Signal Detection Theory

1. Background

Noisy perception

Human perception is not perfect: Sometimes you miss stuff that was there (e.g. you failed to spot a friend that was standing right in front of you). At other times you perceive stuff that was not there (e.g. you could swear you heard your phone ringing when eagerly awaiting that call). Perception is inherently noisy. This is directly visible when you turn off the light, close your eyes, and lay your hands over your eye lids. Everything should be pitch-black right now, but instead you're probably seeing a colourful array of bright dots and blobs. This is noise occurring inside the visual system, caused by spontaneous firing of neurons. In addition to this "internal noise", perception is further hindered by external noise. You may have had the experience of disappointment when finally the bus appears on the horizon, and a moment later it turns out to be a lorry. Your detection of relevant objects (buses) is muddled by irrelevant signals (lorries) – in other words: noise.

Noise becomes more serious in many other real life situations: The radiologist is looking for tumors in noisy X-ray images, the baggage screener is looking for weapons and explosives in cluttered images of bags, and the battlefield operator is looking for enemy activity in complex sattelite, radar, or sonar images. All these stimuli can be thought of as signals that need to be differentiated from noise that is present in the environment as well as in the observer, and that may lead to wrong decisions. Also in many laboratory experiments the observer's task is to detect a stimulus (e.g. the presence of a light, a color, a shape, a tone, or a smell). For science to determine how well sensory mechanisms perform, it needs tools to determine how well such systems distinguish signals from noise.

In 1954 Tanner and Swets formulated exactly such a tool: *Signal Detection Theory* (SDT). After a war-struck world, SDT was developed to quantify the quality of reception of radio signals under noisy circumstances, but it turned out a useful tool in many other areas of technology and science. The strength of SDT strength is its ability to quantify how well a signal can be distinguished from noise by a "receiver", or, in our case, an observer. Moreover, it is able to distinguish the *perceptual sensitivity* for signals from any *biases* an observer might have in responding to such signals.

2. SDT Basic concepts

Sensitivity

Imagine you're a baggage screener looking for explosives in baggage at the airport (Figure 1). In the beginning, as a trainee, you're not that good at distinguishing explosives from toiletries and the like. With training you get better at picking out the dangers and ignoring the rest. In other words, with training you get more *sensitive* to the relevant signals, and learn to ignore the irrelevant signals (noise). The bags themselves have not changed. What has changed is the *internal perceptual signal* generated in your brain: This has become stronger, allowing you to make better decisions. The strength of this internal signal is what we refer to as the *sensitivity*, and it's denoted by the symbol *d'* (called d-prime; think of the "d" of *d*etectability, *d*istinguishability, or *d*iscriminability, but also *d*istance between signal and noise). The more sensitive the observer, the larger *d'* is. Íf *d'* = 0 then the observer cannot distinguish signals from noise at all, and performs at chance. Later we will see how you can compute *d'*.

Training is one way to increase your sensitivity as a baggage screener, as you improve internal perceptual processes. Another way is to strengthen the external signal. For example, we can use software that makes all the explosives in the X-ray image orange, and all the rest blue. Now

explosives will really stand out, and your sensitivity will likewise increase. Another way of saying this is that you as an observer will be more sensitive to strong signals than to weak signals. Again, *d'* thus reflects the *perceived* intensity of the signal, which in this case is aided by an increased difference in the external signal.

Figure 1. A baggage screener at work.

Two states, four outcomes: The response matrix

To makes things simple for you as a baggage screener, the world basically consists of two states: Bags with all sorts of normal stuff in them (noise), and bags with dangerous stuff in them, hidden among the normal stuff (signal + noise). The same goes for your responses: You may either classify a bag as dangerous (Yes, an explosive!), or let it pass (No, all fine!). Together, the two states of your world, and the two possible responses lead to four possible outcomes: 1) There is something dangerous in the bag and you say 'yes'. We call this a *hit*. 2) There is nothing dangerous in the bag and you say 'no'. We call this a *correct rejection*. 3) There is something dangerous in the bag and you say 'no'. We call this a *miss*. 4) There nothing dangerous in the bag but you nevertheless say 'yes'. This is what we call a *false alarm*. Both hits and correct rejections are correct responses, while both misses and false alarms are errors.

Below you see Table 1a, containing the two states of the world, the two possible responses, and the four possible outcomes this generates, together with hypothetical values for you as a trainee baggage screener. We call such a table a *response matrix*.

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : 60%	« False Alarm » : 40%
Response "No"	« Miss » : 40%	« Correct Rejection » : 60%

Table 1a. The response matrix, in which performance is expressed as a percentage (0-100%).

It is good to know that the hits and miss percentages are computed over the number of bags with dangerous stuff in them (signals), while the false alarm and correct rejection percentages are computed over the total number of bags without dangerous stuff in them (just noise). For example,

the percentage hits is computed by dividing the number of hits by the total number of bags that contain suspicious stuff (= hits + misses) and multiplying this by 100. However, in scientific use these numbers are often expressed as proportions or rates rather than percentages, thus the hit rate = hits / (hits + misses), which in this case is 0.6. Conversely the correct rejection rate is computed by dividing the correct rejections by the total number of bags that do not contain suspicious stuff (correct reject rate = correct rejections / (false alarms + correct rejections) = 0.6. For ease of reading, we will mainly use percentages in the text, but use proportions in the tables and calculations. This then results in Table 1b.

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : .60	« False Alarm » : .40
Response "No"	« Miss » : .40	« Correct Rejection » : .60

Table 1b. The same response matrix, but now performance is expressed as proportions (or rates, 0-1).

Note that you're not that good: You miss a large proportion, 40%, of the dangerous items. At the same time you pick out 40% of the bags that are just fine, which will annoy the passengers. With training this is likely to improve, and after a few days on the job it looks more like Table 2:

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : .75	« False Alarm » : .25
Response "No"	« Miss » : .25	« Correct Rejection » : .75

Table 2a. The response matrix after training.

Now let's also install new image-enhancing software, and it becomes as in Table 2b:

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : .90	« False Alarm » : .10
Response "No"	« Miss » : .10	« Correct Rejection » : .90

Table 2b. The response matrix after training plus new image software.

As you make more and more correct responses (now 90% instead of the 60% at the start), your sensitivity is obviously increasing! Note that so far in these examples, the number of hits increases with the same percentage as the number of correct rejections increases, but this does not have to be the case. It might also be that you get better at detecting targets in the dangerous stuff (e.g. the hit rate goes from 75% to 90%), but that you do not get any better at rejecting bags that do not contain dangerous stuff (e.g. your correct rejection rate might stay at 75%). We'll now turn to such asymmetric response matrices.

Criterion and response bias

Although you improved a lot through training, we might still argue that you're not doing a very good job. Even though your hit rate is 90%, this still comes with 10% misses. In other words, of every 10 dangerous items, you let one pass. Obviously this needs to be remedied, as one explosive is enough to cause a disaster! Apart from firing you, one solution would be to keep on training, hoping that you can improve even further. But for the time being, you might also go for a different solution: Adjust your decisions as to err on the side of caution. Note that currently you score as many hits as you score correct rejections, and you make as many misses as false alarms. In other words, you try to do your job rather neutrally, or objectively, giving "Yes" and "No" equal chances.

