BioMath - Probability - September 2, 2014

There is a list of formulas on the last sheet of the problem set. Refer to them if you need to.

1. **Sums of random variables**

   Let's remember that for a probability distribution $p(x)$ over the random variable x, the expectation operator is defined as

   $$E\{\ (\cdot)\ \} = \sum_x (\cdot)\ p(x) \tag{1}$$

   where $(\cdot)$ stands for "whatever you're interested in taking the expected value of." For example, when you replace $(\cdot)$ with x, you're looking at the expected value of x; when instead you replace $(\cdot)$ with x^2, you're looking at the expected value of x^2; and so on.

   Because of this definition, it follows that the expectation operator is linear, and therefore

   $$E\{x + y\} = E\{x\} + E\{y\}. \tag{2}$$

   Let's also remember that the mean of a random variable z is defined as $E\{z\}$; and that the variance of a random variable is defined as $E\{z^2\} - E\{z\}^2$.

   • Use equation (2) to show that if z is the sum of two random variables, $z = x + y$, then the mean of z is the sum of the mean of x plus the mean of y.

   • Show that if x and y are independent, then $E\{xy\} = E\{x\}E\{y\}$.

   • Use equation (2) again to show that if x and y are independent, then the variance of z is the variance of x plus the variance of y. For simplicity, you can focus on the case in which $E\{x\} = 0$ and $E\{y\} = 0$.

2. **Diffusion**

   A neuron with a 1 cm long axon releases neurotransmitter into the synaptic cleft, the space between two neurons. The synaptic cleft is half a micron wide. In body temperature salt water, the neurotransmitter has a diffusion constant of $10^{-9}\text{m}^2/\text{sec}$.

   • Roughly speaking, how long does it take for a significant amount of neurotransmitter to diffuse across the synaptic cleft?

• How long would it take for a significant amount of neurotransmitter to diffuse the length of the axon?

3. **More sums of random variables**

   You're doing a development experiment in which on each attempt of your experiment, you get either 1, or, two, or three dividing cells (we'll call the number of dividing cells y), and there are 1, or, two, or three genes of interest expressed (we'll call this number x). The possible outcomes occur with the following joint probability:

y=1	3/18	1/18	3/18
y=2	1/18	2/18	1/18
y=3	3/18	1/18	3/18
	x=1	x=2	x=3

   • What is the distribution of the random variable $z = x + y$? Make sure that youve sorted the axes correctly so the diagonals youre summing along are diagonals with constant z (that is, should y grow from top to bottom or grow from bottom to top? Either choice is fine as long as what you sum over is correct.)

4. **Still more sums of random variables**

   Let's consider again the joint distributions of branches and puncta that we considered in the last problem set. You got P_{joint1} and P_{joint2} from loading `Pjoint.mat` from `distributions.zip` on the wiki. The problem read:

   > You're looking over neurons in a slice. You are interested in the relationship between presynaptic terminals and number of dendritic branches. You've stained for **synapsin**, a marker of presynaptic terminals. For each neuron, you count the number of dendritic branches, and you also count the number of synapsin puncta that are close to that neuron's dendrites. You do the experiment twice, once with a slice from a wild-type mouse, and the nenxt time with a slice from an MHC-knockout mouse. Load `Pjoint.mat`. You'll find two joint probability distributions, representing the outcome of the two experiments. You can plot them with:

```
>> figure(1);
>> imagesc(nbranches, npuncta, Pjoint1); axis xy;
>> xlabel('nbranches'); ylabel('npuncta'); colormap jet;
colorbar

>> figure(2);
>> imagesc(nbranches, npuncta, Pjoint2); axis xy;
```

```
>> xlabel('nbranches'); ylabel('npuncta'); colormap jet;
colorbar
```

$Pjoint1(m,n)$ is the fraction of neurons for which you counted n dendritic branches and m synpasin puncta.

You will remember that one of the distributions, P_{joint1} had independent branches and puncta, and the other, P_{joint2} , had dependent branches and puncta. Yet they both had the same marginal distributions $P(n_{\text{branches}})$ and $P(n_{\text{puncta}})$.

