Surprisal: Another Foundation for Statistics

Jim Ramsay

1. Motivation and Overview

- The theory of probability is the main mathematical foundation of statistics.
- But students and other people find probability hard to understand.
- They cannot reliably manipulate probabilities, especially when they become extreme.
- Under independence, most properties of nature that we experience add, but probabilities multiply.
- Most measures of size are unbounded, but probability is bounded above by 1.
- Would statistics be easier to understand if we change probability into something more “natural”?

- Surprisal is a measure of how surprising, attention-getting, or informative an event is.
- We use the symbol $!(A)$ for the surprisal of event A.
- Surprisal is proportional to minus the logarithm of probability

$$!(A) = -C \log P(A)$$

  where $C > 0$ is an arbitrary constant.
- We see how statistical theory might look if probability were replaced by surprisal.

2. The basic properties of surprisal

Some notation for events

- $A, B, ...$ indicate events.
- $\sim A$ indicates that event A does not occur.
- AB indicates that both A and B occur, but in no particular order.
- $A \vee B$ indicates that either A or B or both occur.
- $A|B$ indicates that A occurs when B has already occurred.
- $A \subset B$ indicates that if A occurs, then B also occurs, or that A is contained within B.

The properties of probability $P(A)$

- $0 \leq P(A) \leq 1$, probability is bounded below by 0 and above by 1
- $P(A, \sim A) = 0$: impossible events occur with probability 0
- $P(A \vee \sim A) = 1$: certain events occur with probability 1
- $P(AB) = P(A|B)P(B) = P(B|A)P(A)$: the probability of the joint occurrence of A and B can be decomposed in two ways into products of probabilities of single occurrences

Events A and B be are defined to be *independent* if $P(A)B) = P(A)$: knowing that B has occurred tells us nothing about whether B will occur. When independent prevails, we derive that $P(AB) = P(A)P(B)$.

The properties of surprisal $!(A)$

From $!(A) = -C \log P(A), C > 0$ we derive that

- $!(A) \geq 0$: surprisal is not negative.
- $!(A, \sim A)$ is, technically, illegal, but in fact can be taken to be infinity.
- $!(A \vee \sim A) = 0$: events that are certain to occur have zero surprisal.
- $!(AB) = !(A|B) + !(B) = !(B|A) + !(A)$.

If A and B are independent, then $!(AB) = !(A) + !(B)$.

Surprisal is a magnitude

- Like mass, distance, heat, energy and so on, surprisal is a measure of size that add or subtract for quantities on the same scale.
- The human brain adds and subtracts much more accurately and efficiently than it multiplies and divides.
- Because the unit of measurement of surprisal is arbitrary, as it is for other magnitudes, only ratios of surprisal remain unchanged when we change the unit.
- Zero has a definite meaning for surprisal, as well as for other magnitudes. We can't change its origin.
- Like other magnitudes, surprisal can be arbitrarily large.
- Surprisal will combine with other magnitudes through either multiplication or division.
- Surprisal is a *ratio* scale type quantity.

A standard unit for surprisal

- All magnitudes are linked to nature by defining a standard quantity that can be experimentally produced to a relatively arbitrary amount of precision, and that is given magnitude 1.
- Let A be an event that occurs with probability 1/2. Coin tosses come close, but physics can come up with better things.
- Let $!(A) = 1$. The unit of this *standard surprisal* is minus the logarithm to the base 2 of $P(A)$,

$$!(A) = -\log_2 P(A).$$

- Five straight heads will get the attention of most of us, and has standard surprisal 5.
- Conversion of the base 10 logarithm to the base 2 logarithm is easy: Divide by 0.3.

$$\log_{10} x = (\log_2 x)/0.3$$

- Conversion of the natural logarithm to the base 2 logarithm is nearly as easy: Divide by 0.693, or roughly 0.7.
- The standard surprisal of the famous $P = 0.05$ is 4.3, a bit over the surprisal for four heads. $P = 0.01$ translates into $! = 20/3 = 6.67$.

