2 Experimental Design and the Basics of Statistics: Signal Detection Theory (SDT)

Contents

2.1 The Classic Scenario of SDT 13
2.2 SDT and the Percentage of Correct Responses 17
2.3 The Empirical d' 19

What You Will Learn in This Chapter

What is a good performance measure? The most frequently used measure is the percentage of correct responses. Here, we will see that percent correct confuses two variables, namely, sensitivity and criterion, and should therefore be used with care. We will introduce the sensitivity measure d', which turns out to be a crucial term in much of statistics.

2.1 The Classic Scenario of SDT

Assume we are in a yellow submarine cruising through the ocean. It can be quite dangerous to hit a rock and for this reason, the submarine is equipped with a sonar device. Sonar waves are emitted and their reflections are recorded by a receiver. These reflections are combined to form what we call the "sonar measure." If there is a rock, the sonar measure is larger than when there is no rock. However, the situation is noisy and hence even under the very same rock or no-rock condition the sonar measure varies quite a bit across recordings (Fig. 2.1).

A probability distribution corresponds to each of the two conditions, rock vs. no-rock, that indicates how likely it is that a certain value of the sonar measure, indicated on the

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_2

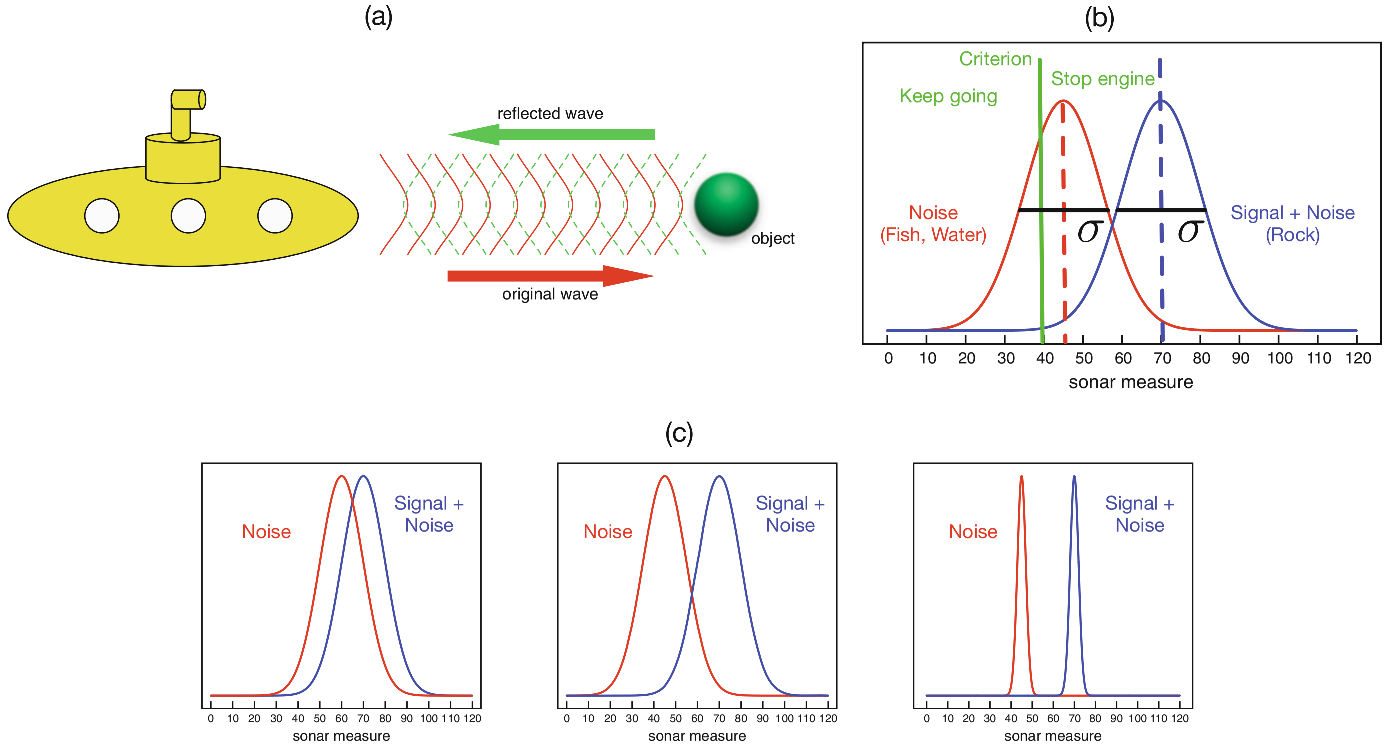


Fig. 2.1 (**a**) A submarine sends out sonar signals and measures a response based on the echoes. The sonar measure is noisy, i.e., for the very same rock, the echoes can differ from sonar measure to sonar measure. For example, fish may change the reflectance. The same is true when there is no rock. How to decide whether or not there is a rock? (**b**) The classic scenario of SDT. To both the rock and no-rock condition belongs a probability distribution, which shows how likely it is to receive a sonar measure at various values. SDT assumes that these probability distributions are Gaussians.

Fig. 2.1 (continued) In this example, the mean value of the sonar measure is 70 when a rock is present and 45 when there is no rock. The spread of a distribution is described by its standard deviation σ. Here, $\sigma = 10$. If we record a sonar measure of 80 it is much more likely that there is a rock than not. A sonar measure of 57.5 is equally likely from a situation with a rock and a situation without a rock. To make a decision, a criterion is needed. When the sonar measure is larger than the criterion, we decide to stop the engine (conclude there is a rock), otherwise we continue cruising (conclude there is no rock). Where to set the criterion is our choice. If we are conservative, we set the criterion to