Statistics 581, Problem Set 3 Solutions

Wellner; 10/23/2002

1. A. Ferguson, ACILST, page 24, problem 4. One strategy for evaluating the integral

$$I = \int_1^\infty \frac{1}{x}\sin(2\pi x)dx = .153...$$

by Monte Carlo approximation is as follows. Write the integral, by a change of variable $y = 1/x$, as

$$I = \int_0^1 \frac{1}{y}\sin\left(\frac{2\pi}{y}\right)dy$$

and approximate I by

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i}\sin\left(\frac{2\pi}{Y_i}\right)$$

where $Y_1, \ldots, Y_n$ is a sample from the uniform distribution on $[0,1]$. How well does this approximation work? Does $\hat{I}_n$ converge to I almost surely?

B. Suppose that I in A is generalized to

$$I_\alpha \equiv \int_1^\infty \frac{1}{x^\alpha}\sin(2\pi x)dx$$

for $\alpha > 0$ (so that the integral I in part A is I_1). Construct the corresponding Monte-Carlo estimator $\hat{I}_{n,\alpha}$ of I_α. For what values of α will the estimator $\hat{I}_{n,\alpha}$ converge to I_α? (Use the same change of variables as in A.)

C. For what values of α will we have

$$\sqrt{n}(\hat{I}_{n,\alpha} - I_\alpha) \to_d \text{something}\,?$$

For those values of α for which this holds, find "something".

Solution: A. By the chance of variable $x = 1/y$ it follows that

$$I = \int_1^\infty \frac{1}{x}\sin(2\pi x)dx = \int_0^1 \frac{1}{y}\sin(2\pi/y)dy\,.$$

Thus a natural estimator is

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{Y_i} \sin\left(\frac{2\pi}{Y_i}\right)$$

where $Y_1, \ldots, Y_n$ is a sample from the uniform distribution on $[0, 1]$. Unfortunately, however,

$$\begin{aligned}
E|\frac{1}{Y_1}\sin(2\pi/Y_1)| &= E\left\{\frac{1}{Y_1}|\sin(2\pi/Y_1)|\right\} = \int_0^1 \frac{1}{y}|\sin(2\pi/y)|dy \\
&= \int_1^\infty \frac{1}{x}|\sin(2\pi x)|dx = \sum_{k=3}^\infty \int_{(k-1)/2}^{k/2} \frac{1}{x}|\sin(2\pi x)|dx \\
&\geq \sum_{k=3}^\infty \frac{1}{k/2}\int_{(k-1)/2}^{k/2} |\sin(2\pi x)|dx \\
&= \sum_{k=3}^\infty \frac{2}{k}\frac{1}{\pi} = \frac{2}{\pi}\sum_{k=3}^\infty \frac{1}{k} = \infty\,.
\end{aligned}$$

Hence the strong law of large numbers fails, and it is not at all clear that $\hat{I}_n$ will be a reasonable estimator of I. Here are plots of the functions being integrated on $[1, \infty)$ and on $(0, 1]$:

