Speed-run Through Social Stratification

36-313, Statistics of Inequality and Discrimination

15 September 2022, Lecture 6

Contents

So far, we've looked at how to describe the distribution of quantitative traits like income and wealth across an entire population, common patterns in those distributions, and how to quantify the level of inequality across the population. We are presently going to switch focus a bit to group differences, that is to comparisons across populations or across sub-populations. In principle we could divide a population up in total arbitrary ways — right-handed vs. left-handed, or people whose Social Security numbers have "13" as the middle two digits versus the rest of us. In practice, the interesting and important comparisons line up with social divisions that are already widely recognized in the society. Social scientists who study **stratification**[1] have developed some useful concepts for talking about particular social divisions, and for organizing different types of social division, but these aren't necessarily familiar to everyone taking this course, hence today's lecture.

[1] In Latin, a *stratum* (plural *strata*) is a layer of something, like dirt or paint. Geological strata are layers of rocks laid down on top of each other. In speaking of "social stratification", the metaphor is that different social groups are like layers of different minerals stacked one upon another.

1 Types of stratification variables

There are two basic distinctions we can draw between *kinds* of social division, or the variables which track it. One, which is familiar to us as statisticians, is between variables which are purely **qualitative**, **categorical** or **nominal** on the one hand, versus those which are **quantitative** or at least **ordinal** on the other. Variables of the first type divide people into distinct types, but on the basis of some kind of qualitative distinction. (The different types usually have names that everyone in the society knows, hence[2] "nominal".) An example might be which political party you are affiliated with (or, in a one-party state, whether you are a member of *the* party). Variables of the second type rank people based on their having more or less of something — sometimes we can apply a precise, numerical value to the "more" or "less", other times it's a bit fuzzier, but it's still about ranking. Income is an outstanding example of a very numerical variable; "education" isn't *as* numerical, but we can and do still plainly talk about who's more educated than whom.

The other great distinction is between **ascribed** and **achieved** statuses. A status is **ascribed** to you by others because of factors completely beyond your control, primarily having to do with your birth and family of origin. Age is an outstanding example of an ascribed status, but there are many others, as we'll see[3]. **Achieved** or **attained** statuses, by contrast, are ones which are at least partly the result of what you have *done*. Income and education levels are also examples of achieved statuses[4].

We can thus organize our thinking about how society is divided up into qualitative versus quantitative divisions, and ascriptive versus achieved divisions. Let's look over some important kinds of stratification from this point of view.

1.1 Quantitative, achieved stratification

Income and **wealth** are outstanding examples of quantitative, achieved stratification variables. Income is a *flow* of money in over some unit of time (often a year); wealth is a *stock* of money accumulated. It's usually a mistake to compare them directly. While these are obviously related, it's important to bear in mind that they're distinct. Many high income people also have large debts, and so small or even negative net worth[5]. Against that, some ill-paid people might be debt-free and even own some assets, so they'd have lower incomes but higher net worths than well-paid debtors. I should also clarify here that when we talk about "income", we usually mean income in money, but this can be less than total compensation or consumption: farmers who grow the food they eat, for instance, have a higher level of consumption than their monetary income indicates[6], as do people who (say) receive free or subsidized services or products from the employer, their worshipers, or their government. Some economists put a lot of effort into trying to come up with a cash equivalent for these factors, and so estimating the distribution of consumption and income-equivalents.

Another achieved variable which clearly admits of ranking and comparisons of "more" or "less" is, as I said above, education. We are quite familiar with the idea that some people are more educated than others. I would, however, suggest that this is more of an ordinal and less of a numerical variable than income. It's very common, for instance, to categorize people by their education level as "less than high school", "high school graduate", "some college", "college graduate", "graduate or professional degree". These levels are ranked but not numerical. A common numerical measure is "years of schooling", but there's no real reason to think that the difference between the first and second year of schooling should be as big, in any sense, as the difference between the twelfth and thirteenth[7].

[2] From the Latin *nomen* "name".

[3] Saying a status is "ascribed" can make it sound *arbitrary*, as though you could have a different ascriptive status if everyone else would just cooperate. Sometimes that's the case, but the example of age should help make it clear that people can have very good reasons to ascribe certain statuses and not others, and indeed that some ascriptions can just be true.

[4] Saying that some dimension of inequality is about achievement doesn't necessarily mean that it's *fair*. If nothing else, many people would just not have had the chance to try for certain achievements!

[5] Recent graduates of expensive universities in high-salary jobs, for instance.

[6] Similarly for hunters. During the years when one of my uncles lived in the mountains of New Mexico, if you just looked at what he *bought*, you would have thought he hardly ever ate meat, and not realized that he usually had a separate freezer full of vension.

[7] Note that both these ways of measuring education, either as ordered levels or as a numerical magnitude, ignores differences

Closely related to education is the notion of different forms of metaphorical "capital" people can possess. Non-metaphorically, "capital" originally meant the resources of a business enterprise — either the money used to fund it ("financial capital"), or the plant and equipment it used to make money ("physical capital"). Around the 1950s, some economists introduced the notion of **human capital**, to refer to an individual worker's marketable skills (Mincer 1974). If you have some skill, and it lets you get a higher income than you'd have without that skill, then you have human capital, and the extra income is regarded, by economists, as the "return" on your human capital. Skill at plumbing and skill at programming are, in this perspective, both forms of human capital; this should suggest that it's multi-dimensional, since "Alice is a better programmer than Dave" makes sense, and "Babur is a better plumber than Chandra" makes sense, but "Alice is better at programming than Babur is at plumbing" doesn't, really[8].

Social capital refers to having valuable social connections: some combination of having strong, trustworthy ties[9], *and* having wide-ranging ties which bring you lots of information, or make you the crucial connection between different social circles. In a social networks course, we might look at different ways of quantifying this; here it's enough to say that someone with lots of close friends and allies *and* a wide-ranging acquaintance clearly has more social capital than an isolated loner. Social capital doesn't translate into income as directly as human capital, but it is nonetheless a valuable thing to have, and unequally distributed.

