Probability in linguistic typology: course handout

Elena Maslova

July 17-19, 2006

Part I

Quantitative typology: the why and wherefore

1 Why quantitative (statistical) typology?

How to calculate the probability of an existing exception being present in a sample?
Counting techniques. The basic idea is to count the number of outcomes corresponding to an occurrence of an event and to divide it by the total number of possible outcomes. In our case, the number of samples containing an exception should be divided by the total number of all samples of the same size. The most important idea for all such calculations is that if one operation can be performed in n_1 different ways and the other, in n_2 different ways, then the total number of ways in which both can be carried out is $n_1 n_2$.
Permutations: How many different orderings are possible for a collection of n items?

$$P_n = n!$$

.
Permutations of r items selected from a set of n elements: If we select r items from a collection of n items, the number of possible permutations is

$${}_nP_r = \frac{n!}{(n-r)!}$$

Combinations: The number of possible combinations of r items from a collection of n (if the order is irrelevant) is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Note that this is the total number of possible samples of size r from a population of size n.

Distinguishable vs. indistinguishable objects. From the point of view of a specified typology, languages that belong to the same type are **indistinguishable**; only types are **distinguishable**. In particular, we do not care which particular languages represent a given type in our sample; we are only interested in the total number of such representatives.

Example 1. A universal with a single exception. Assume there is a single exception from a universal among the total of N languages. The number of samples of size n containing this single exception is equal to the total number of samples of size $n-1$ from the other $N-1$ languages:

$$\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!}$$

To obtain the probability $P(E)$ of finding the exception, we divide this expression by $\binom{N}{n}$ and get:

$$P(E) = \frac{(N-1)!n!(N-n)!}{N!(n-1)!(N-n)!} = \frac{n}{N}$$

What if there are k exceptions? Using the techniques already introduced, we can easily calculate the probability p_k of there being no exceptions in the sample, hence the probability $q_k = 1 - p_k$ of at least one exception.

n	q_1	q_{10}	q_{60}
50	0.01	0.08	0.39
100	0.02	0.15	0.63
300	0.05	0.39	0.95
600	0.10	0.63	1

Table 1: **The probability q_k of finding at least one of k exceptions in a random sample of size n (for $N = 6000$)**

We can achieve some degree of confidence as far as low probability of occurrence is concerned, yet not for absolute universals.

This opens the possibility of new types of questions: if a statement of likelihood less than 0.01 is interesting, than why not about less than 0.10?

2 Dependencies

A special, particularly famous new type of information has to do with dependencies between different language properties ("correlations").

Example 2. Implicational universals of comparative constructions. Leon Stassen introduces a typology of comparative constructions, which comprises, among other types, EXCEED-comparative and Separative Comparative. These seem to be linked to the basic word order:

Basic word order:	SVO	Other
EXCEED-comparative	20	0
Other	15	75

Basic word order:	V... or ...V	...V...
Separative comparative	31	1
Other	42	36

Example 3. Word order correlations: adpositions. The following figures (based on Dryer's data) appear to indicate a dependency between the relative order of verb and object and the locus of adpositions (post- vs. prepositions):

	VO	OV
Postpositions	9	87
Prepositions	62	5

Part II

Statistical data and statistical inference. Inferences and explanations in typology

1 Basic description

1.1 Distribution

A basic description for a set of statistical data invokes “individuals” and “variables”. For typology, “individuals” would generally mean “languages” (with some important qualifications), and “variables” would mean“typological parameters”. Most typological parameters are **categorical variables**, that is, they classify languages into several groups, or categories (such as, e.g. the existence of nasal vowels); some are **quantitative**, that is, each language is assigned a numerical value (e.g. the total number of vowels). The **distribution** of a variable is the set of possible values and their frequencies.

Example 4. Basic word order. Russel Tomlin classifies all languages into six “basic word order” types, based on the most frequent relative order of lexical subject (S), lexical object (O) and the verb (V) in finite clauses, and obtains the following distribution (in a set of 402 languages):

	absolute	relative
SOV	180	0.45
SVO	168	0.42
VSO	37	0.09
VOS	12	0.03
OVS	5	0.01
OSV	0	0.0
Total	402	1

This is how often the categorical variable “basic word order” takes each of its values in Tomlin’s sample of languages.

Example 5. Morphological complexity. Johanna Nichols counts the total number of morphemes coding grammatical relations in a pre-defined set of distinct grammatical constructions . This is a quantitative variable which can, in principle, take any integer value from 0 to 27. Its distribution in the set of languages explored by Nichols is as follows:

Complexity	Number of languages	Frequency
2	7	0.032
3	4	0.018
4	19	0.086
5	13	0.059
6	31	0.140
7	23	0.104
8	32	0.144
9	20	0.090
10	20	0.090
11	18	0.081
13	8	0.081
14	5	0.036
15	3	0.023
16	1	0.014
Total	210	1

1.2 Mean

The distribution of a quantitative variable allows for a much broader range of relative descriptive measures than a categorical variable.

The **mean value** of a variable is the sum of the values observed in all observations divided by the number of observations:

$$(1) \quad \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

If v_j denote the values of the variable, and n_j denotes the number of observations where v_j is observed, then

$$(2) \quad \bar{x} = \frac{1}{n}\sum_{j=1}^{k} v_j \cdot n_j,$$

where k is the total number of possible values.

For example, the mean value of Nichol's measure of morphological complexity is 8.1.

1.3 Order statistics and median

The sample median is the most broadly used order statistic. Assume we list all outcomes Y_j ($j = 1, ..., n$) so that the observed value never decreases ($Y_j \leqslant Y_{j+1}$). Then Y_1 is the smallest order statistic, Y_n, the largest order statistic. If n is odd, then the sample median is the middle observation, Y_k for $k = (n+1)/2$; if n is even, then it is any value between Y_k and Y_{k+1}, where $k = n/2$. x%-percentile is Y_k for the closest $k \leqslant \frac{x}{100}n$.
Exercise: Find the median, the 25%-percentile and the 75%-percentile for Nichols' complexity sample.

1.4 Variance

The measure of variability of the variable in a sample is the **sample variance**. The sample variance, s^2, is defined as

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - x)^2}{n-1}$$

For example, the sample variance of morphological complexity, as examined by Nichols is 9.24.

2 Inferences

The fundamental idea of inferences from statistical data is that if we repeat a certain experiment multiple times, the results would be roughly similar; put it the other way round, a single experiment (e.g. one sample) gives us some information about what would happen in other experiments of the same sort, and this why its result are of interest.

For plain descriptive quantitative typology, this can be translated as follows. Assume we have defined a population in a certain way; it is a finite set of similar individuals with an unknown distribution of cross-linguistic variables. Further, we have chosen ("sampled") a subset of n elements of this set, following certain rules of randomness (for our purposes, let us assume that all individuals had exactly the same probability of making it into our sample), and analyzed the distribution of our variables in this sample. Under these conditions, if n is reasonably large, we can believe that most of the other possible samples of the same size and chosen in the same way from the same population will have roughly the same properties, and this gives us the possibility to infer the properties of the population from the properties of a sample. At this purely descriptive level, the validity of such inferences is guaranteed simply by randomness of sampling.

Genuinely linguistic problems arise if we wish to draw inferences not about a concrete finite population of languages, but about a language as a general phenomenon. For theoretical typology the issue is to construe each specific language as a "trial" of the same phenomenon; and here is where lots of empirical problems arise.

Two types of empirical observations that create problems: genetic similarities and differences between geographical areas.

Example 6. Genetic groupings as individuals. Johanna Nichols devised a three-layer system for this type of sampling (this system also has an area-oriented component, not discussed here).

In her terminology (24-25), the *family* is a genetic grouping of a time depth around ca. 2500-4000 years, like the older branches of Indo-European (i.e. Balto-Slavic). The *stock* corresponds in effect to the highest level reconstructible by the comparative method (the estimated time depth is 5000-8000 years). Each family (present within one of the pre-defined geographical areas) is represented in the sample by a single language, which thus assigns the "type-values" to the whole family. Some effort has been made to exclude languages representing non-dominant types, i.e. languages considered atypical of their families were avoided (there were, however, some exceptions from this rule). Assuming that the intra-family frequencies of dominant typological values are close to unity, the naturally high probability of representing a family by its dominant value was additionally increased by non-randomness introduced in the sampling procedures: on average, a language with a dominant value had a greater chance of being included in the sample (although how exactly these probabilities differ, is hard to estimate).

Each stock containing six or less families is represented by all its families; for stocks that have branched into more families, there was an upper limit of six families. These families have been apparently also selected not quite randomly: first, they have been distributed between geographical areas, and secondly, an attempt was made to cover "the known typological range" of each stock. Thus, while for families the potential effects of intra-family variability were decreased by the sampling procedure, the intra-stock variability was effectively increased by the sampling procedure. Insofar as stocks contain more than one family, the typological variables defined at this level are various descriptive measures of the corresponding family-based distributions (like the mean values, the total numbers of represented types, the frequency of dominant type, etc.). Here, for example, are values for the stock-level variable corresponding to the sample mean of morphological complexity:

Stock	Lgs	Mean complexity
Afroasiatic	4	11.8
Niger-Kordofanian	6	7.3
Indo-European	5	9.4
Uralic-Yukaghir	4	10.5
Pama-Nyangan	6	9.7
Austronesian	6	7.5
Uto-Aztecan	4	9
Penutian	5	10.6

Example 7. Areas as individuals. Matthew Dryer (1989) divided the world into six large areas, with the basic idea that a linguistic preference can be established only if it is supported by statistical data from each area. His raw data on representation of basic word orders is summarized in the following table.

	Africa	Eurasia	Australia	North America	South America	Total
SOV	22	26	19	26	18	111
SVO	21	19	6	6	5	57
Other	2	7	5	28	8	50
Total	45	52	30	60	31	218

Each statement of the form "Type A occurs more frequently than type B" can be evaluated as true or false for each area separately. In other words, each area functions as an individual for which a binary variable is defined. For example, for the statement "SVO occurs more frequently than SOV", individuals-areas are characterized by the following values:

$f(SVO) > f(SOV)$	Africa	Eurasia	Australia	North America	South America
	Yes	Yes	Yes	Yes	Yes

Dryer's idea is that the probability of obtaining five "Yes" by chance (that is, under the assumption that "Yes" and "No" are equally likely, and the values for all areas are mutually independent) equals $\frac{1}{32}$. Accordingly, a linguistic preference for SVO over SOV is considered as established.

On the other hand, the corresponding values for the binary variable “SOV occurs more frequently than all non subject-initial orders” assumes the following values:

$f(SOV) > f(other)$	Africa	Eurasia	Australia	North America	South America
	Yes	Yes	Yes	**No**	Yes

Accordingly, this statement does not reflect, according to Dryer, a genuine linguistic preference.

Is Dryer’s estimate of the probability of establishing a “wrong” linguistic preference correct?

Exercise questions:

1. A linguist wants to establish the distribution of a certain cross-linguistic variable. She selects 100 languages at random from a complete list of languages, and the grammars of 81 of them are found in the library. What is the population under study?
2. In the same situation, it turns out that only 56 of 81 grammars actually describe the phenomenon under analysis in a satisfactory fashion. The same question, what is the population?
3. A linguist selects at random a single language from each genetic stock. What is the population under study? Assuming the variable under analysis has an extremely rare value, has the probability of finding this value changed (as opposed to a random sample)? If yes, then how and why?

