5. Sums of Random Variables and Long-Term Averages

- 5.1 Sums of Random Variables
  - $X_1, X_2, \ldots, X_n$: a sequence of random variables
  - $S_n = X_1 + X_2 + \ldots + X_n$

- Mean and Variance of S_n
  - Mean

$$E[X_1 + X_2 + \ldots + X_n] = E[X_1] + \ldots + E[X_n]$$

regardless of statistical independence

- Variance

  Step 1. Variance of $Z = X + Y$

  - ✓ $$\begin{aligned} \mathrm{VAR}[Z] &= E[(Z - E[Z])^2] \\ &= E[(X + Y - E[X] - E[Y])^2] \\ &= E[\{(X - E[X]) + (Y - E[Y])\}^2] \\ &= E[(X - E[X])^2 + (Y - E[Y])^2 + (X - E[X])(Y - E[Y]) \\ &\quad + (Y - E[Y])(X - E[X])] \\ &= \mathrm{VAR}[X] + \mathrm{VAR}[Y] + \mathrm{COV}(X, Y) + \mathrm{COV}(Y, X) \\ &= \mathrm{VAR}[X] + \mathrm{VAR}[Y] + 2\,\mathrm{COV}(X, Y) \end{aligned}$$

- Note : The variance of a sum is not necessarily equal to the sum of the individual variances.

Step 2.

$$\mathrm{VAR}(X_1+X_2+\cdots+X_n)=E\left\{\sum_{j=1}^{n}(X_j-E[X_j])\sum_{k=1}^{n}(X_k-E[X_k])\right\}$$

$$=\sum_{j=1}^{n}\sum_{k=1}^{n}E[(X_j-E[X_j])(X_k-E[X_k])]$$

$$=\sum_{k=1}^{n}\mathrm{VAR}(X_k)+\sum_{\substack{j=1\\ j\neq k}}^{n}\sum_{k=1}^{n}\mathrm{COV}(X_j,X_k)$$

- cf) If $X_1, X_2, \ldots, X_n$ are independent r.v.'s, then
  $\mathrm{COV}(X_j, X_k) = 0$ for $j \neq k$ (cov can be negative)
  and $\mathrm{VAR}(X_1 + X_2 + \ldots + X_n) = \mathrm{VAR}(X_1) + \ldots + \mathrm{VAR}(X_n)$

❖ pdf of sums of independent r.v.'s.

- $X_1, X_2, \ldots, X_n$: n independent r.v.'s
- $S_n = X_1 + X_2 + \ldots + X_n$
- pdf of S_n ?

- Step 1.
  the $n = 2$ case
  i.e., $Z = X + Y$ and X, Y are independent r.v.'s

- $\Phi_Z(\omega) = E[e^{j\omega Z}]$

$$= E[e^{j\omega(X+Y)}]$$

$$= E[e^{j\omega X} e^{j\omega Y}]$$

$$= E[e^{j\omega X}]E[e^{j\omega Y}]$$

$$= \Phi_X(\omega)\Phi_Y(\omega)$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x') f_Y(z - x') dx'$$

$= f_X(x) * f_Y(y)$: convolution of the pdf's of X and Y.

$$\Phi_Z(\omega) = \mathcal{F}\{f_Z(z)\} = \mathcal{F}\{f_X(x) * f_Y(y)\}$$

$$= \Phi_X(\omega)\Phi_Y(\omega)$$

- Step 2.

$S_n = X_1 + X_2 + \ldots + X_n$ and $X_1, X_2, \ldots, X_n$ are independent.

$$\Phi_{S_n}(\omega) = E[e^{j\omega S_n}] = E[e^{j\omega(X_1+X_2+\cdots+X_n)}]$$
$$= E[e^{j\omega X_1}]\cdots E[e^{j\omega X_n}]$$
$$= \Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega)$$

$$f_{S_n}(x) = \mathcal{F}^{-1}\{\Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega)\}$$

cf) $\mathcal{F}^{-1}\{\Phi_{S_n}(\omega)\} = \dfrac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_{S_n}(\omega)e^{-j\omega S_n}d\omega$

- Discrete r.v.'s (i.e., integer-valued r.v.'s)

