Statistical Inference

Statistical inference is the process through which inferences about a population are made based on certain statistics calculated from a sample of data drawn from that population.

From: Principles and Practice of Clinical Research (Third Edition), 2012

Related terms:

Microarray, Biomarkers, Felbamate, Kurtosis, Protein, Gene, Dependent Personality Disorder, DNA, Regulator Gene

View all Topics

Learn more about Statistical Inference

Statistical Inference

Petter Laake, Morten Wang Fagerland, in Research in Medical and Biological Sciences (Second Edition), 2015

Statistical inference is important in order to analyze data properly. Indeed, proper data analysis is necessary to interpret research results and to draw appropriate conclusions. In this chapter, three basic statistical concepts are presented: effect estimate, confidence interval, and *P*-value, and these concepts are applied to the comparisons of proportions, means, and medians. Regression models are the most commonly used method in medicine and the biological sciences to describe the relationship between an outcome variable and one or more exposure variables. This chapter will demonstrate how to perform linear regression, logistic regression, median regression, Poisson regression, and Cox regression analyses. Examples will be given for all methods.

> Read full chapter

The Human Auditory System

DeborahAnn Hall , KarimaSusi , in Handbook of Clinical Neurology, 2015

Statistical inference

Statistical inference refers to the process of drawing conclusions from the model estimation. When computing the GLM, a □ value is estimated for each regressor (i.e., column in the design matrix). □ values can be used to compare regressors and compute activation maps by creating *t* statistics and equivalent *z* scores for each voxel in normalized brain space. The null hypothesis for fMRI images is that all □s are zero (i.e., that none of the regressors has an effect on the MR signal in the area being scanned). SPM provides a number of different levels of statistical inference for drawing conclusions about the □ estimates (i.e., voxel-level and cluster-level; Friston et al., 1996b). The most commonly used is the voxel-level inference which tells us the likelihood of obtaining at least one voxel whose statistic exceeds the alpha threshold (e.g., $P < 0.05$). For example, Warren and Griffiths (2003) identified pitch- and location-sensitive regions using a voxel-level threshold of $P < 0.05$ with a small volume correction.

> Read full chapter

An Overview of Statistics in Education

S. Sinharay, in International Encyclopedia of Education (Third Edition), 2010

Statistical Inference

Statistical inference consists in the use of statistics to draw conclusions about some unknown aspect of a population based on a random sample from that population. Some preliminary conclusions may be drawn by the use of EDA or by the computation of summary statistics as well, but formal statistical inference uses calculations based on probability theory to substantiate those conclusions. Statistical inference can be divided into two areas: estimation and hypothesis testing. In estimation, the goal is to describe an unknown aspect of a population, for example, the average scholastic aptitude test (SAT) writing score of all examinees in the State of California in the USA. Estimation can be of two types, point estimation and interval estimation, depending on the goal of the application. The goal of hypothesis testing is to decide which of two complementary statements about a population is true. Two such complementary statements may be: (1) the students of California score higher on an average on SAT writing than the students of Texas, and (2) the students of California score lower on an average on SAT writing than the students of Texas. Point estimation is discussed in the statistics section of the encyclopedia. Details on

interval estimation and hypothesis testing, and power analysis, which play a key role in hypothesis testing are also discussed in the statistics section of the encyclopedia. Often, an investigator has to perform several hypothesis tests simultaneously. For example, one may want to compare the SAT critical reading scores of several pairs of schools belonging to a geographical region. The article on multiple comparison in the statistics section of the encyclopedia, discusses how to handle such a situation in an appropriate manner.

> Read full chapter

An Introduction to Biostatistics: Randomization, Hypothesis Testing, and Sample Size Estimation

LAURA LEE JOHNSON, ... PAUL S. ALBERT, in Principles and Practice of Clinical Research (Second Edition), 2007

3.1 The Goals of Statistical Inference

Statistical inference is the procedure through which inferences about a population are made based on certain characteristics calculated from a sample of data drawn from that population. In statistical inference, we wish to make statements not merely about the particular subjects observed in a study but also, more importantly, about the larger population of subjects from which the study participants were drawn. In the beta-interferon/MRI study, we wish to make statements about the effects of beta-interferon, not only in the 14 participants observed in this study but also in all patients with RRMS. Similarly, in the felbamate monotherapy study, we want to make a decision about the effectiveness of felbamate for all patients with intractable partial epilepsy. Statistical inference can be contrasted with exploratory data analysis, where the purpose is to describe relationships in a particular data set without broader inference. Inferential techniques attempt to describe the corresponding characteristics of the population from which the sample data were drawn.