Part III
Probability

1 Sample space, outcomes, events

“Experiment” is the process of obtaining an observed result. A performance of an experiment is called a **trial**, and an observed result is an **outcome**. The set of possible outcomes is called the **sample space** (one one outcome is supposed to occur on any given trial).

Example 8. A word order experiment. Suppose we are interested in the order of major clausal constituents, SOV. Then, a possible outcome of an experiment is a set of orders possible in the language. There are $2^6 - 1$ possible outcomes: in principle, any of six possible word orders can either possible or impossible, yet it cannot be the case that all orders are disallowed. The sample space is a set of all possible subsets:

$$\{\{SOV, SVO, VSO, VOS, OSV, OVS\}, \{SOV, SVO, VSO, VOS, OSV\}, ...\{OVS\}\}$$

.

Example 9. Simple word order flexibility experiment. If we are interested in the word order flexibility, a possible measure would be the total number of possible orders. The appropriate sample space, then, is $\{1, 2, 3, 4, 5, 6\}$.

Example 10. Simple basic word order experiment. If we are interested in only in the dominant (basic) word order, the sample space consists simply of all six possible orders

$$\{SOV, SVO, VSO, VOS, OVS, OSV\}$$

.

Example 11. Area-oriented basic word order experiment. If in Example 10 we wish to control for geographical area (as Dryer does), we would note, along with the dominant order, the area in which language is spoken. Assuming there are five major areas (as in Dryer’s study), the sample space would contain 30 pairs $< order, area >$:

$$\{< SVO, Africa >, < SVO, Eurasia > ...\}$$

.

Example 12. Nichols’ complexity. Nichols examines nine different construction types; in each construction type, the relation between constituents can be coded in three locations: on the head constituent, on the dependent constituent, or by a free marker. This gives 27 possible loci for grammatical markers. The result of experiment is a sequence of 27 ”Yes/No” answers (corresponding to the presence/absence of a marker). The sample space consists of 2^{27} possible outcomes. If, however, we are interested in morphological complexity (the total number of markers in all constructions), the sample space consists of 28 integers from 0 to 28.

A sample space is **finite** if it consists of a finite number of outcomes, and it is **countably infinite** if its outcomes can be put into one-to-one correspondence with the positive integers. A **discrete** (countable) sample space is a sample space that is either finite or countably infinite. Otherwise, it is a **continuous** sample space.

Example 13. Word order flexibility: continued. A possible simple measure of word order flexibility is the actual frequency of the most frequent ("dominant") word order, which can be obtained by text counts: in a set of similar texts in different languages, we can count the total number of independent transitive clauses containing all three relevant constituents (N), and the number of such clauses with the dominant order (n). The outcome of an experiment is the ratio $\frac{n}{N}$. In this case, the sample space can be thought of as continuous, i.e. as a set of all real numbers $(0, 1(6); 1]$. A more thorough study can include frequencies of all possible word orders. Then, the sample space would be a five-dimensional space, with the values along each dimension ranging from 0 to 1, with the additional constraint that their sum is less or equal to one (the space is five-dimensional because the sixths value is uniquely determined by the other five).

An **event** is a subset of the sample space. An event has occurred if it contains the outcome that occurred. An **elementary event** contains a single outcome. The whole sample space is a special kind of event (**sure event**), and the empty set is the **"null event"**. If the intersection of two events is empty, they are referred to as **"mutually exclusive"**. More than two events are mutually exclusive if they are pairwise mutually exclusive.

To illustrate the concept of event, consider Example 10. An event "S precedes O" is the following subset of outcomes $\{SOV, SVO, VSO\}$. For a sample space of Example 8, an event "S always precedes O" would contain outcomes which do not include either of orders VOS, OSV, OVS. There are $2^3 - 1 = 7$ outcomes in this event:

$$E = \{\{SOV, SVO, VSO\}, \{SOV, SVO\}, \{SOV, VSO\}, \{SVO, VSO\}, \{SOV\}, \{SVO\}, \{VSO\}\}$$

If the sample space is continuous, an event would be defined as a region in this space. So, if we measure the frequency of dominant order (Example 13), we can define an event like "Word order is flexible" if the frequency of the dominant order is less than 0.50, i.e $E = (0.16(6), 0.5)$.

2 Probability

2.1 Definition

A set function that associates a real value $P(A)$ with each event A is called a probability set function, and $P(A)$ is called the probability of A if the following properties are satisfied:

$$0 \leqslant P(A) \text{ for every } A \tag{3}$$

$$P(S) = 1 \text{ (S is the sure event)} \tag{4}$$

$$P(\bigcup_{i=1}^{\infty} A_i = \sum_{i=1}^{\infty} P(A_i) \text{ if } A_1, A_2 \ldots \text{ are mutually exclusive events.} \tag{5}$$

2.2 Basic properties

- If A is an event and A' is its complement, then $P(A) = 1 - P(A')$.
- For any event A, $P(A) \leqslant 1$.
- For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- For any three events,

  $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

- If $A \subset B$ then $P(A) \leqslant P(B)$.
- **Boole's inequality** For any sequence of events $A_1, A_2, ...$

  $$P(\bigcup_{i=1}^{\infty} A_i \leqslant \sum_{i=1}^{\infty} P(A_i)$$

  A similar result holds for finite unions.

- **Bonferroni's inequality** For any sequence of events $A_1, A_2, ...$

  $$P(\bigcap_{i=1}^{\infty} A_i \geqslant 1 - \sum_{i=1}^{\infty} P({A'}_i)$$

3 Conditional probability

3.1 Definition

The conditional probability of an event A, given the event B is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

if $P(B) \neq 0$. Relative to the sample space B, conditional probabilities satisfy the original definition.

Example 14. Type and Area. In Nichols (1986), 15 of languages of North America are classified as head-marking, whereas the other eight North-American languages in the sample are not head-marking. What is the conditional probability that a randomly chosen language from this sample turns out to be head-marking if it is already known that it is spoken in North America?

Of the other languages in her sample, only two are classified as head-marking. What is the conditional probability that a language from this sample is from North America if we already know that it is head-marking?

3.2 Multiplication theorem

For any events A and B,

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

Example 15. Sampling without replacement. Imagine there is a genetic group containing 40 languages, 10 of which have SVO as their basic order (figures adjusted from Tomlin's sample data for Afroasiatic). What is the probability that the two first randomly selected languages will be SVO?

	A_1	$A_1^{'}$	
A_2	$10 \cdot 9$	$30 \cdot 10$	$10 \cdot 39$
$A_2^{'}$	$10 \cdot 30$	$30 \cdot 29$	$30 \cdot 39$
	$10 \cdot 39$	$30 \cdot 39$	$40 \cdot 39$

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{10}{40} \cdot \frac{9}{39} = 0.06$$

4 Total probability and Bayes

If $B_1, B_2, ..., B_k$ is a collection of mutually exclusive and exhaustive ($B_1 \cup \ldots \cup B_k = S$) events, then for any event A,

$$P(A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

Example 16. Genetic sampling. In many typological studies, the probability of a language being represented in the sample depends on the total number of languages in the same genetic group (and/or on its branching structure). Let n_i be the pre-determined number of languages from the i-th group to be included in the sample, N_i, the total number of languages in this group. The conditional probability of a language being represented if it belongs to the i-th group is

$$p_i = \frac{n_i}{N_i}.$$

The probability that a randomly selected language belongs to the i-th group is $\frac{N_i}{N}$, where $N = \sum_i N_i$. Then, the probability of a language being represented is

$$P = \sum_i p_i \cdot \frac{N_i}{N} = \sum_i \frac{n_i}{N_i}\frac{N_i}{N} = \frac{1}{N}\sum_i n_i$$

Now let M_i be the number of languages of a certain type in the i-th group ($M = \sum_i M_i$), and assume we have decided to chose exactly one language from each group ($n_i = 1$). What is the probability P' of a random language of this type being represented in our sample?

Bayes' Rule For a set of exhaustive mutually exclusive events B_i, and for each $j = 1, \dots, k$:

$$(6) \quad P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Example 17. Comparative constructions and Bayes rule. The following table gives the probabilities that a languages has SOV as its basic word order under the conditions of known primary type of comparative construction (according to Stassen (1985)), and the probabilities of these comparative types.

	$P(SOV\|C_i)$	$P(C_i)$
C_1 = Separative	0.88	0.3
C_2 = Allative	0.14	0.06
C_3 = Locative	0.67	0.11
C_4 = Exceed	0	0.19
C_5 = Conjoned	0.56	0.17
C_6 = Particle	0.21	0.18

The Bayes rule gives us the probabilities of comparative type under the condition that we know that the language's basic word order is SOV, e.g.

$$P(Separative|SOV) = P(C_1|SOV) = \frac{P(C_1)P(SOV|C_1)}{P(SOV)} = \frac{0.88 \cdot 0.3}{0.45} = 0.59$$

The knowledge of the basic word order makes the hypothesis of Separative comparative type nearly twice more likely.

Example 18. Type and family. Assume that a half of all languages belong to a certain type T; also, we assume that, for a certain genetic classification, 0.9 of all languages in each group belong to a single type (in the corresponding typology). What is the probability of a language chosen at random belonging to T? If we have chosen a language from a certain family and it turned to belong to T, what is the probability that another language randomly chosen from the same group would also belong to T?

Part IV
Random variables

1 Random variables and typological parameters

A **random variable** (say, X) can be defined as a function over a sample space, which associates a real number with any possible outcome. The probability distribution function of a random variable, then, is a function $f(x)$, which assigns to each number x the probability of the corresponding outcome. Thus, in principle, one can define a random variable for any parameter of typological variation, and it can be useful in a broad range of contexts. However, if the parameter is essentially categorical, one has to be cautious, since some important properties of the resulting random variable can turn out to depend on how exactly it is defined.

A notable exception is constituted by binary parameters, for which a random variable can be defined straightforwardly: one value (interpreted as a "positive outcome") would be associated with "1" and the other ("negative outcome"), with "0". Such a variable is known as **Bernoulli variable**. If the probability of the positive outcome is p and the probability of the negative outcome is $q = 1 - p$, then the **probability distribution function** of the variable is

$$(7) \quad f(x) = p^x q^{1-x}$$

If X is a random variable, and $f(x)$, its probability distribution function, then the **expected value** of X is

$$(8) \quad \mu = E(x) = \sum_x x f(x)$$

For example, the expected value of a Bernoulli variable is equal to the probability of its positive outcome ($\mu = 1 \cdot p + 0 \cdot q = p$).

Note that the expected value of a random variable is conceptually related to the familiar mean (the average) of a distribution in a sample:

$$(9) \quad \bar{x} = \frac{x_1 + x_2 + ...x_n}{n} = \frac{1}{n}\sum_1^n x_i$$

For a Bernoulli variable, for example, x_i can assume two values, 0 and 1. If the outcome "i" is observed in n_i trials ($n_0 + n_1 = n$), this can be rewritten as

$$(10) \quad \bar{x} = \sum_0^1 i \cdot \frac{n_i}{n} = \frac{n_1}{n},$$

that is, the mean value equals the relative frequency of the positive outcome.

The second important property of a random variable is its **variance**, which provides a measure of the variability (or dispersion) in the distribution. The variance is defined as the expected value of the function $u(x) = (x - \mu)^2$, i.e.

(11) $\sigma^2 = \mathrm{Var}[X] = E[(X - \mu)^2]$.

The **standard deviation**, σ, is the positive square root of the variance.

