Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Statistical Methods in Bioinformatics

Ten things you need to know about and understand

David A. Stephens

[1]Department of Mathematics and Statistics, McGill University

d.stephens@math.mcgill.ca



http://www.math.mcgill.ca/dstephens/MCB/Stats2012.pdf

The Topics

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

1. Randomness and Uncertainty
2. Probability
3. Statistical Summary
4. Hypothesis Testing
5. Multiple Testing Corrections
6. Bootstrap and Randomization Procedures
7. Regression and Classification
8. Clustering
9. Hidden Markov Models
10. Monte Carlo

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Randomness and Uncertainty

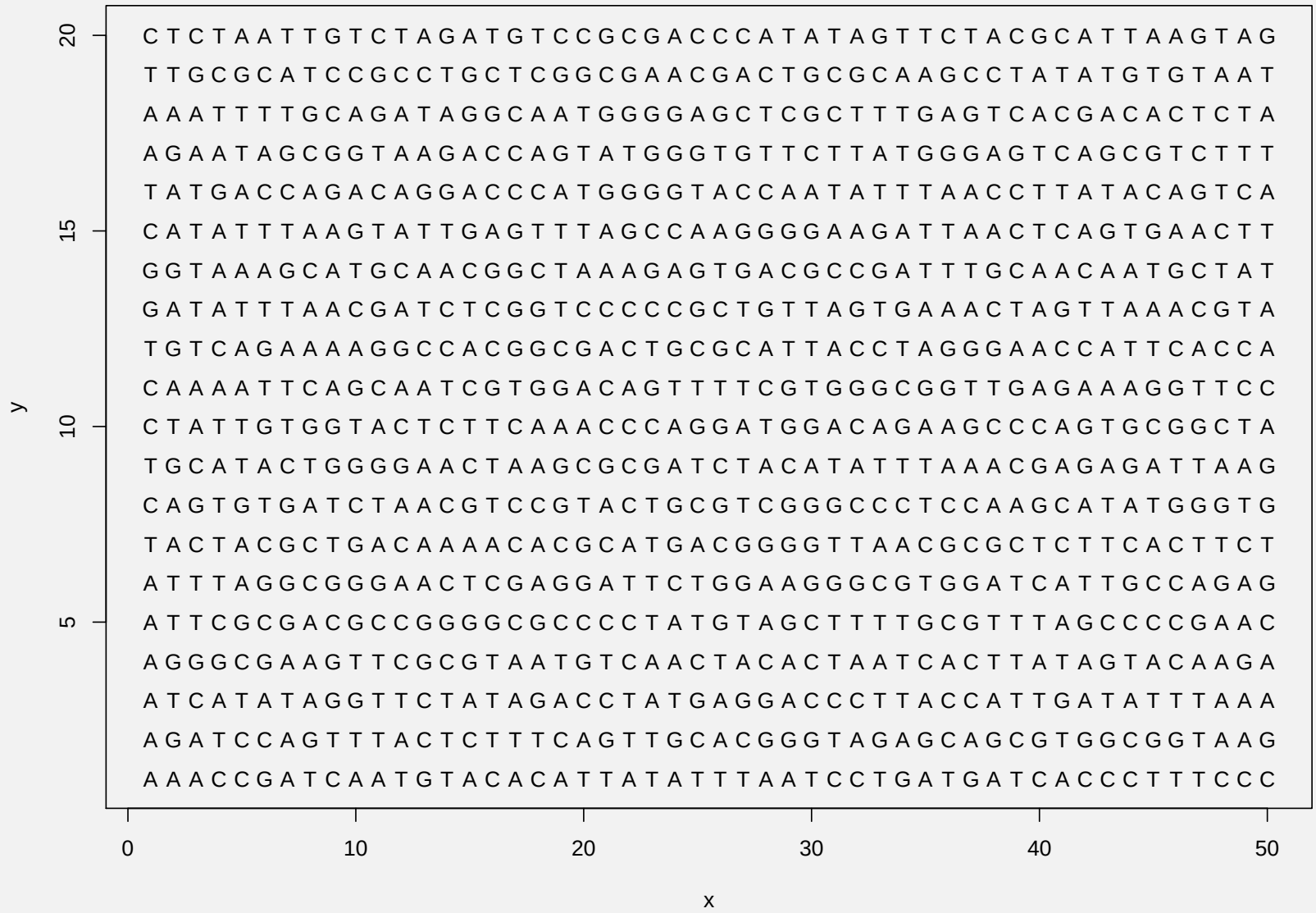
By carrying out experiments to measure some phenomenon, we gather relevant but imperfect information about the real-world.

We may observe data that leads us to suspect that some important feature may be present in the data generation model.