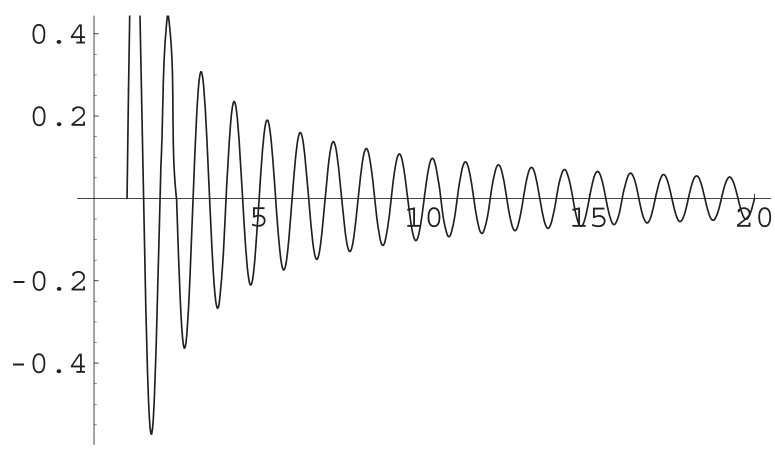

Figure 1: Plot of $x^{-1}\sin(2\pi x)$ on $[1, \infty)$

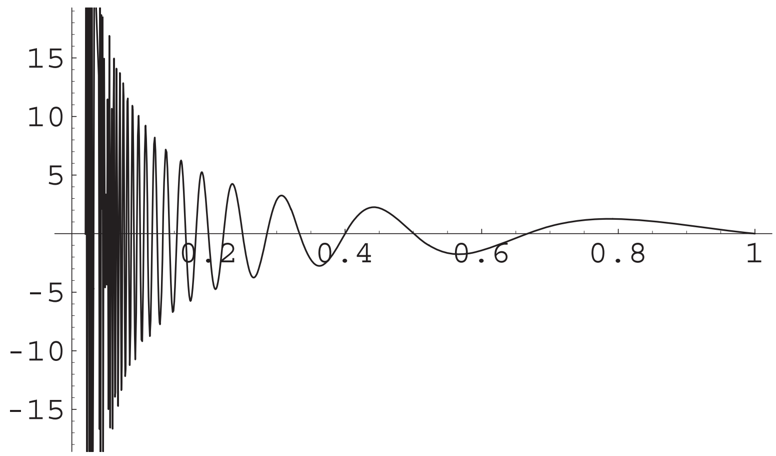

Figure 2: Plot of $y^{-1}\sin(2\pi/y)$ on $(0,1]$

B. For I_α we can write, by the change of variables $x = 1/y$,

$$I_\alpha \equiv \int_1^\infty \frac{1}{x^\alpha}\sin(2\pi x)dx = \int_0^1 y^{\alpha-2}\sin(2\pi/y)dy\,.$$

In this case

$$E|Y^{\alpha-2}\sin(2\pi/Y)| \leq EY^{\alpha-2} = \int_0^1 y^{\alpha-2}dy = \frac{1}{\alpha-1}y^{\alpha-1}\Big|_0^1 < \infty$$

if $\alpha - 1 > 0$ or $\alpha > 1$. Thus by the strong law of large numbers it follows that

$$\hat{I}_{n,\alpha} \equiv \frac{1}{n}\sum_{i=1}^n Y_i^{\alpha-2}\sin(2\pi/Y_i) \rightarrow_{a.s.} E(Y^{\alpha-2}\sin(2\pi/Y)) = I_\alpha$$

if $\alpha > 1$. See figure xx for a plot of I_α for $.15 \leq \alpha \leq 5$

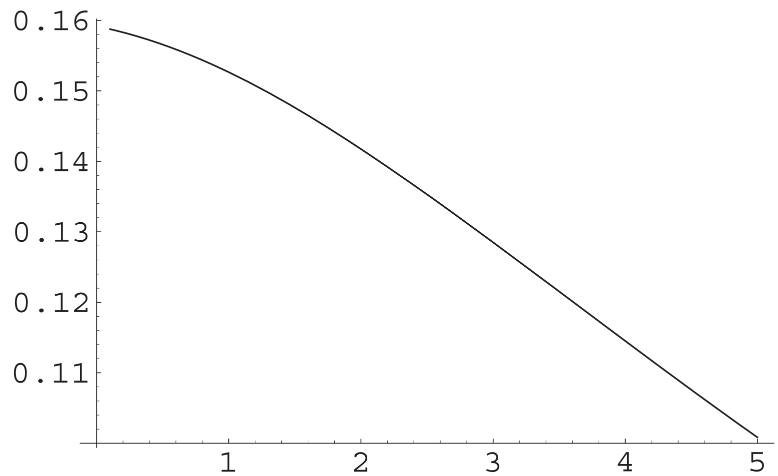

Figure 3: Plot of I_α

C. Note that

$$E|Y^{\alpha-2}\sin(2\pi/Y)|^2 \ \leq \ EY^{2(\alpha-2)} = \int_0^1 y^{2\alpha-4}dy = \frac{1}{2\alpha-3}y^{2\alpha-3}\Big|_0^1 < \infty$$

if $2\alpha - 3 > 0$, and hence if $\alpha > 3/2$. Hence if $\alpha > 3/2$, it follows from the CLT that

$$\sqrt{n}(\hat{I}_{n,\alpha} - I_\alpha) \rightarrow_d N(0, \sigma^2)$$

where

$$\sigma^2 \equiv \sigma^2(\alpha) = E\left(Y^{2(\alpha-2)}(\sin(2\pi/Y)^2\right) - I_\alpha^2 .$$

Here is a plot of $\sigma^2(\alpha)$ as a function of α for $\alpha \in [1.6, 5]$.

