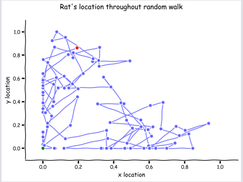
Neuromatch Academy: Statistics Refresher - Summary Sheet[1]

Statistics (W0D5T1) - Probability Distributions

Random Walk

Stochastic models can be used to create models of behaviour. As an example, imagine that a rat is placed inside a novel environment, a box. We could try and model its exploration behaviour by assuming that for each time step it takes a random uniformly sampled step in any direction (simultaneous random step in x direction and random step in y direction)



Discrete Distributions Differentiation

A simple way to model random behaviour is with a single **Bernoulli trial**, that has two outcomes, $Left, Right$, with probability $P(Left) = p$ and $P(Right) = 1 - p$ as the two mutually exclusive possibilities (whether the rat goes down the left or right arm of the maze).

The binomial distribution simulates n number of binary events, such as the $Left, Right$ choices of the random rat in the T-maze is given by:

$$P(k|n,p) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k (1-p)^{n-k} \tag{1}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2}$$

where, p is the probability of turning left, n is the number of binary events, or trials, and k is the number of times the rat turned left. The term $\binom{n}{k}$ is the binomial coefficient.

The Poisson distribution is a 'point-process', meaning that it determines the number of discrete 'point', or binary, events that happen within a fixed space or time, allowing for the occurrence of a potentially infinite number of events. The Poisson distribution is specified by a single parameter λ that encapsulates the mean number of events that can occur in a single time or space interval. The formula for a Poisson distribution is:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{3}$$

where λ is a parameter corresponding to the average outcome of x.

Statistics (W0D5T1) - Probability Distributions

Continuous Distributions

We do not have to restrict ourselves to only probabilistic models of discrete events. While for discrete outcomes we can ask about the probability of an specific event ("what is the probability this neuron will fire 4 times in the next second"), this is not defined for a continuous distribution ("what is the probability of the BOLD signal being exactly 4.000120141..."). Hence we need to focus on intervals when calculating probabilities from a continuous distribution.

If we want to make predictions about possible outcomes for a continuous distribution we can use the integral

$$\int_{x_1}^{x_2} P(x)dx \tag{4}$$

where $P(x)$ is a probability density function.

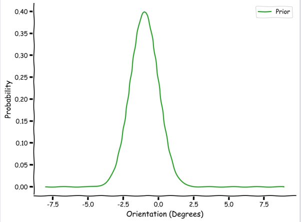
Gaussian Distribution

The most widely used continuous distribution is probably the Gaussian (also known as Normal) distribution. It is extremely common across all kinds of statistical analyses. Because of the central limit theorem, many quantities are Gaussian distributed. Gaussians also have some nice mathematical properties that permit simple closed-form solutions to several important problems.

The equation for a Gaussian probability density function is:

$$f(x;\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{5}$$

where $\mu = -1$ is the mean, $\sigma = 1$ is the variance and x is the random variable for orientation (degrees).



Statistics (W0D5T2) - Statistical Inference

Basic Probability

Probability has to be in the range 0 to 1 $P(A) \in [0,1]$ and the complementary can always be defined as

$$P(\neg A) = 1 - P(A).$$

When we have two variables, the **conditional probability** of A given B is

$$P(A|B) = P(A \cap B)/P(B) = P(A,B)/P(B)$$

while the **joint probability** of A and B is

$$P(A \cap B) = P(A,B) = P(B|A)P(A) = P(A|B)P(B)$$

We can then also define the process of **marginalisation** (for discrete variables) as

$$P(A) = \sum P(A,B) = \sum P(A|B)P(B)$$

where the summation is over the possible values of B.

As an example if B is a binary variable that can take values $B+$ or $B0$ then

$$P(A) = \sum P(A,B) = P(A|B+)P(B+) + P(A|B0)P(B0).$$

For continuous variables marginalization is given as

$$P(A) = \int P(A,B)dB = \int P(A|B)P(B)dB.$$

Markov chains

The Markov property specifies that you can fully encapsulate the important properties of a system based on its current state at the current time, any previous history does not matter. It is memoryless.

As an example imagine that a rat is able to move freely between 3 areas: a dark rest area ($state = 1$), a nesting area ($state = 2$) and a bright area for collecting food ($state = 3$). Every 5 minutes (timepoint i) we record the rat's location. We can use a ''categorical distribution'' to look at the probability that the rat moves to one state from another.

