Remember when we did a continuum of hypotheses, and we called it "beast mode"? Well, I looked up some gamer terms, and above beast mode is God mode. Although I don't think what we are about to do is worthy of divinity, it is another step up. I am going to call it

A Multi-Dimensional Continuum of Hypotheses

And when we are done deriving Bayes Theorem in this case, we could go even further and apply it to Bayesian conjugates for distributions with two parameters, such as μ and σ that characterize bacteria survival times.

(Also, what I am about to do <u>will not be on any exam</u>. I promised a course that didn't have calc as a prereq. We introduced the calc we needed as it came up. But this new subject takes us into multi-variable calculus!! That is just too much new material to do more than brief exposure.

Let's build this up carefully. Suppose you have a model that has two parameters in it. The model might be

$$\underset{\text{pressure}}{p} = a\overset{\text{time}}{t} + b$$

the parameters

Or suppose you have a model that has two <u>independent</u> hypotheses:

H	$\sim H$
paper was written by Hamilton	paper was written by Madison

The other independent hypothesis is the year of authorship

E	$\sim E$
authored in 1787	authored in 1788

So there are 4 mutually exclusive possibilities!

H, E $H, \sim E$

$\sim H, E$ $\sim H, \sim E$

We are going to need a new version of Bayes Theorem!

To make things more general, I am going to make H and $\sim H$ part of a list of N_H hypotheses,

$$H_0, H_1, H_2, \ldots, H_{N_H-2}, H_{N_H-1}$$

and E part of a list of N_E hypotheses,

$$E_0, E_1, E_2, \ldots, E_{N_E-2}, E_{N_E-1}$$

Bayes theorem in this case comes from writing

$$P(H_i, E_j \cap data)$$

$$= P(H_i, E_j / data) P(data)$$

OR A 2ND WAY

$$P(H_i, E_j \cap data)$$

$$= P(data / H_i, E_j) P(H_i, E_j)$$

THEN WE EQUATE THOSE TWO WAYS OF WRITING $P(H_i, E_j \cap data)$ AND GET

$$P(H_i, E_j / data) P(data)$$

$$= P(data / H_i, E_j) P(H_i, E_j)$$

THIS IS ALL OLD HAT. WE SOLVE:

$$P(H_i, E_j / data) = \frac{P(data / H_i, E_j) P(H_i, E_j)}{P(data)}$$

FINALLY WE EXPAND THE DENOMINATOR AND GET

$$P(H_i, E_j / data) = \frac{P(data / H_i, E_j) P(H_i, E_j)}{\sum_{k=0}^{N_H-1} \sum_{l=0}^{N_E-1} P(data / H_k, E_l) P(H_k, E_l)}$$

OF COURSE, WE ARE SOMETIMES SLOPPY (AS YOU HAVE NOW SEEN IN MULTIPLE CONTEXTS), AND WE USE i AND j IN THE DENOMINATOR, EVEN THOUGH THAT i AND j HAS NOTHING TO DO WITH THE i AND j IN THE NUMERATOR. THEN WE HAVE,

$$P(H_i, E_j / data) = \frac{P(data / H_i, E_j) P(H_i, E_j)}{\sum_{i=0}^{N_H-1} \sum_{j=0}^{N_E-1} P(data / H_i, E_j) P(H_i, E_j)}$$

We are not in "God-mode" yet. We just have a fancy way of covering $N_H \cdot N_E$ discrete, mutually-exclusive pairs of hypotheses:

$$P(H_i, E_j | data) = \frac{P(data|H_i, E_j) P(H_i, E_j)}{\sum_{i=0}^{N_H-1} \sum_{j=0}^{N_E-1} P(data|H_i, E_j) P(H_i, E_j)}$$

The Multi-Dimensional Continuum

Imagine that H_i represents the chance that a continuous variable μ is between μ_i and $\mu_i + \Delta\mu_i$ and E_j represents the chance that a continuous variable σ is between σ_j and $\sigma_j + \Delta\sigma_j$

Then

this eqn. is approximate (!) because we just used one corner of the volume to estimate the volume's height

$$P(\mu_i, \sigma_j | data) \Delta\mu_i \Delta\sigma_j = \frac{P(data|\mu_i, \sigma_j) P(\mu_i, \sigma_j) \Delta\mu_i \Delta\sigma_j}{\sum_{k=0}^{N_H-1} \sum_{l=0}^{N_E-1} P(data|\mu_k, \sigma_l) \cdot P(\mu_k, \sigma_l) \Delta\mu \Delta\sigma}$$

a width

another width

In the numerator, μ_i, σ_j, $\Delta\mu_i$ and $\Delta\sigma_j$ can be whatever you like. In the denominator, because the sums have to range over all possible values of μ and σ, we must have

$$\Delta\mu = \frac{\mu_{max} - \mu_{min}}{N_H} \quad \text{and} \quad \Delta\sigma = \frac{\sigma_{max} - \sigma_{min}}{N_E}$$

Of course, if μ and σ are continuous, the only way this is going to be a precise expression is if N_H and N_E are large. We express this by writing the denominator as

$$\lim_{N_H \to \infty} \lim_{N_E \to \infty} \sum_{k=0}^{N_H-1} \sum_{l=0}^{N_E-1} P(data|\mu_k, \sigma_l) \cdot P(\mu_k, \sigma_l) \Delta\mu \Delta\sigma$$

There is a special symbol from multivariable calculus that capture this double-sum idea, and it is

$$\int_{\mu_{min}}^{\mu_{max}} \int_{\sigma_{min}}^{\sigma_{max}} P(data|\mu, \sigma) P(\mu, \sigma) \, d\mu \, d\sigma$$

