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

Goals

Last time: Introduced ideal observer for a Yes/No task: high or low dot density?

The accuracy and reliability of 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 focused on the notion of **ideal observer**. The "ideal" has a model of the inherent uncertainty (i.e. "external variability"), and makes optimal decisions given that variability or uncertainty. We stated that the ideal observer should choose the hypothesis (e.g. the switch setting in our prototypical light discrimination task) with the highest posterior probability, given the data (photon count). This decision process is an example of Bayesian inference. Bayesian theories of vision provide quantitative models of the information available in a task.

It is important to distinguish between uncertainty inherent to the task from uncertainty specific to the biology when drawing conclusions about the underlying neural mechanisms of the brain from behavioral/psychophysical data. Because of biological limitationsm, such as internal variability, humans are typically not ideal observers. But suppose we are approximately ideal at some task. Then that the pattern of errors would largely reflect the uncertainty in the task itself. If this is the case, then our simplest conclusion about the underlying neural mechanism is that it behaves like an ideal observer, i.e. as a very efficient utilizer of the information available. The ideal observer would then provide a good quantitative model of human perceptual behavior. However, ideal performance by a human observer limits our ability to draw conclusions about the neural mechanisms, which are can be revealed by sub-ideal behavior. In actuality, human perceptual performance is near optimal for some tasks, and not for other tasks. Hecht et al. argued that the variation in the proportion of hits was largely due to photon fluctations, with only smaller contributions from limitations of the human observer, suggesting that the variability was due to a high efficiency of photon transduction. Once we account for photon loss in the periphery, humans are almost ideal. What about other tasks?

We'd like to further develop our tools of signal detection theory, and extend them to perceptual decisions more generally, so that we can quantitatively compare humans to ideal observers. We call this comparison ***ideal observer analysis***. The ideal can be used as a benchmark to measure the performance of humans, as well as machines, and even single neurons for more complicated problems like pattern detection.

Today: Classical SDT and d'

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!) We will:

o 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'. To do this, we will introduce the (standard) Gaussian approximation, and apply it to variability in light levels.

o Introduce other tasks. In particular, the two-alternative forced-choice task

o Understand how to quantitatively compare human and ideal performance.

o Measure your own statistical efficiency in a 2AFC task

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

Signal Detection Theory: Gaussian model

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 non-Gaussian fixed distribution) 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. However, as we will see later (when we measure statistics on natural images), the Gaussian assumption/approximation for image random variables is a bad approximation. It is always important to test this assumption. We'll first show that the Gaussian approximation provides a good approximation to the Poisson distribution.

Some terminology. We've adopted the convention of treating a switch set to high dot density (or on average, brighter light) as a "signal". Similarly, we can think of the low switch settings as playing the role of "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.

What does i.i.d. mean?

Gaussian approximation for signal and noise distributions

As the mean *a* of a Poisson distributed random variablegets large, the frequency of occurrance of can be well approximated by the Gaussian distribution:

$p(X = x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}$. The mean or expectation of X is: $E(X) = \mu$,

and the variance is: $\mathrm{var}(X) = \sigma^2$

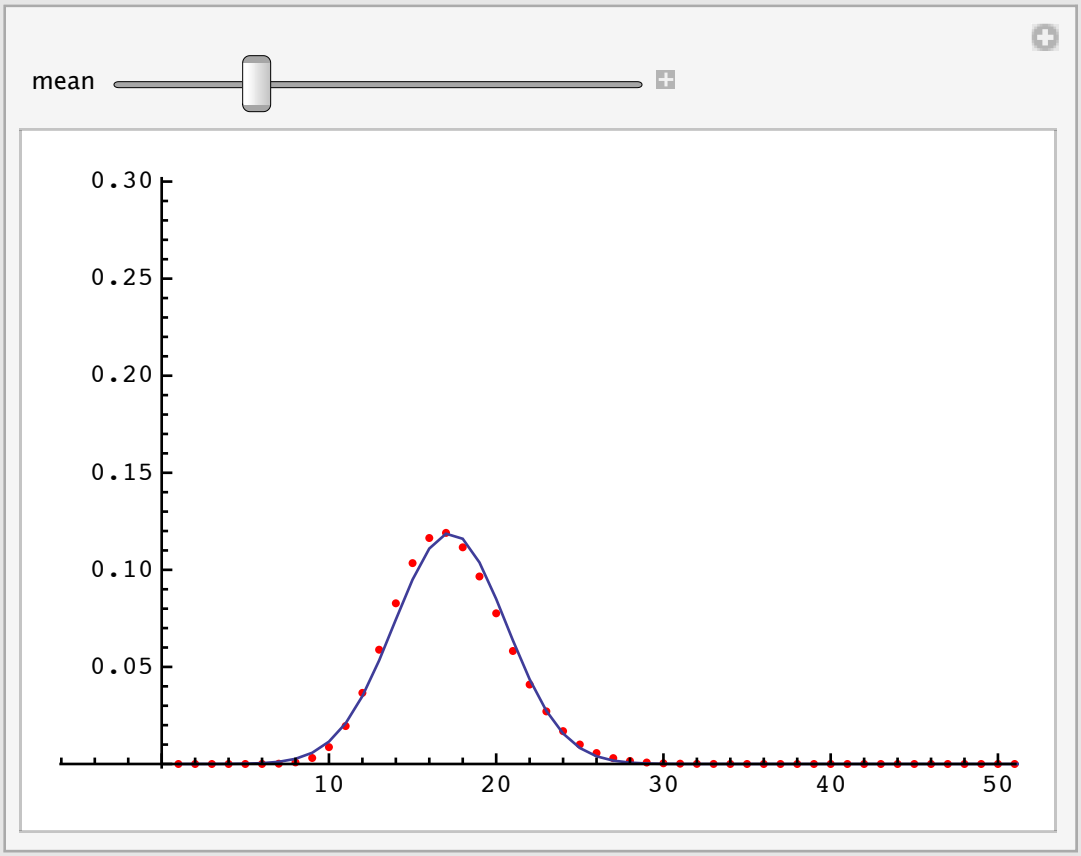
This approximation is useful to estimate probability values for large *a*. If *a* is large enough, the probability of negative values (which is meaningless for a Poisson distribution) is very small. For computational convenience and for later generality, we will usually use the Gaussian approximation.

Let's compare the forms of the Poisson and Gaussian distributions:

In[17]:=

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

Out[17]=



Try comparing the Poisson and Gaussian approximation when the mean is much smaller, like a=4.

■ Note: Discrete vs. continuous distributions

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

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). 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 probability of X=x, for any particular value ($x = \pi$, or x = 3.1, for example) is actually zero! 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 region, dx as:

$$p(x<X<x+dx) \sim p(x)dx.$$

More on this later.

Summarizing performance for an ideal observer

An ideal observer can be characterized by its signal-to-noise ratio, or by its performance in a task, such as Yes/No.