We can model this as a Markov chain, so the animal is only in one of the states at a time and can transition between the states. We want to get the probability of each state at time $i + 1$.

$$\begin{aligned} P(state_{i+1} = 1) = \\ P(state_{i+1} = 1|state_i = 1)P(state_i = 1) \\ +P(state_{i+1} = 1|state_i = 2)P(state_i = 2) \\ +P(state_{i+1} = 1|state_i = 3)P(state_i = 3) \end{aligned}$$

[1] 't Hart et al., (2022). Neuromatch Academy: a 3-week, online summer school in computational neuroscience. Journal of Open Source Education, 5(49), 118. https://doi.org/10.21105/jose.00118

Statistics (W0D5T2) - Statistical Inference

Statistical inference and likelihood

If we do not know the parameters μ, σ that generated the data, we can try to **infer** which parameter values (given our model) gives the best (highest) likelihood. This is what we call statistical inference: trying to infer what parameters make our observed data the most likely or probable.
A generative model (such as the Gaussian distribution from the previous tutorial) allows us to make predictions about outcomes.
After we observe n data points, we can evaluate our model (and any of its associated parameters) by calculating the **likelihood** of our model having generated each of those data points x_i.

$$P(x_i|\mu,\sigma) = \mathcal{N}(x_i,\mu,\sigma) \tag{6}$$

For all data points $\mathbf{x} = (x_1, x_2, x_3, ...x_n)$ we can then calculate the likelihood for the whole dataset by computing the product of the likelihood for each single data point.

$$P(\mathbf{x}|\mu,\sigma) = \prod_{i=1}^{n} \mathcal{N}(x_i,\mu,\sigma) \tag{7}$$

While the likelihood may be written as a conditional probability ($P(x|\mu,\sigma)$), we refer to it as the **likelihood function**, $L(\mu,\sigma)$. This slight switch in notation is to emphasize our focus: we use likelihood functions when the data points $\mathbf{x}$ are fixed and we are focused on the parameters.
Our new notation makes clear that the likelihood $L(\mu,\sigma)$ is a function of μ and σ, not of $\mathbf{x}$.

Maximum likelihood

Implicitly, by looking for the parameters that give the highest likelihood in the last section, we have been searching for the **maximum likelihood** estimate

$$(\hat{\mu},\hat{\sigma}) = \text{argmax}_{\mu,\sigma} L(\mu,\sigma) = \text{argmax}_{\mu,\sigma} \prod_{i=1}^{n} \mathcal{N}(x_i,\mu,\sigma). \tag{8}$$

We want to do inference on this data set, i.e. we want to infer the parameters that most likely gave rise to the data given our model. Intuitively that means that we want as good as possible a fit between the observed data and the probability distribution function with the best inferred parameters. We can search for the best parameters manually by trying out a bunch of possible values of the parameters, computing the likelihoods, and picking the parameters that resulted in the highest likelihood.

Statistics (W0D5T2) - Statistical Inference

Bayesian Inference

For Bayesian inference we do not focus on the likelihood function $L(y) = P(x|y)$, but instead focus on the posterior distribution:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{9}$$

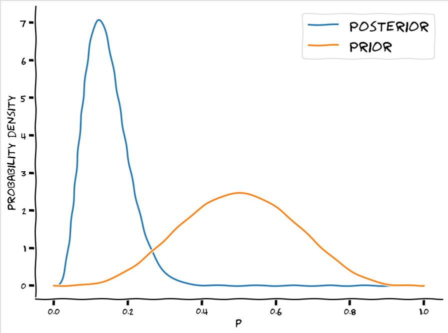
which is composed of the **likelihood** function $P(x|y)$, the **prior** $P(y)$ and a normalising term $P(x)$ (which we will ignore for now).
While there are other advantages to using Bayesian inference (such as the ability to derive Bayesian Nets, see optional bonus task below), we will start by focusing on the role of the prior in inference.

Conjugate priors

Bayesian inference can be used for any likelihood distribution, but it is a lot more convenient to work with **conjugate** priors, where multiplying the prior with the likelihood just provides another instance of the prior distribution with updated values.
For the binomial likelihood it is convenient to use the beta distribution as a prior

$$f(p;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{10}$$

where B is the beta function, α and β are parameters, and p is the probability of the rat turning left or right. The beta distribution is thus a distribution over a probability.
Given a series of Left and Right moves of the rat, we can now estimate the probability that the animal will turn left.
Using Bayesian Inference, we use a beta distribution **prior**, which is then multiplied with the **likelihood** to create a **posterior** that is also a beta distribution, but with updated parameters.


POSTERIOR
PRIOR
PROBABILITY DENSITY
P