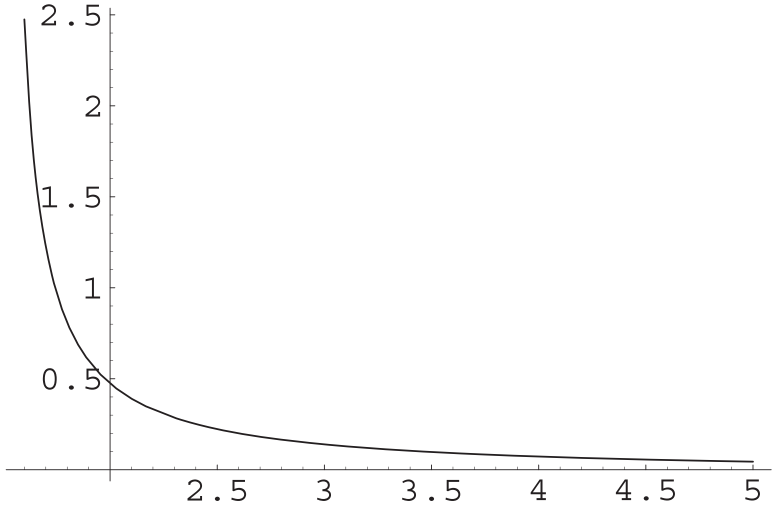

Figure 4: Plot of σ_α^2

2. Ferguson, ACILST, page 34, problem 1 (modified slightly)
A. Suppose that $X_1, X_2, \ldots$ are i.i.d. in R^2 with distribution giving probability θ_1 to $(1,0)'$, probability θ_2 to $(0,1)'$, θ_3 to $(0,0)'$ and θ_4 to $(-1,-1)$ where $\theta_j \geq 0$ for $j = 1,2,3,4$ and $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$. Find the limiting distribution of $\sqrt{n}(\overline{X}_n - E(X_1))$ and describe the resulting approximation to the distributon of $\overline{X}_n$.
B. Suppose that $X_1, \ldots, X_n$ is a sample from the Poisson distribution with parameter $\lambda > 0$: $P(X_1 = k) = \exp(-\lambda)\lambda^k/k!$, $k = 0, 1, \ldots$. Let $Z_n = (1/n)\sum_{i=1}^n 1_{[X_i=1]}$. What is the joint asymptotic distribution of

$$\sqrt{n}((\overline{X}_n, Z_n)' - (\lambda, \lambda e^{-\lambda})')?$$

C. Let $p_1(\lambda) \equiv P_\lambda(X_1 = 1)$. What is the asymptotic distribution of $\hat{p}_1 \equiv p_1(\hat{\lambda}_n)$ where $\hat{\lambda}_n = \overline{X}$?
D. What is the joint asymptotic distribution of $(Z_n, \hat{p}_1)$ (after centering and rescaling)?

Solution: A. Now

$$E(X_1) = \theta_1(1,0)' + \theta_2(0,1)' + \theta_3(0,0)' + \theta_4(-1,-1)' = (\theta_1 - \theta_4, \theta_2 - \theta_4),$$

while

$$
\begin{aligned}
E(XX') &= \theta_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \theta_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \theta_3 \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \theta_4 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \theta_1 + \theta_4 & \theta_4 \\ \theta_4 & \theta_2 + \theta_4 \end{pmatrix} .
\end{aligned}
$$

