Many scenarios generate data that fit into two mutually exclusive categories. Some examples are industrial quality control testing, clinical trials, and signal processing. This monograph presents a very brief overview of sequential testing and the sequential probability ratio test, and demonstrates some empirical results via simulation.

As motivation, imagine a clinical trial where the study drug is expensive, the monitoring of the participant is expensive, and the study drug side effects are very undesirable. In this case, we want to make a decision on whether the study drug is effective as quickly as possible, and by enrolling as few people as possible, subject to some prespecified Type I and II error probabilities.

One way this problem can be approached is to invoke the Bayesian paradigm. Here we choose to accept or reject the null hypothesis at the mth step if, respectively, the posterior probability of our null (alternative) model is less (greater) than some prespecified marker, to be calibrated. That is, we accept H_0 if

$$g_{0m} \equiv P(\mathrm{H}_0|\mathbf{X}_m) = \frac{P(\mathbf{X}_m|H_0) \times P(H_0)}{P(\mathbf{X}_m|H_0) \times P(H_0) + P(\mathbf{X}_m|H_1) \times P(H_1)} > d_0$$

where d_0 is the to-be-calibrated decision threshold. Define $g_0 = \mathrm{P}(\mathrm{H}_0)$ and $g_1 = 1 - g_0 = \mathrm{P}(\mathrm{H}_1)$. We can rewrite the above probability statement in terms of densities using Bayes' Rule to yield the decision rule

$$\text{Accept } \mathrm{H}_0 \text{ if:} \quad \Lambda_m \equiv \frac{f_1(\mathbf{X}_m)}{f_0(\mathbf{X}_m)} \leq \frac{g_0(1-d_0)}{g_1 \times d_0}$$

Formatting the relation in terms of the alternative hypothesis derivation yields a similar result for rejection.

To decide the decision thresholds, define γ_0 and γ_1 as decision rules which are functions of the *a priori* Type I and Type II error probabilities of the test, and define sets $\mathcal{A}_1 = \{\mathbf{x}|\Lambda_m \geq \gamma_1\}$ and $\mathcal{A}_0 = \{\mathbf{x}|\Lambda_m \leq \gamma_0\}$, subsets of the sample space $\mathcal{X}^m$. Looking at the power of the test, we have that

$$1 - \beta = P_{\mathrm{H}_1}(\mathcal{A}_1) = \int_{\mathcal{A}_1} f_1(\mathbf{x})d\mathbf{x} = \int_{\mathcal{A}_1} \frac{f_1(\mathbf{X}_m)}{f_0(\mathbf{X}_m)} f_0(\mathbf{x})d\mathbf{x} =$$

$$\int_{\mathcal{A}_1} \Lambda_m f_0(\mathbf{x})d\mathbf{x} \geq \gamma_1 \int_{\mathcal{A}_1} f_0(\mathbf{x})d\mathbf{x} = \gamma_1\alpha$$

Thus it follows that we have constraint $\gamma_1 \leq (1-\beta)/\alpha$. A similar power calculation under the null yields constraint $\gamma_0 \geq \beta/(1-\alpha)$. Setting β and α will give the desired decision thresholds. Wald, in his original formulation[1] of sequential testing, recommended the conservative approach of setting the decision bounds equal to the constraints. I adopt that approach for all that follows. We can now back-solve to find d_0 and d_1, if we wish, but since we are testing in terms of the likelihood ratio, and not the posterior density, it makes more sense to simply use the decision thresholds derived above, and ignore calibration of the original decision boundaries d_0 and d_1.

Finally, we would like to know how many trials it will take, on average, before we make a decision. Here I invoke some of the terminology and results from the field of stochastic processes. Let

[1] Wald, A. Sequential Tests of Statistical Hypotheses. Annals of Mathematical Statistics 16 (1945), no. 2, 117–186.