It is often easier to calculate the variance using the following equation:

(12) $\mathrm{Var}(x) = \sum_x (x - \mu)^2 f(x) = E[x^2] - \mu^2$

The variance of a Bernoulli variable:

(13) $\sigma^2 = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p = p \cdot (1 - p) = pq$

Note that the maximum variance is achieved for $p = q = 1/2$, and the variance tends to zero as any of these values tends to 1. This conforms well with the intuitive concept of linguistic variability: the more probable one of the values, the less the cross-linguistic variability (in the extreme case of one of the probabilities being equal to one, we have a universal, i.e. no variability at all).

Example 19. Alignment as a random variable. Obviously, it would be useful to extend this measure of variability to non-binary parameters. This approach is used by Johanna Nichols to estimate and compare the amount of variability for different typological parameters. For example, she divides alignment mechanisms into three numerical "gross types", so that "1" is assigned to various split systems, "2" to accusative and neutral, and "3", to ergative mechanisms. Based on her data, we can assign the following probability distribution function to this variable (this takes into account all potential "locations" of an alignment mechanism, namely, nouns, pronouns, and verb agreement): $f(1) = 0.07, f(2) = 0.81, f(3) = 0.12$. The expected value of this variable is, then:

$$\mu = 0.07 + 2 \cdot 0.81 + 3 \cdot 0.12 = 2.05$$

The variance is:

$$\sigma^2 = (0.07 + 4 \cdot 0.81 + 9 \cdot 0.12) - 2.05^2 = 0.19$$

Since the assignment of numerical values is arbitrary, we could as easily say that "1" is accusative/neutral, "2" is ergative and "3" is split. Then, the distribution would be $p(1) = 0.81, p(2) = 0.12, p(3) = 0.07$. The expected value would be

$$\mu = 0.81 + 2 \cdot 0.12 + 3 \cdot 0.07 = 1.26$$

And the variance would be much higher:

$$\sigma^2 = (0.81 + 4 \cdot 0.12 + 9 \cdot 0.07) - 2.05^2 = 0.33$$

Thus, if we arbitrarily assign numerical values to non-binary categorical parameters, we can, quite accidentally, heavily influence the results. The only general way of avoiding this problem is **ranking**, that is, we have to order the values so that their probabilities decrease (or rather, never increase), and use the ranks as the values of our random variable.

Example 20. Morphological complexity. Recall that Nichols' measure of morphological complexity amounts to the total number of morphological markers in $N = 27$ potentially possible locations in different constructions. Imagine, as a simplistic model of this phenomenon, that a marker appears in each location with the same probability p, independently of whether other markers are present. What is, then, the probability that the morphological complexity (i.e. the total number of markers in all locations) will be exactly n? The number of ordered arrangements of n locations from N possible locations is $N(N-1)...(N-n+1) = \frac{N!}{(N-n)!}$. The number of different possible orders (permutations) of n locations is $n!$. Thus, the number of ways of choosing an unordered set of n locations is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

The probability of any such set is $p^n(1-p)^n$. Thus, the probability of morphological complexity n, counted for N possible locations is:

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

For $N = 27$ and $p = \frac{1}{3}$, we get a distribution remarkably close to the empirical one obtained by Nichols.

This is an example of **binomial distribution**, the distribution of sum of n independent identical Bernoulli variables.

(14) $f(x; n) = \binom{n}{x} p^x q^x($ Mean: np, Variance: $npq)$

2 Sampling distributions

Example 21. Sampling from Areas. When we sample languages, we usually sample without replacement. This is not very important if a small set of languages is selected from a large language population. However, it does become extremely important for a sampling procedure invented by Dryer.

Assume there are N genera in an area, and M of them belong to a particular type. What is the probability of finding exactly x languages of this type in a sample of size n? The sample space is a collection of all subsets of size n; there are $\binom{N}{n}$ of these. There are $\binom{M}{x}\binom{N-M}{n-x}$ outcomes where the number of M-languages is exactly x. Hence,

$$P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Hypergeometric distribution:

$$f(x;n) = P(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$\text{Mean: } \frac{nM}{N}, \text{ Variance: } n\frac{M}{N}(1-\frac{M}{N})\frac{N-n}{N-1}$$

3 Discrete, continuous, and mixed

Example 22. Greenbergian morphological complexity. Consider a measure of morphological complexity defined as the mean number of morphemes per word. If it is studied typologically, we will consider, for each language L_i, a selection of texts with a total length of N_i words, and note the number of morphemes M_i in this selection. The results of our observations could be represented as a discrete set of n outcomes ($X_i = \frac{M_i}{N_i}$), each of which would probably occur no more than one or two times. However, it is usually appropriate to consider an idealized situation in which X can assume any value in some interval.

One way to study this distribution would be to consider the relative frequency of languages for which X_i is less than x, i.e. to analyze its cumulative distribution function $F(x)$. The cumulative distribution function of a random variable is defined by $F(x) = P[X \leqslant x]$. The simplest "null" hypothesis would be to suggest that $F(x)$ grows proportionally with x, i.e. $F(x) = cx$. Another would be to consider the relative frequencies of outcomes within some small intervals (e.g. $P[0.05 < X \leqslant 0.1] = F(0.1) - F(0.05)$. If our hypothesis is correct, this probability must be $0.05c$.

Under some conditions of regularity, which we will not discuss, the derivative of $F(x)$ of a continuous random variable is called its probability density function, so that the cdf can be represented as

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Obvious conditions are

$$f(x) \geqslant 0$$
$$\int_{-\infty}^{\infty} f(t)dt = 1$$

The expected value of a continuous variable is defined by

$$\int_{-\infty}^{\infty} xf(x)dx$$

Example 23. Mixed typological parameters. Some cross-linguistic variables are best described as "mixed", partly continuous, partly discrete. This is the best approach, for example, if some sort of constraint works as an absolute grammatical constraint in some languages and as a "soft" preference in other languages. Consider, for example, the tendency to drop independent subject pronouns. The frequency of subject pronouns can be either zero (if a language doesn't have any) or positive (if it does). The corresponding form of the cumulative distribution function would be

$$F(x) = aF_d(x) + (1-a)F_c(x)$$

where a is the probability that a language has no subject pronouns, $F_d(x) = 1$, and $F_c(x)$ is the conditional distribution function, under the condition that the language does have subject pronouns.

4 Some special distributions

4.1 Uniform

Uniform discrete (DU(N))

(15) $f(x) = \dfrac{1}{N}\ x = 1, 2, ...N,\ E(X) = \dfrac{N+1}{2}, \mathrm{Var}(X) = \dfrac{N^2-1}{12}$

Uniform continuous (UNIF(a, b))

(16) $f(x) = \dfrac{1}{b-a}\ a < x < b,\ E(X) = \dfrac{a+b}{2}, \mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$

4.2 Geometric and Negative binomial

The number of trials required to obtain the first success (GEO(p)):

(17) $g(x; p) = pq^{x-1}\ x = 1, 2, 3,\ E(X) = \dfrac{1}{p}, \mathrm{Var}(X) = \dfrac{q}{p^2}$

The number of trials required to obtain r successes.

(18) $f(x; r, p) = \dbinom{x-1}{r-1} p^r q^{x-r}\ x = 1, 2, 3,\ E(X) = \dfrac{r}{p}, \mathrm{Var}(X) = \dfrac{rq}{p^2}$

4.3 Exponential

(19) $f(x; \theta) = \dfrac{1}{\theta} e^{-x/\theta}, F(x; \theta) = 1 - e^{-x/\theta}\ x = 1, 2, 3,\ E(X) = \theta, \mathrm{Var}(X) = \theta^2$

4.4 Normal distribution ($\mathbf{N}(\mu, \sigma^2)$)

(20) $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)/\sigma]^2/2}$, $E(X) = \mu, \mathrm{Var}(X) = \sigma^2$

Standard normal ($\mathrm{N}(0, 1)$)

$$z = \frac{x - \mu}{\sigma}$$

(21) $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

Part V
Dependencies and correlations

1 Independent events

Two events, A and B, are called independent if $P(A \cap B) = P(A)P(B)$; an equivalent formulation: $P(A|B) = P(A)$ ($P(B|A) = P(B)$).
A and B are independent if and only if A and B', A' and B, A' and B' are also independent.
The k events $A_1, A_2, \ldots, A_k$ are (mutually) independent if for every subset of distinct indices $j_1, \ldots, j_i$

$$P(A_{j_1} \cap \cdots \cap A_{j_i}) = P(A_{j_1}) \cdots P(A_{j_i})$$

Note that it is not sufficient to verify pairwise independence.

2 Joint distributions

The joint probability density function of k discrete random variables $X_1, \ldots, X_k$ is defined as $f(x_1, \ldots, x_2) = P[X_1 = x_1 \cap \ldots \cap X_k = x_k]$ for all vectors of possible values.

The joint cumulative distribution function is $F(x_1, \ldots, x_2) = P[X_1 \leqslant x_1 \cap \ldots \cap X_k \leqslant x_k]$ for all vectors of possible values.

For continuous random variables, the joint probability density function $f(x_1, \ldots, x_2)$ is a function defined by the following condition

$$F(x_1, \ldots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots t_k) dt_1 \cdots dt_k$$

Example 24. Nominative and ergative mechanisms of case marking. One cross-linguistic Bernoulli variable can be defined as the use of nominative-accusative mechanism for discrimination of core participants of transitive verbs; let's say that $Nom = 1$ if this mechanism is employed in a language and $Nom = 0$ otherwise. Accordingly, $Erg = 1$ if the ergative mechanism is employed and $Erg = 0$ otherwise. The joint distribution of these variables (estimated for the language population) can be represented by the following table:

	$Erg = 1$	$Erg = 0$	
$Nom = 1$	0.06	0.35	0.41
$Nom = 0$	0.11	0.48	0.59
	0.17	0.83	1

Example 25. Sampling for k-ary typology. Suppose we study intra-family distributions for a three-way typology, e.g. "nominative", "ergative" and "other". A family consists of 100 languages, 70 of them are nominative, 10 ergative and the other 20 do not fit into either of the clear cut types. For our study, we select 15 languages at random (without replacement). X_n is the number of nominative languages in our sample, X_e, the number of ergative languages. The joint probability function for the pair X_n, X_e is

$$f(x_n, x_e) = \frac{\binom{70}{x_n}\binom{10}{x_e}\binom{20}{15-x_n-x_e}}{\binom{100}{15}}$$

for all non-negative x_n, x_e, $x_n + x_e \leqslant 15$.

In the general case, suppose we have a (relatively) small population of N items, which are classified into k types (there M_i items of i-th type), and X_i is the number of items of i-th type in a sample of size n. The joint distribution function is the so-called extended hypergeometric distribution (note that there are only $k-1$ variables here):

$$f(x_1, \ldots, x_{k-1}) = \frac{\binom{M_1}{x_1} \cdots \binom{M_k}{x_k}}{\binom{N}{n}}$$

Example 26. Language change. Suppose that a certain linguistic type, A, can change, within a certain time interval, t, into k other types, with probabilities $p_1, p_2, \ldots, p_k$ ($\sum_{i=1}^{k} p_i < 1$). How many instances of each new type will be found in a set of n languages that have been in the state A t years ago? If X_i ($i = 1, \ldots, k$) is the number of instances of i-th new type, $x_{k+1} = n - \sum_{i=1}^{k} x_i$, the number of instances of the initial type, then

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k! x_{k+1}!} p_1^{x_1} \cdots p_k^{x_k} p_{k+1}^{x_{k+1}}$$

where $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$. This is the multinomial distribution (corresponding to the binomial distribution for a single variable). It describes the case of sampling with replacement and provides an approximation for extended hypergeometric distribution if the sample is much smaller than the population.