low values, when we are more risky we set it to higher values. (**c**) How well we can discriminate between the rock and no-rock conditions depends on the overlap between the two probability distributions. The overlap is denoted by d', which is the difference between the mean values divided by the standard deviation. A large overlap means low discriminability, a low overlap means high discriminability. For a fixed standard deviation, d' increases when the mean difference increases (left vs. center figure). For a fixed mean value difference, discriminability increases when the standard deviation decreases (center vs. right figure)

x-axis, is elicited.[1] As is often the case in statistics, we assume that Gaussian distributions adequately describe the situation. A Gaussian distribution is fully determined by its mean value μ and the standard deviation σ, which determines the width of the Gaussian. A large σ means that for the very same rock condition, very different signals can be reflected with a high probability. As the other extreme, if $\sigma = 0$, there is no variability. We always receive the same value of the sonar measure. Hence, σ reflects how noisy the situation is and that is why σ is also called the noise. We call the no-rock condition the *noise alone condition* because there is no signal from the rock and we call the rock condition the *signal plus noise* condition.

How well can we distinguish between the rock and no rock conditions? It depends on the overlap between the Gaussians. If the Gaussians fully overlap, we cannot discriminate between the two cases. The sonar is useless. When there is almost no-overlap, it is easy to discriminate the two situations because a certain sonar measure value originates very likely from only one of the Gaussians. For example, Fig. 2.1b indicates that a sonar measure of 80 occurs from the rock condition with a much higher likelihood than from the no-rock condition. The overlap can be captured by the difference between the means, μ_1 and μ_2, of the two Gaussians divided by the standard deviation σ, which is assumed to be the same for both distributions[2]:

$$d' = \frac{\mu_1 - \mu_2}{\sigma} \tag{2.2}$$

d' is called sensitivity or discriminability and it measures how well we can discriminate between two alternatives, i.e., d' is a measure of the signal $(\mu_1 - \mu_2)$ to noise (σ) ratio. The overlap depends on both the difference of the means and the standard deviation. Hence, d' can be increased by both increasing the mean value difference or decreasing the standard deviation (Fig. 2.1c). Importantly, the notion of sensitivity here is different from the notion in Chap. 1, where it is identical with the Hit rate. In the following, we will use sensitivity only for Hit rate and not d'.

