[4.6] Random Samples

DEF: Random Sample

The set of random vars $X_1, \cdots, X_n$ 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

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

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

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

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

A RS (random sample) 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~~ 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

$$P[x_1 + x_2 \leqslant c] = \int_0^c \int_0^{c-x_2} e^{-(x_1+x_2)} dx_1 dx_2$$

$$= \int_0^c e^{-x_2} \int_0^{c-x_2} e^{-x_1} dx_1 dx_2$$

$$= \int_0^c e^{-x_2} \Big[- e^{-x_1} \Big]_0^{c-x_2} dx_2$$

$$= \int_0^c e^{-x_2} (1 - e^{-(c-x_2)}) dx_2$$

$$= \int_0^c e^{-x_2} - e^{-c} dx_2$$

$$= \Big[- e^{-x_2} - x_2 e^{-c} \Big]_0^c$$

$$= 1 - e^{-c} - ce^{-c}$$

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 (*)

If the probability doesn't make sense for the situation, then it isn't a good model. The model is not representative of what occurs in practice.

EMPIRICAL CDF

$\boxed{\text{DEF:}}$

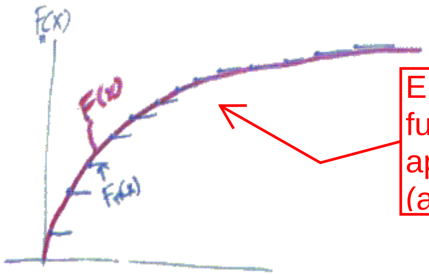
- $x_1, x_2, \cdots, x_n$ is a RS of size n
- $x_i \sim f(x)$
- W = number of variables $x_i \leqslant x$
- W counts the number of successes in n indep Bernoulli trials $W \sim \mathrm{BIN}(n, F(x))$
- Relative frequency of successes on n indep. trials
- $F_n(x) = \dfrac{W}{n}$
- F_n is referred to as the "empirical CDF"
- $F_n(x)$ should be close to $F(x)$ for large n.

– Suppose we have a RS from $x_1, \cdots, x_n$

– Let $y_1 < y_2 < \cdots < y_n$ be the ordered values of the data.

–The empirical CDF based on this data is:

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



Empirical CDF is a step function which approximates the CDF (a flattened S-curve)

Instead of a CDF, show the pdf version of the empirical CDF

Histograms

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

Rationale:

Divide the data into k disjoint intervals,

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

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 $E_1, E_2, \cdots, E_k, E_{k+1}$ 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

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

$Y \sim MULT(n, p_1, p_2, \cdots, p_k)$

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

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
80.90	90.87	91.22	96.35	108.92	112.26	122.71	126.87
127.05	137.96	167.59	183.53	282.49	335.33	341.19	409.97

Choose interval range from 0.15 to 409.97 ⇒ 9 intervals

$I_1 = (0, 50]$

$I_2 = (50, 100]$

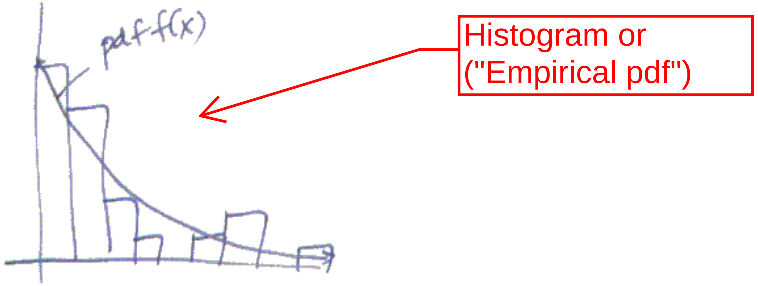
⋮

$I_9 = (400, 450]$

Freq dist of lifetimes of 40 comp.

Limits of I_j	y_j	y_j/n	Height ↙ divide by length of 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	⋮

To get the right scale, you need to make the height of the bars be scaled by the length of the current interval. (height = (yj/n)/length)



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