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

January 27, 2017

Notes from fhh2112

1 Counting

1.1 Estimating distributions

Estimating distributions is a practice in counting. Consider the election example in class. Let y be the candidate an individual voter supports. Let $x_1, \ldots, x_n$ be the characteristics of each individual voter. These characteristics may include age, gender, race, party affiliation, etc. We want to estimate the distribution of $P(y|x_1, \ldots, x_n)$, the probability that a characterized individual supports a certain candidate. In an ideal world, we would obtain data from the entire population and represent the distribution perfectly. However, we are unable to do so when we are limited to small sample sizes of data. For example, if we consider 100 age groups, 2 genders, 5 races, and 3 parties, we need to form estimates across $100 * 2 * 5 * 3 = 3000$ groups. We are unlikely to obtain reliable estimates considering typical surveys collect on the order of 1000 responses.

1.2 Solutions to data sparsity

There are a few ways we can solve the data sparsity issue.

First, we can sacrifice the granularity of our estimates by binning our range of characteristic combinations into fewer groups. For example, instead of considering individuals of each age between 1 and 100 separately, we can consider individuals as part of groups between ages 1-18, 18-29, 30-49, etc.

We can also use more complex models that can be trained on small samples and generalize well. For example, instead of counting, we may choose a regression model.

Another solution is to obtain more data, estimate the relative frequencies in the problem, and use these relative frequencies. The data here should either be representative of the population or appropriately adjusted.

1.3 Margin of error

How many responses do we need to estimate $P(y)$ with a 5% margin of error? The margin of error is equivalent to the standard error which is equivalent to the standard deviation in the sampling distribution of the mean. We used a Binomial random variable to represent the result of votes from n individuals. Each time we simulate this random variable, we obtain an estimate of the true mean. Simulating this random variable many times yields an estimated distribution of the true mean. The standard deviation of this distribution is the margin of error.

We can also solve this problem by hand rather than through simulation. We represent each individual's vote as a Bernoulli random variable. Remember that $Var(X) = p(1-p)$ for a Bernoulli random variable X. We are looking for a relationship between the number of responses, n, the probability of voting for a certain candidate, p, and the margin of error, s. Note that $Var(\alpha X) = \alpha^2 Var(X)$ and $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent.

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\Big(X_i\Big),\ X_i \sim Bernoulli(p)$$

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p)$$

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{np(1-p)}{n^2}$$

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{p(1-p)}{n}$$

$$s = \sqrt{\frac{p(1-p)}{n}}$$

$$n = \frac{p(1-p)}{s^2}$$

As in class, if $p = 0.5$ and $s = 0.05$, then we need $n = 100$ responses. Intuitively, if we want a smaller margin of error, we need more responses. If the probability becomes more fair and tends toward 50%, we need more responses.

1.4 Split, apply, combine

The split, apply, combine method is useful for computing group statistics when all of the data can be loaded into memory. See Figure 1.

1. Split or arrange all observations into their individual groups of interest
2. Apply or compute distributions and statistics within each group of interest
3. Combine or collect results across all groups of interest

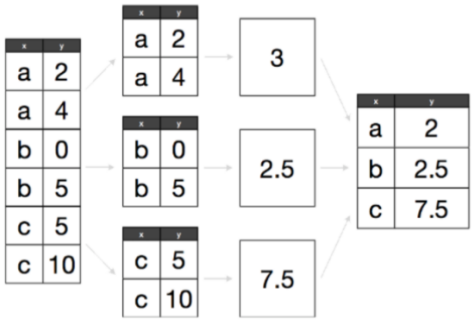


Figure 1. An example of the split, apply, combine method.

1.5 Computing per group average

Consider the following dataset. We wish to compute the average numerical value for each group, A, B, and C. Let N be the number of observations and G be the number of groups. Here, $N = 6$ and $G = 3$.

$$S = \{(A, 2), (A, 4), (B, 3), (B, 9), (C, 10), (C, 12)\}$$

The naive approach is to iterate through all observations, copy a list of values in group A, and iterate through all such group A values to compute an average. Repeat this process for the remaining groups. This method takes GN steps and N space.

A better approach is to iterate through all observations, copy values of each group into separate lists, and compute an average for each list. This method takes N steps and N space.

