

# Statistics 202: Statistical Aspects of Data Mining

**Professor David Mease**

**Tuesday, Thursday 9:00-10:15 AM Terman 156**

**Lecture 7 = Finish chapter 3 and start chapter 6**

## Agenda:

- 1) Reminder about midterm exam (July 26)**
- 2) Assign Chapter 6 homework (due **9AM** Tues)**
- 3) Lecture over rest of Chapter 3 (section 3.2)**
- 4) Begin lecturing over Chapter 6 (section 6.1)**

# **Announcement – Midterm Exam:**

**The midterm exam will be Thursday, July 26**

**The best thing will be to take it in the classroom (9:00-10:15 AM in Terman 156)**

**For remote students who absolutely can not come to the classroom that day please email me to confirm arrangements with SCPD**

**You are allowed one 8.5 x 11 inch sheet (front and back) for notes**

**No books or computers are allowed, but please bring a hand held calculator**

**The exam will cover the material that we covered in class from Chapters 1,2,3 and 6**

# Homework Assignment:

Chapter 3 Homework Part 2 and Chapter 6 Homework is due **9AM** Tuesday 7/24

Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

SCPD students may use email or fax or mail.

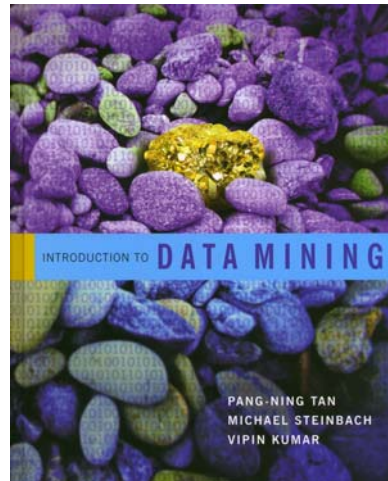
The assignment is posted at

<http://www.stats202.com/homework.html>

**Important:** If using email, please submit only a single file (word or pdf) with your name and chapters in the file name. Also, include your name on the first page.

# Introduction to Data Mining

by  
Tan, Steinbach, Kumar



## Chapter 3: Exploring Data


# Exploring Data

- We can explore data visually (using tables or graphs) or numerically (using summary statistics)
- Section 3.2 deals with summary statistics
- Section 3.3 deals with visualization
- We will begin with visualization
- Note that many of the techniques you use to explore data are also useful for presenting data

# Final Touches

- Many times plots are difficult to read or unattractive because people do not take the time to learn how to adjust default values for font size, font type, color schemes, margin size, plotting characters, etc.
- In R, the function `par()` controls a lot of these
- Also in R, the command `expression()` can produce subscripts and Greek letters in the text
  - example: `xlab=expression(alpha[1])`
- In Excel, it is often difficult to get exactly what you want, but you can usually improve upon the default values

# Exploring Data

- We can explore data visually (using tables or graphs) or numerically (using summary statistics)
- Section 3.2 deals with summary statistics
- Section 3.3 deals with visualization 
- We will begin with visualization
- Note that many of the techniques you use to explore data are also useful for presenting data

# Summary Statistics (Section 3.2, Page 98):

● You should be familiar with the following elementary summary statistics:

-Measures of Location: Percentiles (page 100)

Mean (page 101)

Median (page 101)

-Measures of Spread: Range (page 102)

Variance (page 103)

Standard Deviation (page 103)

Interquartile Range (page 103)

-Measures of

Association: Covariance (page 104)

Correlation (page 104)



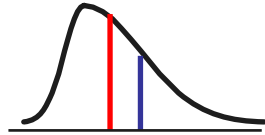
# Measures of Location

- **Terminology:** the “mean” is the average
- **Terminology:** the “median” is the 50<sup>th</sup> percentile
- **Your book classifies only the mean and median as measures of location but not percentiles**
- **More commonly, all three are thought of as measures of location and the mean and median are more specifically measures of center**
- **Terminology:** the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> quartiles are the 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentiles respectively

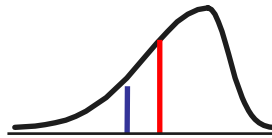
# Mean vs. Median

- While both are measures of center, the median is sometimes preferred over the mean because it is more *robust to outliers* (=extreme observations) and *skewness*

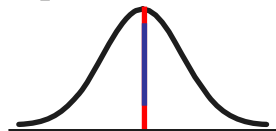
- If the data is *right-skewed*, the **mean** will be greater than the **median**