■ The classic *Standard Additive Gaussian* generative model for signal discrimination

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.

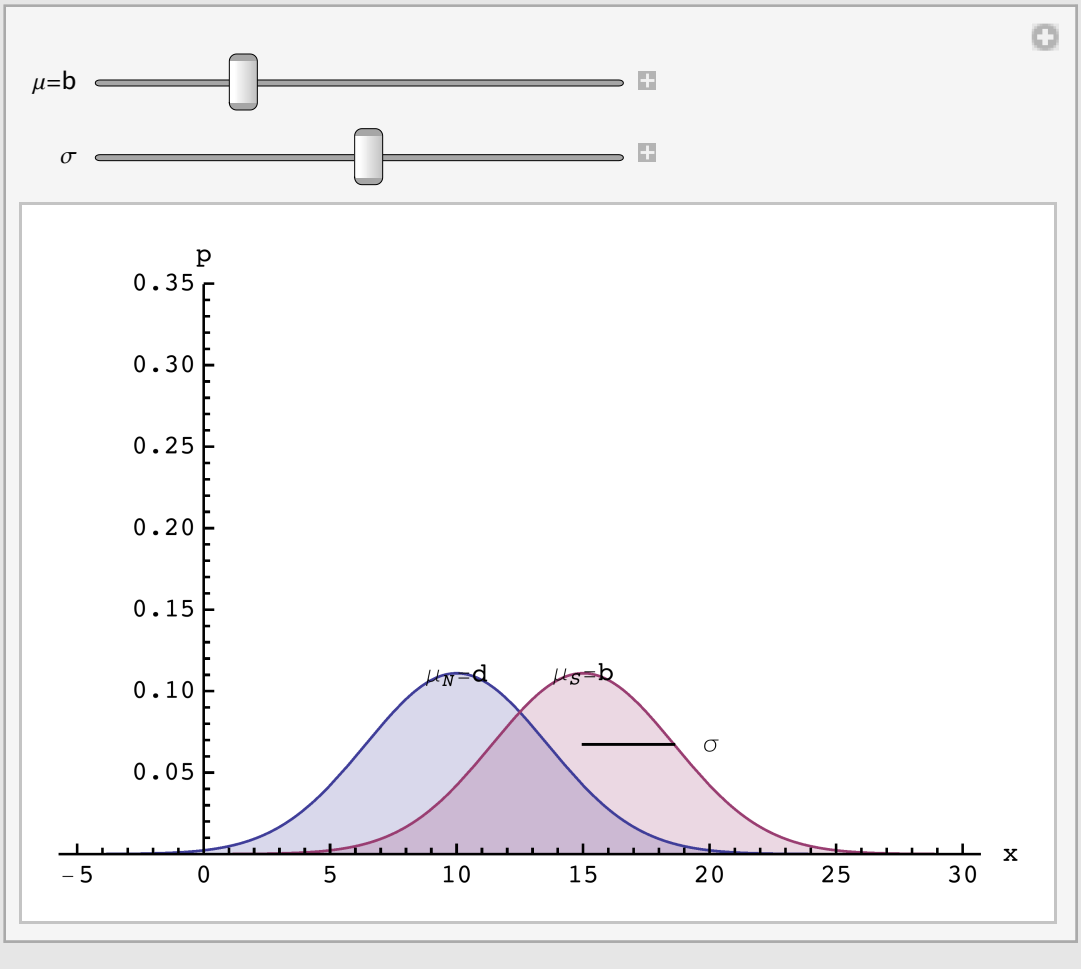
For the photon counting case, **b=highmean**, and **d=lowmean**. ("b for bright" and "d for dim"). Note that the standard deviations of the high and low distributions would, for a Poisson distribution, be different (variance = mean for Poisson). We will assume that for a typical discrimination task, 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[18]:=

```
gauss[x_, mean_, std_] := e^(-(x-mean)^2/(2 std^2)) / (std Sqrt[2 π]);
(*Define our own gaussian distribution*)
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[20]=



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

It is easier to discriminate a difference between bright and dim when the mean difference, **b - d**, is big. But it is also easier if the standard deviation σ is smaller.

By using Gaussian distributions (with equal variances), we can characterize the ideal's signal-to-noise ratio with one number, the "signal-to-noise ratio," defined as d':

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

where b and d are the high and low means, respectively.

This makes intuitive sense. Discrimination should get easier as the difference between the means increases (the "signal" is the difference) or as the spread given by the standard deviation of the additive noise (σ) decreases--hence the term *signal-to-noise ratio*.

But what does this mean in terms of performance?

■ Review criterion for MAP and maximum likelihood decision to minimize error

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. Shorthand for this rule is:

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

Which means "find that value of H (e.g. H = switch high vs. 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: $\log[\frac{p(x|S_H)}{p(x|S_L)}] > 0$?

The previous lecture applied this rule to 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).

Graphically, 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 distributions?

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

■ Performance metrics: hit, false alarm, miss, and correct rejection rates for arbitrary criteria

It is easy to imagine how one might experimentally measure the signal-to-noise 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 possibily measure the signal-to-noise ratio, or d' of a human observer?

Human decisions are based on some hidden, and probably quite complex neural mechanism in the brain. 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".

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:

■ Two ways of being right and two ways of being wrong in a Yes/No task

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, 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)

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

~ 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.

We only need measures of these two because the correction rejection and miss rates are not independent

of the hit and false alarm rates. Show that:

The corresponding correct rejection and miss rates are:

p(correct rejection) = 1 - p(false alarm), and p(miss) = 1 - p(hit), respectively.

How should one compute weighted average for error rate?

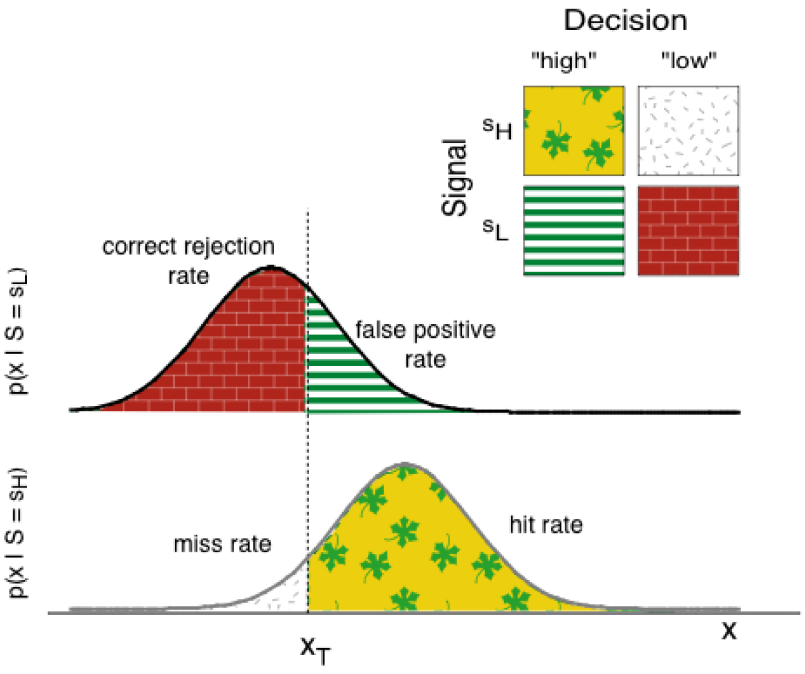
■ A graphical view of the hit, false alarm rate as a function of the criterion

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:

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

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

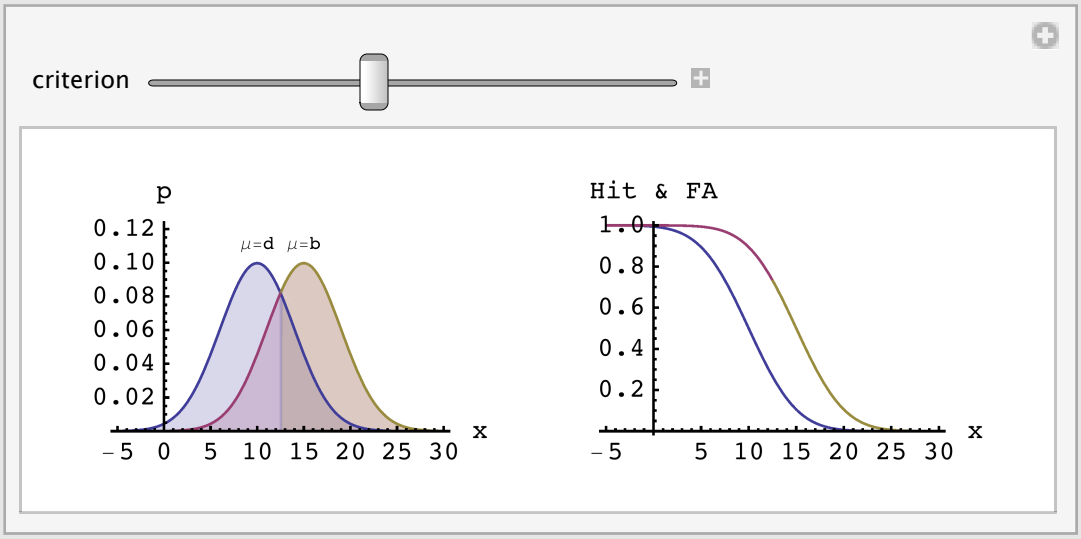
The criterion, and thus the hit and false alarm rates could be determined by the relative costs or benefits (loss or gain) one assigns to a particular choice of hit and false alarm rates. Suppose the criterion is XT, in general not at the cross-over point of the likelihoods (which would only be optimal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



In[21]:=

