"Population" in biology and statistics

Nicola Bertoldi [a,1] , Charles H. Pence [b,*,1]

[a] Louvain Research Institute for Landscape, Architecture, and Built Environment, Université catholique de Louvain, Rue Wafelaerts, 47-51, bte. B3.01.01, 1060, Bruxelles, Belgium
[b] Institut supérieur de philosophie, Université catholique de Louvain, Place du Cardinal Mercier 14, bte. L3.06.01, 1348 Louvain-la-Neuve, Belgium

A B S T R A C T

The development of a biological notion of "population" over the first century of the theory of evolution has been commented upon by a number of historians and philosophers of biology. Somewhat less commonly discussed, however, is the parallel development of the statistical concept of a population over precisely the same period, in some cases driven by the same historical actors (such as Francis Galton and R. A. Fisher). We explore here these parallel developments, first from the perspective of a reconstruction of the historical development of each concept, then with the aid of a digital analysis of a corpus of literature drawn from the journals *Biometrika* and *Journal of Genetics*, between 1900 and 1960. These twin analyses show both points of interesting overlap between these two historical trends as well as points of important divergence. The biological and statistical notions of "population" seem to be relatively clearly distinguishable over these six decades, in spite of the fact that a number of authors contributed clearly to both traditions. The complex interplay of continuity and discontinuity across these two notions of "population" makes them a particularly interesting case study of scientific conceptual change.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Science.*

Few turns of phrase in the history and philosophy of biology have caused both more fruitful research and more confusion and polemic than Ernst Mayr's claim that Charles Darwin "replaced typological thinking by population thinking" (Mayr, 1976; reprinted in Sober, 2006, p. 325). As André Ariew (2008) has noted, the classic presentation of this distinction – Mayr's own – takes it to be a metaphysical distinction. Are types real objects existing in nature, which biologists should study (typological thinking), or are they fictions, emerging from the properties of individuals taken collectively (population thinking)?

To say that this debate has extended far beyond this original context would be an understatement. Mayr's original, metaphysical claim is hotly debated; Bence Nanay has noted that "it is difficult to find an interpretation of this notion that is both unproblematic and does the theoretical work it was intended to do" (Nanay, 2010, p. 91; see also, influentially, Sober, 1980). Population thinking has been discussed across both biological and philosophical literatures, where it has been implicated in unethical research on race (Gannett, 2001), and offered

ample subject matter for at least two insightful doctoral theses (Chung, 2000; Witteveen, 2013).

In what follows, we want to zoom in on a further aspect of this concept that has come in for significant debate.[2] As Margaret Morrison has noted, the relation between "population thinking" and the history of biological practice is something of a strange beast. Mayr himself used it to critique work in early population genetics – including that of a host of self-described arch-Darwinians like R. A. Fisher (Morrison, 2004). This seems, *a priori,* to pose something of a contradiction. How could these figures, so central to the tradition of evolutionary theory, in fact have missed one of Darwin's central insights?

The literature offers us two interrelated efforts to resolve the incompatibility. The first is Morrison's own. She argues that Mayr's critique of "the essentialist program" was largely aimed at a particular view of "the use and interpretation of mathematics in characterizing a population, making it one that is in every sense a 'construct,' dealing with the 'average man' and variation around mean values" (Morrison, 2004, p. 1190). In a similar vein, the geneticist Jody Hey has argued that Mayr's choice of terminology has led to confusion, conflating a statistical sense of "population" that finds its roots in the work of Francis Galton (as later modified by the biometrical school and the early

geneticists of the twentieth century) with the sense of a "natural population" as a biological entity or a unit of evolution. Hey argues that it is the statistical sense that Mayr had in mind, conceptualizing populations in this sense as a "holder of variation," where that variation could be among genes, organisms, or even populations in the non-mathematical sense of biological groups (Hey, 2011, p. 262).

We will return to the arguments of both Morrison and Hey in the following section. For now, the important point should hopefully already be apparent: it is deeply unclear just what the *population* in *population thinking* is supposed to refer to. This ambiguity arises for at least two reasons. First, as we see in both Morrison and Hey, there are at least two notions of "population" at work in contemporary evolutionary theory. The first is a population as a biological entity, an interbreeding, causally related collection of organisms that is standardly taken to be the target of natural selection, genetic drift, migration, and so on. It is this kind of population that can mate randomly or non-randomly, or that can feature in causal explanations, such as when we argue that a *population* of squirrels causes masting behavior in a tree (Potochnik & Brian, 2012, pp. 135–36). The second is a statistical conception of population, as a hypothetical entity, the values of whose parameters we aim to estimate by taking statistics on collected samples. A population in this sense does not, crucially, even have to be a population *of organisms*.

The second reason, we argue, that confusion persists over this question of populations has to do with the historical aspect that we briefly saw in Hey's discussion. As we know, the concept of "population" in biology underwent a radical shift over the course of the late-nineteenth and early-twentieth centuries. Any careful evaluation of population thinking, then, will need to also take into account the question of the history of this idea. This need becomes all the more acute when we add an insight that is not often mentioned in the history and philosophy of biology literature (notable exceptions including the work of Morrison, 2002, 2004 and a brief mention in Winther et al., 2015): the statistical notion of "population" introduced just above was itself also developed during exactly the same period that the biological concept was built – and in some cases by the same authors (Fisher, for instance, playing a prominent role in both stories).

To that end, in this paper we propose a reevaluation of the history of both the biological and statistical concepts of "population," as they changed over the period from 1859 to 1960. We will begin in the following two sections by looking, respectively, at the biological and statistical concepts of "population". The story of the former, though somewhat contested, is reasonably well known; the story of the latter is rarely discussed in the literature. We then turn toward an evaluation of their interactions over this period, both from a conceptual and from an empirical perspective, aided by analysis of the journals *Biometrika* (one of the few venues to have played host to developments in both of these traditions) and *Journal of Genetics*. We find, we argue, that the four concepts of "population" that result – each of the biological and the statistical having undergone a radical shift in meaning between 1900 and 1960 – offer ample material for reflection by historians and philosophers of both biology and statistics, and a story that indicates neither clear "borrowing" nor clear "independence."

1. Populations in biology from Darwin to Dobzhansky

As evidenced by Hey's analysis, looking for the roots of population thinking and the biological aspects of the modern concept of "population" in Darwin's writings leaves one with more questions than answers (Hey, 2011, pp. 256–57). Darwin employed the term "population" rarely, and mainly in a sense close to Thomas Malthus's use of the term in *An Essay on the Principle of Population* (Malthus, 1826). In this context, "population" meant the number of people or organisms living in a specific geographic area, always prone to quantitative fluctuations. Consider, for instance, the following excerpt from the second edition of Darwin's *Descent*: "The decrease of the native *population* of the Sandwich Islands is as notorious as that of New Zealand. It has been roughly

estimated by those best capable of judging that, when Cook discovered the Islands in 1779, the *population* amounted to about 300,000" (Darwin, 1874, p. 186, emphasis added). Clearly, the population mentioned in this sentence is the sum total of the Sandwich Islands' indigenous inhabitants.

While Darwin did not use the word "population" in a sense related to the modern idea of biological populations, the structure of his hypothesis of natural selection as reconstructed, for instance, by Jean Gayon, might make room for the same concept under another name. Per Gayon's reconstruction, selection relies on a tripartite ontology: "variations," "individuals," and "varieties" or "races" are all, to some extent, "units of selection" that interact with natural selection at different levels (Gayon, 1998, p. 64). Leaving aside, for our purposes here, the other notions, consider Darwin's introduction of the concept "variety" in Chapter 2 of the *Origin*. "Variety" is intended to entail "community of descent," but this alone does not distinguish varieties from "monstrosities," which "graduate into varieties" (Darwin, 1859, p. 44). As elsewhere in these early arguments in the *Origin*, the ambiguity in the distinction between varieties and monstrosities provided Darwin with a hinge point in his effort to "subvert," to use Gayon's words, the concept of "variety" itself. He observed that variations such as "the thicker fur of an animal from far northwards" could conceivably be inherited "for at least some few generations" (Darwin, 1859, pp. 44–45). This begins to resemble a definition of varieties as lineages along which potentially useful variations are transmitted from one generation to the next. Can we thus infer that the modern definition of "population" in the biological sense was already present in the *Origin*?

Prima facie, the concept of "variety" as Darwin understood it seems to contain some constitutive elements of the modern concept of "biological population," notably the idea that a population is a group of individuals held together by genealogical relations and the notion that a population is an entity that evolves through the accumulation of variations under natural selection. However, Darwin conceived the genetic relations among the members of a "variety" in a purely diachronic sense, i.e., as *purely* genealogical relations. Furthermore, he could not identify any mechanism capable of accounting both for the emergence of genealogical relations among the members of a given variety and for the accumulation of variations in this same variety. In Gayon's estimation, Darwin's inability to grasp the synchronic and properly genetic aspect of populations is ascribable to his lack of "both a theory of heredity and a statistical approach to populations" (Gayon, 1998, p. 66).

