How Long is Long Enough?
Using Statistics to Determine Optimum Field Length
Suzanne Michelle, June 2009
Covering …

• A little background on my team’s project
• Considerations for summarizing data
• Data analysis
• Analysis and testing opportunities
Me?

• Worked with VisiCalc about 1981, told by my Graduate Supervisor to learn it and teach my fellow classmates how to use it
• Created floppy-disk sized set of relational Lotus123 spreadsheets for a city budget charge-back system in 1986
• Developed and managed various systems for Morgan Bank, W.R.Grace, Perdue Foods, and Hershey Foods, and also steel and roofing materials plants, among others.
• Began working for NYC Transit in 1994 on the Unified General Order System (UGOS)
Covering …

- **A little background on my team’s project**
- Considerations for summarizing data
- Data analysis
- Analysis and testing opportunities
General Orders?

- Run NYC Transit train operations
- Coordinate diversions *from* normal service
- Are planned 8 weeks to 5 years ahead of time depending on work / coordination involved
- Are visible to the riding public as service notice signs (for 1 or more actual operations)
- Affect / are affected by any service disruptions
- Reflect operation dependencies
- Interface with other systems (e.g., Accounting)
UGOS (You-Goes)

• A Calendar System of Work
• A Decision Support System, with History
• A Coordination Tool, between workgroups and departments
• A Reporting Tool
• Forms6i / 10g DB via MS TS / Citrix
Accounting ...

<table>
<thead>
<tr>
<th>Service Plans</th>
<th>EA0300_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP ID</td>
<td>2009RT7925</td>
</tr>
<tr>
<td>Div</td>
<td>TR1</td>
</tr>
<tr>
<td>Plan Status</td>
<td>PD, Draft</td>
</tr>
<tr>
<td>Requestor</td>
<td>TW1, Taylor, W.</td>
</tr>
<tr>
<td>Proj Mgr</td>
<td>TW1, Taylor, W.</td>
</tr>
<tr>
<td>Perm By</td>
<td>TM2, T.2</td>
</tr>
<tr>
<td>Plan Auth</td>
<td>PC2, Chute, Robert</td>
</tr>
<tr>
<td>Message</td>
<td>TE, Trains running express only</td>
</tr>
<tr>
<td>Work Desc</td>
<td>Replace rails, tie blocks and shim plates in 14 Street/Union Square station.</td>
</tr>
<tr>
<td>GO Start/Stop</td>
<td>06/30/2009, 07/13/2009</td>
</tr>
<tr>
<td>Reason</td>
<td>FM, Regular Maintenance</td>
</tr>
<tr>
<td>Options</td>
<td>NB, Both AET and SSS</td>
</tr>
<tr>
<td>FA SP/DR#</td>
<td></td>
</tr>
<tr>
<td>Work Desc</td>
<td>Replace rails, tie blocks and shim plates in 14 Street/Union Square station.</td>
</tr>
<tr>
<td>Mod</td>
<td>05/06/2009, ROCHGHE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Areas</th>
<th>Stations</th>
<th>Time Assn</th>
<th>Operations</th>
<th>Single Trk</th>
<th>Accounting</th>
<th>Works With</th>
<th>Desc/Comm</th>
<th>Covg/Annex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contract</th>
<th>% Alloc</th>
<th>Func No</th>
<th>% Alloc</th>
<th>Job No</th>
<th>% Alloc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RC No</th>
<th>% Alloc</th>
<th>Job No</th>
<th>% Alloc</th>
</tr>
</thead>
<tbody>
<tr>
<td>2561</td>
<td></td>
<td>06269</td>
<td></td>
</tr>
</tbody>
</table>
Operations ...