When should we stop the engine? A decision criterion c is needed. In Fig. 2.1b, the criterion is set at 40, i.e., if the sonar measure is larger than 40, we stop the engine, if

[1] Strictly speaking, the probability is zero that the sonar measure is *exactly* 80. The probability function shows the probability that a value close to 80 occurs (a value from a very narrow interval around 80). Whereas these aspects are crucial for mathematical statistics, they hardly play a role for the basic understanding of statistics.

[2] In Signal Detection Theory (SDT), d' is usually defined by the absolute value of the difference of the means:

$$d' = |\frac{\mu_1 - \mu_2}{\sigma}| \tag{2.1}$$

We use the definition without the absolute values because it is better suited when we apply d' to statistics in Chap. 3.

the sonar measure is smaller than 40, we continue cruising. Where we set the criterion is our choice. If a rock is as likely as a no-rock situation, then the optimal criterion is at the intersection point of the two Gaussians, in the sense that it maximizes the number of correct decisions.

2.2 SDT and the Percentage of Correct Responses

Let us apply SDT to a behavioral experiment. Consider a typical detection experiment. You are looking at a computer screen and either a faint light patch is presented (stimulus present) or the screen is blank (stimulus absent). As in the yellow submarine and HIV test examples, there are four possible outcomes (Table 2.1).

If the stimulus is presented in half of the trials, the percentage of correct responses is computed by the average of the Hit rate and Correct Rejection rate: $\frac{Hit+CR}{2}$. Let us consider "percent correct" in terms of SDT. As in the submarine example, we assume that decision making is noisy. Contrary to the submarine example, we do not know how the perceptual processes are coded in the human brain, i.e., we do not explicitly know the probability distributions. The good thing about SDT is that it is very flexible. For example, we assume that we can focus on one neuron, which codes for the brightness of the stimuli. A value of 0.0 corresponds to a blank screen and a positive value to a light patch with a certain brightness. We use a decision criterion that determines whether we decide for the light patch or the blank. If we are more conservative, we set the criterion to a higher value, i.e., we respond 'light patch present' only when we are quite sure. If we are more risky, we set the criterion to a lower value, i.e., respond 'light patch present' when there is the slightest evidence for a light being present. We can set the criterion to any value we wish. For example to optimise the percentage of correct responses, we can set the criterion at

Table 2.1 There are four outcomes in a classic SDT experiment

		Stimulus present	**Stimulus absent**
Response	**Present**	Hit	False Alarm (FA)
	Absent	Miss	Correct Rejection (CR)

A rock/light patch is present and we decide a rock/light patch is present (Hit). A rock/light patch is present and we decide no rock/light patch is present (Miss). No rock/light patch is present and we decide a rock/light patch is present (False Alarm). No rock/light patch is present and we decide no rock/light patch is present (Correct Rejection) the intersection of the Gaussians. In this case, we respond equally often for 'light patch present' and 'light patch absent' when both stimulus alternatives are presented with the same frequency. If there are more patch absent than patch present situations, we may want to move the criterion towards more "patch absent" responses, i.e., towards the right in this case. In addition, we may want to take the costs of our decision into account. If a "patch absent" response is less rewarded than a "patch present" response, we may want to move the criterion so that more "patch present" responses occur.