Nevertheless, on rare occasions, Darwin seemingly used the term "population" similarly to the modern statistical concept. In the *Variation*, for instance, he seems to make reference to a kind of law of large numbers:

> Let it be assumed that, in a large *population*, a particular affection occurs on an average in one out of a million, so that the *à priori* chance that an individual taken at random will be so affected is only one in a million. Let the *population* consist of sixty millions, composed, we will assume, of ten million families, each containing six members. (Darwin, 1875, p. 1:449, emphasis added)

Darwin then has a colleague calculate the (very small) odds that multiple family members could be independently affected by such a disease. Thus, the large population discussed here is not only a group of families bound by genealogical ties but also a hypothetical entity whose properties can be ascertained through statistical inference or probabilistic modeling.

In cases like these, Darwin seemingly foregrounded the definition of the term "population" that we can find, for example, in Francis Galton's *Natural Inheritance*. In this essay, Galton placed populations as central to his definition of heredity as the study of "the distribution of each faculty among the members of any large group, whether it be a Fraternity or an entire Population" (Galton, 1889, p. 35). Accordingly, he understood a population as a "census officer" would (Galton, 1889, p. 80): a large grouping of individuals (say, households) selected to gather data about

the measurable features of a structure or process in which said individuals all partake (say, household income in a given state). Heredity, then, is a process leading to the emergence of "statistical similarity between successive generations of a people" (Galton, 1889, p. 80).

Consequently, unlike Gayon (1998), we maintain that Galton's concept of "population" resulted from a two-pronged view of populations already found in Darwin: populations are both (1) biological entities characterized by genealogical relations between successive generations and (2) large groupings of individuals *qua* units of statistical analysis. The above-quoted passage from the *Variation* suggests precisely that Darwin held such a view. Moreover, further evidence is provided by various instances in which Darwin did not explicitly mention the term "population" yet still deployed "statistical thinking" *sensu* André Ariew (2022) to study evolutionary phenomena as they emerge from genealogical relations.[3] However, one notable difference remains between the two men. Whereas Darwin stressed the malleability of varieties as evolving populations due to the progressive accumulation of heritable variations, Galton insisted on the remarkable stability of populations across generations. As he explicitly stated in Chapter XII of *Natural Inheritance*, "the characteristics of any population that is in harmony with its environment, may remain statistically identical during successive generations. This is true for every characteristic whether it be affected to a great degree by a natural selection, or only so slightly as to be practically independent of it" (Galton, 1889, p. 192).

Galton thus developed his conception of biological populations in an attempt to answer a very different question than the one addressed by Darwin: why do evolving populations tend to remain stable across time? This difference led Galton to go beyond the limitations that Gayon has identified in Darwin's approach. Galton derived a general equation of heredity grounded in the statistical properties of the normal distribution. This general equation is today known as Galton's "law of ancestral heredity" and suggested, in turn, a theory of descent according to which "descent either was particulate [i.e., traceable to the transmission of unit characters from ancestors to descendants] or acted as if it were so" (Galton, 1889, p. 193).[4]

While Galton therefore, more than Darwin, laid the foundations for the modern view of populations as biological entities, we can identify two significant issues in his characterization of biological populations. First, the theory of heredity underpinning this definition was essentially conceptual and constituted a formal model expressing a minimal condition such "that no theory of descent that failed to satisfy it could possibly be true" (Galton, 1889, p. 192). Second, Galton's statistical approach to (biological) populations needed further development. It has been argued that this inchoateness of the Galtonian statistics of hereditary processes, in turn, played a crucial role in fostering the controversy between the Mendelians and the biometricians (Bertoldi, 2022).

This controversy constituted a pivotal moment in the development of the modern concept of "biological population." It centered precisely on the theoretical and empirical implications for the statistical principles uncovered by Galton and Karl Pearson during the 1880s and the 1890s

of such plant breeding experiments as those of Gregor Mendel (1866) or Wilhelm Johannsen (1903a). Johannsen, for instance, was particularly interested in Galton's "Law of Regression" – the idea that offspring will lie closer to the mean, on average, than their parents – and Karl Pearson's later technical developments of the same idea. In Johannsen's opinion, Pearson's treatment of Galton's law relied on a fundamental theoretical assumption about biological populations: "The population at hand for investigation" should be regarded "as a [single] unit [*als Einheit*] … whether it be the bulk of the inhabitants of a given region [*Bevölkerung*] [or] a stock of animals or plants of a certain kind (a race)" (Johannsen, 1903a, p. 2).[5]

Furthermore, this assumption blended both statistical considerations – populations are represented as distributions of "singular characters" around "a mean, supposedly 'typical' value" (Johannsen, 1903a, p. 4) – and biological ones – "before one treats a population as a [single] unit, one should thus analyze it biologically to gain clarity on its constituent elements" (Johannsen, 1903a, p. 5), to determine whether or not a statistical treatment could possibly apply. While the innovation for which Johannsen is now most often remembered is the experimental construction of "pure line" experimental populations – which might seem extremely distant from "biological populations" in any contemporary sense – he always kept in view that a population "can be a mixture of different types," and thus "has to be *biologically analyzed* before a statistic of heredity is built" (Johannsen, 1903b). We can thus understand Johannsen's approach to hereditary patterns in biological populations as an attempt to grapple with the two issues in Galton's views that we identified above. On the one hand, Johannsen explicitly saw Galton's law of regression in light of the now-mature "statistical theory of regression" (Johannsen, 1903a, p. 12). On the other hand, Johannsen also believed that it was necessary to go beyond the mere application of the theory of regression and conduct "a deeper biological study of hereditary relations" (Johannsen, 1903a, p. 4), eventually culminating in his "pure line" experiments (Gayon, 1998; Meunier, 2016; Roll-Hansen, 1978; Stoltzfus & Cable, 2014), the aim of which was precisely to shed light on the material processes underlying the hereditary patterns identified by biometrical methods of inquiry.

Johannsen's answer to Galton's limitations indicates a further expansion of the semantic field associated with the concept of "biological population" at the dawn of the 20th century, by making populations into units of experimentation – not only in the framework of his experiments on selection in pure lines, but also in the context of his later hybridisation experiments (Johannsen, 1907) *qua* "belated control" for the former (Meunier, 2016, pp. 48–50). It is through these experimental means that Johannsen attempted to deconstruct the biometrical view of heredity and natural selection by uncovering its underlying physiological mechanism.

Johannsen could thus conclude that biological populations generally consisted of a mixture of various distinct and hereditarily stable "typical forms." As argued by Nils Roll-Hansen (1978), the distinction between "average (average character, average value etc.)" and "type" (Johannsen, 1903a, pp. 58–59) that Johannsen developed to analyze his pure line experiments closely approximates the distinction he would later coin between "genotype" and "phenotype" (Johannsen, 1909). Johannsen's analysis thus foreshadowed an essential element of the modern definition of biological populations, later captured by Theodosius Dobzhansky in his concept of "Mendelian population": the members of a population are "*sexual and cross-fertilising individuals which share in a common gene pool*" (Dobzhansky, 1950, p. 405, original emphasis), i.e., a common set of distinct genotypes.

By subordinating statistical analysis to experimental breeding methods, Johannsen thus laid the groundwork for a view of biological populations as materially existing entities bound together by physiological and genetic mechanisms (Gannett, 2003) rather than as

---

[3] *Contra* scholars who have argued the contrary (e.g., Gigerenzer et al., 1990; Porter, 1986; Walsh, 2015), Ariew (2022) has highlighted two instances in which Darwin made a three-pronged use of statistical thinking, i.e., "to *analyze large-scale patterns* that would otherwise be undetected at the level of individuals, *provide empirical tests for hypotheses*, and even *explain* large-scale trends" (Ariew, 2022, p. 216). The first instance is Darwin's discussion of the hypothesis that varieties constitute incipient species in letters written to John Lubbock (Darwin, 1857a) and J. D. Hooker (Darwin, 1857b). The second is the statistical explanation for the evolution of rudimentary characters contained in a letter that Darwin submitted to *Nature* in 1873 (Darwin, 1873).

[4] This work was the basis of Hacking's contention that Galton was the first to have "tamed" chance (Hacking 1990, p. 181). This idea that Galton was the key figure in the development of what has sometimes been called "statistically autonomous explanation" has attracted both support (Ariew et al. 2015, 2017; Depew & Weber, 1995) and critique (Pence, 2022).

[5] Translations from the German are our own.

abstractions created for accounting or statistical purposes. According to Arlin Stoltzfus and Kele Cable, Johannsen's experiments allowed him to understand the normal distributions of quantitative characters described by biometrical models as resulting from "the overlaying of environmental variation on a mixture of types, which is how such distributions are understood today" (Stoltzfus & Cable, 2014, p. 521). In that respect, Johannsen could be said to have anticipated Fisher's (1918) proof of consistency between (biometrical) Darwinism and Mendelism.