Cultural capital is essentially mastery and appreciation of what is recognized locally as important, prestigious forms of cultural knowledge. This can include correct manners[10], fashion sense, etc., but also knowledge and appreciation of the right kind of cultural artifact (movies, music, books, ...)[11]. Displaying the right kind of cultural capital confers prestige and opens doors.

I emphasize the "recognized locally" bit. As I said in lecture, my father's father had an encyclopedic knowledge of classical Persian poetry, and the ability to always come up with an apt quotation from it; this was valuable cultural capital in Afghanistan back in the day, but not much appreciated after he came to this country as a political refugee in the 1970s. Perhaps more relatably, a musical or other artistic "scene" is pretty much *defined* by having a very local sort of cultural capital which outsiders don't appreciate[12]. That being said, lots of societies recognize some sort of "mainstream" high culture, where people will go "oh yeah, that's the fancy stuff", even if they don't necessarily appreciate it themselves. Sometimes this means elevating certain genres of cultural artifacts over others (e.g., classical opera rather than musicals); in the contemporary US it seems to mean enjoying lots of genres, but being picky within every genre (Childress et al. 2021). Regardless of the *content* of cultural capital, it's "capital" because it's a valuable thing some people possess and others don't.

Schooling is, *in part*, about conferring human capital, social capital and cultural capital. We are, for instance, teaching you some skills which command a direct value on the market, which is human capital. We are introducing you to peers who it will be useful to know, at a stage in your lives where you form friendships easily, which is social capital. And we are helping you acquire the habits and tastes which go with high status, which is cultural capital. This isn't *all* a university is about[13], but it's *part* of schooling.

Having talked about economic variables, education and various sorts of capital, we should talk about **class**. This word gets used in two related but confusingly distinct ways. In one sense, a "class" is a group of people

in school quality. You are all here, rather than paying much less tuition at Allegheny Community College or even at Penn State, because you think CMU gives you a better education than either of those schools. I very much hope you're right about that, but that's harder to gather systematic data on, and starts to complicate comparisons. — Another issue with "years of schooling" is how to handle students who are held back a year because of failing classes or disciplinary issues.

[8]On the other hand, "Alice's programming skill raises her income by \$50k/yr, while Babur's plumbing skill raises his income by \$75k/yr" *does* make sense, so you could collapse down to one numerical dimension that way.

[9]As the old joke goes, "A friend is someone who'll help you move; a *good* friend is someone who'll help you move a body."

[10]Strictly speaking, knowing how to express oneself in a way that shows commitment to social justice, or at least doesn't offend those who feel such commitments, is *also* a form of cultural capital. (A lot of people have made this point; the version in Bovy (2017) is unusually thoughtful.)

[11]Strictly speaking, knowing how to express oneself in a way that shows commitment to social justice, or at least doesn't offend those who feel such commitments, is *also* a form of cultural capital. (A lot of people have made this point; the version in Bovy (2017) is unusually thoughtful.)

[12]In case this sounds too art-kid: Knowing Unix, including the `grep` utility, is a marketable skill, and would constitute human capital. Being able to use the *word* "grep" in casual conversation is cultural capital.

[13]If the university is doing right by you, there will be classes whose content never earns you a blessed dime, but which you find yourself thinking about for the rest of your life.

who make their living in the same (or similar) ways, and consequently have similar styles of life. In this sense one could talk about university professors, or corporate executives, or owners of building-trades companies, or retail cashiers, as a class. (You can also imagine that there might be broader- or finer- grained classes, as "small business owner" versus "master plumber".) Members of the same class, in this sense, will tend to have a lot of similarities in income, assets, education, and specific forms of human and cultural capital.

Another way in which "class" is used, however, is as a kind of summary of how high up the economic ladder someone is. A common system of ranks for class in this sense would be: upper, upper-middle, middle, lower-middle, working class, lower class. (Some would add, at this point: under-class.) These are rarely terms used with any precision, especially not "middle class". At one time, you could draw the distinction that those who were "working class" and below didn't have the resources, especially the assets, to *not* go out and work for *someone else* every day, while the middle class(es) and above did, and the upper class had people work for them[14]. One way to try to reconcile the two usages is to think of phrases like "the working class" as a short-hand for "the working class_es_", plural.

Finally, in this quadrant, **status**, **prestige** or **esteem** are variables on which we can rank people, they're ones for which "more" and "less" make sense, but are also achieved. Some status or prestige attaches to occupations — surgeons are higher status than, say, garbage-collectors. Some of it attaches to individual persons. (There are especially esteemed garbage-collectors, and widely despised surgeons.) "Coolness" should be seen as a kind of status or prestige (Heath and Potter 2004). Income *often* tracks prestige, but not always; being, say, a musician or painter does not, for most people, pay very well at all, but it carries prestige and coolness.

1.2 Quantitative, ascribed statuses

The most obvious entry here is, as I said, **age**. In modern American culture, we tend to think of age as just a numerical variable. We do however make a distinction between legal minors and those who have come into their legal majority, and between "seniors" and those younger than them. Some other cultures have maintained a much more elaborate system of "age grades", sometimes with formal initiation ceremonies.

Also in the quadrant of ranked but ascribed variables, we can include one's *parent's* income, wealth, education, and class. As we'll see when we cover the transmission of inequality, these are *very* predictive of people's lives, and certainly rank-able. But the class you were born into is an ascribed variable as far as you're concerned, however much of an achievement it may have been for your parents.

Historically, many societies have had a form of ascriptive status which we might call **rank** or **order**, such as noble or commoner, or distinctions among various grades of noble, or indeed various grades of commoner. As the names suggest, this isn't fully quantitative, but it is certainly ordinal. It is, importantly, distinct from economic status: some commoners might be richer than some nobles, indeed there might be many nobles who were fairly poor, but still had their noble status, and with it various privileges that were categorically unavailable to even the richest commoner[15]. Some of these societies had ways for people to change this sort of status, to become ennobled (or, more rarely, to lose their status of nobility), but this was a rare event. This isn't quite the same as caste, conceptually, because it was usually the case that marriages between nobles and commoners were regarded as legitimate.

