QUALICO 2020, Tokyo, JAPAN, September 10-13, 2021, v-final

Probabilistic frequency applied to Spanish diachronic data

Hiroto Ueda (University of Tokyo)
Antonio Moreno Sandoval (Autonomous University of Madrid)

Keywords: absolute frequency, relative frequency, normalized frequency, probabilistic frequency, Spanish medieval bilabial spellings, combination of preposition and article in Spanish.

Summary

Ueda and Moreno (2017) formulate a reliable and robust frequency based on the binomial probability distribution. On this occasion, we present its mathematical foundations along with its applications to the Spanish diachronic data.

In corpus linguistics, it is usual practice to obtain two-dimensional frequency tables that group the linguistic forms and their variation through the attribute (time, space, style, etc.). This data is obtained by a multiple search (several forms simultaneously) with various attributes (for example, years or geographical areas). The presentation of the quantitative data is usually offered in the form of absolute frequency (AF) with the addition of relative frequency (RF) or normalized frequency (NF) to ensure the feasibility of comparison. However, neither the relative frequency nor the normalized frequency is adequate to compare the figures with very different bases or small bases, as needed. For example, 3 out of 3 (RF: 100%) presents the figure higher than 8 out of 12 (RF: 66.7%), although we intuit that the probability of 3 out of 3 is less important than that of 8 out of 12, and much less important than that of 80 out of 120.

To deal with the problem of lack of reliability and comparability in absolute (AF), relative (RF) and normalized frequencies (NF), we have introduced the concept of binomial probability to calculate "probabilistic frequency" (PF). To get this frequency, first, the expected probability (*e*) is calculated from the absolute frequency (*x*), the basis (*n*) and the desired security

level (*s*) of, for example, 95% or 99% (with 5% or 1% risk). To illustrate, we apply the method of probabilistic frequency to two diachronic questions of the Spanish language. First, we analyze the three variant spellings <u>, <b> and <v> of the lemma «voz» ('voice'). This case is interesting for the history of the language since current Spanish spelling <v> does not represent a labiodental consonant: [voθ], but a bilabial: [boθ], unlike other European languages (Ueda 2018). Next, we discuss combinations of the preposition with the definite article in forms of *del, dela* ('of the'), *al, ala* ('to the'), etc., of which only *del* and *al* have been maintained in modern Spanish, unlike other Romance languages (Ueda 2017).

1. Introduction[1]

In the scientific studies of natural, social or individual phenomena, it is essential to know their frequency to measure their quantitative importance. The frequency is distinguished between absolute (AF) and relative (RF). In our view, the relative frequency should be divided between the partial relative frequency (PRF) and the total relative frequency (TRF). With the partial relative frequency, we refer to the proportion that an element occupies within the sum of treated items, for example, the percentage of the Spanish masculine form of the singular definite article within the five alternative forms: *el, los, la, las, lo* ('the'). On the other hand, if we are interested in the frequency of the English pronoun *we*, in a corpus of political propaganda, compared to other types of documents, we resort to the total relative frequency, which is relativized within the totality of all words used in each document. Usually, the same base of one thousand words or one million words is used.

In this study, we use the term "relative frequency" (RF) for partial relative frequency and "normalized frequency" (NF) for total relative frequency (TRF)[2]. We believe that this distinction is essential since the relative frequency (RF) is the proportion that the element in question occupies within a selected

[1] We thank José Antonio Jiménez Millán, University of Cádiz, Leonardo Campillos, Consejo Superior de Investigaciones Científicas (CSIC), and Hiroshi Kurata, University of Tokyo, for their help in the preparation of this study from their speciality in physics and computer science, natural language processing, and mathematical statistics, respectively.

[2] For 'normalisation', see Evison (2012: 126). McEnery and Hardie (2012: 49) treat equally 'normalized frequency' and 'relative frequency'.

group, with a limited range of 0 to 1 ([0, 1]). In contrast, the normalized frequency (NF) is referred to as the quantitative magnitude appreciated within a population of one thousand or one million words.

Both frequencies have a common characteristic, which is the division by the sum of the frequencies of the selected group (RF), or of the totality of the corpus (NF), which causes a severe problem when comparing the numerical magnitudes, as we will see in the next section 2. We will solve it in the form of probabilistic frequency (PF) in section 3.

2. Three types of word frequency

Let us observe the actual data of the absolute frequency (AF) of the three forms with spelling variation, *uoz, voz, boz,* in 50-year time sequences from 1200 to 1400[3]:

AF	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**3**	**8**	**3**	**11**	**6**
boz	0	3	8	18	35
voz	0	1	1	23	53
Sum	3	12	12	52	94

RF (%)	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**100.0**	**66.7**	**25.0**	**21.2**	**6.4**
boz	0.0	25.0	66.7	34.6	37.2
voz	0.0	8.3	8.3	44.2	56.4
Sum	100.0	100.0	100.0	100.0	100.0

Table 1: Absolute frequency (AF) and relative frequency (RF) of *uoz, boz, voz.*

The absolute frequencies are not comparable since the sums of the three forms are very different in each time range: {3, 12, 12, 52, 94}. For example, the number 3 of *uoz* in 1200 is not directly comparable with the number 8 in 1250. In this case, to compare the absolute frequencies on the same basis, researchers resort to relative frequency (RF), which is calculated by dividing the absolute frequency (AF) by the sum, for example, 3/3 = 1.000, 8/12 = .667. If we multiply the relative frequency by 100, we arrive at the percentage: 1.000 * 100 = 100 (%), 0.667 * 100 = 66.7 (%).

However, neither the relative frequency (RF) nor the percentage (%) is adequate to compare the figures with very distant or reduced bases. For example, 3 in 3 (RF: 1.000, 100%) presents the figure higher than 8 in 12 (RF: 0.667, 66.7%), although we think and intuit that the probability of 3 in 3 is less

[3] The table has been obtained on the site of «CODEA in LYNEAL» (GITHE, 2015), with the selection of Castilla la Vieja region:
http://shimoda.lllf.uam.es/ueda/lyneal/codea.htm [11/25/2019]

important than that of 8 in 12 and much less important than 80 in 120. We believe that the percentage serves to describe the proportion that each case occupies within the set. However, it does not serve to compare each case between several sets with quite different bases (populations)[4]. Later (3.1, 3.2, 3.3), we will look for the solution to the numerical evaluation problem, typical of the relative frequency (RF) and the percentage.