Consequently, Stoltzfus and Cable (2014) have situated Johannsen's research within a "Mendelian-Mutationist synthesis" embodied by the works of William Bateson, Reginald Punnett and T. H. Morgan, falling short of the Modern Synthesis only by not acknowledging how recombination may allow natural selection to shift the average character of a population, even without the emergence of new mutations. This Mendelian-Mutationist synthesis, they argue, was as coherent and complete as what we might dub the "Darwinian-Selectionist synthesis" stemming from the development of Hardy-Weinberg equilibrium (Hardy, 1908; Weinberg, 1908) and the works of Fisher (1930), Sewall Wright (1931) and J. B. S. Haldane (1932). In this other, better-known tradition, Fisher and Wright, despite their differences concerning the conditions that would best favor adaptation, both viewed populations as bound together by the action of shared selective pressures, possibly modulated by random processes such as genetic drift.

As Hey has pointed out, the two concepts of "population" just discussed – the Mendelian-Mutationist and the Darwinian-Selectionist – both influenced Dobzhansky's seminal contributions to modern evolutionary biology. While his early works appear to reflect the influence of Johannsen's idea of a mixture of unmodifiable pure strains (e.g., Dobzhansky, 1927, 1930), later work explicitly refers to Fisher's and Wright's approaches. But by carefully considering the revisions of the "Mendelian population" concept in his later work, we can argue that he went even farther than this, developing a synthesis of the Mendelian-Mutationist concept and the Darwinian-Selectionist one.

On the one hand, Dobzhansky based his definition of "Mendelian population" on the idea of a shared gene pool, thereby acknowledging that the existence of a biological population relies upon a mixture of genetic types. On the other hand, building on Wright's (1931, 1943) previous works on isolation mechanisms, Dobzhansky proposed what we might call a "Mendelian continuum": different kinds of Mendelian populations, varying both in breadth and structural complexity, from local groups of interbreeding individuals all the way to species (Dobzhansky, 1950, p. 405). He thus embedded Johannsen's notion of a mixture of stable and distinct genetic types in the very concept of a panmictic population subject to selective pressures. From this standpoint, genetic differences emerge from individuals freely mating with each other and become more and more pronounced as free matings become less and less probable due to the emergence of isolating factors, geographical or otherwise. Such a synthesis implies a third view of what binds biological populations together: neither hereditary processes nor selective pressures but mating and breeding.

More specifically, as Dobzhansky himself explicitly stated in a 1935 article devoted to "A Critique of the Species Concept in Biology," although evolutionists had been fully aware of the homogenizing effect of sexual reproduction since Darwin, "the development of genetics brought a clarification of the understanding of the mechanism involved" (Dobzhansky, 1935, p. 348). Mendelian populations exist as mixtures of discrete groups of individuals, each of which "represents a definite constellation of genes," i.e., a pure type *à la* Johannsen. In a perfectly panmictic population, the free mating and breeding of individuals from various groups constantly brings about a new equilibrium "in which the different genic constellations become fused into one" (Dobzhansky, 1935, p. 348). By converse, the persistence of distinct groups within a population needs to be regarded as evidence for the presence of barriers to free interbreeding. This consideration directly bears on the problem of the origin of species: since similar individuals appear to be more likely to interbreed than less similar ones, "a stage must exist in the process of

evolutionary divergence, at which an originally panmictic population becomes split into two or more populations that interbreed with each other no longer" (Dobzhansky, 1935, p. 348), thus constituting two separate species. As Dobzhansky himself noted, the "fundamental importance of this stage" had already been emphasized by geneticists such as Bateson[6] and Johannes Paulus Lotsy (see, e.g., Lotsy, 1925), who had coined the term "syngameon" to define a biological species as "'an habitually interbreeding community' of individuals" (Dobzhansky, 1935, p. 349). His original contribution thus consisted in expanding on Lotsy's intuition by moving the emphasis from the absence of interbreeding between distinct species to "the presence of physiological mechanisms making interbreeding difficult or impossible" (Dobzhansky, 1935, p. 349), the genetical foundations of those same mechanisms, and their implications for evolutionary processes by natural selection. Such a shift in emphasis is precisely what led him to define species as "reproductively isolated Mendelian populations" (Dobzhansky, 1950, p. 415).

Accordingly, Dobzhansky's synthesis of the Mendelian-Mutationist notion of "biological population" and the Darwinian-Selectionist one, relied on the identification of isolation mechanisms as a means for developing a common account of three empirical phenomena historically associated with the development of a biological population concept: the sterility of interspecific crossings, the existence of distinct pure lines within groups of individuals found in nature, and the effects of natural selection. By accomplishing such a synthesis, he arguably laid the foundations for what Hey has called "the modern population concept of an interbreeding community of organisms that exists as an evolving entity" (Hey, 2011, p. 257). As Witteveen (2015, 2016) has carefully argued, Mayr's idea of "population thinking" is due in large part to a synthesis between Dobzhansky's use of populations as "a distinctive pattern of reasoning about the dynamics of heredity and variation in biological populations" (Witteveen, 2015, p. 30) and Simpson's prior (and shifting) idea that, in paleontology and systematics, "taxonomists should be mindful not to reify a simple rule of thumb into a definite standard for classification" (which we lack the space to pursue here; Witteveen, 2015, p. 27).

Furthermore, Dobzhansky's synthesis allows us to identify six central elements of this concept. First, biological populations are groups of individuals that occupy a shared range of geographical locations, which are not separated by barriers that impede any exchange between gene pools. Second, they constitute units of descent – their members are genealogically related. Third, biological populations are, at the same time, units of statistical analysis insofar as they can be represented as distributions of characters. Fourth, they embody units of experimentation since they provide the ground for breeding experiments such as those pursued by Johannsen and, later, by Dobzhansky. Fifth, biological populations are characterized by definite mating relationships, or by a common mating system. Sixth, they are evolutionary units to the extent to which they are subject to shared selective pressures. Those six elements jointly encapsulate the richness of the modern biological population concept.

2. Statistical populations from Galton to Fisher

While Galton, as we have already seen above, is undeniably a pioneer in the methods of statistics, his use of the population concept is not the one that we would today expect in a statistical context. As any contemporary statistics textbook will tell you, it is essential to draw a

---

[6] In Dobzhansky's 1935 article, no reference is provided for highlighting Bateson's role in acknowledging the importance of reproductive isolation for the evolution of new species. However, H. Allen Orr (1996) has pointed to Bateson (1909) as containing in germ the quantitative model for the genetics of speciation by sexual isolation and natural selection later developed by Dobzhansky (1934, 1936) himself and H. J. Muller (1940, 1942), known as the "Dobzhansky-Muller model".

distinction between a *population* – the hypothetical object whose *parameters* we would like to know the values of – and a *sample* from that population – the concrete object on which we may compute *statistics* in an effort to infer the values of the population parameters.

This is decidedly not what we see in Galton. As one of the authors has argued elsewhere (Pence, 2022, pp. 23–48), Galton seems to have a deeply *actualist* view of statistical populations. Statistics are useful precisely because they let us understand the properties of the actual citizens of England. If we were to draw out a life-sized statistical distribution, for instance, of the predicted heights of all Englishmen,[7] we could call them forth to take their assigned place under the curve: "The tops of the marshalled row form a flowing curve of invariable proportions; and each element, as it is sorted into place, finds, as it were, a pre-ordained niche, accurately adapted to fit it" (Galton, 1889, p. 66).

It is important to see the peculiar kind of idealization that is at work here. To be sure, describing the distribution of a population in terms of its mean and variance is to construct an approximation or an idealization of that population. But this is, for Galton, only idealized in a very minimal sense: every Englishman could still find his assigned place under the curve – the curve is still taken to serve as a description of the actual, real-world population. For Galton, then, measures like the mean are measures taken on a collection of extant organisms.

One potential explanation for Galton's adherence to this view can be drawn from William Kruskal and Frederick Mosteller's historical retrospective on the development of random sampling. As they note there,

in the 19th century, with its growth of national statistical systems and the contemporary but separate development of sociological studies, one sees two almost polar movements. The first pushes for full coverage, for the complete census; the second cultivates case studies – sometimes called monography – to permit insight and hypothesis formation by intensive investigation …. (Kruskal & Mosteller, 1980, p. 173)

Following this perspective, Galton's view becomes perhaps more explicable. Galton's statistical lodestar remains, throughout his career, Adolphe Quetelet, one of the central architects of the full-census view of population study (see, e.g., Droesbeke, 2021, pp. 238–239, on Quetelet's role in the Belgian census of 1842). Galton's own statistical investigations often work in the same way, as he attempts to send family information cards to literally every English family "of genius" that he can find in order to provide data for his study of intellectual ability (Galton, 1869). In that sense, Galton is a product of his time, a creature of the mid-to-late-nineteenth century demographic and statistical establishment.