Let us change the criterion smoothly and see how the percentage of correct responses changes with d' (Fig. 2.2). We start with a criterion set to 2.0, i.e., at the intersection point. If we now move the criterion a bit from optimal, e.g., to the right, the percentage

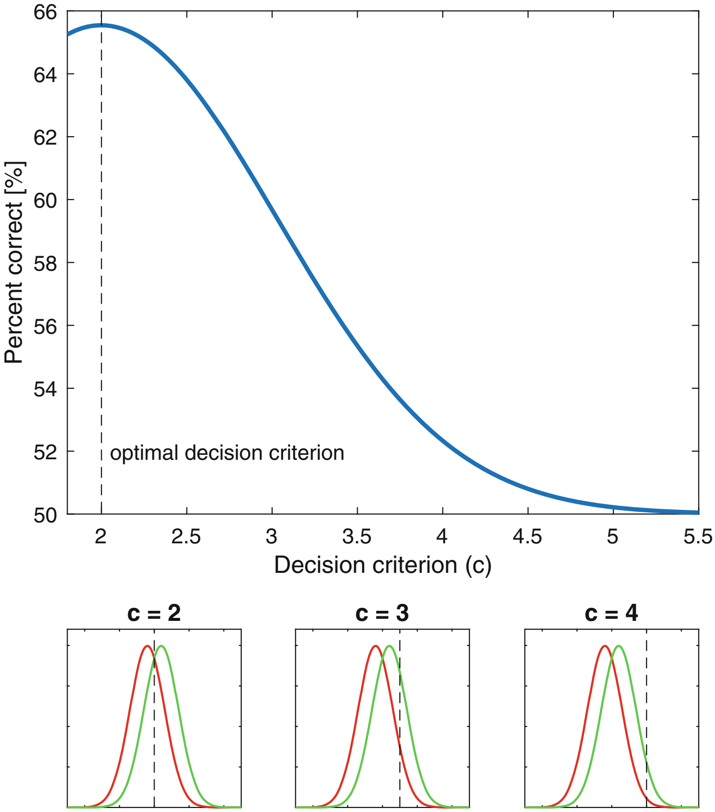


Fig. 2.2 Percent correct depends on the criterion. Let us, first, put the criterion at the intersection of the two Gaussians, which we set at the value 2 (lower left panel). Let us now move the criterion to the right (center and right). The upper panel shows that the percentage of correct responses decreases the further we move the criterion from optimal. If the criterion is set to the most right, we almost always respond for one response alternative, such as "patch absent." In this case, the Hit rate is 0 but the Correct Rejection rate is 1. Hence, percent correct is: $(0 + 1)/2 = 0.5$, i.e., we are at chance level, even though we can, in principle, well discriminate between the two situations

of correct responses deteriorates and it does so the more we move the criterion. If we move the criterion to the far right, performance approaches 50%, i.e., chance level. In this case we always give the same response, say, “light patch absent”, i.e., we have a large *response bias*. Importantly, the discriminability d' has not changed, i.e., our visual abilities are constant. Only our decision behavior has changed.

Hence, the percentage of correct responses confounds decision criterion and discriminability. For a certain performance level, say 75%, we cannot tell whether there is high discriminability and a suboptimal criterion or an optimal criterion with a lower discriminability. For this reason, the percentage of correct responses can be a dangerous measure, which can easily obliterate true effects or lead to false positives. Conclusions based on the percentage of correct responses are based on partial information!

2.3 The Empirical d'

From only the percentage of correct responses, it is impossible to infer the discriminability in an experiment. Aren’t then all experiments hopeless? Surprisingly, one can disentangle discriminability and criterion by separately estimating d' and b, the *bias* of the criterion:

$$d'_{emp} = z(Hit) - z(FA) \tag{2.3}$$

$$b_{emp} = -\frac{z(Hit) + z(FA)}{2} \tag{2.4}$$

To compute d', we simply need to z-transform the Hit and False Alarm rate. The z-transformation is the inverse cumulative Gaussian function. If you are not familiar with the z-transformation just treat it as a function you can find on your computer. b_{emp} tells you how much the current criterion differs from the optimal one, i.e., how far it is from the intersection point of the Gaussians. Hence, b_{emp} measures the bias of the criterion.

