



# **Length *Bias* in Sampling**

**May 10, 2006**

Contact Eric Granz (416.866.7074) for additional information

A. Persuade

B. Inform

C. Entertain

**How many people here today have been asked to consider making a presentation?**

The Data Mining User group needs presenters that will help fulfill a vision -- a vibrant, active user group

BUT.....

*Members are not getting in line to make presentations*

## My mission today

- Lead by example (if I can do it, anyone can do it)
- Demonstrate that you can have fun doing it

# Why *should I* present?

- If not you, then who? (referrals are welcome!)
- You might enjoy the experience
- You might impress others (it could be the ticket to your next job)
- Others may benefit from your knowledge -- what more effective way is there to ensure the message that you have to share is received by the person that needs to hear it?

**Who already knows something about length biased samples?**

# Why the concern about length bias?

- Length bias leads to biased estimates
- Biased estimates lead to wrong conclusions
- Wrong conclusions lead to less than optimal decisions

# What is length bias?

Is best understood with some examples

# Where is it observed?

- Common in Epidemiological studies  
i.e. studies related to the incidence, distribution, and control of disease in a population

# Example: Cancer Screening

- Because of variations in tumor growth rates, more of the cancers with long preclinical phases will be detected when a population is screened
- Cancers most likely to escape detection may be the very cancers that have the greatest likelihood of causing death

# Example: CRM

- Project "Beta"
  - Project financials were based on estimates of Accounts-per-Customer

# Methodology

1. Randomly sample signature cards
2. Check adjoining cards for additional accounts for same customer
3. Record results, repeat
4. Tally accounts and customers → Accounts per Customer

# Customer Base

Last Name	# of Accts
A	2
B	4
C	3
D	4
E	1
F	2
G	7

Last Name	# of Accts
H	1
I	5
J	1
K	1
L	2
M	3
N	6

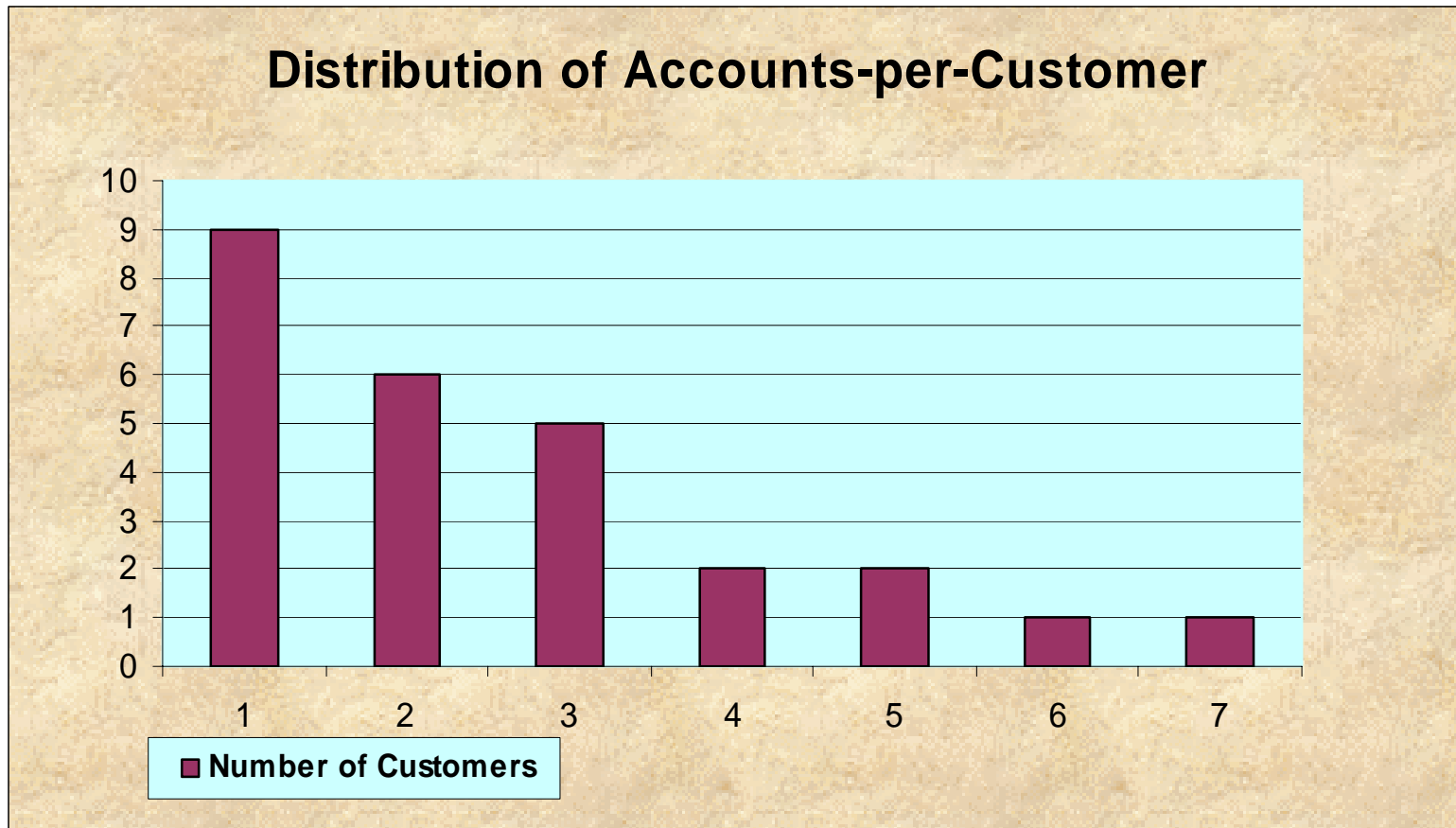
Last Name	# of Accts
O	5
P	2
Q	1
R	3
S	3
T	2

