Review: Statistics, Estimation and Decisions

- December 2008.
- Author: Nathaniel Daw[1,2] & Eero Simoncelli[1,3]
[1]Center for Neural Science, [2]Department of Psychology and
[3]Courant Institute of Mathematical Sciences,
New York University
- Send corrections or comments to nathaniel.daw@nyu.edu

1 Probability Basics

Suppose we perform an experiment, measuring the intensity of a constant light source using a photometer. Each time we make this measurement, we get a slightly different answer. Although the answers will presumably cluster around the "correct" value, there is no way for us avoid the variability in the measurements. The field of **probability** provides an abstract language for describing the uncertainty of these measurements. The field of **statistics** tells us how to take a finite set of measurements and infer something about the world.

1.1 Distributions and their manipulation

The primary entities of probability theory are **random variables** and their associated **probability distributions**. In this example, each light measurement is an instantiation of the random variable x; the probability that on each measurement it will take different values is given by a probability distribution (or density) function (PDF). This is a function mapping possible values of x to their probabilities. The values for x might be discrete (like a coin flip) or continuous (like a reaction time). A fundamental property of probability distributions is that they sum to one over all possible values for x, ie in the discrete case

$$\sum_x \mathbb{P}(x) = 1$$

with $\mathbb{P}(x) > 0$. Here, $\mathbb{P}(x)$ can be thought of like a histogram of long-run frequencies for different measurements x. When x is continuous (like in the limit as the histogram bin size becomes infinitessimally small) the analogous expression is:

$$\int \mathbb{P}(x)dx = 1$$

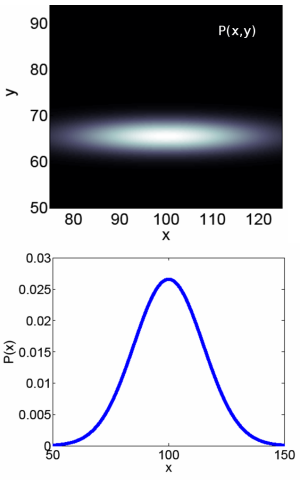

Figure 1: A 2-D joint distribution and one of its 1-D marginals.

and, since the chance that x will take on any particular continuous value is typically infinitessimal, $\mathbb{P}(x)$ actually measures probability **density** rather than probability. Probability is defined by integrating the density over a range of continuous values: $prob(a < x < b) = \int_a^b \mathbb{P}(x)dx$. A potential point of confusion is that probabilities are always less than one (and can be interpreted

like a fractional frequency of occurrence), but densities need not be. For instance, if $\mathbb{P}(x)$ is uniform in the continuous range 0-0.5, then the *density* $\mathbb{P}(x) = 2$ for $0 \leq x \leq 0.5$ (so that the function integrates to 1 over that range) but the *probability* that x takes on a value between 0 and 0.25 is 1/2.

Because they sum to one, PDFs are analogous to functions built up out of chunks of modeling clay. Various transformations of PDFs can shift the clay from bin to bin, but the total amount of clay always stays the same.

Most interesting uses of probability concern reasoning about the relationship between multiple random variables. Suppose that the actual light intensity being measured is y. The relationship between x and y can be summarized by the **joint distribution** $\mathbb{P}(x, y)$, which maps a pair of values x, y to the probability (density) that both will jointly occur (FIgure 1). It is thus like a two-dimensional table of numbers, or a mound of clay piled on a flat surface. We can recover other important distributions by operating on $\mathbb{P}(x, y)$. For instance, if we care only about x we can recover the one-variable **marginal distribution** $\mathbb{P}(x)$ by summing up the probability of each x occurring with different values of y

$$\mathbb{P}(x) = \int \mathbb{P}(x, y) dy$$

like shoveling the modeling clay onto the x axis (FIgure 1). This is just the familiar sum rule from probability theory (i.e., the probability that either of two mutually exclusive events will occur is the sum of their two probabilities), but applied over the whole distribution.