Hence it follows that

$$
\begin{aligned}
\Sigma &= E(XX') - E(X)E(X') \\
&= \begin{pmatrix} \theta_1 + \theta_4 & \theta_4 \\ \theta_4 & \theta_2 + \theta_4 \end{pmatrix} - \begin{pmatrix} (\theta_1 - \theta_4)^2 & (\theta_1 - \theta_4)(\theta_2 - \theta_4) \\ (\theta_1 - \theta_4)(\theta_2 - \theta_4) & (\theta_2 - \theta_4)^2 \end{pmatrix} \\
&= \begin{pmatrix} \theta_1 + \theta_4 - (\theta_1 - \theta_4)^2 & \theta_4 - (\theta_1 - \theta_4)(\theta_2 - \theta_4) \\ \theta_4 - (\theta_1 - \theta_4)(\theta_2 - \theta_4) & \theta_2 + \theta_4 - (\theta_2 - \theta_4)^2 \end{pmatrix} .
\end{aligned}
$$

By the multivariate CLT it follows that

$$
\sqrt{n}(\overline{X}_n - E(X_1)) \rightarrow_d N_2(0, \Sigma) .
$$

For example, if $\theta_j = 1/4$ for $j = 1, 2, 3, 4$, then $E(X_1) = 0$,

$$
\Sigma = \begin{pmatrix} 1/2 & 1/4 \\ 1/4 & 1/2 \end{pmatrix}
$$

so the variances are both $1/2$ and the correlation is $(1/4)/\sqrt{(1/2)(1/2)} = 1/2$. In this case the resulting normal approximation to the distribution of $\overline{X}_n$ is centered at 0 with a variance-covariance matrix $n^{-1}\Sigma$ with Σ as in the last display.

B. Let $W_i \equiv (X_i, Y_i) \equiv (X_i, 1_{[X_i=1]})$. Then the W_i's are i.i.d. with mean $E(W_1) = (\lambda, \lambda e^{-\lambda})'$ and covariance matrix

$$
\Sigma = \begin{pmatrix} \lambda & \lambda e^{-\lambda} - \lambda^2 e^{-\lambda} \\ \lambda e^{-\lambda} - \lambda^2 e^{-\lambda} & \lambda e^{-\lambda}(1 - \lambda e^{-\lambda} \end{pmatrix} . \tag{1}
$$

Hence the multivariate CLT implies that

$$
\sqrt{n}(\overline{W} - E(W_1)) = \sqrt{n}((\overline{X}_n, Z_n)' - (\lambda, \lambda e^{-\lambda})) \rightarrow_d T \sim N_2(0, \Sigma) \tag{2}
$$

where Σ is given in (1).

C. Now $\hat{p}_1 = g(\overline{X}_n)$ where $g(v) = ve^{-v}$. Hence $g'(v) = (1-v)e^{-v}$, $g'(\lambda) = (1-\lambda)e^{-\lambda}$, and $\sqrt{n}(\overline{X}_n - \lambda) \rightarrow_d N(0,\lambda)$ by the CLT (or the first component of the convergence in distribution in part B). Hence it follows from the delta-method that

$$\sqrt{n}(\hat{p}_1 - p_1(\lambda)) = \sqrt{n}(g(\overline{X}_n) - g(\lambda)) \rightarrow_d g'(\lambda)N(0,\lambda) = N(0, \lambda(1-\lambda)^2 e^{-2\lambda}) .$$

D. At this point it is a bit easier to study $(\hat{p}_1, Z_n) = g(\overline{X}_n, Z_n)$ where $g(u,v) \equiv (ue^{-u}, v)$. Then in view of (2) and

$$\nabla g(\lambda, \lambda e^{-\lambda}) = \begin{pmatrix} (1-\lambda)e^{-\lambda} & 0 \\ 0 & 1 \end{pmatrix} ,$$

it follows from the delta-method that

$$\sqrt{n}((\hat{p}_1, Z_n)' - \lambda e^{-\lambda}(1,1)') \rightarrow_d \nabla g(\lambda, \lambda e^{-\lambda}) T \sim N_2(0, \nabla g \Sigma (\nabla g)')$$

where

$$\nabla g \Sigma (\nabla g)' = \begin{pmatrix} \lambda(1-\lambda)^2 e^{-2\lambda} & \lambda(1-\lambda)^2 e^{-2\lambda} \\ \lambda(1-\lambda)^2 e^{-2\lambda} & \lambda e^{-\lambda}(1-\lambda e^{-\lambda}) \end{pmatrix} .$$

