aligned}$$

where the first inequality follows because $|X_n| + |Y_n| \geq |X_n + Y_n|$ by the triangle inequality, the second inequality follows because

$$|X_n| + |Y_n| > \epsilon$$

implies that either $|X_n| > \epsilon/2$ or $|Y_n| > \epsilon/2$, and the third follows because for any events A, B, $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$. It follows that $X_n + Y_n = o_p(1)$.

c) Prove that $o_p(1) + O_p(1) = O_p(1)$.

ANS: Let $X_n = o_p(1)$ and $Y_n = O_p(1)$. By part a, $X_n = O_p(1)$. For any $\delta > 0$, there exist constants B_X, B_Y such that for all n,

$$\begin{aligned} \mathrm{P}(|X_n| \leq B_X) &\geq 1 - \delta/2, \\ \mathrm{P}(|Y_n| \leq B_Y) &\geq 1 - \delta/2. \end{aligned}$$

It follows that for all n:

$$\begin{aligned} \mathrm{P}(|X_n + Y_n| \leq B_X + B_Y) &\geq \mathrm{P}(|X_n| + |Y_n| \leq B_X + B_Y) \\ &\geq \mathrm{P}(\{|X_n| \leq B_X\} \cap \{|Y_n| \leq B_Y\}) \\ &= \mathrm{P}(|X_n| \leq B_X) + \mathrm{P}(|Y_n| \leq B_Y) \\ &\quad - \mathrm{P}(\{|X_n| \leq B_X\} \cup \{|Y_n| \leq B_Y\}) \\ &\geq (1 - \delta/2) + (1 - \delta/2) - 1 \\ &= 1 - \delta. \end{aligned}$$

Question 6 Let $\{X_i\}_{i \geq 1}$ be an iid sequence with finite fourth moments and $\mathrm{E}(X_i) = \mu$. Find constants a, b and $r > 0$ such that

$$n^r \begin{pmatrix} \left(\bar{X}_n - \mu\right) - a \\ \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - b \end{pmatrix}$$

has an asymptotically normal distribution. Is it always true that the limit distribution is non-degenerate?

ANS: First note that if the joint distribution has a non–degenerate limit, then each component must have a non-degenerate limit. We know from Lecture 5 that

$$n^r \left(\left(\bar{X}_n - \mu\right) - a\right)$$

has a non-degenerate limit iff $r = \frac{1}{2}$ and $a = 0$, and from Lecture 6

$$n^r \left[\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - b\right]$$

if $r = \frac{1}{2}$ and $b = \sigma^2$. To establish the joint convergence, note that

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2 \end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} X_i - \mu \\ (X_i - \mu)^2 - \sigma^2 \end{pmatrix}.$$

By the multivariate CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} X_i - \mu \\ (X_i - \mu)^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix},$$

and $\mu_3 = \mathrm{E}(X_i - \mu)^3, \mu_4 = \mathrm{E}(X_i - \mu)^4$. Note that as in class, this still may not guarantee a non-degenerate distribution, but if it does, we can use the CMT to conclude that

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2 + o_p(1) \xrightarrow{d} \mathcal{N}\left(0, \mu_4 - \sigma^4\right).$$

In general, the limit distribution will be degenerate if $\det(\Sigma) = 0$. This can happen if, as in class, $\mu_4 - \sigma^4 = 0$, but there are other examples (of binary variables) that cause this degeneracy. Note by the Cauchy-Schwarz inequality:

$$\begin{aligned} \left|Cov\left(X_i - \mu, (X_i - \mu)^2 - \sigma^2\right)\right|^2 &= \left|\mathrm{E}\left((X_i - \mu)\left[(X_i - \mu)^2 - \sigma^2\right]\right)\right|^2 \\ &\leq \mathrm{E}\left((X_i - \mu)^2\right)\mathrm{E}\left(\left((X_i - \mu)^2 - \sigma^2\right)^2\right) \end{aligned}$$

