Inequalities/Bounds

- **Chebyshev Inequality** (a bound on the probability of how much X can deviate from $E[X] = \mu_X$): $\boxed{P[|X - \mu_X| \geq \varepsilon] \leq \frac{\sigma_X^2}{\varepsilon^2}}$
- **Markov Inequality**: For RVs with nonnegative values, $\boxed{P[X \geq \varepsilon] \leq \frac{E[X]}{\varepsilon}}$
- **The Schwarz Inequality**: i) $|\mathrm{Cov}[X, Y]| \leq \sigma_X \sigma_Y$,

  ii) For real functions $h(X)$ and $g(X)$ of a real rv X,

$$|E[h(X)g(X)]| \leq \sqrt{E[h^2(X)]}\sqrt{E[g^2(X)]}.$$

- **Chernoff Bound** (upper bound on the tail probability):

$$\boxed{P[X \geq a] \leq \min_t \{e^{-at}\theta_X(t)\}}$$

- Example: For $X \sim N(\mu, \sigma^2)$, and $a > \mu$, the Chernoff bound is $P[X \geq a] \leq e^{-(a-\mu)^2/(2\sigma^2)}$.

  $\theta_X(t) = e^{\mu t + 0.5\sigma^2 t^2} \quad \Rightarrow \quad P[X \geq a] \leq \min_t\{e^{-at}\theta_X(t)\} = \min_t\{e^{-at}e^{\mu t + 0.5\sigma^2 t^2}\}$

  Let $g(t) \triangleq e^{-at}e^{\mu t + 0.5\sigma^2 t^2}$. Then, $g'(t) = 0$ gives $t_m = (a-\mu)/\sigma^2$.

  As $g''(t) > 0$, $g(t)$ is minimum at t_m. Then $g(t_m)$ gives the above bound.

- **Bienayme Inequality:**

$$\boxed{P\{|X-a|^n \geq \varepsilon^n\} \leq \frac{E\{|X-a|^n\}}{\varepsilon^n}}$$

$$\text{Hence,} \quad P\{|X-a| \geq \varepsilon\} \leq \frac{E\{|X-a|^n\}}{\varepsilon^n}.$$

  Chebyshev inequality is a special case obtained with $a = \mu$ and $n = 2$.

- **Lyapunov Inequality:**

  Let $\boxed{\beta_k = E\{|X|^k\} < \infty}$ represent the absolute moments of the random variable X. Then for any k, $\boxed{\beta_{k-1}^{1/(k-1)} \leq \beta_k^{1/k}}$

- **The Weak Law of Large Numbers (LLN):**
For iid X_i, $i = 1, \ldots, n$ with a finite mean μ_X, consider the sample mean estimator $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for $\delta > 0$,

$$\boxed{\lim_{n\to\infty} P[|\hat{\mu}_n - \mu_X| < \delta] = 1}.$$

- Since we have $E[\hat{\mu}_n] = \mu_X$ & $\mathrm{Var}[\hat{\mu}_n] = \frac{\sigma_X^2}{n}$, by the Chebyshev inequality, $P[|\hat{\mu}_n - \mu_X| \geq \delta] \leq \frac{\sigma_X^2}{n\delta^2}$, $\lim_{n\to\infty} P[|\hat{\mu}_n - \mu_X| \geq \delta] = 0$, and $\lim_{n\to\infty} P[|\hat{\mu}_n - \mu_X| < \delta] = 1$.

- **The Weak Law of Large Numbers (LLN) - Non-uniform Variance:**
Let X_i be an independent random sequence with constant mean μ and variance σ_i^2 defined for $i \geq 1$. Then if $\lim_{n\to\infty} \sum_{i=1}^{n} \sigma_i^2/n^2 < \infty$, $\hat{\mu}[n] \triangleq (1/n)\sum_{i=1}^{n} X_i \to \mu$ (convergence in probability) as $n \to \infty$.

- For a large enough fixed value of n, the sample mean using n samples will be close to the true mean with high probability. WLLN does not address the question about what happens to the sample mean as a function of n as we make additional measurements.

- **The Strong Law of Large Numbers (LLN):**
  For a sequence of iid X_i, $i = 1, \ldots, n$ with a finite mean μ_X and finite variance,

$$\boxed{P[\lim_{n\to\infty} \hat{\mu}_n = \mu_X] = 1}.$$

- With probability 1, every sequence of sample mean calculations will eventually approach and stay close to μ_X.
- LLN is the theoretical basis for estimating μ_X from measurements.