The best approach is to iterate through each observation and keep a running sum and count of observations in each group. After all observations have been seen, compute the average for each group using the respective running sum and count. This method takes N steps and $2G$ space.

1.6 Computing per group variance

Consider the dataset given previously. We now wish to compute the variance for each group. Define variance as $Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Note that the given formula requires that we know the average. The two pass method first passes through

the data to compute per group average and then passes through the data again to compute per group variance.

We can simplify the formula for variance to develop a one pass method.

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2\overline{x}}{n}\sum_{i=1}^{n}x_i + \overline{x}^2$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\overline{x}^2 \ + \overline{x}^2$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

However, note that for each value x_i, we compute x_i^2. This can cause value overflow, so this method should be avoided when working with large values.

2 Anatomy of the long tail

2.1 Problem description

The distribution of interests among individuals tends to exhibit a long tail. That is, there are few interests that are popular and many interests that are unpopular. The research paper presented in class attempts to answer where this long tail comes from on the subject of movies. Is it because some people only have mainstream interests while others only have niche interests? Or, does everyone have a blend of mainstream interests and niche interests with more variance among niche interests?

2.2 Reading data at scale

Large datasets may exceed available memory. Sampling may result in unreliable estimates for rare groups in the data. Randomly accessing data from disk is prohibitively slow. The best solution is to stream data. We read data one observation at a time and only store the state or values we require to calculate the statistics we are interested in.

2.3 Distribution of the data

Each entry in the dataset includes a user identifier, a movie identifier, and a rating. Thus, in order to compute statistics about each movie/rating, we first have to group entries by movie/rating. We compute these per group statistics using the methods we previously outlined. The data shows that most ratings, regardless of value, are given to the most popular movies. See Figure 2.

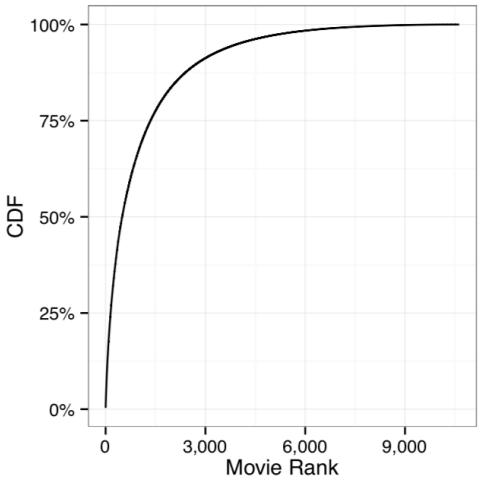


Figure 2. The relationship between number of movie ratings and movie rank.

We also compute each user's eccentricity, or the median rank of each user's rated movies. Note that using the median makes the measure more robust to outliers. A bimodal distribution of eccentricity would suggest two different types of people, those with mainstream interests and those with niche interests. A unimodal distribution of eccentricity would suggest that people generally have a mix of mainstream interests and niche interests. See Figure 3.

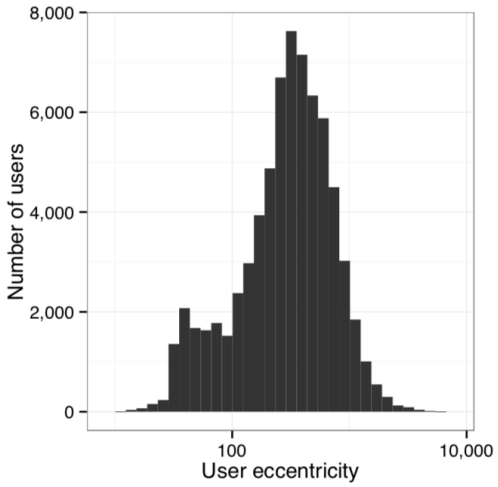


Figure 3. The distribution of user eccentricity.

2.4 Group by operation

The two charts below show the memory requirements for the group by operation under different scenarios. "Distributions" refers to whether we can store full histograms or not. "Statistics" refers to what kind of statistics we can compute. Let N be the number of observations. Let G be the number of distinct groups. Let V be the largest number of distinct values within a group. See Tables 1 and 2.