• For each of the two joint distributions, compute the distribution $P_z(z = n_{\text{branches}}+n_{\text{puncta}})$ as well as the two marginals you had obtained before (namely, $P(n_{\text{branches}})$ and $P(n_{\text{puncta}})$. Now compute the variance of z, and compare it to the sum of the variance of n_{branches} plus the variance of n_{puncta}. Do the two match for one of the two distributions? For which distribution is the variance of z higher?

5. **Sums of continuous random variables**

Consider the continuous probability density for the exponential distribution

$$P(t) = \lambda e^{-\lambda t} \tag{3}$$

• Plot this distribution as a function of time for $\lambda = 0.1$, $\lambda = 1$, and $\lambda = 10$.

• Show that the integral of $P(t)$ over time, running from 0 to infinity, equals 1.

• What is the mean of t under this distribution? The following integration by parts formula, on the right hand side of the implication symbol, will come in handy (the prime below means derivative):

$$[fg]' = f'g + fg' \Longrightarrow \int_a^b fg = [fg]_a^b \int_a^b fg' \tag{4}$$

One of your parts will be t. The other will be $e^{\lambda t}$. You have to decide which one you want to call f and which one you want to call g.

• Now for the fun part! Youve just landed a faculty job, and using a significant fraction of your precious start-up money, youve bought a very expensive piece of machinery. This machine is very sensitive to voltage fluctuations from the power lines, so you buy two surge protectors; when one of them dies, youll immediately replace it with the other. Each protector has a lifetime described by an exponential distribution with rate λ. What is the distribution of the times z at which *both* protectors will have died? Plot this as a function of z. This shape is sometimes called an α function, and is often used as an approximate description of currents through ligand-gated channels.

6. **Convolution**

   The sum of two independent random variables $z = x + y$ has a distribution given by the convolution of the distribution of x and the distribution of y. The convolution formula is

$$p_z(z) = \sum_x p_x(x) p_y(z - x) \tag{5}$$

   • Write a Matlab function that, given vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ that define two distributions over the outcomes in the vector $x = -20 : 0.01 : 20$, computes the convolution of the two distributions, and returns a distribution vector defined over the same space of outcomes x.

   • Now consider $\mathbf{p}_1 \propto e^{-x/0.5}$, as well as the distribution $\mathbf{p}_2 = 1$ for $x = 2$ and zero elsewhere. Remember to normalize your vectors so that $\sum_x p(x) = 1$. What is their convolution? Plot the result as a function of x.

   • Repeat this exercise for $\mathbf{p}_2 = 1/3$ for each of $x = 2, x = 7, x = 12$ and zero elsewhere.

   • Repeat this exercise for $\mathbf{p}_2 = 1/3$ for each of $x = 2, x = 2.8, x = 3.6$ and zero elsewhere. Can you predict what you would expect to see if $\mathbf{p}_2 = 1/3$ for each of $x = 2, x = 2.1, x = 2.2$?

   • What is the mean of $\mathbf{p}_1$? What is the variance of $\mathbf{p}_1$? So if z is the sum of N independent samples of $\mathbf{p}_1$, what is the expected mean of z and variance of z? Call these μ_N and σ_N^2, respectively.

   • Plot the convolution of $\mathbf{p}_1$ with itself. Does the result you get match the intuitions you built up in the previous bullet points? Then convolve your result with $\mathbf{p}_1$again. Do this for each of $N = 10$ iterations. At each iteration, plot the result, and on the same axes, plot a Gaussian with mean μ_N and variance σ_N^2.

Formulae

• Poisson distribution

$$F(k|\mu) = \frac{\mu^k}{k!} e^{-\mu} \tag{6}$$

k is the number of events
μ is the expected (average) number of events

- Exponential distribution

$$D(t|r) = re^{-rt} \tag{7}$$

r is the expected rate of events

- Gaussian distribution

$$D(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \tag{8}$$

μ is the mean
σ is the standard deviation

- Bayes rule

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{9}$$

X and Y are two random variables