

Competition between choices over time

Sean Wallis, Survey of English Usage, University College London
January 7 2010

Measuring choices over time implies competition between alternates. This is a fairly obvious statement. However, some of the mathematical properties of this system are less well known. These inform the expected behaviour of observations, helping us correctly specify null hypotheses.

The proportion of $\{shall, will\}$ utterances where *shall* is chosen, $p(shall | \{shall, will\})$, is in competition with the alternative probability of *will* (they are mutually exclusive) and bounded on a probabilistic scale. The probability associated with each member of a set of alternates $\mathbf{X} = \{x_i\}$, which we might write as $p(x_i | \mathbf{X})$, is **bounded**, $0 \leq p(x_i | \mathbf{X}) \leq 1$, and **exhaustive**, $\sum p(x_i | \mathbf{X}) \equiv 1$.

A bounded system behaves differently from an unbounded one. Every child knows that a ball bouncing in an alley behaves differently than in an open playground. ‘Walls’ direct motion toward the centre. In this section we discuss two properties of competitive choice: the tendency for change to be S-shaped rather than linear, and how this has an impact on confidence intervals.

The S curve

On the floor the only way is up. In an empty lily pond a plant initially colonises the pond at an exponential rate. However the pond is finite, and this places an upper bound on growth. Once the pond starts to fill, the reduction in available light and space causes the rate of growth to slow. This pattern of behaviour may be approximated by the well-known ‘S curve’ (properly known as the sigmoid or *logistic curve*) shown in Figure 1, where p represents the proportion of the pond covered by the plant, and t time.

This curve has the formula $p = 1 / (1 + e^{-kt})$ where k is a gradient constant. Varying k obtains different gradients or growth rates (Figure 1).

The line accelerates from the lower edge and decelerates when approaching the top edge. Rather than attempt to fit observations to an expected straight line, in bounded competitive systems we may use *logistic regression*, i.e. fitting data to a logistic curve.¹

Our example water lily competes with its environment. In other evolutionary models, predator and prey species may compete, or multiple species compete for the same niche. Similarly, two alternate forms, *shall* and *will*, performing the same linguistic function, compete with each other. All other things being equal, if *shall* is in a small minority it will tend to refuse to die out altogether. We say it becomes marked in use, which is another way of saying speakers become aware of using the form due to its rarity. On the other hand, fashionable neologisms or usages may spread very rapidly through the language until they become the norm.

The S curve is merely a mathematical model that predicts behaviour. It is not an iron law and

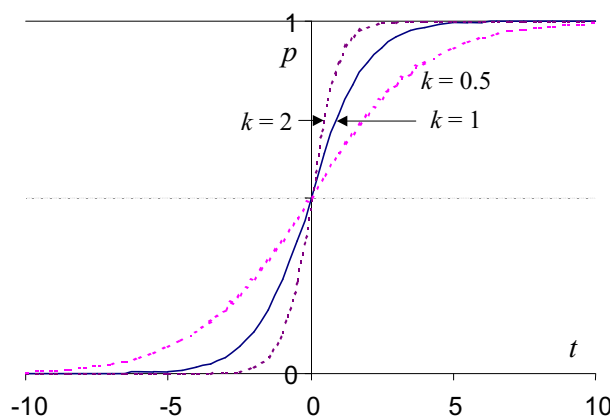


Figure 1: Example logistic curves.

¹ Logistic regression (Hilbe 2009) attempts to fit data to the logistic curve $\text{logit}^{-1}(z) \equiv 1 / (1 + e^{-z})$ where $z = \sum m_i x_i + c$ and m_i is the *logistic coefficient* (weight) for variable x_i ; c is the z -intercept. This formula for z is a straight line, i.e. it assumes that x_i are independent. The logistic function $\text{logit}^{-1}(z)$ maps this line (cf. Figure 3(b)) to an S curve.

does not determine the actual outcome! Neologisms may die out as fast as they spread, for example, if they are driven out by other alternates. If more than two alternate forms are competing, the behaviour may be more complex because two forms may be increasing in usage at the same time. The system may oscillate. Minority species may ‘bounce back’.

This does not undermine the central argument here. *All other things being equal* we may expect this behaviour. We investigate when **observed** behaviour deviates from the **expected** pattern. Unexpected behaviour is worthy of explanation.