To develop a conceptual view of hypothesis testing, we first need to define some terminology. A *statistic* is a descriptive measure computed from data of a sample. For example, the sample mean (average), median (middle value), or sample standard deviation (a measure of typical deviation) are all statistics. A *parameter* is a descriptive measure of interest computed from the population. Examples include population means, population medians, and population standard deviations. The distribution of all possible values that can be assumed by a particular statistic, computed from random samples of a certain size repeatedly drawn from the same population, is

called the *sampling distribution* of that statistic. The goal in statistical inference is to use probability theory to make inferences about population parameters of interest. For example, for the felbamate monotherapy trial, the parameter of interest is the change in daily seizure rates due to felbamate treatment. The statistic is the mean number of seizures per day for participants in the placebo arm minus the mean for participants randomized to the felbamate arm of this trial. Although we cannot observe the population and hence the sampling distribution directly, we can model them based on our understanding of the biological system and the sample that we are studying.

There are two broad areas of statistical inference: statistical estimation and statistical hypothesis testing. Statistical estimation is concerned with best estimating a value or range of values for a particular population parameter, and hypothesis testing is concerned with deciding whether the study data are consistent at some level of agreement with a particular population parameter. We briefly describe statistical estimation and then devote the remainder of this section to providing a conceptual overview of hypothesis testing.

There are two types of statistical estimation. The first type is point estimation, which addresses what particular value of a parameter is most consistent with the data. For example, how do we obtain the best estimate of treatment effect for the beta-interferon/MRI data? Is the best estimate obtained by taking the mean or median reduction in the number of monthly lesions? Depending on the skewness of the data and the exact question of interest, one estimate may be preferable to the other; this is another time to talk with a statistician about the best way to evaluate the effect of interest. The second type of statistical estimation is interval estimation. Interval estimation is concerned with quantifying the uncertainty or variability associated with the estimate. This approach supplements point estimation because it gives important information about the variability (or confidence) in the point estimate. An example would be the statement of the 95% confidence interval for the mean effect of felbamate in the epilepsy clinical trial. This interval gives us an idea of the variability of the treatment effect as well as its size. One can interpret these confidence intervals in a frequentist fashion; in the long term, 95% of similarly constructed confidence intervals will contain the true mean effect. However, one cannot determine whether a particular interval does or does not contain the true mean effect. More loosely one might discuss being 95% confident that the true treatment effect occurs between two stated values, with the caveat of understanding this in a frequentist fashion and not exactly as stated.

Hypothesis testing has a complementary perspective. The framework addresses whether a particular value (often called the null hypothesis) of the parameter is consistent with the sample data. We then address how much evidence we have to reject (or fail to reject) the null hypothesis. For example, is there sufficient evidence

in the epilepsy trial to state that felbamate reduces seizures in the population of intractable partial epilepsy patients?

> Read full chapter

Hypothesis Testing

Julien I.E. Hoffman, in Biostatistics for Medical and Biomedical Practitioners, 2015

Hypotheses

Statistical inference is often based on a test of significance, "a procedure by which one determines the degree to which collected data are consistent with a specific hypothesis…" (Matthews and Farewell, 1996). Hypotheses may be specific, for example, that the slope relating two variables is 1, the line of identity. More often the hypothesis is that two (or more) sample statistics could have been drawn from the same population. The principles to be discussed are illustrated by referring to sample and population means, but apply equally to all other types of statistics. This is the null hypothesis or H_0, and if it is accepted then we believe that the two samples could have come from the same population.

If we choose to reject the null hypothesis, then there is an alternative hypothesis, H_A, that has three forms:

1. ; ; and come from different populations and is bigger or smaller than .
2. ; ; and come from different populations and is smaller than .
3. ; ; and come from different populations and is bigger than .