  - $G_N(z) = E[z^N]$: Probability generating function.
  - $N = X_1 + X_2 + \ldots + X_n$: Sum of independent discrete r.v.'s

$$G_N(z) = E[z^{X_1+\cdots+X_n}] = E[z^{X_1}]\cdots E[z^{X_n}]$$
$$= G_{X_1}(z)\cdots G_{X_n}(z)$$

Sum of a Random Number of Random Variables

$$S_N = \sum_{k=1}^{N} X_k$$

$N \rightarrow$ random number

$X_k \rightarrow$ iid r.v.'s

- N : a random variable independent of the X_k's
- X_k : iid r.v.'s

- The mean of S_N

$$\begin{aligned} E[S_N] &= E[E[S_N \mid N]] \\ &= E[NE[X]] \\ &= E[N]E[X] \end{aligned}$$

- cf) $E[S_N \mid N=n] = E\left[\sum_{k=1}^{n} X_k\right] = nE[X]$

  $\because$ iid

  $\therefore E[S_N \mid N] = NE[X]$

- The characteristic function of S_N

$$E[e^{j\omega S_N} \mid N=n] = E[e^{j\omega(X_1+\cdots+X_n)}]$$
$$= \Phi_X(\omega)^n$$
$$\therefore E[e^{j\omega S_N} \mid N] = \Phi_X(\omega)^N$$

$$\Phi_{S_N}(\omega) = E[E[e^{j\omega S_N} \mid N]]$$

$$= E[\Phi_X(\omega)^N]$$

$$= E[z^N]\Big|_{z=\Phi_X(\omega)}$$

$$= G_N(\Phi_X(\omega))$$

$\rightarrow$Generating function of N at $z = \Phi_X(\omega)$

5.2 The Sample Mean and the Laws of Large Numbers

- X : a random variable, $E[X] = \mu$ is unknown.

- $X_1, X_2, \ldots, X_n$:
  n independent, repeated measurements of X
  $\rightarrow$ X_j's : iid random variables with the same pdf as X

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

: Sample mean $\rightarrow$ a random variable
$\rightarrow$ used to estimate $E[X]$: M_n is the estimator

- Note : Properties of a good estimator.

① $E[M_n] = \mu$

② $E[(M_n - \mu)^2]$ should be small

- $E[M_n] = E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right]$ fixed n

$$= \frac{1}{n}\sum_{j=1}^{n} E[X_j] = \frac{1}{n}\cdot(n\mu) = \mu$$

$\because E[X_j] = E[X] = \mu$ for all j

$E[M_n] = \mu \rightarrow$ unbiased estimator for μ

- Mean square error of sample mean about μ

  = the variance of M_n

- i.e.,

  $$E[(M_n - \mu)^2] = E[(M_n - E[M_n])^2] = \mathrm{VAR}[M_n]$$

cf) $M_n = \dfrac{S_n}{n}, \quad S_n = X_1 + X_2 + \cdots + X_n$

$$\mathrm{VAR}[S_n] = n\,\mathrm{VAR}[X_j] = n\sigma^2$$

$\because$ X_j's are iid random variables.

$$\mathrm{VAR}[(cX)] = c^2\mathrm{VAR}[X]$$

$$\therefore \quad \mathrm{VAR}[M_n] = \mathrm{VAR}\left[\frac{S_n}{n}\right] = \frac{1}{n^2}\mathrm{VAR}[S_n]$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

- Chebyshev inequality

$$P[|M_n - E[M_n]| \geq \varepsilon] \leq \frac{\mathrm{VAR}[M_n]}{\varepsilon^2}$$

$$\Rightarrow P[|M_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow P[|M_n - \mu| < \varepsilon] \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

- Note

For any choice of error ε and probability $1-\delta$, we can select the number of samples n so that M_n is within ε of the true mean with probability $1-\delta$ or greater.