<table>
<thead>
<tr>
<th>Service Plans</th>
<th>EA0300_S</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP ID</td>
<td>2009BMT6347</td>
</tr>
<tr>
<td>Div</td>
<td>BMT</td>
</tr>
<tr>
<td>Plan Status</td>
<td>PF Finished</td>
</tr>
<tr>
<td>Wk</td>
<td>12</td>
</tr>
<tr>
<td>Seq</td>
<td>56</td>
</tr>
<tr>
<td>Yr</td>
<td>2009</td>
</tr>
<tr>
<td>Wk Grp Req No</td>
<td></td>
</tr>
<tr>
<td>Apprvt#</td>
<td>8-12-56</td>
</tr>
<tr>
<td>GO Start/Stop</td>
<td>03/21/2009 03/23/2009</td>
</tr>
<tr>
<td>Reason</td>
<td>TR Track Replacement</td>
</tr>
<tr>
<td>Options</td>
<td>NB Both AET and SSS</td>
</tr>
<tr>
<td>FA SP/DR#</td>
<td></td>
</tr>
<tr>
<td>2 BUSES</td>
<td>MOW Data GO-MOW +1</td>
</tr>
<tr>
<td>Plan Auth</td>
<td>ED1 Ertlitz, David 646-252-5524</td>
</tr>
<tr>
<td>Perf By</td>
<td>CD2 Track Construction Days 2 718-243-3747</td>
</tr>
<tr>
<td>Message</td>
<td>FB Full Shuttle Bus Service in effect</td>
</tr>
<tr>
<td>Code for Service Message: F9 for list</td>
<td></td>
</tr>
<tr>
<td>Mod</td>
<td>03/24/2009</td>
</tr>
<tr>
<td>JGOS</td>
<td></td>
</tr>
</tbody>
</table>

**Areas** | **Stations** | **Time Assn** | **Operations** | **Single Trk** | **Accounting** | **Works With** | **Desc/Cmnt** | **Covg/Annoc** |
|-----------|--------------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|

<table>
<thead>
<tr>
<th>Ln #</th>
<th>Route</th>
<th>Terminals</th>
<th>Signs</th>
<th>#Cars</th>
<th>Hdwy</th>
<th>Crew Req</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>L</td>
<td>8 Avenue</td>
<td>8/12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>Broadway Junction</td>
<td>Av/ Bway Junc</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Operation**
Shortlined to Broadway Junction. S/B operates normal via Q1 to n/o Broadway Junction then to Q2 at Broadway Junction and terminate. N/B in service on Q2 at Broadway Junction, and normal.

**Comments**
4th of 6 weekends.
Open doors onto island platform at Broadway Junction.
Works With (Siblings) ...
Cross-Referenced data ...
UGOS Star Schemas …

- **DR DATA**
  - Other Supporting Code Information
  - Accounting Codes
  - Track Codes

- **SP DATA**
  - People Codes
  - Help Module

- **WT DATA**
  - Boilerplate Codes

- **GO DATA**
  - RCC / ATS DATA

- **SIGN DATA**

- **GT Merge**

**General Universe of UGOS Data**
- DR - Diversion Requests
- SP - Service Plans
- GO - General Orders
- GT - GO Textual Information
- WT - Work Trains

Track, People, and Accounting Codes and the Help Module are supporting sub-modules. "Other Supporting Codes" includes things like Status, Type, Time and Limit Codes.

- Represents POSSIBLE Path (single-ended arrow)
- Represents DEFINITE Path (double-ended arrow, data flows both ways)
UGOS Data Flow and Ownership...

PSR Data

CPM and other Force Account Users

MTA Capital Construction Company

Operations Planning ...

Diversion Requests

Service Plans

MOW

Work Trains (Linden Shop / CPM)

Service Diversion Notices, MTA Website, Other Service Announcements

Sub 'C' GO Writers

RTO

ATS

RCC

Schedule Supplements

OMB (closes the data loop with ridership use vs. GO effect analysis)

Diversion Requestor Pool

Public

...Accounting Data Availability...
Covering …

• A little background on my team’s project
• **Considerations for summarizing data**
• Data analysis
• Analysis and testing opportunities
Three Questions ... 