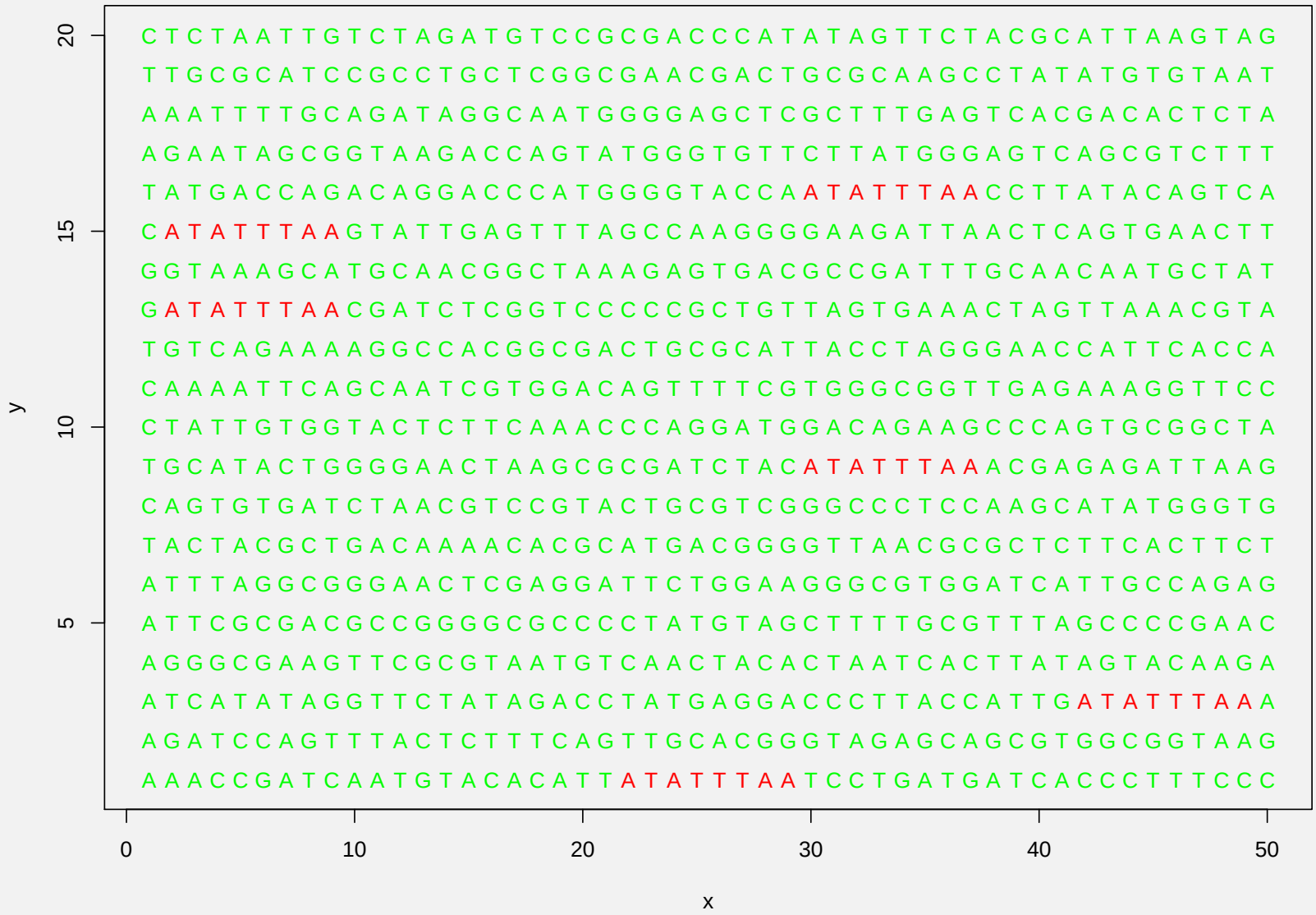
We only obtain a finite number of measurements, so our information is incomplete; we have some remaining uncertainty about the state of the world, and it is likely that what we have observed is not an accurate picture.

Key Question: Is the observed effect genuine, or due purely to chance ?

DNA Sequences



DNA Sequences



Randomness and Uncertainty

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

We collect n observations.

- if n is large, we have learned a considerable amount,
- if n is small, we may have learned little.

We should reflect the sample size in our summary of variability.

$$\text{Sample mean}: \quad \overline{x} = (x_1 + x_2 + \ldots + x_n)/n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Sample variance}: \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\text{Standard Error}: \quad s.e. = \frac{s}{\sqrt{n}}$$

Randomness and Uncertainty

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

As $n \to \infty$

$$\overline{x} \longrightarrow \mu \qquad \text{Population mean}$$

$$s^2 \longrightarrow \sigma^2 \qquad \text{Population variance}$$

$$\frac{s}{\sqrt{n}} \longrightarrow 0$$

Thus the standard error is a measure of uncertainty in our reported mean.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Probability

Probability theory provides the underpinning of all of statistics.

It gives a formal mathematical framework for the assessment of uncertainty in experimental outcomes.

The key components are

- the list of possible outcomes; the sample space
- interesting collections of those outcomes; events
- a way of assigning a numerical value (probability) to each collection.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Probability: The Axioms

Let E and F denote two events, that is, two collections of outcomes of interest. Let Ω denote the entire collection of possible outcomes.

Let P denote the probability function, and $P(E)$ denote the probability that E occurs, that is, the probability that the actual outcome (when we observe it) is one of the collection called E.

Probability: The Axioms

Example

In microarray experiments, gene expression for a large number of genes is measured using fluorescence imaging. The intensity of fluorescence is measured in a spot corresponding to a gene or EST.

The intensity measurement scale typically runs from $2^6 = 64$ to $2^{16} = 65536$. We might wish to assess the chance that the intensity measurement is between $2^8 = 256$ and $2^{10} = 1024$, as this corresponds to biologically significant up-regulation. Thus

- $\Omega = \{64, 65, \ldots, 65536\}$
- $E = \{256, 267, \ldots, 1024\}$

But what is $P(E)$?

Probability: The Axioms

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Theorem

P must exhibit the following properties:

I *$P(E)$ lies between 0 and 1.*

$$0 \leq P(E) \leq 1$$

II *The probability that the actual outcome is in the collection Ω is 1.*

$$P(\Omega) = 1.$$

III *If E and F have no common elements, then the probability the actual outcome lies within either E or F is the sum of the probabilities of E and F.*

$$P(E \text{ or } F) = P(E) + P(F)$$

Probability: The Axioms

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Theorem

P must exhibit the following properties:

I *$P(E)$ lies between 0 and 1.*

$$0 \leq P(E) \leq 1$$

II *The probability that the actual outcome is in the collection Ω is 1.*

$$P(\Omega) = 1.$$

III *If E and F have no common elements, then the probability the actual outcome lies within either E or F is the sum of the probabilities of E and F.*

$$P(E \text{ or } F) = P(E) + P(F)$$

Probability: The Axioms

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Theorem

P must exhibit the following properties:

I *$P(E)$ lies between 0 and 1.*

$$0 \leq P(E) \leq 1$$

II *The probability that the actual outcome is in the collection Ω is 1.*

$$P(\Omega) = 1.$$

III *If E and F have no common elements, then the probability the actual outcome lies within either E or F is the sum of the probabilities of E and F.*

$$P(E \text{ or } F) = P(E) + P(F)$$

Probability: The Axioms

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Theorem

P must exhibit the following properties:

I *$P(E)$ lies between 0 and 1.*

$$0 \leq P(E) \leq 1$$

II *The probability that the actual outcome is in the collection Ω is 1.*

$$P(\Omega) = 1.$$

III *If E and F have no common elements, then the probability the actual outcome lies within either E or F is the sum of the probabilities of E and F.*

$$P(E \text{ or } F) = P(E) + P(F)$$

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Probability Assignment

These mathematical rules do not tell us how to assign $P(E)$.