Memory	Scenario	Distributions	Statistics
N	Small dataset	Yes	General
VG	Small distributions	Yes	General
G	Small number of groups	No	Combinable
V	Small number of outcomes	No	No
1	Large number of groups and outcomes	No	No

Table 1. Group by operation information for arbitrary input data.

Memory	Scenario	Distributions	Statistics
N	Small dataset	Yes	General
VG	Small distributions	Yes	General
G	Small number of groups	No	Combinable
V	Small number of outcomes	Yes	General
1	Large number of groups and outcomes	No	Combinable

Table 2. Group by operation information for pre-grouped input data.

3 Demo

3.1 Command line tools

Become familiar with basic command line tools using the tutorials on the "Installing tools" post on the course website. We saw examples of **ls**, **pwd**, **man**, **echo**, **touch**, **rm**, **cat**, **head**, **uniq**, **tr**, **cut**, **more**, and **less** to name a few. The **man** page for each command provides a guide for usage and various option flags. You should understand basic pattern matching, variable creation, variable referencing, and piping. Also be aware that single quote strings are interpreted literally and double quote strings are partially interpreted.

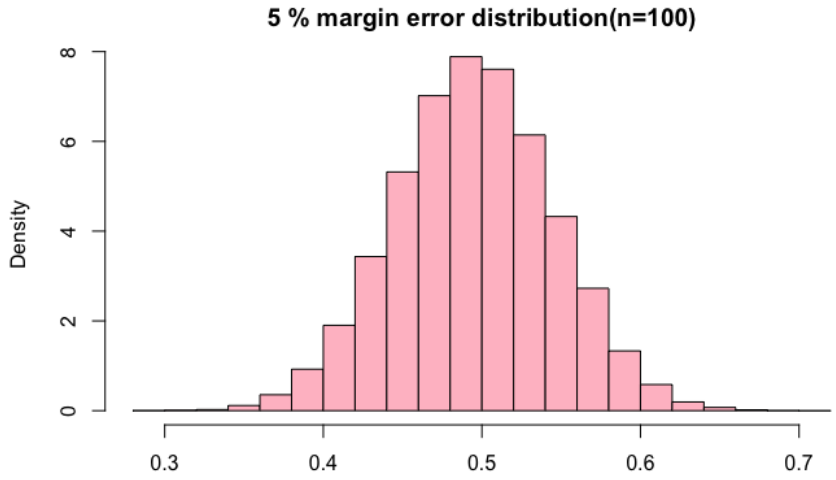
Notes from fh2386

1 Part 1

1.1 Conduct 5% margin of error

5% margin margin of error means 5% standard error or 5% standard deviation in sampling distribution for the mean.
Given 5% margin of error, 100 samples are needed. However,if the standard error is large such as 10%, less samples will need. Following R codes has provided to demonstrate.

```
# flip a coin n times
p=0.5
rbinom(n,1,p)
#estimate p by measuring the fraction of heads
mean(rbinom(n,1,p)
#repeat this 100,000 times
phat<-replicate(1e5,mean(rbinom(n,1,p))
#look at a histogram of the estimates
hist(phat)
#compute the standard deviation of the estimates
sd(phat)
```



1.2 Apply Central Limit Theorem to Binomial Distribution

$$Var(\frac{1}{n}\sum_{i=0}^{n} Xi) = \frac{1}{n^2} Var(\sum_{i=0}^{n} Xi) = \frac{1}{n^2}(\sum_{i=0}^{n} Var(Xi)) = \frac{p(1-p)}{n} \tag{1}$$