From the joint distribution, we can also recover the **conditional** distribution $\mathbb{P}(x|y)$, the probability distribution over x (the measurement) given that y (the true value) takes on any particular value (Figure 2). This is a function of both x and y but is normally viewed as a one-dimensional distribution (over x) for some fixed y:

$$\mathbb{P}(x|y) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(y)}$$

Since $\mathbb{P}(y) = \int \mathbb{P}(x, y) dx$, the division renormalizes the 2D function so that each row sums to one. Thus, by selecting the row of the function corresponding to some particular y we recover a normalized 1D distribution over x, given that particular value for y.

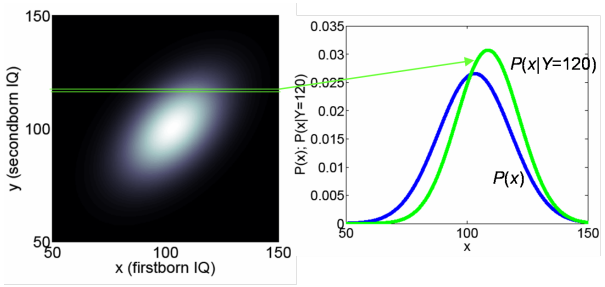

Figure 2: A joint distribution and a conditional distribution. Because the two variables (here, the IQ of a random Swedish male and that of his younger brother) are nonindependent, the joint distribution is tilted and the conditional (green) and marginal (blue) are different.

The definition of the conditional probability is a rearranged version of the familiar product rule of probability (i.e., the probability of two events occurring together is the product of their probabilities, with one conditional on the other)

$$\mathbb{P}(x, y) = \mathbb{P}(y)\mathbb{P}(x|y)$$

though again we apply it to transform a whole distribution. Two random variables are statistically **independent** (like separate rolls of a die) if the above expression simplifies to $\mathbb{P}(x, y) =$

$\mathbb{P}(y)\mathbb{P}(x)$; that is, the two-variable joint distribution separates into the outer product of two one-variable marginal distributions. It follows that for independent variables, the conditional and marginal distributions are equal: $\mathbb{P}(x) = \mathbb{P}(x|y)$. That is, knowing that y takes a particular value doesn't change the distribution over x. Conversely, if x and y **covary** (are not independent), then $\mathbb{P}(x) \neq \mathbb{P}(x|y)$ and conditioning on a value for one variable changes the distribution over the other (Figure 2).

From the definition of the conditional probabilities $\mathbb{P}(x|y)$ and $\mathbb{P}(y|x)$ we can obtain the important Bayes' rule, which describes the relationship between the two:

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

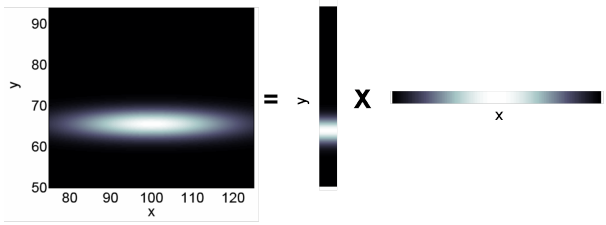

Figure 3: The joint distribution of two indpendent variables factors into the product of two 1-D maginals.

Among other things, this rule exposes the close relationship between the **forward** or **measurement model** implicit in the distribution $\mathbb{P}(x|y)$— which describes how our light meter works to produce a noisy measurement x from some true value y — to an **inference** or **inverse model** $\mathbb{P}(y|x)$, which gives a distribution over y conditional on some measured value x. In interpreting this equation, we normally view the terms as functions of y for a fixed measurement x. The conditional $\mathbb{P}(x|y)$ is also called the **likelihood function**, since (viewed, confusingly, as a function of x) it measures the likelihood of x given different ys. The marginal $\mathbb{P}(y)$ is known as the **prior** probability of y (ie the distribution over y absent any information about x, like the chance that different light intensities will be encountered in the world generally). For any particular x, $\mathbb{P}(x) = \int \mathbb{P}(x|y)\mathbb{P}(y)dy$ is just a normalizing constant. Bayes Rule therefore says that the probability of y given some measurement x is larger for ys that have a high likelihood of producing x, and also for ys that were more likely to begin with.

