Some Elementary Statistics of Data.

Dfn (Sample Variance & Sample standard Deviation):

From a given sequence of RVs $X_1, X_2, \ldots, X_n$, we may obtain another statistic called the n-samples variance or simply the sample variance:

$$T((X_1, X_2, \ldots, X_n)) = S_n^2((X_1, X_2, \ldots, X_n)) := \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

For brevity, we write $S_n^2((X_1, X_2, \ldots, X_n))$ as S_n^2 and its realisation $S_n^2((x_1, x_2, \ldots, x_n))$ based on the realised or observed data $(x_1, x_2, \ldots, x_n)$ as s_n^2.

Sample Standard deviation is the square-root of S_n^2.

$$S_n((X_1, X_2, \ldots, X_n)) = \sqrt{S_n^2((X_1, X_2, \ldots, X_n))}$$

For brevity, we write $S_n((X_1, \ldots, X_n))$ as S_n and its realisation $S_n((x_1, x_2, \ldots, x_n))$ as s_n.

Once again, if $X_1, X_2, \ldots, X_n \overset{IID}{\sim} X_1$, the expectation of the sample variance is:

$$E(S_n^2) = V(X_1)$$

(exercise: show this is the case using properties of Expectations)

Example (1st Ball of Lotto Data): Suppose $X_1, X_2, \ldots, X_{1114} \overset{IID}{\sim}$ de Moivre$(1/40, \ldots, 1/40)$

Go back to Lab 5 and find out

$\overline{x}_{1114} \overset{?}{=}$

$s_n^2((x_1, x_2, \ldots, x_{1114})) \overset{?}{=}$

$s_n((x_1, \ldots, x_{1114})) \overset{?}{=}$

where observed 1st ball data $x = (x_1, \ldots, x_{1114})$

Dfn (Order Statistics).

Suppose $X_1, X_2, \ldots, X_n$ is a sequence of RVs. Then, the n-sample order statistics $X_{([n])}$ is:

$$X_{([n])}((X_1, X_2, \ldots, X_n)) := (X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \text{ such that,}$$

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

For brevity, we write $X_{([n])}((X_1, \ldots, X_n))$ as $X_{([n])}$ and its realisation $X_{([n])}((x_1, x_2, \ldots, x_n))$ as $x_{([n])} = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$.

Thus, we simply sort the data to get the order statistic.

Ex Suppose the outcome of 3 Bernoulli trials is $x = (0, 1, 0)$ then $x_{([3])} = (0, 0, 1)$.

Ex Recall the order statistic of the 1st Ball of Lotto data.

$$x_{([1114])} = (1, 1, 1, \ldots, 40, 40)$$

Ex Suppose the outcome of 5 IID Uniform(0,1) trials is

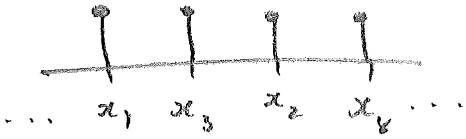
$x = (0.12896\ldots, 0.293858\ldots, 9.8665432, 0.4589321\ldots, 0.7232310\ldots)$

then $x_{([5])} = 0.12896; 0.293658\ldots, 0.4568932\ldots, 0.723210\ldots, 9.8665432\ldots)$

We use sorting algorithms to do our sorting efficiently. Take COSC courses to learn more about sorting.

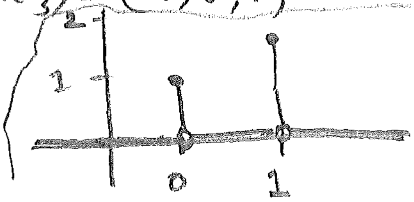
Dfn: EMF or Empirical Mass Function of a sequence of observed data $x_1, x_2, \cdots, x_n$ is the sum of the following indicator functions:

$$EMF((x_1, x_2, \cdots, x_n)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i\}}(x)$$