This sort of inequality is not important in modern American life, because it was largely done away with by our Revolution, however inadvertently (Wood 1992).

[14] A related distinction, still very prominent in American speech, is between "white collar" and "blue collar" jobs. A "white collar" job was one where you could where a white shirt all day without dirtying it, because it was office work without manual labor. A "blue collar" job was one which did involve working with your hands and getting dirty, so you'd wear shirts which wouldn't show stains so much, typically made of blue fabric. ("Pink collar" is a much later play on this distinction, to refer to jobs primarily held by women.)

[15] See, e.g., Cobban (2012) for French examples.

1.3 Qualitative, achieved variables

Political party affiliation belongs here. This has been an important form of social division in many times and places, and is increasingly one in modern America. Party affiliation or membership often runs along family lines, but strictly speaking it's changeable and only something adults are really capable of.

I put **religion** here for similar reasons. Some religions do not allow conversion into the religion, and others make it very difficult, but pretty much every religion recognizes *leaving* the faith. (It may not *accept* it, but that's a different matter.)

1.4 Qualitative, ascriptive variables

1.4.1 Ethnicity

An **ethnicity** is a group of people who perceive themselves as, and are perceived by others as, sharing common habits, customs, traditions, and, usually, a common language, and are typically thought of as also united by shared ancestry. You will notice a lot of hedging in that sentence ("perceived", "thought of"). This is because the truth of this matters is, in this regard, vastly less important than people's beliefs, even or perhaps especially when the truth is demonstrably at odds with common belief. It's a well-worn joke among social scientists that an ethnic group consists of people who share a common mistake about their origins.

It's important to realize that ethnicity is often relative. In the US or Australia, "Chinese" or "Italian" or "Indian" can qualify as an ethnicity, but back home, the difference between a Hakka from coastal southeast China and a Mandarin speaker from Beijing, or between a Bengali and a Tamil, or between a Lombard and a Sicilian, looms much larger. It is possible that "Asian" or "Hispanic" will condense, in the US, into a single ethnicity, especially if something happens to interrupt continued immigration (though I wouldn't personally bet on that).

1.4.2 Race

Originally, the way people thought of races was as something like an umbrella group of ethnicities. The idea was that the human species naturally clustered into a fairly small number of *distinct* groups, biologically distinguishable and separated by ancestry. This is *not* how actual human biology works.

Now, obviously, if I look at a crowd at CMU, or walk down a street in Pittsburgh, I can make a pretty decent guess which continent many people's ancestors lived on 500 years ago. But this is because we've brought together people from widely-separated places fairly recently. When scientists have actually carefully studied the geographic patterns of genetic variation, it turns out that there are no abrupt breaks and no clusters[16], just smooth gradients. (We'll come back to this.) It probably didn't have to be this way. Many other animal species *do* cluster into geographically and genetically distinct sub-species. Indeed, at an early point in our evolutionary history, there were multiple sub-species of *Home sapiens* — not just ours, *Homo sapiens sapiens*, but also the Neanderthals and the Denisovans and possibly others[17]. A world in which there were, say, still surviving Neanderthals would be one in which the concept of "race" would make a lot more *biological* sense than it does in our world, at least as applied to the distinction between Neanderthals and us. A world in which other hominin species, such as *Australopithecus*, had also survived would be one where race made even more biological sense. Again, this is not our world.

1.4.2.1 Race in the Contemporary US

Currently, for official data-gathering purposes, the US recognizes the following races:

[16] If you run genetic data through algorithms which *have* to output clusters, they will indeed report clusters, but that's not exactly *evidence* of clustering, whatever credulous journalists may tell you.

[17] I say "sub-species" rather than species because paleonotological genetics show that *some* of our ancestors were, in fact, Neanderthals and Denisovans, and "can have fertile offspring together" is the criterion for belonging to the same species.

1. Native American or American Indian
2. White
3. Black
4. Asian
5. Other, which in practice is a combination of people who check more than one of the above, and immigrants from countries with very different racial categories.

The rules for who goes into which category are elaborate and somewhat arbitrary[18]. In particular, the rules about "black" call for special comment.

The official racial category of "black" includes African Americans, i.e., American descendants of slaves[19]; many immigrants from the Caribbean and their descendants (sometimes called Afro-Caribbeans); and recent immigrants from sub-Saharan Africa and their descendants. This is a *very* heterogeneous category; some of the recent African immigrant groups, for instance, have average income and education levels above the over-all US average, which African Americans decidedly do not. Historically, of course, "black" in the US has almost always meant the first of these sub-groups, the ones we now call African Americans. Genetically, (almost) everyone in this group has a mixture of African, European, and Native American ancestry. What determined membership in it was what came to be called the "one drop rule", that if you had any identifiable African ancestry, you were black, regardless of how far back it was or how you looked. Most other countries in the Americas imported slaves from Africa, and developed racial hierarchies, but most of the others tended to recognize more grades between black and white; the English colonies in North America, and later the US, tended to make it binary.

I have said that race is an ascribed social status, fixed at birth, and this is mostly accurate but isn't *entirely* true. Precisely because race matters a lot, some people are very motivated to take steps to change their racial status. One way this happens is through presenting[20] oneself as a different race than one was born into. Historically, an important form of this was African Americans moving somewhere new and there "passing as white". It's difficult to know exactly how common this has been, but some estimates (Stuckert 1958) suggest that a large percentage of the white American population actually had African ancestors[21]. Recent studies which survey the same people in multiple years show that at least some of them self-report different races at different times (along with changing self-reports for other supposedly-fixed identities and categories) (Egan 2020). Having said this, however, one way we can tell that race is *supposed* to be an ascribed status is that efforts to change one's race are widely regarded as deceptive, or at best as a sign of being seriously disturbed (Brubaker 2016).