- Classical Approach

  Consider all outcomes equally likely

  Let E contain n_E sample outcomes, and let Ω contain n outcomes. Then

  $$P(E) = \frac{n_E}{n}$$

- Frequentist Approach

  Consider the relative frequency with which E occurs.

  In n hypothetical repeats of the experiment, suppose that E occurs n_E times. Then, if n is large

  $$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$

- Subjective Approach

  Consider your personal degree of belief that E occurs.

Probability Assignment

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

- Easy Case : Coins - Let E be "Heads"





What is $P(E)$?

- Hard Case : Thumbtack - Let E be "Point Up"

What is $P(E)$?

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Probability Assignment

- Easy Case : Coins - Let E be "Heads"





What is $P(E)$?

- Hard Case : Thumbtack - Let E be "Point Up"





What is $P(E)$?

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Statistical Summaries

Use of numerical or graphical methods to summarize the results of experimentation.

- Numerical
  - sample means, variances, standard errors, skewness
  - quantiles, median, min, max
  - correlations
- Graphical
  - one variable: histograms, boxplots, density, cumulative
  - two variables: scatterplot
  - many variables: scatterplot matrices

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing

Hypothesis (significance) testing is the formal statistical procedure for assessing the adequacy of a proposed theoretical model.

It compares the predicted behaviour of a data-related quantity (or statistic) under a proposed model with its observed behaviour in the data sample.

If the predicted and observed behaviour do not coincide, then the proposed model is deemed inadequate; if the observed behaviour is surprising under the proposed model, then proposed model can be rejected.

The statistical formulation formalizes the quantification of evidence.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing

Formally, two models are compared:

$$H_0 \quad : \quad \text{The Null Model}$$
$$H_1 \quad : \quad \text{The Alternative Model}$$

Typically, the null model is a simpler model that is of less scientific interest.

The objective of hypothesis testing is to see whether the model described by H_0 is adequate, or whether it must be rejected in favour of the more complicated model.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing

There are five crucial components to a hypothesis test, namely

- **TEST STATISTIC**
- **NULL DISTRIBUTION**
- **SIGNIFICANCE LEVEL**, denoted α
- **P-VALUE**, denoted p.
- **CRITICAL VALUE(S)**

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing : Strategy

For a data sample $x_1, \ldots, x_n$;

1. consider a pair of competing **hypotheses**, H_0 and H_1
2. define a suitable **test statistic** T.
3. **assume that H_0 is true**, and compute the predicted probability distribution of T; this is the **null distribution**
4. compute the **observed** value of T, $t = T(x_1, \ldots, x_n)$; this is the **test statistic**
5. assess whether t is a surprising observation the distribution of T, that is whether t is in a region of low probability. If it **is** surprising, there is evidence to **reject** H_0; if it is not surprising, we **cannot reject** H_0

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing: Significance Level and Critical Value

What constitutes a "surprising" result ?

This is specified by the **significance level**, α, chosen by considering an acceptable value

$$\alpha = P[T < t_{CR} \text{ or } T > t_{CR}, \text{ given that } H_0 \text{ is TRUE}]$$

that is, the probability that the predicted test statistic is more extreme than some cut-off value t_{CR}, if the null model is in fact correct.

α essentially quantifies how large or small we predict test statistic T to be by implicitly defining the critical value, t_{CR}. α is the maximum allowable false positive probability in the test of H_0.

Typically, a value of $\alpha = 0.05$ or $\alpha = 0.01$ is selected.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing: p-value

The p-**value** of a significance test is the probability p defined by

$$p = P[T < t \text{ or } T > t, \text{ given that } H_0 \text{ is TRUE}]$$

that is, the probability that the predicted test statistic is more extreme than the observed test statistic t, if the null model is in fact correct.

p quantifies the surprisingness of the result, that is, if p is small, the result would be surprising if the null model was in fact true.

If $p \leq \alpha$, we reject H_0.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing: Example

Example

In microarray analysis, replicate arrays yield expression measurements for each of two tissue samples for thousands of genes.

We might have measurements of the expression levels for a single gene

- $x_1, \ldots, x_n$ for tissue A
- $y_1, \ldots, y_m$ for tissue B

and it is of interest to consider whether there is differential expression between to the two tissues.

Hypothesis Testing: Example

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)

We consider the two data samples as being generated from populations with means μ_X and μ_Y respectively.

Then a test of differential expression is a test of the hypotheses

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \\ H_1 &: \mu_X \neq \mu_Y \end{aligned}$$

for which an appropriate test statistic is

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

which is a standardized difference in observed means.

Hypothesis Testing: Example

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)

In this formula s_x^2 and s_y^2 are the sample variances from the two samples.

The null distribution of this test statistic is centered at zero; a large positive or negative value of t leads us to reject H_0, and conclude that there is in fact differential expression.

In fact, any such test statistic can be used. However, the null distribution may not be easy to compute, even if simplifying assumptions are made.

Hypothesis Testing: Confidence Intervals

Hypothesis tests are closely related to confidence intervals, that are regions which capture the estimated value and some measure of uncertainty about a parameter of interest.