But in case of baggage screening it would not hurt if you lowered your threshold for calling a bag dangerous – and thus increased your "Yes, dangerous!" responses a bit more – so that you're more likely to catch the dangerous bags. The response matrix would then look something like Table 3:

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : .999	« False Alarm » : .45
Response "No"	« Miss » : .001	« Correct Rejection » : .55

Table 3. The response matrix after lowering the criterion for what counts as dangerous.

Note that now you miss very few suspicious bags, and you increased you hit rate accordingly. (Note that these are complementary: Since both are calculated of the total number of signals, the hit and miss rates should always add up to 1, or 100%). But this comes at a cost. Since you decided to pick out more bags, the false alarm rate has also gone up (and its complement, the correct rejection rate, has gone down – these also add up to 1, or 100%, as both are calculated over the number of noise events). In contrast to when a neutral criterion has been adopted, the response matrix has now become asymmetrical, heavy on the "yes" responses. This is because the only reason for the higher hit rate is your decision to lower your threshold, or *criterion*, for what counts as a suspicious bag. You have not become more sensitive, you have become more liberal in your response tendency.

In SDT, the criterion for when perceptual evidence counts as a signal is referred to as *c*. Adopting a *liberal* criterion, or low threshold, means that the observer accepts relatively little perceptual evidence (weak signals) for making a positive response, leading to more hits, but also more false alarms. In contrast, when an observer adopts a *conservative* criterion, he or she requires relatively strong perceptual evidence before he or she hits the "Yes!" button. This leads to fewer hits, but also to fewer false alarms, and more correct rejections, together with more misses. For example, Table 4 might be a response matrix for someone in the highest command of state defence, and who has the job of deciding on whether to initiate a nuclear strike (yes), or withhold such a strike (no) in response to enemy activity. You do not want this person to be too trigger-happy, and now correct rejections (i.e. avoiding false alarms) are more important than hits.

	Enemy prepares nuclear strike (signal)	**Normal enemy activity (noise)**
Response "Strike"	« Hit » : 55%	« False Alarm » : 0.1%
Response "No strike"	« Miss » : 45%	« Correct Rejection » : 99.9%

Table 4. A response matrix for a more conservative decision criterion.

Another term that is often used for the criterion is *bias*, and in SDT the symbol for this is β. As we will see later, β is calculated somewhat differently than c, but they essentially express the very same thing, namely to what extent people adopt a decision criterion that is not neutral. With a *liberal criterion*, people are biased towards saying yes, whereas with a *conservative criterion,* they are biased towards saying no. A neutral criterion means no bias. Either way, it is important to realize that biases (criterion shifts) are *independent* of sensitivity. A bias means that responses are shifted *either* towards hits *or* towards correct rejections. Increased sensitivity on the other hand results in increases in *both* hits and correct rejections.

Pay-off matrix

It is important to realize that biases occur more often than not. It is a mistake to assume that observers will be neutral, whether in your experiment or in real life situations. Very often it is in people's own interest to be biased. For example, you as an experimenter may instruct the participant to "detect as many lights as possible". The participant, eager to impress you, will go for as many

hits as possible, and not care about false alarms as much as you perhaps want them to. In real life, it is often not only in the person's self-interest, but also in society's interest if observers are biased, whether your doctor, your baggage screener, or your chief of command in state defence. Biases are typically caused by the *gains* and *losses* that are associated with the four possible outcomes. For example, in case of a baggage screener, a miss can come with a very high price, and thus misses should be avoided virtually "at all cost". In contrast, in the case of the chief of command who is about to launch a nuclear missile, false alarms come with a very high price, and thus false alarms should be avoided. Such gains and losses can be laid out in what is called the *pay-off matrix*. It's basically the same as the response matrix, but now it contains the gains and losses associated with each outcome. For the baggage screener we can put some hypothetical price tags on hits, misses, false alarms, and correct rejections, as in Table 5a.

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : nil	« False Alarm » : Costs a little bit of time each time, plus irritation with the traveler, and thus money (extra personnel, extra security lanes). Price tag: **$4**
Response "No"	« Miss » : Costs many lives and millions of dollars Price tag: **$20M**	« Correct Rejection » : nil

Table 5a. Hypothetical pay-off matrix for baggage screening.

Note that both misses and false alarms come with costs, but the costs of a miss presumably outweigh the costs of a false alarm, leading to a liberal criterion. However, it is always important to realize that the costs may be perceived quite differently for the baggage screener: For them, on an off day, the costs of having yet another irritated customer in front of them may be higher than the way we as society see it. So it is always good to think of the perceived costs rather than actual costs in these types of situation.

The pay-off matrix may not only contain costs, but also benefits. Imagine a manager at the airport, responsible for security. She wants to promote a better safe than sorry attitude in her staff, and she decides to implement a bonus scheme. For every hit (i.e. for every dangerous item found in a bag), baggage screeners will receive an extra $2 on their pay slip. Now the pay-off matrix will look like Table 5b:

	Dangerous stuff (signal)	**Only normal stuff (noise)**
Response "Yes"	« Hit » : **$2**	« False Alarm » : **$4**
Response "No"	« Miss » : **$20M**	« Correct Rejection » : nil

Table 5b. Hypothetical pay-off matrix for baggage screening with a bonus on hits.

As you can imagine, responding "Yes" becomes even more attractive, and indeed, the manager sees that the number of hits goes up. Inevitably though, the false alarms also go up, as this is simply the consequence of people saying yes more often. So the manager tries to think of a better way. Rather than trying to merely increase yes responses, she reasons, I should try to increase *correct* responses. If I boost both hits and correct rejections, she thinks, then both misses and false alarms will go down. So she decides to reward not only hits, but now also correct rejections, say with $0.50 each, as in Table 5c. Now, was this a wise thing to do?

	Dangerous stuff (signal)	Only normal stuff (noise)
Response "Yes"	« Hit » : **$2**	« False Alarm » : **$4**
Response "No"	« Miss » : **$20M**	« Correct Rejection » : **$0.50**

Table 5c. Hypothetical pay-off matrix for baggage screening with a bonus on hits and a (smaller) bonus on correct rejections.

Likelihood of occurrence

At first sight, it seems clever: The manager is trying to boost sensitivity, while maintaining a bias towards caution (as hits deliver $2, while correct rejections only deliver $0.50). However, the manager has failed to take into account another factor that contributes to the pay-off matrix, and that is the *likelihood* of a particular event occurring. It is a fact that bombs, guns etc. are only present in an extremely small minority of bags. Most bags are simply ok. Say (hypothetically!) that normal bags make up 99%, of all bags, with the remaining 1% being dangerous. Thus, every time a baggage screener sees a bag pass by, there is a 99% chance that it is ok. In other words, there is a 99% chance that saying "no" is the correct response (correct rejection). If he is concerned about pay, then this should be taken into account. If he calls it a yes, there is a 1% likelihood of earning $2. But if he calls it a no, there is a 99% likelihood of earning $0.50! Taking these probabilities into account, the pay-off matrix actually looks like Table 6:

	Dangerous stuff (signal)	Only normal stuff (noise)
Response "Yes"	« Hit » : 0.01 * $2 = **$0.02**	« False Alarm » : **$4**
Response "No"	« Miss » : **$2M**	« Correct Rejection » : 0.99 * $0.50 = **$0.495**

Table 6. Hypothetical pay-off matrix for baggage screening which takes probability of occurrence (Expected Value) into account.

