Between-Group and Within-Group Inequality, and Comparison of Distributions

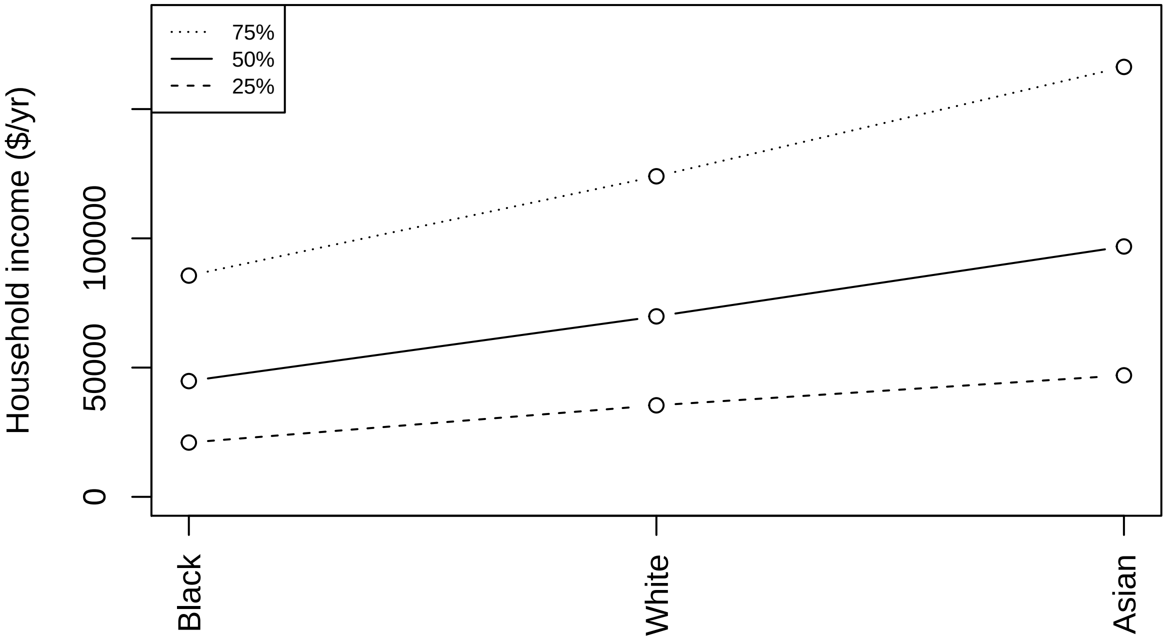
36-313

Lecture 7, 21 September 2021

Contents

Last time, we looked at some ways of measuring differences in typical values across groups, at testing whether differences are non-zero, and at quantifying the uncertainty in those differences. Today we'll continue the theme of looking at differences across groups. Our starting point will be the fact that there is often a fair amount of variation (and hence inequality) *within* groups of interest.

Here, for instance, is a little display where I've marked the median household income, and the 25th and 75th percentiles, by race, for the data-set you're working with in the homework[1]. (The lines here are just "guides to the eye.")



Clearly, there's a big difference in the typical income across these groups. Equally clearly, there's a lot of variation in income within each group. It's also clear, just from this plot, that there have to be other differences in income distribution across these groups, beyond just the typical values.

Between-group versus within-group inequality

If we have a notion of "typical value", and a measure of inequality, we can ask the following questions:

1. How much inequality is there in the typical values of different groups? This is the **between-group** inequality.
2. If we treated each group as its own population, how much inequality would we find in that population? This is the **within-group** inequality.
3. Does the over-all inequality of the population add up to the sum of between-group and within-group inequality?

The last question is a basically mathematical one; when the answer is "yes", it's easier to interpret the answers to the first two questions, about how much between-group and within-group inequality there is. The place where everything works out most nicely is when we take expectations as our notion of "typical value", and variance as our measure of inequality.

[1] I am using the sampling weights, which you don't have to, so your code could be rather simpler.

Analysis of variance

You remember the variance: for any numerical random variable, say Y, it's defined to be the expected squared distance from the mean:

$$\mathrm{Var}\left[Y\right] \equiv \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right]$$

This is equivalent to the expectation of the square minus the square of the expectation:

$$\mathrm{Var}\left[Y\right] = \mathbb{E}\left[Y^2\right] - (\mathbb{E}\left[Y\right])^2$$

The **law of total variance** says that we can always break this up, exactly, into between-group and within-group contributions[2]:

$$\mathrm{Var}\left[Y\right] = \mathrm{Var}\left[\mathbb{E}\left[Y|X\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[Y|X\right]\right]$$

The first part, $\mathrm{Var}\left[\mathbb{E}\left[Y|X\right]\right]$, is the between-group variance in group means. The second part, $\mathbb{E}\left[\mathrm{Var}\left[Y|X\right]\right]$, is the average of the within-group variances. It doesn't matter what X is here, we can condition on any other random variable and this holds.

This is called "analysis of variance" because "analysis", in its oldest sense, means dividing something into its constituent parts[3]. The formula analyzes the total variance of the population into the (average) variance around the expected value in each group, plus the variance of the group averages.

Between-group variance

It's not hard to convince yourself that

$$0 \leq \frac{\mathrm{Var}\left[\mathbb{E}\left[Y|X\right]\right]}{\mathrm{Var}\left[Y\right]} \leq 1$$

The extremes occur either when every group has the same expected value, or when there is no variation within any group around those expected values. Otherwise, this quantity is strictly between 0 and 1.

This ratio is usually written r^2 or R^2. It's sometimes called the "proportion of variance explained", but I think that's mis-leading; there's nothing here about an *explanation*, we've just sub-divided our population[4]. A better phrase would be "proportion of variance predicted" by the group averages.

ANOVA

The law of total variance suggests a way of building very simple but still useful models, which have come to be called ANOVA models. I'll work up to the general idea in stages.

Analysis of variance with one binary category

Suppose we divide the population into two groups (not necessarily of equal size). We pick one group, arbitrarily, as the "reference group", "reference level" or "reference class"; the other is the "contrast" group, level or class. If unit i is the reference class, we write $X_i = 0$, otherwise we write $X_i = 1$. Now here's the model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Here ϵ_i is a variable which we *call* noise, but really it includes everything we can't predict from knowing which group individual i is in. By construction, $\mathbb{E}\left[\epsilon_i|X_i\right] = 0$, but otherwise the distribution of ϵ_i doesn't matter much (at this stage).

[2] The proof isn't hard but we don't need it, so see the appendix if you're curious.

[3] It's the same "ana-" root as in "anatomy".

[4] I can't calculate it with this data, but I guarantee you that the R^2 of income when we classify by astrological sign will be small but not zero. (For that matter, if we number the signs from 1 to 12 and try to predict them from income, r^2 in that direction wouldn't be zero either.)

β_0 is called the **intercept** or **baseline** level of Y. β_1 is called the **contrast** between the two groups. Notice that

$$\mathbb{E}\left[Y|X=x\right] = \beta_0 + \beta_1 x$$

In particular,

$$\begin{aligned} \mathbb{E}\left[Y|X=0\right] &= \beta_0 & (1) \\ \mathbb{E}\left[Y|X=1\right] &= \beta_0 + \beta_1 & (2) \end{aligned}$$

Estimation

If we want to estimate this model, the obvious approach is to set β_0 equal to the sample mean when $X = 0$, and to set β_1 equal to the *difference* in sample means (the sort of difference we worked with last time).

Here, for instance, is a little model where Y = household income in 2020, and X = whether or not the person surveyed[5] had at least a bachelor's degree. (I've rounded to the nearest hundred.)

	Estimate	Std. Error
(Intercept)	70300	700
COLLEGETRUE	68100	1100

The way we read this is that everyone starts out with about 70 k dollars/year. Then those who are college-educated get an extra 68 k dollars/year. Then there's some noise added on to everyone. The important point is that68 k dollars/year isn't the model's estimate of the average income of the college educated, it's the *contrast* between those with and without college education.

The R^2 of this model is 0.084.