- If the data is *left-skewed*, the **mean** will be smaller than the **median**



- If the data is *symmetric*, the **mean** will be equal to the **median**





**Stay on top of the day's big story**  
**Get Breaking News Delivered right to your inbox**

**BREAKING NEWS** New Orleans mayor halts return of residents to city.



[MSNBC Home](#) » [Business](#) » [Personal Finance](#) » [America's Housing Craze](#)
Sponsored by 



**America's HOUSING CRAZE**  
HOT OR NOT? | CITY-BY-CITY DATA | FULL COVERAGE

- Business**
- [Stocks & Economy](#)
  - [Katrina's Cost](#)
  - [Real Estate](#)
  - [Personal Finance](#)
  - [U.S. Business](#)
  - [Intl Business](#)
  - [Oil & Energy](#)
  - [Automotive](#)
  - [Aviation](#)
  - [Food Inc.](#)
  - [CNBC TV](#)
  - [Forbes.com](#)
  - [BusinessWeek](#)
  - [Financial Times](#)

## San Francisco median home price tops \$600K

**Housing sales up even while area struggles to emerge from high-tech slump**

**REUTERS**   
 Updated: 6:14 p.m. ET July 19, 2005

**SAN FRANCISCO** - The median price of a home sold in the San Francisco Bay area topped \$600,000 for the first time in June amid a near record month of sales, analysts said Tuesday.

The median price paid for a home in the region rose to \$610,000 in June, up 18.2 percent from a year earlier and 2.5 percent from May, according to real estate information service DataQuick Information Systems.

**SPECIAL REPORT**



**AMERICA'S HOUSING CRAZE**

Related coverage	City-by-city data
<ul style="list-style-type: none"> <li>• Hot or not?</li> <li>• Affordability index</li> <li>• Mortgage calculator</li> <li>• Full coverage</li> </ul>	<ul style="list-style-type: none"> <li>• Midwest</li> <li>• Northeast</li> <li>• South</li> <li>• West</li> </ul>

# Measures of Spread:

- The *range* is the maximum minus the minimum. This is not robust and is extremely sensitive to outliers.

- The *variance* is 
$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

where  $n$  is the sample size and  $\bar{X}$  is the sample mean. This is also not very robust to outliers.

- The *standard deviation* is simply the square root of the variance. It is on the scale of the original data. It is roughly the average distance from the mean.

- The *interquartile range* is the 3<sup>rd</sup> quartile minus the 1<sup>st</sup> quartile. This is quite robust to outliers.

**In class exercise #22:**

**Compute the standard deviation for this data by hand:**

**2      10      22      43      18**

**Confirm that R and Excel give the same values.**

# Measures of Association:

- The *covariance* between  $x$  and  $y$  is defined as

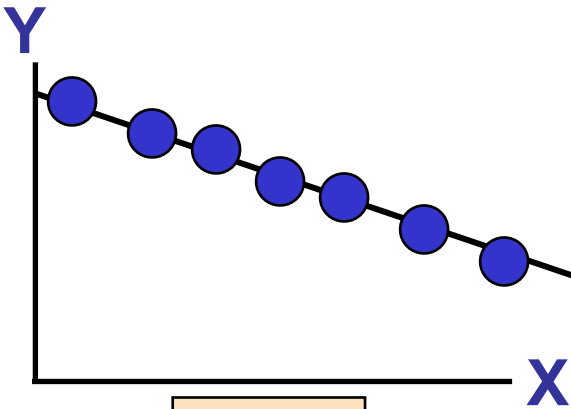
$$\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

where  $\bar{X}$  is the mean of  $x$  and  $\bar{Y}$  is the mean of  $y$  and  $n$  is the sample size. This will be positive if  $x$  and  $y$  have a positive relationship and negative if they have a negative relationship.

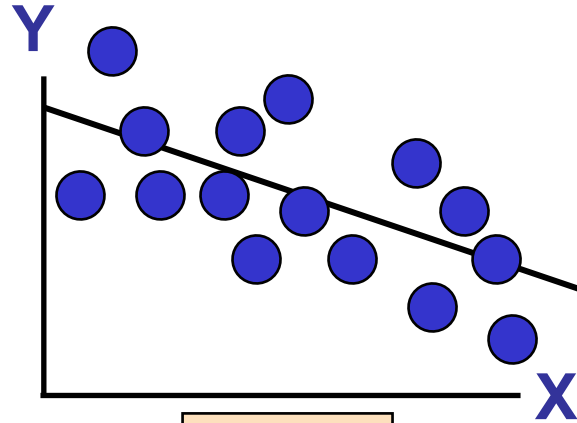
- The *correlation* is the covariance divided by the product of the two standard deviations. It will be between -1 and +1 inclusive. It is often denoted  $r$ . It is sometimes called the coefficient of correlation.