Poisson error bars

Not only is expected behaviour ‘S shaped’, but the expected variance for any observation must also be skewed. Confidence intervals, or ‘error bars’, are conventionally calculated using the Binomial approximation to the Normal (or ‘Gaussian’) distribution with parameters *mean* $\bar{x} \equiv p$, *standard deviation* $s \equiv \sqrt{p(1-p)/n}$.

This approximation is broadly acceptable for a high n and p is not close to the boundary. 95% of the data (Figure 2(a)) will fall within the confidence interval $(\bar{x} - z.s, \bar{x} + z.s)$ where z is the critical value of the Normal distribution (approximately 1.96).

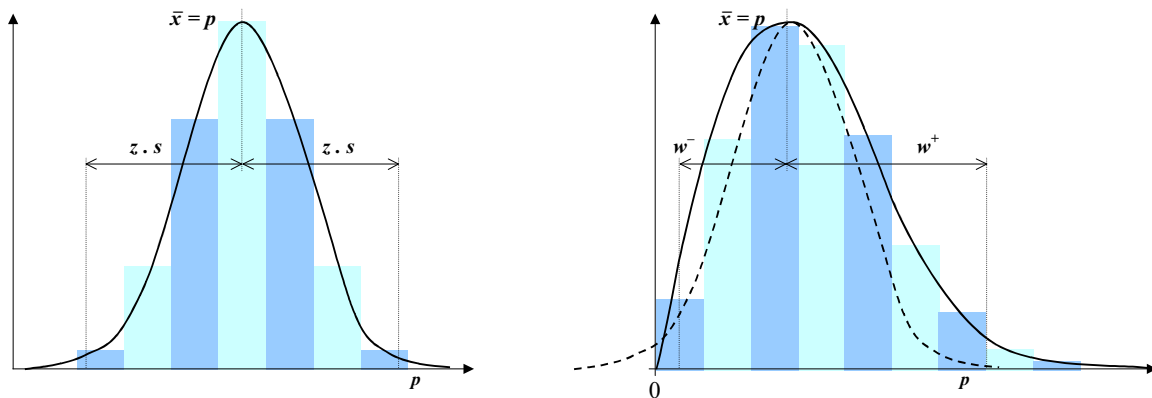


Figure 2: (a) Normal and (b) Poisson distributions.

The Normal distribution ceases to fit Binomial probabilities when p is near the boundary (0 or 1) or n is small. The resulting curve, Figure 2(b), is a ‘Poisson’, or skewed Normal distribution.²

Since p cannot fall below zero, as p approaches 0, error bars also become skewed. This adjusted confidence interval is known as the *Wilson score confidence interval* (Wilson 1927).

This can be written as

$$Wilson\ interval \equiv (w^-, w^+) \equiv \left(p + \frac{z^2}{2n} \pm z \sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}} \right) / \left(1 + \frac{z^2}{n} \right).$$

This formula tends to ‘pull’ the error bars towards the centre in a similar way to the S curve. The centre of the interval, which we might call p' , is a weighted mean of p and 0.5 (weighted by z^2/n).

² Gould (1996) applies this argument to evolutionary biology – that simple organisms may evolve into more complex ones and regress, but cannot be simpler than single-cell organisms. The result is a Poisson distribution of biomass over complexity (1996: 171) due to there being an evolutionary ‘floor’ at cells $c = 1$. Provided *some* species (not necessarily all) evolve into more complex forms, at any point in time the overall distribution will be skewed.

S curves and Wilson intervals

We can sketch the overall behaviour of the system by plotting Wilson intervals for the ‘S’ curve (Figure 3(a)). We have plotted a logistic curve for p ($k = 1$) and added intervals for $n = 10$ and $n = 100$. Observe that with a small n the confidence interval is large and more highly skewed. The difference ($w^+ - w^-$) is greatest for $p = 0.5$.

We have witnessed the asymmetry of the Poisson distribution. The Wilson interval (w^-, w^+) is skewed about p . Newcombe (1998: 870) notes that on a *logit scale* (Figure 3(b)) the interval is symmetric.

