Multiple Hypothesis Testing and the Bayes Factor*

BY [redacted] P.L. 86-36

Secret

Derives some of the main properties of the Bayes Factor and its logarithm and discusses the application of these properties to the classical "two disjoint hypotheses" situation, and—more importantly—to the situation of N hypotheses, n of which are true where n < < N. The object is to reject as many untrue hypotheses as possible while accepting a reasonable percentage of correct hypotheses. Gives two examples of the N hypotheses situation which are of COMSEC (and possibly also general) interest.

In this paper we derive some of the main properties of the Bayes factor and its logarithm in a context which applies to many Agency statistical problems. The Bayes factor arises naturally as a result of an application of the fundamental Neymann-Pearson Lemma of classical hypothesis testing theory. With the "two hypotheses" theory in mind we consider the more important situation of N hypotheses, n of which are true with $n < < N$. Finally, we discuss two examples of the N hypotheses situation which are of considerable COMSEC interest.

1. Consider a list

$$Z = Z_1, \ldots, Z_T$$

of random variables defined on a finite sample space E, an arbitrary member of which is denoted

$$e = e_1, \ldots, e_T .$$

Suppose we have two (disjoint) hypotheses H_1 and H_2 about the list such that each hypothesis completely determines the probability law of Z (denoted P_1 and P_2, respectively). (This is not quite the way the world is around here. This will be discussed later). For notational ease, we write $P_i(e)$ for $P_i(Z = e)$, $i = 1, 2$.

[redacted] P.L. 86-36

NSAL-S-193,398

The problem is to decide which hypothesis is true. The celebrated Neymann-Pearson Lemma tells us how to proceed.
If we set

$$P_1 \text{ (reject } H_1) = a$$
$$P_2 \text{ (accept } H_1) = b$$

and fix a with the hope of minimizing b, our hopes will be realized if we perform a test of the following kind:

$$\text{Accept } H_1 \text{ if } \frac{P_1(e)}{P_2(e)} \geq c$$

$$\text{Reject } H_1 \text{ if } \frac{P_1(e)}{P_2(e)} < c,$$

where e is the observation we are presented with and c is a constant to be determined. This is intuitively quite reasonable. It simply says to accept H_1 if the probability of the observation when H_1 is true is sufficiently greater than the probability of the observation when H_2 is true. The proof is just about this simple. See reference [1]. Actually, if f is any real valued increasing function, then an equivalent procedure is:

$$\text{Accept } H_1 \text{ if } f\left[\frac{P_1(e)}{P_2(e)}\right] \geq f(c)$$

$$\text{Reject } H_1 \text{ if } f\left[\frac{P_1(e)}{P_2(e)}\right] < f(c).$$

The quantity

$$B(e) = P_1(e)/P_2(e)$$

is termed the *factor,* or the *Bayes factor,* in favor of H_1 over H_2. It is often convenient to take for the f above, the natural logarithm *ln.* The terminology is:

$$L(e) = ln\,[P_1(e)/P_2(e)]$$

is the *log factor* or *Bayes score* in favor of H_1 over H_2.

Note that no assumptions about Z (normality, independence, etc.) have been made. Still, it is possible to obtain some interesting results about B and L.

First, note that if e is regarded as an arbitrary point in the sample space rather than a fixed observation, both B and L can be considered random variables. Since

$$a = P_1(B < c)$$
$$b = P_2(B \geq c),$$

we are interested in the distributions of B (and L) when H_1 is true (the "*right case*") and when H_2 is true (the "*wrong case*"). Intuitively, we would like to have these distributions as "far apart" as possible. In practice, it is often useful to know the relationships between the parameters of the two distributions. We now consider some results in this direction. (Subscripts on the expectation operator indicate the probability law used to compute the expectation.)

FACT 1. ([2], Article 53)

a. $\mu_2 \equiv E_2 B = 1$ (Turing's Theorem)

b. $\sigma_2^2 \equiv \mathrm{Var}_2 B = \mu_1 - 1$

c. $E_2(B^n) = E_1(B^{n-1})$. In particular, $E_2(B^2) = E_1(B) = \mu_1$.