Analysis of variance with one more-than-binary category

If the category has more than two levels, say m of them, we introduce $m-1$ multiple **dummy** or **indicator** variables, say $X_1, X_2, \ldots X_{(m-1)}$.
If individual i is in class 0, the reference class, all of these indicator variables are zero. Otherwise, a single one of them "lights up" and becomes 1, indicating which class i belongs to. The model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_{m-1} X_{(m-1)i} + \epsilon_i$$

Now $\mathbb{E}\left[\epsilon|X_1, X_2, \ldots X_{m-1}\right] = 0$. (That is, the noise term has expectation zero, given *all* the Xs.)

Again, the β_js here are *contrasts*, specifically contrasts with the reference level. We can find the expected value for any group by taking β_0 and adding the appropriate contrast.

Estimation

- Find the sample mean for each group.
- Pick one group, it doesn't matter which, as the reference level. Its sample mean is the estimate of β_0, say $\hat{\beta}_0$.
- Subtract $\hat{\beta}_0$ from all of the other group means. What's left are the $\hat{\beta}_j$.

Here, for instance, we apply this to income by race:

[5] To make this data set easier for you to work with, I picked out one person from each multi-person household to represent it. The way I did so *should* have picked the head of the household.

	Estimate	Std. Error
(Intercept)	133600	2700
RACENAMEBlack	-66500	3200
RACENAMENBWA	-55100	4500
RACENAMEWhite	-34800	2800

Again, the reference level of race here is "Asian", so the fact that all of the contrasts are negative indicates that the mean incomes in all these other groups are lower than the mean Asian income. (That the reference level is also the highest-income level is a coincidence.) To find the expected household income level for whites, for instance, we take the baseline or intercept, 1.336×10^5 \$/yr, and add the coefficient for "white", -3.48×10^4 \$/yr, getting 9.88×10^4 \$/yr.

The R^2 of this model is 0.012. That is to say, knowing someone's race lets us predict about 1% of the variance in their household income. So even though the average income levels are differing by tens of thousands of dollars a year across the groups, there's still *even larger* inequality within each group.

ANOVA with cross-classifications

So far, I have been describing what's called "one-way analysis of variance", where we look at different levels of a *single* variable (educational level, race). It's natural to be interested in groups defined by *combinations* of variables — college-educated black people, non-college-educated Asians, etc. When we look at the combinations or intersections of two variables, this is called "two-way analysis of variance". Since three-way analysis of variance basically the same, just with an awkward tangle of notation, I'll stick to two-way ANOVA, and trust you to generalize when needed.

Two-way ANOVA without "interactions"

Let's first consider doing analysis of variance with two binary categories. We pick one level for each categorical (or "classifying") variable to be the reference level for that variable. (It doesn't matter which level is the reference level.) We then write $X_{1i} = 0$ if unit i in the reference level on variable 1, $X_{1i} = 1$ if i is in the reference level on variable 2, and similarly for X_{2i}. Notice that all four combinations of values for X_1 and X_2 are possible. The model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

At this point there are two important issues which I've been able to skip over beforehand.

1. This model might be wrong.
2. Estimating this model is not just a matter of calculating a few sample averages.

As to (1), when we did one-way analyses of variance, that was really just a matter of organizing the within-group averages in a particular way, as baselines-plus-contrasts. The within-group expectations always exist, and then it's a matter of basic algebra that we can take one of them as the reference level and subtract it out from all the others without losing any information. It's not even a model so much as a way of doing the book-keeping. As a representation of what's going on out there in the world, it may be *incomplete*, but it can hardly be *wrong*.

This is very much *not* the case with the two-way ANOVA model I just presented. The reason is that it presumes there's some increment to the average value of Y from variable 1, and another increment on Y from variable 2, and that these two increments just *add up*. Additive models are *mathematically* convenient, and easy for us to interpret, but there's no automatic reason to think they fit reality. The sociologist Aage Sørensen had a great metaphor which really captures what's going on in an additive model, and since my own version in class came out rather mangled, I'll just quote Sørensen (1998), pp. 248–249:

> Additive models can represent peculiar theories. As an example, consider an earnings attainment model. An additive model suggests that each of the independent variables represents some contribution to a person's earnings. A standard sociological earnings model inspired by class analysis might propose a model with class being one independent variable and other independent variables being education, gender, and family background. . . . This model, in fact, proposes a theory where each person receives x dollars from education, y dollars from family background, q dollars from gender, and z dollars from class. All of it adds up to the person's yearly earnings. We can imagine people walking around among pumps in a large gas station getting something from each of the pumps. The picture should be completed by specifying hypotheses about how many dollars each pump provides, and this would give us some idea of the relative importance of pumps that we could teach in courses on getting ahead in society. What one could call the "gas station theory" of earnings has apparently become universally accepted by sociologists, if we infer theories from the models used.[6]

As Sørensen goes on to say, there are times when additive models are useful and appropriate, but we shouldn't just assume that they're right, or even that there's a presumption in their favor.

So much for the weirdness of the gas-station theory of income determination. Let me turn to (2), estimation. With the one-way ANOVA, where we look at different levels of a single classifying variable, taking the sample average for each level and doing the book-keeping was a simple and reliable way of estimating the model. That's no longer true for the two-way ANOVA. Look at the conditional expectations implied by the model:

$$\begin{aligned} \mathbb{E}[Y|X_1 = 0, X_2 = 0] &= \beta_0 && (3) \\ \mathbb{E}[Y|X_1 = 1, X_2 = 0] &= \beta_0 + \beta_1 && (4) \\ \mathbb{E}[Y|X_1 = 0, X_2 = 1] &= \beta_0 + \beta_2 && (5) \\ \mathbb{E}[Y|X_1 = 1, X_2 = 1] &= \beta_0 + \beta_1 + \beta_2 && (6) \end{aligned}$$

We don't know the real expectations, but we can substitute in the sample means on the left-hand side as before. That leaves us with four equations, but only *three* unknowns. As you remember from algebra, when there are more equations than unknowns, there is generally *no* solution[7].

The Ancestors, in the late 1700s, faced similar problems of having more equations than unknowns centuries ago, and they devised a solution: the method of least squares[8]. This is about trying to find the combination of parameters which comes *closest* to matching the data, even though it doesn't *exactly* match the data.

The least-squares estimates of the parameters $\beta_0, \beta_1, \beta_2$ have to be determined together, by solving the following optimization problem:

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \underset{(b_0, b_1, b_2)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i}))^2$$

There is an explicit solution to this kind of least-squares problem in terms of linear algebra (see appendix), so it's computationally quite feasible to find the coefficients which best fit the data. A very fast and reliable numerical algorithm for doing so is instantiated in the R command `lm()`, which is what I've been using.

[6] Sørensen was a sociologist who was critique the customs of his own tribe, but essentially the same sort of model gets used by economists, psychologists, etc., etc.

[7] Let's write m_{00} for the sample mean when $X_1 = X_2 = 0$, etc. An obvious estimate of β_0 is m_{00}, of β_1 is $m_{10} - m_{00}$, and of β_2 is $m_{01} - m_{00}$. But then by the same logic $m_{11} - m_{00}$ should be our estimate of $\beta_1 + \beta_2$, which would imply $m_{01} + m_{10} - 2m_{00} = m_{11} - m_{00}$ or $m_{11} = m_{01} + m_{10} - m_{00}$. Even if the model is completely correct, sampling noise would make it almost certain that this equation will not hold. — Or, again, we should be able to estimate β_1 by $m_{10} - m_{00}$. But we should equally be able to estimate it by $m_{11} - m_{01}$, and sampling noise will guarantee that $m_{10} - m_{00} \neq m_{11} - m_{01}$, even if the model is correct.

[8] Actually, the problem of more equations than unknowns is much older, going back to the earliest mathematical models in astronomy, in ancient Mesopotamia thousands of years ago, which predict the motion of planets through the sky based on parameters about those planets. If the model for the motion of Mars has (say) seven parameters, then once we have more than seven observations of Mars, we've got more equations than unknowns. The ancient practice was to try to pick the best observations, so that there were just as many equations as unknowns, and then solve for the parameters. This of course threw away the information in the other observations, though you might check the solution by seeing whether those parameters could predict other observations. This was pretty much the practice everywhere there mathematical astronomers from ancient Babylonia down through Galileo and Kepler in the 1600s, and beyond. (On selecting observations, see Farebrother (1999).)