At a less individualistic level, there has also been a lot of struggle over what the racial categories should be, and how different groups should be slotted into them. If we go back to the 1800s and even the first half of the 1900s, there was a strong current of opinion in Europe and the Americas that there were, in fact, different races of Europeans[22]. In large part, this took the form of "Nordic" northwestern Europeans claiming racial

[18] As an example, people from the Middle East are classified as white. (Think about where the Caucaus mountains [as in "Caucausians"] are located.) This includes Syrians, Lebanese, Iraqis, Iranians and Afghans. On the other hand, everyone from Pakistan and points east are "Asians", just like (say) people from Cambodia, Korea or Japan. From an official US point of view, two members of the same ethnic group, from opposite sides of the Afghanistan-Pakistan border, could be of different races, even if they were cousins.

[19] I realize that there's a controversial political movement which uses that name, but it *is* a pretty transparent name which clearly describes the relevant population.

[20] For the concept of "presentation of self in everyday life", see the classic and brilliant book of Goffman (1959).

[21] Staples (1999) is a brief, well-written popular account of some of the relevant history and social science. (In the ensuing two decades, inter-racial marriages have become *much* more common, see Alba (2020).) I don't think we have any good data on how common passing is nowadays. Demographers have tried to estimate "racial mobility" by comparing the number of people counted by the Census in each racial group in a given year (say, 2000) to the number of members that group should be expected to have, based on known births in previous years, known immigration and known age-specific survival rates (Perez and Hirschman 2009). These efforts show that there has been a small *net* transition from non-Hispanic white to other categories since around 1980. (It's plausible, but not certain, that much of this is people with mixed ancestry acknowledging it more often.) This sort of apporoach does not, however, let us identify the gross transitions from one category to another. Anecdotally, blacks passing as white was still happening as recently as the 1970s. I myself have two (unrelated!) acquaintances of my generation who discovered, as adults, that a parent or grand-parent had passed.

[22] This is why a classic skeptical history of racist ideas, Barzun (1965), is largely about people we'd all call whites being racist *towards each other*.

superiority over southern and eastern Europeans and the Irish. Starting in the 1920s, immigration to the US was legally limited to quite small levels, but immigration from southern and eastern Europe was *especially* limited, on explicitly racist grounds. Naturally, immigrants to the US from Ireland and southern and eastern Europe contested this, and the emergence of a single "white" category in modern times is, in part, a result of their successful efforts to be re-classified. It is only since 2000 that official US records allow people to identify themselves as members of more than one race, and this, too, is in part the result of persistent cultural and political struggle.

Here I should point out something which you may have already noticed, which is that "Hispanic" does *not* appear on the list of official racial categories. The official position of the US government is that "Hispanic"[23] is not the name of a race, but of an ethnicity, defined by a shared cultural heritage, and that Hispanics can be of any race. There is some logic to this[24], but in everyday life in America, Hispanic is often treated as though it was a race alongside black, white, Native American and Asian. This is why, in studies which use official data sources, you will often see phrases like "non-Hispanic whites" — it's an attempt to compromise between the official classification scheme and some conspicuous social realities. There is also evidence that the fact that all these disparate groups are lumped together in the US as "Hispanic" contributes to the formation of a new and distinct group identity in this country (Mora 2014).

1.4.3 Caste

At this point I should introduce another important kind of social division, namely **caste**. In anthropology or sociology, a caste is a group of people who are restricted to marrying and having children with others of the same caste (or perhaps a small range of closely-allied castes), and where membership in the caste is fixed at birth and passed from parent to child. More precisely, *legitimate* marriage and procreation is restricted to being within the caste; relations outside the caste are scandalous, and the offspring of such relations are always ascribed an unfavorable status. Castes are typically arranged in a hierarchy of prestige, and there are usually economic divisions, with certain occupations being restricted to certain castes. Extremely strong feelings of purity or pollution can attach to caste membership, so that (the wrong sort of) contact between high and low caste individuals may leave the former being seen as, and feeling themselves to be, unclean. The most elaborate and longest-lived caste system is that found in south Asia (especially but not just modern India), going back, in various forms, for thousands of years. But caste divisions, in this sense, have also been found in many other parts of the world, including ones we rarely think of as having them[25].

1.4.3.1 Caste and Race

Now, if you take this definition of "caste", it seems like a pretty good match for how, historically, Americans have understood race. Race was hierarchical, it was inherited and didn't much care about anyone's actual physical features, it was linked to occupation, there were strong feelings of purity and pollution (think of segregated water fountains), and mixed-race children, however common, were scandalous and almost by definition illegitimate[26]. In this sense, American race was a caste system, as some sociologists and anthropologists pointed out as long ago as the 1930s[27] (Warner 1936; Dollard 1937). Whether it still *is* a

[23] It is not my place to take sides on sometimes-heated disputes about whether to use the term "Hispanic", "Latin American", "Latin", "Latino", "Latina/o", "Latinx", "Latine", etc.; "Hispanic" is what's used officially.

[24] There is *something* which ethically Italian Argentines have in common with Dominicans, *mestizo* Puetro Ricans and even with Mayan villagers in Guatemala for whom Spanish is a second language, and which they *don't* have in common with (respectively) Italian Americans from the Bronx, Haitians, Canadian M{'e}tis, or Pueblo Native Americans.

[25] For instance, from around 1000 through the 19th century, parts of western France and northern Spain had a sub-population known as "Cagots", who were in effect a caste of untouchables. (I learned about this case from Appiah (2018), p. 28, who gives good references.) In that situation, there were effectively two castes, the Cagots and everyone else. According to Patterson (1982), pp. 48–51, this is the most common pattern cross-culturally. (Incidentally, Patterson reports that the existence of slavery is only weakly associated with the presence of castes, and that in many societies with both, the slave/free distinction cuts across caste lines.)

[26] Many US states had enforced laws against interracial marriage and "miscegenation" until they were over-turned by the aptly-named Supreme Court case of *Loving vs. Virginia* in 1967.

[27] Isabel Wilkerson's *Caste: The Origins of Our Discontents* (New York: Random House, 2020) is a very recent book-length exposition of this idea, by a distinguished journalist (author of the outstanding popular history Wilkerson (2010)), but I haven't

caste system, or has changed into something else, is less clear. Marrying and having children across racial boundaries is vastly more accepted than it used to be, but there are indications that African Americans form an exception to this, suggesting *that* caste boundary is still in place[28].

read it so I can't comment on it one way or the other.

