Choose the Right Probability Distribution (Computational Biology and Bioinformatics)

Dr. F.J. Rodenburg, L. Outhuis & L. Bikker

2022-08-08

These exercises assume you are familiar with the basics of **probability distributions**. That means that you should know when the following distributions are good approximations or not:

- The normal distribution
- The Poisson distribution
- The binomial distribution

For these questions, you have to **identify the outcome** from a short description of a study, and **choose the appropriate probability distributions** that could be used in the analysis.

Q1 — SNP Frequency

Single nucleotide polymorphisms (SNPs) are alleles differing by a single base, the most common type of genetic variation. Over half a billion SNPs have been reported in human DNA.[1] If researchers want to assess how common a SNP is in a certain demographic, what would be the outcome and what probability distribution could be used to approximate it?

Answer:

Q2 — RNA-Seq

RNA-Seq is a technique to examine the presence and quantity of RNA in a sample, using next generation sequencing. The number of reads can then tell you which genes are expressed more or less at a given time. What kind of outcome does an RNA-Seq experiment yield and what probability distribution can be associated with it?

Answer:

Q3 — Personalized Medicine

A very popular topic in clinical science is that of personalized medicine (i.e., precision medicine). Genomics, transcriptomics and proteomics techniques are used to explore the individual differences in disease-processes. Though it is impossible to design an optimal drug for every patient, often certain archetypes arise, such as non-responders, or highly sensitive individuals. In cases where differential response to a medicine can be traced back to differences at the molecular level, treatment strategies can be 'personalized' based on patient characteristics.[2, see 3 for an example.]

Suppose you have data of known concentrations deemed optimal for a large set of patients and you want to train a model to predict the concentration of a medicine that should be applied to new patients. What would be the outcome and what probability distribution could be used to approximate it?

Answer:

Q4 — Action Potentials

Neurons create electrical pulses that transfer information through the brain, called action potentials. Action potentials are measured in a neuron fixed to a microelectrode. The researcher wants to know how often an action potential is generated in this specific neuron within a specific time frame. What would be the outcome and what probability distribution can this be approximated with?

Answer:

Q5 — GC-Content

To create a database of GC-content of species (% guanine or cytosine bases in DNA), the relative amount of A, C, T & G bases are counted after sequencing experiments. When estimating GC-contents, what probability distribution can be used to approximate it?

Answer:

Q6 — Protein Molecular Weight

Tools like bioinformatics.org/sms/prot_mw.html can compute the molecular weight of a protein based on its amino acid sequence. If you are studying how the molecular weight of proteins relates to their ability to form interactions with other proteins, what would be the outcome and with which probability distribution could it be approximated?

Answer:

References

1. https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/ [Online; accessed 11-August-2022].

2. Goetz, L. H. & Schork, N. J. Personalized medicine: Motivation, challenges, and progress. *Fertility and Sterility* **109**, 952–963 (2018).

3. Belzeaux, R. *et al.* Responder and nonresponder patients exhibit different peripheral transcriptional signatures during major depressive episode. *Translational Psychiatry* **2**, e185–e185 (2012).