Two-way ANOVA without interactions and with more-than-binary categories

Just like two-way ANOVA with binary categories, but with more dummy or indicator variables.

A worked example

Here's what I get from the analysis of variance if I combine the four-fold racial categorization (Asian, black, white, other) with the two-fold educational categorization (bachelor's degree vs. less than a bachelor's degree):

	Estimate	Std. Error
(Intercept)	90500	2700
RACENAMEBlack	-42000	3100
RACENAMENBWA	-31200	4400
RACENAMEWhite	-17200	2700
COLLEGETRUE	66100	1100

The reference level for race is still "Asian" (because R likes to go by alphabetical order), while the reference level for education (`COLLEGE`) is non-college-educated (because, in R, `FALSE < TRUE`). So the intercept here is an estimate of the mean household income for a non-college-educated Asian. The other race contrasts are what would be added to that for other races (or, rather, subtracted), while the educational contrast is what would be added to mean household income for the college-educated. Thus the mean household income for a college-educated black person would be

$$(9.05 \times 10^4) + (-4.2 \times 10^4) + (6.61 \times 10^4) = 1.15 \times 10^5$$

In this way, we could fill out a full 2 × 4 table of expected household income for all the different combinations of race and education. (In fact, that was part of the after-class exercise.)

The R^2 of a model like this is the ratio of between-group variance to total variance. Here, R^2 is 0.0896, i.e., about 9 of the variance in income is, in this model, predicted by the combination of your race and whether or not you're college educated. That's not *nothing*, but it's not a lot, and the reason it's not a lot is that all of these groups have very, very wide income distributions (as we've seen).

It's worth pausing a moment to notice that the educational contrast is *very* large — larger than any of the racial contrasts. (Relatedly, the R^2 for this model isn't much larger than the R^2 for the one-way ANOVA which *just* looked at education.) This is a fairly recent development. If we were to re-do this with data from sixty years ago, in 1960, the educational contrast would be smaller, the racial contrasts relatively larger (and Asians would almost certainly not be the highest-mean-income racial category). Going back to 1900, if we could find the data, the contrast between those with and without college degrees would be even smaller, and the racial contrasts even more stark.

Two-way ANOVA with "interactions"

There's nothing which says we *have* to use an additive model when we've got more than two classifying variables. We could, for instance, use a model like this, for two binary classes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i} + \epsilon_i$$

In the jargon, we say that we **interact** X_1 and X_2 in this model. Mathematically, you can convince yourself that we now have as many equations as unknowns, so we can go back to estimating the βs by just doing some algebra on the means:

$$
\begin{aligned}
\mathbb{E}\left[Y|X_1=0,X_2=0\right] &= \beta_0 && (7)\\
\mathbb{E}\left[Y|X_1=1,X_2=0\right] &= \beta_0+\beta_1 && (8)\\
\mathbb{E}\left[Y|X_1=0,X_2=1\right] &= \beta_0+\beta_2 && (9)\\
\mathbb{E}\left[Y|X_1=1,X_2=1\right] &= \beta_0+\beta_1+\beta_2+\beta_{12} && (10)
\end{aligned}
$$

$$
\begin{aligned}
\beta_0 &= \mathbb{E}\left[Y|X_1=0,X_2=0\right] && (11)\\
\beta_1 &= \mathbb{E}\left[Y|X_1=1,X_2=0\right]-\mathbb{E}\left[Y|X_1=0,X_2=0\right] && (12)\\
\beta_2 &= \mathbb{E}\left[Y|X_1=0,X_2=1\right]-\mathbb{E}\left[Y|X_1=0,X_2=0\right] && (13)\\
\beta_{12} &= \mathbb{E}\left[Y|X_1=1,X_2=1\right]+\mathbb{E}\left[Y|X_1=0,X_2=0\right]-\mathbb{E}\left[Y|X_1=1,X_2=0\right] && (14)\\
&\quad -\mathbb{E}\left[Y|X_1=0,X_2=1\right]
\end{aligned}
$$

Logically, what we've just done is equivalent to a one-way ANOVA on the *combinations* of the two variables.

Substantively or statistically, we're allowing the contrasts between levels of one variable to be modified by the level of another variable. In this model, it no longer makes sense to talk about *the* contrast between $X_1 = 0$ and $X_1 = 1$, we have to say whether $X_2 = 0$ or $X_2 = 1$:

$$
\begin{aligned}
\mathbb{E}\left[Y|X_1=1,X_2=0\right]-\mathbb{E}\left[Y|X_1=0,X_2=0\right] &= \beta_1 && (15)\\
\mathbb{E}\left[Y|X_1=1,X_2=1\right]-\mathbb{E}\left[Y|X_1=0,X_2=1\right] &= \beta_1+\beta_{12} && (16)
\end{aligned}
$$

Symmetrically, contrasts between different levels of X_2 depend on what value X_1 has:

$$
\begin{aligned}
\mathbb{E}\left[Y|X_1=0,X_2=1\right]-\mathbb{E}\left[Y|X_1=0,X_2=0\right] &= \beta_2 && (17)\\
\mathbb{E}\left[Y|X_1=1,X_2=1\right]-\mathbb{E}\left[Y|X_1=1,X_2=0\right] &= \beta_2+\beta_{12} && (18)
\end{aligned}
$$

Again, this model, with interactions, is just a matter of book-keeping on the group means, and so can't be *wrong*. You might ask why we ever bother with analysis of variance models without interactions. Let's look at a little example before answering that question.

A worked example (continued)

	Estimate	Std. Error
(Intercept)	75700	4400
RACENAMEBlack	-25100	4800
RACENAMENBWA	-11700	6100
RACENAMEWhite	-2500	4500
COLLEGETRUE	88700	5500
RACENAMEBlack:COLLEGETRUE	-30100	6600
RACENAMENBWA:COLLEGETRUE	-38800	9400
RACENAMEWhite:COLLEGETRUE	-22500	5600

The way you read this is that everyone starts with the intercept as a base, and then (say) a black college graduate would add in the "black" term, then the "college" term, and then finally the "black:college" term. The fact that the "black:college" term is negative, but smaller than the college term, indicates that college education predicts a *smaller* increase in income for blacks than it does for members of the reference category (namely Asians). The fact that the sum of the college term and the black:college term is still positive indicates that college-educated blacks do have a higher average income than non-college-educated blacks[9].

The R^2 for the model with interactions is 0.0901. Again, even with this more complicated model, there's still a *lot* of variance within each of the compound categories, and just knowing someone's race and education leaves a lot of uncertainty about their income.

You will notice that the R^2 for this model is just a little higher than the R^2 for the model without interactions. Evaluating the predictive performance of models on the same data that was used to fit them is a dubious (though common) practice — it pretty much guarantees that the larger, more complicated model will be declared the winner[10]. It's noticeable, however, that here the increase in R^2 is very, very small, but the model is rather more complicated, even with just two variables and a fairly small number of categories. This illustrates the situation where people *should* prefer to use models without interactions — they're lot easier to interpret, and in this case, there's little or no penalty in accuracy[11].

Analysis of variance for log income

Everything above was done for income, and you might reasonably wonder if the very low R^2 was due to the skewed, heavy-tailed distribution of income that we've seen so often earlier. If we re-do this for *log* income, though, which we know is roughly Gaussianly-distributed, things are not actually much different[12]:

	Estimate	Std. Error
(Intercept)	10.90	0.03
RACENAMEBlack	-0.46	0.03
RACENAMENBWA	-0.23	0.04
RACENAMEWhite	-0.11	0.03
COLLEGETRUE	0.67	0.01

The way you would read this is that being college educated adds 0.67 to log income, which is to say it multiplies income by $e^{0.67} = 1.96$, i.e., it almost doubles income. This model accounts for 9.71% of the variance in log income, because, again, even on a log scale, there is a *lot* of inequality within each of these categories[13].