> Read full chapter

Epidemiology

Martin Prince, in Core Psychiatry (Third Edition), 2012

Chance

Statistical inference involves generalizing from sample data to the wider population from which the sample was drawn. Inferences are made by calculating the probability that, given the size of the sample, chance alone might have accounted for a given observation.

Sampling error and sampling distributions

Chance operates through sampling error. If we wanted to know the average height of boys aged 16 in the UK, we would not go to the trouble of measuring the height of every male of that age. We would instead draw a representative sample from a population register. Random selection of participants should ensure representativeness. However, if the sample was relatively small, say 100 boys, then it is quite likely that, by chance, we would happen to sample those who were on average slightly taller or slightly shorter than 16-year-olds in general. If we repeat the study over and over again, drawing each time a sample of 100 boys, and measuring the mean height on each occasion, we would end up with a *sampling distribution* as in Figure 9.3. The observed means from repeated sampling are normally distributed. This tends to be true even if the trait itself is not normally distributed in the population (the proof is referred to as the central limit theorem). The mean (and median and mode) of the sampling distribution are equal to the population mean; sample estimates for the mean that deviate considerably from the true population mean are observed much less commonly, and are represented in the tails of the distribution. Note that if the size of the samples is increased, then the variance of the means obtained through repeated sampling decreases. This is because larger samples tend to give more precise estimates.

Figure 9.3 •. Sampling distribution for different sample sizes.

Standard errors and confidence intervals

The standard deviation for the sampling distribution is known as the standard error of the mean, and has the property that 95% of sample means obtained by repeated sampling lie ±2 (actually 1.96) standard errors above or below the population mean. This information can therefore be used to construct limits of uncertainty around an observed sample mean, giving the range of likely values for the population mean. These limits of uncertainty are referred to as 95% confidence intervals. The standard error of the mean is the standard deviation of the population (usually estimated as the standard deviation of the sample) divided by the square root of the sample size. Thus, the observed mean height of the sample of 100 boys might be 160 cm with a standard deviation of 40 cm. The standard error of the mean would then be: $40/\sqrt{100} = 4$ cm. The 95% confidence intervals would then be 160 cm ± 4 × 1.96, or 152–168 cm. This would signify that given the sample size and the variance of heights in the sample, there would be a 95% probability that the true mean height of the whole population of 16-year-olds would lie between 152 cm and 168 cm, with only a 5% probability that it lay outside. In descriptive studies, statistical inference therefore allows us to estimate the precision of sample estimates of measures, such as mean anxiety score or prevalence of depression.

In analytical epidemiology, however, we test hypotheses; for example, that those exposed to obstetric complications are more likely to go on to develop schizophrenia than those not so exposed. Statistical inference still works in much the same way as with descriptive studies. Now we are using a *sample* of a certain size to estimate the real relative risk for the association between obstetric complications and schizophrenia *in the general population*. Sampling error may lead us to observe a relative risk (RR) that is lower or higher than the real population effect. We can calculate the standard error of the RR, and use it to construct 95% confidence intervals around our observed value. Again, there will be a 95% probability that the true RR lies within these confidence intervals, and a 5% probability that it lies outside. Thus, in the study of OCs and schizophrenia, an observed RR of 2.0 with 95% confidence intervals of 1.4–2.6 would suggest that a true RR <1.4 or >2.6 would be extremely unlikely (<2.5% probability in each case).

Statistical tests and *p* values

Statistical tests test whether a hypothesis about the distribution of one or more variables should be accepted or rejected. In the case of a hypothesized association between a risk factor and a disease, we can estimate the probability (p) of the observed or an even greater degree of association being observed if the null hypothesis were true, i.e. accounted for by chance alone, there being no association. Conventionally, the threshold for statistical significance is taken to be 0.05. This means that, for a population in which two factors were *not* associated, if the same study with the same sample size were to be repeated 100 times, then on average an association of at least the size observed might be recorded five times. It is important to remember that there is nothing magical about the $p = 0.05$ threshold. It represents nothing more than a generally agreed acceptable level of risk of making what is known as a Type I error, i.e. falsely rejecting a null hypothesis when it is true. The probability of rejecting the null hypothesis when it is indeed false (i.e. detecting a true association) is the study's statistical power. The converse scenario, accepting a null hypothesis when it should have been rejected (i.e. failing to detect a true association), is referred to as a Type II error, and the probability of committing this error is clearly (1 – power) (Table 9.6).