[28] Alba (2020) has an excellent receive view of the evidence on both thee points.

2 Protected characteristics and categories

In US law, since the 1960s, a number of categories and characteristics are "protected", meaning that it's illegal to discriminate on that basis of these characteristics, or membership in these categories. More exactly, any discrimination on these bases is open to legal challenge, and needs to show a compelling justification. What the law enshrines as protected categories are: race, color, religion, sex, national origin, and age. More recent additions have included, in some contexts, disability, genetic information, gender identity and even, sometimes, citizenship. The law recognizes that there are some situations in which it is, in fact, OK to discriminate on these bases: the laws against religious discrimination do not force a Christian church to consider a Muslim or an atheist as a preacher. (The church might get in trouble, however, if it refused to consider hiring a Muslim as a janitor.)

It is notable that these are mostly qualitative, ascribed statuses, but not entirely. (Age is quantitative, religion is at least in-principle attained even if most people stick with whichever religion they were raised in.) Other qualitative, ascriptive characteristics are, legally at least, fair game. Thus there is no American federal law against discrimination on the basis of political party affiliation, or indeed caste[29]. Discrimination on the basis of income, wealth, education, cultural capital, or even class is perfectly[30] legal.

I want to emphasize that everything above is about what the law currently forbids or allows, not about how things *should* be, or even about what the law should be. As a citizen, I have some views about that, but this isn't the place for me to inflict them on you.

[29]Some states have civil rights laws which ban discrimination on the basis of "ancestry", and there is currently a case before the courts in California as to whether this applies to discrimination on the basis of caste in the Indian sense (see, e.g., Anahita Mukherji, "California's Legal Ground In Battling Caste Discrimination Takes Centre Stage In Historic Cisco Case", *The Wire* 10 March 2021, [https://thewire.in/caste/cisco-case-caste-discrimination-silicon-valley-ambedkar-organisations].) Everyone agrees that the federal laws against employment discrimination do not apply in this case, because Brahmin and Dalit Indians are of the same race and national origin.

[30]One wrinkle here is the legal notion of "disparate impact", which says that it's not OK to impose neutral-seeming tests which have the *effect* of harming, or favoring, people with protected characteristics, *unless* there is a good, functional reason for it. Thus the law against discrimination on the basis of national origin is usually held to prohibit discrimination in hiring on the basis of fluency in English and even on one's accent. If you apply for a job as a TV newscaster, the station can reasonably require you to be able to speak English fluently, but they couldn't, legally, reject you for having a Spanish or Chinese accent. (Or, at the very least, you'd have a good case against them, and they'd have to present a very compelling argument as to why a certain accent was a bona fide "business necessity".) Similarly landlords are generally free to raise rent, even if doing so has a disparate impact on member of protected categories.

3 Some Words on Genetic Differences between Social Groups

3.1 Endogamy creates genetic differences

Many social groups are **endogamous**: they mostly, or exclusively, marry and have children with others inside the group. This is part of the definition of "caste", but also very common for race, ethnicity, religion, and class.[31]

If a group is endogamous, and membership in the group is *socially* transmitted from parents to children, then over time there groups will become genetically distinct. This is a basic point, but worth appreciating. Assume that there are just two groups, says "reds" and "blues", and that, at the beginning of time, membership in the different groups is completely independent, statistically, of genes. Because the groups are finite, however, some genetic variants will be more common among reds than among blues, and vice versa. Because of endogamy and transmissible group membership, in the next generation the reds will be children of the original reds, and the blues will be children of the original blues, so the initial genetic differences will, on average, be passed on (this is called a genetic **founder effect**). But in fact a child inherits (on average) only a random half of each parent's genes[32], so there's a sampling process, and some genetic variation is lost from generation to generation. Because this sampling will happen independently in the two groups, they will increasingly diverge over the generations. This in turn means that as time goes on, it will be increasingly easy to infer whether someone is a red or a blue from their genome.

Thus the qualitative argument: endogamy *plus* social transmission of group membership *implies* gene pools that diverge over time. A natural question is how to make this quantitative, i.e., how much divergence should we expect between groups? This is a more complicated question in population genetics, which will involve the size of the founder effects, the population sizes of the two groups[33], the distribution of offspring per parent in each group[34], etc. It's also important to realize that this will happen over generations, and human generations are fairly long[35].

The next page shows a small simulation of what this mechanism can produce over a few generations in small populations. The code is in the .Rmd file for these notes, so you can get a sense of the range of outcomes, and how changing the parameters changes the distribution of outcomes.