What you see here are so-called *expected values* (EVs). The expected value not only takes the nominal amount into account, but also the *likelihood* of that amount being awarded. You can see that by implementing the new payment scheme, the manager has made responding "No" a lot more attractive! So, when constructing a pay-off matrix, also take the probability of events into account.

Taking probabilities into account is very important also in other areas. An often seen example is that of a young woman coming to the doctor with what seems to be a little lump in one of her breasts. The doctor has to decide whether to operate or not. Many people would argue that the doctor better be safe than sorry: Operate! However, first there are the costs of surgery – for society, but of course also for the woman. But second, and perhaps most important here, the likelihood of a young woman having breast cancer (the "prevalence") is extremely small. The "better safe than sorry" strategy here would thus be to follow a more conservative treatment path.

Ideal observer

The optimal decision strategy is thus determined by the gains and losses associated with the different outcomes, which in turn are determined by the probability of occurrence and the nominal value. An observer that adopts the optimal strategy is called an *ideal observer*. Thus, an ideal observer is not necessarily a neutral, or "objective" observer. Rather, it is an observer that optimally takes probabilities, costs and benefits into account when making a decision.

3. Quantifying sensitivity and bias

By now you should have a conceptual understanding of sensitivity and bias (or criterion), and which factors affects them. The strength of SDT is that it allows us to quantify the sensitivity and bias independent of each other. This allows us to objectively compare different observers or different working conditions. The quantification is relatively simple, but it does require basic statistical knowledge.

Step 1: The noise distribution

Below in Figure 2 is a graphical depiction of some noise. It could represent the noise you hear between stations on an old radio, the noise you see on a TV that has not been tuned yet, or the noisy firing of neurons in your brain.

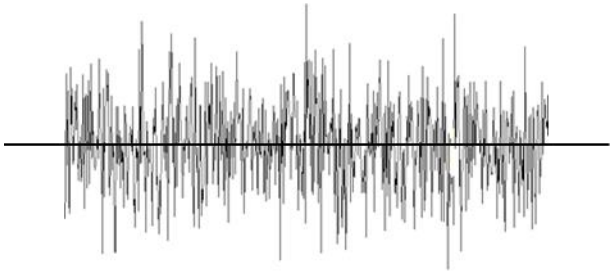

Figure 2. Some noise hovering around a baseline.

As you can see, the noise largely hovers around some average baseline value (the black line), spreading to either side. Occasionally, but less likely, the noise reaches more extreme values, as reflected in the occasional peaks and troughs you see in the pattern, resulting in momentary louder hisses and brighter spots in your perception (or silence and darkness for the lower extremes). Statistically speaking, the noise follows a certain *probability distribution*, depicted in Figure 3:

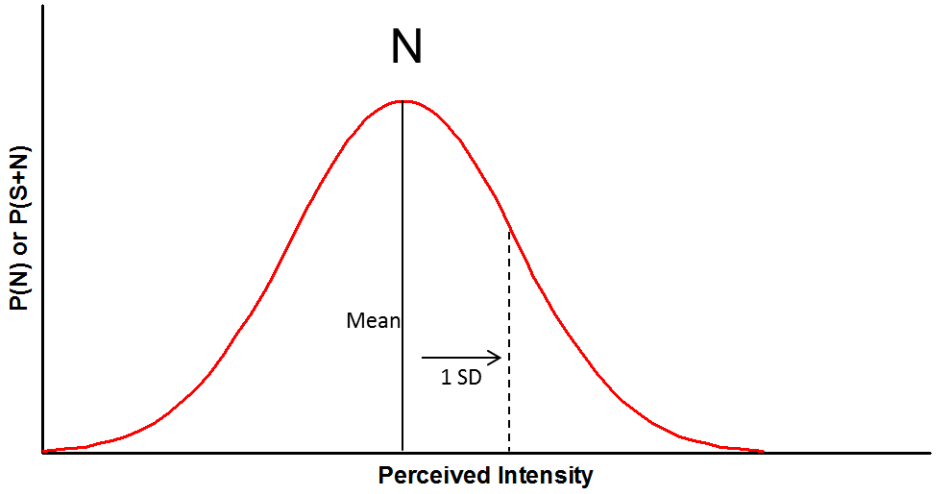

Figure 3. The probability distribution of noise values.

Here the N stands for Noise. On the x-axis is the perceived intensity of the noise, while the y-axis represents the probability of such intensity occurring. The solid vertical line represents the mean intensity (or baseline level). You can see that values around the mean have the highest probability of occurring, whereas values towards the extremes are increasingly less likely. So given many noise events, most of them are likely to occur in the middle (producing a perceived intensity around baseline), whereas only some events will be perceived extremely strongly or extremely weakly.

You may remember from statistics class that bell-shaped probability distributions like this are called *normal* distributions (also called Gaussian distributions), and that almost everything we observe around us follows this distribution (think about the distribution of how tall people are: some people are very small, some people are very tall, but most people will be around average). What we will use here is the *standard normal distribution*, in which the mean is conveniently put at 0. Moreover, the standard normal distribution provides a convenient measure for how extreme a value is: Deviation from the mean is expressed in standard deviations (SD). The more SDs away from the mean, the more extreme a value is. The distance in SDs is called the *z*-score. A *z*-score of 2.5 means that a value is 2.5 SDs away from the mean. The higher the z-score, the lower the likelihood of occurrence.

Step 2: The noise + signal distribution

Now imagine we add a signal to the noise. As the signal is added, average intensity will increase, as shown in Figure 4. In terms of the probability distributions, this looks like Figure 5.

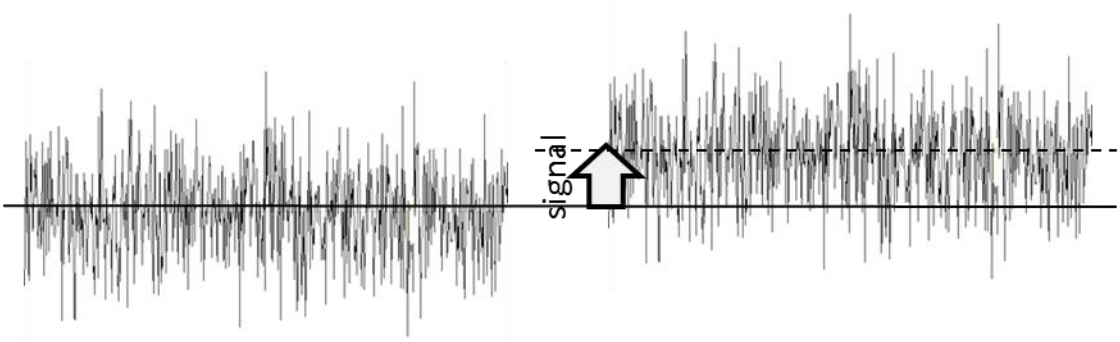

Figure 4. With adding a signal to the noise, overall intensity will increase.

