Computational Vision
U. Minn. Psy 5036
Daniel Kersten
Lecture 4: Ideal Observer Analysis

Goals

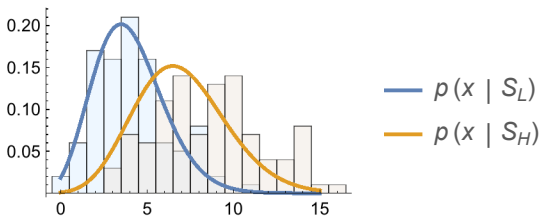
Last time

The accuracy and reliability of human perceptual decisions are limited by two primary sources:

1) inherent uncertainty in the stimulus information for a specific task
2) limitations of the human observer.

Last time we developed a model of **ideal observer.** The ideal observer has a well-defined task (e.g. get the most correct answers in a detection task), and is presumed to have knowledge of the **generative** process that produces the image data. Its ability to make reliable decisions is limited by uncertainty represented by the generative model: 1) the prior p(H); 2) the likelihood p(x|H).

For the dot simulations of photon distributions, the variability is modeled as a Poisson distribution. We can think of a distribution as the limiting case of a histogram, where we have unlimited samples. The spread of the distribution is a measure of the variability in the data (dot count) for a fixed switch setting -- the variability in the generative process. This figure shows the distributions and histograms for high and low settings. These probability distributions of dot counts x, are said to be conditional on the switch setting, and we expressed this as: $p\,(x \mid S_L)$, $p\,(x \mid S_H)$.

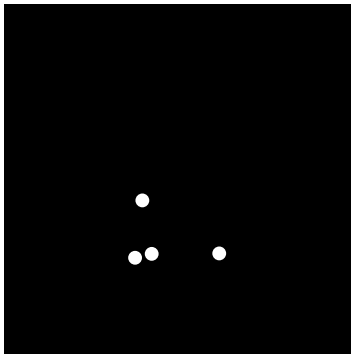


The ideal's job is to take input data and make the best guess regarding some state of the generating process--i.e. make an **inference**. In one demo, the image data was the display of dots, the task is to decide the switch state, high or low, and "best" is defined as minimizing the error rate. This is a called a **yes/no** discrimination task, where the observer is shown only one image in a given “trial”, but it could

have been caused either by the high or low switch setting, and the observer has to decide which. Intuitively, it seems like the best strategy is to count the number of dots, and use this number to decide the state. But how should one determine the criterion that has to be crossed before deciding the switch state is probably high vs. low?

```
dotpattern[mean_] :=
 Table[RandomReal[{0, 1}, 2], {RandomVariate[PoissonDistribution[mean]]}];
display[mean_] := Graphics[{PointSize[0.04], White, Point /@ dotpattern[mean]},
  AspectRatio → 1, Frame → False, FrameTicks → None, Background → GrayLevel[0.],
  PlotRange → {{-0.2, 1.2}, {-0.2, 1.2}}, ImageSize → Small];
```

```
RandomChoice[{display[9], display[4]}]
```

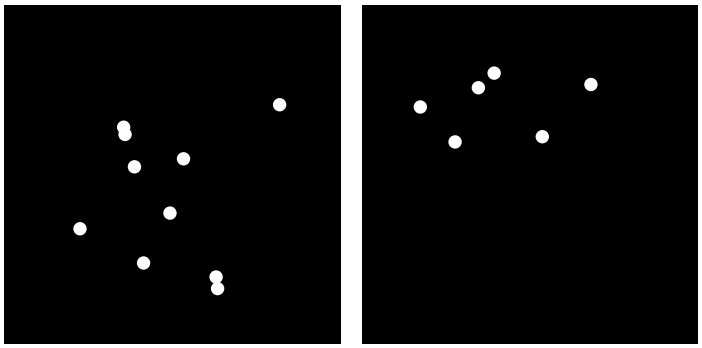


A generalization of the **yes/no** task is *classification*, where given an image input x, the observer has to decide which of N states of the world (S_1,S_2, S_3, ..., S_N) produced image x. For example, S_i could be the name of a person. We will consider the classification task later in the course when we study object recognition.

In the case of a **two-alternative forced-choice** (2AFC) discrimination task (wiki link), an observer is given two images in a given trial--one was caused by the high switch, and one by the low. The only source of information to make the decision is the image data--the dots (i.e. not which pattern was shown first, or whether shown on the left or right). The observer has to decide which of the two was caused by the "high" switch (or "low"). Whether the "signal" is defined as the high or low setting, is a matter of convention, and terminology. We'll treat the high switch setting as the "signal". Psychophysicists like 2AFC because it reduces the problem of varying criterion.

```
dotpattern[mean_] :=
 Table[RandomReal[{0, 1}, 2], {RandomVariate[PoissonDistribution[mean]]}];
display[mean_] := Graphics[{PointSize[0.04], White, Point /@ dotpattern[mean]},
  AspectRatio → 1, Frame → False, FrameTicks → None,
  Background → GrayLevel[0.], PlotRange → {{-0.2, 1.2}, {-0.2, 1.2}}];
```

```
GraphicsRow[RandomChoice[{{display[9], display[4]}, {display[4], display[9]}}]]
```



Appealing to intuition again suggests that the ideal strategy would be to count the dots in each, and decide “signal” for which ever one has the most dots. We’ve seen that this indeed is the ideal strategy for the particular generative model and task of photon discrimination.

Today

In this lecture we complete our introduction to classical signal detection theory (SDT). SDT provides an important set of tools for measuring and modeling the sensitivity of human and neural perceptual decisions. (Later we'll generalize further to "statistical decision theory"--same acronym!) Today will:

- Introduce the (standard) additive signal plus Gaussian noise model, and apply it to variability in light levels.
-Understand how to summarize ideal (and human performance) in the yes/no task in terms of hit and false alarm rates, and to relate these to a sensitivity measure called d’.
- Understand how to quantitatively estimate the “signal-to-noise” ratio “inside the head”, and from there compare human and ideal performance.
- Calculate performance in other tasks. In particular, the two-alternative forced-choice (2AFC) task
- Measure your own absolute statistical efficiency in a 2AFC task

What we learn today will provide the basis for addressing the question: What does the eye see best?

The additive Gaussian noise model for signal detection

Motivation

Most inference modeling is done using Gaussian models of variability. One reason is theoretical convenience. A deeper theoretical reason rests on the Central Limit Theorem, which says that a sum of independently drawn random variables (from a fixed distribution, even non-Gaussian) looks more and more Gaussian the more elements that are in the sum. Empirically, many experiments on human signal detection have been well-fit by assuming Gaussian distributions of the underlying decision variable. However, as we will see later (when we measure statistics on natural images), the Gaussian assumption/approximation for some random variables is a bad approximation. It is always important to test this assumption when possible. We'll first show that a Gaussian approximation provides a good approximation to the Poisson distribution. There are some important subtleties, regarding continuous vs. discrete

distributions, that we learn about below.

Some terminology. We've adopted the convention of treating the high or brighter light as a "signal". Similarly, in the context of detection, we can think of the low switch settings as "noise". We will continue with this here, and use the terms "signal" and "noise". But remember that this is just a convention--the problem is symmetric, and we could be talking about whether a measurement is from hypothesis A vs. hypothesis B ("Not A").