Proof:

a. $\mu_2 = \sum_e \frac{P_1(e)}{P_2(e)} P_2(e) = \sum_e P_1(e) = 1.$

b. $\mu_1 = \sum_e \frac{P_1(e)}{P_2(e)} P_1(e) = \sum_e P_1^2(e)/P_2(e).$

Also,

$$E_2(B^2) = \sum_e \frac{P_1^2(e)}{P_2^2(e)} P_2(e) = \sum_e \frac{P_1^2(e)}{P_2(e)} = \mu_1.$$

Hence, $\sigma_2^2 = \mu_1 - E_2^2 B = \mu_1 - 1.$

c. $$E_2(B^n) = \sum_e \frac{P_1^n(e)}{P_2^n(e)} P_2(e) = \sum_e \frac{P_1^{n-1}(e)}{P_2^{n-1}(e)} P_1(e) = E_1(B^{n-1}).$$

Of course, b is a special case of c. //

FACT 2. $E_1 L - E_2 L \geq 0$ (also, see reference [2], Article 1).

Proof:

$$E_1 L - E_2 L = \sum_e (\ln P_1(e) - \ln P_2(e))\,(P_1(e) - P_2(e)).$$

Consider $(\ln x - \ln y)(x-y)$ for $0 < x, y < 1$.

Then, $x < y \Rightarrow \ln x < \ln y \Rightarrow \ln x - \ln y < 0,$

$x > y \Rightarrow \ln x > \ln y \Rightarrow \ln x - \ln y > 0.$

Hence, each term in the sum is positive. Actually, we have equality iff $P_1(e) = P_2(e)$ for all $e \in E$. //

Now, FACT 2 can be strengthened. For this, we need a

LEMMA. Let $\{p_r\}_1^N$, $\{q_r\}_1^N$ satisfy $p_r > 0$, $q_r > 0$ for all r and $\sum p_r = \sum p_r q_r = 1$. Then.

$$\sum p_r \ln q_r \leq 0 \text{ and } \sum p_r q_r \ln q_r \geq 0.$$

Equality holds iff $q_r = 1$ for all r.

Proof: See [3].

FACT 3. $E_1 L \geq 0$ and $E_2 L \leq 0$, where the inequalities are strict unless $B = 1$.

Proof:

Let $p_e = P_2(e)$, $p_e q_e = P_1(e)$.
Then,

$$E_2 L = \sum_e P_2(e)\, ln \frac{P_1(e)}{P_2(e)} = \sum_e p_e\, lnq_e \leq 0 \text{ by the Lemma.}$$

Also,

$$E_1 L = \sum_e P_1(e)\, ln \frac{P_1(e)}{P_2(e)} = \sum_e p_e q_e\, lnq_e \geq 0 \text{ by the Lemma.}$$

By the Lemma, equality holds iff $q_e = 1$ for all e. That is

$$q_e = \frac{p_e q_e}{p_e} = \frac{P_1(e)}{P_2(e)} = B(e) = 1 \text{ for all } e. \qquad //$$

We now begin adding some assumptions about Z. Recent work by [redacted] [4] makes the following facts more than academically interesting. We will discuss the normality of the log-factor later. P.L. 86-36

FACT 4. If L is normally distributed $N(\mu, \sigma^2)$, then B is said to have a lognormal distribution. In this case,

$$E\,B = e^{\mu+\frac{\sigma^2}{2}},\ E\,B^2 = e^{2\mu+2\sigma^2}.$$

Proof:
Since L is normal, its characteristic function is

$$\phi(t) = E(e^{itL}) = e^{it\mu - \frac{\sigma^2 t^2}{2}}.$$

$$E\,B = E(e^{ln\,B}) = E(e^{L}) = e^{\mu+\frac{\sigma^2}{2}}.$$

Also,

$$E(B^2) = E(e^{2\,ln\,B}) = E(e^{2L}) = e^{2\mu+2\sigma^2}.$$

(Hence, Var $B = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$.) //

DEFINITION. Let X be a random variable and $\phi_X(t) = E(e^{itX})$ its characteristic function. Then the n^{th} cumulant of X, $K_n(X)$, is defined by (if it exists)

$$K_n(X) = i^{-n} \frac{d^n}{dt^n} (\ln \phi_X(t))|_{t=0}.$$

For example,

$$K_1(X) = i^{-1} \frac{d}{dt}(\ln \phi_X(t))|_{t=0} = i^{-1} \frac{d}{dt}(E(e^{itX}))|_{t=0}$$