Opposition to representative sampling – and, by extension, opposition to precisely the novel concept of "population" that we are attempting to trace here – would remain relatively fierce until the 1930s, and would be justified largely on epistemic grounds. For instance, Kruskal and Mosteller quote the statistician Georg von Mayr, speaking in 1895:

I understand that representative samples can have some value, but it is a value restricted to terrain already illuminated by full coverage. One cannot replace by calculation the real observation of facts. A sample provides statistics for the units actually observed, but not true statistics for the entire terrain. (Kruskal & Mosteller, 1980, p. 174)

It would therefore take time – according to some, more time than we might have expected or hoped (Kruskal & Mosteller, 1980, p. 169) – to move from this view to a contemporary notion of "populations."

As a number of authors have noted, it is perhaps surprising that the story does *not* continue with Pearson, the pioneer of mathematical

statistics to have most significantly advanced the rather rudimentary formal tools that Galton had developed. Fisher notes that Pearson had come very close to a novel concept of "population" as early as 1898, but not taken the final, crucial step:

A similar method of obtaining the standard deviations and correlations of *statistics* derived from large samples was developed by Pearson and Filon in 1898. It is unfortunate that in this memoir no sufficient distinction is drawn between the *population* and the *sample*, in consequence of which the formulae obtained [are erroneous] …. (Fisher, 1922c, p. 329n)

Kruskal and Mosteller also note that Pearson might have served this role, largely thanks to his disciplinary position between the uses of statistics in natural science and those in economics or public policy, but again, that he did not in fact do so:

Little interaction occurred between social and economic survey-takers, on the one hand, and statistical analysts in the natural sciences, on the other. Karl Pearson might have formed a bridge between the two statistical worlds, but he did not. (Kruskal & Mosteller, 1980, p. 172)

To see the next real advance in the story, then, we must turn to Fisher. In the very same year that he published his two ground-breaking papers on evolutionary theory, "Darwinian Evolution by Mutations" (Fisher, 1922a) and "On the Dominance Ratio" (Fisher, 1922b), Fisher set his sights on the foundations of statistics. These foundations, he writes, have fallen into "prolonged neglect," chiefly because

it is customary to apply the same name, *mean*, *standard deviation*, *correlation coefficient*, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our means of estimation. (Fisher, 1922c, p. 311)

This confusion traces back to the very fundamentals of statistics: namely, what is the purpose of statistical study? According to Fisher, it is what he calls the "reduction" of data, by which he means something like what Ernst Mach and Pearson called "economy of thought" (Mach [1886] 1914; Pearson, 1900), the representation of a large mass of data by a small number of quantities "which shall adequately represent the whole" (Fisher, 1922c, p. 311).

Straightaway, then, Fisher argues that this goal can only be accomplished if we replace the Galton-Pearson concept of "population":

This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. Any information given by the sample, which is of use in estimating the values of these parameters, is relevant information. (Fisher, 1922c, p. 311)

While such an assertion clearly exhibits the mark of Fisher's infinite-frequentist approach to the foundations of probability, he argues that such an approach is no less applicable when we are measuring probabilities generated by natural systems.

It should be noted that there is no falsehood in interpreting any set of independent measurements as a random sample from an infinite population; for any such set of numbers are a random sample from the totality of numbers produced by the same matrix of causal conditions: the hypothetical population which we are studying is an aspect of the totality of the effects of these conditions, of whatever nature they may be. The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician. (Fisher, 1922c, p. 313)

---

[7] Englishwomen could only find their place if they converted their heights to the "standardized" heights of Englishmen, by multiplying them by a ratio of 13: 12 (Galton, 1889, p. 6).

Fisher therefore moves, in one stroke, from a Galtonian or Pearsonian view of statistical populations to an entirely modern conception, one that sees clearly a distinction between the quantities to be estimated and our estimates of them, as well as a way in which nearly any statistical investigation might be phrased in terms of such sampling of populations, even those that appear at first to lack a straightforward "populational" aspect.

We have quoted Fisher at length not only because of the significance of this insight, but also because of the rapidity with which it reshaped the statistical community. Only a year later, Student (William Gosset), in fairly constant dialogue with Fisher (who was working in Rothamsted on the design of agricultural experiments), published an extensive article attempting to render Fisher's view more practical, pursuing answers to large-scale statistical survey questions in smaller samples (Student, 1923). By nine years later, this view would reach its theoretical culmination in the work of Jerzy Neyman (1934). Here we see not only a fully fleshed-out view of populations and samples, but the application of this theoretical method to a whole host of important statistical problems – among others, the first place where the concept of a confidence interval is discussed in English. Neyman begins the work by explicitly invoking this statistical concept of "population":

> We are interested in characteristics of a certain population, say, π, which it is either impossible or at least very difficult to study in detail, and we try to estimate these characteristics basing our judgment on the sample. (Neyman, 1934, p. 561)

Within a few years, this theoretical work would be applied empirically, perhaps most importantly in the Indian census by Prasanta Chandra Mahalanobis (1937; see also Salsburg, 2001, pp. 169–72). In short, the very shape of a statistical population was, after 1922, largely Fisherian.

While we lack a synthetic source (like Dobzhansky in the previous section) for a set of criteria that can describe the statistical sense of population, we might say that statistical populations are the collections that ground statistical inferences. In some cases, these might be real-world exemplars, as with Galton, while in others these may be hypothetical (or, with Fisher, even infinite) constructions, for which statistical tools let us estimate certain central quantities.

3. Interacting "populations"

Tracing the interaction of these concepts of "population" is not a straightforward enterprise. Here, yet again, we are bequeathed something of a contradiction from the literature. On the one hand, we learn that biological and statistical notions of "population" are *so similar that they may not be clearly distinguishable*. Morrison, for instance, writes of Fisher that "at this point in the development of genetics and biology it is very difficult to distinguish, in any determinate way, what is properly biological from what is strictly mathematical" (Morrison, 2004, p. 1192). On the other hand, we learn elsewhere that these two concepts are *so different that conflating them leads to obvious error*. Hey writes that Mayr "articulated a fuller, more complex meaning of population thinking that has often been confused with the biological concept of a population that emerged in the early parts of the 20th century" (Hey, 2011, pp. 261–62), and Witteveen has argued, as we saw briefly above, that Mayr took two separate lines of argument from Dobzhansky and Simpson and, problematically, "'synthesized,' historicized, and expanded [them] into an all-encompassing typology/population dichotomy" (Witteveen, 2015, p. 21).

It would be helpful, then, to be able to reinforce our historical sketch with a quantitative analysis of (at least a portion of) the literature demonstrating that this change has in fact taken place, and helping us to better understand what this shift meant for practitioners in evolutionary biology and statistics. If we could detect where in the literature these various notions of biological and statistical populations were in use and when, we would better be able to understand what we might call the

"cartography" of concepts of population. How do these changes in "population thinking" both in biology and statistics actually develop over time? Are they indeed correlated, or are they actually distinct and independent conceptual shifts? Answering these empirical questions would be an important first step to knowing which historical and philosophical conclusions could be justified on the basis of the literature in biology and statistics.

To this end, we build here on prior work (Bertoldi et al., 2024) that has explored the content of the journal *Biometrika*, one of the central sources for both work on statistical and on biological populations in the period, roughly from 1900 to 1960, during which the developments we have described took place. In this section, we will begin by offering an empirical confirmation that the shift that we have described did indeed take place, by appealing both to *Biometrika* and, for reasons that we will discuss shortly, supplementing this analysis with an evaluation of the *Journal of Genetics*. We will then turn to a few ways in which we can make our understanding of the shift more profound by seeing its impacts on the published literature in these two journals. While this only amounts to two, English-language journals, they published a significant number of papers and were both scientifically and socially important for the development of both statistical and biological "populations."

3.1. Demonstrating the shift in "population"

While telling the full tale of the history of the journal *Biometrika* would take us too far afield for our purposes (interested readers can consult, e.g., Provine, 1971, pp. 62–64; Porter, 2004, pp. 268–69), it may be briefly summed up as the journal founded when the biometricians sought a new venue to print complex, mathematically sophisticated articles in statistical evolutionary theory without the intervention of the Royal Society and the early geneticists, especially William Bateson. This means that *Biometrika* is unusual in containing early work on a pre-Mendelian biological sense of "population" linked to demography (see section 1), early work on a pre-Fisherian statistical sense of "population" linked to complete sampling (section 2), and later work on the post-Fisherian statistical sense of "population," as concepts of "random sampling" are developed (section 2). (The later biological sense of "population," corresponding to Mendelian populations, is notably absent; more on this in a moment.)

These three senses of population can be demonstrated extremely clearly in a text analysis of the corpus of articles published in *Biometrika* from its founding in 1901 until 1960, a total of 1629 articles. To that end, we will describe here the ways in which these notions appear in a topic model that we and colleagues have developed and described elsewhere (Bertoldi et al., 2024).[8] First, however, we should pause to briefly introduce topic modeling, as it will be our primary analytic tool in what follows.