As a result, the number of sampling can be calculated by following way

$$S = SD(\frac{1}{n}\sum_{i=0}^{n} Xi) = \sqrt{\frac{p(1-p)}{n}} \tag{2}$$

$$n = \frac{p(1-p)}{S^2} \tag{3}$$

2 Part 2

2.1 Why Counting

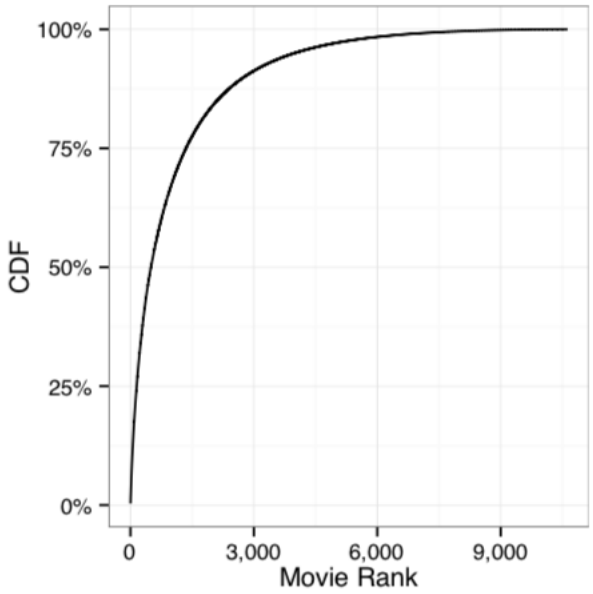
Traditionally difficult to obtain reliable estimates due to small sample sizes or sparsity.
for example, (100age x 2sex x 5race x 3 party)=3,000 groups As we know, 100 samples are needed to conduct 5% margin error. For 3,000 groups, we need 100x3,000=300,000 samples. Apparently, it is almost impossible to conduct such experiment.
Potential solution
1.Combine large observations into fewer groups,which can cause of missing other important info.
2.Come up with more sophisticated methods generated by small samples.
3.**Obtain larger data samples** by other ways and then count and divide to make estimates through its relative frequencies.
Good and bad of large data
Pros: move away from complicated and complex models generated by small samples to simpler models on large samples.
Cons: Computationally challenging at large data.

2.2 Learning to count

Claim: Solving the counting problem at scale enables you to investigate many interesting questions in the social sciences.
R functions:
Split: Arrange observations into groups of interest.
Apply: Compute distributions and statistics within each group
Combine: Collect results across groups.
Examples:

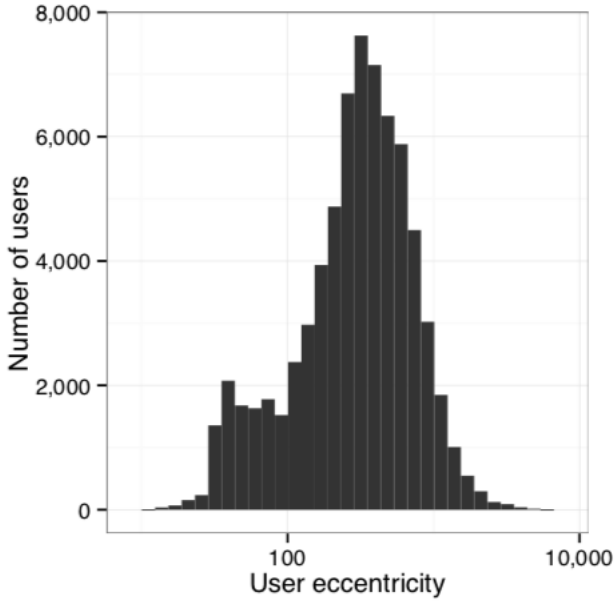
Group	Value
a	2
b	3
a	4
c	10
c	12
b	9

Several ways to approach the solution of finding the average per group.
1. First, find all a's and make a list of values such as (2+4)/2=3 for a's. Then, repeat for each group. With G*N steps and N space.
2.Run through observations and create list for each group.Then compute average each list. With N steps and N space.
3.Create list for each group and keep running average instead of list.
O(n* logn): time for binary search and sorting algorithms.



Movielens plots

According to the graph, majority rankings are from the most popular movies.



Compute median movie rank as user eccentricity. However, when encountering a large data set, median is nearly impossible to find.

Examples:

Ex1.

Take all your movies and look at popularity rank such as following (3,7,100,120,9800)

The median rank as well as eccentricity is 100.

Ex2

user	movie id	rating	ranking
1	37	1.5	8,000
1	43	4.5	10
1	2	3.0	.
.	.	.	.
.	.	.	.
.	.	.	.
2	37	.	8,000

3 Part 3

3.1 The group-by operation

Usually large data sets need large memory store.