Now, let us look at the problem of another type of frequency also used in corpus linguistics in general: normalized frequency (NF), which is calculated by the division of the absolute frequency (AF) by the totality of words (T) counted in each section, multiplied by an appropriate multiplier (*m*):

NF = AF / T * *m*

For example, in the 1200 band of the corpus, 7,736 words have been counted, which is a total of words (T). So, the normalized frequency of *uoz* in 1200 is 3 / 7,736 * 100 000 = 38.8. We recommend using as a multiplier (*m*) the rounded number (100 000) near the maximum of the base (T): 96,059 (in the data set of 1400). We get the lower right table (NF):

AF	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
T	7,736	36,052	40,957	64,999	96,059

NF.	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**38.8**	**22.2**	7.3	16.9	6.2
boz	0.0	8.3	19.5	27.7	36.4
voz	0.0	2.8	2.4	35.4	55.2

Table 2: Absolute frequency (AF) and normalized frequency (NF) of *uoz, boz, voz.*

However, here the normalized frequency (NF) also presents the same lack of comparability in the data of quite different bases, especially with some minimal bases. We cannot help but doubt about the NF figure of *uoz* in 1200, 3 among 7,736 whose NF is 38.8 compared to the NF in the same way *uoz* in 1250, 8 among 36,052, whose NF is 22.2. We wonder if 38.8 is comparable directly with 22.2 in NF.

The essence of the problem is the same in both the relative frequency (RF) and the normalized frequency (NF) in the sense that the two calculate on

[4] Wong (2013: 107) recommends not comparing a percentage of the data with a different size: "Don't compare percentage changes for two entities that are not comparable in size.".

very different bases. Paradoxically, the two frequencies (RF, NF) are used precisely when the bases are different, since if the bases are equal, it is not necessary to resort to these frequencies and, therefore, with the net absolute frequency (AF), we can make the numerical comparison perfectly. The problem occurs when the bases are considerably different.

The problem of the lack of comparability discussed here is solved usually by the juxtaposition of absolute and relative (or normalized) frequencies (Table 1, 2) or by the elimination of the less representative set due to a lack of data (Wong 2013: 107). In reality, juxtaposition is not a solution or evaluation but rather an exposition or description. In the elimination, for example, in the data of the three alternating medieval forms, one would try to exclude the set corresponding to 1200. It is the general practice in numerical analysis. In the baseball sports world, the scores of players with sufficient participation in the matches are calculated. Players who do not pass the previously established participation threshold are excluded from the evaluation. However, we wonder how the threshold is established. We do not know what to do with the 1250 range, where frequencies (total: 36,052) are recorded within the base of almost one-third of 1400 (total: 96,059).

3. Probabilistic frequency

We propose to treat all data without distinction, with standard probability criteria. Our purpose is to look for a new type of frequency, "probabilistic frequency" (PF), that represents the relative value of the absolute frequency (AF) within the set (base) with simple calculations of the probability (Ueda / Moreno Sandoval. 2017). To prove it, we resort to binomial probability. The path to get to know the probabilistic frequency goes through the following three previous steps: (1) security (*s*), (2) expected probability (*e*) and (3) multiplier (*m*).

3.1. Security

We know that, for example, a player who has scored 28 goals in 100 games is more "important" and he has contributed more to the team than the other who has scored 3 goals in 10 games, although the first goal ratio (28 / 100 = 28%) is less than the second (3/10 = 30%). For the degree of importance, we use the concept of "security" (*s*). We start from a few simple and special cases to arrive at the case applicable to frequencies in general.

For calculating the security (*s*), we use the binomial probability[5]. To understand it, we start with some examples as simple as a coin toss, where each face has the expected probability (*e*) of 0.5 (50%).

The following table shows the probabilities of two events (x = 0, 1) of a coin toss (number of trials, n = 1): obverse (x = 1), with value of 1, or reverse, with value of 0 (x = 0). Each event's expected probability (*e*) is 0.5 since there are two possibilities of the same probability: obverse or reverse. In Table 3, each event comes with its own occurrence probability (O), which we have just seen, cumulative probability (C), which is accumulated with each corresponding occurrence probability and security (s = S(x, 1, 0.5)):

X	**Case**	**O(*x*, 1, 0.5)**	**C(*x*, 1, 0.5)**	**S(*x*, 1, 0.5)**
x: 0	(0)	1/2 = 0.5	0.5	0
x: 1	(1)	1/2 = 0.5	0.5 + 0.5 = 1.0	0.5

Table 3: Security (*s*) in one trial (*n* = 1).

The occurrence probability column (O) shows in the first row the probability of reverse (x = 0), with O(0, 1, 0.5) = 1/2 = 0.5 and, in the second row, the obverse (x = 1) with O(1, 1, 0.5) = 1/2 = 0.5. The cumulative probability (C) of x = 0, C(0, 1, 0.5), is 0.5, which is equal to O(0, 1, 0.5), and that of x = 1, C(1, 1, 0.5), is 1.0, which is the sum of O(0, 1, 0.5) = 0.5 and O(1, 1, 0.5) = 0.5. The last cumulative probability (C) is always 1.000.

Now, we define the "security" (*s*) as corresponding to the cumulative probability (C) of x - 1:

$$s = S(x, n, e) = C(x - 1, n, e)$$

(x: occurrence; n: trials; e: expected probability)

The security (*s*) of x = 0, we define it as 0, because there is no cumulative probability (C):

$$S(0, n, e) = 0$$

which means that there is 100% risk probability in occurring x = 0 and x = 1.

[5]. We have consulted the method of binomial test (Ichihara 1990: 18-21; Kiyokawa 1990: 94-95). For the binomial probability distribution, see, for example, Bishir and Drewes (1970: 510-523) and Fleming and Nellis (1994: 93-102).

We consider security (*s*) as the cumulative probability of the occurrence of the immediately preceding case because the sum of the probabilities of the corresponding occurrence and superior cases corresponds to the risk probability. If we toss a coin, the security of the occurrence of 1 (obverse) is 0.5, which is complementary to the risk (not obverse, that is, reverse), which is also 0.5. Therefore, security + risk = 1, which means that there is 0.5 (50%) security of the appearance of the obverse, and there is a 0.5 (50%) risk (reverse) in the case of S(1, 1, 0.5). In other words, if we bet on the appearance of the obverse, there is a 50% risk (and 50% security), which we know and intuit without resorting to the theory of probability.

So far, we have seen a straightforward case in which we throw the coin only once. What happens if we throw the same coin twice? The following table shows the distribution of occurrence probability (O) presented in two trials of tossing a coin (n = 2). There are three possible cases: x = 0, 1, 2, that is, {(0,0)}, {(1,0), (0,1)} and {(1,1)}:

X	**Case**	**O(*x*, 2, 0.5)**	**C(*x*, 2, 0.5)**	**S(*x*, 2, 0.5)**
x: 0	(0,0)	1/4 = 0.25	0.25	0
x: 1	(0,1); (1,0)	2/4 = 0.50	0.25 + 0.50 = 0.75	0.25
x: 2	(1,1)	1/4 = 0.25	0.75 + 0.25 = 1.00	0.75

Table 4: Security (*s*) in two trials (*n* = 2).