Topic modeling is an unsupervised method to help researchers understand the content of a body of text. One simple way to understand these models is to think of them as proposing an abstract model of the "writing" of a document, then using machine learning to deduce the parameters of that model (Blei, 2012). Imagine that we write a document by choosing each of its words in the following way. First, we choose a "topic" from a set (of a size chosen by the investigator), with a given probability fixed for this document. That topic, then, tells us the probability of every word that might come next in the document, so we use it to pick the word that we will insert, and repeat until we have chosen the number of words to be found in the document (without considering their order). Topic modeling assumes that documents were prepared in this way, then infers the values of these probability
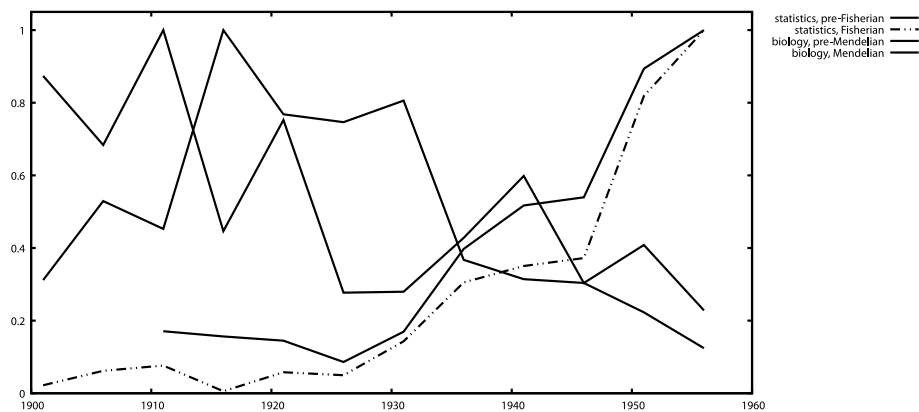
Fig. 1. The prevalence of the topics associated with the concepts of a pre-Fisherian statistical, Fisherian statistical, pre-Mendelian biological, and post-Mendelian biological concept of "population", in *Biometrika* (statistics, pre- and post-Fisherian; biology, pre-Mendelian) and *Journal of Genetics* (biology, Mendelian). Values are normalized so that the highest relative prevalence of each topic during the time period at issue is equal to 1.

distributions. As it turns out, such models, when applied to a variety of texts, produce "topics" that track *topics* in the colloquial sense – they describe the subjects treated in documents (Boyd-Graber et al., 2017).

To interpret these topics, we've performed two (very traditional) kinds of analysis. First, we look at which words are most likely to be chosen by a particular topic (i.e., we look at the twenty words to which the topic gives the highest probability). Then, we look at which documents give the highest probability values for that same topic (i.e., we look at the documents for which the probability distribution over topics gives the highest values to the topic). By analyzing these lists of words and documents, we can develop a sense of what that topic is "about," and then use those topics to trace changes in journal content and emphasis over time.

The topic model of *Biometrika* that we produced contains "population" as one of its twenty most probable words in three different topics – each corresponding to one of the three senses of "population" that we discussed above. One topic (called "A-Value-sample-mean" in Bertoldi et al., 2024) corresponds to the pre-Fisherian statistical sense of population. Other highly probable words in the topic include "value," "sample," "curve," "mean," and "frequency." Largely, this topic concerns two fundamental questions. The first is how to apply Karl Pearson's methodology for curve-fitting in statistics, often called the "method of moments," to real-world data drawn from different case studies. This involves inferring curve properties like means, variances, skew, or kurtosis (Sophister, 1928). Second, papers in this topic consider evaluating associations between variables, including both quantitative and qualitative (binary or categorical) variables (Blakeman & Pearson, 1906; Pearson, 1917). As such, this topic largely traces mathematical statistics in the Pearsonian tradition (Magnello, 1996, 1999a, 1999b, 2005, 2009, 2014; Stigler, 1986, 1999).

A second topic (which Bertoldi et al., 2024 called "B-Sample-population") represents the development of the statistical notion of "population" after Fisher and Neyman. This arrives somewhat late in *Biometrika* for local reasons (including a bitter feud between Pearson and Fisher; Edwards, 1994), but increasingly over the 1930s and 1940s, articles associated with this topic discuss how to select appropriate samples for various practical purposes (e.g., surveys; Smith, 2001), and according to various preferred characteristics (Johnson, 1957; Vagholkar & Wetherill, 1960). Other probable words in this topic include "sample" and "probability," confirming the tight link between the Fisher/Neyman sense of populations and concerns of sampling and probability, as we saw underlined by a host of authors in the last section. (Notably, this topic becomes even more prevalent in the time period after the end of our study, as clinical trial sampling and sequential analyses become increasingly important problems in the pages of *Biometrika*.)

Third and finally, we see a topic (called "A-Age-population") that is the representation of the pre-Mendelian notion of "biological populations". Other top words in this topic include "age," "rate," "birth," "death," and "disease," all terms associated with demographic questions, the primary subject of this vital-statistics tradition. This topic thus seems to cleanly capture the use of statistical methods to analyze variation in characteristics of a particular population as a function of age structure. Among the documents that give the highest probability to this topic, we see the application of mathematical statistics, as a transformative development of early Victorian vital statistics (Magnello, 2006; Magnello & Hardy, 2002), to areas like human and animal evolution (Hacker & Pearson, 1946; Powys, 1905), demography (Elderton, 1914; Karn, 1933), and epidemiology (Martin & Cheeseman, 1938; Stocks, 1924).

We do not, however, see a corresponding topic for Mendelian populations in *Biometrika*. This is entirely expected. As was briefly noted above, the journal was founded explicitly as a *non*-Mendelian publication outlet, and by the time that battle had cooled, had largely become a journal in professional statistics (Bertoldi et al., 2024). We thus need to turn elsewhere to find a comparative dataset for the Mendelian population concept. For this purpose, using precisely the same methodology, we constructed a topic model of *Journal of Genetics* from its founding in 1910 until 1960. This set of documents contains 923 articles, and notably begins a decade later than the *Biometrika* corpus. As it turns out, "population" appears in two topics. One is largely about natural selection in an ecological context and is roughly constant across those fifty years (we will not consider it further here).[9] But the other, which matches "population" with other highly probable terms like "line," "gene," "generation," "selection," "effect," and "genetic," seems to do a much better job of picking out a Mendelian population concept.[10] Highly probable papers for this topic include a number of quantitative studies of genetic inheritance in experimental populations (including, among many others, a long series of studies series commencing with Robertson & Eric, 1952; Reeve & Robertson, 1953), and evaluation of natural selection, gene flow, and evolution in both natural and domestic populations (Diver et al., 1925; James & McBride, 1958).

At least for present purposes, let's assume that these four topics do indeed pick out sets of articles that use each of these four concepts of "population" we described historically in the previous two sections.

---

[9] This may offer a locus for future work interacting with the literature on the role of ecology in understanding populations (e.g., Millstein, 2013).

[10] Data for this topic model (excluding the copyrighted text of the journal articles themselves) can be found at https://doi.org/10.6084/m9.figshare.27987335. Technical details of model construction have been omitted here for readability, but parallel those of Bertoldi et al., (2024).

What do they teach us? This will largely be the subject of the next subsection. But for the moment, we can say that their prevalence shows precisely what we would expect. Consider the graph in Fig. 1. The pre-Fisherian sense of population, while a bit under-represented in the early days of *Biometrika* (likely because of the increased prevalence of non-statistical, biological articles in these years), reaches its peak in 1915 and then steadily declines until practically disappearing by 1960. The same is the case for the pre-Mendelian sense of population. Correspondingly, the Mendelian and Fisherian senses of population practically do not appear in their respective journals prior to 1930, and then rapidly increase in frequency.

In short, the results of topic modeling confirm precisely the same transition that we have described historically above: concepts of pre-Mendelian and pre-Fisherian "populations" are rapidly and almost completely replaced in the span of a few decades from, roughly, 1930 to 1950.

3.2. On populations and "population"

Readers may already have noted that our analysis is tightly focused on the *word* "population," as it appears in *Biometrika* and *Journal of Genetics*. Some of those readers[11] may thus be concerned that we have conflated the term "population" with the concept of *population* (here in italics to distinguish a *sensu lato* version of the notion that extends beyond the bounds of the term itself), and have therefore presented a merely partial view of *population*, which is ostensibly the target of our analysis. Above, in introducing the history of the concept, we already mentioned that "variety" and "race" occasionally play the same terminological role; a reviewer noted that "breeding," "interbreeding," and "species" also might be involved in these discussions during the period that we analyze.

This, then, constitutes an empirical claim about the bounds of *population* that we can test in the models that we have constructed. Are these other terms (we will stick with these five for now) present with a significant probability in topics other than the four that we have selected for analysis above? If they are, do those topics look like they might also capture important aspects of *population*? And if both those things are true, do these same articles also *not* include the term "population," so that they have plausibly been missed by the analysis that we have presented above? If all of these things are true, then our four topics above will not have successfully captured the appropriate concept of *population*.