1.2 Sums of independent random variables

If x and y are independent random variables distributed as $\mathbb{P}(x)$ and $\mathbb{P}(y)$, respectively, and we define a new random variable z as the sum of one sample from each: $z = x + y$, then it is easy to show that the distribution $\mathbb{P}(z)$ is the convolution of $\mathbb{P}(x)$ and $\mathbb{P}(y)$. This is because to obtain the probability of any particular $z = Z$, we must sum over the joint probability $\mathbb{P}(x)\mathbb{P}(y)$ for all pairs x, y that sum to Z: $\mathbb{P}(z = Z) = \int dx \mathbb{P}(x = X)\mathbb{P}(y = Z - X)$, which is just the convolution.

1.3 Transformations of distributions

Given some monotonic function $f(x)$ of a random variable x, e.g. $y = x^3$, we may ask what is the distribution of y, $\mathbb{P}(y)$, in terms of $\mathbb{P}(x)$. The answer to this question arises from the fact that the distributions must both be normalized, therefore, the transformation from $\mathbb{P}(x)$ to $\mathbb{P}(y)$ must preserve density. Locally, $|\mathbb{P}(y)dy| = |\mathbb{P}(x)dx|$; rearranging terms and substituting $y = f(x)$ we have $\mathbb{P}(y) = \mathbb{P}(x)/|(df(x)/dx)|$. To evaluate this expression as a density over y,

for each y we must evaluate the right-hand side at the point $x = f^{-1}(y)$. This produces the rather more confusing expression $\mathbb{P}(y) = \mathbb{P}(f^{-1}(y))/|(df(f^{-1}(y))/dx)|$. That is, the density $\mathbb{P}(y)$ is the original density $\mathbb{P}(x)$, rescaled inversely by the local "slope" of $f(x)$ at that point, $df(x)/dx$, which measures how "stretched out" is the density at x when transformed to $f(x)$ (Figure 4). This is all evaluated at the point that maps to y,$x = f^{-1}(y)$.

One particularly useful $f(x)$ for such transformations is the **cumulative density function** (CDF), which is defined as $f(x) = \int_{-\infty}^{x} \mathbb{P}(x)dx$: that is, the total probability mass at values x or smaller. Since the slope of the CDF, $df(x)/dx$, is just $\mathbb{P}(x)$ itself, transforming x by its CDF produces a uniform distribution over $[0, 1]$. This can be useful for gain control of experimental measurements, and also (by reversing the transformation) can be used to generate random variables with an arbitrary distribution, by generating a uniform $[0, 1]$ variable then transforming it according to the inverse CDF of the desired density.

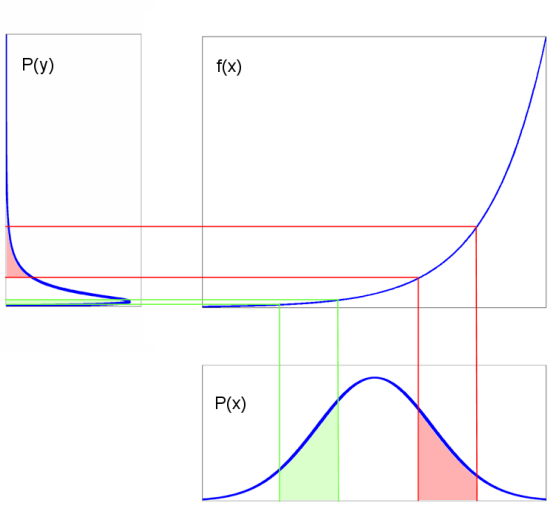

Figure 4: Transformation of a distribution: Equal amonts of probability density (green and pink) in the original distribution map to equal amounts of density in the transformed distribution, but are spread out uneventy according to the slope of the transforming function.

