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

Daniel Hong (sh3266)

January 20, 2017

1 Mentions Beyond Lecture Slide Content

First, we discuss conditional probability with an example of a demographic survey shown in *slide 2*. Note any comments or disclaimers denoted with asterisk * explaining any estimates for the research. When the margin of estimate is high, the uncertainty is also high.

$$p(y|x)$$

$$uncertainty = \sqrt{p(1-p)/N} \approx \sqrt{\sigma/N}$$

We see that as N (sample size) increases, uncertainty decreases.

Unlike in the real world, computers can compute experiments many times. For example, through a programmed simulation of coin flipping, we can visualize the outcome with a histogram and determine an empirical probability.

slide slide 5 states a problem with categorical binning (polling problem). Some bins are more populated than others; significant distribution cannot be derived from bins with small samples. Later slides suggest potential solutions. One approach is to develop a sophisticated model to generalize multiple bins and group them in order to have fewer bins.

Split/Apply/Combine techniques in R are introduced, but we will cover them in more detail later (demonstrated in class during lecture 3). It is always better to use apply functions than for loops, especially for large data. Iterating through a massive matrix is extremely expensive. Instead, R will split the data set, apply a desired function, such as mean(), to subsets, then combine the results to return.

Different ways of computing the mean and the running mean of a data set were demonstrated in class. Given a table consisting of two columns (bin type column and value column), populating each group first, as shown below, before computing the mean is more efficient than naively iterating through the table to find values corresponding to each bin to calculate the mean.

$$a : 2, 5, 6, 10, 3; b : 1, 66, 33; c : 4, 20, 6...etc.$$

We compute the running mean by calculating the sum of each bin as we iterate through the list and dividing by the total occurrence thus far.

$$mean : \bar{x}, var : \frac{1}{N} \sum (x_i - \bar{x})^2$$

Term "long tail" is a reference to the end tails of a distribution curve. This describes the phenomenon of there being few popular things and a lot of unpopular things.

Finally, we discuss some solutions to dealing with large data sets that exceed memory. Random sampling is deprecated because significant extreme points are likely to get filtered out. Loading data from the disk is too slow. One potential approach is to stream data in.