- These are both very sensitive to outliers.

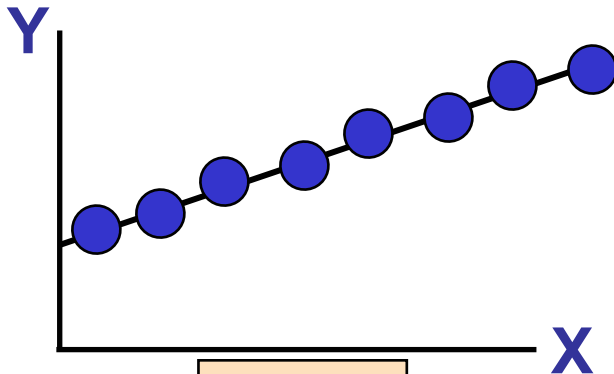
# Correlation ( $r$ ):



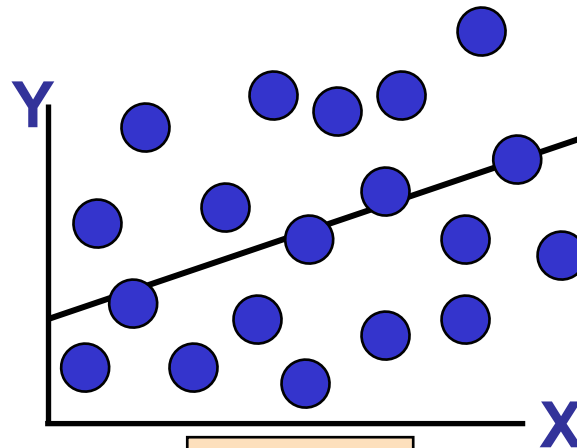
$$r = -1$$



$$r = -.6$$



$$r = +1$$



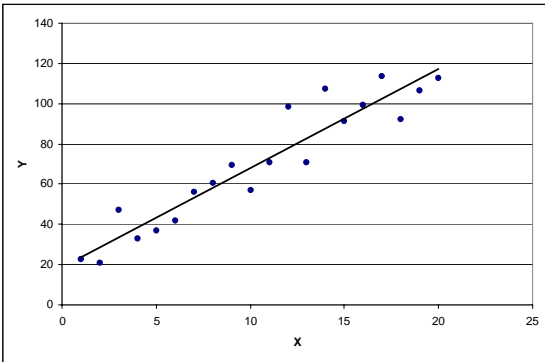
$$r = +.3$$

# In class exercise #23:

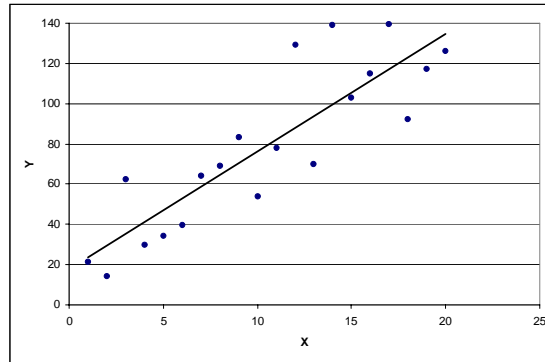
Match each plot with its correct coefficient of correlation.

Choices:  $r=-3.20$ ,  $r=-0.98$ ,  $r=0.86$ ,  $r=0.95$ ,  
 $r=1.20$ ,  $r=-0.96$ ,  $r=-0.40$