[9] You *could* interpret this as "blacks see less benefit for completing college than Asians do", or even "society rewards blacks less for education attainment than it does Asians", but some cautions are in order. First, we're applying a very crude educational threshold, of "finished a bachelor's degree" versus "didn't finish a bachelor's degree". It's quite possible that college-educated Asians are more educated than college-educated blacks, and/or that college-educated Asians studied subjects that prepared them for higher-income careers, and/or that they attended more impressive, and perhaps also genuinely better, schools. (All of those differences might, in turn, point to other sources of social injustice, but they'd be different kinds of injustices than people with equally-good educations getting different rewards for their skills.) Second, this is *household* income, not individual income, and there are differences in household structure; if the returns to education are the same across racial groups, but college-educated Asians are more likely to have a college-educated partner than are college-educated blacks, we'd expect to see a pattern like this. Third, many other factors go into income (e.g., the local cost of living), and those are not necessarily balanced between the two groups. (Asians disproportionately live in larger cities, with higher costs of living, while blacks are [still] disproportionately rural, though I am not sure if that's true of college-educated blacks.) I could go on, but this will, I hope, start to suggest some of the need for adjustments or controls when it comes to making fair comparisons, a topic we'll turn it in the subsequent lectures.

[10] A better approach would be a **hold-out**: Pick, say, 90% of the data at random, use it to estimate both models, and see which one better predicts the other 10% of the data. We will return to this notion of hold-out comparisons, and the related idea of **cross-validation**, later in the course. Social scientists who study inequality don't use crsso-validation very often, but it's become standard practice in statistics since the 1970s (Stone 1974; Geisser 1975; Geisser and Eddy 1979; Wahba 1990), and it transformed machine learning when the computer scientists borrowed it from the statisticans in the 1990s. Eventually, I'm sure, social science will catch up.

[11] It's also true that the standard errors for the coefficients are generally smaller when the model doesn't have any interactions, and people sometimes regard this as an advantage. But precisely answering the wrong questions doesn't help us learn about the world.

[12] I am dropping the small proportion of records (1.1%) with 0 income. If I replace those values with some trivial positive amount, say one cent, it doesn't change things very much.

[13] If I allow full interactions for log income, R^2 only goes up to 9.79%.

Changes in the global distribution of income in recent decades, and between- versus within- country inequality

We've been looking at the components of inequality within a single country (the US) in a single year (2020, so help us). Conceptually, at least, there's nothing stopping us from looking at how inequality for the whole human population, across the world, decomposes into a combination of income inequality *within* countries, plus inequality in typical incomes *between* countries[14]. Gathering the necessary data, and making it comparable over time and across countries, has been a huge undertaking, and sometimes involves a certain amount of approximation, interpolation and sheer guess-work. Nonetheless, the main outlines now appear reasonably clear[15].

Very roughly, then, the situation was as follows.

Before, roughly 1700, inequality in every civilized society was about as high as was compatible with keeping most of the population alive. Moreover, average income levels were quite similar across civilized societies. Obviously some people were richer than others — it was better to be the king of England, or the emperor of India or China, than one of their peasants — but most people were peasants, and there wasn't much to choose between being an English or an Indian or a Chinese peasant. Within-country inequality was high, and between-country inequality low.

Between 1700 and 1800, there was a slow rise in average incomes in western Europe and north America, as compared to the rest of the world. This meant that between-country inequality began to rise and become important for the first time. But even in 1800, the average person in England or America wasn't *much* better off than the average person in India or China[16].

Between 1800 and (say) 1950, there was a steady rise in between-country inequality, as western Europe, north America, and later the Pacific Rim (Japan, Australia, New Zealand) became *much* richer than any previous societies. This was mostly about the industrial revolution and its consequences; here, what matters is that it lead to a huge increase in between-country inequality. By 1950, the material standard of living of the average person in England or America was certainly *much* higher than the average standard of living in most of the rest of the world.

Within the industrialized countries, inequality tended to rise from the beginning of industrialization until roughly the beginning of the First World War in 1914. It then either fell or stayed low for several decades, say until roughly 1980, or even 1990 in some countries. The pattern was very different in the non-industrialized countries. These tended to become even more unequal that the industrialized countries, and to *not* have the suppression of inequality in the middle of the 20th century, for a number of reasons[17]. The exception were the countries which were taken over by Communist governments, which did, indeed, suppress income inequality[18].

[14] For annoying technical reasons, if we measure inequality using the Gini index rather than variance, as is usually done, we don't get an exact decomposition analogous to $\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathrm{Var}[Y|X]]$. (There is no "law of total Gini index".) But we *can* look at trends in the world-wide or "global" Gini index, the Gini we'd compute if we said everyone in each country had the typical income for that country, and the population-weighted average of individual countries' Gini indices.

[15] In what follows, I am synthesizing a bunch of different studies. One of the leading figures in the effort to make this kind of comparison possible is the economist Branko Milanovic, who has written a number of books on the subject; of these, I most strongly recommend Milanovic (2016), but they're all good. Prof. Milanovic also writes a thought-proviking weblog which often touches on these matters, [http://glineq.blogspot.com/].

[16] Pomeranz (2000) argues that as late as 1800, the richest *parts* of China were as prosperous as the richest parts of western Europe. (He suggests that India and Japan really were comparatively behind.) There is some argument about this among economic historians, but it's a serious case by a well-regarded scholar.

[17] In the first place, almost everywhere outside of the industrialized countries was colonized by one or another of the industrialized countries. While it is genuinely unclear whether imperialism was a *net* benefit to the imperialist *countries*, it was certainly very beneficial to the individual *imperialists*, who set up economic institutions which funneled money and resources towards them, and lots of those insitutions stayed in place after the imperialists left. (South Africa is a classic case.) Second, competition from industrial products tended to crush artisan and hand-craft production, creating an economic system lop-sided towards poor peasants, miners, etc., and rich land-owners, merchants, etc. (Imperialists certainly helped this along where they could, as with the British attacks on textile manufacturing in India, but it also happened even where imperialists *didn't* do this, and it's exactly what economic theory predicts (Fujita, Krugman, and Venables 1999).)

[18] Of course, those governments also killed tens of millions of people by inducing famines, and "not being able to get enough to

So, to sum up, the whole period from, roughly, 1700 to the late 20th century — say 1950, say 1980 — was one of *increasing* between-country inequality. The latter part of it, in the 20th century, saw *decreasing* within-country inequality in the rich, industrialized countries, and in the Communist non-industrialized countries.

The big change that has occurred over my lifetime, say since 1980, is a rise in within-country inequality, and the reduction in between-country inequality. Simply put, economic growth has been very rapid, and widely shared, in a lot of countries in Asia (and to a lesser degree elsewhere). Economic growth in India and China, in particular, has been extremely rapid and has done a *lot* to reduce between-country inequality, because the already-rich countries have grown more slowly[19]. Along side this, within-country inequality has tended to go up in most countries, both in the rich democracies and in India and China. So we have seen a rise in within-country inequality in most countries, and a decline in between-country inequality.

It's important not to exaggerate this. The average income level in the US, western Europe, Japan, etc., is still much higher than the average income level in China or India, and that will continue to be the case for many decades to come. But the gap has narrowed, and for between-country inequality to have become *less* important over time is a reversal of a centuries-old trend.

Q-Q plots

One common visual device for comparing two distributions is a **quantile-quantile** plot, or **Q-Q plot**. (You may have encountered these in other courses already.) We make these by plotting the quantiles of group 0 on one axis against the same quantiles of group 1 on the other axis. In symbols: say $Q_0(p)$ is the p quantile of group 0, $\mathbb{P}(Y \leq Q_0(p)) = p$, and likewise for $Q_1(p)$, $\mathbb{P}(Y \leq Q_1(p)) = p$. We make up a sequence of probabilities between 0 and 1, say $0 < p_1 < p_2 < \ldots < p_m < 1$, and for each p_i, we plot the point $(Q_0(p_i), Q_0(p_i))$ until we have m points for our m quantiles in the plot. In R, the function `qqplot()` conveniently does this, including making up a sensible sequence of p_is.

