Forecasting Techniques

Date: 26 July 2017
Presenter: Chad Reynolds
Email: chad.reynolds@dau.mil
forecast

/ˈfɔrˌkast/  

verb

gerund or present participle: forecasting

predict or estimate (a future event or trend).
"rain is forecast for eastern Ohio"

synonyms: predict, prophesy, prognosticate, foretell, foresee, forewarn of
"they forecast record profits"
Agenda

- Basic Forecasting
- World of Big Data
- Inferential Statistics
BASIC FORECASTING
Basic Example

• Example: Using basic information about your lifestyle to predict your life expectancy
Collect Data

• Online surveys contain between 7 – 27 questions
  – Demographic: Birth Date, Sex
  – Lifestyle: Tobacco / Alcohol Consumption, Time Seated per Day
  – Health: BMI, Sun Exposure, Exercise, Brushing / Flossing
  – Diet: Vegetarian, Processed Food, Coffee Consumption
  – Environment: Air Pollution, City vs. Country Life
  – Family: Hereditary Conditions, Age of Grandparents
  – Disposition: Optimistic vs. Normal vs. Pessimistic
Results

Your Personal Death Date is...
Friday, September 09, 2067
Seconds left to live...
1,581,815,715

Your Armageddon's day is Aug 22 2067
Time remaining: 18290 days, 21 hours, 4 minutes, 4 seconds.

Your fateful day will come in year 2046
Lessons Learned

• Common questions between all of the surveys (cost drivers)
• Surveys with the fewer amount of questions (data points) had the largest amount of variability in the results
• Predicting with such specificity does a disservice; better to provide a likely range of results
WORLD OF BIG DATA
How Companies Learn Your Secrets

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
Step 1: Gather Consumer Data

• Keep tabs on everything you buy
  – Credit Card Purchases
  – Loyalty Programs
  – Coupon Use
  – Fill out a Survey
  – Mail in a Refund
  – Call Consumer Help Line
  – Open an Email Sent by the Company
  – Visit Company Website
Step 2: Gather Even More Data

- Age
- Marital Status & how many children
- Part of town where you live (How long it takes to drive to the store)
- Estimated Salary & job history
- Dates when you moved
- What credit cards you carry
- Magazines you read
- Websites you visit
- If you’ve ever declared bankruptcy or filed for divorce
- The year you bought your house
- Where you went to college
- What topics you talk about online
- Reading habits
- Charitable giving
- Political leanings
- Number of cars you own
- Etc.
Step 3: Data Analysis

• This is the difficult step
  – Economists
  – Mathematicians
  – Statisticians
  – Cost Estimators
  – Price Analysts

• Look for
  – Patterns in the data
  – Predictive indicators
  – Cost Drivers
Step 3: Data Analysis

• Pregnancy Indicators in the Target example:
  – Larger quantities of unscented lotion around the beginning of the second trimester
  – Large purchases of calcium, magnesium and zinc supplements in the first 20 weeks
  – Scent-free soap, extra large bags of cotton balls, hand sanitizers and washcloths close to their delivery date
Step 4: Create Predictive Model

• Model in the Target Example
  – Able to identify 25 products when analyzed together can be used to assign each shopper a “pregnancy prediction” score
  – In addition, can predict the due date within a very small window

• Hypothetical Example
  – Jenny Ward, age 23, living in Atlanta, GA
  – In March purchased:
    • Cocoa-butter lotion
    • A purse large enough to double as a diaper bag
    • Zinc and magnesium supplements
    • A bright blue rug
  – 87% chance she is pregnant and her delivery date is in August
Step 5: Test Model and Recalibrate

- Every model can be improved
- Requires analysis of several results over a period of time
Cost and Software Data Reports

Comprised of two components:

1. Contractor Cost Data Reporting (CCDR)
   - Focused on the collection of **actual** total contract costs.
   - Subdivided into standard categories for cost estimating purposes (e.g., by CBS, functional categories, and resource elements).

2. Software Resources Data Reporting (SRDR)
   - Collects software-specific data including software size, effort, activities and schedule data.
CSDR Policy - Current

• Reporting required on MDAP/MAIS contracts and subcontracts:
  – Over $50M.
  – Optional between $20M and $50M.