Gaussian approximation for signal and noise Poisson distributions

As the mean *a* gets large, the frequency of occurrence of a Poisson distributed random variable, X, can be approximated by the expression for a Gaussian distribution:

$$p(X = x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}, \text{ where x is a real number between } -\infty \text{ and } +\infty$$

As we saw with the Poisson distribution, we can ask Mathematica to provide the expression:

```
PDF[NormalDistribution[μ, σ], x]
```

$$\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}$$

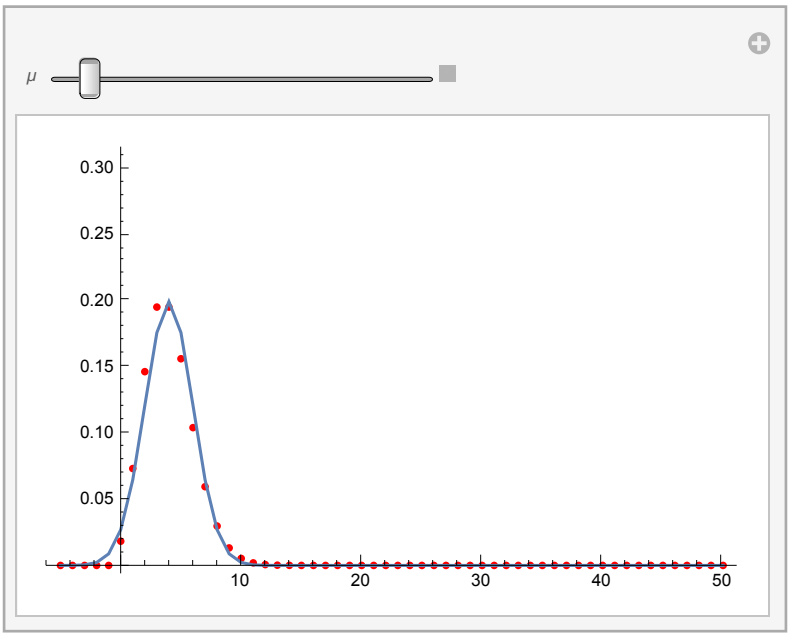
► 1. Use Mean[] and Variance[] to show that:
the mean or expectation of X is : $E(X) = \mu$, and the variance is : $\text{var}(X) = \sigma^2$

For a discrete distribution over random variables $x_1,...,x_N$, with corresponding probabilities $p_1,...,p_N$, $\sum_{i=1}^{N} p_i = p_1+p_2...+p_N = 1$. For a continuous distribution, $\int_{-\infty}^{\infty} p(x)\,dx=1$. To evaluate the area under the gaussian distribution, PDF[NormalDistribution[μ,σ],x], try executing: Integrate[PDF[NormalDistribution[μ,2],x],{x,-Infinity,Infinity}]:

► 2.
```
Integrate[PDF[NormalDistribution[μ, σ], x], {x, -Infinity, Infinity}];
Simplify[%, {Re[σ^2] ≥ 0, σ > 0}]
```

The gaussian approximation is useful to estimate probability values for large μ. If μ is large enough, the probability of negative values (which is meaningless for a Poisson distribution) is very small. For computational convenience and for later generality, we will usually use the Gaussian approximation. Let's compare the forms of the Poisson and Gaussian distributions:

```
Manipulate[
 ndist = NormalDistribution[μ, √μ];
pdist = PoissonDistribution[μ];
 p1 = Table[{x, PDF[pdist, x]}, {x, -5, 50}];
g1 = ListPlot[p1, PlotStyle → RGBColor[1, 0, 0]];
 p2 = Table[{x, PDF[ndist, x]}, {x, -5, 50}];
g2 = ListPlot[p2, Joined → True, PlotRange → {{-5, 50}, {0, .3}}];
 Show[{g1, g2}, PlotRange → {{-5, 50}, {0, .3}}], {{μ, 10, "μ"}, 2, 40, 1}]
```



Compare the Poisson with the Gaussian approximation when the mean is small, like μ=4.

Note: Discrete vs. continuous distributions

The Poisson distribution represents probabilities of a random variable taking on integer values. The distribution is "discrete". In contrast, the Gaussian distribution is continuous. A continuous distribution is represented by a probability *density* meaning that we use it to characterize the probability of a Gaussian random variable falling within a certain range of real numbers.

We can interpret this approximation in two ways. We can discretize the continuous Gaussian function (as above) to give us a set of probabilities (over integers) that closely match those of the corresponding Poisson distribution, and make sure that the discrete sum is one (a fundamental requirement for a probability distribution). This really isn't a gaussian distribution anymore, because we are only defining probability at integer values of the random variable.

Alternatively, as the photon count gets high, we can treat light intensity as a continuous quantity (abandoning our quantized notion of light magnitude). In this latter case, we would treat the random variable X (light intensity) as being a continuous variable with a continuous probability distribution or "density". Then, because there is an infinite number of possible values over any finite range, the probabil-

ity of X=x, for any particular value (x = π, or x = 3.1, for example) is actually zero! We will often use upper case (X) for a random variable that isn't assumed fixed, and lower case (x) for some fixed, e.g. measured, value of X.

To fix this, we treat p(X) as a density (as in mass density in physics), rather than a probability (as in mass). Then we can put a non-zero number on the probability of X taking on a value x in some small range, dx as: p(x<X<x+dx) ~p(x)dx. More on this later.

► 3. To appreciate the differences between a discrete distribution and a continuous density, evaluate PDF[NormalDistribution[36,6],x] and PDF[PoissonDistribution[36],x] for values of x = 30 and 30.01. Explain what you see. Is PDF[NormalDistribution[36,6],30] the probability of X=30?

The generative model for signal discrimination: additive gaussian noise

Let's approximate our photon inspired model with a view towards generalization. We will express the generative model as an "additive gaussian model". This is a standard form used for all kinds of detection tasks, for visual and auditory patterns, as well as non-perceptual decisions. We can model the shift of the peak of the distribution as an additive offset to the mean of a Gaussian. Then we have:

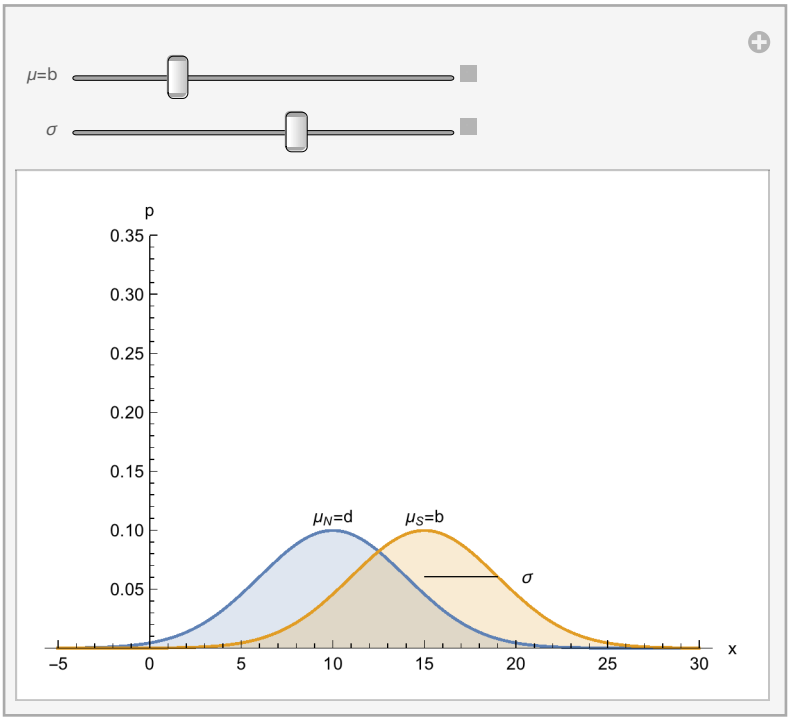
$H = S_H$: x = b + noise;
$H = S_L$: x = d + noise;