We can use a Q-Q plot to compare two theoretical distributions to each other (if we can calculate their quantiles functions). If we have two samples, we use the sample quantiles.

There are a few key things to notice about Q-Q plots.

1. *Equal distributions go up the main diagonal or 45-degree line.* If two distributions are exactly the same, their quantiles will be exactly the same, and so $Q_0(p) = Q_1(p)$ for all p.
2. *Two samples from the same distribution should fall near the main diagonal.* If our two groups are really just two samples from the same distribution, then $Q_0(p) \approx Q_1(p)$, because they differ only by sampling noise[20].
3. *If the Q-Q plot from two samples falls near the main diagonal, the distributions are close.* I will omit the argument here, since this is a question on HW3.
4. *Q-Q plots tell us when one variable is "probably larger" than another, or "stochastically dominates" another.*

The last point deserves a little elaboration.

We say that one random variable X **stochastically dominates** another random variable Y when the *complementary* CDF of X is bigger than the CDF of Y, $\mathbb{P}(X \geq a) \geq \mathbb{P}(Y \geq a)$ for all levels a, and

eat so you starve to death" can be seen as a very extreme form of income inequality indeed. So perhaps we should say that Communist governments suppressed the right tail of the income distribution, while amplifying the left.

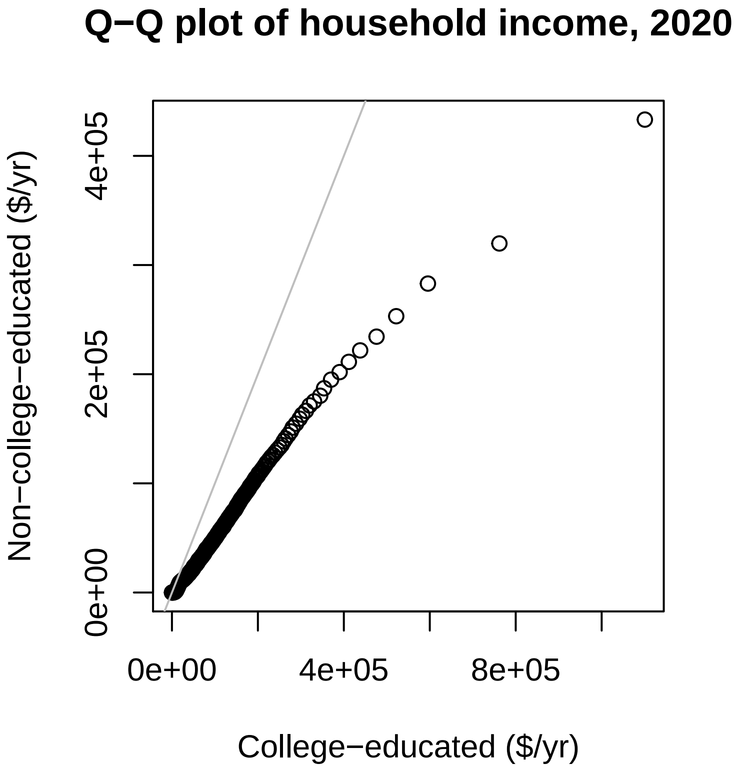
[19] To some extent, this is exactly what economic theory would predict. If a country starts out poor, but gains access to more productive technology, it should be able to grow rapidly by adopting that technology. Once a country is already rich and "at the technology frontier", it should only grow as fast as technology itself improves (Solow 1970). This sort of basic growth theory does not, however, explain why it took so long for economic growth to *start* in so many countries, which is why development economics is a separate and difficult field.

[20] There are actually ways of calculating *how far* from the diagonal two samples from the saame distribution should fall, based on the sample sizes, related to the Kolmogorov-Smirnov test for identity of distribution. This is sometimes useful but not so important for our purposes.

$\mathbb{P}(X \geq a) > \mathbb{P}(Y \geq a)$ for at least some a. Because the quantile function is the inverse of the CDF, an equivalent statement is that $Q_X(p) \geq Q_Y(p)$ for all p, and $Q_X(p) > Q_Y(p)$ for at least some p. This is precisely what we see when the Q-Q plot is either entirely above or entirely below the diagonal (depending on which variable we put on the horizontal axis). We sometimes write this as $X \succ Y$. This doesn't mean that *every* value drawn from X will be larger than *every* value drawn from Y, but it does mean that if you have to bet on which is going to be larger, you should bet on X.

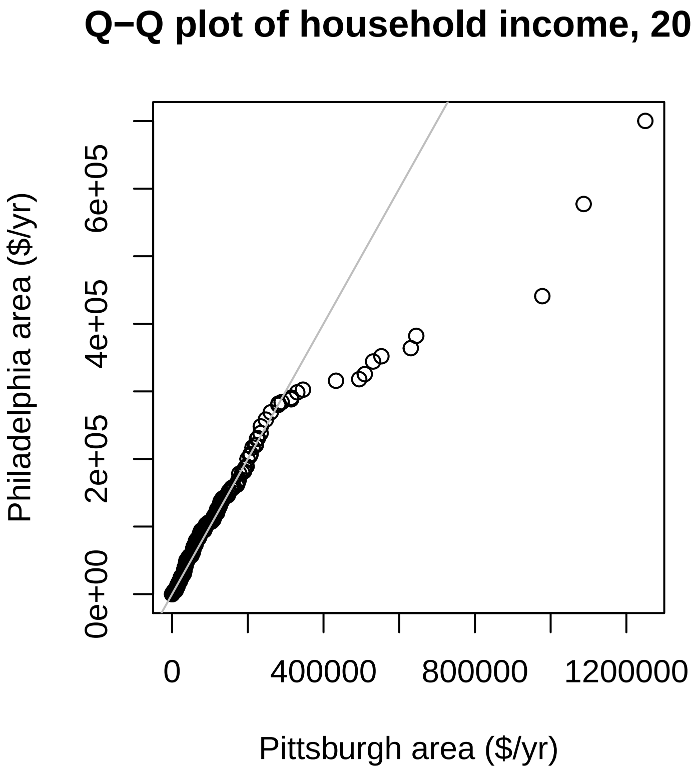
In terms of a Q-Q plot, $X \succ Y$ will show up in every point being below (or at) the main diagonal — $Q_0(p) \geq Q_1(p)$ for all p.

Having said all this, let's look at some examples[21].



We see that the college-educated income distribution stochastic dominates the non-college-educated distribution. This should not be a surprise, especially not considering what we saw from analyses of variance. But those analyses were compatible with a situation where (say) the higher mean income of the college-educated is due to a small number of people who have *really* high incomes. This Q-Q plot confirms that that's not the case, the whole distribution is higher for the college-educated.

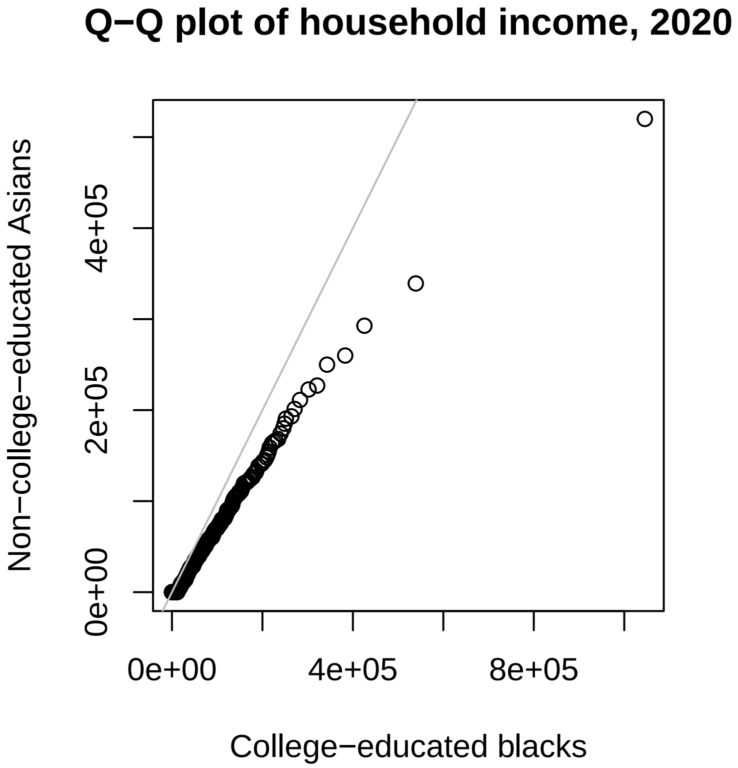
[21] The basic `qqplot()` function in R isn't designed to accommodate sampling weights, so you'll see in the code that I've written my own replacement which does. It doesn't make a *lot* of difference, but it's a little bit sounder statistically, though my version is definitely uglier. I'd suggest sticking to `qqplot()` in the homework.