```
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[25]=



■ Manipulating criterion shifts to affect hit and false alarm rates (with no affect on sensitivity)

A light is flashed, the photon counter indicates x photons received. The decision rule is:

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 rejection 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.

■ Modify log likelihood rule to allow for various decision criteria

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 dots in just the bottom half of the screen.

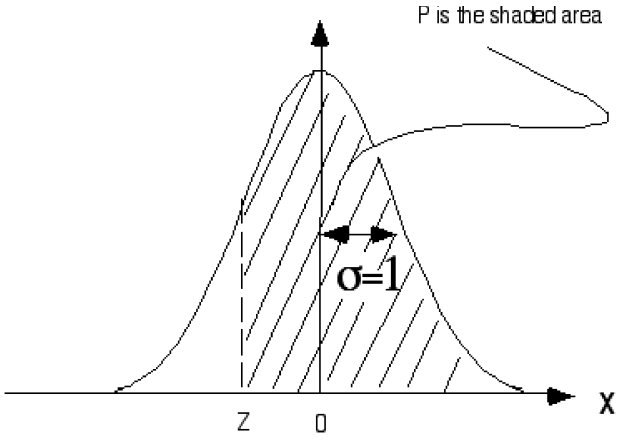
■ Relationship between signal-to-noise ratio (d') and hit and false alarm rates.

We are now ready to show how to estimate the signal-to-noise ratio "inside an observer's head" using only performance measures of hit and false alarm rates. 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), 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:

$$d' = z(PFA) - z(PH)$$

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):



P=$\int_z^\infty$gauss[x, 0, 1] dx. And z(p) is the inverse.

There is no simple formula for z, but there are good closed-form approximations. *Mathematica* doesn't give the direct formula for the z-score, but it does supply the inverse of the **erf[]** function, which comes from the engineering (rather than statistics) tradition:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$$

As you will use in Assignment #1, the z-score function is related to the InverseErf function by:

In[26]:=
```
z[p_] := Sqrt[2] InverseErf[1 - 2 p];
```

Humans vs. ideals: Modeling internal variability of the human observer

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 is optimal, but with the wrong generative model--i.e. a different state of the world. The d' for human, is determined by hit and false alarm rates, or equivalently by:

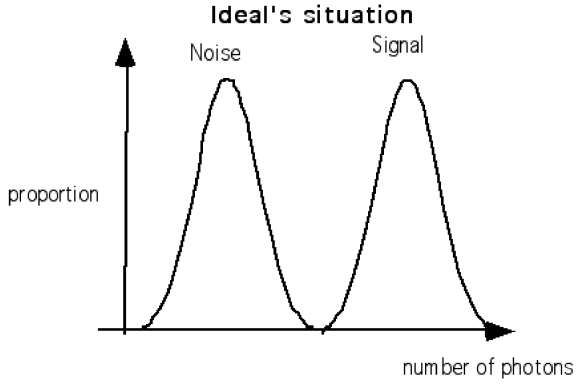
$$d' \text{ for human} = \frac{b^h - d^h}{\sigma^h},$$

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

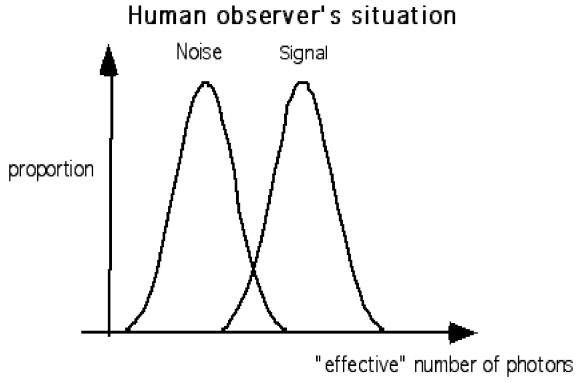
How good is the model? One way of testing it is to plot hit and false alarm rates for human decisions and compare them to this "sub-ideal" that has additive gaussian noise with equal variances. Surprisingly often, the linear, gaussian generative model fits are quite good. But first, lets see how we can 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 us try to understand 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. bigger standard deviation than the ideal is coping with.