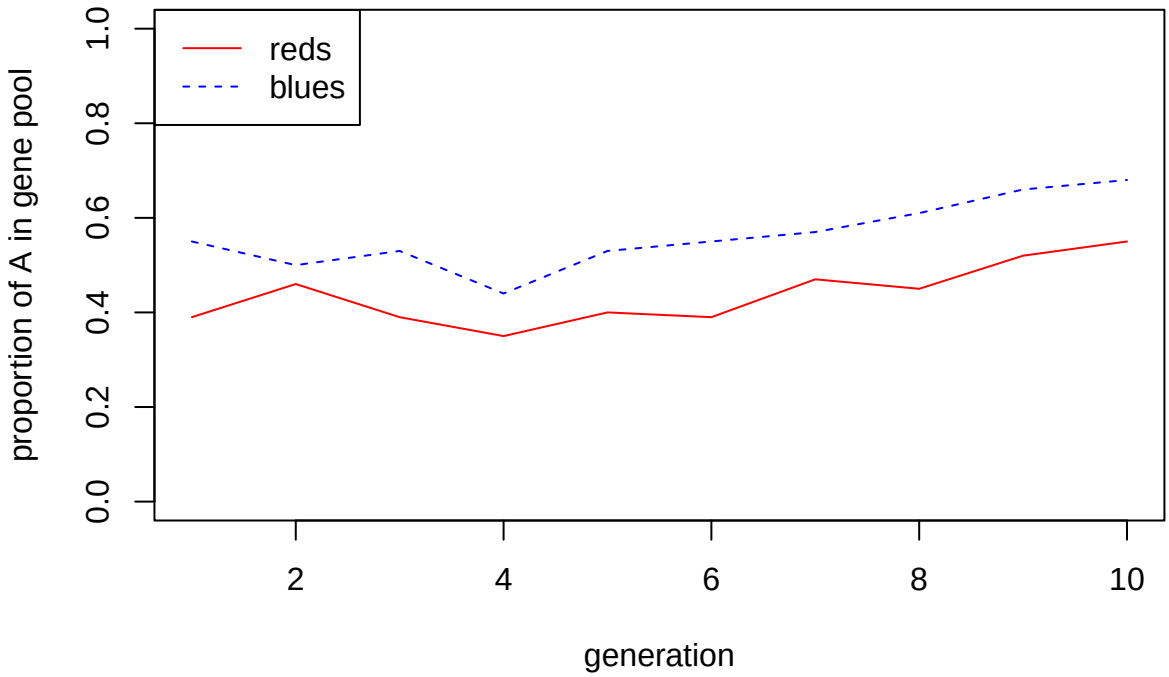
[31] Some people also use "endogamy" to refer to a tendency for people with similar values of *quantitative* traits, like income or education levels, to marry each other. Others describe that as **assortative marriage**. In animals, a (positive) correlation between quantitative traits of mates is called **assortative mating**.

[32] There are some exceptions, like mitochondrial DNA and sex chromosomes, but this isn't a genetics course.

[33] The smaller the groups, the greater the reduction in genetic variation due to sampling in each group.

[34] A higher average number of offspring will imply less sampling, but if the distribution is very skewed, the "effective" population size is reduced.

[35] In modern conditions, many adolescents are biologically capable of reproducing by age 14 or even earlier, but such early development requires levels of nutrition which were historically extremely rare. In pre-industrial conditions, the age of reproductive maturity might be closer to 18. (This needs to be distinguished from age of *marriage*, which might be much younger.) Historically, as now, many people do not just reproduce as soon as they're biologically capable, so taking a human generation as 20 or 25 years is a reasonable rule of thumb. When the simulation below shows 10 generations, then, you should imagine a passage of 250 years.



A small demonstration of how endogamy and cultural inheritance can create genetic differences between social groups. We are looking at one "locus" in the genome, where there are just two genetic variants, called "A" and "a" (geneticists have some weird notational conventions. . .), and plotting the proportion of the "A" variant in the gene pools of two endogamous, self-reproducing groups, the "reds" and the "blues". Initially, the "reds" and "blues" are divided without regard to genetics, but the children of reds are all reds, and the children of blues are all blues, so the gene pool of each generation in each group is a sample *of the group's gene pool in the parents' generation. Thus the gene pools at this locus diverge. Here, the "A" variant becomes somewhat more common among blues than among reds, but if you re-run the simulation, you'll see that it comes out the other way half of the time. Here, we haven't built up much divergence after 10 generations, but remember that the same kind of divergence will happen at each locus in the genome, and modern gene scans can pick out hundreds of thousands of loci, each of which would give* some *information about group membership.*

(To keep things simple, I have fixed the population size for each group at 100, and made each generation a uniform sample, with replacement, from the previous generation. [That is, I'm not simulating the genetic details of sexual reproduction, so pretend this is just looking at passing mitochondrial DNA from mothers to daughters.] By tinkering with the simulation, you can explore what happens if you change the with population sizes, have non-uniform distributions of offspring among parents, and how sex complicates everything.)

Quantitative precision will also require working back in some of the complicating factors that I omitted in sketching the argument above. Two paragraphs ago, I assumed that endogamy is perfect (there are no "mixed marriages" or children without marriage...), that group membership is 100% inherited in a purely social/cultural way (no one born red ever becomes blue or vice versa), that no natural selection acts on the genes, and that there are no mutations[36]. If the endogamy is imperfect, or if there is some migration between the groups, that will tend to make the groups more genetically similar. The consequences of natural selection will depend on whether the selection acts in the same way on the two groups (e.g., both get selected for resistance to locally-common diseases) or if they are subject to different selection pressures (e.g., the groups have different ecological niches or are exposed to subtly different environments)[37]. All of these factors can of course vary over time. Many social divisions are not very endogamous for very long, so they won't create large genetic differences, but strong, sustained endogamy *can* create big genetic differences.

So far, I have been talking only about genetic differences between social groups. But social groups are usually, indeed often, un-equal in various outcomes like income, education, or health. Now put this together: endogamy plus social inheritance makes groups genetically distinct; social groups have unequal outcomes; therefore, *genes can predict social outcomes*, even when the genes are functionally, causally, completely irrelevant to the outcomes. If gene A is more common among blues than among reds, and reds have higher levels of education for purely cultural reasons, gene A will predict educational attainment. At least, it will in a population of reds and blues; transporting the same genetic predictions to a different set of populations would work much less well.

Since this is not a course in quantitative population genetics, however, I will not draw this out any further, but just summarize that genetic differences between social groups are exactly what we should expect from the combination of endogamy and *social* transmission. We should even expect individual-level social outcomes to be *predictable* from genomes, whether or not the genes are in any sense *causes* of those outcomes.

3.2 "The history and geography of human genes"

I mentioned above that when scientists have actually investigated the distribution of genetic variations across human populations, they don't find discrete clusters, as we'd expect if there were distinct races. Instead, there are three outstanding findings:

1. Most of genetic variance is *within* populations, rather than *between* them[38];
2. The *average* gene frequencies in local populations do vary across space, but they vary smoothly, without abrupt breaks into distinct types;
3. There is some correlation across genes (if one population's gene pool contains more than the usual amount of gene A, then on average it will have more than the usual amount of gene B, too), but few of these correlations between genes are very strong.