In estimation of a parameter θ, a 95 % Confidence Interval is the set of possible constants c for which the null hypothesis

$$H_0 : \theta = c$$

is **NOT** rejected given the data being studied at the $\alpha = 0.05$ significance level

More loosely (but incorrectly), a 95 % confidence interval can be interpreted as an interval within which the "true value" of θ lies with probability 0.95.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hypothesis Testing: Types of Test

- Tests for Parameters
  - z-tests, t-tests, Wald tests for coefficients
- Tests for Models
  - ANOVA, F-test
- Goodness of Fit Tests
  - Chi-squared test
- Non-parametric Tests
  - Wilcoxon, Kolmogorov-Smirnov

Hypothesis Testing: Example

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example

In a genomic sequence alignment problem, the BLAST tool is used to assess the alignments by computing a

- test statistic (an alignment score)
- the null distribution (based on large sample approximations)
- a p-value and an E-value, where E is the expected number of hits that a database search would reveal for which the test statistic would be more extreme than the observed test statistic.

In fact

$$E = NKe^{-\lambda y_{\max}}$$

for certain parameters K and λ estimated from the database.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Multiple Testing Corrections

The multiple testing corrections are used when several independent statistical tests are being performed simultaneously, as in microarray analyses.

They are necessary because, while a given significance level α may be appropriate for each individual comparison, it is not appropriate for the set of all comparisons.

The idea of a multiple testing correction is to control the rate of false positive results when many significance tests are being carried out.

In order to avoid a lot of spurious positives, when the null hypothesis is rejected when it is actually true, the α value needs to be lowered to account for the number of comparisons being performed.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Multiple Testing Corrections

When carrying out K significance tests, the simplest and most conservative approach is the Bonferroni correction, which sets the significance level for each test

$$\alpha_B(K) = \frac{\alpha}{K}$$

where α is a significance level that would be assumed for a single test.

This correction is usually too extreme, and is somewhat of an over-correction. More advanced corrections have been proposed that have better testing properties.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Multiple Testing Corrections: FDR

An alternative approach to multiple testing procedures is based on the assessment of the False Discovery Rate (FDR).

Suppose that K tests are to be carried out. Then

$$FDR = \frac{\text{Number of False Test Rejections}}{\text{Total Number of Test Rejections}}$$

if the denominator is greater than zero.

In any practical example, we do not know the numerator. However, we can design test procedures that guarantee to limit the FDR to be below some fixed value (say 10 %).

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Bootstrap and Randomization Procedures

Resampling techniques such as the bootstrap standard error, confidence intervals, and distributions for any statistic.

By resampling observations from the observed data, the process of sampling observations from the population is mimicked.

In bootstrap resampling,

- B new samples, of the same size as the observed data, are drawn with replacement from the observed data.
- The statistic of interest is first calculated using the observed data and then recalculated using each of the new samples, yielding a bootstrap distribution.
- The resulting replicates are used to calculate the bootstrap estimates of bias, mean, and standard error for the statistic.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Permutation tests

The central idea of permutation tests uses rearrangements of the data (permutations) to compute the null distribution in a hypothesis test.

- Often a null hypothesis imposes some form of symmetry on the observed data.
- Using permutation procedures, the sampling distribution of the test statistic under the null hypothesis is computed by forming all (or many) of the permutations that respect the symmetry calculating the test statistic for each.
- These pseudo-test statistics are used to form the null distribution, and hence compute test p-values etc.

Permutation tests: Example

Example

In a microarray analysis, suppose 10 and 8 measurements for two tissue samples are available, with data means

$$5.149 \qquad \text{and} \qquad 15.033.$$

We wish to test whether there is a mean difference using permutation methods. There are

$$\binom{18}{8} = 43758$$

possible relabellings of the samples consistent with the null hypothesis

$$H_0 : \mu_X = \mu_Y$$

so we can compute the exact null distribution by computing the chosen test statistic for each possible relabelling.

Permutation tests: Example

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)

Observed Test statistic:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} = -7.99$$

Permutation test statistic: computed by re-allocating at random 10 observations from 18 to tissue A and remaining 8 to tissue B, and then recomputing t, for all possible re-allocations.

Permutation tests: Example

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

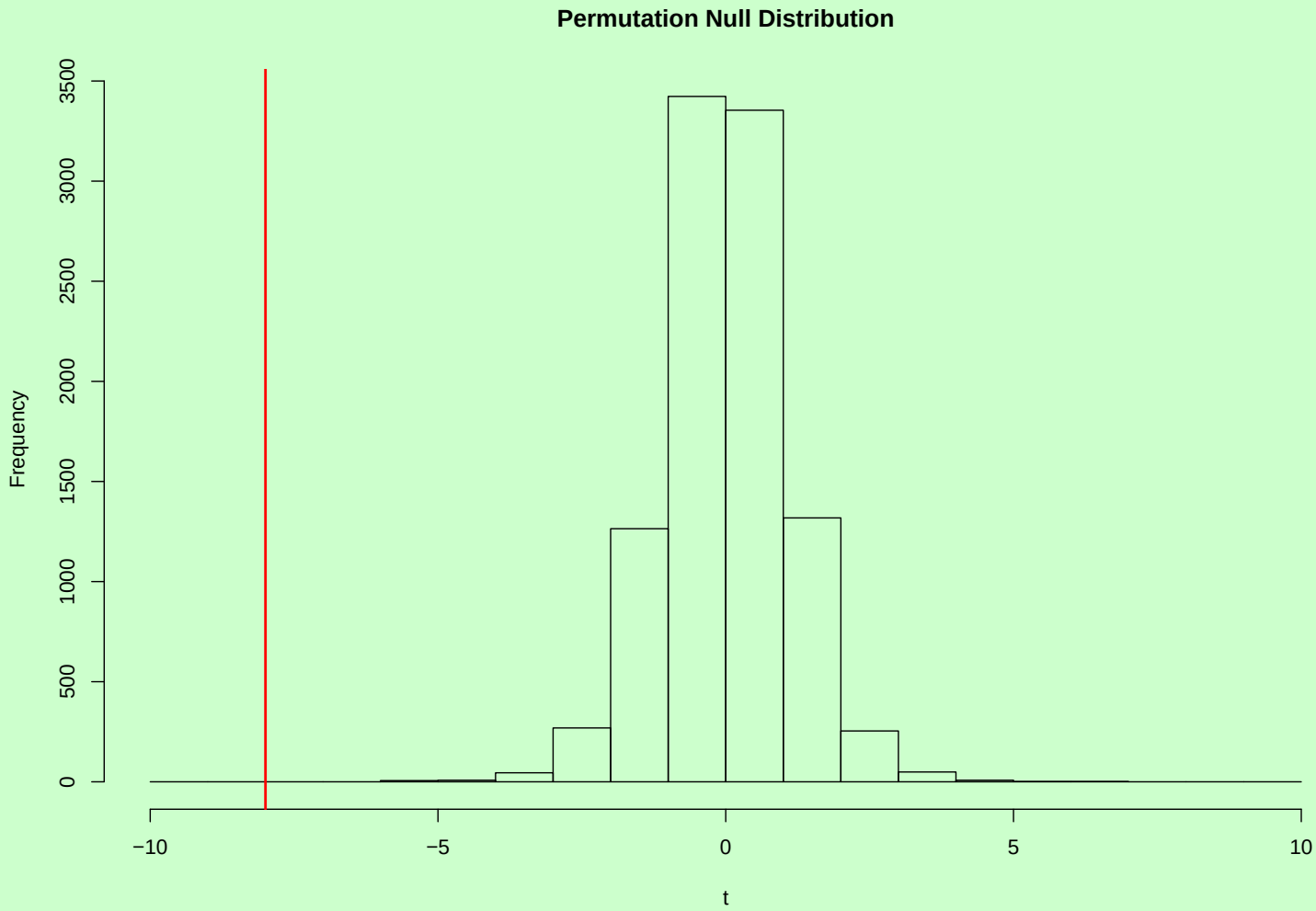
Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)



Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Regression and Classification

Regression involves modelling the variability in a variable as a function of a collection of other variables.

- response variable, Y
- predictors, $X_1, \ldots, X_d$

We fit a model for Y as a function of the X variables that captures the systematic component of variation, with an additional random component.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Regression and Classification

Two cases

- Normal Linear Regression
  - continuous data
  - normal (Gaussian) assumption for random variability
- Generalized Linear Model
  - categorical data
  - needs other probabilistic model for random variability

Regression and Classification

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

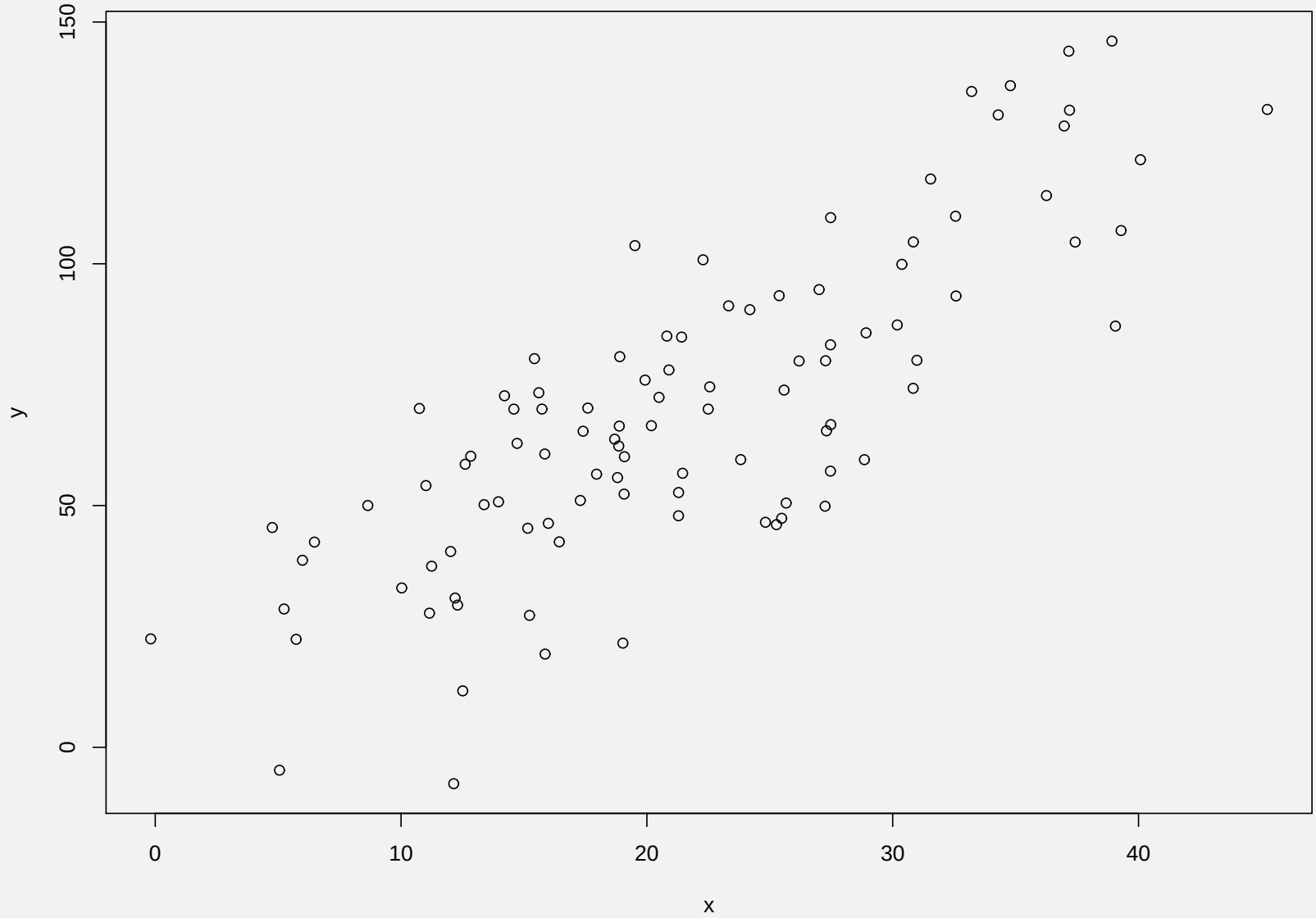
Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo



Regression and Classification

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

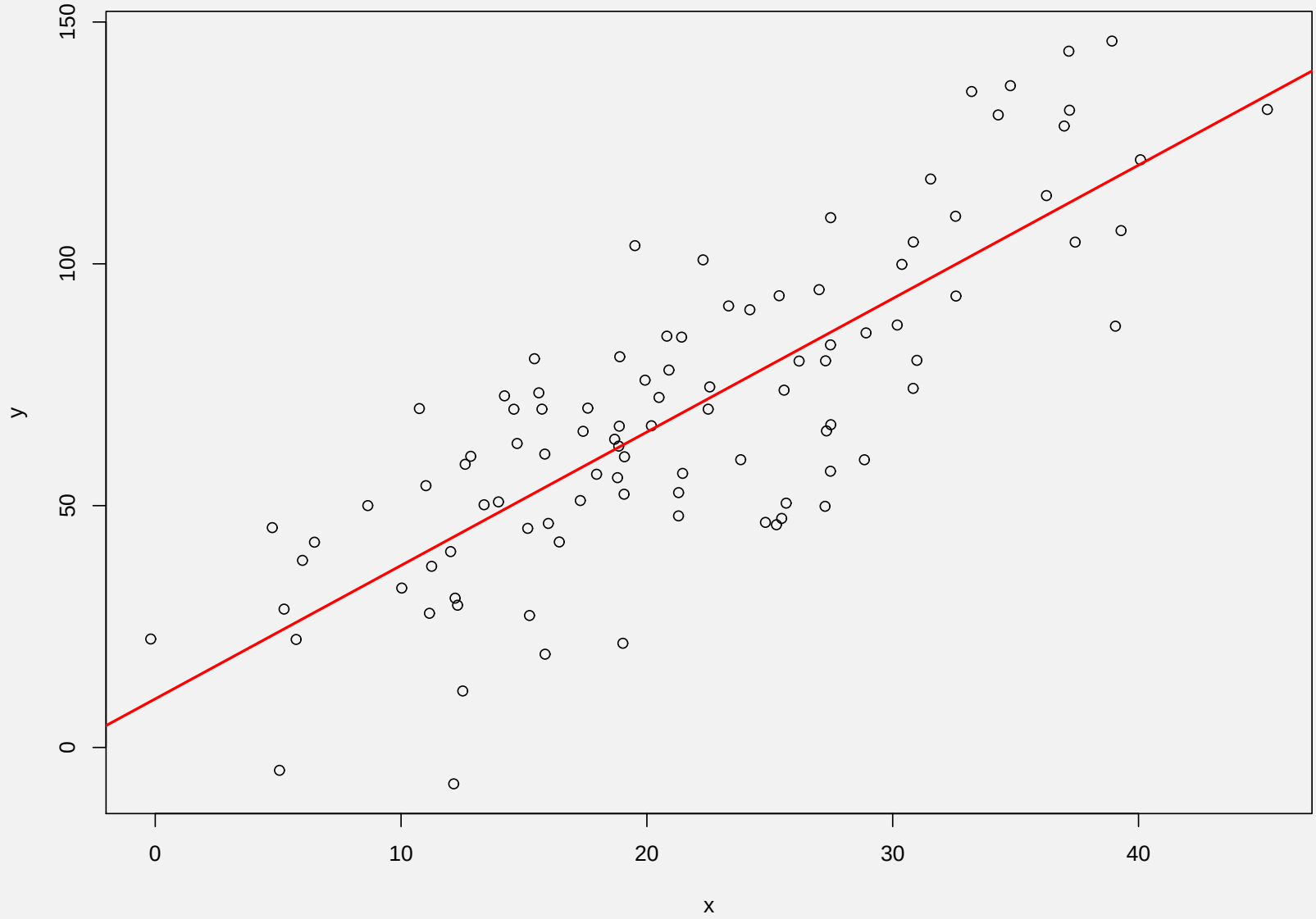
Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo



Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Regression and Classification

If the response variable is categorical, we might be interested in performing a classification exercise

- $Y = 0$ or $Y = 1$ correspond to two classes (healthy/affected, normal/tumour)
- access to some labelled cases
- case status known for some individuals, unknown for others
- objective is to label the unknown cases

