

## 4.6 Random Samples

### DEF: Random Sample

The set of random vars  $X_1, \dots, X_n$  is said to be a random sample of size  $n$  from a population with density function  $f(x)$  if the joint pdf has the form

$$f(x_1, x_2, \dots, x_n) = \underbrace{f(x_1) \cdot f(x_2) \cdots f(x_n)}_{\text{all of the } f\text{'s are the same.}}$$

(In other words, the R.V. are independent and follow a common dist.)

Ex: The lifetime of a certain type of light bulbs is assumed to follow the exponential density

$$f(x) = e^{-x} I_{(0, \infty)}(x)$$

A RS<sup>←</sup> of size 2 is obtained. Then

$$f(x_1, x_2) = e^{-(x_1+x_2)} \prod_{i=1}^2 I_{(0, \infty)}(x_i)$$

Suppose the total lifetime of the two bulbs turned out to be  $\frac{1}{2}$  year?

One may wonder whether this result is reasonable given the assumed model

If the model isn't appropriate, another should be chosen. We can find how appropriate by computing probs

$$\begin{aligned}
 P[X_1 + X_2 \leq c] &= \int_0^c \int_0^{c-x_2} e^{-(x_1+x_2)} dx_1 dx_2 \\
 &= \int_0^c e^{-x_2} \int_0^{c-x_2} e^{-x_1} dx_1 dx_2 \\
 &= \int_0^c e^{-x_2} (-e^{-x_1}) \Big|_0^{c-x_2} dx_2 \\
 &= \int_0^c e^{-x_2} (1 - e^{-(c-x_2)}) dx_2
 \end{aligned}$$

$$\begin{aligned}
 &= \int_0^c e^{-x_2} - e^{-c} dx_2 \\
 &= -e^{-x_2} \Big|_0^c - e^{-c} x_2 \Big|_0^c \\
 &= 1 - e^{-c} - ce^{-c}
 \end{aligned}$$

For  $c = .5$ ,  $P(X_1 + X_2 \leq .5) = .09$

It is unlikely to find the total lifetime of two bulbs to be  $\frac{1}{2}$  year or less, if the true pop model is given by (\*)

## EMPIRICAL DIST    Empirical CDF

DEF -  $X_1, X_2, \dots, X_n$  RS of size  $n$

-  $X_i \sim f(x)$

-  $W =$  number of variables  $X_i \leq x$

-  $W$  counts the number of successes in  $n$  indep. Bernoulli trials  $W \sim \text{BIN}(n, p)$   $p = F(x)$

- Relative frequency of success on  $n$  indep trials is

$$F_n(x) = \frac{W}{n}$$

$F_n$  is referred to as the empirical CDF

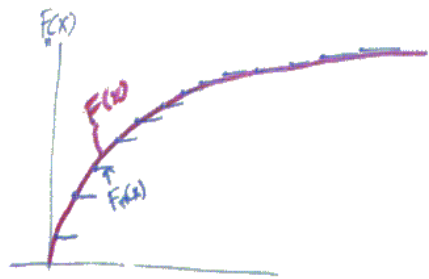
$F_n(x)$  should be close to  $F(x)$  for large  $n$

- We have data from a RS,  $X_1, \dots, X_n$

- Let  $y_1 < y_2 < \dots < y_n$  be the ordered values of the data.

The Empirical CDF based on this data is

$$F_n(x) = \begin{cases} 0, & x < y_1 \\ i/n, & y_i \leq x < y_{i+1} \\ 1, & y_n \leq x \end{cases}$$



### Histograms

Similar to Empirical CDF, but compares a histogram with the pdf.

### Rationale:

Divide the data into  $k$  disjoint intervals,

$$\text{say } I_j = (a_j, a_{j+1}], \quad j=1, \dots, k$$

The relative frequency  $f_j$  with which an observation falls into  $I_j$  gives a rough indication of what values the pdf  $f(x)$  might have over the interval

Define  $E_1, E_2, \dots, E_k, E_{k+1}$  as

$E_j$  occurs iff  $x_i \in I_j$  for some  $i$

$E_{k+1}$  occurs iff  $x_i$  is not in any  $I_j$

DEF:  $Y_j$  = Number of variables that fall into  $I_j$

$$Y = (Y_1, Y_2, \dots, Y_k)$$

$$Y \sim \text{MULT}(n, p_1, p_2, \dots, p_k)$$

$$p_j = F(a_{j+1}) - F(a_j) = \int_{a_j}^{a_{j+1}} f(x) dx$$

EX: Observed lifetimes (in months) of a RS  
of 40 electrical parts (ordered)

0.15 2.37 2.90 7.39 7.99 12.05 15.17 17.56  
 22.40 34.84 35.39 36.38 39.52 41.07 46.50 50.52  
 52.54 58.91 58.93 66.71 71.48 71.84 77.66 79.31  
 80.70 90.87 91.22 96.35 108.92 112.26 122.71 126.87  
 127.05 137.90 167.59 183.53 282.49 335.33 341.19 409.97

Choose interval range from 0.15 to 409.97  $\Rightarrow$  9 intervals

$$I_1 = (0, 50]$$

$$I_2 = (50, 100]$$

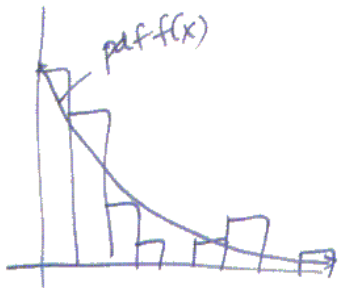
:

$$I_9 = (400, 450]$$

Freq dist of lifetimes of 40 comp.

Limits of $I_j$	$y_j$	$y_j/n$	Height Interval
0-50	15	0.375	$0.375/50 = .0075$
50-100	13	0.325	:
100-150	6	0.150	:
150-200	2	0.050	:
200-250	0	0.000	:
250-300	1	0.025	:
300-350	2	0.050	:
350-400	0	0.000	:
400-450	1	0.025	:

$\downarrow$  divide by length of Height Interval  
 $0.375/50 = .0075$



A smooth curve through the tops of the rectangles would provide a direct approx to the pdf

Things that would affect the picture

- number of intervals
- length of "
- sample
- range of data
- randomness (different for a different sample)  
↑ "sampling error"

This one looks pretty good!

Note: Sampling must be "with" replacement in order for the def of RS to apply

Note that if the pop. is quite large, then the def of RS will be approx correct if the sampling is without replacement