Although the the separation between these two distributions and their standard deviations are not directly measurable in a human subject, we can measure the hit and false alarm rate to estimate d' :

```
d' for human = z (false alarm rate for human) - z ( 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 (as in the experiment below). (The ideal's d' could be calculated from its hit and false alarm rates but this usually isn't as convenient--but *Monte Carlo* simulations 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 is defined as the number of samples (e.g. photons in our light discrimination example) 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). If d' is estimated from hit and false alarm scores, it can be shown that:

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

(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 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: Yes/No & 2AFC

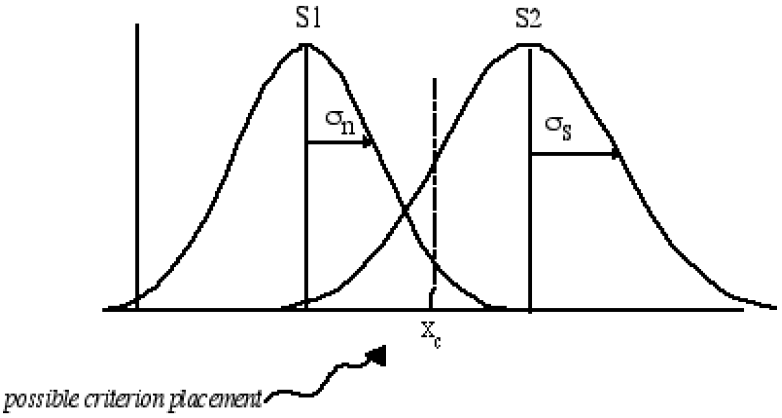
Testing our assumptions: The Receiver Operating Characteristic (ROC) for a Yes/No task

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:

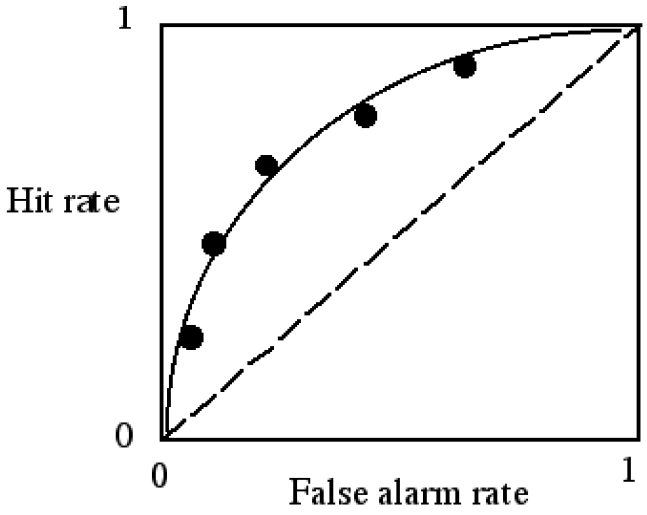
{(hit rate 1, false alarm rate 1), (hit rate 2, false alarm rate 2), ..., (hit rate n, false alarm rate n)}

all from one stimulus condition (i.e. from one 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:

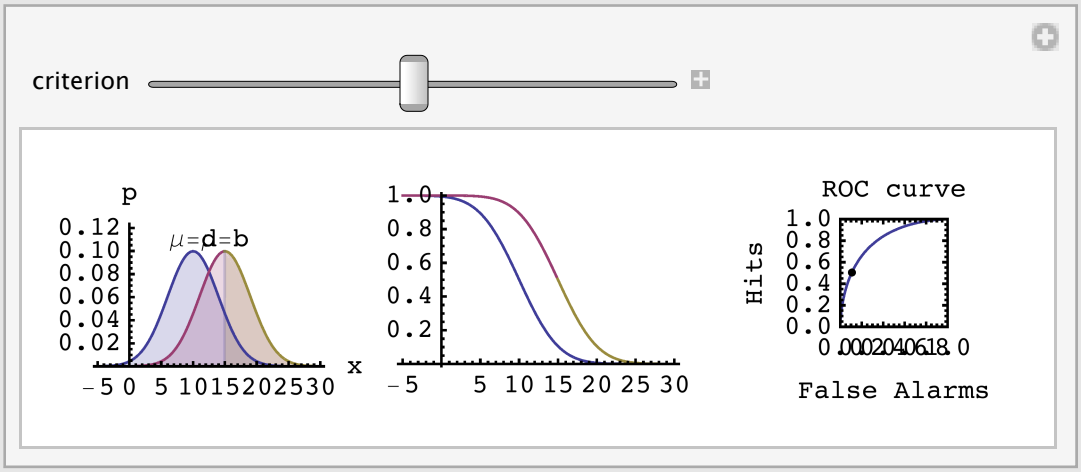


If we plot the hit rate vs. false alarm rate data on a graph as the criterion xc varies, we get something that looks like:



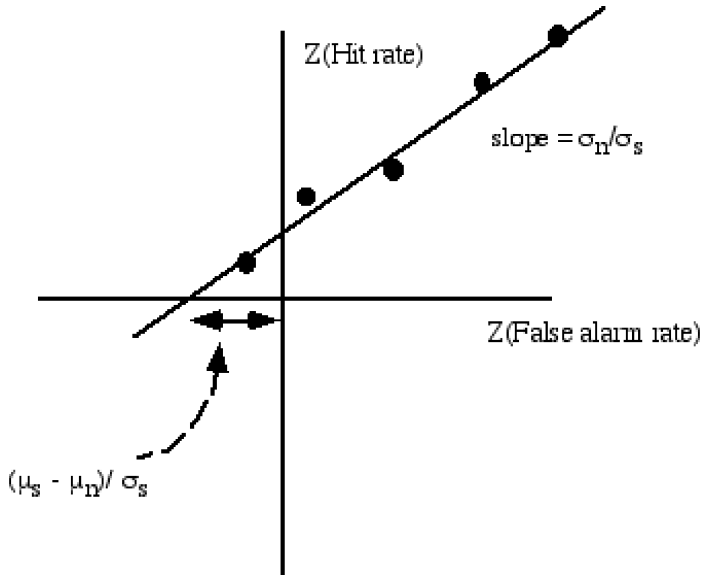
```
b = 15; d = 10; sigma = 4;
ndistd = NormalDistribution[d, sigma];
ndistb = NormalDistribution[b, sigma];
max = PDF[ndistb, b];