1.4 Expectation, moments, and covariance

Given some function $f(x)$of a random variable, as above, it can also be useful to compute the average or **expected value** of $f(x)$, weighted according to the probabilities $P(x)$. This is written $\mathbb{E}[f(x)]$ and defined as $\int \mathbb{P}(x)f(x)dx$. If you think of $f(x)$ as like a vector of values for different x (e.g., for x^2, $[1, 4, 9, ...]$) then, geometrically, taking an expectation is like a dot product between those and a vector of probabilities for the different values of x (e.g., for a fair die, $[1/6, 1/6, 1/6, ...]$). That is, expectation is a projection. From the linearity of this operation follow useful properties for manipulating expectations, e.g., $\mathbb{E}[f(x) + g(x)] = \mathbb{E}[f(x)] + \mathbb{E}[g(x)]$ and $\mathbb{E}[cf(x)] = c\mathbb{E}[f(x)]$.

Some important examples are the **mean** $\mathbb{E}[x]$ (that is, the average value of x itself) and the **variance** $\mathbb{E}[(x - \mathbb{E}[x])^2]$, which measures the expected average spread of the measurements, quantified as the squared distance around their mean. For a joint distribution $\mathbb{P}(\overrightarrow{x})$ over N multiple random variables, which we write with vector notation $\overrightarrow{x} = [x_1, x_2, ...]$, the mean $\mathbb{E}[\overrightarrow{x}]$ averages vectors according to their probability, which amounts to computing the mean for each element separately. The analogue of the variance is the **covariance** matrix $\mathbb{E}[(\overrightarrow{x} - \mathbb{E}[\overrightarrow{x}])(\overrightarrow{x} - \mathbb{E}[\overrightarrow{x}])^T]$, which (since this is an outer product) is an NxN matrix.

In two dimensions, and assuming $\mathbb{E}[\overrightarrow{x}] = \overrightarrow{0}$ to simplify the notation (for the full expression, substitute $(x_1 - \mathbb{E}[x_1])$ for x_1 throughout, and similarly for x_2), the covariance matrix has the

form:

$$\begin{bmatrix} \mathbb{E}[x_1^2] & \mathbb{E}[x_1 x_2] \\ \mathbb{E}[x_1 x_2] & \mathbb{E}[x_2^2] \end{bmatrix}$$

where the diagonal terms measure the spread along x_1 and x_2 marginally, and the offdiagonal term captures the shape or tilt of the distribution. If x_1 and x_2 are independent, then the offdiagonal terms will be zero, because the factorization $\mathbb{P}(x,y) = \mathbb{P}(x)\mathbb{P}(y)$ allows separating the expectation of the product into the product of expectations $\mathbb{E}[x_1]\mathbb{E}[x_2]$. Conversely, if x_1 tends to be high (relative to its mean) when x_2 is high, then they tend to be both positive or both negative at the same time so the average value of their product is positive, and the distribution will tilt right.

Finally, in multiple dimensions we might define a sort of generalized variance as the spread along any direction specified by some unit vector $\overrightarrow{u}$. In particular, if we take $\mathbb{E}[\overrightarrow{x}] = \overrightarrow{0}$ (or redefine $\overrightarrow{x}$ as $\overrightarrow{x} - expect[\overrightarrow{x}]$), then project measurements $\overrightarrow{x}$ onto $\overrightarrow{u}$, ie $\overrightarrow{u}^T\overrightarrow{x}$, this will result in a 1D distribution with variance $\mathbb{E}[(\overrightarrow{u}^T\overrightarrow{x})^2] = \mathbb{E}[\overrightarrow{u}^T\overrightarrow{x}\overrightarrow{x}^T\overrightarrow{u}] = \overrightarrow{u}^T\mathbb{E}[\overrightarrow{x}\overrightarrow{x}^T]\overrightarrow{u}$ which, finally, equals $\overrightarrow{u}^T C\overrightarrow{u}$. That is, the covariance matrix contains all the information necessary to compute the spread (extent, generalized variance) of the distribution along any direction.