- Let n approach infinity, then

$$\lim_{n \to \infty} P[|M_n - \mu| < \varepsilon] = 1$$

❖ Ex. 5.9

- A voltage of constant, but unknown, value is measured.
- $X_j = v + N_j$
  - v → **Constant voltage**
  - N_j → **A noise voltage of zero mean and standard deviation of 1(μV)**
- Assume N_j's as independent r.v.'s
- Find n so that M_n is within $\varepsilon = 1\mu\text{V}$ of the true mean with probability $1-\delta = 0.99$ or greater.

$$1-\frac{\sigma^2}{n\varepsilon^2} = 1-\frac{1}{n} = 0.99 \quad \therefore n = 100$$

Weak Law of Large Numbers

- $X_1, X_2, \ldots$: a sequence of iid r.v.'s
  with finite mean $E[X] = \mu$, then for $\varepsilon > 0$

$$\lim_{n \to \infty} P[|M_n - \mu| < \varepsilon] = 1$$

  - With high probability, the sample mean for a large enough fixed value of n is close to the true mean.

Strong Law of Large Numbers

- $X_1, X_2, \ldots$: a sequence of iid r.v.'s
  with finite mean $E[X] = \mu$ and finite variance.

$$P[\lim_{n\to\infty} M_n = \mu] = 1$$

  - With probability 1, every sequence of sample mean will eventually approach and stay close to $E[X] = \mu$

5.3 The Central Limit Theorem

- Central Limit Theorem
  - S_n : the sum of iid random variables with finite mean $E[X] = \mu$ and finite variance σ^2
  - $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$: The zero-mean, unit variance random variable
  - cf) mean of $S_n = n\mu$ variance of $S_n = n\sigma^2$

    $\mathrm{VAR}[cX] = c^2\mathrm{VAR}[X]$

    $$\therefore \mathrm{VAR}\left[\frac{X}{\sigma}\right] = \frac{1}{\sigma^2}\mathrm{VAR}[X] = 1$$

$\therefore\ Z_n \to$ zero-mean and unit variance.

Then

$$\lim_{n\to\infty} P[Z_n \le z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

: cdf of Gaussian with zero mean and unit variance.

Gaussian Approximation for Binomial Probabilities

- $I_A(\zeta) = \begin{cases} 0 & \text{if } \zeta \text{ not in } A \\ 1 & \text{if } \zeta \text{ in } A \end{cases}$

  → indicator function : random variable

  → proof

  $p_I(0) = 1 - p$ and $p_I(1) = p$

  : Bernoulli random variable

- $X = I_1 + I_2 + \ldots + I_n$: Binomial r.v.

$$P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for} \ \ k = 0, \ldots, n.$$

→ The binomial r.v. is a sum of iid Bernoulli random variables

→ a binomial r.v. X with mean np and variance $np(1-p)$

- A random variable Y : Gaussian r.v. with mean np and variance $np(1-p)$

- For n large,

$$P[X=k] \cong P\left[k-\frac{1}{2}<Y<k+\frac{1}{2}\right]$$

→ interval of unit length about ***k***

$$=\frac{1}{\sqrt{2\pi np(1-p)}}\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} e^{-(x-np)^2/2np(1-p)}dx$$

- Approximation :

The integral → the product of the integrand at the center of the interval of integration ($x = k$) and the length of the interval of integration (one)

$$\Rightarrow P[X = k] \cong \frac{1}{\sqrt{2\pi np(1-p)}} e^{-(k-np)^2 / 2np(1-p)}$$

Proof of the Central Limit Theorem

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu)$$

$$\Rightarrow \quad \Phi_{Z_n}(\omega) = E[e^{j\omega Z_n}]$$

$$= E\left[\exp\left\{\frac{j\omega}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu)\right\}\right]$$

$$= E\left[\prod_{k=1}^{n} e^{j\omega(X_k-\mu)/\sigma\sqrt{n}}\right]$$

$$= \prod_{k=1}^{n} E\left[e^{j\omega(X_k-\mu)/\sigma\sqrt{n}}\right]$$

$$= \{E[e^{j\omega(X_k-\mu)/\sigma\sqrt{n}}]\}^n$$

cf) $E\,[e^{j\omega(X-\mu)/\sigma\sqrt{n}}]$

$$= E\left[1 + \frac{j\omega}{\sigma\sqrt{n}}(X-\mu) + \frac{(j\omega)^2}{2!(\sigma\sqrt{n})^2}(X-\mu)^2 + R(\omega)\right]$$

$$= 1 + \frac{j\omega}{\sigma\sqrt{n}} E[X-\mu] + \frac{(j\omega)^2}{2!(n\sigma^2)} E[(X-\mu)^2] + E[R(\omega)]$$

$$E[X-\mu] = 0, \quad E[(X-\mu)^2] = \sigma^2$$

$$\therefore E[e^{j\omega(X-\mu)/\sigma\sqrt{n}}] = 1 - \frac{\omega^2}{2n} + E[R(\omega)]$$