M be the number of trials before we make a decision. In the language of stochastic processes, it is a Stopping Time. By Wald's equation (a result on random sums of random variables) we have that the expectation of the log of the LRT (i.e. the product of the densities) at the random stopping time M is

$$\mathrm{E}_{\mathrm{H}_i}[\log(\Lambda_m)] = \mathrm{E}_{\mathrm{H}_i}\left[\sum_{i=1}^{M} \log\left(\frac{f_1(x_i)}{f_0(x_i)}\right)\right] = \begin{cases} \mathrm{E}_{\mathrm{H}_i}[M] \times D(f_1||f_0) & i{=}1 \\ -\mathrm{E}_{\mathrm{H}_i}[M] \times D(f_0||f_1) & i{=}0 \end{cases}$$

where the operator $D(f_1||f_0)$ is the Kullback-Liebler divergence, or the expected value of the logarithm of the density ratios. This means that

$$\mathrm{E}_{\mathrm{H}_i}[M] = \begin{cases} \mathrm{E}_{\mathrm{H}_i}[\log(\Lambda_m)]/D(f_1||f_0) & i{=}1 \\ \mathrm{E}_{\mathrm{H}_i}[\log(\Lambda_m)]/D(f_1||f_0) & i{=}0 \end{cases}$$

Now we have the expected number of trials until we make a decision (or expected number of widgets destroyed, items sampled, etc.), and we can apply the results.

Example 1:
Imagine a scenario where we wish to test whether the difference in a blood component in malaria patients changes with the administration of an expensive medication. The drug side effects include fatigue, dizziness, and some stomach pain. If we assume that the level of the blood component is roughly normally distributed, then we have hypotheses

$$\mathrm{H}_0 : \ X \ \sim \ N(0,1) \ \text{ or } \ \mathrm{H}_1 : \ X \ \sim \ N(\theta,1)$$

with $\theta \neq 0$. First, I show that the sequential probability ratio test (SPRT) in this setting can be expressed as a rule which says to stop sampling at the m such that $\sum_{i=1}^{m} X_i$ passes outside a region defined by a pair of linear boundaries. For X Gaussian, the likelihood ratio is

$$\Lambda_m = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}}{\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}} = \frac{\prod_{i=1}^{m} e^{-\frac{(x_i-\theta)^2}{2}}}{\prod_{i=1}^{m} e^{-\frac{x_i^2}{2}}}$$

And the log-likelihood is

$$\log(\Lambda_m) = \log(f_1(\mathbf{x})) - \log(f_0(\mathbf{x})) = \sum_{i=1}^{m} x_i^2/2 - \sum_{i=1}^{m} (x_i-\theta)^2/2 = \theta \sum_{i=1}^{m} x_i - \frac{m\theta^2}{2}$$

The likelihood meeting a decision boundary B is equivalent to the sum of the observations reaching equality with :

$$\sum_{i=1}^{m} x_i = \frac{\log(\text{Bound } B)}{\theta} + \frac{m\theta}{2}$$

As was shown above, the optimal constrained decision is to reject if the SPRT statistic is greater than or equal to $\gamma_1 = (1-\beta)/\alpha$, and accept H_0 if the SPRT is less than or equal to $\gamma_0 = \beta/(1-\alpha)$.

Now let M denote the random number of samples needed to make a decision between H_0 and H_1 (a "hitting time"). Above, I used Wald's optimal stopping time theorem to show that

$$E_{H_i}[M] = \frac{E_{H_i}[\log(\Lambda_m)]}{D(f_1||f_0)}$$

Since I assume that I am making my decision at the boundary where the SPRT hits some γ, I can write the numerator of the term as just a linear combination of the decision boundaries and the false positive and false negative error probabilities:

$$\begin{aligned} E_{H_0}[\log(\Lambda_m)] &\approx (1-\alpha)\log(\gamma_0) + \alpha\log(\gamma_1) \\ E_{H_1}[\log(\Lambda_m)] &\approx (1-\beta)\log(\gamma_1) + \beta\log(\gamma_0) \end{aligned}$$

For the denominator, I note that the KL-divergence is not symmetric in general, however here the formula reduces to

$$E_{H_i}[\log(f_1/f_0)] = \frac{1+(0-\theta)^2}{2} - \frac{1}{2} = \frac{\theta^2}{2}$$