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}}], {{c, b, "criterion"}, 0, 30}]
```



So 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 get something like:



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:

$$Z(\text{hit rate}) = \frac{X_c - \mu_s}{\sigma_s} \tag{1}$$

$$Z(\text{false alarm Integrate}) = \frac{X_c - \mu_n}{\sigma_n} \tag{2}$$

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

$$Z(\text{hit rate}) = \frac{\sigma_n}{\sigma_s} Z(\text{false alarm rate}) - \frac{\mu_s - \mu_n}{\sigma_s} \tag{3}$$

(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 flash), and the other the noise (e.g. low 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. In the spatial version, the signal could be randomly positioned on the left of the computer screen with the noise on the right, or vice versa. One can show that for 2AFC:

$$d' = -\sqrt{2}\ z(\text{proportion correct}) \tag{4}$$

As before, the Z-score can be calculated from the inverse of a standard mathematical function called Erf[] to get Z from a measured P.

In[40]:=
```
z[p_] := Sqrt[2] InverseErf[1 - 2 p];
```

where Z(*) is the z-score for P_c , the proportion correct. And then,

```
dprime[x_] := N[-Sqrt[2] z[x]]
```

Exercise: Prove $d' = -\sqrt{2}\ z(\text{proportion correct})$. See Homework Assignment #1.

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.

$$r = r_L - r_R \tag{5}$$

For example as you will see in Assignment 1, 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 probabililty 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)

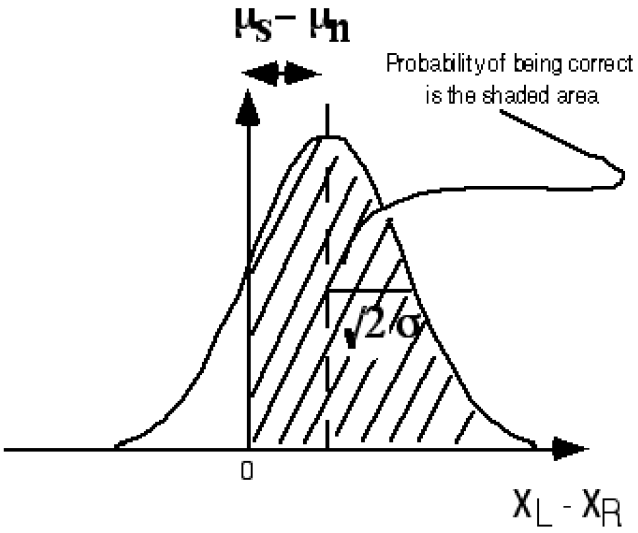
What is the probability distribution of r? Well, from probability rules (see: ProbabilityOverview.nb),

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: $r = r_L - r_R = x_L - x_R$, and $\mu 2 - \mu 1 = \mu_s - \mu_n = b - d$)

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

Sometimes, *sensitivity is operationally defined as 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.

Statistical efficiency: competing with the ideal observer in a 2AFC task

■ Set up the mini-experiment

Let's develop the dot density experiment you piloted last time. We'll make two improvements. First we'll turn it into a two-alternative forced-choice experiment. You'll experience how this makes the task subjectively easier, and reduces the problem of criterion fluctuation. Second, we will calculate your and the ideal's proportion correct, turn these into d's, and from there calculate your statistical efficiency. In 1997, Barlow reported an efficiency near 50%. How good are you?

As above, let's define a Poisson distribution with a mean of **mean**, with a function to draw a sample from this distribution.

For most psychophysical experiments it is a good idea to give the observer some practice trials. For these, let highmean = 300 for the "high" setting, and lowmean =200 for the "low" setting. Set **numtrials=10**; This will give you easy trials to get the hang of it

For the actual measurements, you want to make the task hard enough so that mistakes are made (why?). So make the task harder by setting highmean = 220 for the "high" setting, and lowmean =200 for the "low" setting. Set **numtrials=100**;

```
z[p_] := Sqrt[2] InverseErf[1 - 2 p];
dprime[x_] := N[-Sqrt[2] z[x]];
dotsize = 0.01;
numberofphotons[mean_] := RandomInteger[PoissonDistribution[mean]];
highmean = 220; lowmean = 200;
data = {{"Was I Correct?", "Was Ideal Correct?"}};

numtrials = 10;
```

```
blank = Graphics[{PointSize[dotsize], Black, Point /@ {{}, {}}},
    AspectRatio → 1, Frame → False, FrameTicks → None,
    Background → GrayLevel[0.0], PlotRange → {{-0.2, 1.2}, {-0.2, 1.2}}];
flash = blank;
```

```
CreateDocument[Dynamic[flash], ShowCellBracket → False,
  WindowSize → {300, 300},
  WindowMargins → {{Automatic, 0}, {Automatic, 0}}, WindowElements → {},
  Background → Black, NotebookFileName → "Flash Display"];
```

```
twoflashes := Module[{tempmean},
  Table[whichflash = RandomInteger[{0, 1}];
    If[whichflash == 1, leftnumsample = numberofphotons[highmean],
     leftnumsample = numberofphotons[lowmean]];
    If[whichflash == 0, rightnumsample = numberofphotons[highmean],
     rightnumsample = numberofphotons[lowmean]];

    leftsample = Table[RandomReal[{0, 1}, 2], {leftnumsample}];
    flash = Graphics[{PointSize[dotsize], Red, Point /@ leftsample},
      AspectRatio → 1, Frame → False, FrameTicks → None,
      Background → GrayLevel[0.0],
      PlotRange → {{-0.2`, 1.2`}, {-0.2`, 1.2`}}];
    Pause[.25]; flash = blank; Pause[.25];

    rightsample = Table[RandomReal[{0, 1}, 2], {rightnumsample}];
    flash = Graphics[{PointSize[dotsize], Red, Point /@ rightsample},
      AspectRatio → 1, Frame → False, FrameTicks → None,
      Background → GrayLevel[0.`],
      PlotRange → {{-0.2`, 1.2`}, {-0.2`, 1.2`}}];
    Pause[.25]; flash = blank;

    myanswer = ChoiceDialog["Did the signal (high density) appear on",
      {"First?" → 1, "Second?" → 0}, WindowSize → {300, 80},
      WindowMargins → {{Automatic, 0}, {Automatic, 330}}];

    If[myanswer == whichflash, WasICorrect = 1, WasICorrect = 0];
    idealanswer = If[leftnumsample > rightnumsample, 1, 0];
    If[idealanswer == whichflash, WasIdealCorrect = 1, WasIdealCorrect = 0];
    data = Append[data, {WasICorrect, WasIdealCorrect}], {numtrials}];
 ]
```

■ Execute a trial

Now, randomly turn the switch to "high" or "low", draw a sample, and then input your response (1 for "high" and 0 for "low). Execute the next cell 100 times.

