Lecture Notes for Ec240a (Second Half), Fall 2016

Bryan S. Graham

10/18/2016

Chapter 1

Empirical relationships

1.1 Joint probability distributions

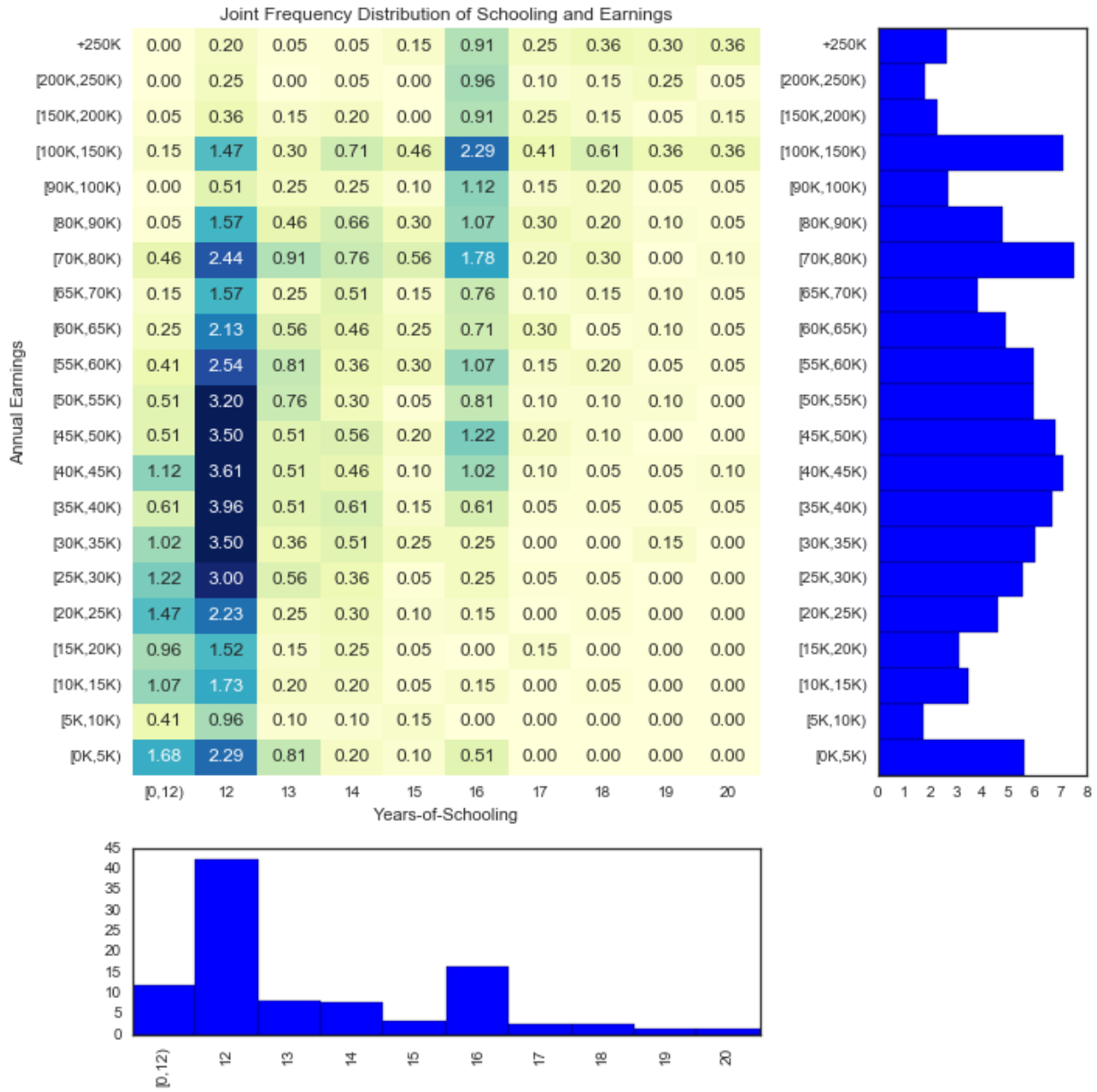
Figure 1.1.1 reports the joint frequency distribution of annual earnings and years of completed schooling across 1,969 white male respondents to the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY79 is a nationally representative sample of individuals born between the years 1957 and 1964 and resident in the United States in 1979. Annual earnings correspond to an average of earnings across the 1997, 1999, 2001 and 2003 calendar years. As a convenient shorthand, I will refer to this sample as a random one from the population of American white males aged 40 (in 2000).

On the left-hand side of the table are a total of $J = 21$ different earnings levels, $Y \in \mathbb{Y} = \{y_1, \ldots, y_J\}$ (reported in 2010 CPI-U prices). The first bin includes all respondents with earnings from \$0 to (less than) \$5,000 per year. The second those with earnings from \$5,000 to (less than) \$10,000 a year and so on. The higher income bins include a wider range of earnings levels, reflecting the fact that respondents in these bins are less numerous. The final bin includes all respondents with earnings greater than or equal to \$250,000 per year. In what follows we will proceed "as if" earnings is a discretely-valued random variable that takes values equal to the midpoints of the earnings bins listed in Figure 1.1.1 (we will discuss how to treat the top earnings bin below).