If the pair of discrete random variables has the joint pdf $f(x_1, x_2)$, then the marginal pdf's of X_1 and X_2 are

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2)$$
$$f_2(x_2) = \sum_{x_1} f(x_1, x_2)$$

The marginal pdf for any single variable in a multinomial distribution is binomial.

3 Independent random variables

Random variables $X_1, \ldots, X_2$ are independent if for every $a_i < b_i$,

$$P[a_1 \leqslant X_1 \leqslant b_1 \cap \cdots \cap a_k \leqslant X_k \leqslant b_k] = \prod_{i=1}^{k} P[a_i \leqslant X_i \leqslant b_i]$$

For discrete random variables, $f(x_1, \ldots, x_k) = \prod_i f_i(x_i)$.

For a pair of random variables X_1, X_2, the support set is the set of all pairs of values (x_1, x_2) for which $f(x_1, x_2) > 0$. They can be independent only if the support set is a Cartesian product.

Example 27. Dominant orders. Consider two random variables: variable X is the frequency of transitive clauses with S before O, and variable Y is "S precedes O in the basic word order". Are these variables independent? Why?

4 Covariance and correlations

4.1 Covariance

The covariance of a pair of random variables X and Y:

$$\mathrm{Cov}(X,Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

Also:

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$$

and $\mathrm{Cov}(X,Y) = 0$ if the variables are independent.

$$\mathrm{Var}(X_1+X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1,X_2)$$

$$E(X_1+X_2) = E(X_1) + E(X_2)$$

Example 28. OV and adpositions. (adapted from Dryer (1989)). The joint distribution of two cross-linguistic Bernoulli variables, *Post* and *OV*, is represented in the following table:

	$Post = 1$	$Post = 0$	
$OV = 1$	0.53	0.03	0.56
$OV = 0$	0.06	0.38	0.44
	0.59	0.41	1

$$E(OV) = 0.56, E(Post) = 0.59, E(OV \cdot Post) = 0.53$$

$$\mathrm{Cov}(OV, Post) = E(OV \cdot Post) - E(OV) \cdot E(Post) = 0.2$$

Example 29. Morphological complexity in nominal and clausal constructions. Nichol's considers her morphological complexity measure as a characteristic of the language as a whole, yet the actual complexity points come from two classes of constructions, nominal and clausal. It would be interesting to figure out whether or not the two random variables, the morphological complexity of NP (C_{NP}) and the morphological complexity of clause (C_{Cl}) are independent. Here is the summary of the data needed to answer this question:

$$E(C_{NP}) = 2.4, E(C_{Cl}) = 5.71, E(C_{NP} \cdot C_{Cl}) = 14.56$$

4.2 Correlation coefficient

The correlation coefficient of X and Y is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

The random variables X and Y are correlated if $\rho \neq 0$. If ρ is the correlation coefficient, $-1 \leqslant \rho \leqslant 1$, and $\rho = \pm 1$ if and only if $Y = aX + b$ with probability 1 for some $a \neq 0$ and b.

Example 30. Correlation between OV and Post. (continuation of Ex. 28). Recall that the covariance of OV and $Post$ is $\text{Cov}(OV, Post) = 0.2$. In order to calculate the correlation coefficient, we need to know variances of the two variables:

$$\text{Var}(OV) = 0.56 \cdot (1 - 0.56) = 0.25, \text{Var}(Post) = 0.59 \cdot (1 - 0.59) = 0.24$$

The correlation coefficient is:

$$\rho = \frac{\text{Cov}(OV, Post)}{\sigma_{OV}\sigma_{Post}} = 0.83$$

Example 31. Correlation of morphological complexity in nominal and clausal constructions. (continuation of Ex. 29) Calculate the correlation coefficient for C_{NP}, C_{Cl}, if

$$\text{Var}(C_{NP}) = 1.1, \text{Var}(C_{Cl}) = 6.4$$

5 Conditional distributions; conditional expectation and variance

If X_1, X_2 are random variables with joint pdf $f(x_1, x_2)$, then the conditional pdf of X_2 given $X_1 = x_1$ is

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

for $x_1 : f_1(x_1) > 0$ and zero otherwise.

The notions of expectation and variance can be extended to the conditional framework:

$$E(Y|x) = \sum_y y f(x|x)$$

$$\text{Var}(Y|x) = E\{[Y - E(Y|x)]^2|x\} = E(Y^2|x) - [E(Y|x)]^2$$

Example 32. Conditional expectation and variance of complexity in clausal constructions. In Exs. 29 and 31 we found that the morphological complexity in NP and in clausal constructions co-vary. Recall that $E(C_{Cl}) = 5.71$ and $\text{Var}(C_{Cl}) = 6.41$. The following table gives the conditional expected values for several values of C_{NP}:

i	$E(C_{Cl} \mid C_{NP} = i)$	$\text{Var}(C_{Cl} \mid C_{NP} = i)$
0	4.44	5.8
1	4.53	6.14
2	5.36	6.75
3	6.05	4.19
4	7	5.6

Exercises

Example 33. Head-marking and dependent-marking. Recall that morphological complexity points come from two major types of "locations" in various constructions, "head" and "dependent". The idea behind the classification of languages into "head-marking" and "dependent-marking" is that there is a strong general tendency to put markers either on the head or on the dependent constituent.

$$E(C_D) = 4.3, E(C_H) = 3.61, E(C_D \cdot C_H) = 13.12$$

What is the covariance of these variables?

Example 34. Head-marking and dependent-marking. (continuation of Ex. 33) Calculate the correlation coefficient for C_D, C_H, if

$$\text{Var}(C_D) = 8.99, \text{Var}(C_{Cl}) = 4.43$$

Part VI

An interim summary

1 Applicability of the logical structure of probability

1.1 Quantitative typology as a descriptive device

Example 35. Population and sample space. A common approach to typological sampling is to use genetic groupings as individuals; the experiment consists in selecting a single language from each genetic grouping at random and establishing its type. Suppose a study identifies a complete set of genetic groupings, $G_1, \dots, G_k$, and defines a variable, V, with two possible values, 0 and 1. The sizes of genetic groups are $N_1, \dots, N_k$, and the actual number of representatives of "$V = 1$" in these groups are $M_1, \dots, M_k$. What is the probability of having x representatives of "$V = 1$" in the sample? What are the expectation and variance of X?

$$f(x) = \sum_{\mathbf{A} \in \mathbf{S_x}} \prod_{i \in \mathbf{A}} \frac{M_i}{N_i} \prod_{j \notin \mathbf{A}} \frac{1 - M_j}{N_j}$$

where $\mathbf{S_x}$ is the set of all sets of x distinct indices.

$$E(X) = \sum_{i=1}^{k} \frac{M_i}{N_i}$$

$$\mathrm{Var}(X) = \sum_{i=1}^{k} \frac{M_i(1 - M_i)}{N_i^2}$$

Do we expect that different samples of this type will exhibit similar frequencies of type "$V = 1$"? What if study not the complete set of genetic groupings, but select r genetic groupings at random?

What is the sample space in this experiment? How can it be re-defined?

1.2 Correlations and implications

Michael Cysouw suggested that the class of typological phenomena referred to as implicational universals cannot be established by analysis of statistical data, i.e. it cannot be given a probabilistic sense. Is this really the case?

Definition. One value of binary variable A is **marked** with respect to another binary variable, B, if Var(B|A = 1) ≠ Var(B|A = 0). Then, the marked value of A if the one for which the conditional variance of B is less.

Definition. A dependency between A and B can be said to be **symmetric** with respect to A if neither value of A is marked with respect to B.

Aside from the case of independence ($P_B(1|A = 1) = P_B(1|A = 0)$), a symmetric dependency means that ($P_B(1|A = 1) = 1 - P_B(1|A = 0)$. Assuming that the notion of "positive" value can be given the same linguistic sense for both parameters (for example, "head-final" for OV and GenN), this means that the probability that both parameters have the same value does not depend on the value of A. That is, if we introduce a third variable, D, such that $D = 1$ if $A = B$ and 0 otherwise, then $P_D(1|A = 1) = P_D(1)$, i.e. D and A are independent.

Now a specific hypothesis of implicational universal can be formulated as follows:

A joint distribution of A and B counts as an implicational universal $A \to B$ if the following conditions are simultaneously met:

$$\mathrm{Var(B|A = 1)} < \mathrm{Var(B|A = 0)} \ \& \ P_B(1|A = 1) > P_B(1) \ \& \ \mathrm{Var}(A|B = 1) \geqslant \mathrm{Var}(A|B = 0)$$

There is an important distinction between the concept of absolute implicational universal and its extension to the case of statistical universals. An absolute implicational universal $A \to B$ is always equivalent to $B \to A$. For a stochastic implicational universal, this is not the case. More specifically, there are two statistically distinguishable types of dependencies that count as implicational universals $A \to B$ according to our definition: strong unidirectional implication: $A \to B$ and $B \to A$, and weak unidirectional implication: $A \to B$, but neither value of B is marked with respect to A.

Exercise: Here are two examples of “implicational” joint distributions; which is weak and which is strong?

	$A = 1$	$A = 0$
$B = 1$	0.26	0.32
$B = 0$	0.04	0.38

	$A = 1$	$A = 0$
$B = 1$	0.57	0.19
$B = 0$	0.06	0.18

2 Extending the population to infinity

Example 36. ‘Language (in)dependence’. (continuation of Ex. 18) Consider two languages L_1, L_2 that are genetically related at time depth t, i.e. t years ago their common ancestor, L, split into two language communities, L_1^* and L_2^*, and L_i is a descendant of L_i^*. The evolutions from L_1^* to L_1 and from L_2^* to L_2 can be considered as ‘trials’. Are these trials independent? Is the following true (if A_i is the event “L_i represents type A”)?

$$P(A_1 \cap A_2) = P(A_1) \cap P(A_2)$$

Example 37. “A language birth”. Suppose at some point in history there are N languages. Suppose there is a typology that classifies them into k types, so that i-th type is represented by n_i languages ($\sum_{i=1}^{k} n_i = N$). The likelihood of a language community splitting into two communities does not depend on the type to which the language belongs. Whichever community splits first, this adds a new language to the population, and this new language belongs to the same type as the ancestor language. After this, $N' = N + 1$, and $n_i' = n_i + 1$ for some $i = x$. What is the probability of the first new language added to the population being of type x? Does this pose a problem for construing of history as a procedure of random sampling?

Example 38. Areas as distinct populations. Recall that Matthew Dryer (1989) proposed to divide the world into five large (continental) areas, and to study the distributions of cross-linguistic variables for each area separately. His idea can be now reformulated as follows: if the non-linguistic processes in language populations work as a random sampling from the space of available possibilities (of sorts), then the cross-linguistic distributions must be similar for different populations (which can then be viewed as "samples" from the same infinite population). Dividing the world into large areas gives us five relatively large "samples" (he also samples "genera", rather than languages, but we will disregard this for now).

His raw data on representation of basic word orders is summarized in the following table.

	Africa	Eurasia	Australia	North America	South America	Total
SOV	22	26	19	26	18	111
SVO	21	19	6	6	5	57
Other	5	7	5	28	8	53
Total	48	52	30	60	31	221

Dryer's original idea was to distinguish two area-based 'events' for each pair of types ('more frequent' and 'less frequent'), which has some obvious disadvantages in terms of statistical testing (We will discuss it in more details later). On the other hand, it can hardly be the case that the non-linguistic processes work as 'random sampling' for some parameters but not for others. So, the fact that Dryer's test gives negative results for many parameters can as well be viewed as the negative answer to the question of whether these processes constitute an appropriate 'random sampling'. Does it mean that the linguistically significant probabilistic interpretations are doomed?