For arbitrary input data:

Memory	Scenario	Distributions	Statistics
N	Small dataset	Yes	General
V*G	Small distributions	Yes	General
G	Small # groups	No	Combinable
V	Small # outcomes	No	No
1	Large # both	No	No

N = total number of observations
G = number of distinct groups
V = largest number of distinct values within group

Combinable: means that all data is needed to compute median.
Streaming: is required when full data-set exceeds available memory. It reads data one observation at a time, storing only needed state. It is also useful for computing a subset of within-group statistics with a limited memory footprint such as min, mean, variance but not median, requiring complete data set to compute.
Median rating are used by both Netflix and YouTube.
Mean rating is also utilized by YouTube, using streaming to compute combinable statistics.
uniq-c in R: c(a,b,a,a,b,c) to c(a,b,a,b,c). Only delete next repeated occurrence.

4 Part 4

4.1 Practice of Shell

Shell in Mac OS is called Terminal.
If you use Windows, you can try the built in bash/Ubuntu shell on Windows 10 or you can install it which includes bash and a terminal application by default. Linux also includes a working shell and terminal.
Useful commands in terminal: curl -o [short name] [URL].

Notes from tc2897

1 Introduction

Counting is incredibly important. 80% of Jake's research centers around knowing what to count and how to count it. Even estimating conditional probabilities is just counting!

2 Example Application

Finding the probability of supporting a particular candidate, given certain characteristics about the populace. For example:

$$\Pr(y|age) or \Pr(y|age, race) or \Pr(y|age, race, sex, party...)$$

To solve probabilities like these, we would simply count the number of people in that specific sub-group(people who meet the given characteristics) who support the candidate and divide by the total number of people in that subgroup. **It's just counting.**

2.1 Question

How many responses do we need to estimate p(y) with a 5% margin of error?

2.1.1 Aside

What do we mean by 5% margin of error?

- Mean won't usually vary by more than 5%
- Standard error will be 5% that is, standard deviation of mean of sampling distribution will be 5%

How would we do this practically? (illustrated through coin flips)
Through simulations.

Just simulate flipping a coin 100 times and calculate mean value. Repeat experiment 100000 times and see distribution of means. We notice that uncertainty is highest when the probability of outcomes is evenly split (like here, when p(h) = p(t) = 0.5) because distinctions are clearer in case of skewed probabilities, which makes the mean distribution clearer.

2.2 Relation between sample size and variance

The all-important question of how many responses do we really need to get a specific margin of error can be better answered using the formula derived here.

$$var(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n^2} var(\sum_{i=1}^{n} x_i) \tag{4}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} var(x_i) = \frac{1}{n^2} np(1-p) \tag{5}$$

$$= \frac{p(1-p)}{n} \tag{6}$$

Now looking at standard deviation s,

$$s = \sqrt{\frac{p(1-p)}{n}} \tag{7}$$

$$n = \frac{p(1-p)}{s^2} \tag{8}$$

Where n is the number of samples. Hence,

$$n\alpha\frac{1}{s^2} \tag{9}$$

2.3 The Challenge

If we wanted to split our conditional probabilities even into just 100 age groups, 2 sexes, 5 races and 3 parties- a very basic split-we'd end up with 3000 separate groups and assuming we need a 100 data points to estimate probabilities for each group, we'd need 300,000 responses. Unfortunately, the usual survey only gets a few hundred or few thousand respondents. Hence, the brute force approach is not useful.

2.3.1 Potential Solutions

- Use less groups, by binning small groups into bigger ones. This sacrifices granularity in data for precision
- Develop sophisticated methods that generalize well, like modeling
- Get more data!

3 Diving into Research

We then dove into a research paper on forecasting elections through non-representative sampling. The paper analyzed a poll taken through XBox, with 300,000+ responses. Note: Working with larger samples is mathematically easier and sounds great, but can get computationally taxing and at very large scales, even impossible. So need to always analyze the trade-off between simpler analysis and more computation.

4 Split-Apply-Combine

An effective way to count data is the split-apply-combine method, where you simply arrange observations into groups,calculate their statistics individually and then collect those results.