Table 9.6. Type I and Type II errors

Null hypothesis	**True**	**False**
Accepted	No associationNull hypothesis correctly accepted	True association, but null hypothesis mistakenly accepted-Type II errorProbability = 1 –- power
Rejected	No association, but chance observation mistakenly leads to rejection of null hypothesis-Type I errorProbability = significance	True association, correctly identifiedProbability = power

The relationship between p values and confidence intervals

If we return to the example of the study assessing the association between obstetric complications and schizophrenia, we can see that confidence intervals (CI) convey all of the information given by p values, and more besides. The 95% CI for the RR 1.4–2.6, tell us immediately that the null hypothesis RR of 1.0 is implausible given the observed value of 2.0, and thus that the null hypothesis can be rejected with reasonable confidence. The probability of a true RR of 1.0 is certainly <5% and thus the association is 'statistically significant' at $p<0.05$. However, the confidence intervals also give us a range of plausible values, both upper and lower, for the true RR.

> Read full chapter

Public Health Surveillance

Chris Poulin, ... Craig Bryan, in Artificial Intelligence in Behavioral and Mental Health Care, 2016

Next-Generation Inference

Statistical inference accuracy is critical to next-generation artificial intelligence systems and medical predictive analytics more generally. Powered by recent advances in applied mathematics, we have reached an exciting place in the convergence of understanding of learning and pattern recognition systems.

A key unifying idea is attention to invariants and regularities across abstract processes. These processes can be either representational (Deep Learning), functional (Compressed Sensing), or algorithmic (Meta-learning). Briefly, these three areas of learning are:

- *Meta-learning*: Learning about (machine) learning systems to produce more efficient machine learning as a runtime and discipline. What one learns is presumably regularity in features or representations about learning systems and their input (features) data (Brazdil, Carrier, Soares, & Vilalta, 2008).
- *Deep Learning*: A prominent form of hierarchical machine learning that uses many deep layers of abstract representation, inspired by human visual processing capabilities. It is currently the dominant classifier approach (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012), and has a long history (Hinton, Sejnowski, & Ackley, 1984).
-

Compressed Sensing/Sampling: A recent development related to finding better than Nyquist-Shannon rates for measuring information in a signal, using invariants used in encoding (Candès & Plan, 2010).

Given common theoretical intersection points in the machine learning space, we might speak of high-level concept of "Relative Networks," a term that captures the many common aspects of modern learning (Figure 9.11).

