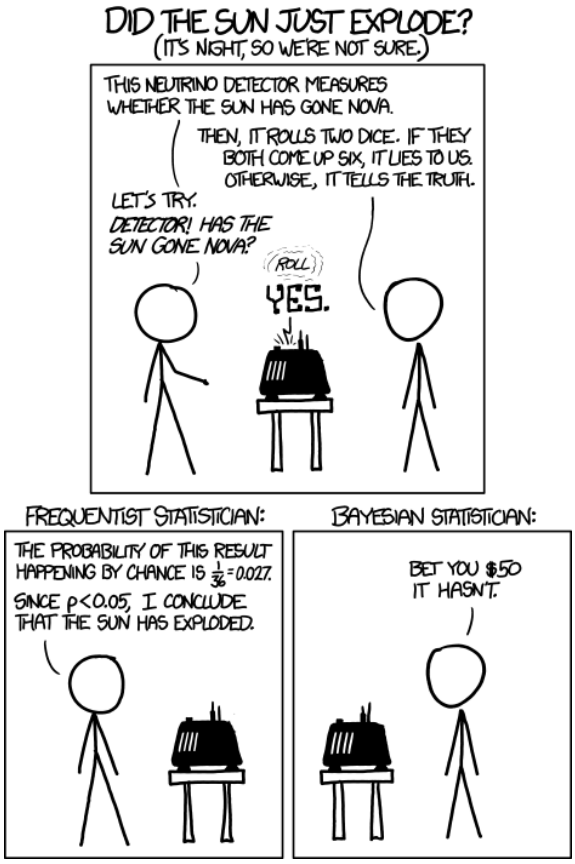
Sensory processing and probabalistic codes.

In this section we will begin to explore the idea that the brain performs Bayesian inference on data. However, we will begin by looking at a cartoon `https://xkcd.com/1132/`



So; will the Bayesian win his or her bet? I will start by annoyingly over explaining the joke; the frequentist knows the detector only lies one time in 36, so he or she assumes that the detector isn't lying. However, the Bayesian takes into account the unlikeness of the detector's statement when estimating the probability its statement is true. In other words, the frequentist is only paying attention to the evidence, the Bayesian also considers how likely the situation the evidence points to is. Randall Munroe's point is that the sort of mistake the frequentist is making is actually made very often, but by illustrating it in this extreme example, the sun is very unlikely to explode, he shows how ridiculous a mistake it is.

Recall Bayes's rule; if X and Y are two random variables then

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1}$$

In this case lets use D to denote the random variable, with $D = w$ denoting a warning from the detector; let S be the state of the sun, with $S = n$ for the sun exploding and $S = f$ for the sun staying fine, that is, not exploding. Now the cartoon tells us that

$$P(D = w|S = n) = \frac{35}{36} \tag{2}$$

and since this is close to one, the silly frequentist assumes the sun has exploded. Let's assume the probability the sun has exploded is one in a million; it is actually much less than that, so $P(S = n) = 10^{-6}$. Now, what we want to know is $P(S = f|D = w)$, the probability the sun hasn't exploded given that detector has warned that it has, now

$$P(S = f|D = w) = \frac{P(D = w|S = f)P(S = f)}{P(W)} \tag{3}$$

The numerator we can calculated easily:

$$P(D = w|S = f)P(S = f) = \frac{1}{36}(1 - 10^{-6}) \tag{4}$$

because the detector will only say the sun is exploding when it isn't when two dice show sixes, that is, one time in 36. The denominator requires marginalizing:

$$P(D = w) = P(D = w, S = n) + P(D = w, S = f) \tag{5}$$

and, using $P(X, Y) = P(X|Y)P(Y)$ this gives

$$\begin{aligned} P(D = w) &= P(D = w|S = n)P(S = n) + P(D = w|S = f)P(S = f) \\ &= \frac{35}{36}10^{-6} + \frac{1}{36}(1 - 10^{-6}) \end{aligned} \tag{6}$$

Substituting all this in to Bayes's rule gives:

$$P(S = f|D = w) \approx 0.999965 \tag{7}$$

The point of this section is that the brain is more like the Bayesian than the frequentist!