First, let's look at *Biometrika*. The only one of those five variant terms that appears in the top twenty words for any topic within the model is "race," which is the 11th most probable term in a topic that Bertoldi et al. (2024) refer to as "A-Skull-Measurement." This is a topic covering almost exclusively eugenic and craniometric papers, and thus its invocation of "race" is a eugenic one, not one in the sense of *population*. It thus seems that no other language for *population* is present in *Biometrika* at any significant frequency.

The story for *Journal of Genetics* is a bit more complicated. Set aside the two topics that we have already discussed above (both the one that we analyzed as invoking the Mendelian notion of population and the one that we set aside as concerning natural selection and ecology), as well as "interbreeding," which does not appear in the top twenty most probable words in any topic. In that topic model, five topics include at least one of those five words in their top twenty most probable terms.[12] Let's consider each quickly in turn. One topic invokes "breed," because it also includes extensive discussion of cattle. This is thus an agricultural sense of "breed" as a descriptive term, not a discussion of *population*. A second

uses both "breed" and "race," though its top words are "blue," "white," "single," "double," "cream," "yellow," and "green," leading to a natural interpretation that these are words used to describe morphological characteristics, and for this reason only incidentally appear near "breed" and "race." A third uses only "variety," but the rest of its terms are almost exclusively from botany; the topic describes botanical crossing experiments. Two more use "species"; they both seem to describe cellular-biological and biochemical experiments on chromosomes and the earliest phases of development. None of these five topics seem to be important to capturing *population*.

To confirm this interpretation, we can look at the cross-topic correlation values. The appearance of the topic that we have selected as capturing Mendelian populations in a document is only significantly correlated with the appearance of one other topic, one on recombination and linkage experiments in mice. Put differently, none of the terms that invoke variant terminology for *population* seem to be plausibly connected to the one topic that we can be fairly certain does indeed discuss the concept.

In short, while one must always be cautious that word- or term-based methodologies do actually capture the concept that we intend them to, all evidence seems to indicate that in this corpus we have, first, indeed properly targeted the notions of population that we have intended to and, second, not left aside other places in the corpus where this concept is discussed.

3.3. Shedding light on the concept(s) of population

Let's return to the results we presented in section 3.1, which are useful for more than simply confirming our pre-existing historical intuitions. We can also use the perspective on the literature that we have provided here to evaluate historical claims about this period, and even to direct potential lines of future inquiry into the nature of and relationships between these various concepts of "population".

Immediately, we can see that Morrison's contention that "the biological" and "the mathematical" are largely inseparable during this period (Morrison, 2004) is not borne out by our work here. Topic models in these two journals, at least, can quite clearly distinguish "biological" from "statistical" notions of "population", largely, of course, by virtue of the correlation between "population" and either biological or statistical concepts that travel with them. In that sense, these journals offer indirect support for Hey's argument that conflation of these different notions would therefore be problematic and is something to be avoided.

Digging a bit deeper, let's see whether our analysis succeeds or fails at recognizing that Mayr is *not* using a Mendelian notion of "population" when he discusses population thinking. Put differently, can we prove that our models are *not* making the kind of conflation that Hey cautions us against? Topic models, in addition to describing the content of an extant corpus, can also be used to classify new documents. In essence, we can ask the following question: had one of Mayr's publications on population thinking been published in *Journal of Genetics*, would our model have mistakenly believed that it was using a Mendelian concept of "population"? The answer is no. We used the topic model for *Journal of Genetics* to classify Mayr's "Darwin's Impact on Modern Thought" (Mayr, 1995); the probability of this document choosing a word from the topic associated with Mendelian populations is a mere 0.000005, one in two-million. Informally, we could say that Mayr is, by this metric, very clearly *not* using a Mendelian population concept. Of course, this result is nothing more than illustrative – we have not prepared or exhaustively analyzed a corpus of Mayr's writing on population thinking – but it underlines the idea that in fact these concepts are fairly easily distinguishable and offers yet more support for Hey's worry that they are being conflated.

Moving beyond Mayr, instead of considering documents and their relation to topics, we might ask which *authors* are most associated with each of the four topics that we have studied here. Which authors, that is, are most likely to discuss each of these four "population" concepts? For a

---

[11] As well as an anonymous reviewer, whom we thank for encouraging us to explore the issues raised in this subsection.

[12] Notably, none of them ranks higher than the eighth-most probable term for each of those topics.

Table 1

The authors that published the largest number of articles in the topics that correspond to each sense of "population." Listed are the top ten authors in strict order, then (following the "*and*") other authors from the top twenty-five that are of particular historical or contemporary interest.

Pre-Mendelian Biological	Mendelian Biological	Pre-Fisherian Statistical	Fisherian Statistical
Pearson, Karl	Robertson, Forbes W.	Pearson, Karl	Freeman, P. R.
Leslie, P. H.	Maynard Smith, J.	Pearson, Egon S.	Page, E. S.
Bailey, Norman T. J.	Reeve, E. C. R.	David, Florence N.	Armitage, Peter
Greenwood, Major	Michie, Donald	Student	Rao, J. N. K.
Becker, Niels G.	Bateman, K. G.	Wishart, John	Jennison, Christopher
Maynard, G. D.	Robinson, Roy	Isserlis, L.	Sobel, Milton
MacDonell, W. R.	Robertson, Alan	Lee, Alice	Wei, L. J.
Stocks, Percy	Falconer, D. S.	Geary, R. C.	Demets, David L.
Seber, George A. F.	Schmidt, Johs	Haldane, J. B. S.	Whitehead, John
Elderton, Ethel M.	Sismanidis, A.	Neyman, J.	Mehta, Cyrus R.
and	*and*	*and*	*and*
Pearson, Egon S.	Haldane, J. B. S.	Elderton, W. Palin	Turnbull, Bruce W.
Brownlee, John	Dunn, L. C.	Johnson, Norman L.	Sedransk, J.
Pearl, Raymond	Charles, Enid	Hartley, H. O.	Berry, Donald A.
Snow, E. C.	Hogben, Lancelot	Yule, G. Udny	Cox, D. R.

brief summary, see Table 1. A number of features of the table are worthy of comment. First, Karl Pearson's dominance over the "early" period of *Biometrika* is evident: he was interested in both sorts of "population," and he is by far the leading author for both of the "pre-" topics. The "Fisherian Statistical" column is practically a who's-who of contemporary statistical study, and this in a variety of different guises: from traditional mathematical statistics to biostatistics and actuarial science.

A few names stand out for the role that they play in other historical narratives. G. Udny Yule, perhaps best known today as having written a kind of "proto-synthesis" of biometry and Mendelism (Yule, 1902a, 1902b; see Tabery, 2004; Pence, 2022, pp. 115–19) reveals himself here as a more straightforward statistician, though one who was quite comfortable in the pre-Fisherian tradition (an account confirmed in his obituary; Yates, 1952). In addition to the American biostatistician Raymond Pearl, we also see in the "pre-Mendelian biological" camp two authors who wrote papers on correlation coefficients that would later be crucially important to the early Fisher: John Brownlee and E. C. Snow (Brownlee, 1910; Snow, 1910; see Pence, 2022, pp. 139–40). The "Mendelian biological" camp contains a number of authors familiar to students of the Modern Synthesis, such as John Maynard Smith and Haldane, as well as a few authors (like Enid Charles and Lancelot Hogben, both known as medical/demographic statisticians) who might be seen as offering us continuity between multiple senses of "population."

While we lack the space here to pursue this data in the detail that it might deserve, it can provide us with a helpful way into a variety of deeper philosophical, historical, and sociological studies of the nature and spread of these four concepts. We see in this case all the hallmarks of a complex transition in the history of science: points of commonality between the biological and the statistical (Karl and Egon Pearson, and the Eldertons, for instance), as well as points of radical rupture (such as the much more contemporary nature of the Fisherian-statistical population concept, which appears almost to divorce it from its history from the model's perspective). The relative ease with which the topic model separates statistical and biological notions of "population" seems to forestall any quick argument that these concepts are used identically in *Biometrika* or *Journal of Genetics*. But the non-trivial overlap between the authors at work on both notions, as well as the historical story linking the two, also stands in the way of any narrative of complete separation. Digital analyses can help us nuance these stories, by giving us access to "minor" figures and broad trends that are hard to detect via close

reading.

At the very least, we have ample reason to agree with Hey that the two ideas of biological and statistical population "have an interesting historical connection" (Hey, 2011, p. 262). We would hasten to add that the complexity and interest of that historical connection – or perhaps better, the historical *transformation* that these concepts underwent in the first half of the twentieth century – makes them assuredly deserving of further scholarly focus.

CRediT authorship contribution statement

Acknowledgments

Data availability

Data generated for this paper are publicly available; see footnote 9.

References

Ariew, A. (2008). Population thinking. In M. Ruse (Ed.), *The Oxford handbook of philosophy of biology* (pp. 64–86). Oxford: Oxford University Press.

Ariew, A. (2022). Charles Darwin as a statistical thinker. *Studies in History and Philosophy of Science, 95*, 215–223.