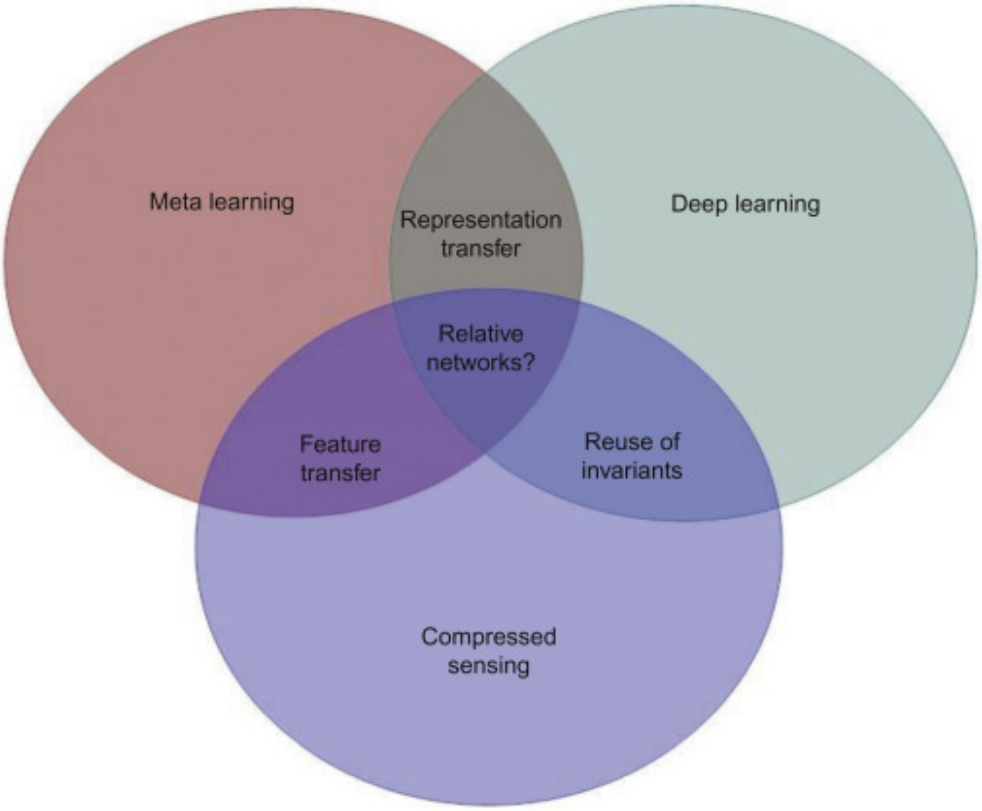

Figure 9.11. Emerging themes in machine learning.

The recurring theme for relative networks is the repeated use of structured signal similarity and intersection (A ∩ B) and the unstructured noise dissimilarity/symmetric difference (A Δ B). It is a focus on invariant representations, and *their own* meta-spatio-temporal relationships as further invariants, that will allow learning to be better understood. This is a true meta-learning, based upon "ontology of degree" or "positionals" to paraphrase the scientist Vannevar Bush. We are currently building systems of this type of brain-like model representation, to enable downstream medical decision.

> Read full chapter

Bioinformatics of Behavior: Part 2

M.A. O'Brien, ... M.F. Miles, in International Review of Neurobiology, 2012

2.3 Utilizing statistical inference for determining significant differential expression

Statistical inference is used to examine gene expression data across biological replicates to isolate significant changes, beyond what would be expected by random chance. Multiple reviews have addressed issues of statistical analysis of microarray data (Kerr & Churchill, 2007; Kim, Lee, & Sohn, 2006; Reimers, 2005). Customary statistical analyses, such as the *t*-test or ANOVA, simply tests whether the mean expression level of a gene between treatment groups is significantly different, when taking variance of measurement into consideration. A *p*-value is calculated to assess the probability of obtaining a test statistic as extreme as the one observed and is compared to a predefined significance level, □. These statistical approaches become problematic when we apply them to the field of gene expression analysis, due to the large number of genes being tested in parallel. With multiple comparisons occurring simultaneously, a significance level deemed acceptable for testing of a single gene, may result in an unacceptable number of false positives. Consider comparing the mean expression level of 1000 genes at one time. If the common significance level, □ = 0.05, was chosen for each test, one would expect, just by random chance, for a possible 50 genes that falsely rejected the null hypothesis to come through the analysis. Since gene expression analyses survey the entire transcriptome, they present an extreme multiple testing issue.

One commonly used method to balance significance and power in statistical analyses is to set an acceptable level for the expected proportion of false positives among the genes declared as differential, also known as a false discovery rate (FDR) (Storey & Tibshirani, 2003). Each hypothesis test can then be associated with a *q*-value, which is the minimum FDR at which the particular test may be called significant. A popular method for statistical filtering of data that utilizes FDR is the significance analysis of microarrays (Tusher, Tibshirani, & Chu, 2001). This method takes into consideration that expression of genes correlate in an unknown manner. An empirical distribution can be created by permuting, or randomizing the data, multiple times and determine how many genes come through as differentially expressed by chance. This will provide an estimate of the FDR for the genes reported to be differentially expressed, put into context of the actual data. The genes that come through the statistical filtering may prove to be influential in mediating the neurobiological process being examined.