$E[R(\omega)] \rightarrow$ neglected as $n \rightarrow \infty$

$$\therefore \Phi_{Z_n}(\omega) \cong \left\{1 - \frac{\omega^2}{2n}\right\}^n$$

$$= 1 - n\frac{\omega^2}{2n} + \frac{n(n-1)}{2!}\left(\frac{\omega^2}{2n}\right)^2 + \frac{n(n-1)(n-2)}{3!}\left(\frac{\omega^2}{2n}\right)^3 - \ldots$$

$$\rightarrow e^{-\frac{\omega^2}{2}} \quad \text{as} \quad n \rightarrow \infty$$

- The Characteristic function of a zero-mean, unit variance Gaussian r.v.

5.4 Confidence Interval

- Sample mean estimator (r.v.)

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

- Sample variance (See Prob. 21)

$$V_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - M_n)^2$$

- Instead of estimating E[X], we attempt to specify an interval of values that is highly likely to contain the true value of the parameter
- That is, find an interval $[l(\mathbf{X}), u(\mathbf{X})]$ such that

$$P[l(\mathbf{X}) \le \mu \le u(\mathbf{X})] = 1 - \alpha$$

  - $(1-\alpha)\times 100\,\%$ confidence interval that depends on the pdf of the X_j's
  - $(1-\alpha)$: confidence level α : error

- Case 1: X_j's Gaussian; unknown Mean and Known Variance

For a Gaussian r.v. M_n and $\alpha = 2Q(z)$

$$1 - 2Q(z) = P[-z \le \frac{M_n - \mu}{\sigma / \sqrt{n}} \le z]$$

$$= P[M_n - \frac{z\sigma}{\sqrt{n}} \le \mu \le M_n + \frac{z\sigma}{\sqrt{n}}]$$

From $\alpha = 2Q(z_{\alpha/2})$ and known σ

$$(M_n - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le M_n + z_{\alpha/2}\sigma/\sqrt{n})$$

$\rightarrow$ $(1-\alpha)\times 100\,\%$ confidence interval for μ

- Case 2 : X_j's Gaussian; Mean and Variance unknown

From $V_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - M_n)^2$

Confidence interval for μ

$$(M_n - \frac{zV_n}{\sqrt{n}}, M_n + \frac{zV_n}{\sqrt{n}})$$

Probability for the interval

$$P[-z \le \frac{M_n - \mu}{V_n / \sqrt{n}} \le z] = P[M_n - \frac{zV_n}{\sqrt{n}} \le \mu \le M_n + \frac{zV_n}{\sqrt{n}}]$$

For r.v. W

$$W = \frac{M_n - \mu}{V_n / \sqrt{n}} = \frac{\sqrt{n}(M_n - \mu)/\sigma}{V_n / \sigma}$$

$$= \frac{(M_n - \mu)/(\sigma/\sqrt{n})}{\{[(n-1)V_n^2/\sigma^2]/(n-1)\}^{1/2}} \quad : \quad \frac{Gaussian \quad r.v.}{\begin{array}{l} Chi \;\; square \quad r.v.with \\ (n-1)\, degrees \;\; of \;\; freedom \end{array}}$$

W has a student's t-distribution with (n-1) degrees of freedom (from Example 4.38)

$$f_{n-1}(y) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)\sqrt{\pi(n-1)}}\left(1 + \frac{y^2}{n-1}\right)^{-n/2}$$

The probability

$$P[M_n - \frac{zV_n}{\sqrt{n}} \le \mu \le M_n + \frac{zV_n}{\sqrt{n}}] = \int_{-z}^{z} f_{n-1}(y)dy = 1 - 2F_{n-1}(-z)$$