This is a situation in which we have two estimators of $P_\lambda(X_1 = 1) = p_1(\lambda)$, namely the MLE $\hat{p}_1 = p_1(\hat{\lambda})$ and the empirical (or "plug-in" estimator $Z_n = \#\{i \le n : X_i = 1\}/n$. Note that the ratio of the asymptotic variance of $\hat{p}_1$ to the asymptotic variance of Z_n is

$$ARE(\hat{p}_1, Z_n) \equiv \frac{\lambda(1-\lambda)^2 e^{-2\lambda}}{\lambda e^{-\lambda}(1-\lambda e^{-\lambda})} = \frac{(1-\lambda)^2 e^{-\lambda}}{(1-\lambda e^{-\lambda})} < 1$$

for all $\lambda > 0$. See the figure below

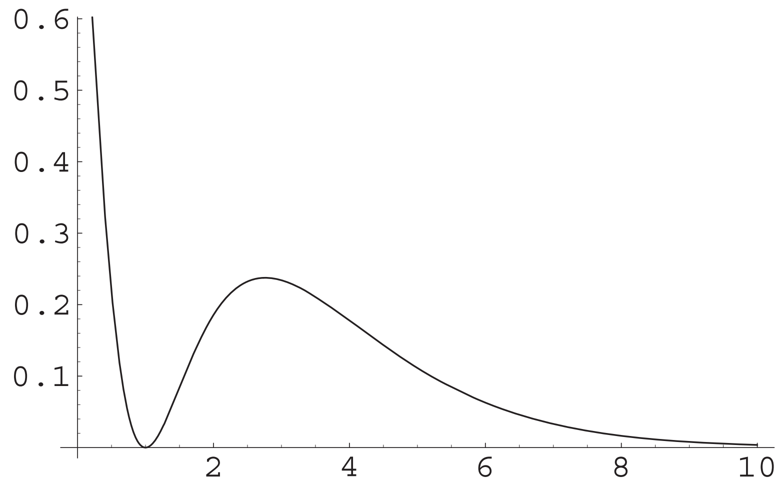

Figure 5: ARE of MLE relative to Plug-In

3. Suppose that X is a random variable with finite fourth moment; $E|X|^4 < \infty$. Then $\mu_4 = E(X-\mu)^4$ is the fourth central moment of X. The ratio $\mu_4/\sigma^4 \equiv \kappa$ is the *kurtosis* of X (or of the distribution function F of X), and $\gamma_2 \equiv \mu_4/\sigma^4 - 3$ is called the *excess of kurtosis*; note that for any $N(\mu, \sigma^2)$ random variable, $\gamma_2 = 0$. Investigate the value of γ_2 for various classical distributions (t_r, uniform, bernoulli, Poission(λ), ...). How big can γ_2 be? How small can γ_2 be?

   Solution: Note that $\mu_4^{1/4} = \{E(X-\mu)^4\}^{1/4} \ge \{E(X-\mu)^2\}^{1/2} = \sigma$ by Liapunov's inequality. Thus $\mu_4/\sigma^4 \ge 1$ always, or $\gamma_2 \equiv \mu_4/\sigma^4 \ge -2$ with equality if $X = \pm 1$ with probability $1/2$ each: then $\mu = 0$, $\sigma^2 = 1$, $\mu_4 = 1$, and $\gamma_2 = -2$.
   For $X \sim N(0,1)$, $\gamma_2 = 0$ since $EX^4 = 3$.
   For $X \sim t_r$, $r > 4$, $\gamma_2 = 6/(r-4) \nearrow \infty$ as $r \searrow 4$; $\gamma_2 \searrow 0$ as $r \nearrow \infty$.
   For $X \sim \text{Gamma}(\alpha, \beta)$, $\gamma_2 = 6/\alpha \nearrow \infty$ as $\alpha \searrow 0$.
   For $X \sim \text{Poisson}(\lambda)$, $\gamma_2 = 1/\lambda \nearrow \infty$ as $\lambda \searrow 0$.
   For $X \sim \text{Bernoulli}(p)$, $\gamma_2 = (1-p)^2/p + p^2/(1-p) - 3$ which $= -2$ when $p = 1/2$, and $\nearrow \infty$ when $p \to 0, 1$.