$$= i^{-1} E(i X e^{itX})|_{t=0} = E X.$$

Similarly, [5],

$$E X^2 = K_2(X) + K_1^2(X) \text{ (i.e., Var } X = K_2(X))$$

$$E X^3 = K_3(X) + 3K_2(X) K_1(X) + K_1^3(X), \text{ etc.}$$

Also, by definition, the expansion for $\ln \phi_X(t)$ is

$$\ln \phi_X(t) = \sum_{k=1}^{\infty} K_k(X) (it)^k / k!$$

These ideas lead to the following important

FACT 5. [6]. The cumulants of the distribution of L satisfy

$$K_1 - \frac{K_2}{2!} + \frac{K_3}{3!} - \frac{K_4}{4!} + \ldots = 0 \text{ if } H_1 \text{ is true, and}$$

$$K_1 + \frac{K_2}{2!} + \frac{K_3}{3!} + \ldots = 0 \text{ if } H_2 \text{ is true.}$$

Proof:
In the right case,

$$\phi_1(t) = \sum_e P_1(e) \exp\left\{ it \ln \frac{P_1(e)}{P_2(e)} \right\}.$$

Hence,

$$\phi_1(i) = \sum_e P_1(e) \frac{P_2(e)}{P_1(e)} = \sum_e P_2(e) = 1.$$

Now,

$$\ln \phi_1(t) = \sum_{k=1}^{\infty} K_k \cdot (it)^k / k!.$$

From the expression above for $\phi_1(i)$, we have

$$ln\ \phi_1(i) = 0 = \sum_{k=1}^{\infty} K_k \cdot (-1)^k/k!$$

A similar proof works for the wrong case (and for continuous distributions). //

FACT 6. If X is a normal random variable, then $K_n(X) = 0$ for $n > 2$.

Proof: $ln\ \phi_X(t) = i\,\mu\,t - \dfrac{\sigma^2 t^2}{2}$. //

FACT 7. If L is normally distributed $N(\mu_1, \sigma_1^2)$ in the right case and $N(\mu_2, \sigma_2^2)$ in the wrong case, then

$$\sigma_1^2 = 2\mu_1 \text{ and } \sigma_2^2 = -2\mu_2.$$

Proof: This follows immediately from the preceding two facts; however, we give the following proof (which does not require the introduction of the concept of cumulant):

Under H_1,

$$\phi_L(t) = E(e^{iLt})$$

$$= \sum_e P_1(e) \exp\left\{ i\,t\,ln\left[\frac{P_1(e)}{P_2(e)}\right]\right\}$$

Hence,

$$\phi_L(i) = \sum_e P_1(e)\frac{P_2(e)}{P_1(e)} = 1$$

Also,

$$L \sim N(\mu_1, \sigma_1^2) \Rightarrow$$

$$\phi_L(i) = e^{\mu_1 ii - \frac{\sigma_1^2 i^2}{2}} = e^{-\mu_1 + \frac{\sigma_1^2}{2}}$$

Hence,

$$ln(1) = 0 = ln\ \phi_L(i) = -\mu_1 + \sigma_1^2/2$$

Hence,

$$2\mu_1 = \sigma_1^2.$$

Similarly, under H_2,

$$\phi_L(-i) = \sum_e P_2(e)\frac{P_1(e)}{P_2(e)} = 1, \text{ and}$$

$$L \sim N(\mu_2, \sigma_2^2) =$$

$$\phi_L(-i) = e^{\mu_2 i(-i) - \frac{\sigma_2^2 (-i)^2}{2}} = e^{\mu_2 + \frac{\sigma_2^2}{2}}$$

Hence,

$$0 = \mu_2 + \frac{\sigma_2^2}{2}$$

Hence,

$$-2\mu_2 = \sigma_2^2. \qquad //$$

FACT 8. Let $L \sim N(\mu_1, \sigma_1^2)$ in the right case and $N(\mu_2, \sigma_2^2)$ in the wrong case. Then:

a. $E_1 L = 1/2\, ln(E_1 B)$.
b. $\sigma_1^2 = 1/3\, ln(E_1 B^2)$.
c. $\sigma_2^2 = ln(E_2 B^2) = ln(E_1 B)$.