where noise is a Gaussian distributed random variable with mean, $\mu = 0$, and standard deviation σ. For the photon counting case, **b=highmean**, and **d=lowmean**. ("b for bright" and "d for dim"). Note that the standard deviations of the high and low distributions would, for a Poisson distribution, be different (variance = mean for Poisson). We will assume that for a typical discrimination task, the distributions are quite close together, so the standard deviations are almost equal. The assumption of Gaussian distributions with equal variance is common, because it simplifies calculations, but more importantly because in many practical cases of discrimination, the approximation is pretty good.

Here is a plot of the theoretically predicted histograms for a signal (high) mean of 15, a noise (low) mean of 10, and a standard deviation of 4 for each:

In[1]:=
```
gauss[x_, μ_, σ_] := PDF[NormalDistribution[μ, σ], x];
b = 15; d = 10; sigma = 4; max = gauss[0, 0, sigma];
Manipulate[
 Plot[{gauss[x, d, sigma2], gauss[x, b1, sigma2]}, {x, -5, 30}, AxesLabel → {"x", "p"},
  Filling → Axis, PlotRange → {0, max + 0.25}, Epilog → {Text["μS=b", {b1, 0.11}],
    Text["σ", {b1 + sigma2 * 1.4, (Exp[-.5] / (Sqrt[2.0 * Pi] * sigma2))}],
    Line[{{b1, (Exp[-.5] / (Sqrt[2.0 * Pi] * sigma2))}, {b1 + sigma2,
       (Exp[-.5] / (Sqrt[2.0 * Pi] * sigma2))}}], Text["μN=d", {d, 0.11}]}],
 {{b1, b, "μ=b"}, d, 30}, {{sigma2, sigma, "σ"}, 1, 6}]
```

Out[3]=



The signal-to-noise ratio: d', a summary statistic for ideal performance

Whether a measurement x comes from the signal or the noise is less uncertain if the difference in means, b-d, is big. But it is also less uncertain if the standard deviation σ is smaller. By using Gaussian distributions (with equal variances), we can define the ideal's signal-to-noise ratio with one number called d':

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

where b and d are the high and low means, respectively. This makes intuitive sense. Inspecting the expression shows that d' will increase if the difference between the means, b - d, gets bigger, or if the denominator σ, representing uncertainty in the noise, gets smaller. d' is called "d prime".

What is the relationship between d' and performance?

The inference model for signal discrimination given additive gaussian noise

How does the ideal observer make a decision as to whether the low or high light was flashed? Earlier we derived a criterion starting from the assumption that we wanted to *maximize the posterior probability* (p(H|x) over H:

1) $\underset{H}{\text{argmax}}\ p(H|x)$

Which means "find that value of H (e.g. H = “switch high” vs. H= “switch low”) which makes p(H|x) the biggest". From here, we showed that if the prior probabilities over the hypotheses were the same ($p(H=S_H) = p(H=S_L)$), this was equivalent to *maximizing the likelihood*:

2) $\underset{H}{\text{argmax}}\ p(x|H)$

Because we are considering only two hypotheses, we could reformulate the decision strategy to testing whether the ratio, $\frac{p(x|S_H)}{p(x|S_L)}>1$? This in turn, is equivalent to testing: $\log[\frac{p(x|S_H)}{p(x|S_L)}]>0$.

Earlier we applied the likelihood ratio rule to our dot density version of light intensity discrimination. We showed the decision could be based on whether the photon count was bigger than a particular criterion (call it X_T) determined by two light level means (b and d).

Similar to what we observed for the Poisson example last time, the ideal that minimizes its error rate makes its decision by deciding "high" if the measurement x is right of the cross-over point on the above plot (i.e. x where $\frac{p(x|S_H)}{p(x|S_L)} = 1$)

This minimizes the probability of error, but how is error related to the decision criterion and the properties of the distributions?

Let's consider the more general case, where the criterion isn't necessarily at the cross-over point.

Hit, false alarm (positive), miss, and correct rejection rates.

It is easy to imagine how one might experimentally measure the signal-to-noise (d’) ratio for light discrimination for the ideal observer—we just collect histograms under the two conditions ($H=S_L$ and $H = S_H$), approximate them by Gaussian distributions (giving us two conditional probability distributions), and assuming the standard deviations are close, use these Gaussian fits to estimate d'. We simulated doing this kind of thing in the last couple of lectures.

But how could we possibly measure the signal-to-noise ratio, or d' inside the head of a human observer?

Human decisions are based on some hidden, and probably quite complex neural mechanism in the brain (cf. Yang & Shadlen, 2007). It seems like we'd need to have access to a neural response that behaves like the ideal's decision variable, but is consistent with human performance (which is usually sub-ideal). This is an interesting scientific problem, but let's see if we can put a number on human d', without "going inside the box"--just based on behavioral performance.

To answer this question, let's take a look at an alternative way of estimating d' for the ideal observer. The ideal observer (or receiver) for light intensity discrimination has two ways of being right and two ways of being wrong:

Being correct:

it can score a

hit (e.g. says "high" when the switch was set to high)

or a

correct rejection (e.g. says "low" when the switch was set to low)

Being incorrect:

it can suffer a

false alarm (also called "false positive") (e.g. says "high" when the switch was set to low)

or a

miss (or "false negative") (e.g. says "low" when the switch was set to high)

(Statisticians talk about a similar distinction in terms of Type I (false positive) and Type II (false negative) errors).

Rates

Average performance in a yes/no task is completely characterized by calculating the proportions of two of the four. Hit and false alarm rates can be treated as estimates of conditional probability distributions on *responses:* p(response | switch setting, H). For example,

$$\textbf{hit rate} = \frac{\#\,\text{times observer says "high when switch was set to high"}}{\#\,\text{times switch was set to high}}$$

~ p(decide high | switch set on high)

$$\textbf{false alarm rate} = \frac{\#\,\text{times observer says "high" when switch was set to low}}{\#\,\text{times switch was set to low}}$$

~ p(decide high | switch is set to low)

Sometimes, we talk about the average error rate. Since there are two ways of being wrong: Deciding "high" when H = S_L, and deciding "low", when H = S_H. The total error rate is the (weighted) average of the miss and false alarm rates. The error rate is determined by the mean values for the high and low settings. As b increases, the separation between the probability distributions increases, and the overlap decreases, so the error rate decreases. So intuitively, there should be some relationship between d' and error and/or success rates.