• OSD Approved Plans must be included within a contract Request for Proposal (RFP) before release.

• Defense Cost and Resource Center (DCARC):

Cost and Software Data Reporting (CSDR)

CSDR Overview and Policy
  • Introduction and Timeline
  • CSDR 5000.02
  • CSDR Manual
  • CSDR Requirements
  • Defense Federal Acquisition Regulations Supplement (DFARS)
  • WBS - MIL-STD-881C
  • Operating and Support Cost Estimating Guide

Plan Development and Contracting
  • WBS - MIL-STD-881C
  • Planning Forms, RDT, and Instructions
  • CSDR Post-Award Meeting Procedures
  • RFP Language
  • CDRL Examples
  • Cost Application - Plan Development
  • Supporting Documentation Checklist

CSDR Forms and Reporting
  • DEIs and Reporting Forms
  • Validation Process
  • Cost Application - Cost Reporting and Validation
  • CSDR Submit & Review Website

Systems and Applications
  • CSDR Planning and Execution Tool (CPEt)
  • CSDR Submit & Review Website

DACIMS
  • Defense Automated Cost Information Management System (DACIMS)
CSDR Policy - Update

- 2017 NDAA changed the requirement
  - From: CSDRs only required on MDAP/MAIS programs
    - Placed on specific Contracts and Subcontracts over $50M
  - To: all acquisition programs (regardless of ACAT level)
    - Lifecycle cost over $100M

<table>
<thead>
<tr>
<th>NDAA Section</th>
<th>Section 842 -- Amendments Relating to Independent Cost Estimating and Cost Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDAA Provisions</td>
<td>Inserts a new Section 2334(g) that gives DCAPE the statutory authority to issue the policies and procedures for collection of acquisition cost data. Requires that the PM and KO ensure that cost data is collected, in accordance with DCAPE policies, for all acquisition programs over $100M</td>
</tr>
<tr>
<td>Impact</td>
<td>CAPE will update DoD 5000.04-M-1 to reflect new statutory policies and procedures. Services will be responsible for managing collection of cost data for ACAT II and III programs. They are currently in the process of developing policies. These policies will be rolled up into DoD 5000.04- M-1.</td>
</tr>
</tbody>
</table>
Cost Assessment Data Enterprise (CADE)

- An OSD Cost Assessment and Program Evaluation (CAPE) initiative.
- Goal to increase analyst productivity and effectiveness by collecting, organizing and displaying data in an integrated single web-based application.
  - Increase analyst productivity
  - Improve data quality, reporting compliance and source data transparency

http://cade.osd.mil/
Analysis with Limited Data

- Without the use of large amounts of data, we are still able to predict future outcomes
- Inferential Statistics
  - Used to make inferences about the larger population based on the sample
Types of Inferential Statistics

- Make a probability statement about the population mean

**Confidence Intervals**

\[
\bar{y} \pm (t) \left( \frac{s}{\sqrt{n}} \right)
\]

- Make a probability statement about the next observation

**Prediction Intervals**

\[
\bar{y} \pm (t) \sqrt{s^2 + \frac{s^2}{n}}
\]
Election Polls

• Uses sampling to predict the outcome
• Usually states **percentage of likely voters** or an “approval rating”
Election Polls

• Uses sampling to predict the outcome
• Usually states percentage of likely voters or an “approval rating”
• Occasionally lists a margin of error
Election Polls

- Uses sampling to predict the outcome
- Usually states percentage of likely voters or an “approval rating”
- Occasionally lists a margin of error
- There is also an underlining level of confidence of the prediction that is hardly ever reported
Election Polls

- Uses sampling to predict the outcome
- Usually states percentage of likely voters or an “approval rating”
- Occasionally lists a margin of error
- There is also an underlining level of confidence of the prediction that is hardly ever reported
THE PROBABILITY THAT THE POPULATION MEAN WILL FALL BETWEEN A LOWER BOUND AND AN UPPER BOUND