Proof:

a. $ln(E_1 B) = \mu_1 + \frac{\sigma_1^2}{2} = 2\mu_1 = 2E_1 L.$

b. $ln(E_1 B^2) = ln(e^{2\mu_1 + 2\sigma_1^2}) = \sigma_1^2 + 2\sigma_1^2 = 3\sigma_1^2.$

c. $ln(E_2 B^2) = ln(e^{2\mu_2 + 2\sigma_2^2}) = -\sigma_2^2 + 2\sigma_2^2 = \sigma_2^2.$

By FACT 1, $ln(E_2 B^2) = ln(E_1 B)$. //

Definition: If $L \sim N(\mu_1, \sigma_1^2)$ in the right case and $L \sim N(\mu_2, \sigma_2^2)$ in the wrong case, then the *sigma-age* $S = (\mu_1 - \mu_2)/\sigma_2$.

FACT 9. Under the conditions of the above definition,

a. $\mu_1 = 1/2\, ln(E_1 B)$
b. $\sigma_1^2 = ln(E_1 B)$
c. $\mu_2 = -\, 1/2\, ln(E_1 B)$
d. $\sigma_2^2 = ln(E_1 B)$
e. $S = \sqrt{ln(E_1 B)}$

Proof: Sections a–d under FACT 9 follow from FACTS 7 and 8.

$$S = (\mu_1 - \mu_2)/\sigma_2 = \frac{1/2\, ln(E_1 B) - (-1/2\, ln(E_1 B))}{\sqrt{ln(E_1 B)}}$$

$$= \sqrt{ln(E_1 B)} \qquad //$$

This fact says that if L is normal, calculation of $E_1 B$ determines both right and wrong case distributions of L. Finally, we note the following relations between expected scores and the concept of entropy:

$$E_1 L = \sum_e P_1(e)\, ln \frac{P_1(e)}{P_2(e)} = \sum_e P_1(e)\, ln\, P_1(e) - \sum_e P_1(e)\, ln\, P_2(e)$$
$$= -H_1(Z) - \sum_e P_1(e)\, ln\, P_2(e)$$

where $H_1(Z)$ is the entropy of Z assuming that H_1 is true.
Similarly,

$$E_2 L = \sum_e P_2(e)\, ln \frac{P_1(e)}{P_2(e)} = -\sum_e P_2(e)\, ln\, P_2(e) + \sum_e P_2(e)\, ln\, P_1(e)$$
$$= H_2(Z) + \sum_e P_2(e)\, ln\, P_1(e).$$

If the size of the sample space is N, and $P_2(e) = \frac{1}{N}$ for all $e \epsilon E$, then,

$$E_1 L = -H_1(Z) + ln\, N$$

$$E_2 L = ln\, N + \frac{1}{N}\sum_e ln\, P_1(e).$$

Hence,

$$H_1(Z) = ln\, N - E_1 L.$$

2. It has already been remarked that the situation of two simple hypotheses is somewhat unreal from our point of view. A situation closer to reality is the following. We have a random list $Z = Z_1, \ldots, Z_T$ defined on a sample space E, an arbitrary member of which is denoted $e = e_1, \ldots, e_T$. We have N hypotheses $H_1, \ldots, H_N$ about Z with n of them being true, where $1 < n < < N$. We assume that each H_i determines two probability laws P_i and P_{-i} for Z:

$$P_i(e) = P(Z=e \mid H_i \text{ is true})$$
$$P_{-i}(e) = P(Z=e \mid H_i \text{ is not true}).$$

(In many COMSEC applications, it is intuitively reasonable to take P_{-i} to be the same for all i.) We want to eliminate as many wrong hypotheses as possible while accepting a reasonable fraction of correct hypotheses. To accomplish this, we test each H_i against $-H_i$ using the theory of the preceding section. That is, we form

$$L_i = ln\{P_i(e)/P_{-i}(e)\}.$$

If we can assume that L_i is normally distributed in both right and wrong cases, then from FACT 9, there exists

$\mu_i > 0$ such that with $\sigma_i^2 = 2\mu_i$,
$L_i \sim N(\mu_i, \sigma_i^2)$ if H_i is true and
$L_i \sim N(-\mu_i, \sigma_i^2)$ if H_i is not true.

