BCBio 444: Bioinformatics Analysis

Karin S. Dorman

Department of Statistics
Department of Genetics, Development & Cell Biology
Program in Bioinformatics & Computational Biology
Iowa State University

October 10, 2017

Probability

- **random experiment**: **outcome**, **sample space**, **event**, **probability measure**, **iid**.
- **probability model**: a *random experiment* with *probability measure* that mimics the data generation process of a *scientific experiment* while making assumptions, including some you may wish to disprove.

Probability Examples 1

1. If the nucleotides in a sequence $\boldsymbol{S}$ are iid and uniformly distributed, what is

$$P(\boldsymbol{S} = \text{ACGAA}) = ?$$

2. Markov Chain example: Given transition matrix $[[0.6, 0.2, 0.1, 0.1, [0.1, 0.1, 0.8, 0], [0.2, 0.2, 0.3, 0.3], [0.1, 0.8, 0, 0.1]]$ and $S_1 \sim \text{Multinoulli}(0.25, 0.25, 0.25, 0.25)$, what is

$$P(\boldsymbol{S} = \text{ACGAA})?$$

Probability Examples 1

1. If the nucleotides in a sequence $\boldsymbol{S}$ are iid and uniformly distributed, what is

$$P(\boldsymbol{S} = \text{ACGAA}) = ?$$

Example

I may also write: Assume a sequence $\boldsymbol{S} = (S_1, S_2, S_3, S_4, S_5)$ of length 5 is produced under model

$$S_i \overset{\text{iid}}{\sim} \text{Multinoulli}(0.25, 0.25, 0.25, 0.25).$$

Then, $P(\boldsymbol{S} = \text{ACGAA}) = 0.25^5$.

2. Markov Chain example: Given transition matrix $[[0.6, 0.2, 0.1, 0.1, [0.1, 0.1, 0.8, 0], [0.2, 0.2, 0.3, 0.3], [0.1, 0.8, 0, 0.1]]$ and $S_1 \sim \text{Multinoulli}(0.25, 0.25, 0.25, 0.25)$, what is

$$P(\boldsymbol{S} = \text{ACGAA})?$$

Probability Examples 2

3. You collect sequence data from permissive and nonpermissive cells, run a program to count the total number of A, C, G, and T.
    1. What is an outcome of this random experiment?
    2. What is the sample space of this experiment?
    3. Using the notation provided above, what must be true of the data for the relative G/C content to be higher in the permissive cells of the observed sample?
    4. What is the probability of the event mentioned in 3 as $n_n = n_p = n \to \infty$ and nucleotides are iid Multinoulli$(0.25, 0.25, 0.25, 0.25)$?
    5. How can you estimate this probability using simulation? If I told you $n = 100$? If I told you $n \sim$ Poisson(300)?

Probability Examples 2

3. You collect sequence data from permissive and nonpermissive cells, run a program to count the total number of A, C, G, and T.
   1. What is an outcome of this random experiment?

Example

Cell type	A	C	G	T
Nonpermissive cells	26	27	21	26
Permissive cells	30	22	23	25

   2. What is the sample space of this experiment?
   3. Using the notation provided above, what must be true of the data for the relative G/C content to be higher in the permissive cells of the observed sample?
   4. What is the probability of the event mentioned in 3 as $n_n = n_p = n \to \infty$ and nucleotides are iid Multinoulli$(0.25, 0.25, 0.25, 0.25)$?
   5. How can you estimate this probability using simulation? If I told you $n = 100$? If I told you $n \sim$ Poisson(300)?

Probability Examples 2

3. You collect sequence data from permissive and nonpermissive cells, run a program to count the total number of A, C, G, and T.
   1. What is an outcome of this random experiment?
   2. What is the sample space of this experiment?

Example

The collection of all sets of counts for A, C, G, and T in permissive and nonpermissive samples. Formally,

$$\Omega = \left\{ (\boldsymbol{n}_p, \boldsymbol{n}_n) : n_{hi} \in \{0, \mathbb{Z}^+\}, h \in \{p, n\}, i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \right\}.$$

3. Using the notation provided above, what must be true of the data for the relative G/C content to be higher in the permissive cells of the observed sample?
4. What is the probability of the event mentioned in 3 as $n_n = n_p = n \to \infty$ and nucleotides are iid Multinoulli$(0.25, 0.25, 0.25, 0.25)$?
5. How can you estimate this probability using simulation? If I told you $n = 100$? If I told you $n \sim$Poisson(300)?

Probability Examples 2

3. You collect sequence data from permissive and nonpermissive cells, run a program to count the total number of A, C, G, and T.
   1. What is an outcome of this random experiment?
   2. What is the sample space of this experiment?
   3. Using the notation provided above, what must be true of the data for the relative G/C content to be higher in the permissive cells of the observed sample?

Example

$$\left\{ \frac{n_{p\mathrm{G}} + n_{p\mathrm{C}}}{n_p} > \frac{n_{n\mathrm{G}} + n_{n\mathrm{C}}}{n_n} \right\},$$

where $n_n = n_{n\mathrm{A}} + n_{n\mathrm{C}} + n_{n\mathrm{G}} + n_{n\mathrm{T}}$ is the total nucleotide count in nonpermissive cells, and n_p the total in permissive cells.

4. What is the probability of the event mentioned in 3 as $n_n = n_p = n \to \infty$ and nucleotides are iid Multinoulli$(0.25, 0.25, 0.25, 0.25)$?
5. How can you estimate this probability using simulation? If I told you $n = 100$? If I told you $n \sim$ Poisson(300)?