The ‘logit’ is the inverse logistic function, i.e. $\text{logit}^{-1}(p) = 1 / (1 + e^{-p})$, and can be expressed as $\text{logit}(p) \equiv \log(p / 1 - p) = \log(p) - \log(1 - p)$. On this surface the logistic curve for p (with $k = 1$) becomes the straight line $y = x$.

The differences either side of p are equalised. Newcombe shows that, provided $p \in (0, 1)$,

$$\text{logit}(p) - \text{logit}(w^-) = \text{logit}(w^+) - \text{logit}(p).$$

Imagine Figure 3(b) as if it were plotted on the surface of a U-shaped ‘trench’ (Figure 4), where time flows along the trench. Looking down on this trench from above reveals Figure 3(a). We have a system bounded by ‘walls’ at $p = 0$ and 1 , directing motion inwards to $p = 0.5$.

In conclusion, a logistic scale provides a model of competition over time within a system by summing numerous local interactions (random variation, growth, decline) into a pattern of overall behaviour that is bounded probabilistically.

This has two consequences: a steady change adopts a logistic curve rather than a straight line, and variation about the change also tends to be skewed towards the centre, causing it to adopt a Poisson rather than a Normal distribution.

References

- GOULD, S. J. 1996. *Life’s Grandeur*. London: Random House.
 HILBE, J. M. 2009. *Logistic Regression Models*. Chapman & Hall/CRC Press.
 NEWCOMBE, R.G. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**: 857-872.
 WILSON, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**: 209-212.

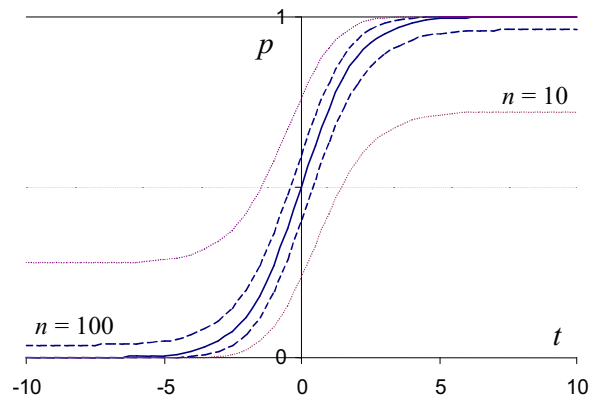


Figure 3(a). Logistic curve ($k = 1$) with Wilson score intervals for $n = 10, 100$.

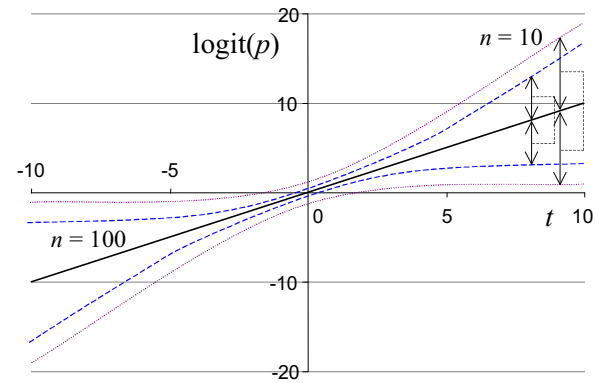


Figure 3(b). Graph (a) on a logit scale.

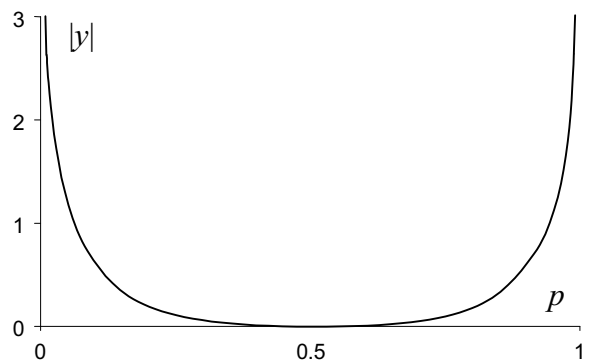


Figure 4. Logit cross-section³ folding an infinite plane into a probabilistic trench.

³ The logit curve over p is the inverse of the logistic curve in Figure 3(a). The tangent at the intercept $(y, p) = (0, 0.5)$ is $4p$, so we can adjust for time by subtraction: $y = \text{logit}(p) - 4p + 2$. Plotting absolute values of y obtains Figure 4.