Lecture 2: Introduction to Counting, Spring 2019
Columbia University

Fan Huang

February 1, 2019

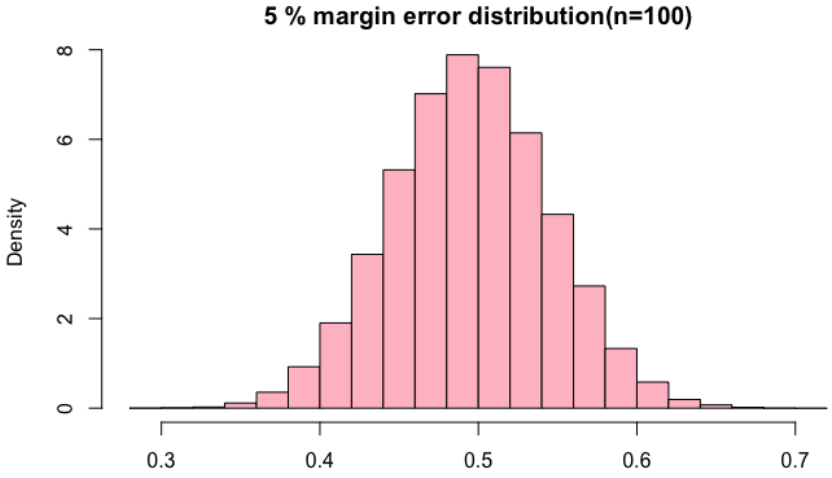
1 Part 1

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 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
p_hat- replicate(1e5,mean(rbinom(n,1,p))
# look at a histogram of the estimates
hist(p_ hat)
# compute the standard deviation of the estimates
sd(p_hat)
```



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 100×3,000=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.

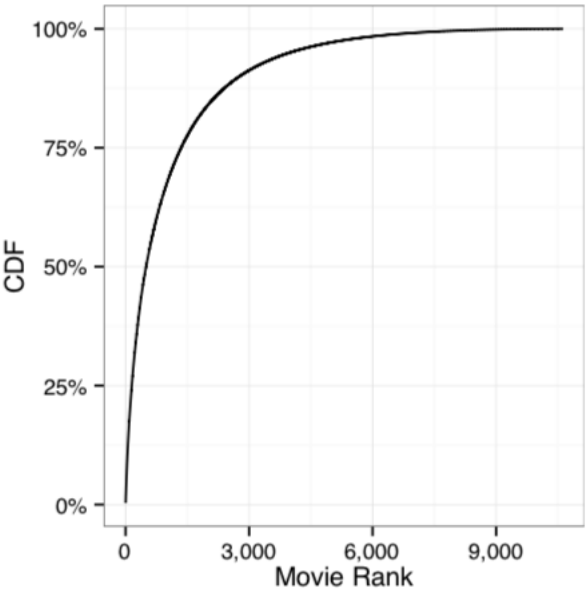
Examples:

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

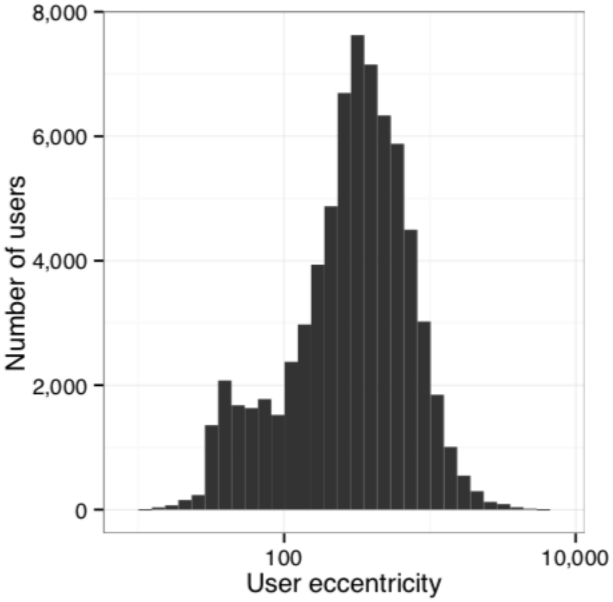
Several ways to approach the solution of finding the average per group.
1. First, find all a's and make a list of values such as (2+4)/2=3 for a's. Then, repeat for each group. With G*N steps and N space.
2.Run through observations and create list for each group.Then compute average each list. With N 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.

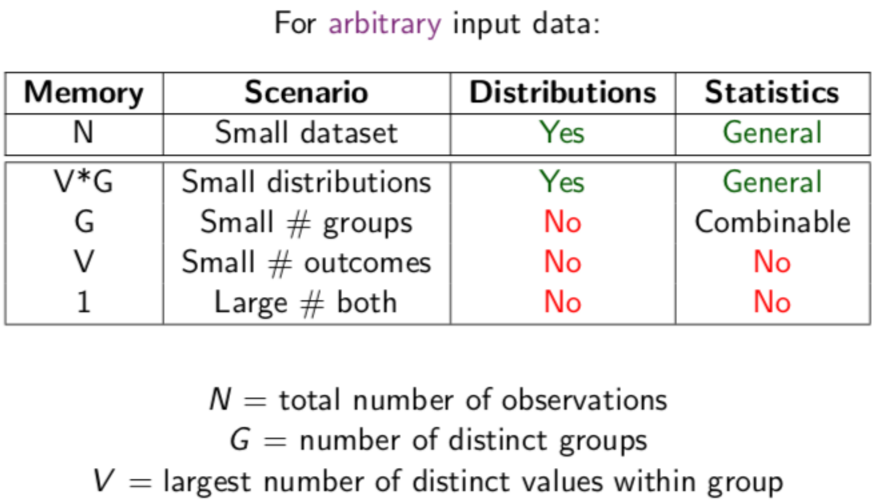
Ex2.

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].