4. Ferguson, ACILST, page 34, problem 6. Let $Z_1, Z_2, \ldots$ be i.i.d. continuous random variables. We say a record occurs at k if $Z_k > \max_{i<k} Z_i$. Let $R_k = 1$ if a record occurs at k, and let $R_k = 0$ otherwise. Then $R_1, R_2, \ldots$ are independent Bernoulli random variables with $P(R_k = 1) = 1 - P(R_k = 0) = 1/k$, for $k = 1, 2, \ldots$. Let $S_n = \sum_{k=1}^n R_k$ denote the number of records in the first n observations. Find $E(S_n)$ and $Var(S_n)$, and show that $(S_n - E(S_n))/\sqrt{Var(S_n)} \rightarrow_d N(0,1)$.

Solution: First we compute

$$E(S_n) = \sum_{k=1}^n E(R_k) = \sum_{k=1}^n \frac{1}{k} = \log n + \gamma + o(1)\,,$$

where $\gamma = .5772157\ldots$ is Euler's constant, and

$$\begin{aligned} Var(S_n) &\equiv \sigma_n^2 \\ &= \sum_{k=1}^n Var(R_k) = \sum_{k=1}^n \frac{1}{k}(1 - 1/k) \\ &= \sum_{k=1}^n \frac{1}{k} - \sum_{k=1}^n \frac{1}{k^2} = \log n + \gamma - \pi^2/6 + o(1)\,. \end{aligned}$$

Now let $X_{nk} \equiv R_k - 1/k$. Then $E(X_{nk}) = 0$ and $Var(\sum_{k=1}^n X_{nk}) = Var(S_n) = \sigma_n^2$. To check the Lindeberg condition we compute

$$\begin{aligned} &\frac{1}{\sigma_n^2} \sum_{k=1}^n E\{X_{nk}^2 1_{[|X_{nk}| > \epsilon\sigma_n]}\} \\ &\quad = \frac{1}{\sigma_n^2} \sum_{k=1}^n E\{|R_k - 1/k|^2 1_{[|R_k - 1/k| > \epsilon\sigma_n]}\} \\ &\quad \leq \frac{1}{\sigma_n^2} \sum_{k=1}^n E\{|R_k - 1/k|^2 1_{[1 > \epsilon\sigma_n]}\} = 1_{[1 > \epsilon\sigma_n]} \rightarrow 0 \end{aligned}$$

since $|R_k - 1/k| \leq 1$ and $\sigma_n \sim (\log n)^{1/2} \rightarrow \infty$. Thus the Lindeberg condition holds and we conclude that

$$\frac{S_n - E(X_n)}{\sqrt{Var(S_n)}} \rightarrow N(0,1)\,.$$

Does the Liapunov CLT work in this case? To check this, we compute

$$\begin{aligned}\gamma_{nk} &\equiv E|R_k - 1/k|^3 = (1/k)|1-1/k|^3 + (1-1/k)|0-1/k|^3 \\ &= (1/k)\{(1-1/k)^3 + (1/k)^2(1-1/k)\} \le 2/k\,.\end{aligned}$$

Hence

$$\gamma_n \equiv \sum_{k=1}^n \gamma_{nk} \le 2\sum_{k=1}^n \frac{1}{k} \sim 2\log n \qquad \text{as} \quad n \to \infty\,.$$

Thus it follows that

$$\frac{\gamma_n}{\sigma_n^3} \le \frac{O(2\log n)}{(\log n + O(1))^{3/2}} \to 0\,,$$

and hence the same conclusion follows from the Liapunov CLT.