1.5 The amazing Gaussian

A particularly important form of distribution is the Gaussian. In one dimension, it is the familiar bell curve, given by $\mathbb{P}(x) \propto \exp[-(x-\mu)^2/2\sigma^2]$, for mean μ and variance σ^2. (The square root of the variance, σ, is the standard deviation.) Here, and throughout, we omit the normalizing constant $1/(\sqrt{2\pi}\sigma)$ implied by the constraint that $P(x)$ sums to one.)

As a joint distribution over $n$multiple random variables $\overrightarrow{x} = [x_1, x_2, ...]$, the expression for the Gaussian is

$$\mathbb{P}(\overrightarrow{x}) \propto \exp[-(\overrightarrow{x} - \overrightarrow{\mu})^T C^{-1}(\overrightarrow{x} - \overrightarrow{\mu})/2]$$

with mean $\overrightarrow{\mu}$ and covariance C.

Gaussians have a number of properties that make them exceptionally useful, among them:

1. The sum of two Gaussian random variables is itself Gaussian. In particular, if x and y are Gaussians with means μ_x and μ_y and variances σ_x^2 and σ_y^2 then their sum is Gaussian with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. These results can be obtained from the convolution theorem, together with the fact that the Fourier transform of a Gaussian is also a Gaussian. From these results it also follows that the average of N draws from x (ie their sum, divided by N) has mean μ_x and variance decreasing with N, i.e. σ^2/N.

2. In fact, the sum or average of random variables with any distribution, so long as they are independent and identically distributed is also Gaussian, in the limit as you sum more and more of them. (This is called the "Central Limit Theorem.")

3. If $\overrightarrow{x}$ is distributed as a multidimensional Gaussian, then its conditional and marginal distributions are also Gaussians. That is, if one slices through a Gaussian (by conditioning on a particular value for some of its variables, and renormalizing what remains) or sums over some of the variables in the Gaussian to compute a marginal distribution over the remaining variables, then in both cases the resulting distributions are still Gaussians.

2 Statistical Estimation

The previous section was about abstract mathematical descriptions of probability. Now we imagine ourselves in a more practical context, in which we've made some observations (i.e., experimental measurements), from which we want to estimate some quantity. In general, each time we make a measurement, it comes out differently. This unpredictability might arise either from aspects of the environment that are beyond our control (eg., stray electromagnetic radiation), or from unobservable fluctuations within the system itself (eg., flow of individual ions through channels in a membrane), or from uncertainties introduced by the measurement process. We refer to the quantity we're trying to measure as the "signal", and all these sources of uncertainty as "noise". We describe the noise using random variables.

Consider an example in which we wish to measure the brightness of a constant light source. Our measurement is corrupted by the quantal nature of light, by disturbances in the medium (air) through which the light must propagate, and by inacuraccies in our measurement device. To make the problem simpler, it is often assumed that such measurement uncertainties are combined additively along with the "true" value to yield the measurement:

$$\mathbb{P}(m|b) = b + n \tag{1}$$

Here n is a random variable that represents the combination of three sources of uncertainty mentioned above. The right side of equation (1) is known as the "likelihood" function: it tells us the likelihood of our measurements given a particular value of b. Now, the problem of estimation is to *invert* this equation: We want estimate of b, given a finite set of measurements $\{m_k\}$. We'll notate our estimate as $\hat{b}(\{m_k\})$, with the "hat" indicating that this is not the true b, and the parentheses indicating that it is a function of the data. More compactly, we can also bundle the m_k's into a vectxor $\vec{m}$.

Before discussing specific estimators, it should be intuitively obvious that we'll want to minimize the error in our estimates – that is, the difference between the estimate and the true value. We decompose this error into two distinct pieces:

Bias : This is the average error in the estimator:

$$B(b) = \mathbb{E}_n\left[\hat{b}(\vec{m}) - b\right]$$

An estimator that is (on average) equal to the true value is called **unbiased**.

Variance : This is simply the variance of the error:

$$V(b) = \mathbb{E}_n\left[(\hat{b}(\vec{m}) - b - B(b))^2\right]$$

Notice that the mean squared error is just the sum of the squared bias and the variance.