with equality iff

$$(X_i - \mu)^2 - \sigma^2 = d(X_i - \mu),$$

for some constant d. There will generally be two values of X_i which solve

$$(X_i - \mu)^2 - d(X_i - \mu) - \sigma^2 = 0.$$

which yields

$$X_i - \mu = d \pm \frac{\sqrt{d^2 + 4\sigma^2}}{2}.$$

Taking expectations gives

$$\mathrm{E}(X_i - \mu) = d + (2p - 1)\frac{\sqrt{d^2 + 4\sigma^2}}{2} = 0.$$

The case seen in class corresponds to $p = \frac{1}{2}, d = 0$.

Question 7 (Computational Question) Let $\{X_i\}_{i \geq 1}$ be an iid sequence such that $X_i \sim Bernoulli(p)$.

That is:

$$X_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

a) Show that

$$\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \xrightarrow{d} \mathcal{N}(0,1),$$

and that

$$\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}} < p < \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right) \to 1-\alpha,$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\alpha$ quantile of the normal distribution.

ANS: In Question 4 we showed that if $p \in (0,1)$, then

$$\sqrt{n}\left(\bar{X}_n - p\right) \xrightarrow{d} \mathcal{N}(0, p(1-p)).$$

By the Strong Law of Large Numbers,

$$\bar{X}_n \xrightarrow{a.s.} \mathrm{E}(X_i) = p.$$

By the continuous mapping theorem with $g(x) = \frac{1}{\sqrt{x(1-x)}}$,

$$\frac{1}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \xrightarrow{a.s.} \frac{1}{\sqrt{p(1-p)}}.$$

By Slutsky's theorem, we conclude that

$$\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \xrightarrow{d} \frac{1}{\sqrt{p(1-p)}}\mathcal{N}(0, p(1-p)) \overset{d}{=} \mathcal{N}(0,1).$$

Next note that

$$\begin{aligned} &\mathrm{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}} < p < \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right) \\ =&\mathrm{P}\left(\left|\bar{X}_n - p\right| \le z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{P}\left(\left|\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathrm{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \leq z_{1-\frac{\alpha}{2}}\right) - \mathrm{P}\left(\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} < -z_{1-\frac{\alpha}{2}}\right) \\
&\to \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(-z_{1-\frac{\alpha}{2}}\right) = 2\Phi\left(z_{1-\frac{\alpha}{2}}\right) - 1 = 1-\alpha,
\end{aligned}
$$

by the CLT and the definition of convergence in distribution.

b) In problem set 2, we constructed a confidence interval of width 2ϵ which contained p with probability at least $1 - \frac{1}{4n\epsilon^2}$. We now consider the confidence interval

$$
\left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}\right]
$$

based on the normal approximation in part a). How large must n be to ensure the width of this confidence interval is at most 0.2, when $1 - \alpha = 0.95$?

ANS: The width of the confidence interval is

$$
2z_{0.975}\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}} \leq 2z_{0.975}\sqrt{\frac{1}{4n}} = \frac{z_{0.975}}{\sqrt{n}}.
$$

This is at most 0.2 if

$$
n \geq \left(\frac{z_{0.975}}{0.2}\right)^2 \approx 96.0365,
$$

so we need a sample size of at least $n = 97$ to guarantee a with of no greater than 0.2.

c) Simulate n iid draws from this distribution with $p = 0.4$, for each of $n = 25, 50, 100$. Compute the confidence intervals for each n based on your simulated data. Does the true value of p lie inside the confidence interval? Repeat this exercise 1000 times for each value of n, (though you don't need to display the results of each replication). For each value n, report the proportion of your replications for which the true value of p lies in your confidence interval. Does this vary much with n? Why? What is the advantage of using this confidence interval instead of one where the width is selected based on Chebyshev's inequality?