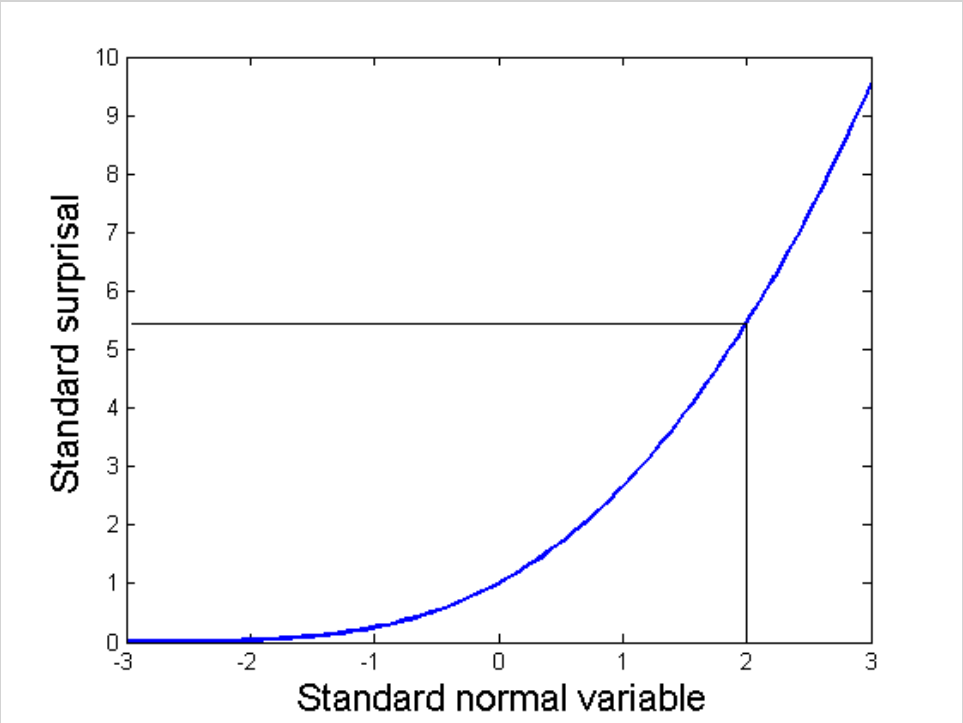
Standard surprisal and information

- The event of r heads, or any other event equivalent to a pre-defined sequence of 0'1 and 1's, requires r *bits* of information to define.
- Standard surprisal is therefore a measure of *information*, and is used in the mathematical theory of information, which is central to digital signal processing and computer design.

Surprisal functions, distribution functions, and hazard functions

- Let x be a real number, and let $A(x)$ be the event that an observed value $X \leq x$.
- The probability distribution function $F(x)$ defines the probability of this event $A(x)$ as a function of x.
- $!(x)$ is a decreasing function tending to zero, sometimes called the *survivor function*.
- $!(\sim (A(x))$ is an increasing function with a lower limit of 0.
- Moreover, $d!(\sim (A(x))/dx$ is the *hazard function*.

Standard surprisal as a function of probability of a standard normal variable exceeding a value. The probability of exceeding 2 is about 5 1/2 heads in a row.



3. Estimating surprisal from discrete data

Binary data

- If we observe an event A occurring r times in n tries, estimate $!(A)$ by $\log_2 n - \log_2 r$.
- This may seem awkward, but recall that counts are very often modelled in statistical work by constructing a linear model for the logarithm of their expected value, and, in this case, the model is a difference.

Multinomial data

- Suppose that K events are possible, and that the numbers of times n_k that event k is observed have a multinomial distribution with probabilities $p_k(\theta)$ determined by a parameter vector θ. Then the negative log likelihood of for these frequencies and parameter θ is

$$-\ln L = C \sum_{k}^{K} !(k|\theta) n_k$$

  and the maximum likelihood estimate of θ maximizes the cosine of the positive angle between the data and the model expressed in surprisal terms.

4. Surprisal for continuous data

- The probability density function $f(x)$ for a continuous random variable x is not a probability, but, at least formally, $f(x)dx$ is.
- The surprisal analogue is a function $s(x)$ called the *surprisal density function* indicating the density of surprisal for values of x.
- Surprisal density is related to probability density by

$$s(x) = -C \log f(x) + D$$

  where D is an arbitrary constant and where the logarithms to the base 2 would also be "natural".
- That is, surprisal density is on an *interval* scale.
- Those nasty normalizing constants in $f(x)$ disappear from the scene.

Gaussian surprisal

- If Gauss had thought in terms of surprisal, which would have been most uncharacteristic for him, he would have proposed

$$s(x|\mu) = [\frac{(x-\mu)}{\sigma}]^2 + 2\log\sigma$$

  as the surprisal density.

- Maximum likelihood estimation of a parameter vector θ defining expectation $\mu_i(\theta)$ with Gaussian error would have involved minimizing

$$SSE = \sum_i^N [x_i - \mu_i(\theta)]^2$$

Scaling issues for surprisal density

- Surprisal density $s(x)$ should be invariant with respect to changes in the unit for x. This may require revising the standard expression for the probability density function.
- Gaussian surprisal density is scale invariant with respect to x since re-scaling σ shifts by a constant.
- Surprisal density for the gamma distribution

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x/\tau}$$

leads to, using the natural logarithm,

$$s(x) = (\alpha - 1)[\log x - \log \Gamma(\alpha)] - \frac{x}{\tau}$$

- On the other hand, the unit for the vertical axis, surprisal itself except for a possible shift, is arbitrary.

Maximum likelihood and Bayesian estimation

- This is even simpler: we minimize

$$S(\theta|x) = \sum_i^N s(x_i|\theta),$$

  or the total surprisal density of the data. Again, note that normalizing constants are irrelevant.

- Let $r(\theta)$ be a prior surprisal function indicating how surprised we would be by various values of θ.

- Posterior surprisal is clearly

$$t(\theta|x) = S(\theta|x) + r(\theta)$$

Expectation in surprisal terms

- The expectation of $g(x)$ is defined as

$$E[g(x)] = \int g(x) f(x) dx$$

- But instead, we can make an equivalent definition: $E[g(x)] = \mu_g$ minimizes

$$\int [g(x) - \mu_g]^2 e^{s(x)} dx$$

- The solution is

$$E[g(x)] = \frac{\int g(x) e^{s(x)} dx}{\int e^{s(x)} dx}$$

- Exponentiation of surprisal density is permitted because $s(x)$ is scale invariant along the x-axis.

- This makes obvious the equivalency between expectation, which an idempotent operator, and projection in the metric of $e^{s(x)}$.
- That is, expectation is a Hilbert Space concept, and tied, whether we like it or not, to Gaussian surprisal.