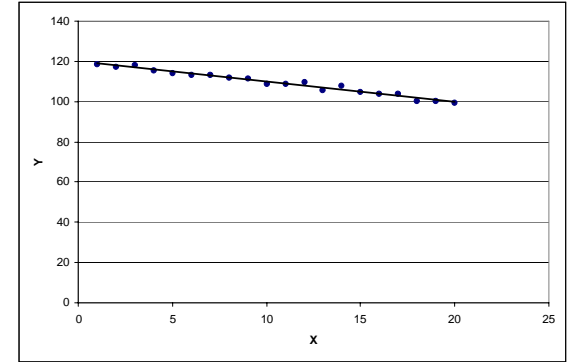
**A)**



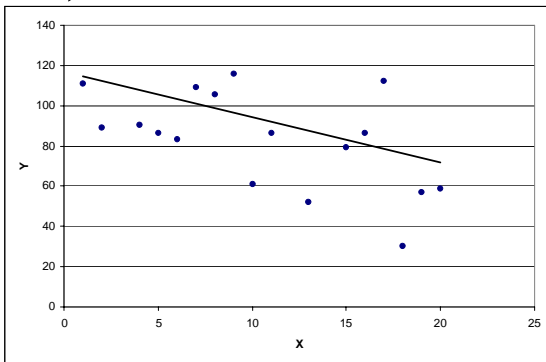
**B)**



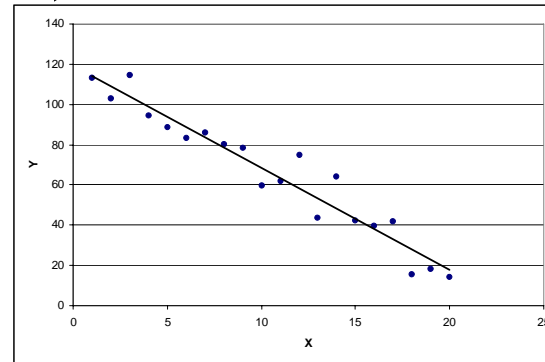
**C)**



**D)**



**E)**



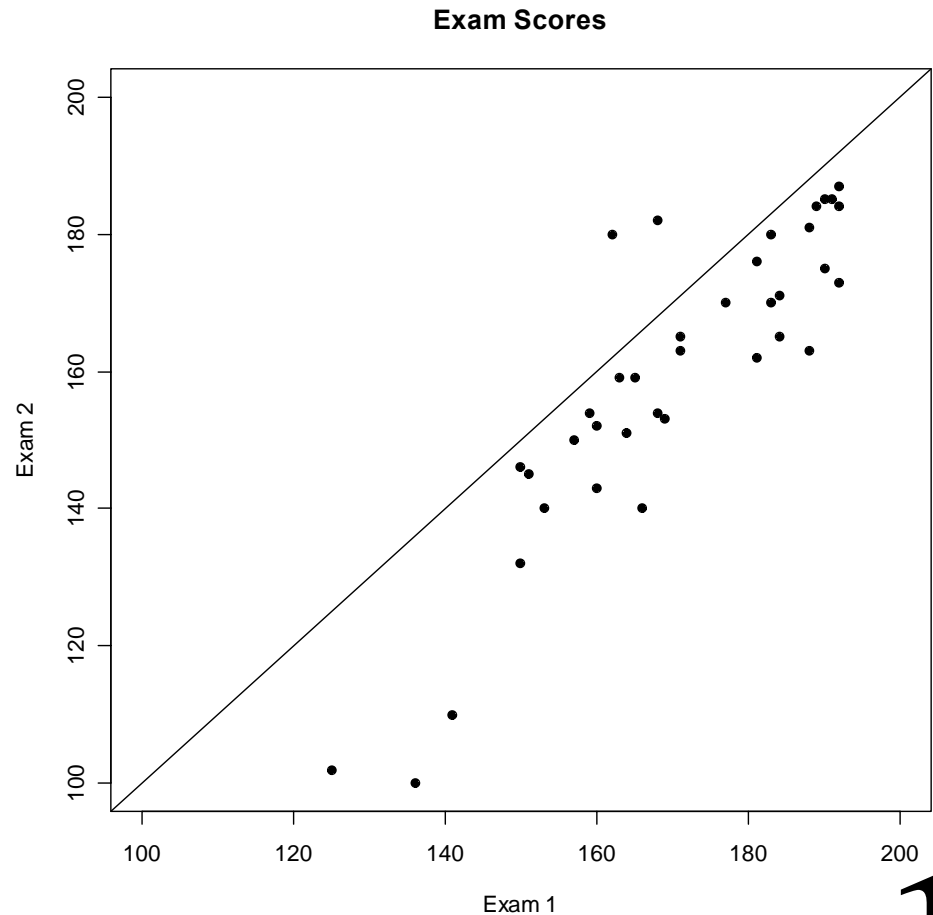


In class exercise #24:

Make two vectors of length 1,000,000 in R using `runif(1000000)` and compute the coefficient of correlation using `cor()`. Does the resulting value surprise you?