5. Suppose that $X_1, \ldots, X_n$ are independent $N(0,1)$ random variables, and let $Y_i = X_i^2$, for $i = 1, \ldots, n$. Thus $\sum_1^n Y_i \sim \chi_n^2$.
(a) Show that $\sqrt{n}(\overline{Y}_n - 1) \to_d N(0, \text{“}\textit{something}\text{”})$, and find “something”.
(b) Show that for each $r > 0$, $\sqrt{n}(\overline{Y}_n^r - 1) \to_d N(0, V^2(r))$ and find $V^2(r)$ as a function of r.
(c) Show that

$$\frac{\sqrt{n}(\overline{Y}_n^{1/3} - (1 - 2/(9n)))}{\sqrt{2/9}} \to_d N(0,1)\,.$$

Does this agree with your result in (b)?
(d) Make normal probablility plots to compare the approximations in (a) and (c). [The transformation in (c) is called the “Wilson-Hilferty” transformation of a χ^2 random variable.]

Solution: (a) Since the Y_i's are i.i.d. with $E(Y_i) = 1$ and $Var(Y_i) = E(X_i^4) - E(X_i^2)^2 = 3 - 1 = 2$, it follows from the CLT that

$$\sqrt{n}(\overline{Y}_n - 1) \to_d Z \sim N(0,2)\,.$$

(b) For $g(x) = x^r$ we have $g'(x) = rx^{r-1}$. Hence by the g'-theorem

$$\begin{aligned}\sqrt{n}(\overline{Y}_n^r - 1) &= \sqrt{n}(g(\overline{Y}_n) - g(1)) \\ &\to_d g'(1)Z = rN(0,2) = N(0, 2r^2)\,.\end{aligned}$$

Thus $V^2(r) = 2r^2$.
(c) When $r = 1/3$, we find from (b) that

$$\sqrt{n}(\overline{Y}_n^{1/3} - 1) \to_d (1/3)Z \sim N(0, 2/9)\,.$$

Hence it follows that

$$\begin{aligned} & \sqrt{n}(\overline{Y}_n^{1/3} - (1 - 2/(9n)) \\ = \quad & \sqrt{n}(\overline{Y}_n^{1/3} - 1) + (2/9\sqrt{n}) \\ \to_d \quad & N(0, 2/9) + 0 = N(0, 2/9)\,. \end{aligned}$$

Hence

$$\frac{\sqrt{n}(\overline{Y}_n^{1/3} - (1 - 2/(9n)))}{\sqrt{2/9}} \to_d N(0, 1)$$

in complete agreement with (b). (The added term $(2/9n)$ gives a higher order approximation to the mean.)
(d) Here are some plots showing the effect of taking the 1/3 power.