• How to better summarize disparate data types and sources?
• How to allow on-the-fly data extraction (varying columns, criteria, purposes / audiences)?
• How to find / generate appropriate sample data for testing?
Our Solution: 2-fold ... VIEWS

- **“Count” views** across child tables
  - One record per PK
  - Columns for standard selections (type, status, date-related) and at least 1 column per child type counted (sometimes more – denormalized, to suite analysis needs)

- **“Aggregate” views**, one per select star
  - Use count views to analyze lengths, data combinations, to find test samples
  - Often done when design begins (but we did not do it – benefit now is we know the data very well)
Our Solution: 2-fold … WITH

• Using Count view summary data to select actual data …for function, procedure, and error handling purposes

• Using Oracle analytic functions, aggregate the data (one record per type of child record summarized) for a reporting / extract view
Create View VW_SPALLCNTS (CSPID, cPlanStatus, cYear, CNTTIMES, CNTEXCPS, CNTTEXT, CNTADJ, CNTWW, CNTDRs)
AS select CSPID, cPlanStatus, cYear,
    (select count(*) from SP3TA Three
     where Three.CSPID = Base.CSPID
     and Three.CDATETYPE <> 'AE'),
    (select count(*) from SP3TA Three
     where Three.CSPID = Base.CSPID
     and Three.CDATETYPE = 'AE'),
    (select count(*) from SP4DescFour
     where Four.CSPID = Base.CSPID
     and Four.CTEXTTYPE <> 'CN'),
    (select count(*) from SP5Adj Five
     where Five.CSPID = Base.CSPID),
    (select count(*) from SP6WWs Six
     where Six.CSPID = Base.CSPID),
    (select count(*) from DRXRef Sxtn
     where Sxtn.CSPID = Base.CSPID),
from SP0SERVICEPLANS Base;
In SQL Developer …
Covering …

- A little background on my team’s project
- Considerations for summarizing data
- **Data analysis**
- Analysis and testing opportunities
MAIN Question: How long should the Aggregate column be?

We looked at the summary data using Excel, across all columns, in each “All Count” view.

ROUGH Answer: 1000 characters was a nice round number …

• Too large for Accounting Data (but so what?)
• Too small for Text Data (but we knew we could not aggregate all text and a summary field existed in one child table)
• Just right for almost all other data, with error traps in place (but how to find the records that could cause error?)
with DATA as (select cyear, cspid, cntww,
  (case when cntww >50 then '10 = 51+'
         when cntww >45 then '09 = 46-50'
         when cntww >40 then '08 = 41-45'
         when cntww >35 then '07 = 36-40'
         when cntww >30 then '06 = 31-35'
         when cntww >25 then '05 = 26-30'
         when cntww >20 then '04 = 21-35'
         when cntww >15 then '03 = 16-20'
         when cntww >10 then '02 = 11-15'
         when cntww >5 then '01 = 6-10'
         when cntww >0 then '00 = 1- 5'
         else '00 = 0' end) as Catg
  from vw_SPALLCNTS)
select catg as grp,
       count(cspid) as cnt_sp, sum(cntww) as sum_ww
from DATA
  group by catg order by catg;
A Simple Bell Curve …