Importantly, d'_{emp} does not change when the criterion (and thus response bias) changes. However, d' is criterion free but not model free. There are three assumptions:

1. The underlying probability distributions are Gaussians.
2. The Gaussians have the same variance.
3. The criterion is not changed during the measurements.

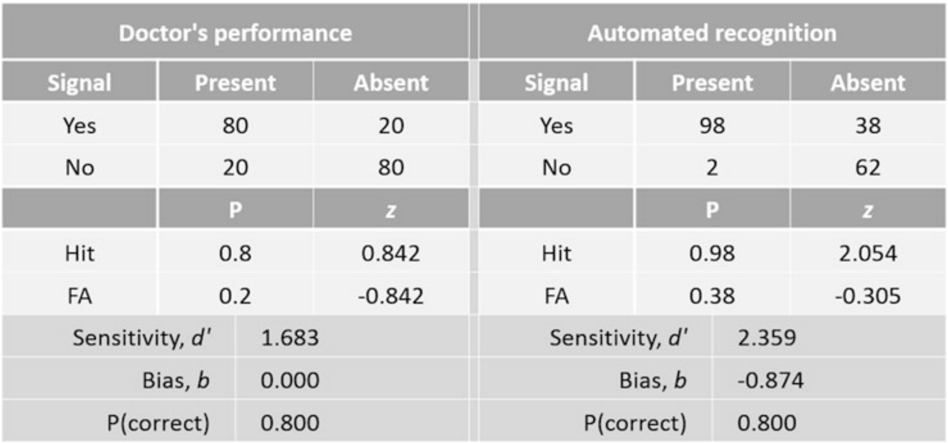
Assumption 1 is crucial since we compute the z-transform, i.e., the inverse Gaussian distribution, which only makes sense when the data are Gaussian distributed. Assumption 1 is often fulfilled. Assumption 2 is usually fulfilled since the stimulus alternatives are similar. Assumption 3 is crucial and is not easy to check.

Attention

The term “sensitivity” is used in two different ways:

1. Mainly in the medical literature, sensitivity is the same as Hit Rate.
2. In SDT, sensitivity corresponds to discriminability, i.e., $d' = z(Hit) - z(FA)$. We will use the term discriminability rather than sensitivity in the context of SDT.

Example 1 (Automated System) Performance (Hit, FA, Miss, CR) of both a doctor and an artificial intelligence system (AI) for diagnosing a disease is shown in Fig. 2.3. The overall percentage of correct responses is 80% for the two. d' is almost as simple to compute as percent correct: one only needs to z-transform Hit and FA. As it turns out, performance in terms of d', contrary to the percentage correct, is strongly different. Is the doctor or the AI system better? Usually, discriminability is an inbuilt, hard to change, characteristic of a system. Your eyes are as good as they are. On the other hand, changing the decision criterion is easy, one just needs to respond more often for one than the other alternative. Obviously, the AI system is strongly biased towards “yes” responses, thereby avoiding misses but leading to a higher false alarm rate. Hence, the AI’s system criterion is far from being optimal. Setting the criterion to optimal strongly increases performance in terms of percent correct.



Doctor's performance		
Signal	Present	Absent
Yes	80	20
No	20	80
	P	z
Hit	0.8	0.842
FA	0.2	-0.842
Sensitivity, *d'*	1.683	
Bias, *b*	0.000	
P(correct)	0.800	

Automated recognition		
Signal	Present	Absent
Yes	98	38
No	2	62
	P	z
Hit	0.98	2.054
FA	0.38	-0.305
Sensitivity, *d'*	2.359	
Bias, *b*	-0.874	
P(correct)	0.800	

Fig. 2.3 Doctor vs. Machine. The percentage of correct responses is identical for both the doctor and the machine. d' is almost as simple to compute as percent correct: one only needs to z-transform Hit and FA and subtract. Performance in terms of d', i.e, discriminability, is strongly different. Is the doctor or the AI system better? Obviously, the AI system is strongly biased towards “yes” responses, which avoids Misses but also leads to a higher False Alarm rate. The criterion for the AI system is far from optimal. Courtesy: Mark Georgeson