Last Name	# of Accts
U	1
V	3
W	2
X	1
Y	1
Z	1

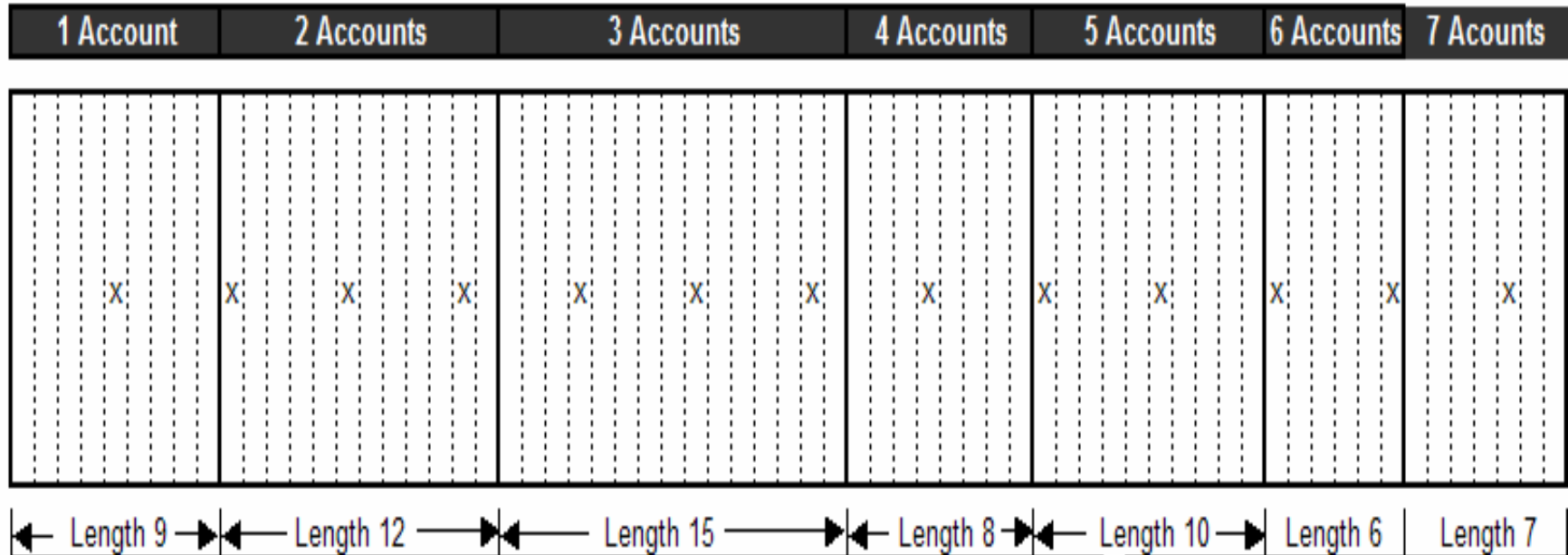
## Customer Base

Number of Accounts	Number of Customers	Total Number of Accounts	
1	9	9	
2	6	12	
3	5	15	
4	2	8	
5	2	10	
6	1	6	
7	1	7	
<b>Total</b>	<b>26</b>	<b>67</b>	<b>2.58</b>

## Customer Base



## Customer Base



x Denotes sample selection

# Customer Base

Number of Customers	Number of Accounts	Total Number of Accounts	
1	1	1	
3	2	6	
3	3	9	
1	4	4	
2	5	10	
2	6	12	
1	7	7	
<b>13</b>	<b>28</b>	<b>49</b>	<b>3.77</b>

# Other examples

- Ever wondered why it takes the bus so long to arrive?
- Ever wonder why your talkative friend always seems to be on the phone when you call him/her?

# Other examples

- How would you estimate the average term to maturity for a loan portfolio?

# What did we learn?

Why did a simple well-known challenge from the domain of epidemiological research escape the notice of almost everyone involved in a Bank-wide project?

## What did we learn?

I did, you can too

## What did we learn?

Seek

Discover

Propagate