► 4. Show that we only need measures of hit and false alarm rates because they are simply related to the correction rejection and miss rates.

The corresponding correct rejection and miss rates are:

p(correct rejection) = 1 - p(false alarm),

and

p(miss) = 1 - p(hit), respectively.

► 5. How should one compute weighted average for error rate?

Graphical view of the hit, false alarm rate, ...

For a probability density (continuous distribution) function (i.e. a "PDF"), say p(x), the probability of a measurement X falling within a certain range is given by the area under the density over that range:

$$P(x1 < X < x2) = \int_{x1}^{x2} p(x)\, dx$$

The probability that X falls anywhere left of X = x_T is:

$$P(X < x_T) = \int_{-\infty}^{x_T} p(x)\, dx$$

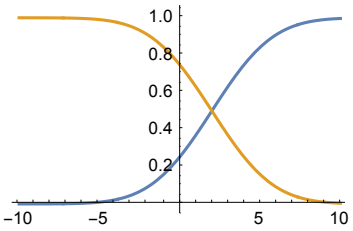
$P(X < x_T)$ is called the *cumulative probability distribution function of* x_T. The probability of X falling to the right of x_T is:

$$P(X > x_T) = \int_{x_T}^{\infty} p(x)\, dx$$

Note that $P(X > x_T) = 1 - P(X < x_T)$, because the total area under any probability density has to equal 1 by definition. See Exercise above.

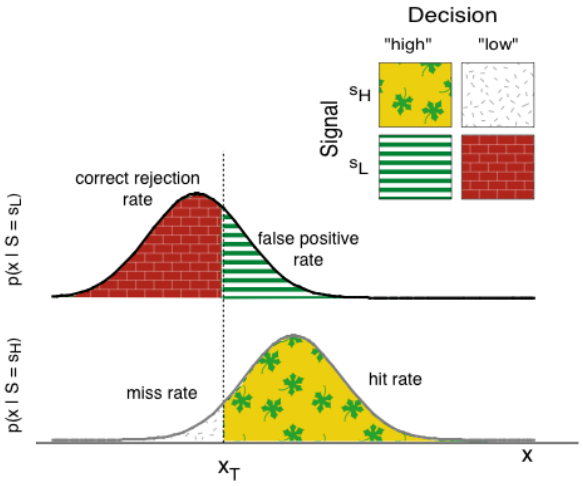
Here are plots of: $\int_{x_T}^{\infty} p(x)\, dx$ and $\int_{-\infty}^{x_T} p(x)\, dx$, for $p(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma} = \frac{e^{-\frac{(x-2)^2}{2\cdot 3^2}}}{\sqrt{2\pi}\,3}$.

```
Plot[{CDF[NormalDistribution[2, 3], x], 1 - CDF[NormalDistribution[2, 3], x]},
 {x, -10, 10}, ImageSize → Small]
```



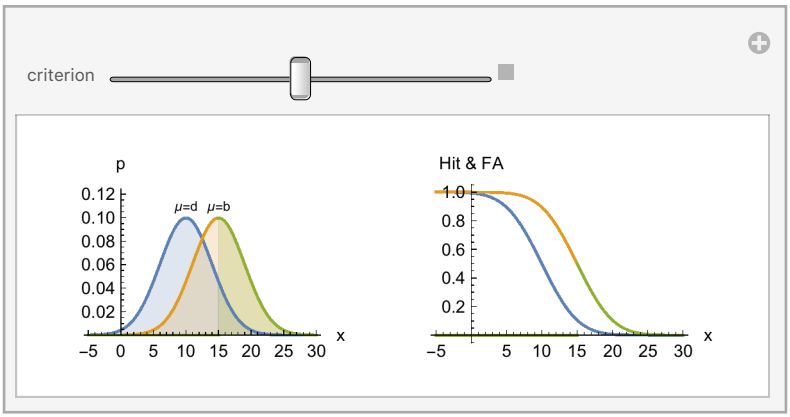
We use 1-CDF[NormalDistribution[μ,σ],x] below.

The criterion, and thus the hit and false alarm rates could be determined by the relative costs or benefits (loss or gain) one assigns to a particular choice of hit and false alarm rates. Suppose the criterion is x_T, in general not at the cross-over point of the likelihoods (which would only be optimal for constant prior probabilities and the goal of minimizing average error). Then the hit rate (P_H) is determined by the area under the (signal or high) curve to the right of x_T. The false alarm rate (P_{FA}) is given by the area under the ("noise" or low) curve to the right of x_T.



In[4]:=
```
b = 15; d = 10; sigma = 4; max = 1.1;
ndistd = NormalDistribution[d, sigma];
ndistb = NormalDistribution[b, sigma];
max = PDF[ndistb, b];
Manipulate[
 g1 = Plot[{PDF[ndistd, x], PDF[ndistb, x],
    (UnitStep[x - c] * Max[PDF[ndistb, x], PDF[ndistd, x]])}, {x, -5, 30},
   AxesLabel → {"x", "p"}, Filling → Axis, PlotRange → {0, max + 0.025},
   Epilog → {Text[Style["μ=b", 7], {b, 0.11}], Text[Style["μ=d", 7], {d, 0.11}]}];
 g2 = Plot[{1 - CDF[ndistd, x], 1 - CDF[ndistb, x],
    (UnitStep[x - c] * (1 - CDF[ndistb, x]))},
   {x, -5, 30}, AxesLabel → {"x", "Hit & FA"}];
 GraphicsGrid[{{g1, g2}}], {{c, b, "criterion"}, 0, 30}]
```

Out[8]=



Criterion shifts affect hit and false alarm rates, but don't affect basic sensitivity, because d' is fixed.

Consider again our decision rule:

if x > X_T guess "high switch caused the intensity measured"

if x<=X_T guess "low switch caused the intensity measured"

In general, where the criterion gets placed depends on the decision goal. One could have other goals (than minimizing error) that would determine where to put the criterion level. Put yourself in the place of an ideal (not a MAP observer) with the following constraints:

If you were slapped on the wrist every time you said "high", you might never say high--you would never get any hits. This in effect pushes the criterion far to the right.

If you liked chocolates as much as I do, and received a sweet every time you said high, you might always say high, even if you thought the signal was not presented-- after all, why not be optimistic? You would have many false alarms. This pushes the criterion far to the left.

So the goal doesn't have to be determined by maximizing the proportion of correct responses (minimizing error), it can be determined by other criteria, which in turn modify the decision rule. These other factors can be incorporated into a pay-off matrix (see Green and Swets, 1974).

In that we haven't changed the means or standard deviations, d' hasn't changed. Hit and correct rejec-

tion rates trade-off against each other. You can get more hits but at the expense of making more false alarm mistakes. (Recall correction rejection rate = 1 - hit rate). It seems like there should be some way to calculate d' from the hit and false alarm rates. We'll see how to do that shortly. Later we will look at formalizing and generalizing the notions of costs and benefits as statistical decision theory. For example, the cost to an error in an estimate of illumination can be low as compared to a cost in the error of face identification.