Part VII
Introduction to random processes

1 Markov processes

A Markov process is specified by a set of states, $S = \{s_1, s_2, ..., s_r\}$; each move (called a *step*) consists of moving from one state to another (or remaining in the same state). The probability of moving from state s_i to state s_j is denoted by p_{ij} and does not depend on the previous history of moves. These probabilities are called *transition probabilities*. The probability of remaining in state s_i is denoted by p_{ii}.

Example 39. An example of Markov chain. Consider a three-way word-order typology: there are the verb-initial type-state (V_i), the verb-final type-state (V_f), and all other languages (V_m); imagine that for a 1000-years "step", the probabilities of finding the language in a certain state depending on its state at the previous step are as follows:

$$P = \begin{array}{c} V_i \\ V_m \\ V_f \end{array} \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/8 & 3/4 & 1/8 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

1.1 Transition matrices. Probabilities for n steps

(22) $p_{jk}^{(2)} = \sum_{i=1}^{r} p_{ji} p_{ik}$

(23) $p_{jk}^{(n)} = \sum_{i=1}^{r} p_{ji}^{(n-1)} p_{ik}$

(24) $\mathbf{P}^n = \mathbf{P}^{n-1}\mathbf{P}$

(25) $p_{jk}^{(n)} = \sum_{i=1}^{r} p_{ji}^{(m)} p_{ik}^{(n-m)}$

(26) $\mathbf{P}^n = \mathbf{P}^m \mathbf{P}^{n-m}$

Distribution after n steps

(27) $\mathbf{u}^{(n)} = \mathbf{u}\mathbf{P}^n$

Example 40. The Ehrenfest two-urn model. (Ehrenfest Model) The Ehrenfest model has been used to explain diffusion of gases, yet it can also be construed as a model of totally random type-shifts. We have two urns and n balls (for the simplest example, let us take $n = 4$). At each step, one of the four balls is chosen at random and moved from the urn that it is in into the other urn. The states are the number of balls in the first urn. The transition matrix is then:

$$P = \begin{array}{cc} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

2 Absorbing Markov chains

A state s_i of a Markov chain is called *absorbing* if it is impossible to leave it (i.e., $p_{ii} = 1$). A Markov chain is *absorbing* if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state (not necessarily in one step). In an absorbing Markov chain, a state which is not absorbing is called *transient.*

Example 41. Nominal conjunction as absorbing state. At the present time, languages can be classified into two types, those that have a nominal conjunction construction and those that do not (a comitative (WITH-like) construction is used instead). A possible hypothesis is that the state with nominal conjunction is absorbing, i.e. once a language acquires a nominal conjunction construction, it never loses this construction type. The transition matrix thus looks as follows:

	+Conj	-Conj
+Conj	1	0
-Conj	α	$1 - \alpha$

Example 42. Rise and fall of language families. Consider the following simple model of the rise and fall of language families. We consider the number n of languages in a family ($n = 1, ..., M$) as the states of the process. There is a certain probability $\lambda(n)$ of a single new language appearing by virtue of language split (where n is the current number of languages, $\lambda(M) = 0$). There is another probability $\mu(n)$, of a single language disappearing. Obviously, if the last language disappears, there can be no more splits, so $\lambda(0) = 0$. $n = 0$ is the absorbing state.

In an absorbing Markov chain, the probability of absorption is 1.

3 Ergodic Markov chains

3.1 Introductory

A Markov chain is called *ergodic* (or *irreducible*) if it is possible to reach every state from every other state (not necessarily in one step).

Example 43. A reducible chain of reciprocity. Languages can be classified into five major types depending on how they express reciprocity: N — languages without conventionalized reciprocal constructions, B – languages with reciprocal constructions where reciprocal participants must be referred to in two distinct syntactic slots, R – languages where reciprocity is conventionally expressed by the reflexive construction, I – languages with a non-reflexive reciprocal construction, and M, languages with both reflexive and non-reflexive constructions. There are reasons to assume that there are no diachronic path from any of the last three states to either of the first two; that is, once a language evolves a reciprocal construction that does allow for all reciprocal participants to be referred to in a single slot, it will always have such a construction (possibly another one). The corresponding transition matrix can look like this:

$$\mathbf{P} = \begin{array}{c} N \\ B \\ R \\ I \\ M \end{array} \begin{pmatrix} 0.7 & 0.1 & 0.1 & 0.1 & 0 \\ 0 & 0.7 & 0.1 & 0.1 & 0.1 \\ 0 & 0 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0.1 & 0.2 & 0.7 \end{pmatrix}$$

This chain is neither ergodic (it is impossible, e.g., to get from R to B), nor absorbing (there are no states from which there is no way out). As far as its behaviour is concerned, it can be represented as a combination of two chains, one absorbing and one ergodic.

First, let us define a new type-state $U = \{R, I, M\}$, which includes all states that can be reached from every other state. In the new transition matrix, this state is absorbing:

$$\mathbf{P_A} = \begin{array}{c} N \\ B \\ U = \{R, I, M\} \end{array} \begin{pmatrix} 0.7 & 0.1 & 0.2 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 1 \end{pmatrix}$$

In other words, after some time the process will be absorbed by the generalized state, U. The original changed is thereby *reduced* to an ergodic chain:

$$\mathbf{P_E} = \begin{array}{c} R \\ I \\ M \end{array} \begin{pmatrix} 0.7 & 0 & 0.3 \\ 0 & 0.7 & 0.3 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$$

3.2 A first look at the long-term behaviour

The fundamental property of ergodic chains is that if we consider a sequence of transition matrixes for n steps of the process, with increasing n, i.e. $\mathbf{P}, \mathbf{P}^1, \mathbf{P}^2, \mathbf{P}^3$, etc., the rows of the matrix gradually become more and more similar to one another and, ultimately, identical. Consider, for example, the powers of the ergodic "reciprocal" matrix from our example:

$$\mathbf{P} = \begin{pmatrix} 0.7 & 0 & 0.3 \\ 0 & 0.7 & 0.3 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.52 & 0.06 & 0.42 \\ 0.03 & 0.55 & 0.42 \\ 0.14 & 0.28 & 0.58 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.41 & 0.13 & 0.47 \\ 0.06 & 0.47 & 0.47 \\ 0.16 & 0.31 & 0.53 \end{pmatrix}$$

$$\mathbf{P}^3 = \begin{pmatrix} 0.33 & 0.18 & 0.49 \\ 0.09 & 0.42 & 0.49 \\ 0.16 & 0.32 & 0.51 \end{pmatrix}$$

$$\mathbf{P}^4 = \begin{pmatrix} 0.28 & 0.22 & 0.49 \\ 0.11 & 0.39 & 0.49 \\ 0.16 & 0.33 & 0.51 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.25 & 0.26 & 0.5 \\ 0.13 & 0.37 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} 0.22 & 0.28 & 0.5 \\ 0.14 & 0.36 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^6 = \begin{pmatrix} 0.21 & 0.3 & 0.5 \\ 0.15 & 0.35 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^7 = \begin{pmatrix} 0.19 & 0.31 & 0.5 \\ 0.15 & 0.35 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^8 = \begin{pmatrix} 0.19 & 0.31 & 0.5 \\ 0.16 & 0.34 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^9 = \begin{pmatrix} 0.18 & 0.32 & 0.5 \\ 0.16 & 0.34 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^{10} = \begin{pmatrix} 0.18 & 0.32 & 0.5 \\ 0.16 & 0.34 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^{11} = \begin{pmatrix} 0.17 & 0.33 & 0.5 \\ 0.16 & 0.34 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^{12} = \begin{pmatrix} 0.17 & 0.33 & 0.5 \\ 0.16 & 0.34 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^{13} = \begin{pmatrix} 0.17 & 0.33 & 0.5 \\ 0.17 & 0.33 & 0.5 \\ 0.17 & 0.33 & 0.5 \end{pmatrix}$$

3.3 The unique stationary solution

The fact that $\mathbf{P}^n$ tends to a limiting vector as n approaches ∞ guarantees that there is a unique probability vector with the following property:

$$\mathbf{wP} = \mathbf{w}$$

We can find the limiting vector w for the verb-based word order typology of Example 39 from:

$$w_1 + w_2 + w_3 = 1$$

and

$$(w_1 w_2 w_3) \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/8 & 3/4 & 1/8 \\ 0 & 1/2 & 1/2 \end{pmatrix} = (w_1 w_2 w_3)$$

These relations lead to the following four equations in three unknowns:

$$\begin{gathered} w_1 + w_2 + w_3 = 1 \\ \frac{1}{4}w_1 + \frac{1}{8}w_2 = w_1 \\ \frac{1}{2}w_1 + \frac{3}{4}w_2 + \frac{1}{2}w_3 = w_2 \\ \frac{1}{4}w_1 + \frac{1}{8}w_2 + \frac{1}{2}w_3 = w_3 \end{gathered}$$

If the equations are solved, we obtain the unique solution:

$$\mathbf{w} = (\frac{1}{9}\ \frac{2}{3}\ \frac{2}{9})$$

Exercises

1. Calculate the limiting probabilities for the ergodic reciprocal chain from Ex. 43.
2. Show that for any two-state process,

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

   the limiting vector is:

$$\mathbf{w} = (\frac{\beta}{\alpha+\beta}\ \frac{\alpha}{\alpha+\beta})$$

3.4 Time spent in a state

As n approaches infinity, the time (the number of steps) the process spends in each state becomes proportional to the limiting probability of this state.

Let $\mathbf{P}$ be the transition matrix for an ergodic chain. Let $\mathbf{A}^n$ be the matrix defined by

$$\mathbf{A}^n = \frac{\mathbf{I} + \mathbf{P} + \mathbf{P}^2 + \cdots + \mathbf{P}^n}{n+1}.$$

Then $\mathbf{A}^n \to \mathbf{W}$, where $\mathbf{W}$ is a matrix all of whose rows are equal to the unique fixed probability vector $\mathbf{w}$ for $\mathbf{P}$.

Assume that we have an ergodic chain that starts in state s_i. Let $X(m) = 1$ if the m-th step is to state s_j and 0 otherwise. Then the average number of times in state s_j in the first n steps is given by

$$H^{(n)} = \frac{X(0) + X(1) + X(2) + \cdots + X(n)}{n+1}$$

But $X(m)$ takes on the value 1 with probability $p_{ij}^{(m)}$ and 0 otherwise. Thus $E(X(m)) = p_{ij}^{(m)}$, and the ij-th entry of $\mathbf{A}^n$ gives the expected value of $H^{(n)}$, that is, the expected proportion of times in state s_j in the first n steps if the chain starts in state s_i.

Example 44. Reconstructing transition matrices. Imagine there is a pair of binary parameters known to correlate in the language population, e.g. VO vs. OV and Postpositions vs. Prepositions: the values of the first parameter are roughly equiprobable, almost (but not all) VO languages are prepositional, and almost (but not all) OV languages are postpositional. Construct a Markov process for this typology, in such a way that its limiting vector predicts the observed properties of the distribution. Check whether you succeeded by calculating the limiting vector for your transition matrix. What linguistic hypotheses are incorporated into your transition matrix?

Part VIII

Non-linguistic random processes in the language population

1 Birth-and-death process

Assume there is a certain probability of community split ("language birth"), p_b, and a certain probability of population shift ("language death"), p_d within a short time interval, Δt.

Exercise questions.