Ariew, A., Rice, C., & Rohwer, Y. (2015). Autonomous-statistical explanations and natural selection. *The British Journal for the Philosophy of Science, 66*, 635–658. https://doi.org/10.1093/bjps/axt054

Ariew, A., Rohwer, Y., & Rice, C. (2017). Galton, reversion and the quincunx: The rise of statistical explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences, 66*, 63–72. https://doi.org/10.1016/j.shpsc.2017.08.001

Bateson, W. (1909). Heredity and variation in modern lights. In A. C. Seward (Ed.), *Darwin and modern science* (pp. 85–101). Cambridge: Cambridge University Press.

Bertoldi, N. (2022). "Batesonian Mendelism" and "Pearsonian biometry": Shedding new light on the controversy between William Bateson and Karl Pearson. *History and Philosophy of the Life Sciences, 44*(4), 49. https://doi.org/10.1007/s40656-022-00528-5

Bertoldi, N., Lareau, F., Pence, C. H., & Malaterre, C. (2024). A quantitative window on the history of statistics: Topic-modelling 120 years of Biometrika. *Digital Scholarship in the Humanities, 39*, 13–29. https://doi.org/10.1093/llc/fqad072

Blakeman, J., & Pearson, K. (1906). On the probable error of mean-square contingency. *Biometrika, 5*(1/2), 191. https://doi.org/10.2307/2331660

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77. https://doi.org/10.1145/2133806.2133826

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval, 11*(2–3), 143–296. https://doi.org/10.1561/1500000030

Brownlee, J. (1910). The significance of the correlation coefficient when applied to Mendelian distributions. *Proceedings of the Royal Society of Edinburgh, 30*, 473–507. https://doi.org/10.1017/S0370164600030935

Chung, C. J. (2000). *Essence, variation, and evolution: An analysis of Ernst Mayr's distinction between 'typological' and 'population' thinking.* Minneapolis: University of Minnesota. Ph.D. Thesis.

Darwin, C. (1857a). Darwin correspondence project. Letter no. 2123 https://www.darwinproject.ac.uk/letter/?docId=letters/DCP-LETT-2123.xml. (Accessed 20 August 2024).

Darwin, C. (1857b). Darwin correspondence project. Letter no. 2130 https://www.darwinproject.ac.uk/letter/?docId=letters/DCP-LETT-2130.xml. (Accessed 20 August 2024).

Darwin, C. (1859). *On the origin of species* (1st ed.). London: John Murray.

Darwin, C. (1873). On the males and complemental males of certain cirripedes, and on rudimentary structures. *Nature, 8*, 431–432.

Darwin, C. (1874). *The descent of man and selection in relation to sex* (2nd ed.). London: John Murray.

Darwin, C. (1875). *The variation of animals and plants under domestication* (2nd) (2nd, Vol. 2). London: John Murray.

Depew, D. J., & Weber, B. H. (1995). *Darwinism evolving: Systems dynamics and the genealogy of natural selection*. Cambridge, MA: Bradford Books.

Diver, C., Boycott, A. E., & Garstang, S. (1925). The inheritance of inverse symmetry in *Limnaea peregra*. *Journal of Genetics, 15*(2), 113–200. https://doi.org/10.1007/BF02983106

Dobzhansky, T. (1927). Studies on the manifold effect of certain genes in *Drosophila melanogaster*. *Zeitschrift für induktive Abstammungs- und Vererbungslehre, 43*(1), 330–388. https://doi.org/10.1007/BF01741014

Dobzhansky, T. (1930). Genetical and environmental factors influencing the type of intersexes in *Drosophila melanogaster*. *The American Naturalist, 64*(692), 261–271. https://doi.org/10.1086/280315

Dobzhansky, T. (1934). Studies on hybrid sterility. I. Spermatogenesis in pure and hybrid *Drosophila pseudoobscura*. *Zeitschrift für Zellforschung und Mikroskopische Anatomie, 21*, 169–233. https://doi.org/10.1007/BF00374056

Dobzhansky, T. (1935). A critique of the species concept in biology. *Philosophy of Science, 2*(3), 344–355. https://doi.org/10.1086/286379

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics, 21*(2), 113–135. https://doi.org/10.1093/genetics/21.2.113

Dobzhansky, T. (1950). Mendelian populations and their evolution. *The American Naturalist, 84*(819), 401–418.

Droesbeke, J. J. (2021). *Adolphe Quetelet: Passeur d'idées*. Bruxelles: Académie Royale des Sciences, des Lettres et des Beaux–Arts de Belgique.

Edwards, A. W. F. (1994). R. A. Fisher on Karl Pearson. *Notes and Records of the Royal Society of London, 48*(1), 97–106.

Elderton, E. M. (1914). Note on infantile mortality and employment of women. *Biometrika, 10*(1), 193. https://doi.org/10.2307/2331750

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh, 52*, 399–433.

Fisher, R. A. (1922a). Darwinian evolution by mutations. *Eugenics Review, 14*, 31–34.

Fisher, R. A. (1922b). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh, 42*, 321–341. https://doi.org/10.1016/S0092-8240(05)80012-6

Fisher, R. A. (1922c). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A, 222*(602), 309–368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences* (1st ed.). London: Macmillan.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Gannett, L. (2001). Racism and human genome diversity research: The ethical limits of "population thinking". *Philosophy of Science, 63*(3), S479–S492.

Gannett, L. (2003). Making populations: Bounding genes in space and in time. *Philosophy of Science, 70*(5), 989–1001. https://doi.org/10.1086/377383

Gayon, J. (1998). *Darwinism's struggle for survival: Heredity and the hypothesis of natural selection*. Cambridge: Cambridge University Press.

Gigerenzer, G., Swijtink, Z., Porter, T. M., Daston, L., & Lorenz, K. (1990). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Hacker, H. P., & Pearson, H. S. (1946). The growth, survival, wandering and variation of the long-tailed field mouse, *Apodemus sylvaticus*. *Biometrika, 33*(4), 333. https://doi.org/10.2307/2332198

Haldane, J. B. S. (1932). *The causes of evolution*. London: Longmans, Green and Company.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*. https://doi.org/10.1126/science.28.706.49

Hey, J. (2011). Regarding the confusion between the population concept and Mayr's "population thinking". *Quarterly Review of Biology, 86*(3), 253–264.

James, J. W., & McBride, G. (1958). The spread of genes by natural and artificial selection in closed poultry flock. *Journal of Genetics, 56*(1), 55–62. https://doi.org/10.1007/BF02984720

Johannsen, W. (1903a). Ueber Erblichkeit in Populationen und in reinen Linien. *Ein Beitrag zur Beleuchtung schwebender Selektionsfragen*. Jena: Gustav Fischer.

Johannsen, W. (1903b). Om Darwinismen, set fra Arvelighedslærens Standpunkt [About Darwinism, seen from the point of view of the science of heredity]. *Tilskueren, translated by Nils Roll-Hansen. BSHS Translations*. https://www.bshs.org.uk/bshs-translations/johannsen.

Johannsen, W. (1907). Does hybridisation increase fluctuating variability? In W. Wilks (Ed.), *Report of the third international conference on genetics* (pp. 98–112). London: Spottiswoode & Co.

Johannsen, W. (1909). *Elemente der exakten Erblichkeitslehre*. Jena: Gustav Fischer.

Johnson, N. L. (1957). Optimal sampling for quota fulfilment. *Biometrika, 44*(3/4), 518. https://doi.org/10.2307/2332882

Karn, M. N. (1933). A further study of methods of constructing life tables when certain causes of death are eliminated. *Biometrika, 25*(1/2), 91. https://doi.org/10.2307/2332264

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique, 48*(2), 169–195. https://doi.org/10.2307/1403151

Lotsy, J. P. (1925). Species or linneon? *Genetica, 7*(5–6), 487–506. https://doi.org/10.1007/bf01676287

Mach, E. (1886), 1914. In *The analysis of sensations and the relation of the physical to the psychical* (5th ed.). Chicago and London: Open Court.

Magnello, M. E. (1996). Karl Pearson's Gresham lectures: W. F. R. Weldon, speciation, and the origins of Pearsonian statistics. *British Journal for the History of Science, 29*(1), 43–63.

Magnello, M. E. (1999a). The non-correlation of biometrics and eugenics: Rival forms of laboratory work in Karl Pearson's career at University College London, Part 1. *History of Science, 37*, 79–106.

Magnello, M. E. (1999b). The non-correlation of biometrics and eugenics: Rival forms of laboratory work in Karl Pearson's career at University College London, Part 2. *History of Science, 37*, 123–150.

Magnello, M. E. (2005). Karl Pearson's paper on the chi square goodness of fit test (1900). In *Landmark writings in western mathematics 1640-1940* (pp. 724–731). Elsevier Science.

