Privacy Worksheet

Question Sheet

Zak Varty

The questions on this sheet are designed to let you test your own understanding of the course content on privacy. Some questions will test basic notions, while others will encourage you to think more deeply about some of the concepts introduced this week.

Question 1: Definitions of Privacy Terms

Define the following terms:

1. A k-anonymous database
2. Data supression
3. Data generalisation
4. Background-knowledge attack
5. Homogeneity attack

Question 2: A Teacher's Gradebook (K-anonymity)

A teacher keeps the following gradebook of her student's grades. Missing values are indicated be an asterisk (*).

student_id	dob	gender	score	grade
2166433	2000-05-18	F	76	A
2124771	1999-12-29	M	63	B
2197243	2000-04-06	M	48	F
2135974	2000-05-29	*	70	A
2176719	2000-04-14	M	65	B

student_id	dob	gender	score	grade
2130069	1999-12-17	*	61	B
2199235	2000-05-20	F	63	B
2153174	2000-04-21	M	38	F
2199376	2000-04-23	F	54	C
2168752	1999-12-08	F	59	C

a) Why is this gradebook not 2-anonymous with respect to `student_id`, `dob`, `gender` and `grade`?

b) Suggest how suppression could be used to achieve 2-anonymity with respect to `student_id`, `dob`, `gender` and `grade` in this gradebook. Can you find multiple ways of doing this?

c) Suggest how data generalisation could be used to achieve 2-anonymity for student grades in this gradebook.

d) How do the your assumptions about the missing values in the gender attribute impact the k-anonymity of this data set?

e) Why might the teacher wish to use each of suppression and generalisation to anonymise this database?

f) Give examples of suppressions and generalisations of this data set that lead to 3- 4- and 5-anonymity.

Question 3: Estimating prevalence of extra-marital affairs (Randomised response)

A sociology researcher is interested in estimating p, the population proportion of people in monogamous relationships who are currently or have in the past engaged in an affair. The researcher has established that her sampling frame is representative of her target population and is able to sample individuals uniformly at random from her sampling frame.

When conducting the survey the researcher has ten statement cards, which she asks the participants to shuffle and select one to secretly read. Eight of these cards are printed with statement (A) "I am currently having or have previously had an affair", while the remaining two cards are printed with statement (B) "I have never had an affair". The respondent then tells the researcher whether the statement that they read was "TRUE" or "FALSE".

Let C be the event that an individual has cheated on their partner, A be the event of being shown card (A) and T be the event of responding "TRUE". For each event E denote its compliment by E^C.

a) Draw a probability tree to describe this randomised response survey.

b) Hence or otherwise calculate $\Pr(T)$ and $\Pr(T^C)$.

c) Using your answer to part (b), suggest a method of moments estimator $\hat{P}$ for the population proportion of unfaithful partners, p.

d) Calculate the expectation and variance of the estimator $\hat{P}$ when using a survey with n responses. Use these to show that $\hat{P}$ is a consistent estimator of p.

e) Based on a set of 120 survey responses in which 34 respondents replied "TRUE", calculate a point estimate $\hat{p}$ for p. Calculate the standard error of this estimate and give an approximate 95% confidence interval for p.

Question 4: The downside of randomised response

A survey company is deciding whether to use randomised or direct responses in a survey to estimate the proportion of voter who support a controversial news broadcaster.

In the suggested randomised response version of the survey, participants are asked to toss a fair coin before answering. If the coin lands heads up then they respond honestly. If the coin lands heads down then they toss the coin a second time. If on the second toss the coin lands heads up then they answer in support of the broadcaster and otherwise answer in opposition to the broadcaster.

a) Under this randomised response mechanism, derive an expression for the probability of a response against the broadcaster in terms of the true proportion of people p who are truly in favour the broadcaster.

b) Use your answer to the previous question to construct an estimator $\hat{P}$ for p under this randomised response scheme with n respondents. In your answer, let the response $Y_i = 1$ if the i^{th} respondent declares against the broadcaster and $Y_i = 0$ otherwise.

c) State the method of moments estimator $\tilde{P}$ under a direct response survey with responses $Z_1, \ldots, Z_n$ which take the value 1 when the respondent is against the broadcaster and the value 0 when the respondent is in support of the broadcaster.

d) Calculate and compare the expectation and variance of the estimators $\tilde{P}$ and $\hat{P}$.

e) Describe why randomised response is used for sensitive questions but not in all survey responses. You should link your answer to your findings in part (d).