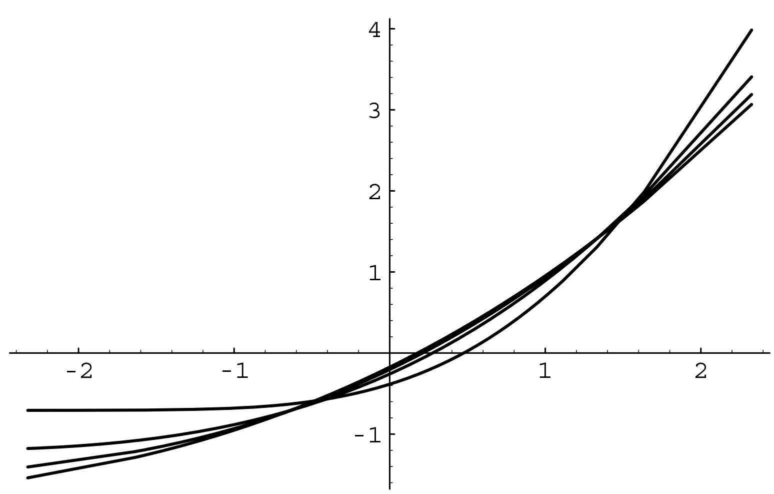

Figure 6: Q-Q plot for $\sqrt{n/2}(\overline{Y}_n - 1)$, $n = 1, 3, 5, 7$

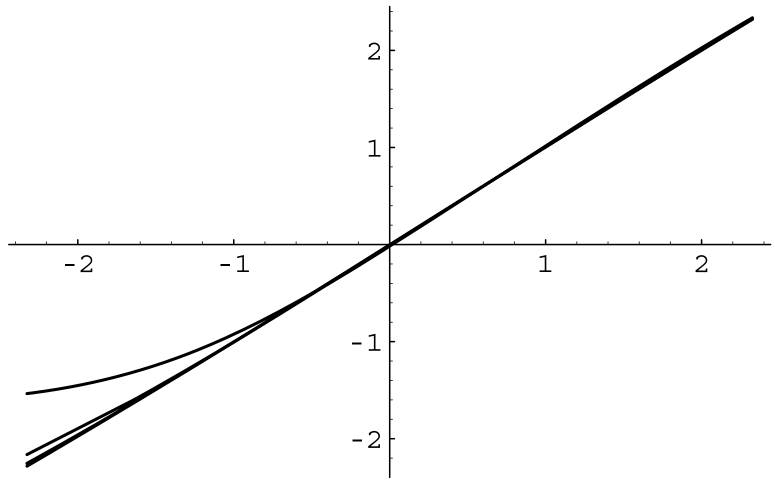

Figure 7: Q-Q plot for $\sqrt{9n/2}((\overline{Y}_n)^{1/3} - (1 - 2/(9n)))$, $n = 1, 3, 5, 7$

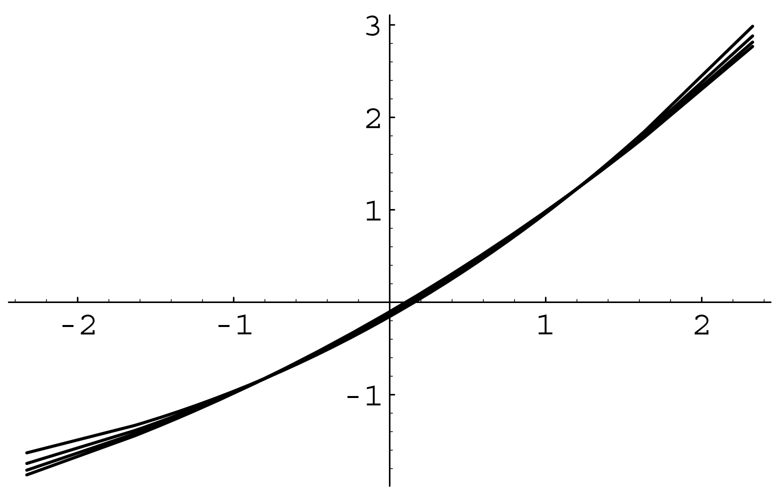

Figure 8: Q-Q plot for $\sqrt{n/2}(\overline{Y}_n - 1)$, $n = 9, 13, 17, 21$

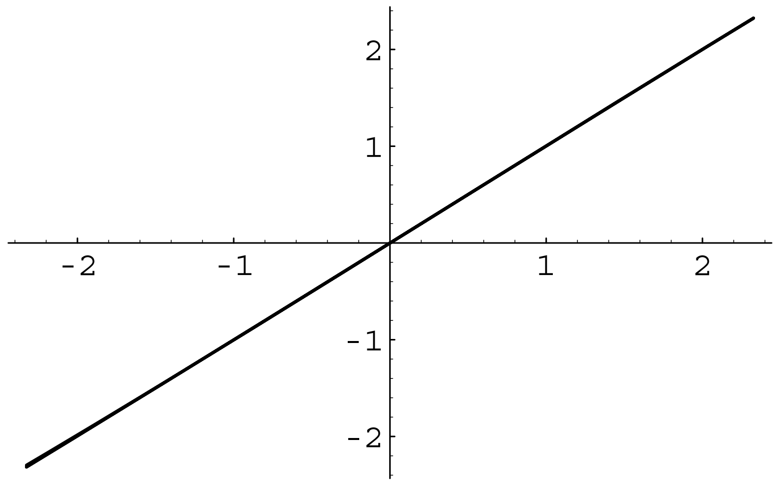

Figure 9: Q-Q plot for $\sqrt{9n/2}((\overline{Y}_n)^{1/3} - (1 - 2/9n))$, $n = 9, 13, 17, 21$