The Central Limit Theorem

- Theorem: Let $X_1, \ldots, X_n$ be n mutually independent r.v.'s with CDF's $F_{X_1}(x_1), \ldots, F_{X_n}(x_n)$, and $\bar{X}_k = 0$, Var$[X_k] = \sigma_k^2$. Denote

$$s_n^2 \triangleq \sigma_1^2 + \ldots + \sigma_n^2.$$

If, for a given $\varepsilon > 0$ and a sufficiently large n, $\sigma_k < \varepsilon s_n$ $\forall k$, then the normalized sum $Z_n \triangleq (X_1 + \ldots + X_n)/s_n$ converges to the standard Normal CDF, i.e., $\boxed{\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z) = 1 - Q(z).}$

Central: CDF converges to normal CDF around the center (mean).

- Theorem: Let $X_1, \ldots, X_n$ be iid r.v.s with $\bar{X}_i = 0$ and Var$[X_i]$$= 1$, $\forall i$. Then $Z_n \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ tends to the normal in the sense that its characteristic function Φ_{Z_n} approaches to the characteristic function of $N(0,1)$, i.e., $\boxed{\lim_{n\to\infty} \Phi_{Z_n}(w) = e^{-\frac{w^2}{2}}}$.

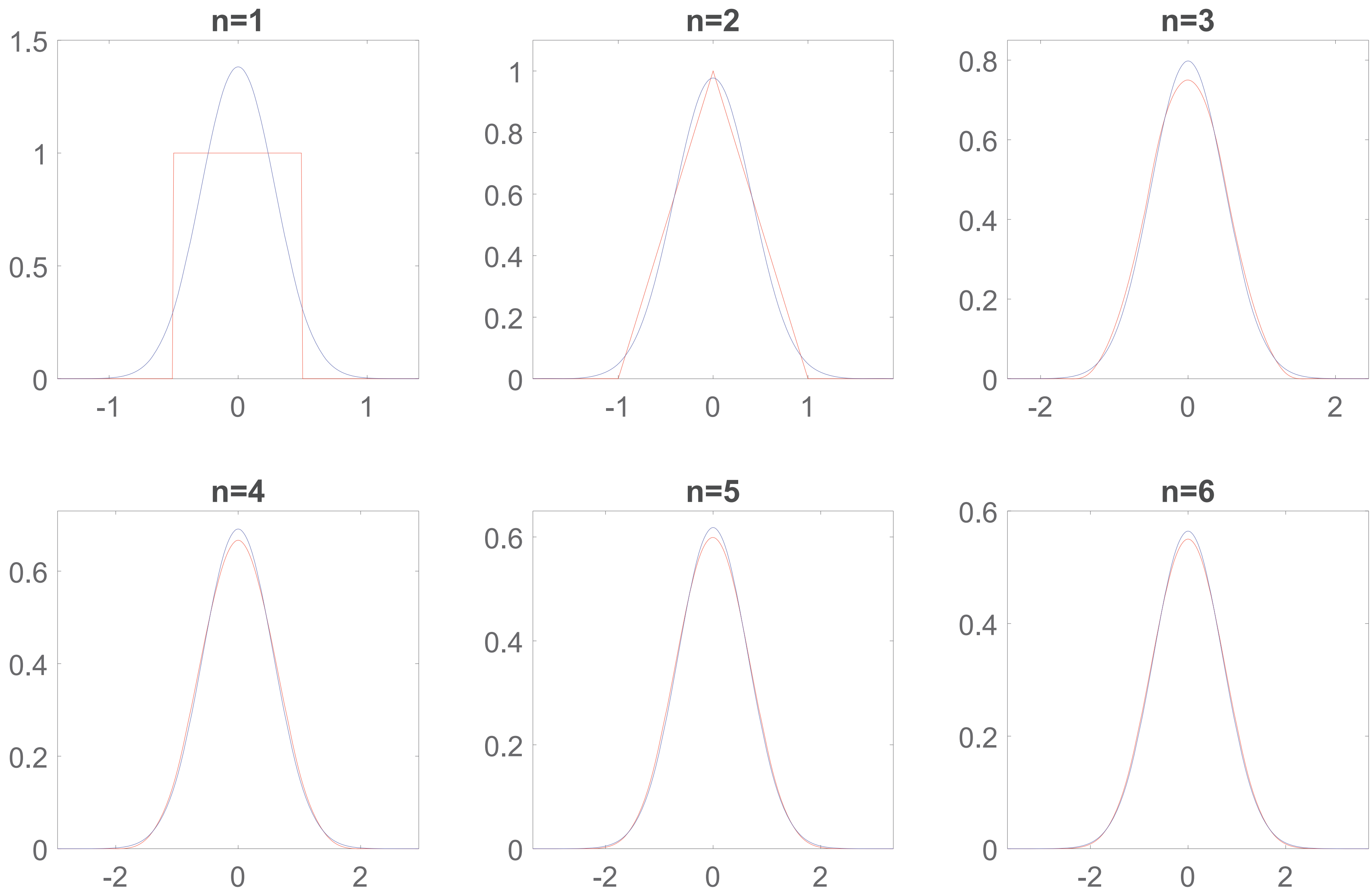

Fig. (An illustration of the CLT) Comparison of the Normal pdf $\mathcal{N}(0, n/12)$ (dotted line) and the pdf of $Y = \sum_{i=1}^{n} X_i$ (solid line) where $\{X_i\}$ are iid uniform r.v.'s within $[-0.5, 0.5]$

- Example: The time between events in a certain random experiment is i.i.d. exponential random variables with mean m seconds. Find the probability that the 1000th event occurs in the time interval $(1000 \pm 50)m$.

X_j = the time between events (inter-arrival time)
S_n = the (occurrence or arrival) time of the nth event

$$S_n = X_1 + X_2 + \ldots + X_n.$$

Exponential: $E[X_j] = m$ and $\mathrm{Var}[X_j] = m^2$