Bayesians versus frequentists

Commenting on the response to his cartoon, Randall Monroe says this[1]

> Hey! I was kinda blindsided by the response to this comic.
>
> I'm in the middle of reading a series of books about forecasting errors (including Nate Silver's book, which I really enjoyed), and again and again kept hitting examples of mistakes caused by blind application of the textbook confidence interval approach.

[1] see http://www.explainxkcd.com/wiki/index.php/1132:_Frequentists_vs._Bayesians

> Someone asked me to explain it in simple terms, but I realized that in the common examples used to illustrate this sort of error, like the cancer screening/drug test false positive ones, the correct result is surprising or unintuitive. So I came up with the sun-explosion example, to illustrate a case where naïve application of that significance test can give a result that's obviously nonsense.
>
> I seem to have stepped on a hornet's nest, though, by adding 'Frequentist' and 'Bayesian' titles to the panels. This came as a surprise to me, in part because I actually added them as an afterthought, along with the final punchline. (I originally had the guy on the right making some other cross-panel comment, but I thought the 'bet' thing was cuter.)
>
> The truth is, I genuinely didn't realize Frequentists and Bayesians were actual camps of people - all of whom are now emailing me. I thought they were loosely-applied labels - perhaps just labels appropriated by the books I had happened to read recently for the standard textbook approach we learned in science class versus an approach which more carefully incorporates the ideas of prior probabilities.
>
> I meant this as a jab at the kind of shoddy misapplications of statistics I keep running into in things like cancer screening (which is an emotionally wrenching subject full of poorly-applied probability) and political forecasting. I wasn't intending to characterize the merits of the two sides of what turns out to be a much more involved and ongoing academic debate than I realized.
>
> A sincere thank you for the gentle corrections; I've taken them to heart, and you can be confident I will avoid such mischaracterizations in the future!
>
> At least, 95.45% confident.

Since the rules of probability are fixed it seems odd that there are different 'camps'; but there are certainly two different approaches to statistics described by frequentist and Bayesian. The difference comes down to different interpretations of what we are talking about when we talk about randomness. For a frequentist a random event is truly random, like putting a Geiger counter beside a radioactive source and an observation is a deterministic process which definitively checks whether how many decays a Geiger counter has counted.

For a Bayesian randomness is a representation of our ignorance. An example might be a weather prediction; when I am intending to cycle to Bath I check the weather on my phone and it might say that there is a 27% chance of rain at 5pm. This doesn't mean that there is a random process that 27% of the time will cause rain and the other 73% cause not rain. It means that according to the measurements made by the weather service the state of the weather is such that, based on previous experience, it will rain 27% of the time. This

situation can change, it might start raining now and this new evidence would make it more likely it will rain at 5pm; the situation would also change if the weather service decided helping me avoid rain was a priority so more weather detectors and satellite imaging resources should be directed at predicting rain between Bristol and Bath; this would change the predictions without changing the actual chance of rain.

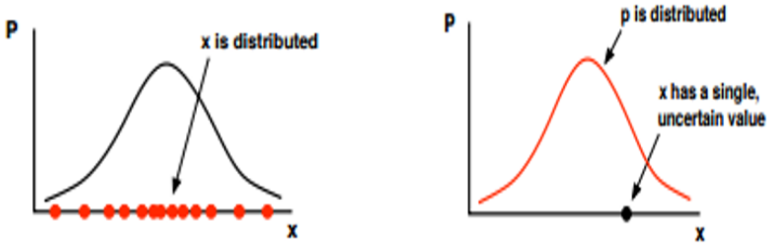


Figure 1: Frequentist versus Bayesian. The left picture illustrates a frequentist interpretation, the right, a Bayesian one. [Pictures from Rosalyn Moran]

In an extreme description, in the frequentist picture a random variable will have many values, they will be different because the process that produces the outcome is random. In the Bayesian picture, there is just one thing being observed; the state of some variable, the probability distribution models our knowledge of its value. This situation is illustrated in Fig. 1. The two pictures, of course, intersect, for example, in the Bayesian picture our partial knowledge of variable is sometimes attributed to noise in the sensors we use to measure it, this noise is a random event in the frequentist sense.

In this way there isn't an argument between frequentists and Bayesians about the laws of probability; rather, frequentist and Bayesian statistics describe different approaches to modelling different types of problems. The problem the brain faces is a Bayesian one, the brain gathers evidence about the world through the senses, these are unreliable and limited but the brain has to infer the state of the world from the information they provided.