1. If there are n languages at t_0, what is the expected number of languages at $t_1 = t_0 + \Delta t$?
2. What is the probability of the total number of languages increasing by k?
3. What is the probability that total number of languages decreasing by k?

In order to describe the long-term effects of birth-and-death, we'll have to switch to processes with continuous time (instead of discrete "steps"). For this, the transition probabilities are replace by the probabilities of change for any time interval, from t to $t+s$, $P_{jk}(s)$. $P_{jk}(s)$ is the conditional probability of the state E_k at time $t+s$ under the condition that the process was in the state E_j at time $t < t+s$. It is assumed that this probability depends only on the length of the time interval, but not on the specific value of t. Such a process is called time-homegenous. The continuous-time counterpart of the equations describing probabilities for n steps is Kolmogorov-Chapman equation:

(28) $$P_{jk}(x+y) = \sum_i P_{ji}(x)P_{ik}(y)$$

Example 45. Feller-Arley process. Switching to a process with continuous time can be though of, very informally, decreasing Δt. As the time interval approaches 0, the probability of the population size changing by more than one tends to 0, too. This gives the Feller-Arley model, a model of linear birth-and-death process, a continuous-time Markov process with the following transition rates (Feller 1971:454-457, Srinivasan & Mehata 1978):

$$\begin{array}{lll} q_{n,n+1} = n\lambda & \text{for} & n > 0 \\ q_{n,n-1} = n\mu & \text{for} & n > 0 \\ q_{n,m} = 0 & \text{for} & n \geq 0, m \geq 0, m \neq n \pm 1, \end{array}$$

where λ and μ are probability densities for birth and death respectively. Probabilities $p_n(t|1)$ for a language to have n descendants by the end of time interval t are given by the following expressions:

(29) $$p_n(t|1) = (1-a)(1-b)b^{n-1} \quad \text{for } n > 1,$$

(30) $$p_0(t|1) = a,$$

where $a = (\mu e^{(\lambda-\mu)t)} - \mu)/(\lambda e^{(\lambda-\mu)t)} - \mu)$, $b = \lambda a/\mu$.

Probabilities for a population with initial size N_0 to have n members by the end of time interval t are:

$$(31)\quad p_n(t|N_0) = a^{N_0} b^n \sum_{j=0}^{\min(N_0,n)} \binom{N_0+n-j-1}{n-j}\binom{N_0}{j}(\frac{1-a-b}{ab})^j.$$

The mean value of population size $N(t|N_0)$ and the variance are given by the following equations:

$$(32)\quad \mathrm{Exp}[N(t|N_0)] = N_0 e^{(\lambda-\mu)t},$$

$$(33)\quad \mathrm{Var}[N(t|N_0)] = N_0 \frac{\lambda+\mu}{\lambda-\mu} e^{(\lambda-\mu)t}(e^{(\lambda-\mu)t} - 1).$$

The expected number of ancestor languages that will have at least one descendant is:

$$(34)\quad K(t|N_0) = N_0(1 - p_0(t|1)).$$

Example 46. Birth-and-death process and $\mathcal{V}(t)$. Assume we are interested in one particular aspect of $\mathcal{V}(t)$, the frequency of a certain type. At t_0, it is represented by m languages, of the total of n languages. How will the frequency change at $t_1 = t_0 + \Delta t$?

2 A simple contact-based model

Assume there are n languages in a certain area, n_X of them are of type X and n_Y are of type Y ($n_X + n_Y = n$). Assume, further, that there is a certain probability γ of a language being influenced by another language (from the same area). If a language is influenced by a language of the same type, the corresponding property does not change. If it is influenced by a language of the other type, it changes with the probability α after a certain time interval (taken as a "step").

What is the expected number of languages of type X after a single time step?

Part IX

Language change as a random process. Ergodic hypothesis

1 Types as states of a random process

Example 47. Life-times of types. Consider the following question: if a language is in a certain typological state E (e.g. it has SOV as its basic word order or a nominative-accusative transitive construction), what is the probability $p_s(t)$ that it will still be in the same state after some time t? Let's begin with considering this question in terms of discrete 'time steps', i.e. we want to describe $p^{(k+l)}(E|S(k) = E)$ (the probability that the language is in the state E at the $k + m$-th step given that it was in this state at k-th step. To put it in other terms, we want to describe the probability that exactly x steps will pass before the first type-shift. Let X, the number of steps before the first type-shift be our random variable.

From what we know about languages, this probability cannot depend on how long the language has already spent in this state before the k-th step (the speakers have no access to this information). Thus, the process has a so-called "no-memory" property:

$$P(X > j + k|X > k) = P(X > k)$$

The only discrete distribution that has this property is the geometric distribution, often described in terms of the number of trials before the first success (for a given probability of success, p). For an event $[X = x]$ to occur, it is necessary to have exactly $x - 1$ successful transmissions (from step i to step $i + 1$) followed by a single success. Thus,

$$g(x; p) = p(1 - p)^{(x-1)}$$

(for our example, p is the probability of change).

If we switch to continuous time model, the no-memory property is reformulated as follows:

$$P(X > a + t|X > a) = P(X > t)$$

This property is satisfied if and only if X has the exponential distribution:

$$f(x; \lambda) = \lambda e^{-\lambda x} \; x > 0$$

where λ is the rate of change (the expectation of the life time is $1/\lambda$ and variance $1/\lambda^2$).

Example 48. The number of changes within a time interval. Let $X(t)$ denote the number of language changes within a given time interval $[0, t]$. It seems reasonable to assume that the probability that a change will occur within a given short time interval $[t, \Delta t]$ is approximately proportional to Δt, and that the occurrences of events in non-overlapping time intervals are independent; if the probability of two events occurring within the same Δt is negligible if $\Delta t \to 0$, than $X(t)$ follows the Poisson distribution:

$$P_n(t) = P[X(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

,

with $E(n) = \lambda t$ and $\mathrm{Var}(n) = \lambda t$.

The same model can be applied to various kind of "external hits" affecting a language or its subsystem (including language contact).

2 Randomness in language change: propagation

Example 49. Propagation via communication. Imagine a community including of n_1 speakers with one variant of a sociolinguistic variable (v_1) and n_2 speakers with the other variant (v_2).

Assuming that each speaker is equally likely to communicate with any other speaker, what is the probability that an act of communication involves speakers from different groups, given that it does take place?

What is the probability of a v_1-speaker communicating with a v_2-speaker?

It may be assumed that the probability that a speaker will switch to the other variant is a function of the number of times it encounters this other variant (along with other parameters). How will this probability depend on the values of n_1 and n_2?

3 Randomness at the level of mental grammars

Reanalisis changes the underlying structure of a pattern, but does not involve immediate visible modifications. For a reanalysis to take place, a subset of the tokens of a particular constructional type must be open to the possibility of multiple analyses. One of them is old, hence applicable to all tokens, the other is new and applicable to a subset of tokens. A fundamental feature of reanalysis is that it is distinct from and must happen before *actualization*, i.e. before any visible consequences of the novel analysis occur.

Example 50. Reanalysis: first stage. The deterministic approach would be to assume that, given a fully specified linguistic environment, including the basis for reanalysis (i.e. the pattern open to multiple analyses) and all other relevant aspects of grammar, the novel analysis either necessarily arises or does not arise at all. This would be simply pushing randomness to the previous process of language change (which creates one or another environment), let alone the problem of certain randomness within the environment (different speakers have slightly different inputs). The probabilistic approach would be to assume that there is a certain probability of the novel analysis being created within a mental grammar, which is quite likely a function of certain parameters of the linguistic environment, including, for example, the relative frequency of the tokens open for multiple analysis among all occurrences of the construction type. An interesting property of this stage is that all that happens happens within each mental grammar: the rise of a novel analysis in a mental grammar is invisible to other speakers and thus cannot influence other mental grammars. In other words, the events of a novel analyses being created in different mental grammars are independent of one another; there is a single probability of a novel analysis, say $\alpha(\theta)$, where θ stands for the relevant parameters of the environment. What is, then, the distribution of the number of speakers having the novel analysis?

Example 51. Reanalysis: actualization. If the novel analysis is present in the mental grammar, it can be actualized in a certain novel visible properties. Assume that, in a linguistic environment characterized by certain values parameters θ_1, a passive construction is analyzed as ergative with the probability $\alpha_1 = \alpha(\theta_1)$. One implication of this behavior might be, e.g., dropping the novel subject (A) in the context of clausal conjunction (when it is co-referent with the A/S of the first clause), which would be impossible under the passive analysis. Let us assume that the conditional probability of doing this, for a speaker with double analysis, in the appropriate context is β.

Obviously, the occurrences of such constructions must increase the likelihood of novel analysis, and the higher the relative frequency d of tokens compatible with the novel analysis only, the higher must be the probability of the ergative analysis. Let us assume a very simple dependency:

$$\alpha = \alpha_1 \text{ if } d < \alpha_1$$

$$\alpha = d \text{ if } \alpha_1 < d < \frac{1}{2}$$

$$\alpha = 1 \text{ if } d \geqslant \frac{1}{2}$$

How the process of reanalysis will develop for different values of α_1 and β?

Example 52. Complexity measures and performance preferences. Hawkins develops various measures of syntactic complexity, which are supposed to account both for language-internal performance preferences and for some aspects cross-linguistic variation. For example, a certain construction can be compatible with a range of different syntactic slots, yet the resulting patterns will be characterized by different values of complexity; an example of this is the hierarchy of accessibility to relativization, where the complexity of pattern appears to increase as we move from the higher points in the hierarchy to the lower points:

S > DO > IO > Gen

In performance, the speakers would avoid complex constructions, and so the frequency of occurrence, within a single language, will decrease as the complexity grows. That is, in contexts which allow for several coding options, the likelihood for a speaker to avoid a construction will increase with its complexity.

$$f(\mathit{RelS}) = \rho > f(\mathit{RelO}) = \frac{1}{2}\rho > f(\mathit{RelIO}) = \frac{1}{3}\rho > f(\mathit{RelGen}) = \frac{1}{4}\rho$$

On the other hand, there are reasons to believe that the probability of a construction surviving in a language depends on its frequency. A simple model would assume that if the frequency f of a construction falls below a certain threshold, its transmission to the next generation of speakers is not certain, but occurs with a certain probability $\alpha(f) < 1$.

Part X

Random processes and cross-linguistic distributions

1 Regular differences between types of populations

1.1 Targeting different genetic depths

Example 53. Basic word order: "languages" vs. "genera". Tomlin and Dryer study essentially the same cross-linguistic variable ("basic word order"), yet in different populations: Dryer (with some qualifications) chooses a single language per genus (ca. 3500-4000 years of time depth), and Tomlin uses a much shallower type of genetic grouping (ca. 1000 years). The major differences between the results are summarized in the following table:

	Absolute frequencies		Relative frequencies	
	Tomlin (*m*-languages)	Dryer (genera)	Tomlin (*m*-languages)	Dryer (genera)
SOV	180	111	0.45	0.5
SVO	168	57	0.42	0.26
Other	54	53	0.13	0.24
Total	402	221	1	1

In Tomlin's population SOV and SVO are nearly equiprobable (all other types together being much less frequent), whereas in Dryer's population, SVO is almost twice less frequent than SOV, and other types together are almost as frequent as SVO. Wherefore the difference?

1.2 Limiting cases: maximum mobility vs. minimal mobility

Let us begin by considering two limiting cases in terms of the overall rate of change along the parameter. In the framework we have discussed, these limiting cases can be defined in terms of the number of steps needed to achieve stabilization of the distribution (the fixed vector of the transition matrix). These limiting cases need not be realistic.