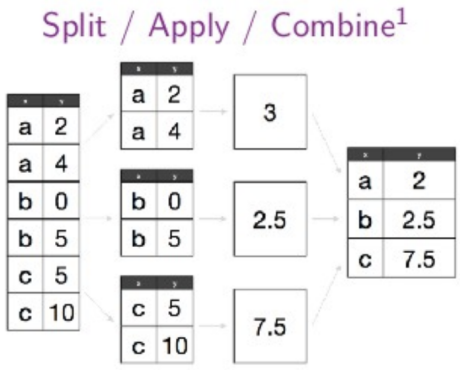


Figure 1: Illustration of Split-Apply-Combine From Lecture Slides, originally from The Split-Apply-Combine Strategy for Data Analysis by Hadley Wickam

5 Methods of Computation

We can often compute the same statistic in various different ways. Taking the example of computing averages, we saw how the computation space and time can vary greatly based on how we calculate the same statistic. Ways of computing the mean of each group:

1. The Dumbest: Find all entries associated with a group and make a list of their values. Then get the average by summing up this list. Repeat for each group.
   Time: G x N where N is number of observations and G is number of groups
   Space: N (worst case is all observations are in one group so list is N elements long)
2. Less Dumb: Create a list of each group. Run through observations, adding each observation to the list of the group to which it belongs. In the end, find the average for each list(group).
   Time: N
   Space: N
3. Least Dumb: Keep a running average for each group that is, store two numbers per group - the sum of all observations and the count of observations of each group.
   Time: N
   Space: 2G (approximately equal to G)

6 Research Paper on Long Tails

6.1 Overview

We then moved onto a research paper, "Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes" which explored whether the 'long tail' in inventory, or the massive list of rarely purchased goods arises from a few people having completely eclectic tastes, while others are completely mainstream or everyone having partially mainstream and partially eclectic tastes.

The two datasets used for this experiment were the Netflix and MovieLens data sets, which varied in both granularity and sample size.
A set of graphs was computed, which helped us visualize the distribution of data.

6.2 Visualizing the Data

We saw graphs for ratings vs number of ratings and a density vs rating graph to help us understand how the ratings were distributed.
We also saw a cumulative rating fraction graph that illustrated what fraction of the ratings were given to movies ranked in order of decreasing ratings. The steep decline in the slope beyond a certain point showed that most ratings were given to a select few movies, while most movies received very few ratings(the long tail).

We then delved into the users and their preferences, studying the median rank of users' rated movies. This was defined as the eccentricity of the user and showed that most users were partially mainstream and partially eccentric in their tastes.

6.3 Additional Insights

If size of data is larger than available memory, we can:

- Use random sampling of the data, but this results in unreliable estimates for rare groups
- Randomly access the data from the disk- gives us 1000x storage but takes 1000x the time.
- Using streaming - read one observation at a time, storing only what you need.

Streaming algorithm is important!

6.4 More Practical Examples

We also touched on other examples, like Youtube videos, and explored how we could handle data of that scale and what could and could not be computed from that on a computer with 8Gb Ram

For pre-grouped input data:

Memory	Scenario	Distributions	Statistics
N	Small dataset	Yes	General
V*G	Small distributions	Yes	General
G	Small # groups	No	Combinable
V	Small # outcomes	Yes	General
1	Large # both	No	Combinable

N = total number of observations
G = number of distinct groups
V = largest number of distinct values within group

For arbitrary input data:

Memory	Scenario	Distributions	Statistics
N	Small dataset	Yes	General
V*G	Small distributions	Yes	General
G	Small # groups	No	Combinable
V	Small # outcomes	No	No
1	Large # both	No	No

N = total number of observations
G = number of distinct groups
V = largest number of distinct values within group

Figure 2: From Lecture Slides

7 Introduction to Command Line

We dove into the command line operations. Jake mentioned that it's worthwhile to know how to work with the command line largely because how quickly one can explore data through it, without having to load it anywhere or read into anything.

- Moving around our disk - ls to list files in directory, pwd to find present working directory, moving into and out of directories with cd
- Distinguishing between the effects of using different types of quotes or no quotes
- Working with variables
- Working with pipes and understanding input/output streams as well as cat to display contents
- Running Scripts

7.1 Playing Around with CitiBike Data

We then conducted a basic exploratory data analysis of publicly available Citibike data to experiment with the operations we'd learnt.
We did some counting and pattern matching. We counted the number of distinct genders and looked up addresses similar to a particular pattern using grep among other things and even plotted a histogram on the command line interface!

8 Important Formulae

$$var(\lambda x) = \lambda^2 var(x) \tag{10}$$

$$var(x_1 + x_2) = var(x_1) + var(x_2) \tag{11}$$