Summary of modified log likelihood rule:

Thus we can see that one could derive a simple modification of the log likelihood rule:
Rather than testing:

$$\log[\frac{p(x|S_H)}{p(x|S_L)}] > 0?,$$

instead we decide using:

$$\log[\frac{p(x|S_H)}{p(x|S_L)}] > k?, \text{ (equivalent to } \frac{p(x|S_H)}{p(x|S_L)} > e^k \text{)}$$

where k is a function of the costs and benefits. Optimal decisions are based on the value of likelihood ratio. This ratio (or any monotonic function of it) is called the *decision variable*. The photon (or dot) count is a decision variable. More generally, a decision variable may or may not lead to optimal performance. For example, you may base your decision on counting only half of the dots in our dot flash demo.

Relationship between signal-to-noise ratio (d') and hit and false alarm rates.

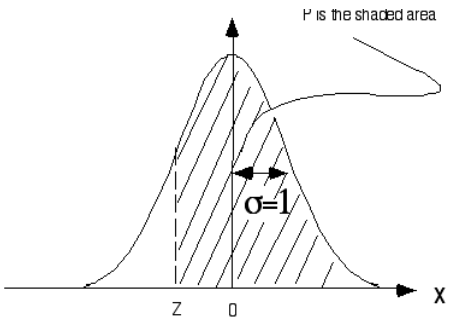
We are now ready to see how to estimate the signal-to-noise ratio "inside an observer's head" using only performance measures of hit and false alarm rates. We need this to compare human d' with an ideal d'.

We noted that the signal-to-noise ratio d' can be estimated from the means and standard deviation. But if we don't have access to those numbers (as happens in psychophysics), we need a method to go from performance (e.g. hit and false alarm rates) back to a function of a hypothetical mean and standard deviation.

It turns out that with a bit of mathematics, one can show that d' can be obtained from the hit and false alarm rates using the following formula:

$$d' = z(PFA) - z(PH) \quad (1)$$

where z(p) is the z-score given a probability p. The figure below illustrates the relationship between probability P and z using the standard normal density (i.e. gaussian with zero mean and a standard deviation of 1):



A z-score function can be obtained as the negative of the inverse of the normal cumulative distribution, with mean and standard deviation equal to zero, and one, respectively:

```
z[p_] := -InverseCDF[NormalDistribution[0, 1], p]
```

This section has provided us with a key result.

In order to estimate the d' "inside the head", we can do a yes/no experiment to estimate the proportions of hits and false alarms, run these through the z[ ] function, and take the difference, which is equal to d'.

Statistical efficiency of human vision: Humans vs. ideals

The basic idea: human observers are sub-ideal, but may be "ideal-like"

The idea is to model the human signal discrimination as being "ideal-like" in assuming that human decisions respect an implicit generative model:

$H = S_H$: x= b^h + noise';
$H = S_L$: x = d^h+ noise';

where b^h, d^h, and noise' are the equivalent states of the world (corresponding to the two effective means of the human observer) that could give rise to the human's d', as measured from hit and false alarm rates:

$$d'_{human} = z(PFA_{human}) - z(PH_{human})$$

It is *as if* the human visual system does Bayesian inference (is ideal), but has the wrong, internalized generative model--i.e. a different state of the world. The previous section showed that d' for a human observer can be determined by hit and false alarm rates, which would correspond to some internal values of b^h, d^h, and σ^h , which we can't directly measure. But we can estimate the relationship represented by:

$$d' \text{ for human} = \frac{b^h - d^h}{\sigma^h},$$

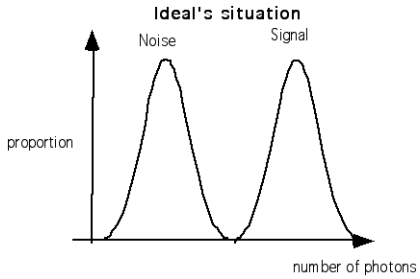
There is indeterminacy in these "implicit" variables, b^h, d^h, and σ^h (the standard deviation of an equivalent noise') -- there is an infinite family of combinations of b^h, d^h, and σ^h which give the same d'.

How good is the linear additive gaussian noise model for human detection behavior? One way of testing it is to plot hit and false alarm rates for human decisions and compare them to this "sub-ideal" that has additive gaussian noise with equal variances. Surprisingly often, the model fits are quite good. But first, lets see how we can make an absolute comparison of performance.

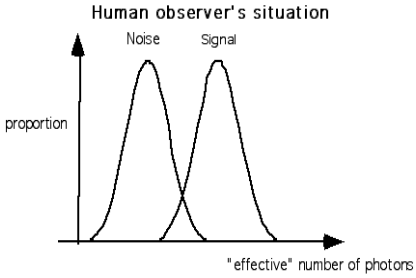
Comparing ideal and human performance

For the light discrimination problem, the physics of the experiment determines the generative model, i.e. the mean levels d, b and the standard deviation. We have seen that the ideal's performance is characterized by one number called the sensitivity d'. Now that we understand the limitations on the performance of an ideal observer, let's how to compare human performance to the ideal. Even if the ideal is making near perfect discriminations, the human observer may not be doing so well because of other sources of

uncertainty. For example, the ideal may be contending with the following situation:



We can't "see" or directly measure the distributions that the human observer is using to make the decision, but we can suppose that it is based on distributions that are in effect much closer together:



Or they could be noisier--i.e. human behavior may be consistent with a bigger standard deviation than the ideal is coping with. Although we can not measure the human's d' by measuring the separation between these two distributions and their standard deviations (they are not directly measurable), we've seen that the human (or ideal) observer's performance can be determined by the hit and false alarm rate: $d' \text{ for human} = z(\text{false alarm rate for human}) - z(\text{hit rate for human})$

Statistical efficiency

Given the means to compute d' for the ideal and for the human observer in the same task, we can compare them. Usually we calculate the ideal's d' from the signal-to-noise ratio, and the human's from the hit and false alarm rate in a yes/no task or from the proportion correct in a 2AFC task. (The ideal's d' could be calculated from its hit and false alarm rates but this usually isn't as convenient--but it might serve as a good way to double-check that you are doing the right calculations.)

With these two d's in hand we can compare the performances of the two observers. One way is in terms of *statistical efficiency*. Efficiency can be defined as the ratio (usually squared) of the two signal-to-noise ratios above:

$$\text{Statistical efficiency} = (d' \text{ for human} / d' \text{ for ideal})^2. *$$

In our case of photon or dot density discrimination, it is possible to show that this ratio is equal to the number of data samples (i.e. # of photons or dots) required by the ideal divided by the number of samples required by the human, when they are performing equivalently (e.g. same hit and false alarm rates).

Statistical efficiency is an old concept in statistics, going back to Sir Ronald Fisher. When measuring photon counts, statistical efficiency is called “quantum efficiency”. The sensitivities of light sensors can be represented by their quantum efficiencies at various wavelengths of light. Statistical efficiency is one,

among several, measures of information capacity that are used in many areas, including neuroscience and psychology.

*(It is the reciprocal of this if the d' represents the physical signal to noise ratios at threshold. See Kersten and Mamassian, 2008)

Historical note -- *Quantum Efficiency accounting for the missing information*

In 1962, Horace Barlow reported results on the measurements quantum efficiency for light discrimination (rather than absolute detection) under low light (scotopic) conditions similar to those of Hecht et al., and came up with a figure for QE of about 10%. That is, the human observer behaved like an ideal observer who was only receiving one out of every ten photons. Where was the photon loss coming from? Like we saw for Hecht et al., Barlow traced the losses to reflection, scatter and absorption by the optic media, and losses due to photons falling in the spaces between the rods, and an imperfect isomerization efficiency. Recall that a figure of 10% is close to what one would predict from Hecht et al.'s experiment.