This time the expected probability (*e*) of obverse is also 0.5. The occurrence probability (O) column shows that the O of 0 obverse occurrences, O(0, 2, 0.5), is 0.25 (reverse, reverse) = (0, 0), that is 1 of 4 cases. The total cases are 4, because there are 4 following cases: {(0, 0), (0, 1), (1, 0), (1, 1)}. The probability of 1 occurrence of obverse, {(obverse, reverse) , (reverse, obverse)}; {(1, 0), (0, 1)} is O(2, 2, 0.5) = 0.5 (2 of 4 cases). And finally, the probability of 2 occurrences of obverse, (obverse, obverse), (1, 1), O(2, 2, 0.5) is 0.25, which occurs 1 of 4 cases. The cumulative probability (C) column presents the probabilities added from 0 to 2 in each occurrence: x = 0, 1, 2.

The security column (S) corresponds to the previous case of cumulative probability (C). The last security (S) of x = 2 is S(2, 2, 0.5) = 0.75, which represents a considerable increase over the previous experiment, in which the coin was only thrown once: 0.5 (n = 1), which means that the probability of 2 in 2 (S = 0.75) is much more "significant" (essential, important) than that of 1 in 1 (S = 0.5), although both are equal to 100% cumulative probability (C). However,

security (*s*) still does not reach more than 0.75 (75%), which means that there is 0.25 (25%) of risk, which can occur in pure randomness. Now, we need the three parameters: *x*: occurrences, *n*: total of the trials and *e*: expected probability (E):

s = S(2, 2, 0.5*)* = C(1, 2, 0.5) = 0.75

In the same way, the security(*s*) of *x* = 1 is:

s = S(1, 2, 0.5*)* = C(0, 2, 0.5) = 0.25

Let us look at the experiment of three trials (*n* = 3):

X	**Case**	**O(*x*, 3, 0.5)**	**C(*x*, 3, 0.5)**	**S(*x*, 3, 0.5)**
x: 0	(0,0,0)	1/8 = .125	.125	0
x: 1	(1,0,0), (0,1,0), (0, 0, 1)	3/8 = .375	.500	.125
x: 2	(1,1,0), (1,0,1), (0,1,1)	3/8 = .375	.875	.500
x: 3	(1,1,1)	1/8 = .125	1.000	.875

Table 5: Security (*s*) in three trials (*n* = 3).

The security (*s*) of the last occurrence (*x* = 3) has increased to 0.875 and therefore the risk has now decreased into 0.125 = 1 - 0.875.

s = S(3, 3, 0.5*)* = C(2, 3, 0.5) = 0.875 (87.5%)

If we bet that the obverse does not come out 3 times in 3 trials, there is an 87.5% chance of winning the bet, which is the security (*s*); and the risk of losing the bet is 12.5%. Therefore, we should increase security to at least 95% (0.95) and, if possible, up to 99%, with 5% or 1% risks, respectively. In this way, we lose the bet only 1 of 20 times (5%) or 1 of 100 times (1%).

Let us look at the experiment of 10 trials (*n* = 10):

X	**O(*x*, 10, 0.5)**	**C(*x*, 10, 0.5)**	**S(*x*, 10, 0.5)**
x: 0	.001	.001	.000
x: 1	.010	.011	.001
x: 2	.044	.055	.011
x: 3	.117	.172	.055
x: 4	.205	.377	.172
x: 5	.246	.623	.377
x: 6	.205	.828	.623
x: 7	.117	.945	.828

x: 8	.044	.989	.945
x: 9	**.010**	**.999**	**.989**
x: 10	**.001**	**1.000**	**.999**

Table 6: Security (*s*) in ten trials (*n* = 10, *e* = 0.5).

Finally, when *x* = 9, we obtain security (*s*) by S(9, 10, 0.5*)* = 0.989, greater than 95%, and S(10, 10, 0.5*)* = 0.999, greater than 99%, which means that we can present the figure of 9 between 10 with security (*s*) greater than 95%, and 10 between 10 with security (*s*) greater than 99%. Actually, when tossing the coin 10 times, if the obverse of the coin comes out 9 times, the total occurrence probabilities less than 9 {0, 1, 2, ..., 8} adds up to 98.9%, which is quite significant. That is, with the 98.9% security, we can affirm that 9 out of 10 is significant (important). It is significant in the sense that the frequency of 9 times or 10 times of obverse out of 10 trials of coin toss occurs only with the risk probability of 0.010 + 0.001 = 0.011 (1.1%). In the same way, we can affirm that 10 out of 10 has security of 0.999 (99.9%). Compare the cases of 1 in 1 (50% security), 2 in 2 (75%), 3 in 3 (87.5%), and now, 10 in 10 (99.9%).

So far, we have seen the mathematical behavior of security (*s*), which depends on the three parameters: *x*: occurrences, *n*: total and *e*: expected probability. We have observed its movement according to *x* and *n*. Now let us see what security (*s*) is presented according to the change in the expected probability (*e*). The following table shows the security (*s*) of the occurrences (*x*) of events endowed with the expected probability (*e*) of 0.1, for example, taking out the card "1" within the ten cards of {1, 2, ..., 10}, with replacement[6]:

---

[6] We use the binomial probability (B) to obtain the security (*s*):

$$B(i, n, e) = {}_nC_i \, e^i \, (1 - e)^{n - i}$$

$$s = S(x, n, e) = \sum^{[i=0,\, x-1]} B(i, n, e) = \sum^{[i=0,\, x-1]} {}_nC_i \, e^i \, (1 - e)^{n-i} = \sum^{[i=0,\, x-1]} n! \,/\, [i! \, (n - i)!] \, e^i \, (1 - e)^{n - i}$$

(*s*: security, *x*: occurrence, *n*: number of trials, *e*: expected probability)

This formula of security (*s*) is complicated, so we use the Excel BINOMDIST function instead (see Appendix, Program-A):

$$s = S(x, n, e) = \text{BINOMDIST}(x - 1, n, e, 1)$$

X	O(*x*, 10, 0.1)	C(*x*, 10, 0.1)	S(*x*, 10, 0.1)
x: 0	.349	.349	.000
x: 1	.387	.736	.349
x: 2	.194	.930	.736
x: 3	.057	.987	.930
x: 4	**.011**	**.998**	**.987**
x: 5	**.001**	**1.000**	**.998**
x: 6	.000	1.000	1.000
x: 7	.000	1.000	1.000
x: 8	.000	1.000	1.000
x: 9	.000	1.000	1.000
x: 10	.000	1.000	1.000

Table 7: Security (S) in ten trials (*n* = 10, *e* = 0.1).

For example, when $x = 5$, $n = 10$ and $e = 0.1$, S(5, 10, 0.1*)* turns out to be 0.998, that is, the sum of the occurrence probability (O) of $x = 0, 1, 2, 3, 4$ is 0.998. Therefore, when setting the security standard (*s*) at 0.99 (99%) of the occurrences correspond to 0, 1, 2, 3, 4. There is almost never 5 onwards (5, 6, 7, ...) and there is a low probability of 0.01 (1%).