> Read full chapter

Statistical Methods

Thomas D. Gauthier, Mark E. Hawley, in Introduction to Environmental Forensics (Third Edition), 2015

5.2.1 Population Parameters and Sample Statistics

Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data. In most cases, it is not practical to obtain all the measurements in a given population. For example, if we were interested in knowing the average concentration of arsenic in the top two feet of soil at a one-acre site and each measurement required a 100 gram sample, we would have to collect and analyze 37 million samples[1] to know the true average. This, of course, is impractical. As a tradeoff, we accept some uncertainty in our estimate of the true average in exchange for making fewer measurements.

The *population* consists of all the conceivable items, observations, or measurements in a group. In this example, the population consists of the total number of 100 gram quantities of soil contained in the top two feet of the one-acre site (i.e., 3.7×10^7 items). A *sample* is a subset of observations or measurements used to characterize the population. In the previous example, we might collect and analyze twenty 100-gram quantities of soil to estimate the average arsenic concentration. Thus, the sample would consist of those twenty measurements.

In this case, the *population parameter* of interest is the arithmetic mean or average of the 3.7×10^7 arsenic measurements. Parameters used to describe characteristics of the underlying population are usually represented by Greek letters. The arithmetic mean, denoted by the Greek letter μ (mu), is a measure of central tendency. Another parameter of interest is the standard deviation, a measure of the dispersion or variability in the population, denoted by the Greek letter (sigma).

Estimates of population parameters derived from a subset of the measurements in a sample drawn from the underlying population are called *sample statistics*. Latin letters are used to represent sample statistics. For example, the sample mean is denoted by (x-bar) and the sample standard deviation is denoted by *s*.

The arithmetic average or mean of the population, μ, is equal to the sum of all observations, *xi* (where x_i is the i'th observation), divided by the total number of conceivable observations, N.

(5.1)

Because we never really know the true population mean (unless we sample N times for all x_i), our best estimate of this value is the sample mean. The sample mean is equal to the sum of n values in the sample divided by the number of values.

(5.2)

One way to characterize the dispersion or variability in a population is to note the lowest and highest measurements, but this yields no information about how the data are distributed in relation to the mean. A better measure of dispersion is to see how the values vary, on average, in relation to the mean value. The average of the square of the deviations about the mean is called the mean square deviation or the variance. The variance is denoted by the Greek letter 2 (sigma squared) and is defined in Equation 5.3.

(5.3)

The population standard deviation is equal to the square root of the variance. It is also known as the root mean square deviation.

(5.4)

The sample standard deviation, s, is an estimate of the population standard deviation and is defined in Equation 5.5.

(5.5)

Notice that the denominator in the square root term in Equation 5.5 is $n-1$ instead of n. This is because one degree of freedom2 is used to estimate the arithmetic mean () in the sample.

The Central Limit Theorem states: If a variable x has a distribution with a mean μ, and a standard deviation , then the sampling distribution of the mean (), based on random samples of size n, will have a mean approximately equal to μ and a standard deviation () for which:

(5.6)

and will tend to be normal as the sample size, n, becomes large (Kachigan, 1991). The standard deviation of the population divided by the square root of n is known as the standard error of the mean (SEM) and is an important parameter for estimating confidence limits. What is meant by a "normal" distribution is defined in section 5.2.2 confidence limits are defined later in the text.

> Read full chapter

Protein complex identification from AP-MS data

Zengyou He, in Data Mining for Bioinformatics Applications, 2015

7.3.2 The statistical inference method

Statistical inference deduces properties of data sets from a set of observations and hypotheses. When the data set is a graph, the clustering objective is to find a partition model that best fits the graph based on the connectivity patterns of vertices. In the context of protein complex identification, Bayesian inference is widely adopted in which observations (bait–prey graph) are used to estimate the probability that a given hypothesis is true.

There are two basic ingredients in Bayesian inference: the observed evidence and a statistical model with some parameters. Bayesian inference starts by writing the likelihood that the observed evidence is generated by the model for a given set of parameters. The inference is performed to find parameters that maximize the posterior distribution of the parameters given the model and the evidence. Graph clustering can be considered to be a specific example of Bayesian inference problem, where the evidence is the graph structure and a hidden partition model that one wishes to infer along with some parameters.