Barlow later went one step beyond Hecht et al.. He concluded (Barlow, 1977) that there was still a residual inefficiency even after taking into account all the above causes, which he calculated as accounting for only 80% of the photon loss. He was left with about 50% of human discrimination efficiency due to limitations in the brain's ability to "count" point events. That is, for example, if 100 photons are incident on at the cornea of the eye, about 20 of these are reliably transduced and this information is sent to the brain. But he argued, the brain deals with this average of 20 photons with 50% efficiency--that is, the ideal's "brain" could discriminate just as well with only an average of 10 photons. Barlow made this latter conclusion by a clever argument involving a psychophysical experiment in which he had observers discriminate differences in dot density (rather than photon density) on a CRT screen. The idea was that although the presence of a photon at the retina does not necessarily make it to the brain, a dot will.

Psychophysical tasks & techniques

Testing our assumptions: The Receiver Operating Characteristic (ROC)

Let's now go back to the question: *How good is the linear additive gaussian noise model for human detection behavior?*

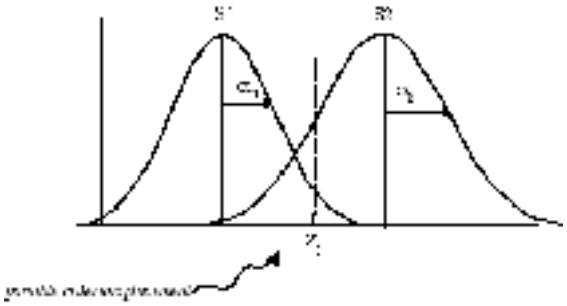
Although we can't directly measure the internal distributions of a human observer's decision variable, we've seen that we can measure hit and false alarm rates, and thus d'. But one can do more, and actually test to see if an observer's decisions are consistent with Gaussian distributions with equal variance. If the criterion is varied, we can obtain a set of n data points, each coming from a block of trials:

{(hit rate 1, false alarm rate 1), (hit rate 2, false alarm rate 2), ..., (hit rate n, false alarm rate n)}

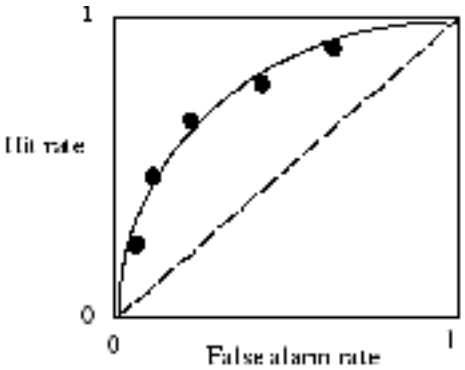
and all from one experimental condition (i.e. from one fixed signal-to-noise ratio, call it d_{ideal}'). This is because as the hit rate varies, so does the false alarm rate (see the above figures showing how hit and false alarm rates relate to area under the signal and noise distributions.). One could compute the d' for each pair and they should all be equal for the ideal observer. Of course, we would have to make a large

number of measurements for each one--but on average, they should all be equal.

To get meaningful and equal d's for each pair of hit and false alarm rates assumes that the underlying relative separation of the signal and noise distributions remain unchanged and that the distributions are Gaussian, with equal standard deviation. We might know this is true (or true to a good approximation) for the ideal, but we have no guarantee for the human observer. Is there a way to check? Suppose the signal and noise distributions look like:

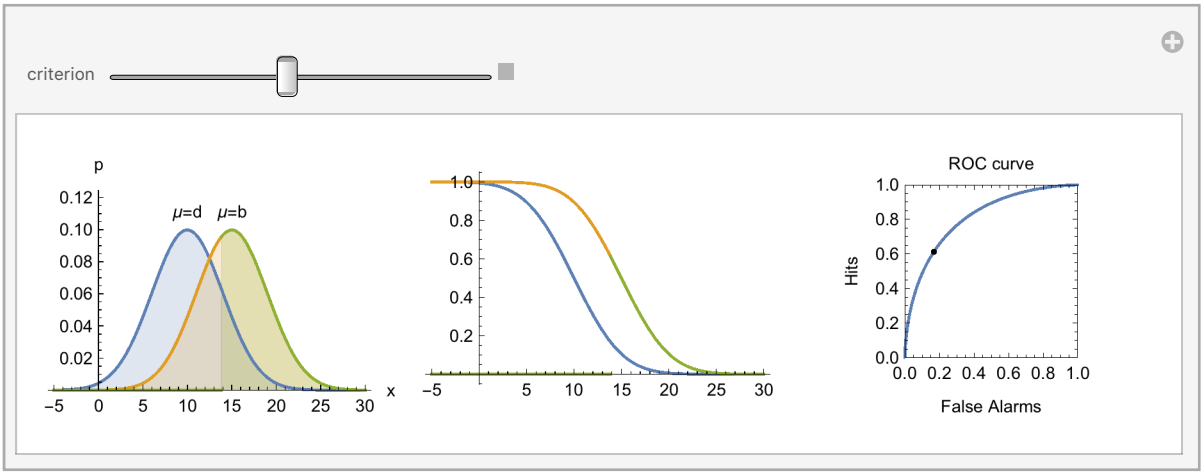


If we plot the hit rate vs. false alarm rate data on a graph as the criterion xc varies, we get something that looks like:



```
b = 15; d = 10; sigma = 4;
ndistd = NormalDistribution[d, sigma];
ndistb = NormalDistribution[b, sigma];
max = PDF[ndistb, b];

Manipulate[
 g1 = Plot[{PDF[ndistd, x], PDF[ndistb, x],
    (UnitStep[x - c] * Max[PDF[ndistb, x], PDF[ndistd, x]])}, {x, -5, 30},
   AxesLabel → {"x", "p"}, Filling → Axis, PlotRange → {0, max + 0.025},
   Epilog → {Text["μ=b", {b, 0.11`}], Text["μ=d", {d, 0.11`}]}];
 g2 = Plot[{1 - CDF[ndistd, x], 1 - CDF[ndistb, x],
    (UnitStep[x - c] * (1 - CDF[ndistb, x]))}, {x, -5, 30}];
 g3 = ParametricPlot[{{1 - CDF[ndistd, x], 1 - CDF[ndistb, x]}}, {x, -100, 100},
   FrameLabel -> {{"Hits", ""}, {"False Alarms", "ROC curve"}},
   PlotRange → {{0, 1}, {0, 1}}, Frame → True, AspectRatio → 1,
   Epilog → {Point[{1 - CDF[ndistd, c], 1 - CDF[ndistb, c]}]}];
 GraphicsGrid[{{g1, g2, g3}}, ImageSize → Large], {{c, b, "criterion"}, 0, 30}]
```



One can do some math to show that:

the area under the ROC curve is equal to the proportion correct in a two-alternative forced-choice experiment (see Green and Swets).