```
twoflashes
```

■ Display the data

```
data // MatrixForm
```

Was I Correct?	Was Ideal Correct?
1	1
0	0
0	1
1	1
1	1
1	1
0	1
1	1
1	1
0	1

■ Analyze the data

Let's drop the table heading stored in row 1, and then transpose the matrix so that the columns become the rows:

```
data2 = Transpose[Drop[data, 1]]
```

```
{{1, 0, 0, 1, 1, 1, 0, 1, 1, 0}, {1, 0, 1, 1, 1, 1, 1, 1, 1, 1}}
```

Let's use a combination of Map[ ] and Count[ ] (used earlier to make histograms) to count up all occurrences of an event type. So the total for myhits is:

```
myproportioncorrect =
  N[Map[Count[data2[[1]], #] &, {1}] / Dimensions[data2][[2]]][[1]];
idealproportioncorrect =
  N[Map[Count[data2[[2]], #] &, {1}] / Dimensions[data2][[2]]][[1]];

mydprime = dprime[myproportioncorrect];
idealdprime = dprime[idealproportioncorrect];
mystatisticalefficiency = Round[100 * (mydprime / idealdprime) ^ 2];
```

```
Print[
  Style[
   Grid[{{"my prop correct", "ideal's prop correct", "my d'",
      "ideal's d'", "my efficiency (%)"},
     {myproportioncorrect, idealproportioncorrect, mydprime,
      idealdprime, mystatisticalefficiency}}, Frame → All], 9]];
```

my prop correct	ideal's prop correct	my d'	ideal's d'	my efficiency (%)
0.6	0.9	0.358287	1.81239	4

To get a reliable estimate, you need at least 100 or more trials.

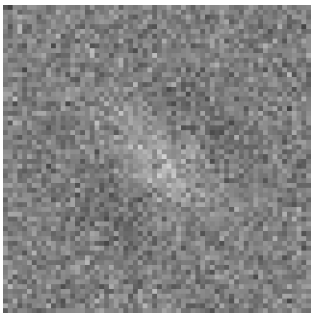
■ Adaptive procedures for finding thresholds using 2AFC or yes/no

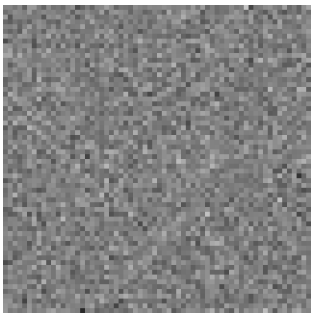
What can you do if the human psychophysical observer is making lots of mistakes or alternatively getting all the trials right? The above method is called the "method of constant stimuli", because (although the stimuli really aren't constant), the conditions (highmean and lowmean) are. Adaptive or tracking methods are more efficient. The idea is to have a computer program automatically hunt for that threshold (e.g. highmean is adjusted) so that the observer is getting a prescribed proportion correct (e.g. 75%). There have been a number of advances in the art of efficiently finding values of a signal which produce a certain percent correct in a psychophysical task such as 2AFC. For more on this, see the QUEST procedure of Watson and Pelli (1983) and the analyses of Treutwein (1993).

Next time

Probability overview

From dots to image intensity patterns: What does the eye see best?





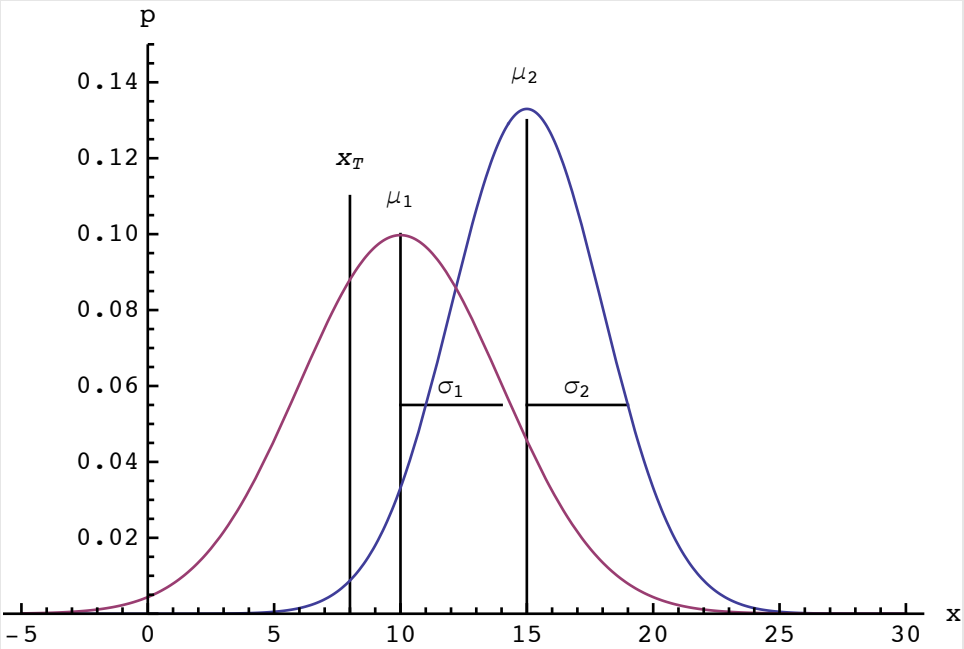
- **Computing the ideal observer for patterns**
- **Comparing psychophysical performance for pattern detection with properties of visual neurons in the brain**

Appendix

Figure code

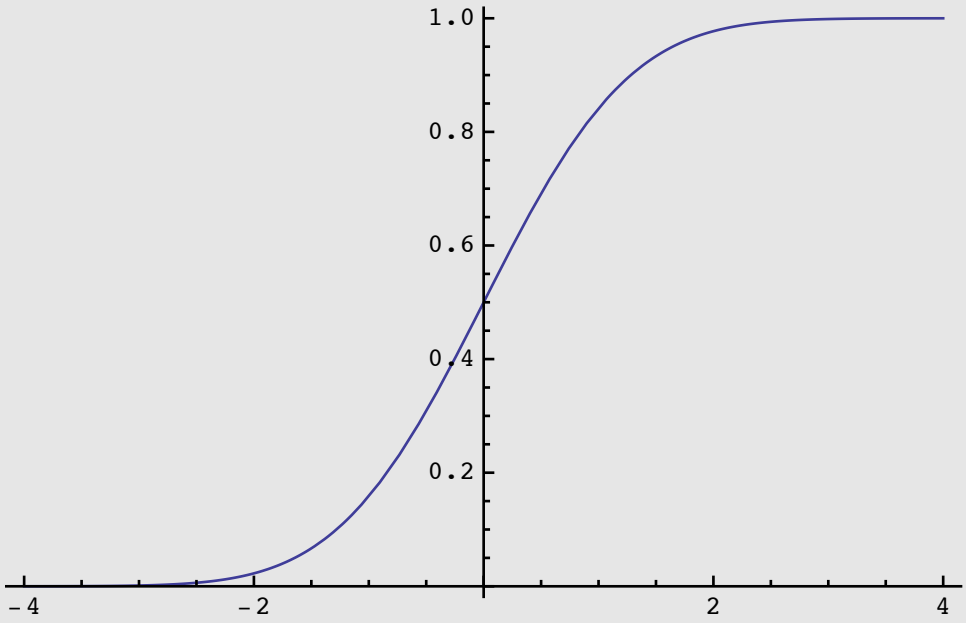
```
gauss[x_, μ_, σ_] := PDF[NormalDistribution[μ, σ], x]
```