Probability Examples 2

3. You collect sequence data from permissive and nonpermissive cells, run a program to count the total number of A, C, G, and T.
   1. What is an outcome of this random experiment?
   2. What is the sample space of this experiment?
   3. Using the notation provided above, what must be true of the data for the relative G/C content to be higher in the permissive cells of the observed sample?
   4. What is the probability of the event mentioned in 3 as $n_n = n_p = n \to \infty$ and nucleotides are iid Multinoulli$(0.25, 0.25, 0.25, 0.25)$?

Example

$$P\left(\left\{\frac{n_{pG} + n_{pC}}{n_p} > \frac{n_{nG} + n_{nC}}{n_p}\right\}\right) = 0.5.$$

5. How can you estimate this probability using simulation? If I told you $n = 100$? If I told you $n \sim$ Poisson(300)?

Statistics

- **(Statistical) Sample: simple random sample**
- **(Statistical) Population:** set of similar items or events on which we would like to make inference.
- **Statistical inference:** parameter **estimation**, **hypothesis testing**.
- **Random variable:** A real number computed on that outcome of a random experiment, a function that operates on outcomes.
- **Statistic:** A real number computed from the data in a sample, a function that operates on data (a type of random variable). There are **estimators** and **test statistics**.
- ***p*-value:** The probability of data *as or more extreme* than the observed data. Also, the probability we are wrong in concluding that a null hypothesis H_0 (defined later) is not correct.
- **Vocabulary**: significance level, bias, variance, type I error, type II error

Statistics – Hypothesis Testing

- Start with a generic null hypothesis and identify *test statistics* sensitive to the truth of the null or relevant alternative hypothesis.
- Finalize a specific null hypothesis H_0, a *probability model* that mimics the data generation process and *test statistic* T.
- Identify values of the *test statistic* that are inconsistent with H_0.
- Collect data and compute the *observed test statistic* t. Compute the *p-value*:

$$P(\{T \text{as or more extreme than } t\} \mid H_0).$$

Statistics – Resampling

(General) Algorithm: Mimic the randomness/uncertainty of the random experiment using a computer.

- **Input**: the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$, a large number $B \in \mathbb{Z}^+$ for the number of times to repeat, and a model H_0 (constructed and confirmed with a biologist).
- Loop B times: at iteration i
  - Generate a **simulated** data set $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$
  - Compute and store the test statistic: $T^{(i)} = T(\boldsymbol{x}^{(i)})$.
  - (If simulator directly makes $T^{(i)}$, then only one step.)
- Compute the *observed test statistic*: $t = T(\boldsymbol{x})$.
- **Output:** Compute the p-value as the proportion of simulation samples where $T^{(i)}$ is as or more extreme (shows more signal) than the observed test statistic t.

$$P(\{T \text{ as more more extreme than } t\} \mid H_0)$$

$$\approx \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\left\{T^{(i)} \text{ as or more extreme than } t\right\},$$

Python Programming

- Benefits of programming instead of manual processing.
- Pros, cons of Python.
- Name two recommended coding practices.

Machine Learning

- ML emphasis: prediction, general-purpose, automatic, big data
- ML types: supervised (classification, regression), unsupervised, clustering, reinforcement, semi-supervised
- ML vocabulary: training, validation, testing, feature importance, feature correlation, feature selection, complexity, linear separability, margin, bagging, cross-validation.
- ML metrics: accuracy, entropy, gini index, type I error (FP), type II error (FN), precision (1 - FPR) = TP/(TP+FP), recall/TPR/sensitivity = TP/(TP+FN)), specificity = TN/(FP+TN), confusion matrix, ROC (TPR vs. FPR), Matthew's correlation coefficient (MCC)
- Methods: *k*-NN, decision tree, linear classifier, random forest

ML Examples

- Which are supervised classification? spam filtering, face detection, spoken language detection (English, French, Chinese, *etc.*), yield prediction, Netflix recommendation

Models

- **Multinoulli**:
  $S_1, S_2, S_3, \ldots, S_n \overset{\text{iid}}{\sim}$ Multinoulli(p_A, p_C, p_G, p_T) or
  $X_1, X_2, X_3, \ldots, X_m \overset{\text{iid}}{\sim}$
  Multinoulli$(p_A, p_C, p_D, p_E, p_F, p_G, p_H, p_I, p_K, \ldots)$, IID codons.
- **Markov chain**: transition matrix
- **Permutation model**.
- ***k*-Nearest Neighbors** (supervised – classification). Pros, cons
- **Decision tree** (supervised – classification or regression): tree depth, gini index, overfitting
- **Linear classifier** (supervised –classification): $f(\boldsymbol{x}) = w^T \boldsymbol{x} + b$, if $f(\boldsymbol{x}_i) \geq 0$, then predict $y_i = 1$, otherwise $y_i = -1$.
- **Random forest** (supervised – classification or regression): bootstrap aggregating (bagging), randomness (feature bagging)

Tasks

- Detect mutation comparing two samples with and without.
- Detect motif comparing two samples with and without.
- Classify sequences as permissive/nonpermissive.
- Classify yield into $\{\text{Very Low}, \text{Low}, \text{High}, \text{Very High}\}$.
- Find coding genes, introns, exons, horizontal gene transfer events, transcript start sites (TSS), CpG islands, unusually frequenty k-mers in single set of sequence data.

Test statistics

- Sums of squared deviations.
- Odds ratio, ratios of odds ratios.
- Whatever you like that is responsive to truth of H_0!

Algorithms/Recipes

Some algorithm are *specific* enough to implement given input, output, steps, and little else. Others, such as hypothesis testing and resampling, are more *general* and require additional user intervention, for example, to choose a model.

- Hypothesis testing (general).
- Resampling (general).
- Training, testing, validation, predicting (general).
- *k*-NN Algorithm (specific).
- Decision tree (specific).
- Random forest (specific).
- Dr. F's ORF finder (specific).