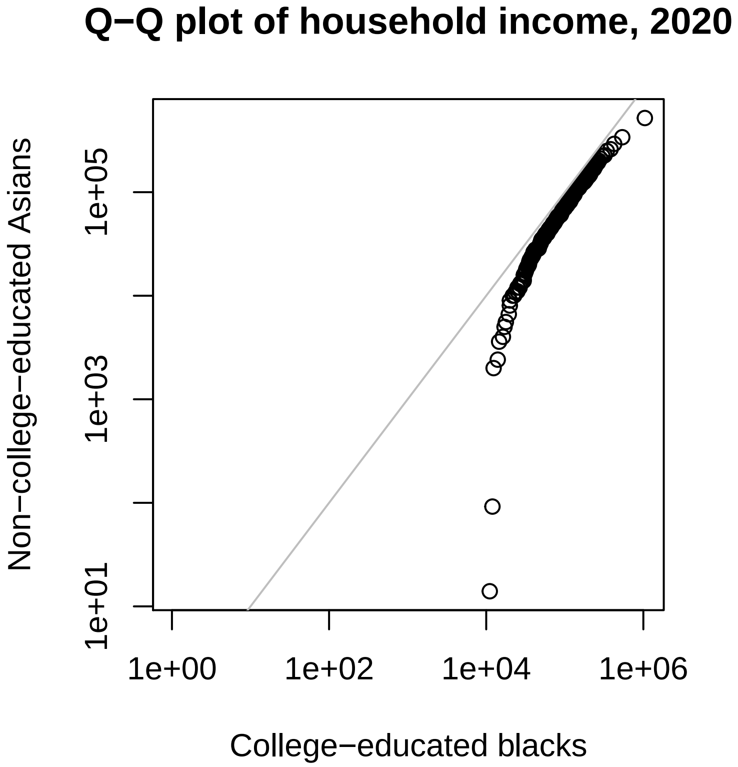


I have been hammering away at racial and educational comparisons because they're socially important, they arouse a lot of emotional heat, and they're easy to grasp, but I want to emphasize that there's nothing *statistically* special about them. One of the variables in the data set, `METAREA`, records which "metropolitan statistical area"[22] each respondent lives in, so we can compare (for instance) the income distribution for the Pittsburgh area versus the Philadelphia area. We can see that they're actually extremely similar, though Pittsburgh has something of an advantage (?) at the very highest quantiles (and, if you zoom in, a less striking advantage at the very lowest).

As a final example of Q-Q plots, let's compare the distribution of college-educated blacks to non-college-educated Asians. This is of some interest because we saw earlier that while Asians have the highest mean income, and blacks have the lowest mean income (among the groups we're considering), college-educated blacks have a higher mean income than non-college-educated Asians. Again, is this due to a small number of high-income, highly-educated black people, or does it apply more broadly across the distribution?

[22] These are essentially cities, or linked clusters of cities, *plus* their suburbs and environs. They're defined by studying commuting patterns, to try to ensure that each MSA follows actual patterns of trade and contact, and not just legal boundaries. So the Philadelphia MSA, while focused in southeastern Pennsylvania, actually includes parts of the state of New Jersey and Delaware, etc.





(The second plot is the same as the first, but with a log scale on both axes.)

What these plots confirm is that the income distribution for college-educated blacks stochastically dominates the income distribution for non-college-educated Asians, though there's a range where the two sets of quantiles nearly line up.

Relative distribution

Q-Q plots are a traditional visual way of comparing distributions, and ANOVA is a traditional quantitative way, but limited to expectation values, rather than the whole distribution. A very striking idea which gives us a quantitative way of comparing whole distributions is the **relative distribution method** due to Handcock and Morris (1998), which I like too much to leave un-explored here.

Suppose we've got two samples from two populations, say $y_1, y_2, \ldots y_{n_1}$ from population 1 and $z_1, z_2, \ldots z_{n_2}$ from population 2. What would happen if these distributions were really the same? Well, one thing we could do is this: find the CDF for population 1, say F_1. Now, for each z_i, we calculate the **relative value** $r_i = F_1(z_i)$. This says where z_i would fall in the distribution of population 1. More exactly, it tells us at what *quantile* of population 1 we'd find z_i.

The reason this is interesting is that if the two populations had the *same* distribution, then the r_i should look like a sample from a uniform distribution on the unit interval, $\text{Unif}(0, 1)$. After all, if the distributions are the same, 50% of the z_is should be $\leq$ the population 1 median, 75% should be $\leq$ the population 1 3rd quartile, and so forth. So the relative values should be uniformly distributed[23] if the two distributions are equal. But, conversely, if the two distributions are *not* equal, there should be an excess of relative values in some parts of the unit interval, and a corresponding deficit of r_i values in other places.

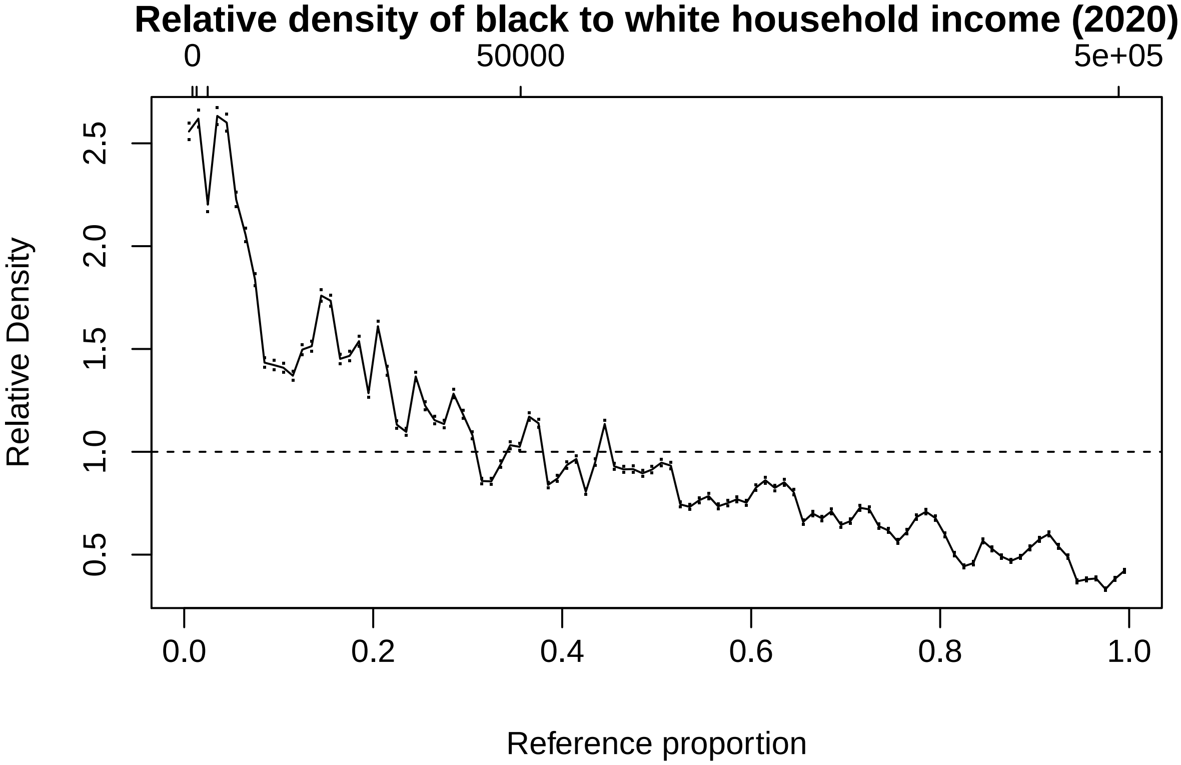
So, fact 1: the relative values r_i are uniformly distributed if, and only if, the two distributions are the same.

