

Data Mining Techniques
Chapter 5:
The Lure of Statistics: Data Mining Using Familiar Tools

Occam's razor	2
A look at data I	3
A look at data II	4
Measuring response I	5
Measuring response II	6
Measuring response III	7
Measuring response IV	8
Multiple comparisons	9
Chi-square test	10
Data mining and statistics (differences)	11

Occam's razor

- The simpler explanation is (usually) the preferable one.
- Statistical context: null hypothesis assumes that differences among observations are due simply to chance.
- p-value: probability the null hypothesis is true or probability the observed variation could be explained by chance alone;
 - technically incorrect definition, but OK for intuition.
- q-value: confidence (flip-side to p-value);
 - not a widely used term in my experience.

© Iain Pardoe, 2006

2 / 11

A look at data I

- Bar charts (qualitative or categorical data) and histograms (quantitative data);
 - note that the "histogram" on p127 is usually just called a bar chart; histograms generally graph quantitative data into intervals (bins).
- Time series: graph using line charts with time on horizontal axis.
- Standardized values: $Z = (Y - \text{mean}) / \text{std. dev.}$
 - e.g., if mean height is 5'9" and std. dev. is 3" how unusual is someone who is 6'? What about someone who is 6'6'?
 - example in book on daily service stops;
 - central limit theorem: sample mean is approximately normal;
 - example in book on stops if use weekly averages.

© Iain Pardoe, 2006

3 / 11

A look at data II

- Distributions: smoothed population histograms (e.g., normal bell-curve).
- From standardized values to probabilities:
 - normal (or t) tables;
 - one-tail or two-tail.
- Cross-tabulations for categorical data:
 - counts, proportions (e.g., Table 5.1, p136).
- Sample statistics:
 - range, mean, median, (mode), variance, standard deviation.
- Correlation and regression.

© Iain Pardoe, 2006

4 / 11

Measuring response I

- Is current (champion) or challenger offer better?
- Challenger offer sent to 100k prospects: 5% response.
- 95% confidence interval for population response rate: sample mean $\pm 1.96 \times$ std. error:
 - sample mean = 0.05, std. error = $\sqrt{(0.05 \times 0.95)/100000} = 0.000689$;
 - confidence interval is (0.0486, 0.0514).
- If 5.1% response for champion offer sent to 900k:
 - sample mean = 0.051, std. error = $\sqrt{(0.051 \times 0.949)/900000} = 0.000232$;
 - 95% confidence interval is (0.0505, 0.0515);
 - overlaps, so challenger offer no different.
- If 4.8% response for champion offer sent to 900k:
 - sample mean = 0.048, std. error = $\sqrt{(0.048 \times 0.952)/900000} = 0.000225$;
 - 95% confidence interval is (0.0476, 0.0484);
 - no overlap, so challenger offer better.

© Iain Pardoe, 2006

5 / 11

Measuring response II

- Alternative: rather than seeing if two intervals overlap, calculate interval for “difference of proportions” and see if it includes zero.
- Std. error formula on p143 incorrect; should be:

$$\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$$

- If 5.1% response for champion offer sent to 900k:
 - difference = -0.001 , std. error = $0.000727 = \sqrt{(0.051(0.949))/900000 + (0.05(0.95))/100000}$;
 - 95% confidence interval is $(-0.0024, 0.0004)$;
 - includes zero, so challenger offer no different.
- If 4.8% response for champion offer sent to 900k:
 - difference = 0.002 , std. error = $0.000725 = \sqrt{(0.048(0.952))/900000 + (0.05(0.95))/100000}$;
 - 95% confidence interval is $(0.0006, 0.0034)$;
 - excludes zero, so challenger offer better.

© Iain Pardoe, 2006

6 / 11

Measuring response III

- Use larger samples to have more confidence in results.
- If 4.8% response for champion offer sent to 50k:
 - difference = 0.002, std. error = $0.00118 = \sqrt{(0.048(0.952))/50000 + (0.05(0.95))/100000}$;
 - 95% confidence interval is $(-0.0003, 0.0043)$;
 - includes zero, so challenger offer no different;
 - significant difference not detected because sample size too small.
- What the confidence interval really means:
 - ensure random samples, beware bias.

© Iain Pardoe, 2006

7 / 11

Measuring response IV

- Size of test and control for an experiment (power calculations).
- Example: how large should N be in a difference of proportions test to detect a 0.002 response rate difference?
 - assume equal sample size in control and test;
 - Std. error formula on p147 incorrect; should be:

$$\frac{0.002}{1.96} \stackrel{\text{set}}{=} \sqrt{\frac{p(1-p)}{N} + \frac{(p+d)(1-p-d)}{N}}$$

- Final answer is incorrect too; it should be:

$$N = \frac{0.096796}{0.00102^2} = 92963.$$

© Iain Pardoe, 2006

8 / 11

Multiple comparisons

- Formulas up to now are based on only one comparison.
- Bonferroni correction divides sig. level by number of comparisons being made:
 - e.g., 2 tests each with sig. level 5% has an overall sig. level of approximately 10%;
 - practical implication: if doing k comparisons, do each test with significance level $5/k$ to ensure 5% overall.

© Iain Pardoe, 2006

9 / 11

Chi-square test

- Test for significant differences between row and column categories in a cross-tabulation.

	actual response			expected response		difference	
	yes	no	total	yes	no	yes	no
Example: champion	43,200	856,800	900,000	43,380	856,620	180	-180
challenger	5,000	95,000	100,000	4,820	95,180	-180	180
total	48,200	951,800	1,000,000	48,200	951,800		

- Test statistic (note no square root as on p151):

$$\chi^2 = \frac{180^2}{43380} + \frac{-180^2}{856620} + \frac{-180^2}{4820} + \frac{180^2}{95180} = 7.85.$$

- Significant difference since p-value = 0.005:
 - Excel: =CHIDIST(7.85, 1);
 - degrees of freedom = $(r - 1)(c - 1)$.
- Another example: regions/starts from table 5.1.

© Iain Pardoe, 2006

10 / 11

Data mining and statistics (differences)

- No measurement error in basic (DM) data (arguable).
- There is a lot of data:
 - more is better than less (in general);
 - computer power;
 - oversampling.
- Time dependency pops up everywhere.
- Experimentation is hard.
- Data is (sometimes) censored and truncated.

© Iain Pardoe, 2006

11 / 11