ANS: While it is possible to compute the confidence intervals as stated in part b, and check if they contain p, (see code) we may also note that for any k,

$$
\left|\bar{X}_n - p\right| \leq k\sqrt{\frac{\bar{X}_n\left(1-\bar{X}_n\right)}{n}}
$$

$$\iff \bar{X}_n \in \left[\frac{2np + k^2}{2k^2 + 2n} \pm \frac{\sqrt{(2np + k^2)^2 - 4n\left(k^2 + n\right)p^2}}{2k^2 + 2n}\right]$$
$$\iff \bar{X}_n \in \left[\frac{2np + k^2}{2k^2 + 2n} \pm \frac{\sqrt{k^4 + 4nk^2p\left(1 - p\right)}}{2k^2 + 2n}\right]$$

where the inclusion follows by squaring both sides of the inequality and solving the resulting quadratic, and the last expression is a simplification. This equivalent formulation produces a fixed interval for each n that does not vary with the value of $\bar{X}_n$. We then check if $\bar{X}_n$ lies in this interval for each simulation replication.

Note that when $p = 0.4$, the confidence interval based on the normal approximation has coverage probability *reasonably* close to 0.95 for all sample sizes $n \geq 25$. Therefore, the coverage probability remains fairly close to 0.95 in the simulations. The exact coverage probability is indicated in the following plot:

