The *P*-Value and Probability Distributions

Please attach a cover page (-10%).

Problem 1 (30%):
Given a total of *N* items, *K* of which are type-1 items, the probability to observe a subset of *k* type-1 items in a subset of *n* items selected from the total *N* items without repetitions is given by the hypergeometric probability distribution

$$P(k;N,K,n) = \binom{N}{n}^{-1}\binom{K}{k}\binom{N-K}{n-k}.$$

The corresponding *p*-value is defined as the probability to observe the subset of *k* or more than *k* type-1 items in a subset of *n* items selected from the total *N* items without repetitions. Use combinatorial proofs to show that the three formulations below are all mathematically equivalent formulations of the *p*-value:

a) $\sum_{i=k}^{K}\binom{N}{n}^{-1}\binom{K}{i}\binom{N-K}{n-i}$;

b) $1-\sum_{i=0}^{k-1}\binom{N}{n}^{-1}\binom{K}{i}\binom{N-K}{n-i}$;

c) $\sum_{i=k}^{n}\binom{N}{n}^{-1}\binom{K}{i}\binom{N-K}{n-i}$.

Problem 2 (20%):
Which one of the three mathematically equivalent formulations of hypergeometric distribution-based *p*-value is computationally advantageous? Explain.

Hints: Consider the computing time, which is related to the number of operations that need to be carried out in the computation. Consider also computing precision, which is related to the precision in which numbers can be defined by using scientific notation with a pre-defined limited number of digits for the base and separately for the exponent. (Note that the IEEE standard for floating point arithmetic uses scientific notation for the representation of numbers).

```
(* BIOEN 3070/6070: Introduction to Statistics for Bioengineers *)
```

```
(* © Orly Alter 2016 *)
```

```
(* Assignment 3: The P-Value and Probability Distributions *)
```

```
(* General Commands *)
```

```
Clear["Global`*"]
```

```
? ScientificForm
```

ScientificForm[*expr*] prints with all real numbers in *expr* given in scientific notation.
ScientificForm[*expr*, *n*] prints with numbers given to *n*-digit precision. ≫

```
(* Problem 3 (15%): Define a function that calculates the hypergeometric distribution-
  based P-value by using the built-in binomial coefficient. Explain. *)
```

```
? :=
```

lhs := *rhs* assigns *rhs* to be the delayed value of *lhs*. *rhs* is maintained in
an unevaluated form. When *lhs* appears, it is replaced by *rhs*, evaluated afresh each time. ≫

```
? Binomial
```

Binomial[*n*, *m*] gives the binomial coefficient $\binom{n}{m}$. ≫

```
(* Problem 4 (15%): Define a function that calculates the hypergeometric distribution-
  based P-value by using the Mathematica built-
  in function of the hypergeometric distribution. Explain. *)
```

```
? HypergeometricDistribution
```

HypergeometricDistribution[*n*, n_{succ}, n_{tot}] represents a hypergeometric distribution. ≫

```
? PDF
```

PDF[*dist*, *x*] gives the probability density function for the symbolic distribution *dist* evaluated at *x*.
PDF[*dist*, {x_1, x_2, …}] gives the multivariate
probability density function for a symbolic distribution *dist* evaluated at {x_1, x_2, …}.
PDF[*dist*] gives the PDF as a pure function. ≫

```
? CDF
```

CDF[*dist*, *x*] gives the cumulative distribution function for the symbolic distribution *dist* evaluated at *x*.
CDF[*dist*, {x_1, x_2, …}] gives the multivariate cumulative
distribution function for the symbolic distribution *dist* evaluated at {x_1, x_2, …}.
CDF[*dist*] gives the CDF as a pure function. ≫

```
(* Problem 5 (20%): Reproduce the results of Table 1 in Lee et al., PLoS One (2012). *)
```

```
(* Problem 6 (50% Extra Credit): Explain, by using a combinatorial proof,
why the two P-values in the 2nd row of Table 1 in Lee et al.,
PLoS One (2012) are identical. *)

(* Problem 7 (50% Extra Credit):Reproduce the results of Table 1 in Tavazoie et al.,
Nat Genet (1998). *)
```