Fact 2: because the relative values are all in $[0, 1]$, it's very straightforward to estimate their CDF, and even to estimate their probability density. There are lots of tools for non-parametric density estimation which we could apply here. (If all else fails, there's always the histogram.) This distribution is called the **relative distribution**. Let's call the pdf of the relative distribution, a.k.a. the relative density, g.

If $g(r) = 1$, then values close $Q_Y(r)$ are equally common whether we're considering population 1 or population 2. If $g(r) > 1$, then values to $Q_Y(r)$ are *more* common in population 2 than in population 1. If $g(r) < 1$, then population 2 is relatively depleted around $Q_Y(r)$. If $g(r) = 1$ everywhere, then the two distributions are equal.

The `reldist` package (Handcock 2016) provides an extremely handy way of estimating the relative distribution from two samples. It also provides tools to *adjust* the relative distribution to account for covariates, something we'll return to in later lectures.

[23] Until further notice, when I say "uniformly distributed", I'll mean "uniformly distributed on $[0, 1]$".



Here, I've plotted black versus white household income, taking white as the reference class. The slightly-jagged black line is the estimate of the relative density g. (The little dots around it aren't printing errors, but confidence intervals — there's enough data here that the confidence intervals are actually very narrow!) The dashed horizontal line at 1 shows what we'd anticipate is the two distributions are equal. Unsurprisingly, the relative density is above one at low quantiles, and less than 1 at high quantiles. What may be somewhat surprising is that there's a range, say from about the 30th percentile to about the 60th (in the vicinity of \$50k/yr), where the relative density is actually pretty close to 1, and that even at the lowest is more like 0.5 than, say, 0.01. It is *also* absolutely true that blacks are, compared to whites, much more likely to be at the very bottom of the income scale.

(Also, notice from the upper horizontal axis that incomes here "only" go up to about half a million dollars a year — there are sufficiently few households above that that they don't show up much in this sort of survey of the general population.)

Appendix: Proof of law of total variance

The trick turns on the "law of total expectation", $\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|W]]$.

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 && (19)\\
&= \mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 && (20)\\
&= \mathbb{E}\left[\mathrm{Var}[Y|X] + (\mathbb{E}[Y|X])^2\right] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 && (21)\\
&= \mathbb{E}[\mathrm{Var}[Y|X]] + \mathbb{E}\left[(\mathbb{E}[Y|X])^2\right] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 && (22)
\end{aligned}$$

Now it'll be clearer if I introduce the symbol $A \equiv \mathbb{E}[Y|X]$. (Remember that X is a random variable, so

$\mathbb{E}[Y|X]$ is a random variable.) The last line becomes

$$\begin{aligned} \mathrm{Var}[Y] &= \mathbb{E}[\mathrm{Var}[Y|X]] + \mathbb{E}\left[A^2\right] - (\mathbb{E}[A])^2 & (23) \\ &= \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[A] & (24) \\ &= \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathbb{E}[Y|X]] & (25) \end{aligned}$$

Appendix: Least-squares estimation

We have n observations $y_1, \ldots y_n$, and, for each $i \in 1:n$, a vector of p **covariates** or **features**, say $x_{i1}, x_{i2}, \ldots x_{ip}$. The squared-error objective function is

$$L(b_0, b_1, \ldots b_p) = \sum_{i=1}^{n} (y_i - (b_0, + \sum_{j=1}^{p} b_j x_{ij}))^2$$

To solve this problem means to find the coefficients or parameters $(b_0, b_1, \ldots b_p)$ which minimizes the objective function L. To do so, it helps to first re-write L more compactly, using matrix algebra. Define $\mathbf{y}$ to be the $n \times 1$ matrix of the y_is, and $\mathbf{x}$ to be the $n \times (p+1)$ matrix whose first column is all 1s, and where the other columns contain the x_{ij}s. Finally, write $\mathbf{b}$ for the $(p+1) \times 1$ matrix of coefficients. Then

$$L(\mathbf{b}) = (\mathbf{y} - \mathbf{x}\mathbf{b})^T(\mathbf{y} - \mathbf{x}\mathbf{b})$$

(You can check that this really is the same objective function, starting with verfying that it's a 1×1 matrix, i.e., a scalar.)

Taking the derivative with respect to $\mathbf{b}$,

$$\frac{\partial L}{\partial \mathbf{b}} = 2\mathbf{x}^T(\mathbf{y} - \mathbf{x}\mathbf{b})$$

Setting this to zero at the optimum, $\hat{\beta}$, gives

$$\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}\hat{\beta}$$

Solving,

$$\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$$

When you run `lm()`, the command essentially just sets up and chugs through this piece of algebra.

(In practice, if p is large, inverting the $(p+1) \times (p+1)$ matrix can be more computationally time-consuming than solving for $\hat{\beta}$ directly from the previous equation, but that's an implementation detail.)

Historical aside

While the math above all is all very compact and straightforward now, it really was not when the Ancestors first worked out least squares around 1800. At the time, there was no notion of "vector", "matrix", "matrix multiplication" or "inverse of a matrix". All of those only developed gradually, over the course of the 19th century (in part because people realized least-squares problems were important and there needed to be a better way). You will find it a character-building exercise to set $p = 2$, explicitly differentiate $L(b_0, b_1, b_2)$ with respect to each b_i, set the derivatives to zero, and solve. The modern mathematical concepts and notations I used to find the solution in a few lines all developed gradually over the 1800s; part of the motivation for doing so was that people knew least-squares problems were important, and felt there had to be a better way (Farebrother 1999). I should add that even once were able to set up an equation like $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ in that form, inverting a $(p+1) \times (p+1)$ matrix by hand was not for the faint of heart, the careless, or the easily bored. If you read about ANOVA in an old-fashioned statistics textbook, or a research-methods text from another field, you will find a lot of formalism about setting out special tables, &c., &c., which was really about avoiding having to directly invert a matrix when all computing had to be done using pencil, paper, and the brains of the East African Plains Ape.

Least squares vs. one sample mean

Suppose we have $y_1, \ldots y_n$, and want to solve the following least-squares problem:

$$m = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - b)^2$$

This would be an intercept-only model. You can convince yourself that the solution is actually equal to the sample mean, $n^{-1} \sum_{i=1}^{n} y_i$. One way to do this would be to use the general matrix-algebra formula ($\mathbf{x}^T \mathbf{y}$ will be the sum of all the y_i, $\mathbf{x}^T \mathbf{x}$ will be n). Another would be to directly differentiate with respect to b and set the derivative to zero.

Least squares and mutually exclusive categories

If the X variables are a collection of mutually-exclusive indicators for groups, it seems like we should get something equivalent to the within-group sample means. This is in fact correct, but I will leave it as an exercise to show it.