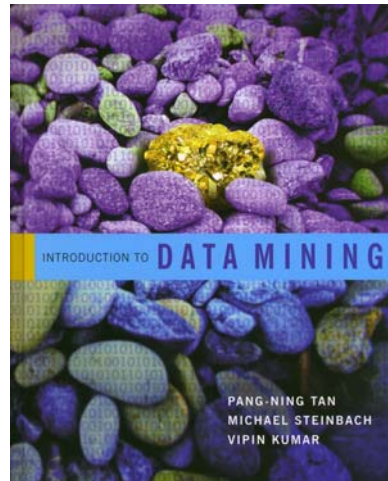
## In class exercise #25:

What value of  $r$  would you expect for the two exam scores in [www.stats202.com/exams\\_and\\_names.csv](http://www.stats202.com/exams_and_names.csv) which are plotted below. Compute the value to check your intuition.



# Introduction to Data Mining

by  
Tan, Steinbach, Kumar



## Chapter 6: Association Analysis

# What is Association Analysis:

- Association analysis uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction

- Examples:

{Diaper} → {Beer},  
{Milk, Bread} → {Eggs, Coke}  
{Beer, Bread} → {Milk}

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Implication means co-occurrence, not causality!

# Definitions:

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## ● Itemset

- A collection of one or more items
- Example: {Milk, Bread, Diaper}
- k-itemset = An itemset that contains k items

## ● Support count ( $\sigma$ )

- Frequency of occurrence of an itemset
- E.g.  $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

## ● Support

- Fraction of transactions that contain an itemset
- E.g.  $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$

## ● Frequent Itemset

- An itemset whose support is greater than or equal to a *minsup* threshold

# Another Definition:

## ● Association Rule

–An implication expression of the form  $X \rightarrow Y$ , where  $X$  and  $Y$  are itemsets

–Example:

$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

# Even More Definitions:

## ● Association Rule Evaluation Metrics

### –Support (s)

=Fraction of transactions that contain both X and Y

### –Confidence (c)

=Measures how often items in Y appear in transactions that contain X

## ● Example:

{Milk, Diaper}  $\Rightarrow$  Beer

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

## In class exercise #26:

Compute the support for itemsets {a}, {b, d}, and {a,b,d} by treating each transaction ID as a market basket.

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}



## In class exercise #27:

Use the results in the previous problem to compute the confidence for the association rules  $\{b, d\} \rightarrow \{a\}$  and  $\{a\} \rightarrow \{b, d\}$ . State what these values mean in plain English.

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

## In class exercise #28:

Compute the support for itemsets {a}, {b, d}, and {a,b,d} by treating each customer ID as a market basket.

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

## In class exercise #29:

Use the results in the previous problem to compute the confidence for the association rules  $\{b, d\} \rightarrow \{a\}$  and  $\{a\} \rightarrow \{b, d\}$ . State what these values mean in plain English.

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

In class exercise #30:

The data [www.stats202.com/more\\_stats202\\_logs.txt](http://www.stats202.com/more_stats202_logs.txt) contains access logs from May 7, 2007 to July 1, 2007. Treating each row as a "market basket" find the support and confidence for the rule

Mozilla/5.0 (compatible; Yahoo! Slurp;  
http://help.yahoo.com/help/us/ysearch/slurp)→  
74.6.19.105

# An Association Rule Mining Task:

- **Given a set of transactions  $T$ , find all rules having both**
  - **support  $\geq$  minsup threshold**
  - **confidence  $\geq$  minconf threshold**
  
- **Brute-force approach:**
  - **List all possible association rules**
  - **Compute the support and confidence for each rule**
  - **Prune rules that fail the minsup and minconf thresholds**
  - **Problem: this is computationally prohibitive!**

# The Support and Confidence Requirements can be Decoupled

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$  (s=0.4, c=0.67)  
 $\{\text{Milk, Beer}\} \rightarrow \{\text{Diaper}\}$  (s=0.4, c=1.0)  
 $\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}$  (s=0.4, c=0.67)  
 $\{\text{Beer}\} \rightarrow \{\text{Milk, Diaper}\}$  (s=0.4, c=0.67)  
 $\{\text{Diaper}\} \rightarrow \{\text{Milk, Beer}\}$  (s=0.4, c=0.5)  
 $\{\text{Milk}\} \rightarrow \{\text{Diaper, Beer}\}$  (s=0.4, c=0.5)

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

# Two Step Approach:

## 1) Frequent Itemset Generation

= Generate all itemsets whose support  $\geq$  minsup

## 2) Rule Generation

= Generate high confidence (confidence  $\geq$  minconf ) rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Note: Frequent itemset generation is still computationally expensive and your book discusses algorithms that can be used

### In class exercise #31:

Use the two step approach to generate all rules having support  $\geq .4$  and confidence  $\geq .6$  for the transactions below.

Table 6.2. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}