The DQM Package

Brian Connolly, Pawel Matykiewicz, Robert Faist,
and John Pestian

March 2014

Given a set of integer pairs, DQM calculates two measures:

1. the probability that the first elements in the pairs are similar to a given distribution,[1] and
2. the probability the ratio of the first and second elements of the pairs are similar, given that a success in the first element is greater than a success in the second.

Similarity is defined statistically: two sets of numbers are similar if they are derived from the same distribution. If it is more likely the distributions are different, DQM also calculates whether it is more likely the dis-similaries are due to general differences, or if there exists a single outlying number or pair. Further, if an outlier is shown to be responsible for the dis-similarity, DQM determines whether the outlying number or pair's ratio is an excess or deficit.

As discussed in [1], these types of similarity measures are useful in the context of detecting inter-hospital differences in patient database queries. For example, suppose one queries patients from several hospitals that satisfy certain criteria. If they operate in a similar manner with the same patient pool, then the query will identify more or less the same percentage of patients from the hospital's total population. The percentages will not be exactly the same, as the number of patients are subject to statistical fluctuations. However, significant differences can indicate differences in quality of care, reporting of care, or even trivial differences in database structure.

The measures of similarity described above can be applied directly to detecting differences in the number of queried patients and encounters. If patient counts are the first elements in the pairs, then calculating the similarity in the percentages of patients pulled from the total populations is the same as calculating the similarity of the distribution of queried patients across hospitals to the distribution of patients before the query.

Also, if the query returns total encounters (i.e., total patient contacts) over the number of patients, then under the assumption the hospitals operate similarly, the hospitals will have similar encounters to patients ratios. We could

[1] In the document, the word "distribution" is synonymous with "probability mass function".

then invoke calculation #2 to evaluate the similarities of the ratios given the probability of an "encounter" always be greater than a "patient". (There cannot be more patients than patient encounters.)

The manual is organized as follows. Section 1 contains a guide to quickly begin calculating numbers with DQM, along with a detailed description of the input/output. Section 2 gives a description of the subroutines used in the executable for those interested in implementing them elsewhere. Appendices A.1 and A.2 contain derivations of the similarity measures.

Any questions or comments should be directed to Brian Connolly,
brian.connolly@cchmc.org

1 Quick Start

We then explain, step-by-step, how to run dqm.pl and interpret its output.

1. The DQM package consists of a single perl script, dqm.pl, which has a single dependency: the Statistics::ROC package. The ROC package can be found through CPAN (http://search.cpan.org/ hakestler/Statistics-ROC-0.04/lib/Statistics/ROC.pm), and installed by typing:

   ```
   sudo cpanm Statistics::ROC
   sudo cpanm --look Statistics::ROC
   perl Makefile.PL
   make install
   exit
   ```

2. Test that dqm.pl executes properly by typing the command

   ```
   ./dqm.pl
   ```

   which should render a description of the expected input:

   ```
   dqm.pl [patient counts 1] [patient counts 2] ... [patient counts N
       ] [encounter counts 1] [encounter counts 2] ... [encounter
       counts N] ... [gold standard counts 1] [gold standard counts
       2] ... [gold standard counts N]
   ```

   With this out the the way, we are ready to run dqm.pl on data.

3. DQM calculates two things: (1) if a set of counts are consistent with a given ratio and (2) if a set of ratios are consistent with one another. These calculations are made under two assumptions: (1) the counts fluctuate according to Poisson statistics and (2) that the ratios are generally less than 1:1. For example, suppose the patient counts for three hospitals A, B and C are 85, 81 and 102, and the total patient encounters are 181, 285 and 152, respectively. We further suppose the queries will be made with patient pools of 9000, 8000, and 10,000 patients, respectively. That is, the ratio of patient counts from the hospitals are expected to be 9:8:10 if they operate similarly. DQM returns (1) the probability of the query returning patients in hospitals A, B and C with probabilties 9/(9+8+10),

8/(9+8+10), and 10/(9+8+10), respectively, and (2) the probability that 181/85, 285/81, and 152/102 are consistent with one another. Typing

```
./dqm.pl 85 81 102 181 285 152 9 8 10
```

returns

```
PATIENT COUNTS: [85,81,102] ENCOUNTER COUNTS: [181,285,152] GOLD
    STANDARD: [9,8,10]
P( Patients Same | Patients=[85,81,102] ) = 0.96240 > 0.5, they're
     similar!
P( Encounters/Patients Same | Patients=[85,81,102], Encounters
    =[181,285,152] ) = 0.00050, all Encounters/Patients ratios are
     different from each other
```

Note that the number of hospitals/PATIENT COUNTS/ENCOUNTER COUNTS/GOLD STANDARD counts are deduced by the number of arguments; if the number of arguments is not divisible by 3 then dqm.pl returns an error.