In Ref. [6], a Bayesian approach is proposed to identify protein complexes from AP-MS data. To illustrate this method, we use the sample data set in Chapter 6 (Figure 6.2) as an example. As shown in Table 7.1, the original AP-MS data are first transformed into a binary purification matrix $\mathbf{U}$ with the size of $R \times N$, where R is the number of bait proteins and N is the number of all proteins that have once appeared in the purifications. Then, the corresponding adjacency matrix $\mathbf{M}$ in Table 7.2 is defined by $\mathbf{M} = \mathbf{U}^T\mathbf{U}$, which is a symmetric $N \times N$ matrix. The ijth element, $\mathbf{M}_{ij}$, is the number of purifications in which protein i and protein j cooccur.

The element $\mathbf{M}_{ij}$ in the adjacency matrix can be regarded as the number of distinct "paths" between protein i and protein j discovered by the AP-MS experiment. For example, there are three paths between protein A and protein C and no path that directly connects protein B and protein H. However, it is possible to reach protein H indirectly from protein B through their neighbors. For instance, B can connect with H via the path B → D → H. The number of distinct paths between two proteins via another protein can be directly obtained by the matrix product $\mathbf{MM}$. More generally, the number of paths from protein i to protein j of length l on the graph corresponds to the ijth element of the matrix M^l. Therefore, the number of distinct paths with different lengths can be used to measure the "similarity" between two proteins. Based on this observation, the von Neumann diffusion kernel is used to evaluate the likelihood of two proteins belonging to the same complex in Ref. [6]:

(7.1)

where is a parameter (the diffusion factor) to make the effect of longer paths decay exponentially. The kernel can be normalized into [0, 1] in the following way:

(7.2)

Because the elements of von Neumann kernel matrix are between 0 and 1, this makes $\mathbf{S}_{ij}$ suitable as a probabilistic measure for evaluating the likelihood of two proteins belonging to the same complex.

To identify protein complexes, a binary matrix **Z** for protein complex membership is defined as well. Each entry z_{ci} in **Z** is a random variable, which indicates the membership of the ith protein in the cth complex. Note that the number of protein complexes is unknown in advance and one protein may belong to multiple complexes. The task here is to infer the unknown protein membership matrix **Z** from the observed AP-MS data.

Because the actual number of protein complexes is unknown, an infinite latent feature model is employed for protein complex membership identification [6]. Initially, the method starts with a finite model of C complexes, and then takes the limit as $C \to \infty$ to obtain the prior distribution over the binary matrix **Z**.

If each protein belongs to a complex c with probability πc, then the conditional probability is a product of binomial distributions:

(7.3)

where is the number of proteins in the cth complex.

If the prior distribution of π is a beta distribution $beta(/C, 1)$ with a model parameter , then conditional distribution for any z_{ci} is:

(7.4)

where represents the set of all entries in **Z** except z_{ci}, and is the number of proteins (excluding the ith protein) in the cth complex.

If we let $C \to \infty$, then the conditional distribution of z_{ci} becomes

(7.5)

for any c such that .

For the c's with , the number of new complexes associated with this protein has a Poisson distribution:

(7.6)

where νi is the expected number of new complexes.

For a given protein complex membership matrix **Z**, the inner product of two protein column vectors *zi*T*zj* can be used to check if protein *i* and protein *j* belong to the same complex. That is, two proteins are in the same complex if . Then, the likelihood can be evaluated as:

(7.5)

where **S** is the normalized von Neumann kernel matrix obtained from the AP-MS data.

According to the Bayes' theorem, the posterior distribution of the protein complex membership matrix **Z** is $P(Z|S)$, which is proportional to $P(S|Z)P(Z)$, where $P(S|Z)$ is given in (7.5), and $P(Z)$ is defined by the infinite latent feature model [6]. To carry out the inference, a Gibbs sampler with the following steps is used.

(1) Initialize **Z** randomly.

(2) For $t = 1$ to T(a)According to , sample *zci* for each *i* and each *c* with .(b)According to , sample the number of new complexes for each *i*.(c)Save the sample **Z**

> Read full chapter