The interpretation is that this is the volume under the two-dimensional sheet

$$P(data|\mu, \sigma) P(\mu, \sigma)$$

Often, the lower and upper limits of integration μ_{min}, μ_{max}, σ_{min}, and σ_{max} are either infinite or unspecified. In that case, we write the denominator as

$$\iint P(\text{data}|\mu,\sigma)P(\mu,\sigma)\,d\mu\,d\sigma$$

Also, in the numerator, we leave the subscripts off, since now μ and σ can be anything, not some discrete set of values that becomes larger and larger as $N_H \to \infty$ and $N_E \to \infty$

Then instead of

$$P(\mu_i,\sigma_j|\text{data})\Delta\mu_i\Delta\sigma_j = \frac{P(\text{data}|\mu_i,\sigma_j)P(\mu_i,\sigma_j)\Delta\mu_i\Delta\sigma_j}{\sum_{k=0}^{N_H-1}\sum_{l=0}^{N_E-1} P(\text{data}|\mu_k,\sigma_l)\cdot P(\mu_k,\sigma_l)\Delta\mu\Delta\sigma}$$

we have

$$P(\mu,\sigma|\text{data})\Delta\mu\Delta\sigma = \frac{P(\text{data}|\mu,\sigma)P(\mu,\sigma)\Delta\mu\Delta\sigma}{\iint P(\text{data}|\mu,\sigma)P(\mu,\sigma)\,d\mu\,d\sigma}$$

Compare this with Donovan & Mickey Eq. 12.10 on p. 180.

It is the exact same except that they don't have the $\Delta\mu$ and $\Delta\sigma$ multiplying the numerators.

You can of course cancel those off in my equation, but I prefer to keep them around because they aid in the interpretation of the probability density.

Now I will cancel $\Delta\mu\Delta\sigma$ off since it is in the numerator on both sides, and we know how to put it back if we want to restore the interpretation of the probability density.

Bayesian Conjugates with Gaussian Data

On the last four pages, we derived

$$P(\mu, \sigma | \text{data}) = \frac{P(\text{data}|\mu, \sigma) P(\mu, \sigma)}{\iint P(\text{data}|\mu, \sigma) P(\mu, \sigma)\, d\mu\, d\sigma}$$

We did this in the midst of chapter 12 because our likelihoods in chapter 12 involve two parameters. Our canonical example is bacteria survival times:

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

survival time

mean survival time

standard deviation of survival times

It turns out that the algebra is a little tidier if we let $\tau = \frac{1}{\sigma^2}$ and write the likelihood as

$$P(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}}\, e^{-\tau(x-\mu)^2/2}$$

The version of Bayes' Theorem at left becomes

$$P(\mu, \tau | \text{data}) = \frac{P(\text{data}|\mu, \tau) P(\mu, \tau)}{\iint P(\text{data}|\mu, \tau) P(\mu, \tau)\, d\mu\, d\tau}$$

Let us remember the goal of Chapters 10-12. It is to find magical priors called "Bayesian conjugates" whose magical property is that the functional form of the posterior (which is the product of the likelihood and the prior) is the same (but with different parameters) as the functional form of the prior. So that is what we are seeking for the likelihood at left.

And now I just have to write the magical prior down, and then do the nasty algebra to show that it is conjugate to the Gaussian likelihood. Without further ado:

$$\text{"prior"} = P(\mu,\tau) = P(\mu|\tau)\,P(\tau)$$

where

$$P(\mu|\tau) = \sqrt{\frac{n_0\tau}{2\pi}}\; e^{-n_0\tau(\mu-\mu_0)^2/2}$$

and

$$P(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\tau^{\alpha-1} e^{-\beta\tau}$$

This formula for $P(\tau)$ is the same one that showed up in Chapter 11 for the prior that is conjugate to Poisson distributions was introduced.

I am not going to do the derivation in this handout, and to be frank, I haven't done it yet myself. In Michael I. Jordan's Stat 260 handout for February 8th, 2010, on p.6 he says that if the

and might not find time to ☹

likelihood for n new data points x_i is Gaussian and independently and identically distributed (i.i.d.) and the prior is

$$x_i \,|\, \mu,\tau \sim \mathcal{N}(\mu,\tau) \quad \text{i.i.d.}$$
$$\mu \,|\, \tau \sim \mathcal{N}(\mu_0, n_0\tau)$$
$$\tau \sim \mathrm{Ga}(\alpha,\beta)$$

then the posterior has the same form with new parameters

$$\mu \,|\, \tau, x \sim \mathcal{N}\left(\frac{n\tau}{n\tau+n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau+n_0\tau}\mu_0 \quad , \quad n\tau+n_0\tau\right)$$
$$\tau \,|\, x \sim \mathrm{Ga}\left(\alpha+\frac{n}{2} \quad , \quad \beta+\frac{1}{2}\sum(x_i-\bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right)$$

In Jordan's notation the curly $\mathcal{N}$ represents a "normal" distribution, and is what we call a Gaussian. The "Ga" represents a gamma distribution and is identical to the gamma distribution Donovan and Mickey have in Eq 11.8 on p.159 and that was also in my Bayesian conjugates handout.

I do not trust these formulas (Jordan has grad students write the course notes he posts), let alone claim to understand them. I would need to do the derivation myself.

What you can count on in this handout is everything up to but not including this last page of claims.

But if you want a serious challenge, you can verify the rest ☺