Σ (sigma) is a unit of deviation from the mean.
Sample Analysis … “Works With” …

<table>
<thead>
<tr>
<th>GRP</th>
<th>CNT_SP</th>
<th>SUM_WW</th>
<th>PerCnt</th>
<th>PerAccum</th>
<th>PerTot</th>
<th>PerAccTot</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 = 0</td>
<td>15393</td>
<td>0</td>
<td>34.44</td>
<td>34.44</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>00 = 1-5</td>
<td>12935</td>
<td>38206</td>
<td>28.94</td>
<td>63.38</td>
<td>13.15</td>
<td>13.15</td>
</tr>
<tr>
<td>01 = 6-10</td>
<td>8914</td>
<td>68775</td>
<td>19.94</td>
<td>83.33</td>
<td>23.67</td>
<td>36.82</td>
</tr>
<tr>
<td>02 = 11-15</td>
<td>3783</td>
<td>47925</td>
<td>8.46</td>
<td>91.79</td>
<td>16.50</td>
<td>53.32</td>
</tr>
<tr>
<td>03 = 16-20</td>
<td>1728</td>
<td>30712</td>
<td>3.87</td>
<td>95.66</td>
<td>10.57</td>
<td>63.89</td>
</tr>
<tr>
<td>04 = 21-35</td>
<td>737</td>
<td>16757</td>
<td>1.65</td>
<td>97.31</td>
<td>5.77</td>
<td>69.66</td>
</tr>
<tr>
<td>05 = 26-30</td>
<td>334</td>
<td>9284</td>
<td>0.75</td>
<td>98.05</td>
<td>3.20</td>
<td>72.86</td>
</tr>
<tr>
<td>06 = 31-35</td>
<td>182</td>
<td>5938</td>
<td>0.41</td>
<td>98.46</td>
<td>2.04</td>
<td>74.90</td>
</tr>
<tr>
<td>07 = 36-40</td>
<td>115</td>
<td>4361</td>
<td>0.26</td>
<td>98.72</td>
<td>1.50</td>
<td>76.40</td>
</tr>
<tr>
<td>08 = 41-45</td>
<td>83</td>
<td>3554</td>
<td>0.19</td>
<td>98.90</td>
<td>1.22</td>
<td>77.63</td>
</tr>
<tr>
<td>09 = 46-50</td>
<td>55</td>
<td>2634</td>
<td>0.12</td>
<td>99.03</td>
<td>0.91</td>
<td>78.53</td>
</tr>
<tr>
<td>10 = 51+</td>
<td>435</td>
<td>62367</td>
<td>0.97</td>
<td>100.00</td>
<td>21.47</td>
<td>100.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>44694</td>
<td>290513</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
So the REAL pattern is ...

with ALIAS as (select [identifier], [SumItem],
    (case when SumItem >50 then '10 = 51+'
         when SumItem >45 then '09 = 46-50'
         when SumItem >40 then '08 = 41-45'
         when SumItem >35 then '07 = 36-40'
         when SumItem >30 then '06 = 31-35'
         when SumItem >25 then '05 = 26-30'
         when SumItem >20 then '04 = 21-35'
         when SumItem >15 then '03 = 16-20'
         when SumItem >10 then '02 = 11-15'
         when SumItem >5  then '01 = 6-10'
         when SumItem >0  then '00 = 1- 5'
         else '00 = 0' end) as Category
    from CountView)

select Category,
    count(identifier) as CatCount, sum(SumItem) as CatSum
from ALIAS
    group by Category order by Category;
Covering …

• A little background on my team’s project
• Considerations for summarizing data
• Data analysis
• **Analysis and testing opportunities**
Examples and Uses …

• For UGOS, counts of text and counts of service adjustments are in newsletter article … analysis was done for each type of record where records were flattened into a single data entry.

• SOME types of data flatten in several ways, each important to a different user group or purpose, e.g., station codes versus station text.

• We’ve worked out a way to add in and aggregate new types of data as these become known, but length and “fit” analysis has to be done for each type of data, to avoid “overstuff” errors (our self-imposed 1,000 char limit).
Uses for Testing ...

```sql
with data as (select cSPID from vw_SPALLCNTS
  where (cntCont+CntFunc+CntJobs+CntRCNs) >3
  and cYear >'2006')
select cspid||' ': '||substr(get_Accounting(cSPID,'A'),1,100) as useful
  from data where rownum <2;

2007IRT8077: CNs:C-33293 JNs:15705 RNs:2832 FN:500

with data as (select cSPID from vw_SPALLCNTS
  where (cntCont+CntFunc+CntJobs+CntRCNs) =0
  and cYear >'2006')
select cspid||': '||substr(get_Acctg(cSPID,'A'),1,100) as useful
  from data where rownum <2;

2007IND7996: No Acctg Data

select GET_ACCOUNTING('2007IRT2524','P') as useful from dual;
  Bad Type Entered

select GET_ACCOUNTING('2007I24','A') as useful from dual;
  Bad Key Length
```
Questions?

Suzanne Michelle

Suzanne.Michelle@nyct.com