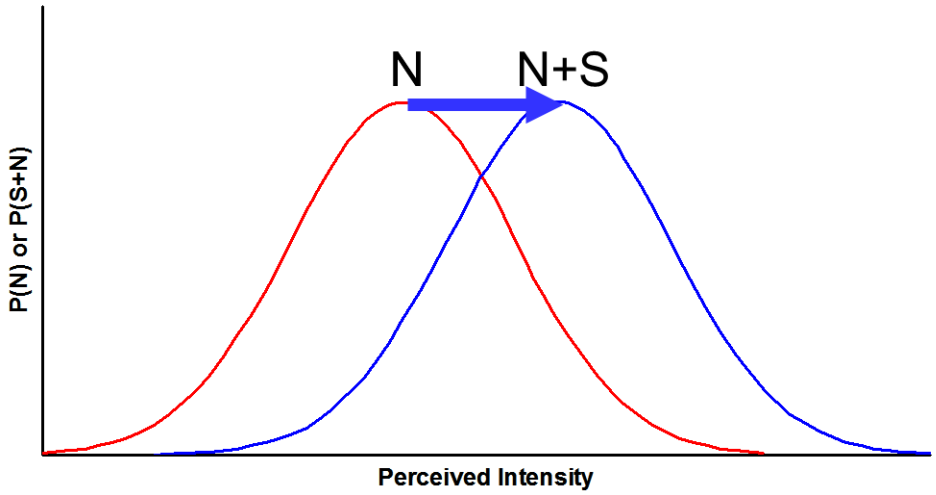

Figure 5. The probability distributions of noise values (N) and noise + signal values (N+S).

As you can see, the noise, and noise + signal follow similar distributions, but shifted in terms of mean perceived intensity. Yet, at the same time, there is also considerable overlap here between the distributions. There is thus *uncertainty*: A moderately strong percept could be a signal, but there is also a fair chance that it belongs to noise. Given the uncertainty, and given that the observer needs to make a decision, we need some criterion. For perceived intensities stronger than the criterion, the observer then decides "yes, signal present". For perceived intensities weaker than the criterion, the observer then decides "no, signal absent, just noise". In Figure 6 you see the two distributions again, together with a criterion that has been placed in the form of a vertical line. In SDT, the criterion,

also called threshold or bias, is denoted with a *c*, or β. The vertical line represents a particular perceived intensity (i.e. the point on the x-axis). Everything above this perceived intensity is taken as a signal, and everything below it is taken as noise. Figure 6 shows a *neutral* criterion: it lies exactly in between the noise and noise+signal distributions.

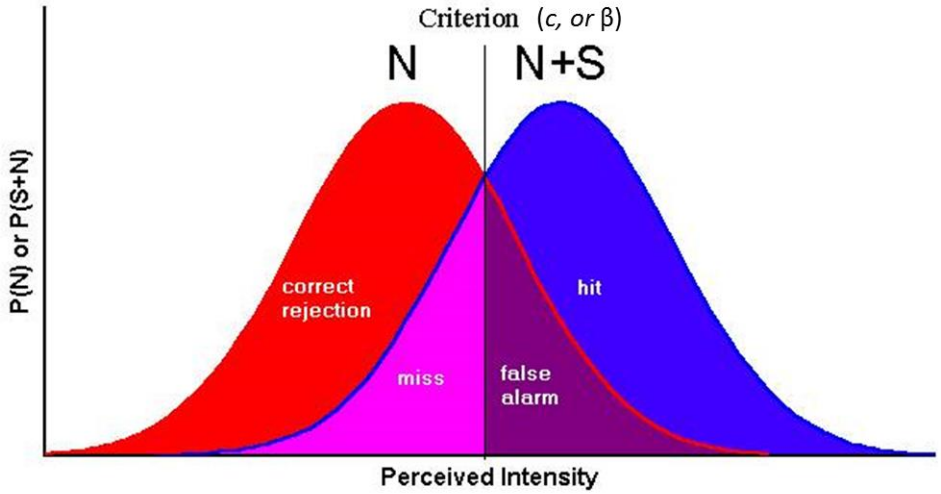

Figure 6. The noise and noise+signal distributions, together with a neutral criterion (the vertical line).

The different areas under the curve represent the probabilities of the various outcomes in the response matrix for this criterion. On the right side, you see the blue area, which represents the proportion of *hits* given that the signal exceeds the criterion. It is thus part of the noise+signal distribution, and it is good to know that this blue area therefore extends *behind* the dark purple area representing the *false alarm* rate. This latter area represents the proportion of false alarms given that it is noise which exceeds the criterion, and it is thus part of the noise distribution. On the left side, you see the red area, which represents the proportion of *correct rejections* given that there is no signal. It is therefore part of the noise distribution and extends behind the magenta area representing the *misses*. Note that the misses belong to the noise+signal distribution, as this is the proportion of responding "no" despite the fact that there was a signal. However, in these cases the signal is too weak and does not exceed threshold, resulting in a miss.

Now imagine we increase the strength of the signal. In other words, we increase the distance between the noise and the noise + signal. In terms of the distributions, this looks like Figure 7.

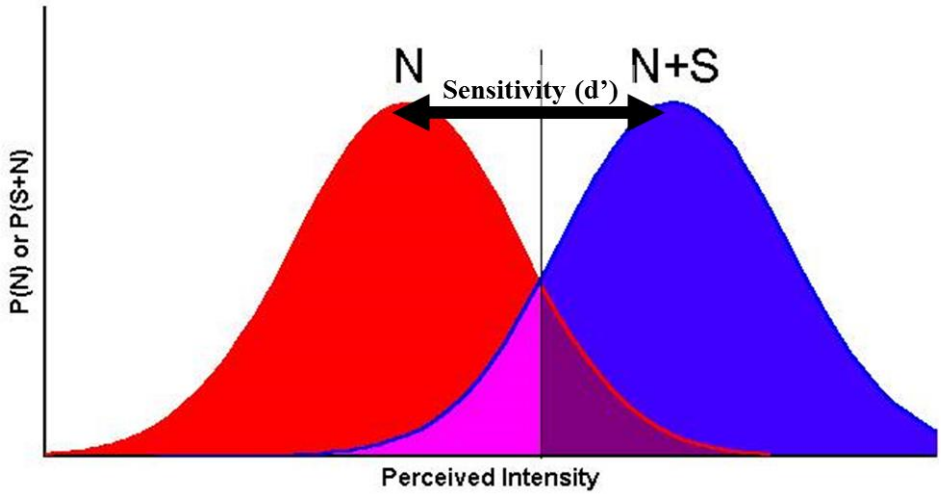

Figure 7. For a stronger signal (or higher sensitivity) the distance between the noise and noise+signal distributions increases.

