Lecture 2: Introduction to Counting
Modeling Social Data, Spring 2017
Columbia University

Husam Abdul-Kafi

January 27, 2017

1 Counting

Regular Statistics is basically just conditional probability: trying to find $P(y|x)$. If we have a small sample size, we're going to have large expected error.

The uncertainty we get is inversely related to the square root of the sample size:

$$\sqrt{\frac{p(1-p)}{N}}$$

The problem gets worse when we condition on more variables. For example, if we condition on multiple features such as $P(y|x_1, x_2, ...)$

Concrete example: 100 ages, 2 sexes, 5 races, 3 parties $\rightarrow$ 3000 groups. So, if we want to have a good enough N for each group, we need LOTS of samples.

One thing to do: bin features e.g. age between 18-24 is a group.

Another solution is to make a huge, non-representative study - e.g. poll all XBox users (mostly 9-17 year old guys). This causes a new problem: computation is very hard on such a large sample.

New Framework: split/apply/combine: split the data into groups, apply the computation (e.g. mean), combine the groups to get a sample-wide statistic.

Examples:

Bad Way:

1. Scan through X searching for a
2. Compile list of y-values
3. compute mean
4. repeat for b, c, ... etc.

Time: N*G (# of samples times # of groups)
Space: N

Better Way:

1. Scan through X
2. Compile list of y-values for each value of x
3. compute mean

Time: 2N
Space: 2N

Even Better:

1. Scan through X
2. Sum up total for each group and a counter
3. compute mean

Time: 2N
Space: 2G