Labelled cases are known as training data

Regression and Classification

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

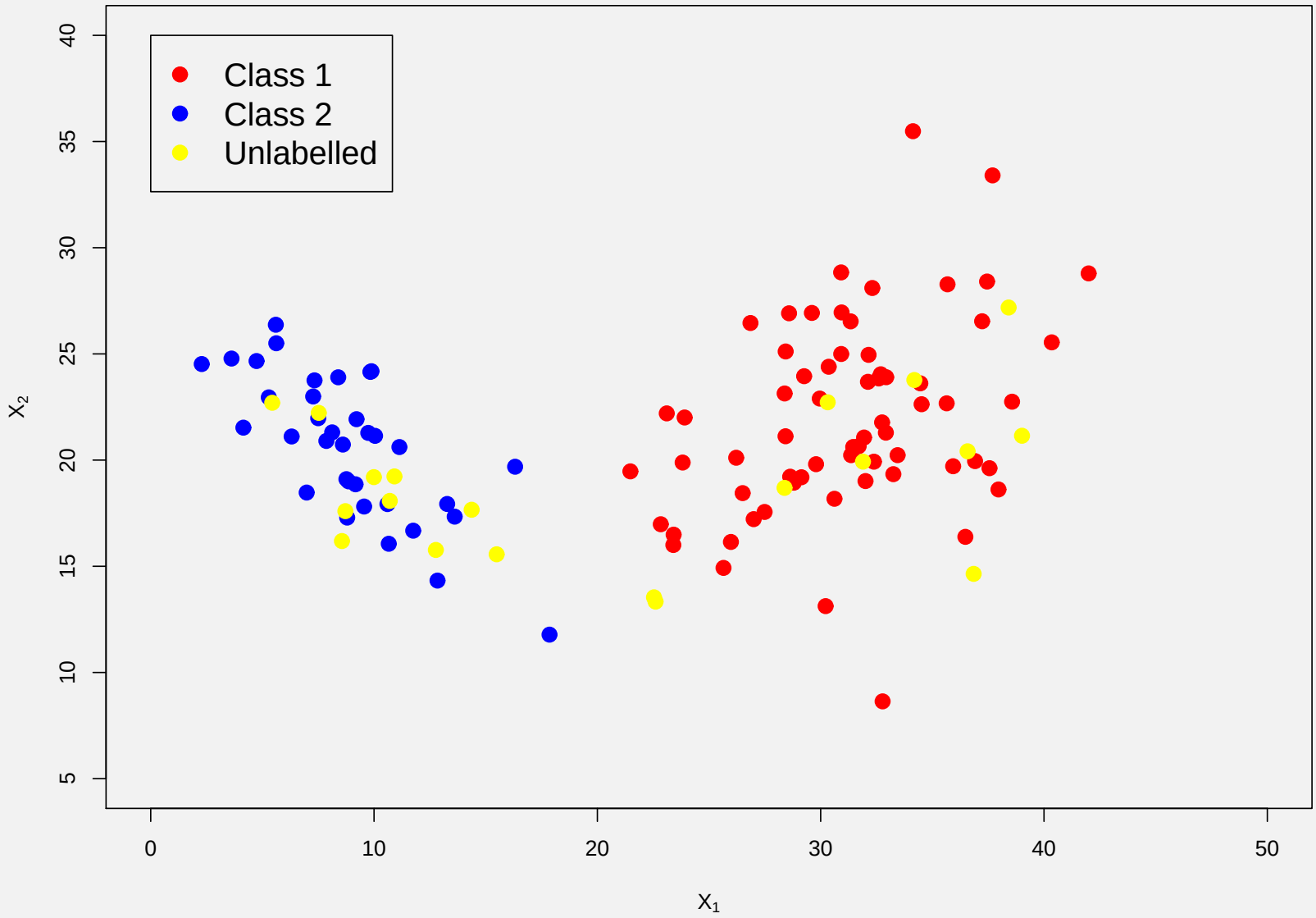
Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo



Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Regression and Classification: Example

Example

Microarray data provide large amounts of information on gene expression. An objective is to uncover a signature expression profile that allows discrimination between tissue types.

In cancer studies it may be useful to identify characteristic signatures of different tumour types; such studies are statistically challenging

- Usually many thousands of genes
- Usually very few cases
- In a famous example (Golub et al., 1999)
  - ALL/AML tumours
  - $\sim$ 7000 genes
  - 72 cases

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Regression and Classification

Links to other forms of multivariate analysis.

- Principal Components Analysis (PCA)
- Least-squares and Partial Least Squares
- Discrimination

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Clustering

Cluster analysis involves searching for groups (clusters) in the data, in such a way that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar.

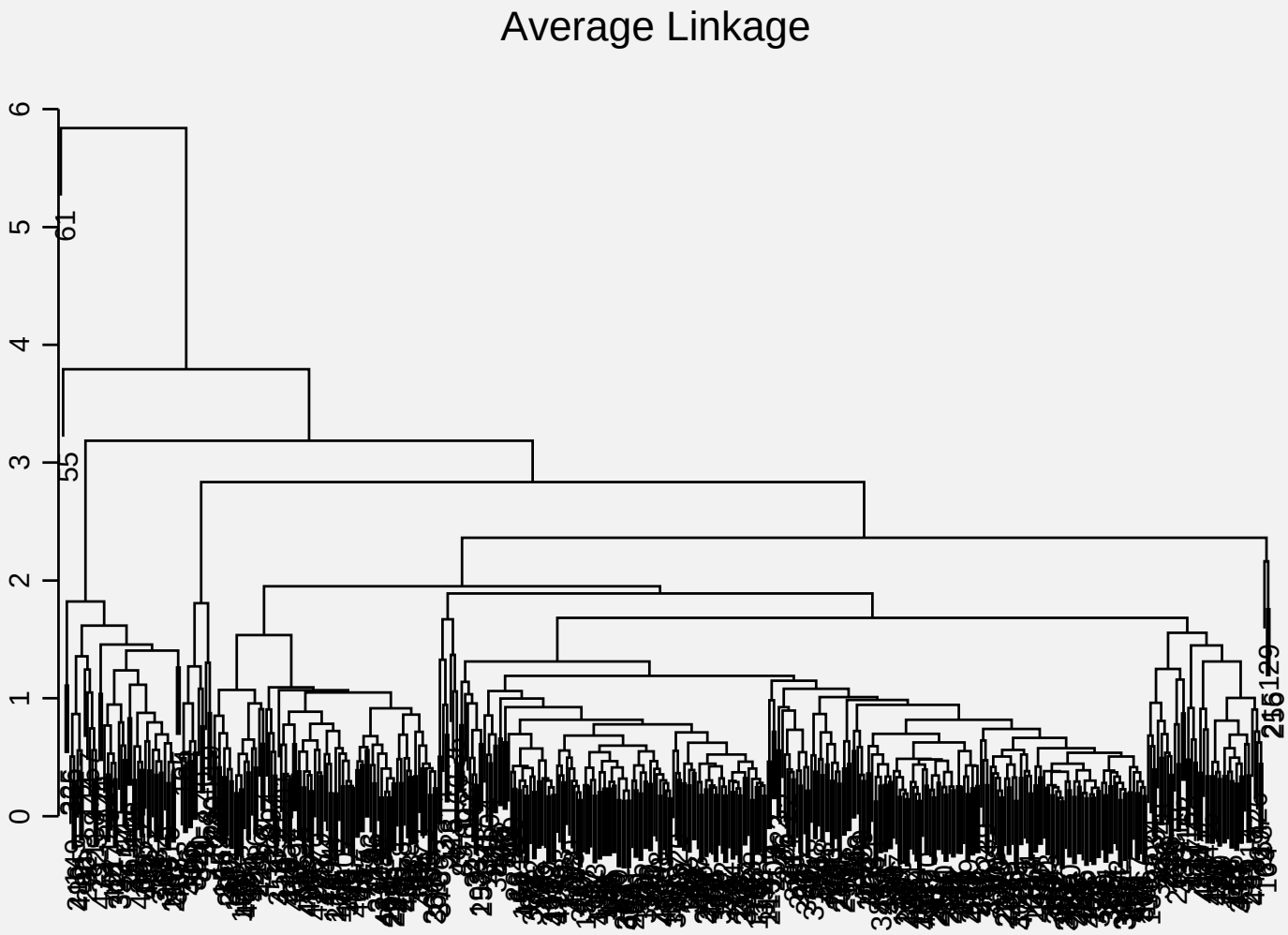
There are two basic sorts of algorithm

- Partitioning algorithms
- Hierarchical algorithms

Both use the basic concept of similarity, but have different rules for defining the subgroups.