Then,

$$P(\text{accept } H_i \mid -H_i) = 1 - F\left(\frac{c_i - (-\mu_i)}{\sigma_i}\right) \equiv b_i$$

$$P(\text{reject } H_i \mid H_i) = F\left(\frac{c_i - \mu_i}{\sigma_i}\right) \equiv a_i,$$

where F is the $N(0, 1)$ distribution function and c_i is the threshold for the test (the c which appears in the statement of the Neymann-Pearson Lemma). We fix $a_i = a$ (the same a for all i) and can then solve for the c_i above. Then the theory of section 1 indicates that we have minimized b_i given the fixed value a. In particular, if we take $a = 1/2$, as is often done, then $c_i = \mu_i$ and b_i depends upon

$$\frac{\mu_i - (-\mu_i)}{\sigma_i},$$

which is the sigma-age as defined in section 1. It is this appearance of the sigma-age which makes the concept important.

Now, after testing all of the hypotheses H_i as above, the expected number of wrong hypotheses ["Expected Wrong Case Survivors," E(WCS)] accepted is (since we assume $n < < N$)

$$\mathrm{E(WCS)} \cong \sum_{i=1}^{N} b_i$$

and the expected number of correct hypotheses ["Expected Right Case Survivors," E(RCS)] accepted is

$$\mathrm{E(RCS)} = a\,n.$$

Now, in order to determine E(WCS) as above, it is necessary to determine all of the b_i's and sum them. This would cost almost as much as doing the actual testing of the hypotheses. Hence, from a COMSEC point of view, the above expression for E(WCS) is not practically useful. We need another method to estimate E(WCS). The method usually employed is as follows (for simplicity, assume we have fixed $a = 1/2$). We find an approximation μ to the average of the μ_i's

$$\mu \doteq \frac{1}{N}\sum_{i=1}^{N} \mu_i .$$

Then we form

$$b = 1 - F\left(\frac{\mu - (-\mu)}{\sqrt{2\mu}}\right)$$

and take as an estimate to E(WCS)

$$\mathrm{E(WCS)} \doteq Nb.$$

In general, let $\mu_i = E_i L_i$ denote the expected value of L_i computed assuming that H_i is true, and $\mu_{-i} = E_{-i} L_i$ the expected value computed assuming H_i is not true. Similarly for $\sigma^2_{-i} = \mathrm{Var}_{-i} L_i$. Then the quantity

$$\{E(\mu_i) - E(\mu_{-i})\}/\sqrt{E(\sigma^2_{-i})}$$

is some sort of approximation to an expected sigma-age. We call it the *essential sigma-age*. The procedure is quite questionable, and there seems to be room for considerable investigation, theoretical and empirical, in this area. In the examples of section 4, we indicate how to determine the approximation μ.

3. In this section, we say the little that it seems to be possible to say about normality of the log factor. Normality of L is often assumed in general due to the fact that, in practice, it often turns out that B is a product of random variables. Then L is a sum of random variables, and if these random variables may be assumed to be independent and identically distributed with finite variances, then a central limit theorem will imply the approximate normality of L (see reference [5] page 431, Theorem 4A). Actually, less stringent requirements may be made of the random variables and normality in the limit may still be implied (again see reference [5] page 431, Theorem 4B).

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted]

~~SECRET~~

P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

EO 1.4.(c)
P.L. 86-36

P.L. 86-36

EO 1.4.(c)
P.L. 86-36

[redacted]

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted] ~~SECRET~~ P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

[redacted]

EO 1.4.(c)
P.L. 86-36

REFERENCES

[1] Hogg and Craig, *Introduction to Mathematical Statistics* (2nd ed., The Macmillan Co., New York, 1965).

[2] *Collected Papers on Mathematical Cryptology.*

[3] Hardy, Littlewood and Polya, *Inequalities* (Cambridge University Press, 1967).

P.L. 86-36

[4] [redacted], "The Strength of the Bayes Score, *S12 Informal Note #283* (8 September 1970).

[5] Parzen, *Modern Probability Theory and Its Applications* (John Wiley and Sons, Inc., New York, 1960).

[6] Good and Toulmin, "Coding Theorems and Weight of Evidence," *Jour. Inst. Maths Applics*, 4 (1968), 94–105.

[7] [redacted]

[8] [redacted]

EO 1.4.(c)
P.L. 86-36