Measuring Risk using Distribution Functions

Alan Karr

April 25, 2006

Summary

- Risk of a masked data set *M* is measured by a distribution function $F_M(t)$
- Compare using stochastic ordering

$$M_1 \preceq M_2 \text{ if } F_{M_1}(t) \leq F_{M_2}(t) \text{ for all } t$$

  so still can define frontier
- Lots of
  - Tools (reliability, Chebyshev systems, …)
  - Relationships (e.g., integrals of convex functions)

Example 1

- r = record-level measure of risk (e.g., probability of re-identification)
- $F_M(t) = \frac{1}{n}\sum_{x \in M} 1\{r(x) \leq t\}$

Example 2

- K = intruder's knowledge = unobserved random variable
- P = agency prior on K
- $r(M, K)$ = risk of masked data M when intruder's knowledge is K
- $F_M(t) = \int 1\{r(M, k) \leq t\} dP(k)$

Example 3

- Multiple risk measures $R_1(\cdot), \ldots, R_J(\cdot)$
- Weights $w_1, \ldots, w_J$
- Consider $F_M = \sum_j w_j \varepsilon_{R_j(M)}$
- Many measures are of the form $\int f(t) dF_M(t)$
- Example: $f(t) = t \to R(M) = \sum w_j R_j(M)$

Questions

- With complex partial order, is frontier sufficiently small to make a difference?
- Are there corresponding distribution function-valued measures of utility?
  - Lame (at least, lamish) example: analogs of Example 3
- Are there computational tools?
  - Example: given (discrete) F and G, is there an efficient algorithm to determine whether $F(t) \leq G(t)$ for all t?