Now how do we decide on an estimator? As with most such questions, the answer is "it depends on the problem". But it is worth knowing about three particular estimators that are most commonly used, and which are built upon each other. First, suppose that all we know about

our problem is the likelihood function of equation (1). In this case, the simplest choice is to choose the value of b that makes the measurements most likely:

$$\hat{b}_{\mathrm{ML}}(\vec{m}) = \arg\max_b \mathbb{P}(\vec{m}|b)$$

This is known as the **Maximum Likelihood** estimator (MLE).

Take the example of light measurement, and assume that n is zero-mean, Gaussian distributed, with variance σ^2. Assume we make N measurements, and that these are statistically independent. Then the likelihood function is a product of the individual likelihoods:

$$\begin{aligned}\mathbb{P}(\vec{m}|b) &= \prod_k \mathbb{P}(m_k|b) \\ &= \prod_k \exp[-(m_k - b)^2/2\sigma^2]\end{aligned}$$

To compute the estimate, we could maximize this expression, but it's simpler to maximize the log:

$$\begin{aligned}\hat{b}(\vec{m}) &= \arg\max_b \log \mathbb{P}(\vec{m}|b) \\ &= -\arg\max_b \sum_k (m_k - b)^2/2\sigma^2\end{aligned}$$

Taking the derivative of the righthand expression with respect to b and setting equal to zero gives:

$$\sum_k 2(m_k - \hat{b}(\vec{m})) = 0$$

or

$$\hat{b}(\vec{m}) = \frac{1}{N}\sum_k m_k$$

After all that, the answer is quite simple: take the average of the measurements! [Now, verify that this is unbiased, and compute the variance.]

Now we consider a more sophisticated estimator. Suppose we had some knowledge of the values that b could assume. For example, we might know that it must lie within a particular range. Or perhaps some values, while possible, are extremely unliky to occur in the real world. This kind of knowledge may be represented with a probability distribution on b, known as the **prior** distribution: $\mathbb{P}(b)$.

Given this knowledge, we can use Bayes' rule to turn the likelihood into the inverse conditional probability, and we can then maximize that:

$$\begin{aligned}\hat{b}_{\mathrm{MAP}}(\vec{m}) &= \arg\max_b \mathbb{P}(b|\vec{m}) \\ &= \arg\max_b \mathbb{P}(\vec{m}|b)\mathbb{P}(b)/\mathbb{P}(\vec{m}) \\ &= \arg\max_b \mathbb{P}(\vec{m}|b)\mathbb{P}(b)\end{aligned}$$

In the last step, we dropped the denominator from the expression because it does not depend on b, and thus has no influence on the maximum. This estimator is known as the **maximum aposteriori** (MAP) estimator.

Finally, we might want to augment the problem by including some sort of cost function (also called a "loss" function) that describes how much penalty is incurred by making each particular error. In general, this is a function of both the true value and the estimated value: $L(b, \hat{b})$. A **Bayesian** estimator attempts to minimize the average (expected) loss:

$$\begin{aligned} \hat{b}_{\text{Bayes}}(\vec{m}) &= \arg\min_{\hat{b}} \mathbb{E}_b[L(b,\hat{b})|\vec{m}] \\ &= \arg\min_{\hat{b}} \int_b L(b,\hat{b})\mathbb{P}(b|\vec{m}) \end{aligned}$$

[Ex: Gaussian linear case. Note that as the number of measurements increases, the estimate gets closer and closer to the ML estimate.].

A special case of this estimator is the **Bayes Least Squares** (BLS) estimator, in which the loss function is just squared error:

$$\begin{aligned} \hat{b}_{\text{BLS}}(\vec{m}) &= \arg\min_{\hat{b}} \int_b (b-\hat{b})^2\mathbb{P}(b|\vec{m}) \\ &= \int_b b\mathbb{P}(b|\vec{m})q = \mathbb{E}_b[b|\vec{m}] \end{aligned}$$

where the second line is achieved by differentiating the first line, setting the result equal to zero, and solving for $\hat{b}$. That is, the BLS estimator is simply the conditional mean of the parameter given the data!