Least squares and overlapping categories

If the Xs indicate category membership but they can overlap, perhaps because of cross-classification (like race × education), we won't just get out something equivalent to the sample means. Think back, above, to the set of equations I got for crossing two binary categories:

$$\begin{aligned}
\mathbb{E}\left[Y|X_1 = 0, X_2 = 0\right] &= \beta_0 & (26)\\
\mathbb{E}\left[Y|X_1 = 1, X_2 = 0\right] &= \beta_0 + \beta_1 & (27)\\
\mathbb{E}\left[Y|X_1 = 0, X_2 = 1\right] &= \beta_0 + \beta_2 & (28)\\
\mathbb{E}\left[Y|X_1 = 1, X_2 = 1\right] &= \beta_0 + \beta_1 + \beta_2 & (29)
\end{aligned}$$

As I said, if we just plug in the sample means on the left-hand side, we can't find a single set of three coefficients which satisfies all four equations. What least-squares will do in this situation is try to find the set of parameters which comes *closest* to satisfying the equations, in the sense of minimizing the squared difference between the two sides. But it will *weight* the equations differently, giving each equation a weight proportional to the number of data points where it applies.

$$
\begin{aligned}
& \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2}))^2 & (30)\\
&= \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{i:x_{i1}=u, x_{i2}=v}(y_i - (b_0 + b_1 u + b_2 v))^2 & \\
&= \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{i:x_{i1}=u, x_{i2}=v}(m_{uv} + (y_i - m_{uv}) - (b_0 + b_1 u + b_2 v))^2 & (31)\\
&= \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{i:x_{i1}=u, x_{i2}=v}(m_{uv} - (b_0 + b_1 x + b_2 v))^2 + (m_{uv} - y_i)^2 - 2(m_{uv} - y_i)(m_{uv} - (b_0 + b_1 x + b_2 v)) & \\
&= \sum_{u=0}^{1}\sum_{v=0}^{1} n_{uv}(m_{uv} - (b_0 + b_1 x + b_2 v))^2 & (32)\\
&\quad + \sum_{u=0}^{1}\sum_{v=0}^{1}\sum_{i:x_{i1}=u, x_{i2}=v}(m_{uv} - y_i)^2 & \\
&\quad -2\sum_{u=0}^{1}\sum_{v=0}^{1}(m_{uv} - (b_0 + b_1 x + b_2 v))\sum_{i:x_{i1}=u, x_{i2}=v}(m_{uv} - y_i) &
\end{aligned}
$$

where m_{uv} is the mean value of y when $X_1 = u$, $X_2 = v$, and n_{uv} is the number of such observations. Notice that the second term in the last equation doesn't involve the unknown coefficients b, so, for the purposes of minimizing the squared error, it doesn't matter and we can drop it. Furthermore, the inner-most sum in the third term, $\sum_{i:x_{i1}=u, x_{i2}=v}(m_{uv} - y_i)$, is always exactly zero (why?). So minimizing the least-squares objective function is equivalent to minimizing

$$\sum_{u=0}^{1}\sum_{v=0}^{1} n_{uv}(m_{uv} - (b_0 + b_1 x + b_2 v))^2$$

i.e., to doing least squares on the system of equations, but with a different weight, n_{uv}, on each equation, reflecting how often it applies.

— You can check that nothing in the line of argument above really relied on there being just two binary categories, or the categories being merely binary.

So does least squares converge, or what?

Suppose that both Y and the p-dimensional vector X are randomly generated from some (joint) probability distribution. We can then ask about the **expected squared error** of a linear model with intercept b_0 and p-dimensional slope vector b, which would be

$$\mathbb{E}\left[(Y - (b_0 + X \cdot b))^2\right]$$

We can ask what coefficients β_0 and β would minimize the expected squared error. You can check that they're as follows:

$$
\begin{aligned}
\beta_0 &= \mathbb{E}[Y] - \mathbb{E}[X]\beta & (33)\\
\beta &= (\mathrm{Var}[X])^{-1}\mathrm{Cov}[X, Y] & (34)
\end{aligned}
$$

Here $\mathrm{Var}[X]$ is the $p \times p$ matrix of the variances and covariances of the p coordinates of X, and similarly $\mathrm{Cov}[X, Y]$ is the $p \times 1$ matrix giving the covariance of each coordinate of X with Y. In deriving this, we do *not* have to assume that Y is actually linearly-related to X; this is just the best linear approximation to Y using the variables in X.

By the law of large numbers, $n^{-1}L(b_0, b) \rightarrow \mathbb{E}\left[(Y - (b_0 + Xb))^2\right]$. It should thus be very plausible that the least-squares value of $\hat{\beta}$ will converge on the population-optimal value β, and indeed it does. So if the linear model is right, least squares will converge (as $n \rightarrow \infty$) on the right coefficients[24]. If the linear model is wrong, least squares will converge on the best linear approximation.

Further reading

I have tried to say everything I think most important about analysis of variance, R^2, linear models, least-squares estimation, etc., etc. elsewhere (Shalizi 2015). (That document began as course notes for 36-401.) If you want to know more but couldn't stand reading more of me, I recommend Gelman and Hill (2006), and/or Berk (2004).

Analysis of variance goes back to Fisher (1918), though it could probably have been invented much earlier[25]. We will come back, later in the course, to the issue of why a *geneticist* was so interested in a statistical method for partitioning inequality.

On the relative distribution method, the best reference, by far, is Handcock and Morris (1999), though if you want something shorter which omits many details, read Handcock and Morris (1998).

References

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique.* Thousand Oaks, California: Sage.

Farebrother, Richard William. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900.* New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-0545-6.

Fisher, R. A. 1918. "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52:399–433. http://hdl.handle.net/2440/15097.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables. 1999. *The Spatial Economy: Cities, Regions, and International Trade.* Cambridge, Massachusetts: MIT Press.

Geisser, Seymour. 1975. "The Predictive Sample Reuse Method with Applications." *Journal of the American Statistical Association* 70:320–28. https://doi.org/10.1080/01621459.1975.10479865.

Geisser, Seymour, and William F. Eddy. 1979. "A Predictive Approach to Model Selection." *Journal of the American Statistical Association* 74:153–60. https://doi.org/10.1080/01621459.1979.10481632.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, England: Cambridge University Press.

Handcock, Mark S. 2016. Los Angeles, CA. https://CRAN.R-project.org/package=reldist.

Handcock, Mark S., and Martina Morris. 1998. "Relative Distribution Methods." *Sociological Methodology* 28:53–97. https://doi.org/10.1111/0081-1750.00042.

[24] Whether some other estimator than least squares might converge *faster* is another question. A classic result in linear-model theory, called the "Gauss-Markov theorem", essentially says that least squares is (basically) optimal when the noise variables ϵ_i are uncorrelated with each other and $\text{Var}\left[\epsilon|X = x\right] = \sigma^2$, a constant independent of x. If there's non-constant conditional variance and/or correlation in the noise, least squares is still consistent and even unbiased, but other estimators might converge more rapidly. See, e.g., Shalizi (2015), ch. 21.

[25] Milan (2013) is a historical romance novel in which part of the plot involves the hero inventing analysis of variance in 1867, and using it for statistical arbitrage on maritime insurance contracts. Another and larger part of the plot involves the heroine discovering what, in our timeline, we call Mendel's laws of genetics. Both moves are surprisingly plausible given the actual history of science. (I will not be answering questions about any of this.)

———. 1999. *Relative Distribution Methods in the Social Sciences.* Berlin: Springer-Verlag. https://doi.org/10.1007/b97852.

Milan, Courtney. 2013. *The Countess Conspiracy.* Vol. 3. The Brothers Sinister. Femtopress.

Milanovic, Branko. 2016. *Global Inequality: A New Approach for the Age of Globalization.* Cambridge, Massachusetts: Harvard University Press.

Pomeranz, Kenneth. 2000. *The Great Divergence: China, Europe, and the Making of the Modern World Economy.* Princeton, New Jersey: Princeton University Press. https://doi.org/10.2307/j.ctt7sv80.

Shalizi, Cosma Rohilla. 2015. "The Truth About Linear Regression." Online Manuscript. http:///www.stat.cmu.edu/~cshalizi/TALR.

Solow, Robert M. 1970. *Growth Theory: An Exposition.* Oxford: Oxford University Press.

Stone, M. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions." *Journal of the Royal Statistical Society B* 36:111–47. http://www.jstor.org/stable/2984809.

Sørensen, Aage B. 1998. "Theoretical Mechanisms and the Empirical Study of Social Processes." In *Social Mechanisms: An Analytical Approach to Social Theory*, edited by Peter Hedström and Richard Swedberg, 238–66. Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511663901.010.

Wahba, Grace. 1990. *Spline Models for Observational Data.* Philadelphia: Society for Industrial; Applied Mathematics.