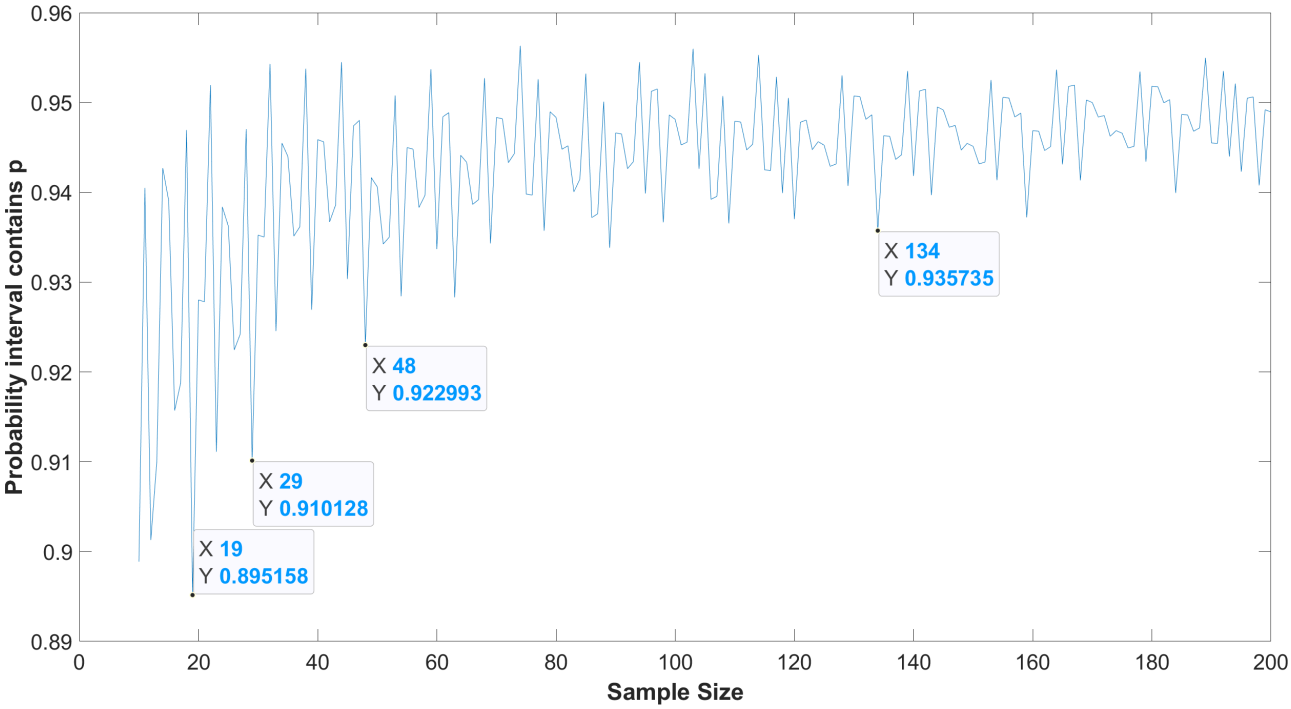

Figure 1: Exact Coverage Probability as a function of sample size, $p = 0.4$

Notice, however, that the coverage probability is erratic. While it is true that as $n \to \infty$, the probability converges to $1 - \alpha$, the actual coverage probability is non-monotonic in the sample size, with sharp drops and increases. The advantage of using this confidence interval, despite its erratic behaviour, is that the interval is narrower while still retaining coverage probability reasonably close to 0.95, the desired probability. An interval based on Chebyshev's inequality would require

$$1 - \frac{1}{4n\epsilon^2} \geq 0.95 \iff \epsilon \geq \sqrt{\frac{5}{n}},$$

which gives an overall width of

$$2\sqrt{\frac{5}{n}} \approx \frac{4.472}{\sqrt{n}},$$

whereas in this problem set the interval is never wider than

$$2z_{0.975}\sqrt{\frac{1}{4n}} = \frac{z_{0.975}}{\sqrt{n}} \approx \frac{1.960}{\sqrt{n}}.$$

We saw that the interval based on Chebyshev's inequality is wider than necessary – the true coverage probabilities were (much) higher than those promised by the bound.

d) In this part we will show that asymptotic confidence intervals do not guarantee finite sample coverage probabilities. Repeat part c) but with $p = 0.99$, and for sample sizes $n = 25, 50, 100, 250, 500, 1000, 2000$. Does the asymptotic 95% confidence interval contain the true parameter with approximately the right probability? Why is this happening? Interpret your simulations in light of the result derived in part a).

ANS: The plot of actual coverage probability for $p = 0.99$ is given below.