From a purely empirical standpoint it can be useful to measure *sensitivity by the area under the ROC curve*. This provides a single summary number, even if the standard definition of d' is inappropriate, for example because the variances are not equal, or the distributions are not gaussian.

We return to our basic question: is there a way to spot whether our gaussian equal-variance assumptions are correct for human observers? If we take the same data and plot it in terms of Z-scores we transform the ROC curve to a straight line:

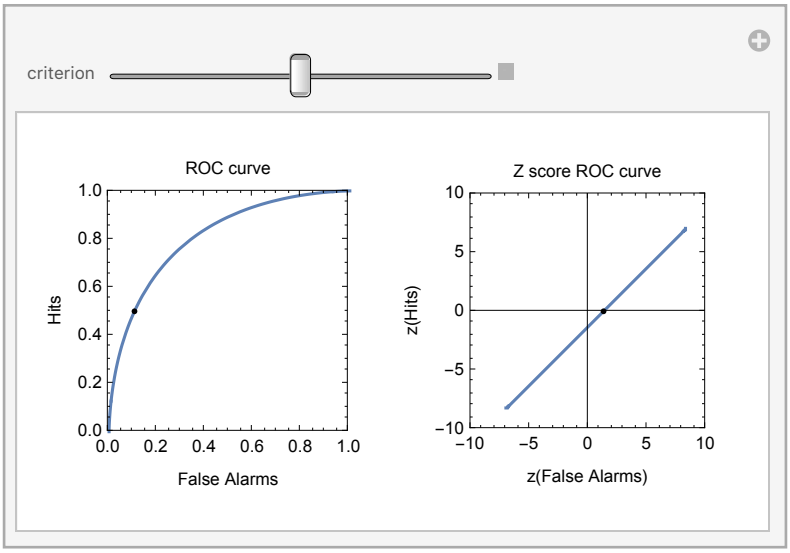
In[9]:=
```
b = 15; d = 10; sigma = 4;
ndistd = NormalDistribution[d, sigma];
ndistb = NormalDistribution[b, sigma];
max = PDF[ndistb, b];
zscore[p_] := -InverseCDF[NormalDistribution[0, 1], p];

Manipulate[
 g3 = ParametricPlot[{{1 - CDF[ndistd, x], 1 - CDF[ndistb, x]}},
   {x, -100, 100}, FrameLabel -> {{"Hits", ""}, {"False Alarms", "ROC curve"}},
   PlotRange → {{0, 1}, {0, 1}}, Frame → True, AspectRatio → 1,
   Epilog → {Point[{1 - CDF[ndistd, c], 1 - CDF[ndistb, c]}]}];

 g4 = ParametricPlot[
   {{zscore[1 - CDF[ndistd, x]], zscore[1 - CDF[ndistb, x]]}}, {x, -100, 100},
   FrameLabel -> {{"z(Hits)", ""}, {"z(False Alarms)", "Z score ROC curve"}},
   PlotRange → {{-10, 10}, {-10, 10}}, Frame → True, AspectRatio → 1,
   Epilog → {Point[{zscore[1 - CDF[ndistd, c]], zscore[1 - CDF[ndistb, c]]}]}];

 GraphicsGrid[{{g3, g4}}], {{c, b, "criterion"}, 0, 30}]
```

Out[13]=



In fact, if the underlying distributions are Gaussian, the data should lie on a straight-line. If they both have equal variance, the slope of the line should be equal to one. This is because:

$$Z\,(\text{hit rate}) \;=\; \frac{X_c - \mu_s}{\sigma_s}$$

$$Z\,(\text{false alarm rate}) \;=\; \frac{X_c - \mu_n}{\sigma_n}$$

And if we solve for the criterion X_c , we obtain:

$$Z(\text{hit rate}) = \frac{\sigma_n}{\sigma_s} Z(\text{false alarm rate}) - \frac{\mu_s - \mu_n}{\sigma_s}$$

(I've switched notation here, where b = μ_s, and d = μ_n). The main point of this plot is to see if the data tend to fall on a straight line with slope of one. If a straight line, this would support the Gaussian assumption. A slope = 1 supports the assumption of equal variance Gaussian distributions.

In practice, there are several ways of obtaining an ROC curve in human psychophysical experiments. One can vary the criterion that an observer adopts by varying the proportion of times the signal is presented. As observers get used to the signal being presented, for example, 80% of the time, they become biased to assume the signal is present. One needs to block trials in groups of, say 400 trials per block, where the signal and noise priors are fixed for a given block.

One can also use a *rating scale* method in which the observer is asked to say how confident she/he was (e.g. 5 definitely, 4 quite probable, 3 don't know for sure, 2, unlikely, 1 definitely not). Then we can bin the proportion of "5's" when the signal vs. noise was present to calculate hit and false alarm rates for that rating, do the same for the "4's", and so forth. The assumption is that an observer can maintain not just one stable criterion, but four---the observer has in effect divided up the decision variable (x) domain into 5 regions. An advantage of the rating scale method is efficiency--relatively few trials are required to get an ROC curve. Further, in some experiments, ratings seem psychologically natural to make. But if there is any "noise" in the decision criterion itself, e.g. due to memory drift, or whatever, this will act to decrease the estimate of d' in both yes/no and rating methods.

Usually rather than manipulating the criterion, we would rather do the experiment in such a way that it does not change. Is there a way to reduce the problem of a fluctuating criterion?

The 2AFC (two-alternative forced-choice) method

Relating performance (proportion correct) to signal-to-noise ratio, d'.

In psychophysics, the most common way to minimize the problem of a varying criterion is to use a two-alternative forced-choice procedure (2AFC). In a 2AFC task the observer is presented on each trial a pair of stimuli. One stimulus has the signal (e.g. high mean, bright flash), and the other the noise (e.g. low mean, dim flash). The order, however, is randomized. So if they are presented temporally, the signal or the noise might come first, but the observer doesn't know which from trial to trial. Or if presented spatially, the observer doesn’t know whether the left or right side as the signal. An ideal strategy is to compute the cross-correlation decision variable for each image (i.e. the dot product between each image vector and an exact template of the signal one is looking for), and pick the image which gives the larger cross-correlation. This strategy will result in a single measurable number, the proportion correct, Pc. One can show that for 2AFC:

$$d' = -\sqrt{2}\, z(\text{proportion correct}) \tag{2}$$

where as before, z(P) is given by the inverse of:

$$P = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{x^2}{2}} dx$$

i.e., the area under the normal distribution to the right of z, which can be calculated as: 1 minus the inverse cumulative normal distribution. The following code calculates z, given P:

```
z[p_]:=-InverseCDF[NormalDistribution[0,1],p];
dprime[x_] := N[-Sqrt[2] z[x]]
```

▶ 6. Exercise: Prove d' = $-\sqrt{2}$ z(proportion correct)

If you want to prove this for yourself, here are a couple of hints--actually, a lot of hints. Let us imagine we are giving the light discrimination task to the ideal observer. We have two possibilities for signal presentation: Either the signal is on the left and the noise on the right, or the signal is on the right and the noise on the left. There are two ways of being right. The observer could say "on the left" when the signal is on the left, or "on the right" when the signal is on the right. For example, for the light detection experiment, a reasonable guess is that all the ideal observer would have to do is to count the number of photons on the left side of the screen and count the number on the right too. If the number on the left is bigger than the number on the right, the observer should say that the signal was on the left. Thus, a 2AFC decision variable would be the difference between the left and right decision variables, where each of these is what we calculated for the yes/no experiment.