Note that in the Gaussian linear case studied above, the BLS is identical to the MAP estimator, since the peak of a Gaussian distribution is the same as the mean. But for non-Gaussian posterior densities, the BLS and MAP are often different.

3 Statistical Decision Theory

The previous section described the problem of estimating the value of some unknown quantity. This is an instance of a more general problem of making a decision based on a set of uncertain measurements. In general, we have some unknown quantity like b relevant to the decision, and a set of measurements $\vec{m}$ that bear on b, producing a posterior distribution $\mathbb{P}(b|\vec{m})$. Finally, we have some set or range of possible decisions d, and a loss function $L(b, d)$ measuring the cost of each decision for each possible true value of b.

Bayesian statistical decision theory simply asserts that we should choose the d that minimizes the expected loss: $\arg\min_d \mathbb{E}_b[L(b,d)|\vec{m}]$, where the expectation is according to the posterior distribution $\mathbb{P}(b|\vec{m})$. The estimation problem can be viewed as a particular subcase of this problem, in which the decisions d are candidate estimates $\hat{b}$, and the decision is which estimate to produce. But more generally, the same statistical framwork applies to many other sorts of decisions such as whether to buy health insurance or where to aim a missile. In estimation, the MAP estimator follows from this strategy if all errors are equally costly ($L(b, d) = 0$ for $b = d$; 1 ow) and the ML estimator arises when, additionally, the prior is flat, so that the likelihood function is proportional to the conditional distribution: $\mathbb{P}(\vec{m}|b) \propto \mathbb{P}(b|\vec{m})$.

3.1 Signal detection: the ideal observer

In perceptual psychology, a restricted estimation/decision problem, known as **Signal Detection Theory** has been used to describe the process by which observers detect stimuli in experiments. This is a particularly simple situation in which formally to examine the tradeoffs involved in decision-making, and we consider the ideal decision theoretic solution before considering how the framework can be used to characterize the behavior of experimental subjects.

Here, we assume that the unknown variable b can take only two possible values, which we will write S and N (for "signal present" and "not present"). We further, usually, assume that the measurement distributions, conditional on a signal being present or absent, $\mathbb{P}(\vec{m}|S)$ and $\mathbb{P}(\vec{m}|N)$, are Gaussian random variables with means μ_S and μ_N, respectively, and variances σ_S and σ_N. Finally, we often further assume that $\sigma_S = \sigma_N$ and $\mu_N = 0$; the idea being that the observation is the same (Gaussian noise) either in the presence or absence of a signal of constant magnitude. The observer's goal is to observe a measurement, and decide whether the signal was present ("S"), or not ("N").

:

Note that since the measurement distributions overlap, it will not be possible to respond with 100% accuracy; instead, an observer can only shift her errors between different categories. There are four possible outcomes of a trial (FIgure 5: in the presence of a signal, the subject can either respond "S" (a **hit**), or "N" (a **miss**), and in the presence of no signal, she may again respond "S" (a **false alarm**, FA) or "N" (a **correct rejection**, CR). A loss function therefore assigns a cost to each of these events (and it suffices to assign nonzero cost only to the two types of errors, miss and FA).

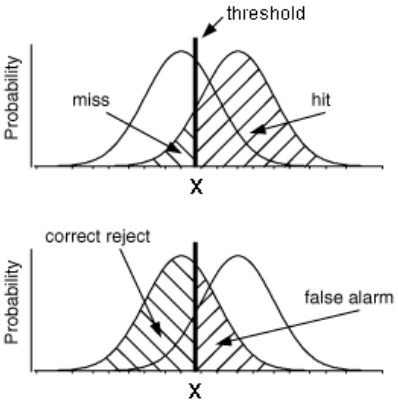

Figure 5: Signal detection (modified from David Heeger): Distributions of measurements and possible outcomes.