A description of the output follows. The first line simply re-iterates the input. PATIENT COUNTS are the patient counts for the query, ENCOUNTER COUNTS are the number of encounters or patient contacts across all patients returned by the query, and GOLD STANDARD defines the expected ratio of patient counts that should be present across the hospitals. Note that the normalization for the GOLD STANDARD is arbitrary;

```
./dqm.pl 85 81 102 181 285 152 90 80 100
```

dqm.pl would return the same results.

The second line is the probability that the PATIENT COUNTS are derived from a parent distribution defined by the GOLD STANDARD. Here, the probability they are similar is ≈ 0.96, which indicates similarity as it is greater than 50%.

The third line is the probability that the Encounters/Patients ratios are the same, given the probability of an 'encounter' is greater than the probability of a 'patient'. The low probability, 0.00050 indicates the ratios 181/85=2.13, 285/81=3.52 and 152/102=1.50 are dis-similar. In other words, it is unlikely that the pairs of numbers are generated with the same relative (binomial) probabilties. The output also indicates the ratios are probably all different. That is, there is no one hospital is responsible for the dis-similarity – there is no 'outlying' hospital.

Note if any of the ENCOUNTER COUNTS are negative, then only the PATIENT COUNTS similarity is calculated. If any of the GOLD STANDARD values are negative then only the Encounters/Patients similarity is calculated.

2 DQM Subroutines

The following is a detailed description of the subroutines within dqm.pl.

data_quality_measure_with_patients_only_pathological
([vector of patient counts], [gold standard vector of patient counts])

The input is two vectors of counts. The output is the probability that the first vector is derived from the distribution defined by the second. Note the second vector has an arbitrary normalization.

which_patient_is_pathological([vector of patient counts], [gold standard vector of patient counts])

If it is determined that it is more probable than not that the counts are derived from a different distribution than the one defined by the gold standard (i.e., data_quality_measure_with_patients_only_pathological returns a number less than 1/2), this function is called to determine if the counts are generally different from the gold standard, or if there is a single outlying count. The output is then a two-element array: the outlying count and whether the count is an excess or deficit. The first number is negative if it is more probable the distribution of counts is generally different from the gold standard; otherwise, the function returns the outlier (the output is 0 if it is the first count, 1 if it is the second count, etc.). The second number is 1 or 0 if there is a deficit or an excess of counts compared to the gold standard, respectively.

data_quality_measure_with_patients_and_encounters_pathological
([vector of patient counts], [vector of encounter counts])

The input is two vectors of counts. The output is the probability that the ratios of their respective elements are derived from the same ratio (Eqn. 5), under the constraint that success in an element in the second vector is more probable than success in the corresponding element in the first vector.

which_patients_and_encounters_is_pathological
([vector of patient counts], [vector of encounter counts])

If it is determined that it is more probable than not that the count ratios are derived from a different ratios (i.e., data_quality_measure_with_patients_and_encounters_pathological returns a number less than 1/2), this function is called to determine if the ratios are generally different, or if there is a single outlying ratio. The output is a two-element array: the outlying ratio and whether the ratio is an excess or deficit relative to the other ratios. The first number is negative if all the ratios are different; otherwise, the function returns which ratio is the outlier (the output is 0 if it is the first ratio, 1 if it is the second ratio, etc.). The second number is 1 if the ratio is a deficit and 0 if it is an excess.

A Appendix

DQM computes two probabilities:

1. the probability that a set of counts $\{A_i\}$ are derived from multinomial probabilities $\{q_i\}$, and

2. the probability that the pairs within a paired set of counts, $\{(A_i, B_i)\}$, are all derived from the same binomial probabilities, given the binomial probability associated with B_i is greater than the probability associated with A_i.

The following sections show the derivation of these probabilities.

A.1 The Probability a Set of Counts are Derived from a Set of Multinomial Probabilities

Suppose N categories (e.g., hospitals). Further suppose the probability of success for the i^{th} category is hypothesized to be q_i, while A_i is the measured number of successes. We then want to calculate the probability that $\{q_i\}$ produced $\{A_i\}$.

To evaluate this probability, we must not only know $\{q_i\}$, but what the probability distribution of successes might look like if it were not $\{q_i\}$..

Iin this latter case, we suppose two possibilities. The distribution is either derived from a distribution, $\{p_i\} \neq \{q_i\}$, or a distribution that satisfies, for some constant C, $\forall i : p_i = Cq_i$ except if, for some k, $i = k$ (e.g., there would be an outlying category or hospital).