Example 2 (Learning) In a learning experiment, observers are shown either a left or right tilted line. Since the difference is small, observers perform poorly. To improve performance, observers train on the task with 10 blocks, each containing 80 trials. We consider the number of correct responses for each of the 10 blocks, by averaging the 80 trials. Performance improves strongly. Does this imply that perception has improved? Improvements can be caused by changes in both discriminability or criterion. For example, training with the stimuli may lead to a decrease of the variance σ of the Gaussians, i.e., people can more clearly discriminate the tilt of the lines. A decrease of the variance leads to an increase in d', i.e., an increase in discriminability (Fig. 2.2). An increase in discriminability can also occur when the means of the Gaussians are pulled apart. Performance can also improve when the decision criterion of the participants is not optimal in block 1. During training, participants may learn to adjust the criterion. A change of *perception* is generally considered to be related to a change of discriminability. When we analyze the data with the percentage of correct responses, we cannot make proper conclusions since we cannot disentangle changes in discriminability from changes in criterion. Hence, for all learning experiments, it is important to plot the results as d' and $bias$.

Example 3 (Sensitivity and Specificity) In Chap. 1, the HIV test had a very high sensitivity and specificity. Determining sensitivity and specificity also depends on a criterion. In fact, it is in no way different than determining percent correct. Recall that sensitivity is Hit rate and specificity is the rate of Correct rejections. Thus, the situation is exactly the same as in the above example with the submarine, except that on the x-axis we have HIV anti-body concentration (as measured by the test). We need a criterion that determines whether or not the test is positive for a certain antibody concentration. Hence, we can increase sensitivity at the cost of specificity and vice versa. Just to mention, (sensitivity + specificity)/2 is percent correct.

Example 4 (Speed-Accuracy Trade-Off) In many experiments, responses are speeded, i.e., observers need to respond as fast as possible. Often, slow observers, e.g., older people, have a higher d' than fast observers, e.g., younger people; a so called speed-accuracy trade-off. So the situation is even more complicated because we need to pit Reaction Times against d' and bias to reach proper conclusions. Experiments with a clear speed-accuracy trade-off are often hard to interpret.

Example 5 (Floor and Ceiling Effects) One additional issue involves so called floor and ceiling effects. In a (non-sense) experiment, the experimenter holds up one hand, with all fingers clearly visible. When asked, all observers correctly identify the 5 fingers, i.e., 100% correct responses. Can we conclude that all observers have the same good eyes, i.e., the same discriminability? Of course not; the task was too simple and for this reason observers performed in a ceiling regime, i.e., close to 100%. Conclusions are useless. Computing d' does not help in this situation because the false alarm rate is 0.0 and d' is infinity.

The very same is true for floor effects, i.e., when performance is close to chance level (50%). It is therefore important to make sure that stimulus alternatives are in a range where the differences between participants can be detected.

Example 6 (Standardized Effects) d' is also often called a standardized effect because the division by σ converts the measurement to units of standard deviation. As a result, d' is insensitive to the original units (e.g., it does not matter whether the original measurements are in meters or inches). Moreover, a standardized effect size can also be insensitive to some experimental variations. For example, if a reaction time experiment can be manipulated to make everything slower by a common factor, then both the difference of means will increase (signal) and the standard deviation will increase (noise) by an equal factor. The ratio, d', will be unchanged.

Take Home Messages

1. Be aware of partial information. The percentage of correct responses confounds discriminability d' and decision criterion c.
2. Be aware of partial information. The same is true for many other measures such as Sensitivity and Specificity in medical tests.
3. You can disentangle discriminability and criterion by using d'_{emp}.
4. d'_{emp} is criterion free but not model free. There is no free lunch.