```
b = 15; d = 10; sb = 3; sd = 4;
p1 = Plot[{gauss[x, b, sb], gauss[x, d, sd]}, {x, -5, 30},
  Background → GrayLevel[1], AxesLabel → {"x", "p"}, PlotRange → {0, 0.15`},
  Prolog → {Text["\!\(\*SubscriptBox[\(μ\), \(1\)]\)", {d, 0.11`}],
    Text["\!\(\*SubscriptBox[\(μ\), \(2\)]\)", {b, 0.143`},
     {0.1`, 0.1`}], Text["\!\(\*SubscriptBox[\(σ\), \(1\)]\)",
     {d + 2, 0.06`}], Text["\!\(\*SubscriptBox[\(σ\), \(2\)]\)",
     {b + 2, 0.06`}], Line[{{b, 0.055`}, {b + 4, 0.055`}}],
    Line[{{d, 0.055`}, {d + 4, 0.055`}}], Line[{{d, 0}, {d, 0.1`}}],
    Line[{{d - 2, 0}, {d - 2, 0.11`}}],
    Text["\!\(\*SubscriptBox[\(x\), \(T\)]\)", {d - 2, 0.12`}],
    Line[{{b, 0}, {b, 0.13`}}]}]
```

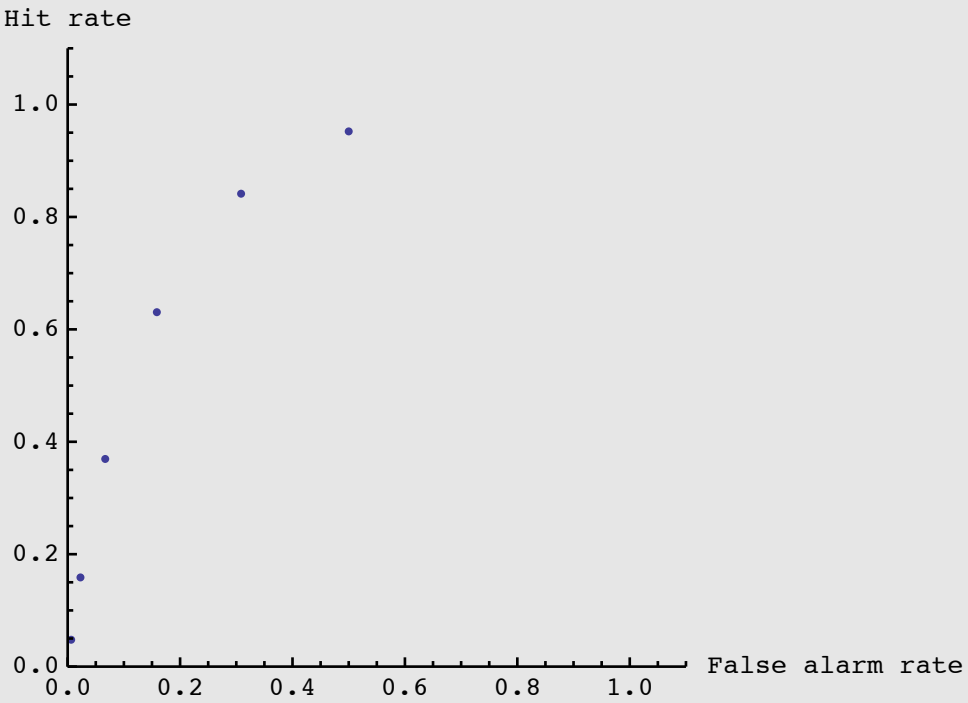



```
z[p_] := Sqrt[2] InverseErf[1 - 2 p];
```

```
cumulgauss[x_, μ_, σ_] := CDF[NormalDistribution[μ, σ], x];
Plot[cumulgauss[x, 0, 1], {x, -4, 4}]
```



```
hitrate = N[Table[1 - cumulgauss[xt, d, sd], {xt, 10, 20, 2}]];
falsealarmrate = N[Table[1 - cumulgauss[xt, b, sb], {xt, 10, 20, 2}]];
ROC = Table[{hitrate[[i]], falsealarmrate[[i]]}, {i, Length[hitrate]}];
ListPlot[ROC, PlotRange → {{0, 1.1`}, {0, 1.1`}},
 AxesLabel → {"False alarm rate", "Hit rate"}, AspectRatio → 1,
 Prolog → AbsolutePointSize[5]]
```




```
dg = ListPlot[z[ROC], AxesLabel → {"Z[False alarm rate]", "Z[Hit rate]"},
   AspectRatio → 1, PlotRange → {{-1, 4}, {-3, 5}},
   Prolog → {AbsolutePointSize[5], Text["x intercept=(μ2 - μ1)/σ2", {2, 4.0}],
     Text["Slope=σ1/σ2", {3.2, 1}]}];
```

```
z[ROC]
```

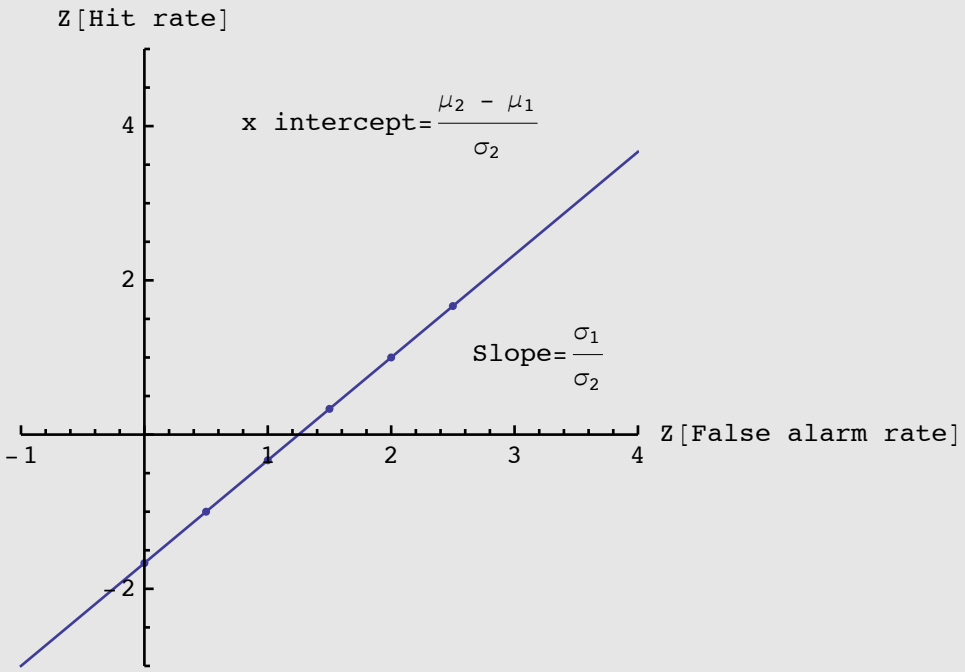
```
{{0., -1.66667}, {0.5, -1.}, {1., -0.333333},
 {1.5, 0.333333}, {2., 1.}, {2.5, 1.66667}}
```