The following table (Table 8) shows the securities (*s*) according to the occurrences ($x = 0, 1, 2, ..., 10$) and with different expected probability ($e = 0.1, 0.2, ..., 0.9$):

***s* = S(*x*, 10, *e*)**	***e*: 0.1**	***e*: 0.2**	***e*: 0.3**	***e*: 0.4**	***e*: 0.5**	***e*: 0.6**	***e*: 0.7**	***e*: 0.8**	***e*: 0.9**
***x*: 0**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
***x*: 1**	**0.349**	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
***x*: 2**	**0.736**	0.376	0.149	0.046	0.011	0.002	0.000	0.000	0.000
***x*: 3**	0.930	0.678	0.383	0.167	0.055	0.012	0.002	0.000	0.000
***x*: 4**	0.987	0.879	0.650	0.382	0.172	0.055	0.011	0.001	0.000
***x*: 5**	**0.998**	0.967	0.850	0.633	0.377	0.166	0.047	0.006	0.000
***x*: 6**	1.000	**0.994**	0.953	0.834	0.623	0.367	0.150	0.033	0.002
***x*: 7**	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013
***x*: 8**	1.000	1.000	**0.998**	0.988	0.945	0.833	0.617	0.322	0.070
***x*: 9**	1.000	1.000	1.000	**0.998**	0.989	0.954	0.851	0.624	0.264
***x*: 10**	1.000	1.000	1.000	1.000	**0.999**	**0.994**	0.972	0.893	0.651

Table 8: Security (*s*) in ten trials (*n* = 10, *e* = [0.1, 0.9]).

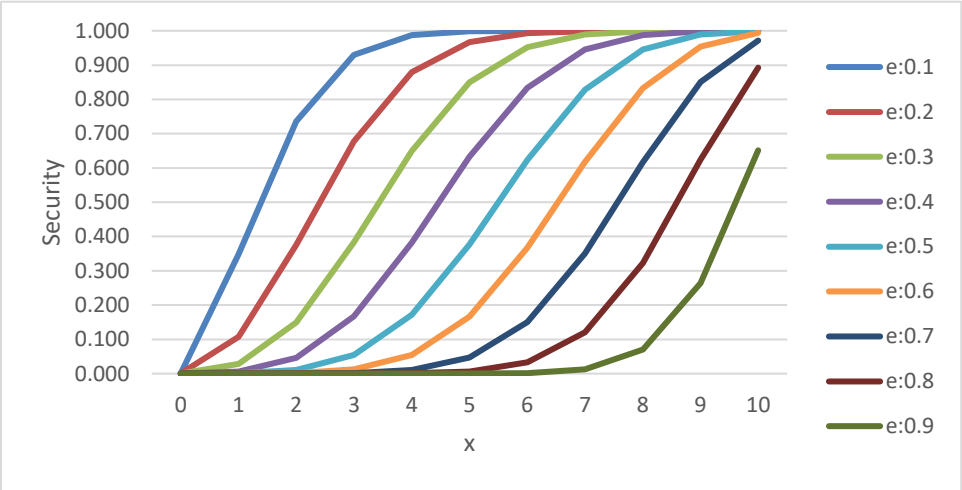

Fig. 1: Security (*s*) in ten trials. *n* = 10, *e* = [0.1, 0.9].

For example, we find 0.349 in the cell of $x = 1$; $e = 0.1$:

S(1, 10, 0.1) = 0.349

When rehearsing 10 times of the event with the expected probability (*e*) of 0.1, the occurrence 1 ($x = 1$) corresponds to the security (*s*) of .349 (34.9%). The case of 2 occurrences ($x = 2$) of the same event corresponds to 0.736 (73.6%):

S(2, 10, 0.1) = 0.736

The concept of security (*s*) applies to absolute frequency tables, for example, Table 1, which we reproduce below (AF)[7]:

AF	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**3**	**8**	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

s	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	*.963	*.981	.181	.019	.000
boz	.000	.181	*.981	.526	.758
voz	.000	.008	.008	.934	#1.000

Table 9: Absolute frequency (AF) and security (*s*).

For example, the frequency 3 in the sum 3 and 8 in 12 with the expected probability of 1/3 (0.333) have the following security:

[7] In the security table (*s*), the sign * corresponds to the figure greater than .95 and the sign #, to the figure greater than .99.

S(3, 3, 1/3) = .963 (96.3%)
S(8, 12, 1/3) = .981 (98.1%)

which exceeds statistically significant 95% ($p < 0.5$), but does not reach 99% ($p < 0.1$). In this way, we can use the security figure to carry out a statistical test: a security test. The function S(8, 12, 1/3) returns the sum of probabilities below the corresponding case ($x = \{0, 1, 2, ..., 7\}$), of occurrence endowed with the same equitable probability of three trials (rows): 1/3 (0.333), based on the assumption that the frequencies occur randomly with the same probability on three occasions, corresponding to the three forms: *uoz*, *boz*, *voz*.

The complementary figure of security with respect to 1 represents the risk (R), which corresponds to p-value of the statictical test (security + risk = 1):

R(3, 3, 1/3) = 1 - S(3, 3, 1/3) = 1 - .963 = .037 (3.7%)
R(8, 12, 1/3) = 1 - S(8, 12, 1/3) = 1 - .981 = .019 (1.9%)

The risk of 8 in 12 with probability 1/3, which is a complementary value of the security (*s*), represents the sum of the probabilities of the corresponding case and the superior cases: $x = \{8, 9, 10, 11, 12\}$, which occupies the upper (right) area of the frequency distribution (Table.10, Fig.2)[8]:

[8] Ichihara (1990: 118) explains the probability (P) used in the ratio test by binomial probability distribution:

P = BINOMDIST($\underline{x}$, *n*, *e*, 1)

On the other hand, our security(*s*) is:

s = S(*x*, *n*, *e*) = BINOMDIST($\underline{x\text{-}1}$, *n*, *e*, 1)

The difference between the two probabilities, P and S in BINOMDIST, is due to the difference of the test object: lower (left) probability in P (Ichihara), and higher (right) probability in risk (*r*), complement of security (*s*): *s* = 1 - *r*.

x	Binom.	Accm.
0	0.0077	0.0077
1	0.0462	0.0540
2	0.1272	0.1811
3	0.2120	0.3931
4	0.2384	0.6315
5	0.1908	0.8223
6	0.1113	0.9336
7	0.0477	0.9812
8	**0.0149**	**0.9961**
9	**0.0033**	**0.9995**
10	**0.0005**	**1.0000**
11	**0.0000**	**1.0000**
12	**0.0000**	**1.0000**

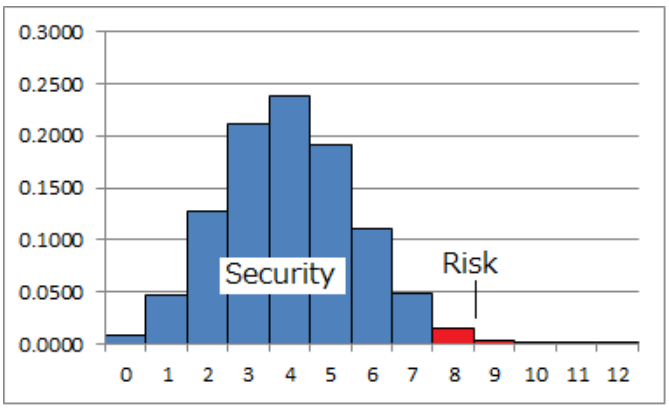

Table 10. Fig. 2: Security and risk in security test (*x* = 8, *n* = 12, *e* = 1/3).