Note how the larger distance between the two distributions leads to less overlap, and thus fewer errors. After all, the magenta and purple areas represent the misses and false alarms. In other words, the observer is more sensitive to this stronger signal. This sensitivity, *d'*, is simply the distance between the noise and noise+signal distributions, and is expressed in *z*-scores (which, remember, are standard deviations). The areas under the curves thus represent the expected performance of an observer (how many hits, correct rejections, false alarms and misses he or she will make) given a certain sensitivity and criterion. Conversely, this means that if we have obtained the proportions of these types of responses (as we can do by measuring the observer), we should be able to estimate the probability distributions and the distance between them.

Step 3: Computing d'

Luckily, these z-scores can be straightforwardly computed from the proportion of responses that people give, Specifically, we use the proportion of hits, p(hits), which is a direct indication of how good people are when the signal is there, and the proportion false alarms, p(false alarms), which is a direct indication of how good people are when the signal is *not* there. This leads to the following equation:

$$d' = z[p(hits)] - z[p(false\ alarms)] \qquad (1)$$

We could also take the correct rejections and the misses into account, but remember that these are complementary to the hits and false alarms, in that p(misses) = 1-p(hits), and p(correct rejections) = 1-p(false alarms). Adding these measures in one way or another would therefore not add any additional information to the computation of d', and using the hits and false alarms is thus sufficient.

It follows from Equation 1 that the larger the proportion hits, and the smaller the proportion false alarms, the larger d' will be. Note that to compute d', p(hits) and p(false alarms) need to be converted to z-scores. Such conversions can be looked up in z tables that can be found on-line or standard statistics text books. There are also on-line tools that will compute the z-score from a proportion straightaway. Or you can compute them using software such as Matlab and Excel. In Matlab the function is z = norminv(p), so *d'* = norminv(p[hits]) – norminv(p[false alarms]). In Excel it is z = norm.s.inv(p), so *d'* = norm.s.inv(p[hits]) – norm.s.inv(p[false alarms]).

So now you know what sensitivity means, how to compute it, and in what unit it is expressed.

Step 4: Computing the criterion, c or β

Finally, let's turn to the criterion. We have seen that a *neutral* criterion lies in the middle, as one would expect. A *liberal* criterion means that people accept weaker evidence for their "yes" response. In other words, they lower their threshold. In our graph this means that the criterion is shifted to the left, as in Figure 8.

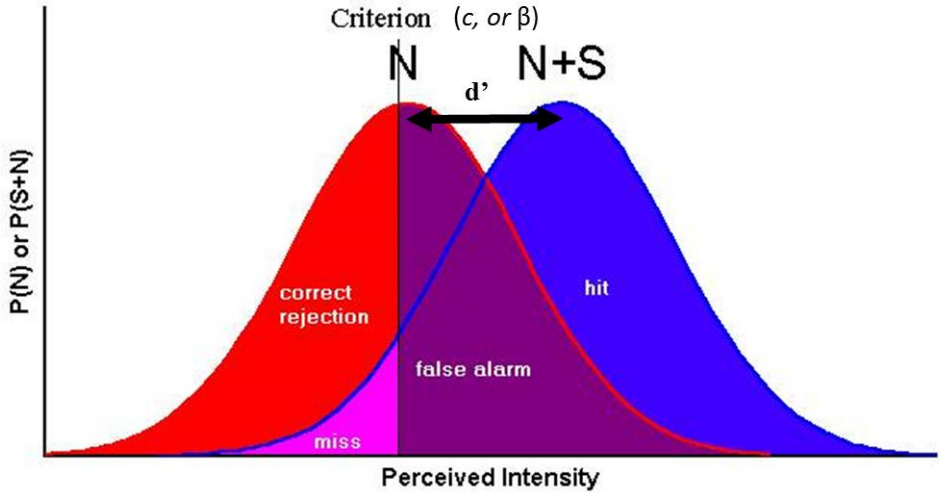

Figure 8. The same distributions, but now with a more liberal criterion (lower threshold).

Note how a lower criterion leads to more hits (the blue area, which extends behind the false alarms, has grown), and fewer misses, but it also leads to substantially more false alarms here. Likewise, observers can adopt a more *conservative* criterion, in which case they demand stronger perceptual evidence before they decide "yes". The criterion then shifts towards the right, as in Figure 9.

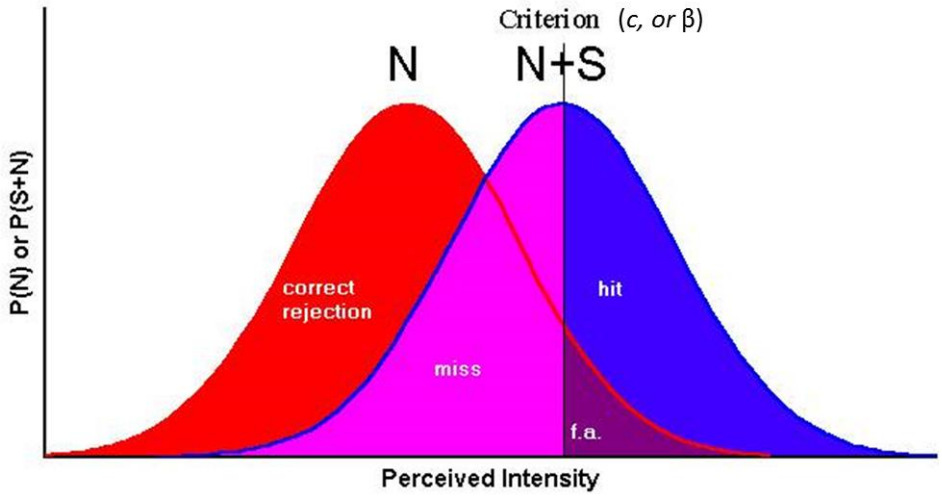

Figure 9. The same distributions, but now with a more conservative criterion (higher threshold).

This results in more correct rejections (again, the red area extends behind the misses) and fewer false alarms. But as you can see, it also results in more misses, and fewer hits. This is inevitable. Note once more that these criterion shifts work independently from sensitivity (*d'*), as the distance between the distributions does not change with criterion shifts.

One formula for computing the criterion, *c* is:

$$c = -(z[p(hits)] + z[p(false\ alarms)])/2 \qquad (2)$$

Here too the formula simply uses the proportion of hits and the proportion of false alarms that the observer scores. The way to understand this formula is by realizing that the more liberal an observer is, the more hits and the more false alarms her will make. Thus the term within brackets (adding hits and false alarms) will increase. This is then preceded by a minus, in order to make liberal correspond to a leftwards shift or lower thresholds (and vice versa, a conservative threshold to a rightwards shift or higher threshold). This then leads to:

If $c = 0$ → neutral criterion If $c < 0$ → liberal criterion If $c > 0$ → conservative criterion

Instead of *c*, many investigators use a slightly different measure for the criterion, or bias, namely *β*. The computation of *β* again makes use of p(hits) and p(false alarms), but in a different way. Going back to Figure 7, we see that if we choose a perfectly neutral criterion, right in between the two distributions (N and N+S), the two distributions will be equal in *height* at that point. The height of the distribution represents the non-cumulative probability. If we would divide the height of the N+S distribution by the height of the N distribution, the result should thus be 1. If we'd move the criterion to the right (make it more conservative), the curve of the N+S distribution at the point of criterion would be higher up than the curve of the N distribution, and the result of dividing the height of N+S by the height of N would be greater than 1. This is illustrated in Figure 10, where the circled points indicate where the criterion crosses the curves describing the two distributions.