Point (1) means that there is a lot of genetic variation within any human population. But there is nonetheless *some* average frequency for each gene[39] in a population, so we can think of those average frequencies as being high-dimensional vectors[40]. Point (2) is saying that those vectors vary smoothly as we move geographically over the surface of the Earth. If you think back to the example above about how endogamous social groups

[36] Since mutations would happen independently in the two groups, adding in mutuation *alone* would only make the groups more genetically distinct.

[37] Thus, for example, the genetic diseases of sickle-cell anemia and thalassemia are both side-effects of mutations which offer (partial) protection against malaria; the former is more common in west African and the latter in Mediterranean populations, both areas where malaria has long beenn a serious endemic disase. It is also important to realize that "environment" or "ecological niche" for humans can reflect culture and technology. The classic example is adult lactose tolerance. Lactose is the primary sugar in milk, and almost all mammals lose the ability to metabolize it as they become adults. Most humans are also lactose-intolerant as adults. There are, however, a number of populations which have long kept large herd animals which can be milked (goats, sheep, and especially cattle), and adult lactose tolerance has evolved multiple times among these groups.

[38] We will look in detail at parsing variance into within-group and between-group components very shortly in the course.

[39] Strictly speaking, each variant at each locus in the genome — what geneticists call **alleles**.

[40] There are roughly 30,000 human genes, almost all of which have multiple alleles, so these are at least 30,000 dimensional vectors.

will become increasingly distinct genetically, point (2) suggests that nearby populations have only been *partially* isolated from each other reproductively[41].

Point (3), the correlations between genes, suggests a way of usefully simplifying the complexity of gene frequencies. If two genes were perfectly correlated (or perfectly anti-correlated), we wouldn't need to track *both* of them, so we could get rid of one coordinate. If there are p dimensions but there were q perfect correlations, we could get away with just $p - q$ dimensions, because the vectors would lie on a $(p - q)$-dimensional plane in the p-dimensional space. The actual correlations aren't perfect, but we can nonetheless hope that the vectors lie *close* to a low-dimensional plane in the high-dimensional space. This has led geneticists, following Cavalli-Sforza, Menozzi, and Piazza (1993), to use dimension-reduction techniques[42] to explore the large-scale patterns of human genetic diversity, and what they reveal about the history of population movements, the expansion of agriculture, etc. These analyses show smooth gradients for genetic differences, not distinct races — but the techniques used don't *presume* such patterns.

[41] If the Yellows live between the Reds and the Greens, what we usually see is that the Reds and the Yellows are more genetically similar than are the Reds and the Greens. But if all three groups were completely endogamous, even if they diverged from a common pool of ancestors, there'd be no reason for the Reds to be more similar to the Yellows than to the Greens — certainly no reason for that to be a repeating pattern. If however there is "gene flow" between neighboring populations — if people sometimes have children with their neighbors — we'd expect to see exactly this pattern.

[42] Specifically, principal components analysis (PCA), a very old (Pearson 1901; Hotelling 1933a, 1933b) but very robust method of dimension reduction; if you're curious for details, see the chapter on PCA in Shalizi (n.d.).

4 Some optional exercises

4.1 Conceptual consolidation

1. In class, I drew a diagram where one axis was attained/ascribed, and the other was quantitative/qualitative, and located different social variables at different points on the diagram. *Without* consulting your notes, draw such a diagram with points for each social variable discussed here.

4.2 Extending the genetic simulation example

1. Run the simulation for two populations of size 100 and 10 generations. Calculate the probability that someone in generation 10 can be correctly classified as red or blue, depending on whether they have gene A or a. (Your answer will depend on how your simulation run turns out, but on my first try I got an accuracy of 63.5%.) *Hint*: First find $\mathbb{P}(red|A)$ and $\mathbb{P}(blue|a)$.
2. Run the simulation 9 more times, but change the gene labels from Aa to Bb, Cc, etc., down to Jj. (Keep the initial probabilities of all the genetic variants equal.) This is like looking at 10 independent loci in the genome. Taking the individuals in the tenth generation, fit a logistic regression which takes as its ten input features whether an individual has gene A or gene a, gene B or b, etc., and predicts whether the individual is red or blue. What accuracy does this logistic regression have for classifying individuals? (On my first try, I got an accuracy of 83%.)
3. Assign each individual a random number of years of education: blues are uniformly distributed on the integers between 10 and 14, reds are uniformly distributed on the integers between 11 and 20. Build a linear model which predicts years of education from the ten genetic features. What's the R^2? (On my first try, I got an R^2 of 0.2.)

(Using linear and logistic regression here has no particular statistical or genetic justification, but they're methods you know, which makes the exercises easier, and which geneticists know, which makes the exercises more realistic.)

5 Further Reading

Distinguishing between ascribed and attained forms of status is a very standard part of the tool-kit of sociology and anthropology; the book by Keister and Southgate (2012) in our readings is a pretty good source, but any introductory textbook on those subjects should cover it.

On ethnicity, nationality and nationalism, Gellner (1983) is essential reading; if you read only one book from this section, make it that one. If you have it in you to read *two* books about nationalism, make the second one Anderson (1991).

Exactly how we should understand "race", the extent to which race and caste are analogous (or race should be understood as a variety of caste), etc., etc., are subjects which fill libraries, particularly in this country[43]. Without pretending to have mastered this literature, I will somewhat arbitrarily recommend Appiah (2018), Loury (2002), Fields and Fields (2012) as recent-ish contributions that strike me as especially insightful and well-written. We will revisit issues related to racial classification, and its role in official statistics, later in the class, but if you are eager to read ahead, Anderson and Fienberg (2000) is very good.

Population genetics is a very important, and fundamentally probabilistic and statistical, branch of science. Gillespie (1998) is a good short introduction at about the mathematical level of these notes. The key figure in the development of large-scale studies of (to borrow the title Cavalli-Sforza, Menozzi, and Piazza (1994)) "the history and geography of human genes" was the late L. L. Cavalli-Sforza; the most accessible account of his work is Cavalli-Sforza (2000).