3.2. Expected probability

We have observed that security (*s*) is obtained by the function of S(*x*, *n*, *e*):

s = S(*x*, *n*, *e*) … (*x*: occurrences, *n*: sum, *e*: expected probability)

However, in analyzing linguistic data, unlike such experiments on the coin toss or the taking out of a card (with replacement), the expected probability (*e*) of events from the beginning is generally unknown. The expected probability (*e*) is precisely our objective for obtaining the probabilistic value, from *x*: occurrences (frequency), *n*: trials (sum) and *s*: security. That is, we want to know what probability is guaranteed in frequency *x* out of sum *n*, with the desired security (95% or 99%). Now the known parameters are *x* (occurrences) and *n* (sum). The security (*s*) is set by the user.

Actually, according to the previous formula, the security (*s*), occurrences (*x*), sum (*n*), and expected probability (*e*) are interdependent; that is, if the three values are known, the remaining one value is mathematically or algorithmically derived from the known three values. For this reason, we elaborate the function E(*x*, *n*, *s*) that returns the expected probability (*e*) from *x* (occurrences), *n* (sum) and *s* (security):

e = E(x, n, s) … (x: occurrences, n: sum, s: security)

The function E(x, n, s) returns the expected probability (e) assumed from an event that occurs x times in n trials with security (s), calculated by binomial probability distribution. With these three parameters, the expected probability (e) is returned by a function E that we will explain later (Appendix; Program-B, Table-A, B):

e = E(5, 10, 0.99) = 0.150.

The following table shows the expected probability (e) of the events of 10 trials (n = 10), according to occurrences (x) from 1 to 10 (x = 1, 2, ..., 10), and with different securities (s): s = 0.95, 0.99, 0.999. In this table we observe that the greater the occurrence (x), the greater is the expected probability (e). For example, in the security (s) of 0.99, the expected probability (e) of x = 1 is 0.001, while that of x = 10 is 0.631[9]:

E(*x*, *n*:10, *s*)	***s*: 0.95**	***s*: 0.99**	***s*: 0.999**
x: 1	0.005	**0.001**	0.000
x: 2	0.037	**0.016**	0.005
x: 3	0.087	**0.048**	0.021
x: 4	0.150	**0.093**	0.050
x: 5	**0.222**	**0.150**	**0.090**
x: 6	0.304	**0.218**	0.141
x: 7	0.393	**0.297**	0.205
x: 8	0.493	**0.388**	0.282
x: 9	0.606	**0.496**	0.376
x: 10	0.741	**0.631**	0.501

Table 11: Expected probability (e) in ten trials (n = 10).

[9] We do not expose the case with x = 0, since the function E always returns 0.

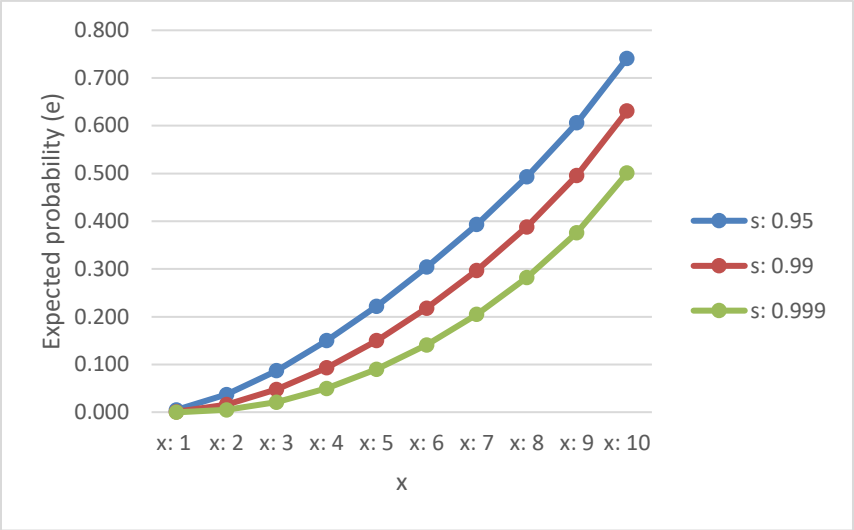

Fig. 3. Expected probability (*e*) in ten trials (*n* = 10).

At the same time, we confirm that the increase in security (*s*) causes a general decrease in the expected probability (*e*). For example, E(5, 10, 0.95) = 0.222, while the same with the security (*s*) of 0.99 is 0.150 and the same with the security of 0.999 is 0.090.

Suppose we have had 2 success (*x* = 2) in 10 experiments (*n* = 10). With this data, however, we cannot expect 20 successes in 100 future experiments. Let us see how the expected probability (*e*) are presented by increasing the number of experiments *n* = 10, 100, 1000, ...:

n	**E(*n**.2, *n*, .95)**	**E(*n**.2, *n*, .99)**
10	.037	.016
100	.137	.116
1,000	.179	.171
10,000	.193	.191
100,000	.198	.197
1,000,000	.199	.199
10,000,000	.200	.200

Table 12: Expected probability (*e*) in *n* tests (*n* = 10, 100, 1000, ...).

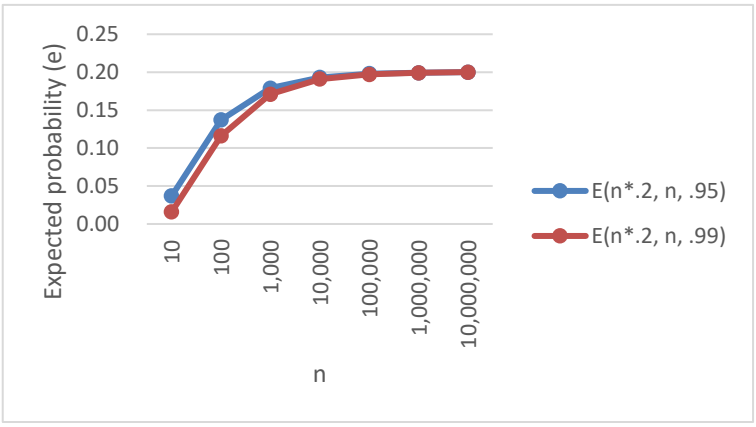

Fig. 4. Expected probability (*e*) in *n* tests (*n* = 10, 100, 1000, ...).