To show this I derive the KLD for two normal distributions p and q, each with unit variance.

$$\int [\log(p) - \log(q)]p(x)dx = \int \left[-\frac{1}{2}(x-\theta_1)^2 + \frac{1}{2}(x-\theta_2)^2 \right] \frac{1}{\sqrt{2\pi}} \exp\{-(x-\theta_1)^2/2\}dx =$$

$$\frac{1}{2}E_1[(x-\theta_2)^2] - \frac{1}{2}E_1[(x-\theta_1)^2] = \frac{1}{2}E_1[(x-\theta_2)^2] - \frac{1}{2}$$

Adding and subtracting θ_1 inside the square in the remaining expectation yields

$$\frac{1+(\theta_1-\theta_2)^2}{2} - \frac{1}{2}$$

As was the claim, this is symmetric with respect to the order of distribution, and so in this case the KLD is identical under both H_0 and H_1. Therefore the expected wait time until a decision is made under the two hypotheses is

$$\begin{aligned} E_{H_0}[M] &\approx \frac{(1-\alpha)\log(\gamma_0) + \alpha\log(\gamma_1)}{\theta^2/2} \\ E_{H_1}[M] &\approx \frac{(1-\beta)\log(\gamma_1) + \beta\log(\gamma_0)}{\theta^2/2} \end{aligned}$$

I now use simulation to show some hitting times for $\theta = 1.75$, $\alpha = 0.05$, and $\beta = 0.01$. I plot the value of the sum $\sum_{i=1}^{m} X_i$ on the y-axis, with the mth trials on the x-axis, and with the decision boundaries plotted as dashed lines. Here the boundary is crossed on the fifth trial.

Figure 1: Red color indicates crossing to decision H_1, blue to H_0

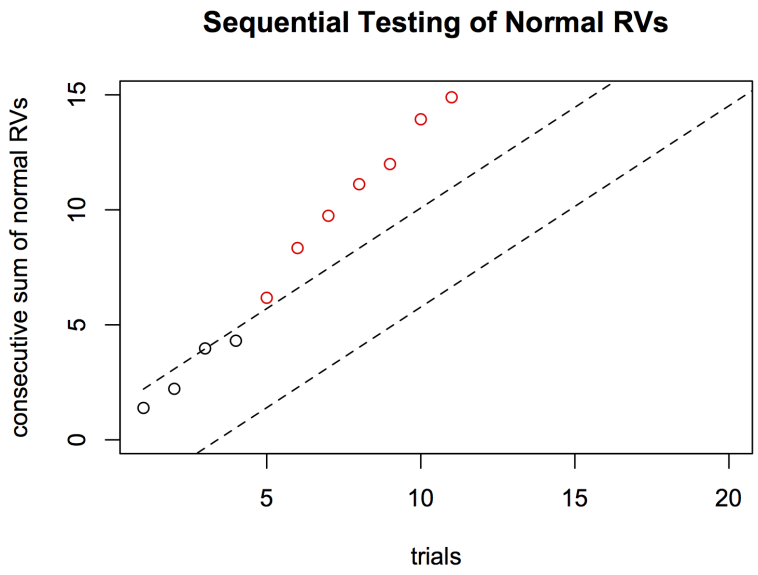
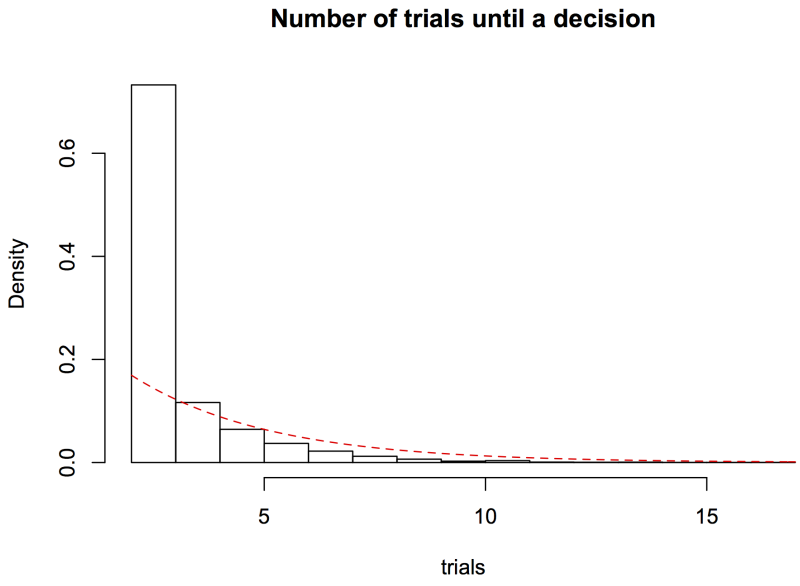