Ex: Toss a coin thrice with $(x_1, x_2, x_3) = (1,0,1)$

Then EMF is: $\frac{1}{3}\left(\mathbb{1}_{\{1\}}(x) + \mathbb{1}_{\{0\}}(x) + \mathbb{1}_{\{1\}}(x)\right)$



Ex: Recall how the dictionary was used to get the relative frequencies for the 40 ball numbers:

$$\frac{1}{1114}\left(\sum_{i=1}^{1114} \mathbb{1}_{\{x_i\}}(x)\right)$$

Dfn: Empirical Distribution Function:

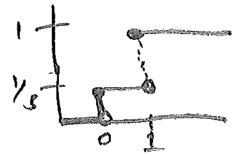
Suppose we have n RVs $X_1, X_2, \cdots, X_n$.

$\hat{F}_n$ is the n-sample empirical distribution function (EDF):

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)}{n}, \text{ where, } \mathbb{1}(X_i \leq x) := \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

Ex: Recall plot for Lotto Data in Lab 5.

Ex: For 3 Bernoulli data $x = (1,1,0)$



Computer generated Random Numbers

Pseudo-Random Numbers

Qn: How do we produce realisations from the most elementary Uniform(0,1) RV. X?
i.e., how to produce samples $(x_1, x_2, \cdots, x_n)$ from $X_1, X_2, \cdots, X_n \overset{IID}{\sim} \text{Uniform}(0,1)$?

Ans: Modular arithmetic & Number theory gives us Pseudo-Random Number Generators.

Qn: What can we do with such samples from Uniform(0,1) RV?

Ans: We can use them to sample (produce realisations) from other more complicated real-world random phenomenon, including:

(i) queues in Operations
(ii) Estimating missing data in Stats NZ's Accommodation occupancy survey.
(iii) help Chch Hospital manage critical care for pre-term babies.
(iv) help D.O.C. with marine bio-reserve management (minimize extinction probs of various marine organisms).
using Coalescent Theory in Stat. Genetics.

SAGE Simul

BUT ⋮ First we need to understand Modular Arithmetic

Modular Arithmetic (Arithmetic Modulo m)

Central theme in number Theory and crucial to machine-implementing objects in Probability Theory. The latter is necessary for computational statistical experiments.

— x —

Arithmetic modulo m is like usual arithmetic, except every time we add or muliply, we also divide by m and return the remainder

Ex: Let modulus $m = 12$, as in hours of analog clock.

we have:

$$8 + 6 = 14 \xrightarrow{\text{mod}(14,12)} = 2$$

Qn: "IF it is 8PM after chores and dinner today, what will be the time in 6 hours from then when I give this course its expected hours per week?"

Ans: 2 A.M.

Arithmetic with integers modulo m is well defined (will see in basic algebra course, but assume here) and has the following properties:

- $a+b = b+a$ (addition is commutative)
- $a \cdot b = b \cdot a$ (multiplication is commutative)
- $a \cdot (b+c) = a \cdot b + a \cdot c$ (distributive)
- If a is <u>coprime</u> to m (i.e., not divisible by any of the same primes) then there is a unique b (mod m) such that

  $a \cdot b = 1$

— x —

See SAGE Interact of W. Stein to appreciate +, • modulo m.

<u>10 minutes</u>

A Simple Pseudo-random number generator

Linear Congruential Generators (LCG)

Algorithm:

Input:
(i) modulus m, $0 < m$
(ii) multiplier a, $0 \leq a < m$
(iii) increment c, $0 \leq c < m$.
(iv) seed x_0, $0 \leq x_0 < m$
(v) number of desired pseudo-random numbers n

Output: $(x_0, x_1, \ldots, x_{n-1})$, The linear congruential sequence of length n.

```
for i = 1 to n-1 do
    x_i ← mod((a x_{i-1} + c), m) ...
end for
return (x_0, x_1, ..., x_{n-1})
```

"gets" (assignment in pseudo-code).

—— x ——

Examples:
(Re) Do all of examples of LCGs in Lab 6.