In the previous table with the condition that the security (*s*) is 0.95 (95%), when obtaining two successes in 10 trials, its expected probability (*e*) is 0.037 (3.7%) and is very far from the probability of success of 0.20 (20%). When *n* = 10,000 it reaches *e* = 0.193 (19.3%). From *n* = 10 000 onwards, the increase in the expected probability (*e*) is reduced. Finally, we obtain *e* = 0.20 (20%) when we reach *n* = 10,000,000. This characteristic of the expected probability (*e*) is important since through it we can appreciate what theoretical probability there is in each case of 2 in 10, 20 in 100, 200 in 1000, and so on. We are struck by the first cases where the magnitude of the base (10, 100) is considerably reduced, which causes the low expected probability: 0.037. This means that when the expected probability (*e*) = .037, two successes in 10 trials has 95% security (*s*). In this sense the function of expected probability (*e*), E(*x*, *n*, *s*), is inverse function of security (*s*), S(*x*, *n*, *e*):

e = E(2, 10, **0.95**) = 0.037
s = S(2, 10, 0.037) = **0.95**

3.3. Multiplier

We calculate the probabilistic frequency (PF) in the following formula:

PF = *e* * *m* … (*e*: expected probability, *m*: multiplier)

The probabilistic frequency (PF) is obtained by the expected probability function E(*x*, *n*, *s*) in combination with the multiplier (*m*).

PF = *e* * *m* = E(*x*, *n*, *s*) * *m*

Conveniently, the amount of the multiplier (*m*) will be constant in rounded form in order to maintain the same standard, for example 1,000 or 1,000,000 in accordance with the magnitude of the data[10]. The following table shows the probabilistic frequency (PF) with 95% security, corresponding to the frequency (*x*: [10, 100]) and the base (*n*: [10, 100])[11]:

PF	n:10	n:20	n:30	n:40	n:50	n:60	n:70	n:80	n:90	n:100
x:10	741	302	193	142	113	93	80	69	62	55
x:20		861	501	361	283	233	198	172	152	137
x:30			905	613	474	387	328	284	251	225
x:40				928	684	554	466	403	355	318
x:50					942	734	612	527	463	414
x:60						951	770	658	576	513
x:70							958	797	694	616
x:80								963	819	723
x:90									967	836
x:100										970

Table 13: Probabilistic frequency (PF). *s* = 0.95, *m* = 1000.

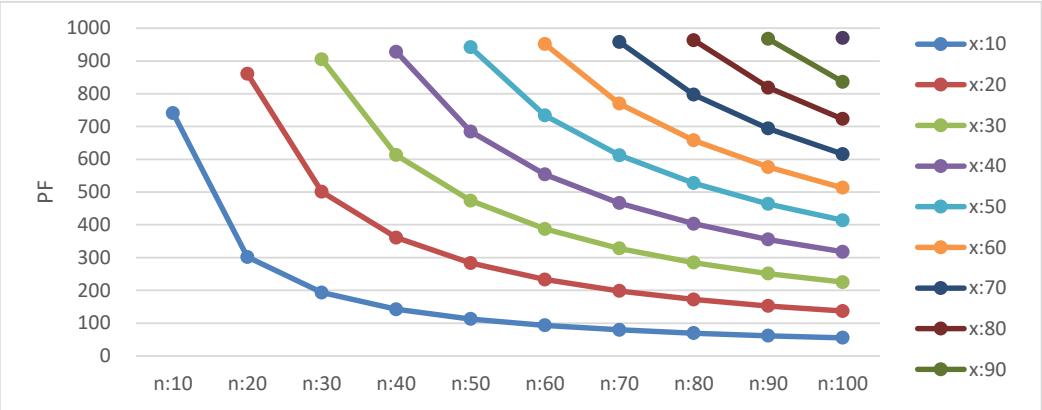

Fig. 5: Probabilistic frequency (PF). *s* = 0.95, *m* = 1000.

[10] When we want to know the probabilistic percentage (PP), that is to say, the percentage guaranteed with the security of 95% or 99%, the multiplier must be 100: PP(3, 3, 0.95) = E(3, 3, 0.95*)* * 100 = 0.368 * 100 = 36.8 (%). This means that the probability of 3 out of 3 is not 100%, but 36.8%, with the security (*s*) of 95% (0.95).

[11] For cases of *n* = [1, 10], [10, 100], [100, 1000], with *s* = 0.95, 0.99, see Table-A and Table-B in Appendix.

The lines of the graph (Fig. 5) show the monotonous downward trend, so we can interpolate the figures between e.g. x:10 and x:20 in n:20 to approximate the probabilistic frequency of e.g. x:12 in n:20 by 302 + (861-302) * 2/10 = 414.

4. Application

4.1. Spelling variants of «voz»

We now turn to verify the usefulness of the probabilistic frequency (PF) applied to two specific cases of Spanish diachrony. The first case deals with variant spellings of Spanish word «voz» ('voice'), which we have explained in section 2, where we have seen that the relative frequency (RF) is not convenient to correctly evaluate the frequency of 3 within 3 at 1200 (=100.0 %), which exceeds the frequency of the same word in 1250 (8 within 12 = 66.7%):

AF	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

RF (%)	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**100.0**	**66.7**	25.0	21.2	6.4
boz	0.0	25.0	66.7	34.6	37.2
voz	0.0	8.3	8.3	44.2	56.4
Sm	100.0	100.0	100.0	100.0	100.0

Table 14a, b: Absolute frequency (AF) and relative frequency (RF) of *uoz, boz, voz*.

The similar situation is also not solved in the normalized frequency (NF) by 100,000 words, where *uoz* in 1200 continues to exceed that of 1250 (38.8, 22.2)[12]:

AF	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
T	7 736	36 052	40 957	64 999	96 059

NF.	**1200**	**1250**	**1300**	**1350**	**1400**
uoz	**38.8**	**22.2**	7.3	16.9	6.2
boz	0.0	8.3	19.5	27.7	36.4
voz	0.0	2.8	2.4	35.4	55.2

Table 15a, b: Absolute frequency (AF) and normalized frequency (NF) of *uoz, boz, voz*.

If we resort to the probabilistic frequency (PF) with the security of 95% or 99% (Table 16a,b), for 1000 words, the relative importance of *uoz* in 1250 stands out (391, 302) in comparison with the same in 1200 (368, 215):

[12] TW: Total of words.

PF: *s* = .95	1200	1250	1300	1350	1400
uoz	**368**	**391**	72	123	28
boz	0	72	391	237	289
voz	0	4	4	324	474

PF: *s* = .99	1200	1250	1300	1350	1400
uoz	**215**	**302**	39	97	19
boz	0	39	302	200	259
voz	0	1	1	282	439

Table 16a, b: Probabilistic frequency (PF) with security (*s*) of 95%, 99%.

The last two tables describe the frequency guaranteed by 95% security (*s*) (Table 16a) or 99% security (Table 16b), which means that 3 in 3 does not represent 100% in any way, but only 36.8% with the security of 95%, or 21.5% with 99% security. Naturally, the probabilistic frequency value is decreased by increasing the security (*s*).