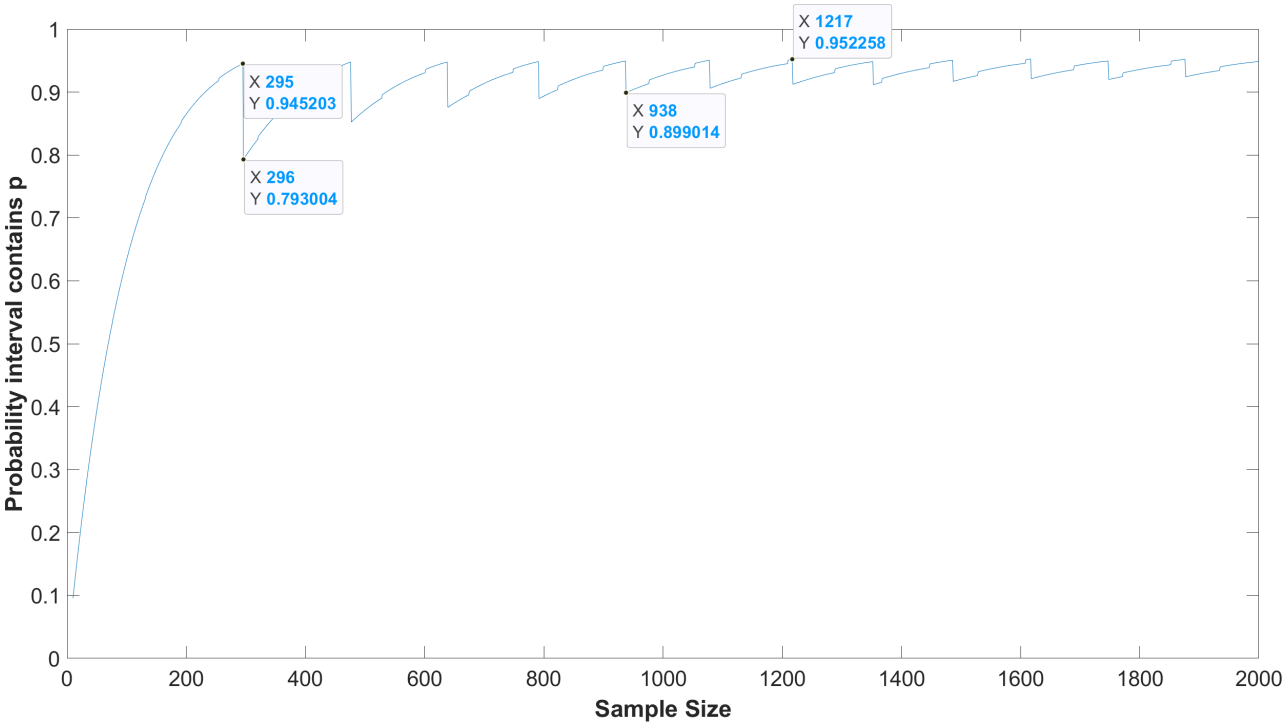

Figure 2: Exact Coverage Probability as a function of sample size, $p = 0.99$

From the plot it is clear that the true coverage probability is far from 0.95 for any $n \leq 150$. For this reason, the true coverage probability is low for $n = 25, 50, 100$. Unless the sample size is above $n \approx 1000$, we can't ensure the coverage probability will exceed 0.90, which is poor for such a large sample size. In the simulations, we notice a non-monotonicity: A coverage probability of 0.871 when $n = 500$ and 0.915 when $n = 250$. It appears from the plot that there are 'lucky' and

'unlucky' values of n. (This has led several authors to modify this natural but erratic asymptotic confidence interval to produce more reliable performance). The main takeaway from this part of the question is that the asymptotic conclusion of part a, while true, does not guarantee good coverage probability even when n is "large". This phenomenon is similar to what we observed in class: Highly skewed distributions do not necessarily produce sample averages that look normal, even when the number of observations used to compute the average is large. A more straightforward observation is that when $p = 0.99$ and $n = 100$, the probability of an entire sample of $X_i = 1$ for $1 \leq i \leq n$ is $(0.99)^{100} \approx 0.366$. For such a sample, the confidence interval is a singleton, $\{1\}$! Code and simulation results are below.

Question 7, Parts C,D

Code for Question 7d: $p = 0.99$

Below are four equivalent solutions to the simulation portion of Problem 7d using 10,000 replications. The first solution uses the fact that the sum of iid bernoullis is binomial. This is by far the fastest to execute, but we won't always be able to take such a shortcut. The next two solutions rely on vectorization (generate all the simulation draws at the start and put them in a matrix), while the last uses a loop, as in the solution to Problem Set 2. The loop takes the longest to run.

```
library(Matrix)
#set parameters
reps <- 10000
sampsizes <- c(25,50,100,250,500,1000,2000)
p = 0.99
alpha = 0.05
cv = (qnorm(1-alpha/2))
cv2 = cv^2

set.seed(476) #for reproducibility
```

Solution 1

```
probs = rep(0,length(sampsizes))
for (n in sampsizes) {
  draws <- rbinom(reps, n,p) # sum of n iid Bernoulli variables is Binomial(n,p)
  L <- n*(2*n*p + cv2 - sqrt(cv2^2 +4*n*cv2*(p-p^2)))/(2*(n+cv2)) # lower bound
  U <- n*(2*n*p + cv2 + sqrt(cv2^2 +4*n*cv2*(p-p^2)))/(2*(n+cv2)) #upper bound
  F <- (draws >= L & draws <= U) #check if binomial sims are between bounds
  probs[match(n,sampsizes)] <- nnzero(F) / reps
}

for (j in sampsizes) {
  cat("Proportion of times p lies in the CI (n = ",j,"): ",
      probs[match(j,sampsizes)],"\n", sep="")
}
```

```
## Proportion of times p lies in the CI (n = 25): 0.2286
## Proportion of times p lies in the CI (n = 50): 0.3928
## Proportion of times p lies in the CI (n = 100): 0.6325
## Proportion of times p lies in the CI (n = 250): 0.9113
## Proportion of times p lies in the CI (n = 500): 0.8725
## Proportion of times p lies in the CI (n = 1000): 0.9244
## Proportion of times p lies in the CI (n = 2000): 0.9481
```