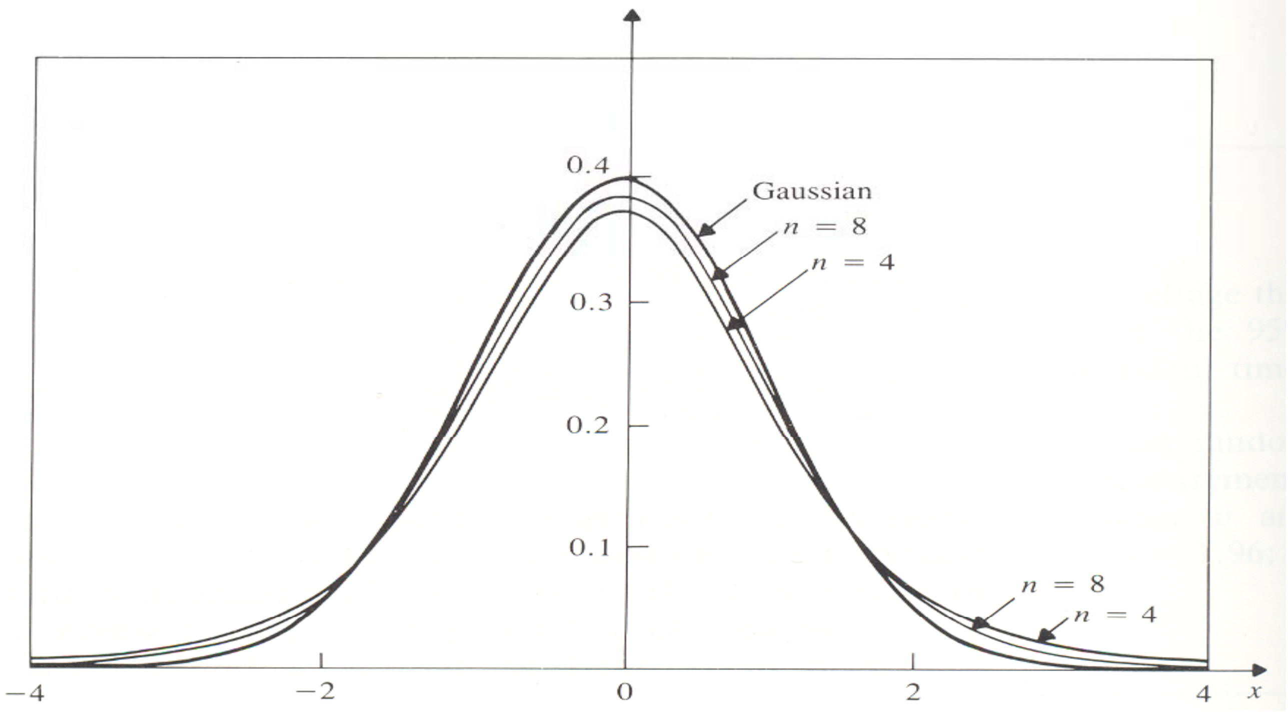
$(1-\alpha)$ confidence interval from $\alpha = 2F_{n-1}(-z_{\alpha/2,n-1})$

$$(M_n - z_{\alpha/2,n-1}V_n/\sqrt{n}, M_n + z_{\alpha/2,n-1}V_n/\sqrt{n})$$

: Confidence interval for the mean μ

Find $z_{\alpha/2,n-1}$ according to the look-up table

- Figure 5.7



< Gaussian pdf and Student's t pdf for n=4 and n=8 >

- Case 3 : X_j's non-Gaussian – mean and variance unknown
  - Use the method of batch means (μ)

  - M independent experiments

  - Sample mean each from n iid observations

  - Gaussian sample mean and $(M_n - z_{\alpha/2,n-1} V_n/\sqrt{n}, M_n + z_{\alpha/2,n-1} V_n/\sqrt{n})$

    by the central limit theorem

-> compute the confidence interval by using *M* sample means

HW: 6, 8, 10, 14, 16, 21, 25, 29, 31, 35