RANDOM VARIABLE: SOME DEFINITIONS

MRT2 §5.1	A variable (X) whose value is a number determined by the outcome of an experiment Can also be considered as a *function* that assigns a real number to each sample point i.e. $\{X(e_1), X(e_2), \ldots \}$
MM3 §4.3	A variable (X) whose value is a numerical outcome of a random phenomenon
WMS5 §2.11	A real-valued function for which the domain is a sample space. [Y: variable to be measured]

.

RANDOM VARIABLE: EXAMPLES

Experiment	Random Variable (and S ----> Y)

Toss 2 coins	Y = Number of "Heads"
	`Sample point          Y`
	`T T                   0`
	`T H                   1`
	`H T                   1`
	`H H                   2`

Turn over cards until get 1st ace	Y = Number of cards down to and including the first ace `Sample point          Y` A 1 $\bar{A}$ A 2 $\bar{A}$ $\bar{A}$ A 3 .. $\bar{A}$ $\bar{A}$ ..... $\bar{A}$ A 49

RANDOM VARIABLE: MORE EXAMPLES

Experiment	Random Variable (and S ----> Y)

Put 3 events in the order in which they occurred
 	
say w.l.o.g. correct order is Event1, Event2, Event3
	Y = Number in correct Position
	`Sample point              Y`
	`Event1 Event2 Event3      3`
	`Event1 Event3 Event2      1`
	`Event2 Event1 Event3      1`
	`Event2 Event3 Event1      0`
	`Event3 Event1 Event2      0`
	`Event3 Event2 Event1      1`

2 of 4 cans filled with water (W)
 	
Guess which 2 contain water
	$\boxed{W}$ $\boxed{}$ $\boxed{W}$ $\boxed{}$ `(w.l.o.g.)`
	Y = Number of correctly identified Cans
	`Sample point              Y`
	`X   X   -   -   1`
	`X   -   X   -   2`
	`X   -   -   X   1`
	`-   X   X   -   1`
	`-   X   -   X   0`
	`-   -   X   X   1`

Chose a word and measure how long it is, i.e., # characters (c's)
 	
probability distribution of Y depends on source (dictionary, article, ..)
	Y = Number of characters in word
	`Sample points             Y`
	`*                         1`
	`**                        2`
	`***                       3`
	`...`
	`...                       ?`

RANDOM VARIABLE: YET MORE EXAMPLES

Experiment	Random Variable (and S ----> R.V.)
Chose 100 single family dwellings. For each, measure how many 1000's of cubic metres of water is consumed in a year	T = Total amount of water consumed by 100 Sample points $\boxed{T}$ $\{C_1, C_2, \ldots, C_{100}\}$ $\sum_{i=1}^{i=100} C_i$ C_i = consumption for i-th randomly selected dwelling
Ditto	$\bar{C}$ = Mean amount of water consumed by 100 Sample points $\boxed{\bar{C}}$ $\{C_1, C_2, \ldots, C_{100}\}$ $\frac{T}{100}$
For a woman who has breast cancer surgery in Québec this year, measure duration of "workup"	Y = Duration (days) of workup Sample points Y $\frac{M}{S}$ 0 $\frac{M}{-S}$ 1 $\frac{M}{--S}$ 2 ... ... M: Mammogram S:Surgery

RANDOM VARIABLE: EVEN MORE EXAMPLES

Experiment	Random Variable (and S ----> Y)
Chose 100 single family dwellings. For each, measure how many 1000's of cubic metres of water consumed in a year	R.V. = Variability (SD) in amount of water consumed by 100 Sample points $\boxed{\text{Random Variable}}$ $\{C_1, C_2, \ldots, C_{100}\}$ $SD\{C_i\}$ SD = Standard Deviation
ditto	R.V. = Variability in amount of water consumed by 100 Sample points $\boxed{\text{Random Variable}}$ $\{C_1, C_2, \ldots, C_{100}\}$ $C_{[75]} / C_{[25]}$ $C_{[75]}$ = 75th in size (low-high); $C_{[25]}$ = 25th in size (low-high)
ditto	R.V. = Variability (CV*) in amount of water consumed by 100 Sample points $\boxed{\text{Random Variable}}$ $\{C_1, \ldots, C_{100}\}$ $\frac{SD[C's]}{mean[C's]}$ * Coefficient of Variation, (usually expressed as %)

Those interested may wish to consult "Lectures in Course 610 -- Nov 1999" accessible from J Hanley's web page. The notes in bold below are excerpts from them.

"Population" : Universe (conceptual or actual) of interest

Why a sample (rather than "Census")

Data not otherwise available

Don't need the precision of a census
(sometimes, a census can actually be less precise)

Reduced costs and time

Testing may be destructive
(In Quality Control, determinations on biological material, ..)
(blood samples, biopsies, ...)

$$ gained from 100% processing may be less than cost of the effort (In financial accounts, telephone billing,)

Can pay more attention to ascertainment and to quality of measurements

If use probability sampling, can measure the reliability of the sample estimates from the sample itself

Some Sampling Designs

SIMPLE RANDOM SAMPLE ("unrestricted random sample")
SYSTEMATIC (RANDOM) SAMPLE
STRATIFIED RANDOM SAMPLE
RATIO ESTIMATES FROM SRS'S
SINGLE-STAGE CLUSTER SAMPLE
MULTI-STAGE SAMPLE

SIMPLE RANDOM SAMPLING

Population contains N units

FORMALLY: SRS is a method of selecting n units out of N such that every one of the $^{N}C_{n}$ samples has an equal chance of being selected

IN PRACTICE, a SRS is drawn unit by unit:

Units are numbered 1 to N

Series of random numbers between 1 and N is drawn from, for example,

a hat, bowl, ...

(in succession, WITHOUT REPLACEMENT)

a table of ("pre-drawn") random numbers

(discarding any number previously drawn)

Units which bear these numbers constitute the sample

How Statistical Inference is Connected to Random Variables

Population of Size N;

Values of some characteristic :V_1, V_2, ... V_N.

Interest is in some function of V_1, V_2, ... V_N. (Parameter)

Measurement (y) on n randomly chosen individuals (SRS)

Order Chosen	Measurement (RV)	
1	y_1	[subscripts 1-n in sample
2	y_2	are different from the
..	..	subscripts 1-N in Population]
n	y_n	

Subscripts 1-n in sample are different from subscripts 1-N in Population [see diagram]

Note: Unless substantial, the Sampling Fraction (n/N) has little impact on reliability of estimate derived from sample.

e.g. N = 5, n = 2

1st chosen \ 2nd chosen	V_1	V_2	V_3	V_4	V_5
V_1		V_1 V_2	V_1 V_3	V_1 V_4	V_1 V_5
V_2	V_2 V_1		V_2 V_3	V_2 V_3	V_2 V_5
V_3	V_3 V_1	V_3 V_2		V_3 V_4	V_3 V_5
V_4	V_4 V_1	V_4 V_2	V_4 V_3		V_4 V_5
V_5	V_5 V_1	V_5 V_2	V_5 V_3	V_5 V_4	

For statistical purposes, since the order in which the units were selected usually doesn't contain extra information, there are 10, rather than 20, distinguishable pairs of V's.

One possible sample pair would be (shown shaded)

$y_1 = V_4$; $y_2 = V_3$

Statisticians often write upper case Y for "possible value" (the R.V.) and y for a specific *realization;* Thus, "Probability(Y = y)"