In summary, we hypothesize three possible origins of $\{A_i\}$:

1. $\{A_i\}$ is derived from $\{q_i\}$,

2. $\{A_i\}$ is derived from the distribution $\{p_i\}$, where $\forall i : p_i \neq q_i$

3. $\{A_i\}$ is derived from the distribution $\{p_i\}$ where $p_i = Cq_i$ except if, for some k, $i = k$.

These three possibilities are hereafter termed the “same”, “all diff” and “k^{th} diff” hypothesis, respectively. Note we do not consider cases where there is more than one “outlier”.

With “similar” and “different” well defined, $P(\text{same}|\{A_i\})$ can be calculated through Bayes' Theorem:

$$P(\text{same}|\{A_i\}) = \frac{P(\{A_i\}|\text{same})P(\text{same})}{P(\{A_i\}|\text{same})P(\text{same}) + P(\{A_i\}|\text{all diff})P(\text{all diff}) + P(\{A_i\}|\mathrm{k^{th}\ diff})\mathrm{P(k^{th}\ diff)}}. \tag{1}$$

Each term is defined in turn. $P(\{A_i\}|\text{same})$ is the likelihood of obtaining $\{A_i\}$ (patient) counts given a set of multinomial probabilities, $\{q_i\}$:

$$P(\{A_i\}|\text{same}) = \frac{\left(\sum_{i=1}^{N} A_i\right)!}{\prod_i A_i!} \prod_i q_i^{A_i}. \tag{2}$$

Note the multinomial distribution is appropriate as each A_i is assumed to be independent and identically Poisson distributed.

$P(\{A_i\}|\text{all diff})$ is similar in form to $P(\{A_i\}|\text{same})$, but the multinomial probabilities are integrated (marginalized over) as they are assume to be unknown:

$$P(\{A_i\}|\text{all diff}) = (N-1)! \int d\vec{p} \frac{\left(\sum_{i=1}^{N} A_i\right)!}{\prod_i A_i!} \prod_i p_i^{A_i} \tag{3}$$

The $(N-1)!$ is the normalization for the integral $\int d\vec{p}$ – that is, it ensures that the term is an average[2].

In the case where all but A_k are derived from a distribution that follows $\{q_i\}$, we need to parameterize the multinomial probabilities such that $p_i = (1-f)q_i$, except when $i = k$ and the multinomial probability is f. $P(\{A_i\}|\mathrm{k^{th}}\text{ diff})$ is then

$$P(\{A_i\}|\mathrm{k^{th}}\text{ diff}) = \mathrm{(N-2)!} \int \mathrm{df} \frac{\left(\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{A_i}\right)!}{\prod_{\mathrm{i}} \mathrm{A_i}!} \mathrm{f^{A_k}} \int \mathrm{d\tilde{q}} \prod_{\mathrm{i \neq k}} \mathrm{((1-f)q_i)^{A_i}}. \tag{4}$$

As for the prior probabilities in Eqn. **??**, it is assumed that there is an even probability that the distribution $\{A_i\}$ is or is not derived from $\{q_i\}$. We then set $P(\text{same}) = 1/2$. Assuming then that “all diff” and “$\mathrm{k^{th}}$ diff” hypotheses are equally probable, and assuming $N_{categories}$ (hospitals), $P(\{A_i\}|\text{all diff}) = P(\mathrm{k^{th}}\text{ diff}) = \left(\frac{1}{2}\right) \times \left(\frac{1}{\mathrm{N_{categories}+1}}\right)$.

A.2 The Probability Pairs of Counts are Generated from the Same Binomial Distribution

Suppose a set of paired counts, $\{(A_i, B_i)\}$. We want to calculate the probability that all the pairs are derived from the same binomial distribution, assuming

[2] [2] for a derivation of the $(N-1)!$ term

the binomial probability of an “B”-like event (e.g., encounter) is greater than the probability of a “A”-like event (e.g., patient). Mathematically, if p_i is the probability of obtaining an “A”-like event, DQM calculates the probability that $\forall i : p_i = p$ (i.e., all the p_i's are the same), given that $p \leq 0.5$. No other restrictions are imposed on p, and so it's marginalize over (integrated). This similarity measure is relevant to evaluating the similarity of patient and encounter counts, as the encounters:patients ratio should be similar if the hospitals are similar. Note the contraint on p is pertinent to evalulating encounters:patients similarity, as there cannot be more patients than encounters.

The probability of similarity can be expressed through Bayes' Theorem:

$$P(\text{same}|\{(A_i, B_i)\}) = \frac{P(\{(A_i, B_i)\}|\text{same})P(\text{same})}{P(\{(A_i, B_i)\}|\text{same})P(\text{same}) + P(\{(A_i, B_i)\}|\text{all diff})P(\text{all diff}) + P(\{(A_i, B_i)\}|\text{k}^{\text{th}}\text{ diff})\text{P}(\text{k}^{\text{th}}\text{ diff})}. \tag{5}$$

Each term is discussed in turn.