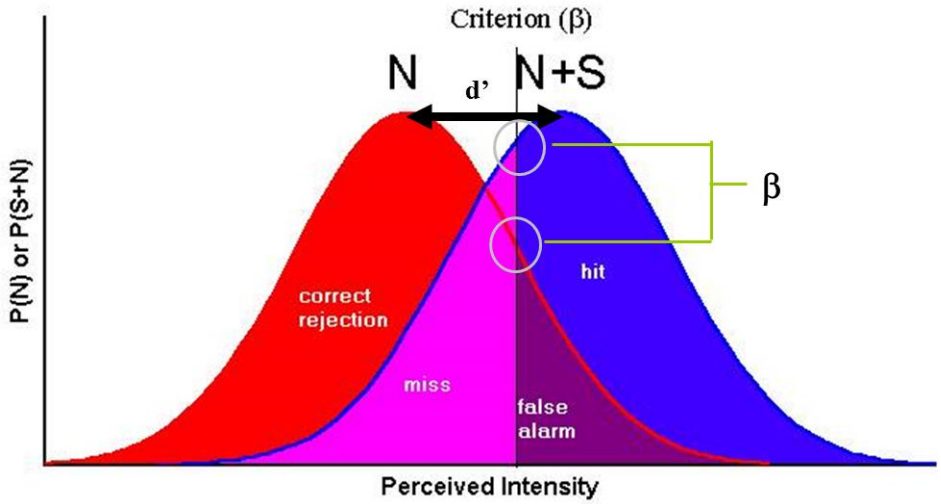

Figure 10. One way to compute the criterion, *β*, is by taking the relative height of the curves at the threshold.

Likewise, if we were to move the criterion to the left (i.e. make it more liberal), then the curve of the N+S distribution would be below that of the N distribution, and the result of dividing the height of N+S by the height of N would be smaller than 1. We can derive the height of the N+S distribution at the criterion point directly from the proportion of *hits* the observer has scored, and the height of the N distribution from the proportion of false alarms, such that:

$$\beta = f[z|p(hits)] \,/\, f[z|p(false\ alarms)] \qquad (3)$$

This equation seems more difficult than it is. The values for the N+S and N curves, *f[z|p(hits)]* and *f[z|p(false alarms)]*, can be looked up in most z-tables. The f(z) represents the non-cumulative probability at that point. Alternatively, software such as Excel and Matlab can be used. All you need to know are the hit and false alarm rates. In Matlab, the function is f(z) = normpdf(z). In Excel it is f(z) = norm.s.dist(z,FALSE). This then leads to:

If $\beta = 1$ → neutral criterion If $\beta < 1$ → liberal criterion *If* $\beta > 1$ → conservative criterion

Example
Table 7 shows the performance of two trainee baggage screeners.

SCREENER A	**Dangerous stuff (signal)**	**Only normal stuff (noise)**
Response "Yes"	« Hit » : 90%	« False Alarm » : 40 %
Response "No"	« Miss » : 10%	« Correct Rejection » : 60%

SCREENER B	**Dangerous stuff (signal)**	**Only normal stuff (noise)**
Response "Yes"	« Hit » : 80%	« False Alarm » : 10%
Response "No"	« Miss » : 20%	« Correct Rejection » : 90%

Table 7. Performance for two baggage screeners (A and B).

We only need the proportions of hits and false alarms (again, misses and correct rejections are complementary). We then look up the z-scores for these proportions in the *z* column of the table provided in the Appendix. Thus, Screener A scores 90% hits and 40% false alarms, and the respective z-scores are 1.28 and -0.25. Then $d' = 1.28 - (-0.25) = 1.53$. For Screener B the scores are 80% hits and 10% false alarms, then $d' = 0.84 - (-1.29) = 2.13$ (you can check this using the z-table in the Appendix). Note that Screener B is more sensitive than Screener A, and thus a better observer, even though B scored fewer hits than A! How is this possible? It's possible, because Screener B scored a lot fewer false alarms. Screener B apparently used a much more conservative criterion than Screener A.

Let's see if we can find evidence for this is when computing the criteria used by these observers. For Screener A, $c = -(1.28 - 0.25) / 2 = -0.26$, which is < 0, so indeed a liberal criterion. For Screener B, $c = -(0.84 - 1.29) / 2 = 0.23$, which is > 0, so indeed a conservative criterion. We can do the same for β, even though this is slightly more complicated. We can use the table in the appendix to look op the function values that come with the z-scores (i.e. *f(z)*, now take the other column). For Screener A, the proportions p(hits) and p(false alarms) are .90 and .40, and we find the values *f(z)* of 0.176 and 0.387 respectively. The criterion is therefore at $\beta = 0.176/0.387 = 0.455$. This value is < 1, so the subject was using a liberal criterion. For Screener B, p(hits) = 0.80, p(false alarms) = 0.10, and the corresponding *f(z)* values are 0.280 and 0.176. The criterion is therefore at $\beta = 0.280/0.176 = 1.6$. This value is >1, so the subject was using a conservative criterion.

The advantage of β over c*: *The Ideal Observer

Computing the criterion as *c* has the advantage of simplicity: You take the same z-scores as the ones you need for computing *d'*. However, computing the criterion as β also has an advantage. It allows you to directly compare the measured criterion of an observer to what would be the *optimal* criterion for that situation. In other words, it allows you to compare your observer to the *ideal observer*. The formula for computing the optimal β is:

$$\beta_{\text{opt}} = \frac{P(\text{N})}{P(\text{S})} \times \frac{V(\text{CR}) + C(\text{FA})}{V(\text{H}) + C(\text{M})} \quad (3)$$

Here P(N) and P(S) are the probabilities of an event being respectively noise or a signal. V(H) and V(CR) are the nominal values of hits and correct rejections, while C(M) and C(FA) are the costs of a miss and a false alarm. So, for example, the higher the chance of a signal, P(S), the lower β_{opt} : The ideal observer becomes more liberal. As another example, the higher the costs associated with a false alarm, C(FA), the higher β_{opt}, : The ideal observer becomes more conservative. You can work this out for all the other terms in the equation too.

4. Visualizing sensitivity and bias: The ROC curve

There is a useful tool for visualizing sensitivity and bias in a single graph. It's called the -called *Receiver Operating Characteristic*, or *ROC curve*. It has this somewhat strange name because it was originally used to characterize radio receivers at the reception end of a noisy connection in World War II. But it has proven very useful in psychology, medicine and other areas of decision making.