In a Bayesian interpretation Bayes's rule

$$P(W|E) = \frac{P(E|W)P(W)}{P(E)} \tag{8}$$

allows us to update our view of the world W based on evidence E. The rule is described as an update rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{9}$$

The *prior*, is our belief about the world before our observation, the likelihood is the probability of the evidence given the state of the world $P(E|W)$, the

evidence is the probability of the evidence $P(W)$ and the posterior, $P(W|E)$, is our new belief about the world given the evidence we have observed.

An example of human perception

As an example of how this might apply to perception, we consider the experiment done by Ernst and Banks [**?**] in which people were asked to judge the height of a block. The height of the block is x and the subjects were asked to judge its height under three conditions, visually v, by touch, that is haptically, h and by vision and touch together hv. In the visual trials noise is added using goggles. The set up is pictured in Fig. 2

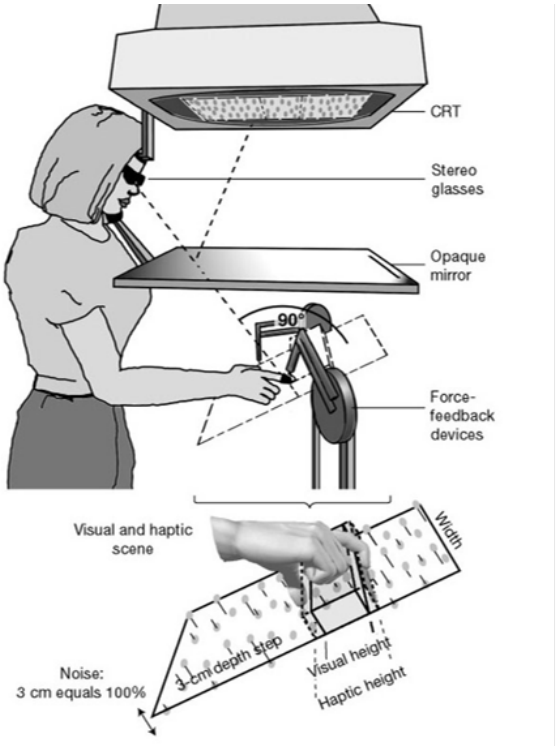


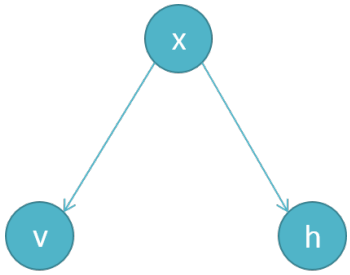
Figure 2: Illustration of the Ernst and Banks experiment. [Pictures from Rosalyn Moran who took it from the paper]

The joint probability is $p(v, h, x)$, it is assumed that V and H are conditionally independent:

$$p(v, h|x) = p(v|x)p(h|x) \tag{10}$$

In mathematics this would be modelled with a Markov chain $V \to X \to H$, in the world of Bayesian inference this is illustrated with a *directed acylcic graph*

or *Bayesian network*:[2]



Either way, the idea is that although there is a relationship between V and H, if the block is big they will both be large, but that this relationship is only because they are both related to X. One consequence of this is

$$p(v, h, x) = p(v|x)p(h|x)p(x) \tag{11}$$

We are interested in the posterior judgement of the height of the block:

$$p(x|v, h) = \frac{p(v, h|x)p(x)}{p(v, h)} = \frac{p(v|x)p(h|x)p(x)}{p(v, h)} \tag{12}$$

It is assumed that the prior is uniform, so $p(x)$ is constant over some range of possible x, so if x is in that range

$$p(x|v, h) \propto \frac{p(v|x)p(h|x)}{p(v, h)} \tag{13}$$

The value of x that maximizes this is called the *maximum a posteriori*.

Now the purpose of the Ernst and Banks experiment is to study sensory fusion, the fusing of two noisy pieces of evidence. Bayesian analyses is used to show the optimal fusing of the data, this is what we will derive now. This is then compared to data to show that the brain approaches the problem in an optimal way.

