

## Sums of Random Variables:

Law of Large Numbers (LLN)  
Central Limit Theorem (CLT)

Many practical experiments rely on vectors of observations (sampling a random waveform or taking measurements in repeated trials). These observations are used to form estimates of parameters of a physical system

"best" conjecture  
about parameters  
based on  
observations

or counting occurrence of some events. We often are interested in adding up observations and normalize by the number of observations (ex. sample mean and sample variance in your HW assignments).

We are interested in empirically evaluating c.d.f. or density, etc.

Today: we will look at sample mean and relative frequency as the estimates of the mean and the probability of an event.

We make statements when the estimates are "good."

We formulate LLN.

This theoretical result demonstrate consistency between probability and statistics.

We then talk about modeling cumulative behavior of observations and state CLT, which approximates pdf of sum of RVs by Gaussian pdf.

Consider a sequence of RVs

$X_1, \dots, X_n$

If the mean of each RV is specified,  $E[X_i] = \mu_i$ , and  $\{\text{cov}(X_i, X_j)\}_{i,j=1}^n$  is known, we can form the sum  $S_n = \sum_{i=1}^n X_i$  and evaluate its mean and variance.

$$E[S_n] = \sum_{i=1}^n \underbrace{E[X_i]}_{\mu_i} = \sum_{i=1}^n \mu_i = n\mu$$

↑  
linearity

↑  
equal mean

$$\text{var}(S_n) = E[(S_n - E(S_n))^2]$$

$$= E\left[\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)^2\right]$$

$$= E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right]$$

$$\stackrel{\text{linearity}}{=} \sum_{i=1}^n \sum_{j=1}^n \underbrace{E[(X_i - \mu_i)(X_j - \mu_j)]}_{\text{cov}(X_i, X_j)}$$

$$\stackrel{\text{uncorrelated}}{=} \sum_{i=1}^n \sigma_i^2 \stackrel{\text{equal variance}}{=} n \cdot \sigma^2$$

### The weak Law of Large Numbers

Consider a RV  $X$  with unknown mean  $\mu$ . Assume that  $n$  measurements of this RV are available  $X_1, X_2, \dots, X_n$ .

independent measurements

To conjecture about  $\mu$  based on the measurements

3/6

form the sample mean

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Since  $\{X_i\}$  is a set of RVs,  $M_1, M_2, \dots, M_n, \dots$  is a set of RVs.

If we evaluate the mean, we want a "good" estimate of the mean.

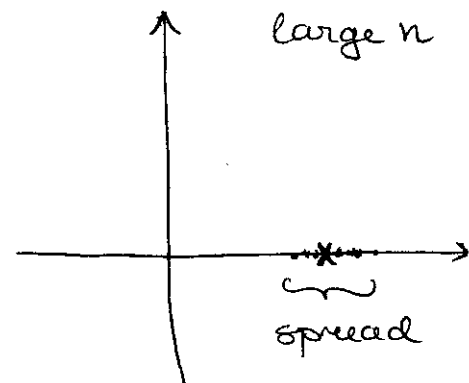
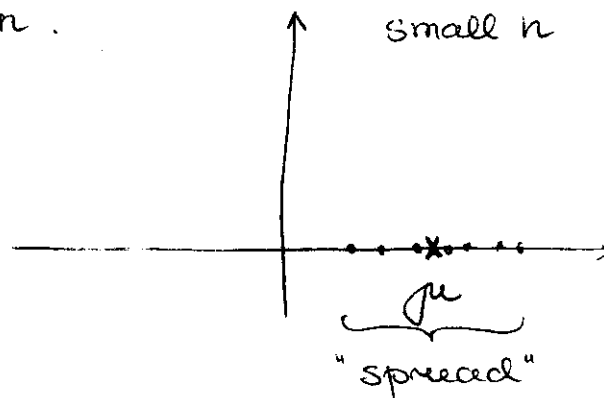
we have to specify what "good estimate" means

A good estimate has to have a set of properties:

(1)  $E[M_n] = \mu$  (estimate is unbiased) or unbiased for large  $n$   
 $\int x: \mu$

(2)  $\text{var}(M_n)$  should be small  
 For  $n \rightarrow +\infty$ ,  $\text{var}(M_n) \rightarrow 0$ .

The estimate is concentrated around the true mean.