```
Fit[z[ROC], {1, x}, x]
```

```
-1.66667 + 1.33333 x
```

```
fg = Plot[Evaluate[Fit[z[ROC], {1, x}, x]], {x, -1, 4},
   PlotRange → {{-1, 4}, {-3, 5}}];
```

```
Show[dg, fg]
```



Probability and statistical sampling

How to use *Mathematica* to generate probability distributions, cumulative distributions and do statistical sampling.

Distributions and densities

The probability of x photons being detected by an ideal detector is given by the Poisson distribution:

```
poisson[x_,a_] := Exp[-a] a^x / Factorial[x];
```

where **a** is the mean.

Exercise: What is the probability of detecting 12 photons if the mean is 10? Generate more than 20 values and then use the function **Apply[Plus,t1]**, to demonstrate that the sum over all values is 1.

Whenever it is convenient in this course, we will make use of predefined functions in *Mathematica*.

Let's define a Poisson distribution with a mean of 20:

```
pdist = PoissonDistribution[20];
```

The probability distribution function (PDF is given by:

```
PDF[pdist,x]
```

$$\frac{20^x}{e^{20}\, x!}$$

You can obtain the mean, variance and standard deviation of the distribution we've defined. Try it:

```
Mean[pdist]
Variance[pdist]
StandardDeviation[pdist]
```

```
20
```

```
20
```

$$2\sqrt{5}$$

The output shows *Mathematica*'s definition of the function. The "If"'s test to make sure that x is not negative and is an integer. The rest of the definition should look familiar. In this lecture, we make use of the fact that the continuous normal density can provide a good approximation to the Poisson distribution when the mean is large enough and if we set the standard deviation to **Sqrt[20]**:

```
ndist = NormalDistribution[20,4.47214];
```

```
N[PDF[pdist,28]]
N[PDF[ndist,28]]
```

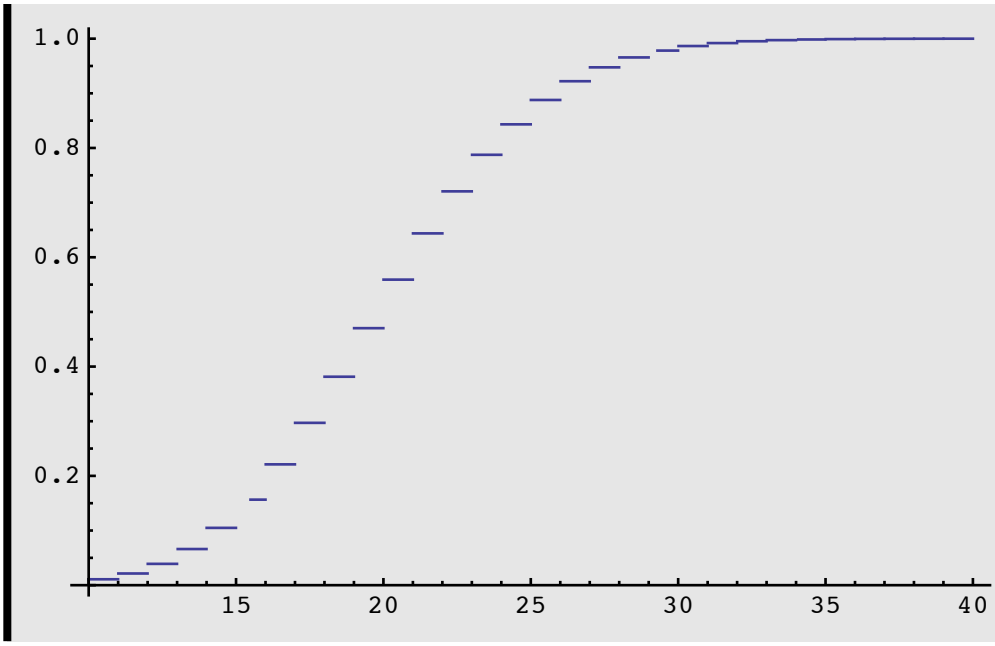
```
0.0181472
```

```
0.0180105
```

The cumulative distribution

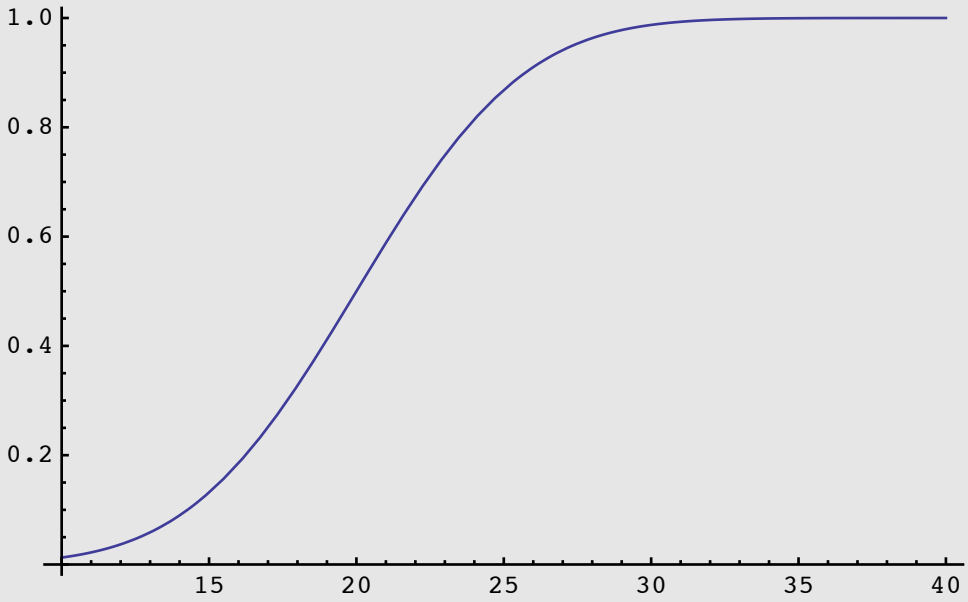
The cumulative distribution gives the probability that the detector signals x<k photons. It is obtained by adding up the probabilities for all values less than k. For the cumulative density function, we integrate over all values less than k. Here is the cumulative distribution for the discrete Poisson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

Treutwein, B. (1993). Adaptive psychophysical procedures: A review. submitted to Vision Research, (email: bernhard@tango.imp.med.uni-muenchen.de)

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. (See: http://vision.arc.nasa.gov/mathematica/psychophysica/)

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