$r = r_L - r_R$

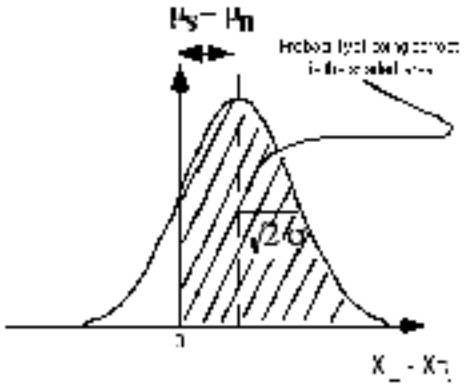
For example, r_L and r_R for the SKE observer would be the dot products of the signal pattern template with observation image vectors on the left and right sides.
So, the probability of being correct is:
pc = p(r>0|signal on left) p(signal on left) + p(r<0| signal on right)p(signal on right)

What is the probability distribution of r? Well, from probability rules (see: ProbabilityOverview.nb),

average(r) = $\mu 2 - \mu 1 = \mu_s - \mu_n$

var(r) = var(r_L) + var(r_R), so $\sigma_r = \sqrt{2}\,\sigma_{r_L} = \sqrt{2}\,\sigma_{r_R}$

(Because the mean of the sum of two independent random variables is the sum of their means and that the variance of the sum is the sum of the variances.)
If the signal is equally likely to appear on the left or the right,the probability of being correct is the area under the curve to the right of zero of the distribution of r:

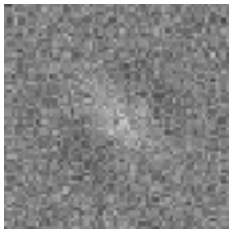
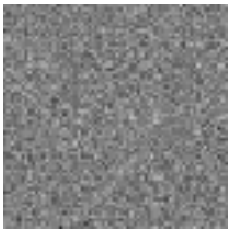


(Note in the above figure: $r = r_L - r_R = \mathbf{x}_L - \mathbf{x}_R$, and $\mu 2 - \mu 1 = \mu_s - \mu_n = b - d$)

Next time

Probability overview: Look at the notebook / pdf

From pixels to patterns: What does the eye see best?

Computing the ideal observer for patterns

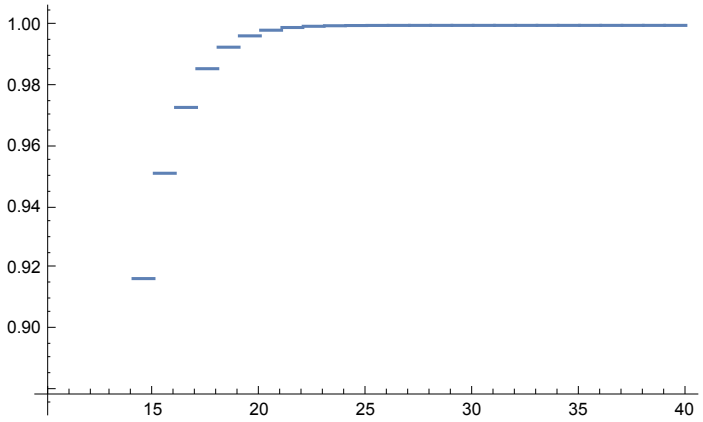
Comparing psychophysical performance for pattern detection with properties of visual neurons in the brain

Notes

Cumulative distributions

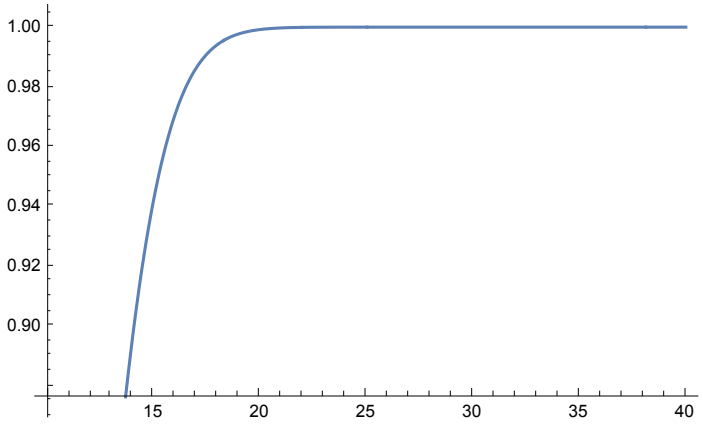
The cumulative distribution gives the probability that the detector signals x<k photons. It is obtained by adding up the probabilities for all values less than k. For the cumulative density function, we integrate over all values less than k. Here is the cumulative distribution for the discrete Poisson distribution with a mean of 20:

```
Plot[CDF[pdist, x], {x, 10, 40}]
```



What is the probability of detecting 50 or less photons when the mean is 20? It is virtually certain-- as you can see from the graph, the probability is almost 1. Here is the plot of the continuous normal distribution with a mean of 20, and a standard deviation of **Sqrt[20]**:

```
Plot[CDF[ndist, x], {x, 10, 40}]
```



References

Applebaum, D. (1996). Probability and Information. Cambridge, UK: Cambridge University Press.

Barlow, H. B. (1962). A method of determining the overall quantum efficiency of visual discriminations. Journal of Physiology (London), 160, 155-168.

Barlow, H. B. (1977). Retinal and central factors in human vision limited by noise. In B. H. B., & F. P. (Ed.), Photoreception in Vertebrates Academic Press.

Barlow, H. B., & Levick, W. R. (1969). Three factors limiting the reliable detection of light by retinal ganglion cells of the cat. *J Physiol, 200*(1), 1-24.

Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis . New York.: John Wiley & Sons.

Jason M. Gold, Craig Abbey, Bosco S. Tjan, and Daniel Kersten (2009) Ideal Observers and Efficiency: Commemorating 50 Years of Tanner and Birdsall: Introduction. JOSA A, Vol. 26, Issue 11, pp. IO1-IO2 doi:10.1364/JOSAA.26.000IO1

Geisler, W. (1989). Sequential Ideal-Observer analysis of visual discriminations. Psychological Review, 96(2), 267-314.

Green, D. M., & Swets, J. A. (1974). Signal Detection Theory and Psychophysics . Huntington, New York: Robert E. Krieger Publishing Company.

Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Ed.), Vision: Coding and Efficiency (pp. 3–24). Cambride: Vision: Coding and efficiency.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. Nature, 423(6941), 752–756. http://doi.org/10.1038/nature01516

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. Nature, 447(7148), 1075–1080. http://-doi.org/10.1038/nature05852

Treutwein, B. (1993). Adaptive psychophysical procedures: A review. Vision Research

Van Trees, H. L. (1968). Detection, Estimation and Modulation Theory . New York: John Wiley and Sons.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. 33, 113-120.

Watson, Andrew B. & Fitzhugh, A., (1990) The method of constant stimuli is inefficient Perception & Psychophysics 47(1), 87-91.