The minimum number of steps required to achieve stabilization is 1: this situation would obtain if all rows in the transition matrix (for a relatively small time interval) were equal. For example, the state the language is in because independent of its previous state after a single 100-year step. Then, the birth-and-death process can have no effect whatsoever, and the difference between populations is inexplicable within our framework (they should be identical as far as cross-linguistic distribution of a mobile parameter is concerned).

The maximum stability is achieved if all states are absorbing, i.e. there can no change at all, and the dependency on the initial state lasts forever. Then, the difference between the populations must be due to the birth-and-death process alone. Is that likely? We can use the properties of the birth-and-death process with continuous time, as discussed before, to answer that question.

The frequency $f(t|N_0, f_0)$ of a linguistic trait with initial frequency f_0 at time t can be represented as a function of two independent variables, corresponding to the size of two populations (for two types).

$$f(t|N_0, f_0) = \frac{N(t|N_0^+)}{N(t|N_0^+) + N(t|N_0^-)}, \tag{35}$$

where $N_0^+ = N_0 f_0, N_0^- = N_0(1 - f_0)$, under the condition that at least one language survives by the end of time interval t. We have already described the distribution of the population-size variables.

The expectation of frequency $f(t|N_0, f_0)$ equals f_0, and its variance can be estimated as:

$$\mathrm{Var}[f(t|N_0, f_0)] \cong \frac{\lambda + \mu}{\lambda - \mu} \frac{f_0(1 - f_0)}{N_0} \frac{e^{(\lambda-\mu)t} - 1}{e^{(\lambda-\mu)t}} \tag{36}$$

Using

$$K(t|N_0) = N_0(1 - p_0(t|1)). \tag{37}$$

where $p_0 = (\mu e^{(\lambda-\mu)t)} - \mu)/(\lambda e^{(\lambda-\mu)t)} - \mu)$, we obtain a simplified estimate:

$$\mathrm{Var}[f(t|N_0, f_0)] = \frac{f_0(1 - f_0)(\lambda + \mu)}{K\lambda} \tag{38}$$

Note that the variance increases as μ approaches λ and decreases as the size of initial population grows. In reality, the estimates of these values (which we do not know) are closely related to one another, so we can, for the sake of simplicity, switch to the birth-only process. This gives us the following rule of thumb:

With a probability more than 0.95, the deviation of frequency $f(t)$ in a descendant population from its value in the ancestor population is less than $\sqrt{\frac{f_0(1-f_0)}{K}}$, where K is the number of genetic groupings of the time depth t.

For our word order example, this means that the difference in the frequency of SOV in two populations can be “accidental”, whereas the difference in the frequency of SVO is extremely unlikely to have been brought about by the birth-and-death process alone.

1.3 Analysis of expectations: apparent time

Example 54. Case alignment in lexical NPs. The figures before the arrows are mean values of relative frequencies of alignment types for several samples of mutually isolated languages (sampling at the deepest depth of genetic affiliation, so called “stocks”. The figures after the arrows are relative frequencies of the same types in a random sample from the population. The boldface highlights the most striking differences.

	Nominative	**Split**	**Ergative**
Consistent	**0.17 → 0.22**	0.02 → 0.05	**0.16 → 0.09**
Differential	0.10 → 0.13	0.02 → 0.01	0.02 → 0.02
Neutral		0.5 → 0.48	

The effect we have discussed for word order is observed here for the nominative-ergative dimension. Again, if, say, the ergative alignment type is less stable than the nominative alignment, i.e. if there are systematic differences in transition probabilities, then there will be more languages that will have changed their alignment type among the descendants of ergative ancestors than among the descendants of nominative ancestors. As a result of this difference, the frequency of ergative languages in the modern language population will have decreased (which is what we actually observe), whereas the sample of isolated languages is more likely to represent an earlier distribution.

Note that, in this example, we also observe a variable for which there is no difference between two populations, the variable associated with the neutral alignment. The most likely explanation within our framework is that the actual distribution has already stabilized, i.e. it is close to the fixed limiting vector of the (yet unknown) transition matrix.

1.4 Actual values of expectations of family-level values

Example 55. Family-internal frequencies of uncharacteristic alignment values. Table 3 represents our estimates of the family-internal frequencies of uncharacteristic values for three "weak" binary variables, [+Nom], [+Erg], and [+Neu]. Here A stands for the positive value of a 'weak' binary variable, the presence of some nominative, ergative, and neutral features; $B = A'$.

$A =$	Nominative	Ergative	Neutral
Frequency of B-languages in A-families	0.14	0.18	0.17
Frequency of A-languages in B-families	0.17	0.03	0.25

Do these figures corroborate our "apparent time" hypothesis for alignment?

2 Divergence rates

Example 56. The probability of divergence. Imagine there is a binary typological variable with the following transition matrix:

$$\mathbf{P} = \begin{pmatrix} 1-p_{01} & p_{01} \\ p_{10} & 1-p_{10} \end{pmatrix}$$

Imagine two languages that are in the same state at the k-th step. What is the probability $P(D)$ of them being in different state at the $k+1$-th step?

(39) $P^{k+1}(D) = 2p^{(k)}(E_0)(1-p_{01})p_{01} + 2p^{(k)}(E_1)(1-p_{10})p_{10}$

Example 57. Divergence rate and p^{k+1}. Continuing Example 56, $p^{(k+1)}(E_1)$ is given by

(40) $p^{(k+1)}(E_1) = p^{(k)}(E_1)(1-p_{10}) + p^{(k)}(E_0)p_{01} = p^{(k)}(E_1)(1-p_{10}) + (1-p^{(k)}(E_1))p_{01}$

Accordingly,

(41) $$p^{(k)}(E_1) = \frac{p^{(k+1)}(E_1) - p_{01}}{1 - p_{10} - p_{01}}$$

Combining (39) and (41), we get the following linear dependency between $p(E_1)$ and $P(D)$ at the same step of the process:

(42) $$P(D) = ap(E_1) + b, \quad \text{where} \;\; a = 2(p_{10} - p_{01}), b = 2p_{01}(1 - p_{10})$$

Example 58. Divergence of Alignment. Suppose we have the following estimates for $P(E_1)$ and $P(D)$ for different alignment variables and different sub-populations. How can we obtain estimates for transition probabilities?

	E_1 = **Neutral**		E_1 = **Nominative**		E_1 = **Ergative**	
	$P(E_1)$	$P(D)$	$P(E_1)$	$P(D)$	$P(E_1)$	$P(D)$
I.	0.85	0.20	0.45	0.26	0.62	0.56
II.	0.11	0.20	0.05	0.13	0.2	0.05

3 Conclusion: empirical evidence for stochastic regularities of language change

We have looked at various kinds of statistical cross-linguistic data in order to figure out how it can be interpreted if language change is viewed as a genuine random process. Do these data provide empirical evidence for this kind of modeling? They do, insofar as the inferences agree with each other: e.g. the large-scale drift (for a rather long time interval) and the divergence rates (for a shallow time depth) indicate the same type of systematic differences in transition probabilities.

Most important is the similarity between the limiting distribution predicted on the basis of transition probabilities (obtained by measuring the divergence rate) and the actual synchronic distribution. We will consider the degree of their similarity in detail later; for now, focus on the most obvious: the variable associated with the neutral encoding, which has obviously stabilized at the limiting distribution determined by the identity of transition probabilities (presumably because the overall rate of change along this dimension is greater). Since the last period in the history of language population could not have been long enough to stabilize the distribution, it means that the transition probabilities have remained constant over a much longer period of time.

Part XI
Limiting distributions, convergence, laws of large numbers

A function of observable random variables, which does not depend on any unknown parameters, is called a **statistic**. The statistic is also a random variable, the distribution of which depends depends on the form of the function and on the distribution of the original random variables. The distribution of a statistic is often called derived, or sampling, distribution (in contrast to the population distribution).

1 Limiting distributions and stochastic convergence

Example 59. Accidental absolute universals. Let $\{T_1, \ldots, T_n\}$ be a (potentially infinite) set of definable linguistic properties ("types"), and $p_i = P(T_i)$, their probabilities (e.g. proportions in a certain population). One way in which p_i might be "random" (and linguistically irrelevant) can be described in terms of the (continuous) uniform distribution: $p_i \sim \mathrm{UNIF}(0, 1)$.

$$F_i(p) = F(p) = p.$$

The proportions $p_i, \ldots, p_n$ for a sequence of randomly selected linguistic properties can then be viewed as a random sample from this uniform distribution. Although this does involve obvious oversimplifications, this can be considered a model of certain aspects of typological research over the decades.

Now let U_n be the largest observed value for the given sequence sequence of $p_1, \ldots, p_n$ (this is called the largest order statistic). The cumulative distribution function of U_n is then:

$$G_n(p) = p^n \quad 0 < p < 1,$$

zero if $p \leqslant 0$ and one if $p \geqslant 1$. As n approaches ∞, $G_n(p)$ approaches 0 for $p < 1$ and 1 for $p \geqslant 1$:

$$(43) \quad G(p) = \begin{cases} 0 & p < 1 \\ 1 & p \geqslant 1 \end{cases}$$

.

A function like that defined by 43, where the distribution of a random variable is concentrated at one value ($c = 1$), is the CDF of a degenerate distribution. If $Y_n \sim G_n(y)$ and if for some $G(y)$

$$\lim_{n\to\infty} G_n(y) = G(y)$$

the sequence $Y_1, \ldots, Y_n$ **converges in distribution** to $Y \sim G(y)$. The distribution corresponding to $G(y)$ is called the limiting distribution of Y_n. If the limiting distribution is degenerate at $y = c$, the sequence **converges stochastically** to c.

Example 60. Rara and rarissima. Let $X_1, ..., X_n$ be a random sample from a Bernoulli distribution, $X_i \sim \text{BIN}(1, p)$, and consider $Y_n = \sum_{i=1}^n X_i$. If we let $p \to 0$ as $n \to \infty$ in such a way that $np = \mu$ for a fixed $\mu > 0$, then Y_n converges in distribution to $Y \sim \text{POI}(\mu)$, i.e. $f(y) = \frac{e^{-\mu}\mu^y}{y!}$ (with the mean and the variance equal to μ).

2 The central limit theorem and normal approximations

2.1 (Recalling) normal distribution

The normal distribution $\text{N}(\mu, \sigma^2)$ with mean μ and variance σ^2 has the probability density function:

$$(44) \quad f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(x-\mu)/\sigma]^2/2}$$

The standard normal distribution ($\text{N}(0, 1)$) results from the following transformation of a normal distribution:

$$(45) \quad z = \frac{x - \mu}{\sigma}$$

$$(46) \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The sum of independent normally distributed variables is normally distributed with $\mu = \sum_i \mu_i$, $\sigma^2 = \sum_i \sigma_i^2$. The special case of a random sample from a normally distributed population ($\text{N}(n\mu, n\sigma^2)$); the **sample mean** is then normally distributed with mean μ and variance σ^2/n.

2.2 The central limit theorem

Central limit theorem. If $X_1, ...X_n$ is a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

is the standard normal distribution, $Z \sim \text{N}(0, 1)$ as n approaches ∞.

Example 61. Birth-and-death effects and the population size. The size S of a genetic grouping of a fixed time-depth t appears to follow the Pareto (power-law) distribution.

$$(47) \quad f(s; \theta, k) = \frac{k}{\theta}(1 + \frac{x}{\theta})^{-k+1}$$

This distribution is as far from normal as can be. However, if we are interested in the influence of the birth-and-death process on the proportions of language types of the populations, we are interested not in the size of a single grouping, but rather in the sum $X_n = S_1 + \cdots + S_n$ of sizes of all n groupings whose ancestors belonged to a certain type.