Solution 2

```
probs = rep(0,length(sampsizes)) # to fill.

for (n in sampsizes) {

  draws <- matrix(rbinom(n*reps, 1,p), ncol = n) #draw bernoullis, arrange in matrix
  means <- rowMeans(draws) #compute sample means
  L <- (2*n*p + cv2 - sqrt(cv2^2 +4*n*cv2*(p-p^2)))/(2*(n+cv2))
  U <- (2*n*p + cv2 + sqrt(cv2^2 +4*n*cv2*(p-p^2)))/(2*(n+cv2))
  F <- (means >= L & means <= U) #check if sample means between bounds
  probs[match(n,sampsizes)] <- nnzero(F) / reps
}
for (j in sampsizes) {
  cat("Proportion of times p lies in the CI (n = ",j,"): ",
      probs[match(j,sampsizes)],"\n", sep="")
}
```

```
## Proportion of times p lies in the CI (n = 25): 0.2172
## Proportion of times p lies in the CI (n = 50): 0.386
## Proportion of times p lies in the CI (n = 100): 0.6239
## Proportion of times p lies in the CI (n = 250): 0.9133
## Proportion of times p lies in the CI (n = 500): 0.8711
## Proportion of times p lies in the CI (n = 1000): 0.9285
## Proportion of times p lies in the CI (n = 2000): 0.9502
```

Solution 3

```
probs = rep(0,length(sampsizes))

for (n in sampsizes) {

  draws <- matrix(rbinom(n*reps, 1,p), ncol = n)
  means <- rowMeans(draws)
  L <- means - cv*sqrt(means*(1-means)/n) # vector of confidence interval lower bounds
  U <- means + cv*sqrt(means*(1-means)/n) # vector of confidence interval upper bounds
  F <- (p >= L & p <= U) #check if p inside confidence interval.
  probs[match(n,sampsizes)] <- nnzero(F) / reps #proportion of intervals containing p
}
for (j in sampsizes) {
  cat("Proportion of times p lies in the CI (n = ",j,"): ",
      probs[match(j,sampsizes)],"\n", sep="")
}
```

```
## Proportion of times p lies in the CI (n = 25): 0.2259
## Proportion of times p lies in the CI (n = 50): 0.4025
## Proportion of times p lies in the CI (n = 100): 0.6401
## Proportion of times p lies in the CI (n = 250): 0.9138
## Proportion of times p lies in the CI (n = 500): 0.8693
## Proportion of times p lies in the CI (n = 1000): 0.9285
## Proportion of times p lies in the CI (n = 2000): 0.9478
```

Solution 4

```
boolean_array = array(NA,c(reps,length(sampsizes))) # to fill
for (i in 1:reps){
  j <- 1
  for (n in sampsizes) {
    draws <- rbinom(n, 1, p)
    mean <- mean(draws)
    ci <- c(mean - cv*sqrt(mean*(1-mean)/n), mean + cv*sqrt(mean*(1-mean)/n))
    boolean_array[i, j] = (p >= ci[1] & p <= ci[2])
    j <- j + 1
  }
}

proportions <- colMeans(boolean_array)

for (j in sampsizes) {
cat("Proportion of times p lies in the CI (n = ",j,"): ",
    proportions[match(j,sampsizes)],"\n", sep="")
}
```

```
## Proportion of times p lies in the CI (n = 25): 0.2251
## Proportion of times p lies in the CI (n = 50): 0.3903
## Proportion of times p lies in the CI (n = 100): 0.6354
## Proportion of times p lies in the CI (n = 250): 0.9197
## Proportion of times p lies in the CI (n = 500): 0.8719
## Proportion of times p lies in the CI (n = 1000): 0.9308
## Proportion of times p lies in the CI (n = 2000): 0.9495
```