In Figure 11 you see a number of such ROC curves. They are plotted in a coordinate system with our two important measures, p(hits) and p(false alarms), on respectively the vertical and horizontal axes. First have a look at observer D. This person lies on the diagonal of the plot, which means that for this person, p(false alarms) = p(hits). But if a person makes as many errors as he gets it right, it means that that person cannot distinguish signal from noise at all. In other words, for observer D, d'= 0. Please note that this goes for every point on the diagonal. For instance, observer A may decide to increase hits, but on the diagonal this will increase false alarms to the same extent.

Contrast this with observer A. If you read from the axes, you see that observer A has a very high proportion of hits, and at the same time a very low proportion of false alarms. In other words, this observers is very good, very sensitive. Thus ROCs that steeply curve into the top left corner of the graph (the 0,1 extreme) signify a very high d'.

Observers B and C have more moderate sensitivities, but they illustrate another aspect of the ROC, namely the bias. Notice that observer B has a higher sensitivity (d'= 2) than observer C (d'= 1), yet at the same time observer C has a higher proportion of hits than observer B. However, observer C also has a very high proportion of false alarms. In other words, observer C has adopted a very liberal criterion. In contrast, observer B scores fewer hits, but also scores very low on false alarms, both indicative of a conservative criterion. So, the more an observer is positioned towards the 0,0 extreme, the more conservative they are – and vice versa, the more they score towards the 1,1 extreme, the more liberal they are. This is illustrated further in Figure 12. Thus, taken together, the position on the ROC provides information on both sensitivity and bias.

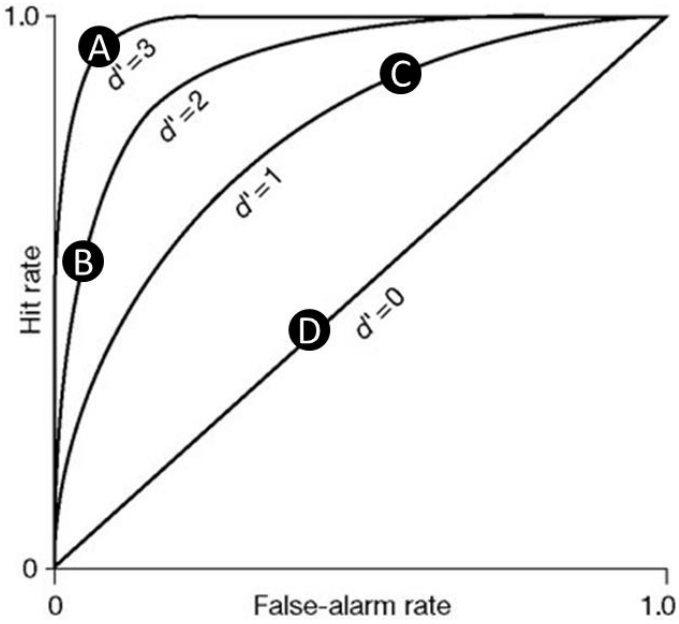

Figure 11. Four different observers (A-D) on four different ROC curves

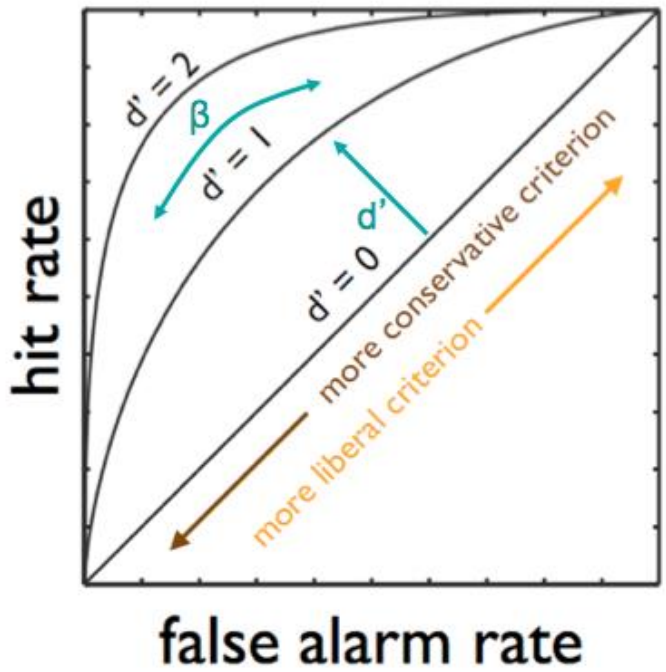

Figure 12. Illustrating the directions in which *d'* and *β* change. For *β* you can also think *c*.