It is assumed that the haptic and visual channels have independent Gaussian noise around the true value, that is $\mathcal{N}(x, \sigma_v^2)$ and $\mathcal{N}(x, \sigma_h^2)$ respectively, so

$$p(v|x) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{(v-x)^2}{2\sigma_v^2}} \tag{14}$$

and

$$p(h|x) = \frac{1}{\sqrt{2\pi\sigma_v^2}} e^{-\frac{(h-x)^2}{2\sigma_h^2}} \tag{15}$$

and we ignore the normalizing factor $p(v, h)$ since it is independent of x, then

$$p(x|v, h) \propto p(v|x)p(h|x) \tag{16}$$

[2]Picture from Rosalyn Moran

Now if we multiply out the two Gaussians the exponent is

$$-\frac{(h-x)^2}{2\sigma_h^2}-\frac{(v-x)^2}{2\sigma_v^2}=-\left(\frac{1}{2\sigma_h^2}+\frac{1}{2\sigma_v^2}\right)x^2+\left(\frac{h}{\sigma_h^2}+\frac{v}{\sigma_v^2}\right)x+\text{other stuff} \quad (17)$$

so if we define σ by

$$\frac{1}{\sigma^2}=\frac{1}{\sigma_v^2}+\frac{1}{\sigma_h^2} \quad (18)$$

and $\bar{x}$ by

$$\bar{x}=\frac{\sigma^2}{\sigma_v^2}v+\frac{\sigma^2}{\sigma_h^2}h \quad (19)$$

this gives

$$-\frac{(h-x)^2}{2\sigma_h^2}-\frac{(v-x)^2}{2\sigma_v^2}=-\frac{(x-\bar{x})^2}{2\sigma^2}+\text{other stuff} \quad (20)$$

In other words, the optimal guess for x is the variance weighted measurements of v and h; the factors σ^2/σ_v^2 and σ^2/σ_h^2 quantify how much the variability of v and h contribute to the variability of $\bar{x}$, and the more they contribute, the lower the weighting is given to the measurement. This is summarized by writing $\lambda_v=\sigma^2/\sigma_v^2$ and $\lambda_h=\sigma^2/\sigma_h^2$, so

$$\lambda_v+\lambda_h=\frac{\sigma^2}{\sigma_v^2}+\frac{\sigma^2}{\sigma_h^2}=\frac{1}{1/\sigma_v^2+1/\sigma_h^2}\left(\frac{1}{\sigma_v^2}+\frac{1}{\sigma_h^2}\right)=1 \quad (21)$$

so λ_v and λ_h are the visual and haptic weights.

In the actual experiment the visual noise is changed, allowing the experimenters to manipulate the two weights. They then do multiple experiments where they ask the subjects to decide which of two blocks is larger and manage through a regression analysis to decide how the subjects are weighting the two pieces of evidence. In other words it is assumed that the subjects are combining the haptic and visual estimates

$$\xi=\mu_v v+\mu_h h \quad (22)$$

where ξ is the subjects estimate of x and μ_v and μ_h are some weighting factors to be estimated from the behavioural data by regression; for details see the paper. They then compare the measured weightings, μ_v and μ_h, to those estimated by the optimal Bayesian model, that is λ_v and λ_h. The result is shown in Fig. 3 and demonstrate Bayesian optimal perception in this aspect of human sensory fusion.

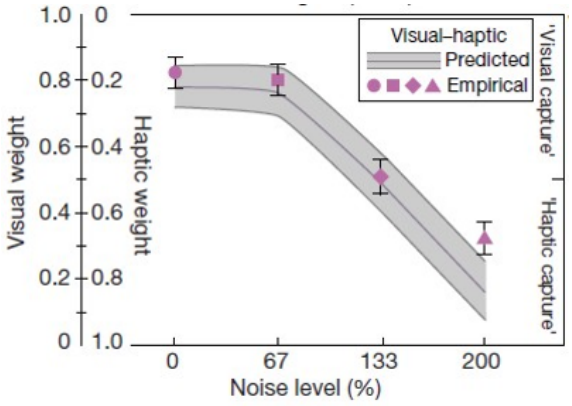


Figure 3: Results of the Ernst and Banks experiment. As the weights are changed by manipulating the visual noise by adjusting the goggles, the weighting of how the visual and haptic information are fused should change according to the grey line; by a clever set of tests it is possible to estimate the true weightings used by the subjects, these are the red features. The match is very good.[Pictures from Rosalyn Moran who took it from the paper]