4.2. Combination of preposition and article

In Romance languages, Rhaeto-Romance, Italian, Portuguese, Catalan, French and Spanish, except for Romanian, there are many contractions of preposition and definite article. Within them, Spanish has only two forms *del* ('of the') and *al* ('to the') and, in other combinations, the two words are separated: *de la* ('of the'), *a la* ('to the'), *en el* ('in the'), and so forth. In the studies of general linguistics and history of the Spanish language in particular, it is explained that the forms of *de el* and *a el* have contracted into *del* and *al* by their frequent usage (Bybee 2007: 330; Elvira 2015: 18).

However, in the CODEA corpus from 1200 onwards there are no separate forms (*de el* and *a el*)[13], the starting point of the supposedly frequent contraction, except in an unusual way in 1500, 1600, 1700. On the other hand, for the other combinations, *a la, a las, de la*, etc., numerous examples are found in both united and separated forms:

AF	1200	1300	1400	1500	1600	1700
de el	6	2	0	**19**	**51**	**16**
del	1920	1829	2247	2858	1358	426
de la	309	110	145	370	451	171
dela	957	992	1303	1590	427	171
T: words	224,708	230,383	261,564	287,380	125,366	52,938

Table 17: Union and separation of preposition and article. Absolute frequency (AF).

[13] https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/codea.htm

We believe that from the historical point of view there have been no such processes of contraction *de el* > *del*, *a el* > *al*. If the forms *la, los, las* with the apheresis of initial *e* (*ela, elos, elas* > *la, los, las*) are due to the combination with preceding prepositions, *dela, delos, delas* > *de la, de los, de las* (Menéndez Pidal, 1926: 331), especially at the combination with the preposition «de», there should have been the united forms of the *dela, delos, delas* before the birth of the current forms of the definite article: *la, los, las*. Naturally, the formation of *del* and *al* must coincide with the united forms: *dela, delos, delas*; *ala, alas*. Therefore, we think that the contracted forms existed from the beginning of the history of the Spanish language and are not products of the contraction of frequent separate forms, in accordance with other Romance languages. These forms have passed the history of separation (*dela* > *de la*), rather than of contraction (*de la* > *dela, de el* > *del*), in which only two forms, *del, al,* remainded unchanged in united form (Ueda 2017).

To make the general observation of the tendency of the separated and united forms, we use the absolute frequencies (AF) just seen (Table 17). However, the total numbers of the words of each subcorpus of every 100 years in CODEA corpus are different (T: words). Therefore, from the absolute frequencies and the total word frequencies, we calculate the following probabilistic frequencies (multiplier = 100 000):

PF	**1200**	**1300**	**1400**	**1500**	**1600**	**1700**
de el	12	2	0	**43**	**318**	**190**
del	8,228	7,637	8,296	9,642	10,356	7,419
de la	1,249	405	481	1,179	3,324	2,835
dela	4,035	4,084	4,757	5,307	3,140	2,835

Table 18: Probabilistic frequency (PF). s = .95, m = 1,000,000 words.

We look at the probabilistic frequency (PF) of *de el* in 1500, where it is significantly lower (= 43) than its PF in 1600 (= 318) and 1700 (= 190). In the absolute frequency (Table 17), the figure of 1500 (19 occurrences) was recorded higher than in 1700 (= 16). However, the trend in the probabilistic frequency is inverse, that is, they are higher in 1600 and 1700, which can be negative evidence of the supposed contraction process since the contraction process assumes that the separate forms would be more frequent in the previous dates than in the later ones.

These observations do not vary either in the following normalized

frequencies (NF), since the bases of division are large and do not offer significant differences in 1600 and 1700:

NF	1200	1300	1400	1500	1600	1700
de el	27	9	0	**66**	**407**	**302**
del	8,544	7,939	8,591	9,945	10,832	8,047
de la	1,375	477	554	1,287	3,597	3,230
dela	4,259	4,306	4,982	5,533	3,406	3,230

Table 19: Normalized frequency (NF) per 1,000,000 words.

When comparing the probabilistic frequency (PF: Table 18) and the normalized frequency (NF: Table 19), the NF of *de el* in 1600 and 1700 (407, 302) are much higher than the PF (318, 190). Despite these differences, the general observation does not vary substantially. The use of the NF does not cause a problem when dealing with the data of large bases but certainly does in the data of the small bases (Table 15b). On the other hand, the PF does not produce problems both in the data of small bases (Table 16a, b) and those of big bases (Table 19). We can affirm that PF is robust in the sense that it applies to both types of data, with small and big bases.

5. Conclusion

Within several probability distributions treated in the statistics course, the binomial distribution is basic and easy to understand even for students of human science with fundamental knowledge of the concept of the probability treated in the high school mathematics. However, its ease of treatment does not necessarily imply a reduction in importance in the frequency analysis. On the contrary, the binomial distribution is essential to approach the world of statistical probability and guarantees its value when calculating the security or risk ratio (p-value) of individual frequency (3.1).

The percentage simply does not guarantee the security of the frequency, for example, the success of 80% is quite doubtful in case of only 4 successes in 5 trials. The probabilistic frequency implies the educational utility so as not to be deceived by the apparently high percentage. We are always cautious about the relative frequency with a reduced base. Now with the probabilistic frequency, we can calculate to what extent of frequency it is allowed to infer with the guaranteed security based on the binomial probability. The success of 4 times in 5

trials does not guarantee 80% success, but only 34%, a statement made with the security of 95%: E(4, 5, 0.95) = 0.343 (34.3%). We have to be careful when an ad only puts the success rate in percentage without base (population) or when we see the base with small numbers.

We believe we have shown that probabilistic frequency offers a solution to the problem found in the absolute, relative and normalized frequencies, widely used not only in corpus linguistics but in all sciences and daily life dealing with the frequency of phenomena in general. The way to evaluate the probabilistic frequency with a high degree of security (95%, 99%) is adequate when observing the figures within the reduced population and/or comparing the figures with highly distant bases.

On the other hand, the expected probability, the basis of the probabilistic frequency, naturally should not be taken as a absolute and constant value, as well as other frequencies treated in this study (absolute, relative and normalized frequency), since all these values are the figures observed in the limited data. If the spelling <u> is presented in 1200, 3 out of 3, its probabilistic percentage (s. = 95%) is only 36.8%. However, this does not mean that it will always be presented in other documents in 1200 with the probability of 36.8%. The expected probability is a value guaranteed with the security of, for example, 95%. It can also present other higher figures, although with less security, or lower figures, with more security. In this sense, the probabilistic frequency should be used to evaluate past results, for example, football or baseball notes, rather than to infer the future success rate in educational methods, medical treatments, sports, etc. We usually analyze historical or present documents rather than predict the future of linguistic situation.

Reference

Bishir, John W. / Drewes, Donald W. 1970. *Mathematics in the Behavioral and Social Sciences.* New York. Harcourt, Brace & World.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press.

Elvira, Javier. 2015. *Lingüística histórica y cambio gramatical.* Madrid. Editorial Síntesis.