$$E[S_n] = nE[X_j] = nm,$$

$$\mathrm{Var}[\mathrm{S_n}] = n\mathrm{Var}[\mathrm{X_j}] = nm^2.$$

Let $Z_n = (S_n - E[S_n])/\sqrt{\mathrm{Var}[S_n]}$. Then, the CLT gives

$$P[950m \leq S_{1000} \leq 1050m] = P\left[\frac{950m - 1000m}{m\sqrt{1000}} \leq Z_n \leq \frac{1050m - 1000m}{m\sqrt{1000}}\right]$$

$$= P\left[-1.58 \leq Z_n \leq 1.58\right] \simeq Q(-1.58) - Q(1.58) = 1 - 2Q(1.58) = 0.8866.$$

Thus, as n becomes large, S_n is very likely to be close to its mean nm.
Hence, we can conjecture that the long-term average rate at which events occur is $\frac{n \text{ events}}{S_n \text{ seconds}} = \frac{n}{nm} = \frac{1}{m}$ events/second.

The Berry-Esseen Theorem

If $E[\mathbf{x}_i^3] \le c\sigma_i^2$, $\forall\, i$, where c is some constant, then the distribution $F_{\bar{x}}$ of the normalized sum

$$\bar{\mathbf{x}} = \frac{\mathbf{x}_1 + \cdots + \mathbf{x}_n}{n}$$

is close to the normal distribution $G(x)$ in the following sense with $\sigma^2 = \sigma_1^2 + \cdots + \sigma_n^2$:

$$\boxed{|F_{\bar{x}} - G(x)| < \frac{4c}{\sigma}} \quad \text{(bound)}.$$

The central limit theorem is a corollary of this theorem because this theorem leads to the conclusion that

$$F_{\bar{x}} \to G(x) \quad \text{as} \quad \sigma \to \infty.$$

Note-1: Whereas the above is the convergence in distribution of $\bar{\mathbf{x}}$ to a normal random variable, the theorem also gives a *bound* of the deviation of $\bar{F}(x)$ from normality.

Note-2: The condition for the theorem is not too restrictive. It holds, for example, if the random variables $\mathbf{x}_i$ are i.i.d. and their third moment is finite.

