

Re: Distribution of a vowel on the page

Source: <http://sci.tech--archive.net/Archive/sci.stat.math/2007-10/msg00222.html>

- *From:* Richard Ulrich <Rich.Ulrich@xxxxxxxxxxxx>
 - *Date:* Mon, 22 Oct 2007 12:03:53 -0400
-

On Sun, 21 Oct 2007 20:36:40 -0500, David Winsemius
<doe_snot@xxxxxxxxxxxx> wrote:

Richard Ulrich <Rich.Ulrich@xxxxxxxxxxxx> wrote in
<news:t9snh31bag2npc1bcrqq3s6ho30frufj0g@xxxxxxxx>:

On Sun, 21 Oct 2007 14:07:26 -0700, luca.pamparana@xxxxxxxx wrote:

Hi everyone,

I have a statistics question. We were doing Poisson distribution in our class today and the discussion topic is whether the number of times a vowel occurs on a line follows Poisson distribution or not. I think that it will not follow Poisson distribution as the probability of a vowel occurring is more than the probability for a consonant occurring. However, most people are of the opinion that it should follow Poisson distribution because regardless of the probability it should still be distributed randomly.

I did some test and took a page of paper from a normal book and did some stats on it. I did not get it to look like Poisson distribution. I am wondering which reasoning here is more reasonable.

In a book in English?

There are two reasons that the number of vowels on a line

Re: Distribution of a vowel on the page

will not follow Poisson. First, Poisson describes a relatively *rare* event when the circumstance is otherwise binomial (like this). Second, the events would have to be *independent*.

It is true that the Poisson is a good approximation to the binomial when the rate parameter is small, but the Poisson distribution may also be good when even when the rate parameter is not small. The question is really only answerable by reference to the data.

I get it, if you are talking about the Poisson rate parameter by itself, since that can be an arbitrary counter.

– Can you show me where a *binomial* rate parameter is near or above 50% and results in Poisson appearance?

But consonants and vowels are structured by words, and words seldom start or end with more than 3 (say) consonants or vowels. What you observe will have almost no instances of more than 6 -- of *either* consonants or vowels. And one of those has $p > 0.5$, so the event should not be enormously rare.

The average number of vowels per line must be around 20–25, so your statement about "almost no instances of more than six" does not make much sense to me. The next sentence makes even less sense. I am guessing that

Ah. I needed to be more clear. I was going on about the *dependency*. If the occurrences are independent, then there will be no "correlation" between *consecutive* occurrences.

Clearly, vowels and consonants in words violate that assumption. If there were a positive correlation (consecutive repetitions), the distribution, if otherwise Poisson, would be over-dispersed (variance too large for the mean). Since there is a negative correlation, it will be under-dispersed.

It occurs to me that if there is a positive correlation of having words with many-versus-few vowels, then the between-word r might tend to offset the negative correlation within words. I don't know how that would work out. But the OP already stated that the empirical distribution he obtained did not look Poisson.

Re: Distribution of a vowel on the page

you took the OP to be asking about was the number of one _specified_ vowel and were not seeing the counts as _per_line_ but per word. At a rate of 25 vowels per line the Poisson is going to be distributed pretty much as a Normal variable with mean=25 and s.d.=5. You should let the data determine whether the added flexibility of a two parameter distribution like the Normal is needed. I think the OP should provide his data and we can look it over.

--

Rich Ulrich, wpilib@xxxxxxxx

<http://www.pitt.edu/~wpilib/index.html>