Figure 2: Hitting Time for 10,000 simulated testing batches



Next I run simulations of 10,000 testing runs, showing that the empirical average number of trials until a hit is 3.0878. The standard deviation was about 1.64. The maximum observed number of steps until a decision was made was 17.

I plug in the value $\theta = 1.75$ to get an expected hitting time under the alternative hypothesis:

$$\mathrm{E}_{\mathrm{H}_1}[M] \approx \frac{(1-\beta)\log(\gamma_1) + \beta\log(\gamma_0)}{\theta^2/2} = \frac{(1-0.01)\times 2.985682 + 0.01 \times -4.553877}{1.75^2/2} \approx 1.9$$

This downward bias of the expected hitting time is either totally acceptable or terrible, depending on the application. I tested a few other values of θ, and found that the difference between the empirical hitting time and the expected hitting time was usually around 1.5. This is fine for many settings; but for a clinical setting or a costly destructive testing applications, having to test another 1 or 2 more units than expected before making a decision could cost thousands of dollars. With qualification, I assert that the bias is pretty good.

Example 2:
Now suppose that we want to classify industrial products from a manufacturing process as either defective or non-defective: a Bernoulli process. Let p be the probability that a product will be defective and assume that the status of each product is independent of the others. Some defective products are unavoidable, so we define two values p_L and p_H, for low and high levels of defective products. The goal is to identify whether the production of defective items is low (i.e. H_0: $p = p_L$) or high (i.e. H_1: $p = p_U$). Here I choose $p_U = 0.4$ and $p_L = 0.2$. The SPRT is formed the same as above, only here I use a Bernoulli distribution instead of a Gaussian. I define X_m as the number of 1s out of my m tests and note that, as above, crossing the decision boundary is equivalent to X_m meeting one of two linear constraints (with the details left to the reader). Then deciding that $p = 0.4$ or 0.2 is the same as having

$$\Lambda_m = \text{Boundary} \iff \log(\Lambda_m) = \log(\text{Boundary } \gamma_i) \iff$$

$$\log(\Lambda_m) = X_m \log\big(\frac{p_u/(1-p_u)}{p_l/(1-p_l)}\big) + m\log\big(\frac{1-p_u}{1-p_l}\big) = \log(\text{Boundary } \gamma_i)$$

i.e., X_m satisfying linear constraint

$$X_m = a + bm$$

with a and b being equal to

$$a = \frac{\log(\text{Boundary } \gamma_i)}{\log\big(\frac{p_u/(1-p_u)}{p_l/(1-p_l)}\big)}$$

and

$$b = \frac{\log\big(\frac{1-p_u}{1-p_l}\big)}{\log\big(\frac{p_u/(1-p_u)}{p_l/(1-p_l)}\big)}$$

Expected stopping times are calculated similarly to the derivation for example 1, and for the specified set of parameters is about 28 for p_L. Here I plot one realization for $p = 0.2$, $\alpha = 0.05$, and

$\beta = 0.05$. I observe that the decision boundary is crossed on the 25th trial (in favor of p_L).

Figure 3: Red color indicates crossing to decision p_U, blue to p_L