Evison, Jane. 2010. "What are the basics of analyzing a corpus?", in O'Keeffe, Anne and McCarthy, Michael (eds.) *The Routledge Handbook of Corpus Linguistics.* pp. 122-135. New York. Routledge.

Fleming, Michael C. / Neillis, Joseph G. 1994. *Principles of Applied Statistics.* London. Routledge.

GITHE (Grupo de Investigación Textos para la Historia del Español): CODEA+ 2015 («Corpus de documentos españoles anteriores a 1800») [en línea] http://corpuscodea.es/ [11/15/2019]

Ichihara, Kiyoshi. 1990. S*tatistics for Bioscience.* (in Japanese.) Tokyo. Nankodo.

Kiyohara, Hideo. 1990. *Introduction to the Study of English Education,* (in Japanese.) Tokyo. Taishukan Shoten.

Kurata, Hiroshi / Hoshino, Takahiro. 2009. *Introduction to Statistical Analysis.* (in Japanese) Tokyo. Shinseisha.

McEnery / Hardie. 2012. *Corpus Linguistics. Method, Theory and Practice.* Cambridge. Cambridge University Press.

Menéndez Pidal, R. 1926-1980. *Orígenes del español.* Madrid. Espasa-Calpe.

Ueda, Hiroto. 2017. "Formación histórica de «del» y «al»", *90 Años de la Academia Boliviana de la Lengua*, La Paz, pp. 147-151.

_____. 2019. "Las grafías bilabiales sonoras <u>, <v> y <b> del español en relación con el fonema /f/ y el paradigma sibilante", Mónica Castillo Lluch / Elena Diez del Corral Areta (eds.) *Reescribiendo la historia de la lengua española a partir de la edición de documentos.* Bern. Peter Lang, pp. 141-174.

Ueda, Hiroto / Moreno Sandoval, Antonio, 2017. *Análisis de datos cuantitativos, para estudios lingüísticos*. https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/doc/numeros-es.pdf [11/15/2017]

Wong, Dona M. 2013. *The Wall Street Journal. Guide to Information Graphics. The Dos & Don'ts of Presenting Data, Facts, and Figures.* New York. W.W. Norton & Company.

Appendix

Program-A: Security

The function S, on receiving *x* (frequency), *n* (total), *e* (expected probability), *sel* (0, 1), returns *s* (security):

```
Function S(x, n, e) 'Security (Ueda 2017)
'(x: frequency, n: sum, e: expected probability)
  If x = 0 Then S = 0 'Definition
  If x > 0 Then S = Application.BinomDist(x - 1, n, e, 1)
End Function
```

Program-A: Security. S. Excel VBA.

Program-B: Expected probability

The function E, on receiving *x* (frequency), *n* (total), *s* (security) returns *e* (expected probability). In the program (Microsoft Excel VBA), we have used the binary search technique with the application of the encounter on the range (maximum-minimum). The sequential search is impracticable for the time it takes to reach the meeting.

```
Function E(x, n, s) 'Expected probability (Ueda 2017)
  '(x: ocurrence, n: trials, s: security)
  Dim i, k, p, mn, mx, c, lw, up: E = 0: k = 0
  If x = 0 Then Exit Function
  p = 10 ^ 6: mn = 0: mx = p: lw = s - 1 / p: up = s + 1 / p
  'p: precision, mn: min, mx: max in binary search, lw(er), up(per)
  For k = 1 To 1000
     i = (mx + mn) / 2 'i: Midpoint between mx and mn
     E = i / p 'E: Candidate of the expected probability
     c = Application.BinomDist(x - 1, n, E, 1)
     If c < lw Then 'If c does not reach lw …
        mx = i 'Lower the maximum of search to the midpoint (i)
     ElseIf c > up Then 'If c exceeds up ...
        mn = i 'Raise the minimum of search to the midpoint (i)
     Else 'If c is between lw and up …
        Exit For 'Exit the loop
     End If
  Next
End Function
```

Program-B: Expected probability. E. Excel VBA.

PF	n:1	n:2	n:3	n:4	n:5	n:6	n:7	n:8	n:9	n:10
x:1	50	25	17	13	10	9	7	6	6	5
x:2		224	135	98	76	63	53	46	41	37
x:3			368	249	189	153	129	111	98	87
x:4				473	343	271	225	193	169	150
x:5					549	418	341	289	251	222
x:6						607	479	400	345	304
x:7							652	529	450	393
x:8								688	571	493
x:9									717	606
x:10										741

PF	n:10	n:20	n:30	n:40	n:50	n:60	n:70	n:80	n:90	n:100
x:10	741	302	193	142	113	93	80	69	62	55
x:20		861	501	361	283	233	198	172	152	137
x:30			905	613	474	387	328	284	251	225
x:40				928	684	554	466	403	355	318
x:50					942	734	612	527	463	414
x:60						951	770	658	576	513
x:70							958	797	694	616
x:80								963	819	723
x:90									967	836
x:100										970

PF	n:100	n:200	n:300	n:400	n:500	n:600	n:700	n:800	n:900	n:1000
x:100	970	440	288	215	171	142	122	106	94	85
x:200		985	619	458	363	302	258	225	200	179
x:300			990	712	563	466	397	347	307	276
x:400				993	768	634	540	470	417	374
x:500					994	806	685	596	528	474
x:600						995	834	724	640	574
x:700							996	854	754	675
x:800								996	870	778
x:900									997	883
x:1000										997

Table-A1, A2, A3: Probabilistic frequency (PF). *s* = .95, *m* = 1000.

PF	n:1	n:2	n:3	n:4	n:5	n:6	n:7	n:8	n:9	n:10
x:1	10	5	3	3	2	2	1	1	1	1
x:2		100	59	42	33	27	23	20	17	16
x:3			215	141	106	85	71	61	53	48
x:4				316	222	173	142	121	105	93
x:5					398	294	236	198	171	150
x:6						464	357	293	250	218
x:7							518	410	344	297
x:8								562	456	388
x:9									599	496
x:10										631

PF	n:10	n:20	n:30	n:40	n:50	n:60	n:70	n:80	n:90	n:100
x:10	631	239	151	110	87	72	61	53	47	42
x:20		794	439	312	243	199	168	146	129	116
x:30			858	559	426	346	291	252	222	198
x:40				891	637	509	426	367	322	287
x:50					912	692	572	489	428	381
x:60						926	733	620	540	479
x:70							936	764	659	582
x:80								944	789	691
x:90									950	809
x:100										955

PF	n:100	n:200	n:300	n:400	n:500	n:600	n:700	n:800	n:900	n:1000
x:100	955	416	271	201	160	133	113	99	88	79
x:200		977	600	441	349	289	247	215	191	171
x:300			985	696	547	452	385	335	297	267
x:400				989	755	620	527	458	406	364
x:500					991	795	673	584	516	463
x:600						992	824	713	629	563
x:700							993	845	744	665
x:800								994	862	769
x:900									995	876
x:1000										995

Table-B1, B2, B3: Probabilistic frequency (PF). *s* = .99, *m* = 1000.