We will have a lot more to say about discrimination against protected categories, and how it can be established, later in the class. The discussion of protected categories and characteristics above, and later in the course, is overwhelmingly about US law, because this is a course at an American university and it's what I'm most familiar with. Mercat-Bruns (2016) provides an interesting comparison of American federal law on these topics with the relevant parts of the legal systems of France and of the European Union. (One interesting aspect of this, to which we will return, is the much greater role of statistics in American anti-discrimination law than in its European counterparts.) As to the rest of the world, I am too ignorant to comment.

I should probably add to these further-reading suggestions when I have time, but this will keep you busy, and if you have questions about specific topics, feel free to ask.

[43] "[T]o any non-American, the most oppressive feature of intercultural relations in America is not that people are racist, but just that they talk and think incessantly about race, even worse than the way the English talk and think incessantly about class" (Heath 2014, ch. 13).

References

Alba, Richard. 2020. *The Great Demographic Illusion: Majority, Minority, and the Expanding American Mainstream*. Princeton: Princeton University Press.

Anderson, Benedict. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Revised. London: Verso.

Anderson, Margo, and Stephen E. Fienberg. 2000. "Race and Ethnicity and the Controversy over the US Census." *Current Sociology* 48:87–110. https://doi.org/10.1177/0011392100048003007.

Appiah, Kwame Anthony. 2018. *The Lies That Bind: Rethinking Identity*. New York: Liveright Publishing.

Barzun, Jacques. 1965. *Race: A Study of Superstition*. Second. New York: Harper; Row.

Bovy, Phoebe Malz. 2017. *The Perils of "Privilege": Why Injustice Can't Be Solved by Accusing Others of Advantage*. New York: St. Martin's Press.

Brubaker, Rogers. 2016. *Trans: Gender and Race in an Age of Unsettled Identities*. Princeton, New Jersey: Princeton University Press. https://doi.org/10.2307/j.ctt1wf4ckd.

Cavalli-Sforza, Luigi L. 2000. *Genes, Peoples, and Languages*. New York: North Point Press.

Cavalli-Sforza, Luigi L., Paolo Menozzi, and Alberto Piazza. 1993. "Demic Expansions and Human Evolution." *Science* 259:639–46. https://doi.org/10.1126/science.8430313.

———. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Childress, Clayton, Shyon Baumann, Craig M. Rawlings, and Jean-Francois Nault. 2021. "Genres, Objects, and the Contemporary Expression of Higher-Status Tastes." *Sociological Science* 8:230–64. https://doi.org/10.15195/v8.a12.

Cobban, Alfred. 2012. *The Social Interpretation of the French Revolution*. Second. Cambridge, England: Cambridge University Press.

Dollard, John. 1937. *Caste and Class in a Southern Town*. New Haven, Connecticut: Yale University Press. https://hdl.handle.net/2027/heb.02855.

Egan, Patrick J. 2020. "Identity as Dependent Variable: How Americans Shift Their Identities to Align with Their Politics." *American Journal of Political Science* 64:699–716. https://doi.org/10.1111/ajps.12496.

Fields, Karen E., and Barbara J. Fields. 2012. *Racecraft: The Soul of Inequality in American Life*. London: Verso.

Gellner, Ernest. 1983. *Nations and Nationalism*. Ithaca, New York: Cornell University Press.

Gillespie, John H. 1998. *Population Genetics: A Concise Guide*. Baltimore: Johns Hopkins University Press.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Anchor Books.

Heath, Joseph. 2014. *Enlightenment 2.0: Restoring Sanity to Our Politics, Our Economy, and Our Lives*. New York: HarperCollins.

Heath, Joseph, and Andrew Potter. 2004. *Nation of Rebels: Why Counterculture Became Consumer Culture*. New York: Harper Business.

Hotelling, Harold. 1933a. "Analysis of a Complex of Statistical Variables into Principal Components [Part 1 of 2]." *Journal of Educational Psychology* 24:471–41. https://doi.org/10.1037/h0071325.

———. 1933b. "Analysis of a Complex of Statistical Variables into Principal Components [Part 2 of 2]." *Journal of Educational Psychology* 24:498–520. https://doi.org/10.1037/h0070888.

Keister, Lisa A., and Darby E. Southgate. 2012. *Inequality: A Contemporary Approach to Race, Class, and Gender*. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511842986.

Loury, Glenn C. 2002. *The Anatomy of Racial Inequality*. The W. E. B. DuBois Lectures. Cambridge, Massachusetts: Harvard University Press.

Mercat-Bruns, Marie. 2016. *Discriminating at Work: Comparing European, French, and American Law*. Berkeley: University of California Press.

Mincer, Jacob. 1974. *Schooling, Experience and Earnings*. New York: Columbia University Press.

Mora, G. Cristina. 2014. *Making Hispanics: How Activists, Bureaucrats, and Media Constructed a New American*. Chicago: University of Chicago Press.

Patterson, Orlando. 1982. *Slavery and Social Death: A Comparative Study*. Cambridge, Massachusetts: Harvard University Press.

Pearson, Karl. 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine* 2 (series 6):559–72. https://doi.org/10.1080/14786440109462720.

Perez, Anthony Daniel, and Charles Hirschman. 2009. "Estimating Net Interracial Mobility in the United States: A Residual Methods Approach." *Sociological Methodology* 39:31–71. https://doi.org/10.1111/j.1467-9531.2009.01220.x.

Shalizi, Cosma Rohilla. n.d. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV.

Staples, Brent. 1999. "The Real American Love Story." *Slate*. https://slate.com/news-and-politics/1999/10/the-real-american-love-story.html.

Stuckert, Robert P. 1958. "African Ancestry of the White American Population." *The Ohio Journal of Science* 58:155–60. https://kb.osu.edu/bitstream/handle/1811/4532/V58N03_155.pdf.

Warner, W. Lloyd. 1936. "American Caste and Class." *American Journal of Sociology*. https://doi.org/10.1086/217391.

Wilkerson, Isabel. 2010. *The Warmth of Other Suns: The Epic Story of America's Great Migration*. New York: Random House.

Wood, Gordon S. 1992. *The Radicalism of the American Revolution*. New York: Vintage.