Since there are only two possible values for b, for any particular measurement, it is convenient to compare them using the **likelihood ratio**

$$LR = \frac{\mathbb{P}(m|S)}{\mathbb{P}(m|N)}$$

. A set of decision rules of particular importance are those obtained by thresholding this quantity: ie, for some threshold θ, respond "S" if $LR > \theta$ and "N" otherwise. Larger θs are more conservative: they increase correct rejections at the expense of also increasing misses. Clearly, the maximum likelihood estimate of b takes this form, for $\theta_{ML} = 1$, so that the subject responds according to whichever likelihood is larger. Additionally, it is easy to show that the MAP estimate (choose whichever response has higher posterior probability) corresponds to

$$\theta_{MAP} = \frac{\mathbb{P}(N)}{\mathbb{P}(S)}$$

. That is, the threshold is corrected by the ratio of prior probabilities of the two events, and becomes more conservative if null events are more common. Finally, the full Bayesian estimator

is

$$\theta_B = \frac{\mathbb{P}(N)}{\mathbb{P}(S)} \cdot \frac{L(FA) - L(CR)}{L(MISS) - L(HIT)}$$

which minimizes expected loss and is more conservative if false alarms are costly relative to misses.

Note that in the traditional equal-variance case, when $\sigma_S = \sigma_N$, the likelihood ratio is a monotonic function of the measurement m, so these rules all simply correspond to thresholds on the measurement itself. For unequal variances, however, this is no longer the case: the LR is non-monotic in the measurement, so a single threshold on the LR corresponds to multiple thresholds on the measurement.

3.2 The laboratory observer

In short, statistical decision theory asserts that how an observer should ideally trade off different sorts of errors in a detection problem depends on how costly they are relative to each other, and uses the tools of statistical inference to estimate the expected cost of different choices. Signal detection theory has also been used as a framework for characterizing the behavior of subjects in psychophysical experiments, in which case their subjective loss functions or what threshold they might adopt are not known. Indeed, researchers in perception are often interested in the perceptual characteristics of, say, the visual system (e.g., how many photons does one need to detect a flash of light?) unconfounded by the various biases or motivations that enter into the decision. Signal detection theory is a framework for understanding how the properties of the sensory system (i.e., in this model, the means and variances of the signals being detected) affect the observable decision behavior.

A useful one-variable summary of the perceptual aspects of a detection problem is:

$$d' = \frac{\mu_S - \mu_N}{\sigma}$$

, that is, the distance between the signal and no-signal means, measured in units of the standard deviation of the noise. Easier detection problems (those with more signal, or less noise) have higher d'.

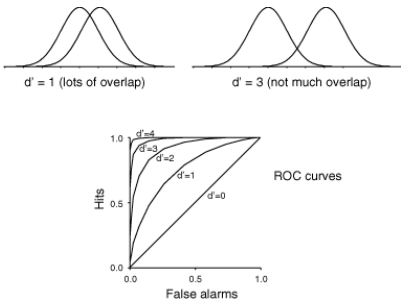

Figure 6: d' and the ROC curve (from David Heeger)

One tool that suggests how to measure d' without making any assumptions about θ is the so-called **Receiver Operating Characteristic** (ROC) curve. An ROC curve (Figure 6) is a plot of the fraction of hits against that of false alarms. Any particular threshold in any particular decision problem corresponds to a point in the ROC graph (a hit rate, and a false alarm rate), as we increase θ, we normally decrease both hits and false alarms. For a particular decision problem — that is, a particular d' — the range of all possible θs sweeps out a curve across the chart, from the upper right hand corner (100% hits and false alarms) to the lower left (100% misses and correct rejections). Easier detection problems define curves closer to the upper-left hand corner (100% hits, no false alarms), and farther from the main diagonal (along which hits and false alarms are equal). Indeed, the area under the ROC curve is a functon of d'. All this suggests two experimental approaches for measuring

d' independent from the decision threshold θ: one is simply to measure the proportion of hits and false alarms across a number of detection trials, and identify the d'on whose ROC curve this performance lies. Another approach (which is less closely bound to the specific form of the noise assumed) is to induce the subject to adopt many different θs, for instance by changing the instructions, measure a point on the ROC curve for each, and use these to approximate the area under the curve.