Confidence Level

WE ARE 90% CONFIDENT THAT THE APPROVAL RATING OF CANDIDATE X IS BETWEEN 64% AND 72%

Population Mean

Lower Bound

Upper Bound
Confidence Interval Geometry

Assumes a normal distribution

Sample Mean

$\bar{y}$

Sample Standard Deviation

$s$
Confidence Interval Geometry

Assumes a normal distribution

LEVEL OF CONFIDENCE

90%

(1 - \(\alpha\))
Confidence Interval Geometry

Assumes a normal distribution

Lower Bound

\[ \alpha/2 \]

(1 - \( \alpha \))

Upper Bound

\[ \alpha/2 \]

LEVEL OF CONFIDENCE

90%

5%

5%

Level of Significance

10%
Confidence Interval Estimates

- We took what we believed to be a representative sample, and our best estimate of the true average (the population mean) is our sample mean.
- Our sample is one of many possible samples, and our sample mean is one of many possible sample means.
Making Probability Statements

- If we took the means of all these samples, the means would form a distribution.
- The mean of the resulting distribution would be the population mean (μ).
- The standard deviation of this distribution is \( \frac{s}{\sqrt{n}} \) also known as the standard error of the mean.
Confidence Interval Equation

- A probability statement about the population mean

\[
\bar{y} = \frac{\sum y_i}{n} \quad \text{Sample Mean}
\]

\[
s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} \quad \text{Sample Standard Deviation}
\]

\[
\bar{y} \pm (t) \left( \frac{s}{\sqrt{n}} \right)
\]

# of Observations
Student’s t-Distribution Table

- Use the table to find your t-value
- Dependent on your degrees of freedom \((n-1)\) and your level of significance \((\alpha)\)

<table>
<thead>
<tr>
<th>d.f.</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
</tr>
<tr>
<td>11</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
</tr>
<tr>
<td>12</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
</tr>
<tr>
<td>13</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
</tr>
</tbody>
</table>

90% Conf. 95% Conf. 98% Conf. 99% Conf.
Student’s t-Distribution Table

- Use the table to find your t-value
- Dependent on your degrees of freedom \((n-1)\) and your level of significance \((\alpha)\)

![Image of Student's t-Distribution Table]

<table>
<thead>
<tr>
<th>d.f.</th>
<th>.10</th>
<th>.05</th>
<th>.02</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
</tr>
<tr>
<td>2</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.225</td>
</tr>
<tr>
<td>3</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
</tr>
<tr>
<td>4</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
</tr>
<tr>
<td>5</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
</tr>
<tr>
<td>6</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
</tr>
<tr>
<td>7</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
</tr>
<tr>
<td>8</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
</tr>
<tr>
<td>9</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
</tr>
<tr>
<td>10</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
</tr>
<tr>
<td>11</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
</tr>
<tr>
<td>12</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
</tr>
<tr>
<td>13</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
</tr>
</tbody>
</table>
Confidence Interval Width

- A smaller/tight confidence interval requires:
  - Collection of more data \((n \text{ gets larger})\)
  - Less confidence in the prediction \((t\text{-value gets smaller})\)
  - Tightly grouped data set \((s \text{ gets smaller})\)

\[
\bar{y} \pm (t) \left( \frac{s}{\sqrt{n}} \right)
\]

- Sample Mean
- Sample Standard Deviation
- \# of Observations
- \(t\)-value
Sample Size Calculator

- First scenario:
  - I want to be 95% confident
  - I want my confidence interval (range) to be ± 3%
  - There are 20,000 voters in my district
  - Need to survey 1,013 people

- Second scenario:
  - I want to be 99% confident
  - I’ve only surveyed 150 people
  - There are 20,000 voters in my district
  - My current confidence interval is ± 10%
Election Polls

- Uses sampling to predict the outcome
- Usually states the percentage of likely voters or an “approval rating”
- Occasionally lists a margin of error
- There is also an underlining level of confidence of the prediction that is hardly ever reported

**Confidence Interval**

\[ \bar{y} \pm (t) \left( \frac{s}{\sqrt{n}} \right) \]
Summary

• The more accurate and timely data collected, the better the forecast
• Lack of data requires using tools such as inferential statistics
• Estimating an exact number / cost is a fool’s errand
• Need to quantify and bound the range of possible outcomes