Magnello, M. E. (2006). Victorian vital and mathematical statistics. *BSHM Bulletin: Journal of the British Society for the History of Mathematics, 21*(3), 219–229. https://doi.org/10.1080/17498430600964508

Magnello, M. E. (2009). Karl Pearson and the establishment of mathematical statistics. *International Statistical Review, 77*(1), 3–29. https://doi.org/10.1111/j.1751-5823.2009.00073.x

Magnello, M. E. (2014). Pearson, Karl: His life and contribution to statistics. *Wiley StatsRef: Statistics reference online*. Hoboken, N. J.: John Wiley & Sons.

Magnello, M. E., & Hardy, A. (2002). *The road to medical statistics*. Amsterdam: Editions Rodopi B. V.

Mahalanobis, P. C. (1937). The need of a sample survey of the growth of population in India. *Sankhyā, 3*(1), 58.

Malthus, T. R. (1826). *An essay on the principle of population* (6th ed.). London: John Murray.

Martin, W. J., & Cheeseman, E. A. (1938). Rural mortality. Its comparative sex incidence. *Biometrika, 30*(1/2), 57. https://doi.org/10.2307/2332223

Mayr, E. (1976). Typological versus population thinking. In *Evolution and the diversity of life: Selected essays* (pp. 26–29). Cambridge, MA: Belknap.

Mayr, E. (1995). Darwin's impact on modern thought. *Proceedings of the American Philosophical Society, 139*(4), 317–325.

Mendel, G. (1866). Experiments in plant hybridization. *Verhandlungen Des Naturforschenden Vereines in Brünn, IV*, 3–47.

Meunier, R. (2016). The many lives of experiments: Wilhelm Johannsen, selection, hybridization, and the complex relations of genes and characters. *History & Philosophy of the Life Sciences, 38*(1), 42–64. https://doi.org/10.1007/s40656-015-0093-7

Millstein, R. L. (2009). Populations as individuals. *Biological Theory, 4*(3), 267–273. https://doi.org/10.1162/biot.2009.4.3.267

Millstein, R. L. (2010a). The concepts of population and metapopulation in evolutionary biology and ecology. In M. A. Bell, D. J. Futuyma, W. F. Eanes, & J. S. Levinton (Eds.), *Evolution since Darwin: The first 150 years* (pp. 61–86). Sunderland, MA: Sinauer.

Millstein, R. L. (2010b). Should we be population pluralists? A reply to Stegenga. *Biological Theory, 5*, 271–275. https://doi.org/10.1162/BIOT_a_00051

Millstein, R. L. (2013). Exploring the status of population genetics: The role of ecology. *Biological Theory, 7*, 346–357. https://doi.org/10.1007/s13752-012-0056-0

Morrison, M. (2002). Modelling populations: Pearson and Fisher on Mendelism and biometry. *The British Journal for the Philosophy of Science, 53*(1), 39–68. https://doi.org/10.1093/bjps/53.1.39

Morrison, M. (2004). Population genetics and population thinking: Mathematics and the role of the individual. *Philosophy of Science, 71*(5), 1189–1200. https://doi.org/10.1086/425241

Muller, H. J. (1940). Bearing of the *Drosophila* work on systematics. In J. S. Huxley (Ed.), *The new systematics* (pp. 185–268). Oxford: Clarendon Press.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. *Biological Symposia, 6*, 71–125.

Nanay, B. (2010). Population thinking as trope nominalism. *Synthese, 177*, 91–109. https://doi.org/10.1007/s11229-009-9641-6

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*(4), 558–625. https://doi.org/10.2307/2342192

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics, 144*(4), 1331–1335. https://doi.org/10.1093/genetics/144.4.1331

Pearson, K. (1900). *The grammar of science* (2nd ed.). London: Adam and Charles Black.

Pearson, K. (1917). On the probable error of biserial η. *Biometrika, 11*(4), 292. https://doi.org/10.2307/2331828

Pence, C. H. (2022). *The rise of chance in evolutionary theory: A pompous parade of arithmetic*. London: Academic Press.

Porter, T. M. (1986). *Karl Pearson: The scientific life in a statistical age*. Princeton, NJ: Princeton University Press.

Porter, T. M. (2004). *The rise of statistical thinking, 1820-1900*. Princeton, NJ: Princeton University Press.

Potochnik, A., & Brian, M. G. (2012). The limitations of hierarchical organization. *Philosophy of Science, 79*(1), 120–140. https://doi.org/10.1086/663237

Powys, A. O. (1905). Data for the problem of evolution in man. On fertility, duration of life and reproductive selection. *Biometrika, 4*(3), 233. https://doi.org/10.2307/2331614

Provine, W. B. (1971). *The origins of theoretical population genetics*. Princeton, NJ: Princeton University Press.

Reeve, E. C. R., & Robertson, F. W. (1953). Studies in quantitative inheritance: II. Analysis of a strain of *Drosophila melanogaster* selected for long wings. *Journal of Genetics, 51*(2), 276–316. https://doi.org/10.1007/BF03023299

Robertson, F. W., & Eric, R. (1952). Studies in quantitative inheritance: I. The effects of selection of wing and thorax length in *Drosophila melanogaster*. *Journal of Genetics, 50* (3), 414–448. https://doi.org/10.1007/BF02986839

Roll-Hansen, N. (1978). Drosophila genetics: A reductionist research program. *Journal of the History of Biology, 11*(1), 159–210.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman and Company.

Smith, T. M. F. (2001). Biometrika centenary: Sample surveys. *Biometrika, 88*(1), 167–243. https://doi.org/10.1093/biomet/88.1.167

Snow, E. C. (1910). On the determination of the chief correlations between collaterals in the case of a simple mendelian population mating at random. *Proceedings of the Royal Society of London - Series B: Containing Papers of a Biological Character, 83*(561), 37–55.

Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science, 47*(3), 350–383. https://doi.org/10.1086/288942

Sober, E. (2006). *Conceptual issues in evolutionary biology* (3rd ed.). Cambridge, MA: Bradford Books.

Sophister. (1928). Discussion of small samples drawn from an infinite skew population. *Biometrika, 20A*(3/4), 389. https://doi.org/10.2307/2332119

Stegenga, J. (2010). "Population" is not a natural kind of kinds. *Biological Theory, 5*, 154–160. https://doi.org/10.1162/BIOT_a_00029

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Stocks, P. (1924). Cancer and goitre. *Biometrika, 16*(3/4), 364. https://doi.org/10.2307/2331713

Stoltzfus, A., & Cable, K. (2014). Mendelian-Mutationism: The forgotten evolutionary synthesis. *Journal of the History of Biology, 47*(4), 501–546. https://doi.org/10.1007/s10739-014-9383-2

Student. (1923). On testing varieties of cereals. *Biometrika, 15*(3/4), 271–293. https://doi.org/10.2307/2331868

Tabery, J. G. (2004). The 'evolutionary synthesis' of George Udny Yule. *Journal of the History of Biology, 37*(1), 73–101. https://doi.org/10.1023/B:HIST.0000020390.75208.ac

Vagholkar, M. K., & Wetherill, G. B. (1960). The most economical binomial sequential probability ratio test. *Biometrika, 47*(1/2), 103. https://doi.org/10.2307/2332963

Walsh, D. M. (2015). *Organisms, agency, and evolution*. Cambridge: Cambridge University Press.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg, 64*, 369–382.

Winther, R. G., Giordano, R., Edge, M. D., & Nielsen, R. (2015). The mind, the lab, and the field: Three kinds of populations in scientific practice. *Studies in History and Philosophy of Biological and Biomedical Sciences, 52*, 12–21. https://doi.org/10.1016/j.shpsc.2015.01.009

Witteveen, J. (2013). Rethinking 'typological' vs. 'population' thinking: A historical and philosophical reassessment of a troubled dichotomy. *Trinity College*. Cambridge: University of Cambridge. Ph.D. Thesis.

Witteveen, J. (2015). A temporary oversimplification': Mayr, Simpson, Dobzhansky, and the origins of the typology/population dichotomy (Part 1 of 2). *Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 20–33. https://doi.org/10.1016/j.shpsc.2015.09.007

Witteveen, J. (2016). A temporary oversimplification': Mayr, Simpson, Dobzhansky, and the origins of the typology/population dichotomy (Part 2 of 2). *Studies in History and Philosophy of Biological and Biomedical Sciences, 57*, 96–105. https://doi.org/10.1016/j.shpsc.2015.09.006

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*(2), 97–159.

Wright, S. (1943). Isolation by distance. *Genetics, 28*(2), 114–138. https://doi.org/10.1093/genetics/28.2.114

Yates, F. (1952). George Udny Yule. 1871–1951. *Obituary Notices of Fellows of the Royal Society, 8*(21), 308–323. https://doi.org/10.1098/rsbm.1952.0020

Yule, G. U. (1902a). Mendel's laws and their probable relations to intra-racial heredity [Part 1]. *New Phytologist, 1*(9), 193–207. https://doi.org/10.1111/j.1469-8137.1902.tb06590.x

Yule, G. U. (1902b). Mendel's laws and their probable relations to intra-racial heredity [Part 2]. *New Phytologist, 1*(10), 222–238. https://doi.org/10.1111/j.1469-8137.1902.tb07336.x