We can evaluate

$$P[|M_n - \mu| > \epsilon]$$

and state the conditions when

$$P[|M_n - \mu| > \epsilon] \xrightarrow{n \rightarrow +\infty} 0$$

First, evaluate  $E[M_n]$  and  $\text{var}(M_n)$

Note  $M_n = \frac{1}{n} S_n$

$X_i \sim \text{iid}$

(independent identically distributed)

Then  $E[M_n] = \frac{1}{n} \cdot n\mu = \mu$

and

$$\begin{aligned} \text{var}(M_n) &= \text{var}\left(\frac{1}{n} S_n\right) = E\left[\left(\frac{1}{n} S_n - \frac{1}{n} E[S_n]\right)^2\right] \\ &= \left(\frac{1}{n}\right)^2 \cdot E[(S_n - E[S_n])^2] \stackrel{X_i \sim \text{iid}}{=} \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Use Chebychev inequality to upperbound the probability  $P[|M_n - \mu| > \varepsilon] \leq \frac{\text{var}(M_n)}{\varepsilon^2}$

$$\stackrel{X_i \sim \text{iid}}{=} \frac{\sigma^2}{n \varepsilon^2} \xrightarrow{n \rightarrow +\infty} 0$$

Summary: We showed that the sequence of probabilities  $P[|M_n - \mu| > \varepsilon] \xrightarrow{\text{as } n \rightarrow +\infty} 0$  under the condition that  $\text{var}(M_n) \xrightarrow{\text{as } n \rightarrow +\infty} 0$ .

Since we are dealing with the sequence of probabilities, we will say that

$$M_n \rightarrow \mu \text{ in probability as } n \rightarrow +\infty.$$

Often it is written as

$$M_n \xrightarrow{P} \mu \text{ as } n \rightarrow +\infty.$$

This is a loose proof of the fundamental result known as the weak LLN.

5/6

Theorem: Consider a sequence  $\underbrace{X_1, \dots, X_n}_{\text{of iid Random}}$  samples with finite mean  $E[X] = \mu$  and  $\text{var}(X) = \sigma^2$ .  
Then for any  $\varepsilon > 0$  if  $\frac{\sigma^2}{n} \rightarrow 0$

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu \text{ as } n \rightarrow +\infty.$$

meaning

$$P[|M_n - \mu| > \varepsilon] \xrightarrow{n \rightarrow +\infty} 0.$$

ex.

Consider the Indicator function

$$X_i = \begin{cases} 1, & Y_i \leq y \\ 0, & \text{otherwise} \end{cases}$$

Let  $\{X_i\}_{i=1}^n$  be independent (not necessarily identically distributed)

$$\underbrace{E[X_i]}_{m_i} = E[I_{Y_i}] = 1 \cdot P[Y_i \leq y] + 0 \cdot P[Y_i > y] = F_{Y_i}(y)$$

$$\begin{aligned} \underbrace{\text{var}(X_i)}_{\sigma_i^2} &= E[(X_i - m_i)^2] = E[(I_{Y_i} - m_i)^2] \\ &= E[I_{Y_i}^2] - m_i^2 \\ &= \underbrace{F_{Y_i}(y) - F_{Y_i}^2(y)}_{< 1} \end{aligned}$$

Then if we form the sample estimate of the edf of  $Y$  as  $\frac{1}{n} \sum_{i=1}^n I_{Y_i}$ , then

$$\lim_{n \rightarrow +\infty} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 = 0$$

and  $P\left[ \left| \frac{\sum_{i=1}^n I_{Y_i}}{n} - \frac{1}{n} \sum_{i=1}^n m_i \right| > \varepsilon \right] \xrightarrow{\text{as } n \rightarrow +\infty} 0$

6/6

By the weak LLN

$$\frac{1}{n} \sum_{i=1}^n (I_{Y_i} - m_i) \xrightarrow{P} 0 \text{ as } n \rightarrow +\infty$$

If  $X_i$  are iid. Then

$$\frac{1}{n} \sum_{i=1}^n I_{Y_i \leq y} \xrightarrow{P} F_Y(y) \text{ as } n \rightarrow +\infty$$

in the following sense

$$P \left[ \left| \frac{1}{n} \sum_{i=1}^n I_{Y_i \leq y} - F_Y(y) \right| > \varepsilon \right] \rightarrow 0 \text{ as } n \rightarrow +\infty$$

for any  $\varepsilon > 0$