z	p(z)	f(z)	z	p(z)	f(z)	z	p(z)	f(z)	z	p(z)	f(z)	z	p(z)	f(z)	z	p(z)	f(z)
-3.00	0.001	0.004	-2.00	0.023	0.054	-1.00	0.159	0.242	0.00	0.500	0.399	1.00	0.841	0.242	2.00	0.977	0.054
-2.98	0.001	0.005	-1.98	0.024	0.056	-0.98	0.164	0.247	0.02	0.508	0.399	1.02	0.846	0.237	2.02	0.978	0.052
-2.96	0.002	0.005	-1.96	0.025	0.058	-0.96	0.169	0.252	0.04	0.516	0.399	1.04	0.851	0.232	2.04	0.979	0.050
-2.94	0.002	0.005	-1.94	0.026	0.061	-0.94	0.174	0.256	0.06	0.524	0.398	1.06	0.855	0.227	2.06	0.980	0.048
-2.92	0.002	0.006	-1.92	0.027	0.063	-0.92	0.179	0.261	0.08	0.532	0.398	1.08	0.860	0.223	2.08	0.981	0.046
-2.90	0.002	0.006	-1.90	0.029	0.066	-0.90	0.184	0.266	0.10	0.540	0.397	1.10	0.864	0.218	2.10	0.982	0.044
-2.88	0.002	0.006	-1.88	0.030	0.068	-0.88	0.189	0.271	0.12	0.548	0.396	1.12	0.869	0.213	2.12	0.983	0.042
-2.86	0.002	0.007	-1.86	0.031	0.071	-0.86	0.195	0.276	0.14	0.556	0.395	1.14	0.873	0.208	2.14	0.984	0.040
-2.84	0.002	0.007	-1.84	0.033	0.073	-0.84	0.200	0.280	0.16	0.564	0.394	1.16	0.877	0.204	2.16	0.985	0.039
-2.82	0.002	0.007	-1.82	0.034	0.076	-0.82	0.206	0.285	0.18	0.571	0.393	1.18	0.881	0.199	2.18	0.985	0.037
-2.80	0.003	0.008	-1.80	0.036	0.079	-0.80	0.212	0.290	0.20	0.579	0.391	1.20	0.885	0.194	2.20	0.986	0.035
-2.78	0.003	0.008	-1.78	0.038	0.082	-0.78	0.218	0.294	0.22	0.587	0.389	1.22	0.889	0.190	2.22	0.987	0.034
-2.76	0.003	0.009	-1.76	0.039	0.085	-0.76	0.224	0.299	0.24	0.595	0.388	1.24	0.893	0.185	2.24	0.987	0.032
-2.74	0.003	0.009	-1.74	0.041	0.088	-0.74	0.230	0.303	0.26	0.603	0.386	1.26	0.896	0.180	2.26	0.988	0.031
-2.72	0.003	0.010	-1.72	0.043	0.091	-0.72	0.236	0.308	0.28	0.610	0.384	1.28	0.900	0.176	2.28	0.989	0.030
-2.70	0.003	0.010	-1.70	0.045	0.094	-0.70	0.242	0.312	0.30	0.618	0.381	1.30	0.903	0.171	2.30	0.989	0.028
-2.68	0.004	0.011	-1.68	0.046	0.097	-0.68	0.248	0.317	0.32	0.626	0.379	1.32	0.907	0.167	2.32	0.990	0.027
-2.66	0.004	0.012	-1.66	0.048	0.101	-0.66	0.255	0.321	0.34	0.633	0.377	1.34	0.910	0.163	2.34	0.990	0.026
-2.64	0.004	0.012	-1.64	0.051	0.104	-0.64	0.261	0.325	0.36	0.641	0.374	1.36	0.913	0.158	2.36	0.991	0.025
-2.62	0.004	0.013	-1.62	0.053	0.107	-0.62	0.268	0.329	0.38	0.648	0.371	1.38	0.916	0.154	2.38	0.991	0.023
-2.60	0.005	0.014	-1.60	0.055	0.111	-0.60	0.274	0.333	0.40	0.655	0.368	1.40	0.919	0.150	2.40	0.992	0.022
-2.58	0.005	0.014	-1.58	0.057	0.115	-0.58	0.281	0.337	0.42	0.663	0.365	1.42	0.922	0.146	2.42	0.992	0.021
-2.56	0.005	0.015	-1.56	0.059	0.118	-0.56	0.288	0.341	0.44	0.670	0.362	1.44	0.925	0.141	2.44	0.993	0.020
-2.54	0.006	0.016	-1.54	0.062	0.122	-0.54	0.295	0.345	0.46	0.677	0.359	1.46	0.928	0.137	2.46	0.993	0.019
-2.52	0.006	0.017	-1.52	0.064	0.126	-0.52	0.302	0.348	0.48	0.684	0.356	1.48	0.931	0.133	2.48	0.993	0.018
-2.50	0.006	0.018	-1.50	0.067	0.130	-0.50	0.309	0.352	0.50	0.691	0.352	1.50	0.933	0.130	2.50	0.994	0.018
-2.48	0.007	0.018	-1.48	0.069	0.133	-0.48	0.316	0.356	0.52	0.698	0.348	1.52	0.936	0.126	2.52	0.994	0.017
-2.46	0.007	0.019	-1.46	0.072	0.137	-0.46	0.323	0.359	0.54	0.705	0.345	1.54	0.938	0.122	2.54	0.994	0.016
-2.44	0.007	0.020	-1.44	0.075	0.141	-0.44	0.330	0.362	0.56	0.712	0.341	1.56	0.941	0.118	2.56	0.995	0.015
-2.42	0.008	0.021	-1.42	0.078	0.146	-0.42	0.337	0.365	0.58	0.719	0.337	1.58	0.943	0.115	2.58	0.995	0.014
-2.40	0.008	0.022	-1.40	0.081	0.150	-0.40	0.345	0.368	0.60	0.726	0.333	1.60	0.945	0.111	2.60	0.995	0.014
-2.38	0.009	0.023	-1.38	0.084	0.154	-0.38	0.352	0.371	0.62	0.732	0.329	1.62	0.947	0.107	2.62	0.996	0.013
-2.36	0.009	0.025	-1.36	0.087	0.158	-0.36	0.359	0.374	0.64	0.739	0.325	1.64	0.949	0.104	2.64	0.996	0.012
-2.34	0.010	0.026	-1.34	0.090	0.163	-0.34	0.367	0.377	0.66	0.745	0.321	1.66	0.952	0.101	2.66	0.996	0.012
-2.32	0.010	0.027	-1.32	0.093	0.167	-0.32	0.374	0.379	0.68	0.752	0.317	1.68	0.954	0.097	2.68	0.996	0.011
-2.30	0.011	0.028	-1.30	0.097	0.171	-0.30	0.382	0.381	0.70	0.758	0.312	1.70	0.955	0.094	2.70	0.997	0.010
-2.28	0.011	0.030	-1.28	0.100	0.176	-0.28	0.390	0.384	0.72	0.764	0.308	1.72	0.957	0.091	2.72	0.997	0.010
-2.26	0.012	0.031	-1.26	0.104	0.180	-0.26	0.397	0.386	0.74	0.770	0.303	1.74	0.959	0.088	2.74	0.997	0.009
-2.24	0.013	0.032	-1.24	0.107	0.185	-0.24	0.405	0.388	0.76	0.776	0.299	1.76	0.961	0.085	2.76	0.997	0.009
-2.22	0.013	0.034	-1.22	0.111	0.190	-0.22	0.413	0.389	0.78	0.782	0.294	1.78	0.962	0.082	2.78	0.997	0.008
-2.20	0.014	0.035	-1.20	0.115	0.194	-0.20	0.421	0.391	0.80	0.788	0.290	1.80	0.964	0.079	2.80	0.997	0.008
-2.18	0.015	0.037	-1.18	0.119	0.199	-0.18	0.429	0.393	0.82	0.794	0.285	1.82	0.966	0.076	2.82	0.998	0.007
-2.16	0.015	0.039	-1.16	0.123	0.204	-0.16	0.436	0.394	0.84	0.800	0.280	1.84	0.967	0.073	2.84	0.998	0.007
-2.14	0.016	0.040	-1.14	0.127	0.208	-0.14	0.444	0.395	0.86	0.805	0.276	1.86	0.969	0.071	2.86	0.998	0.007
-2.12	0.017	0.042	-1.12	0.131	0.213	-0.12	0.452	0.396	0.88	0.811	0.271	1.88	0.970	0.068	2.88	0.998	0.006
-2.10	0.018	0.044	-1.10	0.136	0.218	-0.10	0.460	0.397	0.90	0.816	0.266	1.90	0.971	0.066	2.90	0.998	0.006
-2.08	0.019	0.046	-1.08	0.140	0.223	-0.08	0.468	0.398	0.92	0.821	0.261	1.92	0.973	0.063	2.92	0.998	0.006
-2.06	0.020	0.048	-1.06	0.145	0.227	-0.06	0.476	0.398	0.94	0.826	0.256	1.94	0.974	0.061	2.94	0.998	0.005
-2.04	0.021	0.050	-1.04	0.149	0.232	-0.04	0.484	0.399	0.96	0.831	0.252	1.96	0.975	0.058	2.96	0.998	0.005
-2.02	0.022	0.052	-1.02	0.154	0.237	-0.02	0.492	0.399	0.98	0.836	0.247	1.98	0.976	0.056	2.98	0.999	0.005