Clustering

Hierarchical clustering results can be displayed using a dendrogram



Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hidden Markov Models

In a biological sequence analysis, suppose that the character at a given sequence position can be classified as being in a **region** of one of H possible regions.

Within a region labelled by h the observed nucleotide sequence follows a Markov chain

- A Markov chain is a particular stochastic process model that captures the dependencies between adjacent positions or regions.
- The chain is specified by two quantities
  - a transition matrix P that specifies the probabilities that one character is followed by another in the sequence,
  - a probability vector p_0 that specifies the distribution of the initial character

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hidden Markov Models

To complete the specification, we assume that there is a sequence in parallel to the observed sequence, comprising region label random variables $H_1, H_2, \ldots$ which itself follows a Markov chain governed by transition matrix P_θ with $(i,j)^{\text{th}}$ element

$$P(\text{Region type } i \text{ at time } t \to \text{Region type } j \text{ at time } t+1)$$

The key difference is that the $\{H_t\}$ sequence is unobserved or hidden.

Our objective is to uncover the hidden sequence of regions or states from the observed data.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hidden Markov Models

We use special algorithms to estimate the parameters and the hidden state sequence

- forward/backward algorithms
- Viterbi algorithm
- Baum-Welch algorithm

Code is widely available for a wide range of specific HMMs useful in bioinformatics problems

Hidden Markov Models: CpG Islands

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Durbin et al. (1998))

CG or CpG dinucleotides are common in certain parts of the human genome.

Regions that are CpG rich are called CpG islands, and are typically up to 1000 bases long.

It is of biological interest to determine the locations of CpG islands, as these regions are often associated with, for example, promoter sites.

$$\cdots CGCGTACGCGCGTGAC \cdots$$

Hidden Markov Models: CpG Islands

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Durbin et al. (1998))

From the observed sequence alone, it is not possible to deduce whether a given CG pair is part of a CpG island, or is just a normal dinucleotide pair.

We can use an eight state hidden Markov model (HMM)

- A_+, C_+, G_+, T_+ are the states corresponding to nucleotides within a CpG region
- A_-, C_-, G_-, T_- are the states corresponding to nucleotides outside of a CpG region

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Hidden Markov Models: CpG Islands

Example (Continued)

- we can augment these eight states by two further states for completeness
  - a **begin** state
  - an **end** state
    These two states are "silent" as they do not contribute a nucleotide to the observed nucleotide sequence, but are useful for introducing gaps into the sequence.

Hidden Markov Models: Parameter Estimation

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

For HMMs, we need to estimate

- **transition probabilities**

$$a_{kl} = \Pr(h_i = l | h_{i-1} = k)$$

- **emission probabilities**

$$e_k(b) = \Pr(x_i = b | h_i = k)$$

where x_i is the nucleotide in position i.

Hidden Markov Models: CpG Islands

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)

In the CpG island case, the emission probabilities are all either 1 or zero, for example

$$P(x_i = A|h_i = A_+) = 1 \qquad \Pr(x_i = C|h_i = A_+) = 0$$

and so on.

Parameters in the state transition matrix can be obtained from training data.

Hidden Markov Models: CpG Islands

Example (Continued)

For the CpG example, within the islands

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

Hidden Markov Models: CpG Islands

Example (Continued)

Outside the islands

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Need also to assess probabilities of transitions between region types, for example $A_+ \to C_-$.

Hidden Markov Models

More complicated examples can be handled using the same framework

Example (Globin Proteins)

```
(a)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

(b)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

(c)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D     ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Sequence Alignments to human alpha globin HBA_HUMAN; (a) human beta globin HBB_HUMAN, (b) lupin leghaemoglobin LGB2_LUPLU, (c) nematode glutathione S-transferase F11G11.2.

Hidden Markov Models

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example (Continued)

Hidden states reflect consensus sequence, and observed states differ due to evolutionary forces.

Can also build HMM to infer protein secondary structures

- α-helices
- β-sheets
- insertions/deletions
- loops
- features and sites

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Monte Carlo

Monte Carlo is a simulation approach to studying biological models and systems that allows generation of hypothetical data.

- Allows assessment of proposed model through simulation of possibly complicated systems
- Relies on random number generation techniques
- Requires a stochastic (probability) model for the experimental variability or data generation model.

Monte Carlo

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Example

In statistical population genetics, a common model for the correlation in genomic sequences is the coalescent, which is based on a stochastic process model of evolution, mutation, selection and recombination.

It is a complicated model that is formulated via a tree model for the unobserved relationships. Analytic computations are difficult to carry out, but the model is relatively easy to simulate forward in time.

Therefore properties of the model can be studied using

- forward Monte Carlo simulation of replicate data
- summarization of the simulated data.

Randomness and Uncertainty

Probability

Statistical Summary

Hypothesis Testing

Multiple Testing Corrections

Resampling Procedures

Regression and Classification

Clustering

Hidden Markov Models

Monte Carlo

Summary

- Statistical calculations are an essential part of bioinformatics and computational biology.
- Computational statistical methods such as the bootstrap forms a central part of statistical bioinformatics as they avoid the need for use of toy models.
- Current challenges largely involve the magnitude and diversity of biological data available, and the need for coherent data fusion.