- **Theorem:** The pdf of $Y \triangleq X_1 X_2 \ldots X_n$ for independent continuous and positive r.v.'s X_i with a large n:
For large n, the density of $\mathbf{y}$ is approximately *lognormal*:

$$f_y(y) \simeq \frac{1}{y\ \sigma\ \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (\ln y - \eta)^2 \right\}\ U(y)$$

where

$$\eta = \sum_{i=1}^{n} E[\ln \mathbf{x}_i] \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \mathrm{Var}(\ln \mathbf{x}_i).$$

Proof: The random variable

$$\mathbf{z} = \ln \mathbf{y} = \ln \mathbf{x}_1 + \cdots + \ln \mathbf{x}_n$$

is the sum of the random variables $\ln \mathbf{x}_i$. From the CLT, for large n, $\mathbf{z}$ is nearly $\mathcal{N}(\eta, \sigma^2)$. And $\mathbf{y} = e^{z} \Rightarrow$ lognormal $\mathbf{y}$.
The theorem holds if $\ln \mathbf{x}_i$'s satisfy the conditions for the validity of the CLT.

Entropy, Differential Entropy, and Relative Entropy

- H_X = Entropy of a discrete RV X
  = expected value of uncertainty of the value of X
  = average amount of information required to identify the value of X
- For a discrete RV X with PMF $P_X(x)$, $\boxed{H_X = E[-\log(P_X(x))]}$, i.e., $H_X = -\sum_i P_X(x_i)\log(P_X(x_i))$ where if $\log$ is base 2, unit of H_X is bits.
- Relative entropy of $Y \in \{a_1, \cdots, a_K\}$ with respect to $X \in \{a_1, \cdots, a_K\}$ is $\boxed{H(X;Y) = E_X[\log(\frac{P_X(a_i)}{P_Y(a_i)})]}$, i.e., $H(X;Y) = \sum_{i=1}^{K} P_X(a_i)\log(\frac{P_X(a_i)}{P_Y(a_i)})$ which is nonnegative and equal to zero iff $P_X(a_i) = P_Y(a_i)$ for all i.
  Note: $0\log(0/0) \triangleq 0$, $\log(0/q) \triangleq 0$, $p\log(1/0) \triangleq \infty$.
- Differential entropy of a continuous RV X with pdf $f_X(x)$:
  $\boxed{H_X = E[-\log(f_X(x))]}$, i.e., $H_X = -\int_{-\infty}^{\infty} f_X(x)\log(f_X(x))dx$.
- Relative entropy of continuous RVs Y w.r.t X with pdfs $f_Y(y)$ and $f_X(x)$:
  $\boxed{H(X;Y) = E_X[\log(\frac{f_X(x)}{f_Y(x)})]}$, i.e., $H(X;Y) = \int_{-\infty}^{\infty} f_X(x)\log(\frac{f_X(x)}{f_Y(x)})dx$

Entropy, Differential Entropy, and Relative Entropy

- Kullback-Leibler divergence (= Relative entropy):

  For discrete RVs, $\boxed{D_{\mathrm{KL}}(P_X||P_Y) = H(X;Y)}$

  For continuous RVs, $\boxed{D_{\mathrm{KL}}(f_X||f_Y) = H(X;Y)}$.

  A measure of how one pdf/PMF is different from a second reference pdf/PMF.

  In Bayesian perspective, it is information gained when we revise beliefs from prior pdf f_Y (PMF P_Y) to posterior pdf f_X (PMF P_X).

  In general, $D_{\mathrm{KL}}(f_X||f_Y) \neq D_{\mathrm{KL}}(f_Y||f_X)$ and
  $D_{\mathrm{KL}}(P_X||P_Y) \neq D_{\mathrm{KL}}(P_Y||P_X)$.

Method of Maximum Entropy

- For a discrete RV $X \in \{x_1, \cdots, x_K\}$ with unknown PMF $P_X(x)$, suppose we know $E[g(X)] = c$. Then, the PMF which maximizes the entropy is given by
  $\boxed{P_X(x_i) = Ae^{-\lambda g(x_i)}}$
  where A and λ are chosen to satisfy a valid PMF and $E[g(X)] = c$.

- For a continuous RV X with unknown pdf $f_X(x)$, suppose we know $E[g(X)] = c$. Then, the pdf which maximizes the differential entropy is given by
  $\boxed{f_X(x) = Ae^{-\lambda g(x)}}$
  where A and λ must be chosen to satisfy a valid pdf and $E[g(X)] = c$.