$P(\{(A_i, B_i)\}|\text{same})$ is the likelihood of obtaining $\{(A_i, B_i)\}$ where the binimial probability for each pair is p. As we do not know *a priori* what p is, and so it is integrated (marginalized over) it, assuming $p \leq 0.5$. The likelihood of similarity is then

$$P(\{(A_i, B_i)\}|\text{same}) = \frac{1}{2}\int_0^{0.5} dp \left[\prod_i \binom{A_i + B_i}{A_i} p^{A_i}(1-p)^{B_i}\right] \tag{6}$$

Note the factor of 1/2 is due to the flat prior on p, and the integral can be reduced to an incomplete beta function.

The likelihood that the pairs are derived from different binomial probabilities is similar in form to $P(\{(A_i, B_i)\}|\text{same})$, except the probabilities are integrated separately, as each p_i is independent and unknown:

$$P(\{(A_i, B_i)\}|\text{all diff}) = \left[\prod_i \left(\frac{1}{2}\right)\int_0^{0.5} dp_i \binom{A_i + B_i}{A_i} p_i^{A_i}(1-p_i)^{B_i}\right] \tag{7}$$

We further consider the possiblility that a single pair can be responsible for the dis-similarity. Assuming then the k^{th} pair has a different binomial probability than the others,

$$P(\{(A_i, B_i)\}|\text{k}^{\text{th}}\text{ diff}) = \left[\left(\frac{1}{2}\right)\int_0^{0.5} \text{dp}_\text{k} \binom{\text{A}_\text{k} + \text{B}_\text{k}}{\text{A}_\text{k}} \text{p}_\text{k}^{\text{A}_\text{k}}(1-\text{p}_\text{k})^{\text{B}_\text{k}}\right] \frac{1}{2}\int_0^{0.5} \text{dp} \left[\prod_{\text{i}\neq\text{k}} \binom{\text{A}_\text{i} + \text{B}_\text{i}}{\text{A}_\text{i}} \text{p}^{\text{A}_\text{i}}(1-\text{p})^{\text{B}_\text{i}}\right] \tag{8}$$

where (A_i, B_i) has a binomial probability p if $i \neq k$, and p_k otherwise.

The prior probabilities, $P(\text{same})$, $P(\{A_i\}|\text{all diff})$ and $P(\{A_i\}|\text{k}^{\text{th}}\text{ diff})$, retain the same definitions as in the previous section.

A.3 Isolating “Outlying” Categories (Hospitals)

Provided that it determined that the categories are dis-similar (i.e., $P(\text{same}|\{A_i\}) < 0.5$ or $P(\text{same}|\{(A_i, B_i)\}) < 0.5$), we then want to determine the nature of the dis-similarity. In short, we want to know (1) are all the categories dis-similar, or is there an "outlier" and (2) if the dis-similarity is due to an outlier, does the outlier have an excess or deficit in its counts (atio)?

To determine whether or not the dis-similarity is due to an outlying hospital, we compare each $P(\{A_i\}|\text{k}^{\text{th}} \text{ diff})$'s (P($\{(A_i, B_i)\}|k^{\text{th}}$ diff)'s) with $P(\{A_i\}|\text{all diff})$ $P(\{(A_i, B_i)\}|\text{all diff}))$. If, $\forall k : P(\text{k}^{\text{th}} \text{ diff}) > \text{P}(\{\text{A}_\text{i}\}|\text{all diff})$ ($\forall k : P(\{(A_i, B_i)\}|\text{all diff}) > P(\{(A_i, B_i)\}|\text{all diff})$), then the difference is due to an outlying hospital; otherwise, it is more likely that the counts are different from the gold standard (ratios are different from one another).

If the dis-similarity is then determined to be the result of an outlier, the outlier is determined by finding that k which maximizes $P(\{A_i\}|\text{k}^{\text{th}} \text{ diff})$ $(P(\{(A_i, B_i)\}|\text{k}^{\text{th}} \text{ diff}))$.

References

[1] Connolly B, Faist R, West C, Matykiewicz P, Glauser TA, Pestian J. Choosing the Right Statistical Framework for Your Infographics: A Case Study Involving Database Quality Measures for Multi-Institutional Epilepsy Patient Queries. Submitted to Visible Language. 2014.

[2] Gregory P, Loredo TJ. A new method for the detection of a periodic signal of unknown shape and period. The Astrophysical Journal. 1992;398:146–168.