The CLT tells us that as n grows, the distribution of X_n becomes closer and closer to normal, and its variance decreases proportionally to n. This is the fundamental reason why the birth-and-death effects do not produce strong effects in large populations.

Example 62. Genetically-isolated samples. In Example 35, we considered the number of representatives of a certain type (X) in a sample of containing a single language from each genetic grouping from a certain pre-established set. In particular,

$$E(X) = \sum_{i=1}^{k} \frac{M_i}{N_i}$$

Now, $Y_i = \frac{M_i}{N_i}$ themselves are independent random variables drawn from an unknown distribution $f(y)$ determined by the interaction of random process in the language population. What is the distribution of $Z = E(X)$ and how its parameters depend on the properties of $f(y)$?

Example 63. Sample proportions. Does the CLT apply to sample proportions? Why and how?

2.3 Approximations for binomial distribution

For large n and fixed p, $\mathrm{BIN}(n, p)$ is approximately $Y_n = \mathrm{N}(np, npq)$. This approximation works best when p is close to 0.5 (one guideline is to use the normal approximation when $np \geqslant 5$ and $nq \geqslant 5$.

Example 64. SOV. Assume that the probability that a modern language is in the SOV-state is 0.5. If we randomly select 20 languages, what is the probability that at least nine of them are SOV? The exact probability is

$$P[Y_{20} \geqslant 9] = 1 - P[Y_{20} \leqslant 8] = 1 - \sum_{y=0}^{8} \binom{20}{y} 0.5^y 0.5^{20-y} = 0.7483$$

A normal approximation is

$$P[Y_{20} \geqslant 9] = 1 - P[Y_{20} \leqslant 8] = 1 - \Phi(\frac{8 - 10}{\sqrt{5}}) = 1 - \Phi(-0.89) = 0.8113$$

Continuity correction. Each binomial probability, $b(y; n, p)$, has the same value as the area of a rectangle of height $b(y; n, p)$ and the interval $[y-0.5, y+0.5]$ as its width. The area of this rectangle can be approximated by the area under the probability density function of $Y \sim \mathrm{N}(np, np(1-p))$ (with the same interval as its base):

$$P[a \leqslant Y_n \leqslant b] = \Phi(\frac{b + 0.5 - np}{\sqrt{npq}}) - \Phi(\frac{a - 0.5 - np}{\sqrt{npq}})$$

For example,

$$x = b(7; 20, 0.5) = 0.0739$$
$$x \approx \Phi(\frac{7.5 - 10}{\sqrt{5}}) - \Phi(\frac{6.5 - 10}{\sqrt{5}}) = \Phi(-1.12) - \Phi(-1.57) = 0.0732$$

If the same idea is applied to our original problem, we get an approximation which is much closer to the exact value:

$$P[Y_{20} \geqslant 9] = 1 - P[Y_{20} \leqslant 8] = 1 - \Phi(\frac{8.5 - 10}{\sqrt{5}})1 - \Phi(-0.67) = 0.7486$$

3 Laws of Large Numbers

The Law of Large Numbers. If $X_1, ..., X_n$ is a random sample from the distribution with finite mean μ and variance σ^2, then the sequence of sample means converges to μ.
Bernoulli Law of Large Numbers. The sequence of sample proportions converges stochastically to p as n approaches infinity (there is a fixed p and we consider the random variable W_n, the proportion of successes in a sample of size n ($W_n = \frac{Y_n}{n}$).

Example 65. Language change and population size. Imagine a language in such state S that the probability of a certain change ('mutation') in an individual mental grammar is α. The mutation 'takes off' as a language change if the proportion x of speakers having this mutation (in the community) is higher than a certain threshold $\gamma = 0.9\alpha$. How the probability p that this threshold will be reached for a given generation of speakers changes with the growth of the community?

Part XII

Applicability of statistics and tests of hypotheses

1 Tests of hypotheses: the general idea

A statistical hypothesis is a statement about a distribution. If the hypothesis completely specifies the distribution, it is called a simple hypothesis; otherwise it is called composite.

Example 66. Preference. Suppose we have developed a theory of word-order distributions in any geographically defined language population that predicts that the proportion of SOV languages in a language population is distributed normally with $\mu = 0.5$ and $\sigma^2 = 1/K$, where K is the number of genetic stocks represented in the population. We want to test this hypothesis (H_0) against the alternative hypothesis $H_a : \mu > 0.5$.

The critical region for a test of hypotheses is the subset of the sample space that corresponds to rejecting the null hypothesis. In our example, the critical region can be expressed in terms of the sample proportion of SOV languages, $\hat{p}$; i.e. it will include all samples for which $\hat{p}$ satisfies certain conditions. Since the alternative hypothesis is $\mu < 0.5$, a natural form of the critical region is

$$C = \{(x_1, ..., x_k) | \hat{p} < c\}$$

for some appropriate constant c.

Type I error: Reject a true H_0. $P[TI] = \alpha$ Type II error: Fail to reject a false H_0 ("accepting a false H_0") $P[TII] = \beta$

A test statistic and a critical region are selected in such a way that we would have a small probability of making these two errors. For a simple null hypothesis, the probability of type I error is referred to as the significance level of the test.

The standard approach would be to specify or select some acceptable level of type I error, and then to determine a critical region that would achieve this α (among all possible regions, we would select the one that has the smallest $P(TII)$.

For this illustration, we will use Dryer's genera-based data, disregarding the problems created by genera-based sampling and assuming that there 25 stocks in each large area, i.e. $\sigma^2 = 1/25$ and $\sigma = 1/5 = 0.2$.

	Africa	Eurasia	Australia	North America	South America	Total
SOV	22	26	19	26	18	111
SVO	21	19	6	6	5	57
Other	5	7	5	28	8	53
Total	48	52	30	60	31	221

For $\alpha = 0.05$:

$$c = \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = 0.5 - 1.645 \cdot \frac{0.2}{\sqrt{n}}$$

The following table gives the actual values of $\hat{p}$ and c for all areas; the North-American sample rejects the hypothesis, while all other areal samples fail to do so.

	Africa	Eurasia	Australia	North America	South America
$\hat{p}$	0.46	0.5	0.63	0.43	0.58
c	0.45	0.45	0.44	0.46	0.44
Rejected:	No	No	No	Yes	No

2 Contingency tables and goodness-of-fit

2.1 χ^2-distribution

The most common way of solving this sort of problems invokes the χ^2-distribution, a special case of Gamma distribution.[1]

$$Y \sim \chi^2(v) = \text{GAM}(2, v/2); \quad f(x;v) = \frac{1}{2^{v/2}\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}$$

The fundamental importance of this distribution lies in the fact that if $Z \sim \text{N}(0,1)$, then $Z^2 \sim \chi^2(1)$, which means that it can be used to study deviations from the expected values. If $X_1, ..., X_n$ denotes a random sample from $\text{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$$

$$\frac{n(\bar{X} - mu)^2}{\sigma^2} \sim \chi^2(1)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

The square of an approximately normally distributed test statistic will be approximately distributed as $\chi^2(1)$. The sum of independent chi-square variables is chi-square distributed.

A common test for independence is based on the question whether the probability of a certain value is the same in r samples. Returning to the data of our first example, we can ask whether the probability of SOV is the same in all areas. So our null hypothesis H_0 is now that it is indeed the same; if so, its best estimate is based on the pooled sample: $\hat{p} = 111/221 = 0.5$. From this, we can

[1]

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$$

$$\Gamma(k) = (k-1)\Gamma(k-1)$$

$$\Gamma(n) = (n-1)!, n = 1, 2...$$

$$\Gamma(\frac{1}{2}) = \pi^{\frac{1}{2}}$$

$$X \sim \text{GAM}(\theta, k); \quad f(x;\theta,k) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}$$

calculate the expected numbers of SOV ($\hat{e}_{1j}$) and non-SOV $\hat{e}_{2j}$ ($j = 1, \ldots, 5$) languages, given the size of each sample, and compare them with the actually observed numbers (o_{ij}) by forming the following (standardized) random variables $X_{ij} \sim \chi^2(1)$:

$$x_{ij} = \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

	Africa	Eurasia	Australia	North America	South America	Total
SOV	0.17	0	1.07	0.53	0.4	2.17
Other	0.17	0	1.07	0.53	0.4	2.17
Total	0.34	0	2.14	1.06	0.8	**4.34**

The test statistic $\bar{\chi}^2$ is the sum of these random variables

$$\bar{\chi}^2 = \sum \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2(v)$$

which is approximately χ^2-distributed with $v = r - 1$ degrees of freedom (one degree of freedom is lost for each dimension). Now we will reject the null hypothesis if the actual value of this statistics is more than the $1-\alpha$-th percentile of the χ^2-distribution. For $\alpha = 0.05$ and four degrees of freedom the critical value is 9.49, which means that the test failed to reject the null hypothesis.

2.2 Goodness-of-fit

Example 67. The limiting distribution of alignment. If we have built a model of transition process, which predicts a particular limiting distribution, we can check whether a sample from the language population can be assumed to be drawn from the limiting distribution.

	Nominative	Nom. Diff.	Ergative	Split and Erg.Diff.	Neutral	Total
p_i	0.23	0.15	0.07	0.05	0.50	1
$e_i = 400p_i$	92	60	28	20	200	400
o_i	88	52	36	32	192	400

The value of χ^2 is 11.05 (for $v = 4$), $p = 0.03$, so the hypothesis is rejected for $\alpha = 0.05$, yet would not be rejected for $\alpha = 0.01$.

Example 68. Dynamic independence. The following model of transition processes for basic word order typology is based on the analysis of divergence rates:

E_0	E_1	p_{01}	p_{10}	$\lim_{n\to\infty} p_1^{(n)}$
OS	SO	0.586	0.004	0.993
VS	SV	0.236	0.004	0.983
OV	VO	0.096	0.055	0.635

Here p_{ij} are "unconditional" transition probabilities (nothing is known about the initial state except for the value of the particular variable described). Yet transition events can depend on the values of other variables, in particular, of the other word-order variables. The χ^2

statistic can be used to test how well divergence rates based on the general model predict the divergence rates in samples with fixed values of other variables.

fixed	E_1	sample size	$\hat{p}_1^{(n)}$	$\hat{d}^{(n)}$	$e(D)$	$o(D)$	χ^2
OV	SV	141	0.698	0.149	21.01	17	0.8
VO	SV	146	0.986	0.015	2.19	0	2.19
OV	SO	141	0.924	0.104	14.66	16	0.12
VO	SO	146	0.983	0.031	4.53	1	2.75
VS	SO	36	0.778	0.286	10.3	8	0.51
SV	SO	269	0.996	0.013	3.5	2	0.64

3 Regression

The case of simple linear regression:

$$E(Y|x) = \beta_0 + \beta_1 x$$

The standard approach is the Principle of Least Squares, which says to minimize the sum of the squared deviations:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

which gives

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x}^2)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example 69. Divergence rates. The model based on Markov processes entails a linear dependency between the divergence rate $d^{(n)}$ (the conditional probability of two languages being in different states given that they were in the same state at the previous step) linearly depends on the probability of the states at the same step of the process, $p_1^{(n)}$:

$$d^{(n)} = \beta_1 p_1^{(n)} + \beta_2$$

where the coefficients are defined by transition probabilities. How the linear regression model can be used to estimate the transition probabilities?