On the bottom of the table $K = 10$ different schooling levels $X \in \mathbb{X} = \{x_1, \ldots, x_K\}$ are listed. The first bin includes all respondents with less than 12 years of schooling. This includes all individuals who dropped out of high school (as well as a few who completed even less schooling). The highest schooling level corresponds to 20 years, which generally coincides with completing a Ph.D.

What can we learn about the relationship between schooling and earnings from this table? In total, there are $J \times K = 21 \times 10 = 210$ cells in the table. In the $(j, k)^{th}$ cell is the

Figure 1.1.1: Joint frequency distribution of annual earnings at age 40 and schooling among white males



SOURCE: National Longitudinal Survey of Youth (1979) and author's calculations. Figure rendered using Python (iPython Notebook `Ec240a_Python_Notebook_1.ipynb`)
NOTES: Joint frequency distribution of annual earnings and schooling in a sample of 1,969 white male NLSY79 respondents. Units from the military and poor non-Black, non-Hispanic subsamples are excluded. Earnings equals real average earnings across the 1997, 1999, 2001 and 2003 calendar years in 2010 prices (CPI-U Index). If earnings is missing for one or more years, the average is computed over non-missing observations. All units with complete information on years of schooling completed by age 28, earnings, and Armed Forces Qualification Test (AFQT) percentile score are included. The target subsample includes 2,439 respondents, 81 percent (i.e., 1,969) of which are complete cases. Frequencies are presented as percentages for readability.

percentage of respondents (out of the $N = 1,969$ in sample) with exactly $X = x_k$ years of schooling and $Y = y_j$ dollars of earnings. Let $i = 1, \ldots, N$ index the $N = 1,969$ respondents in our sample and $\mathbf{1}(\cdot)$ denote the indicator function. The $(j,k)^{th}$ cell in the table equals

$$\hat{p}(y_j, x_k) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(Y_i = y_j)\,\mathbf{1}(X_i = x_k). \tag{1.1.1}$$

For example, if we look at the $(j,k) = (1,1)$ cell we see that 1.68 percent of our sample has both less than 12 years of completed schooling and earnings between \$0 and (less than) \$5,000 a year (the numbers in the table coincide with (1.1.1) multiplied by 100 to convert them into percentages).

The NLSY79 was collected using a complex sampling methodology, but for pedagogical purposes we may imagine it is instead a simple random sample from the population of white males in the United States aged 40 in the year 2000. Specifically consider a list of all individuals in our target population. Imagine our sample of 1,969 NLSY79 respondents was selected by drawing names at random from this list. We allow for the possibility that an individual may be sampled twice (i.e., we sample with replacement), although in practice this is unlikely to occur. Now imagine a sequence of larger and larger simple random samples. The limit of (1.1.1)

$$\lim_{N\to\infty} \hat{p}(y_j, x_k) = p(y_j, x_k) = \Pr(Y = y_j, X = x_k). \tag{1.1.2}$$

gives the ex ante probability (of the event) that a random draw from the population of white males aged 40 in the year 2000 will have earnings $Y = y_j$ and schooling $X = x_k$. We call $p(y,x)$ for $y \in \mathbb{Y}$ and $x \in \mathbb{X}$ the **joint probability mass function (pmf)** of (Y,X). For all (y,x) with $y \notin \mathbb{Y}$ and/or $x \notin \mathbb{X}$ we set $p(y,x) = 0$. Note that our interpretation of probability is a frequentist one.

In practice $\hat{p}(y_j, x_k)$ maybe arbitrarily far from $p(y_j, x_k)$. For example, our sample might, by chance, include a large number of high earning dropouts. However, under weak conditions, which we will study in subsequent chapters, the limit (1.1.2) exists. For the remainder of this Chapter I will equate the frequencies reported in Figure 1.1.1 with their limits.

If we sum (1.1.2) over all possible j,k pairs we get

$$\sum_{j=1}^{J}\sum_{k=1}^{K} p(x_j, y_k) = 1,$$

since the sum of probabilities attached to a non-overlapping partition of the event space

must equal one.

We put some of these heuristics on a firmer footing with the following definition.

Definition 1.1.1. (PROBABILITY DISTRIBUTION) Let Ω denote a finite sample space. A function $p : \Omega \to \mathbb{R}$ is a **probability distribution function** if (i) $0 \leq p(a) \leq 1$ for all $a \in \Omega$ and (ii) $\sum_{a \in \Omega} p(a) = 1$.

A probability distribution function attaches a probability to each possible outcome of a random process. Definition 1.1.1 applies to finite sample spaces. The more careful arguments made in this book will generally involve only discrete probability. The extension of probability theory to infinite sample spaces is non-trivial and mathematically demanding (Billingsley (1995) is a standard reference). While we will encounter infinite sample spaces in subsequent chapters, I will generally avoid many of the technical details involved in a careful treatment of them.

In (1.1.2) our sample space Ω includes all $J \times K = 210$ possible configurations of earnings and schooling. As another example, consider a sequence of 3 die roles. For a single role there are six possible outcomes, and hence $6^3 = 216$ possible sequences of six die roles. The sample space here is $\{1, 2, 3, 4, 5, 6\}^3$. Since each sequence occurs with equal probability the probability distribution will assign probability $1/216$ to each. Our notion of probability is a frequency one: $p(a)$ gives the frequency of the event $a \in \Omega$ across many independent (i.e., an infinite number of) replications of our experiment. In (1.1.2) the experiment corresponds to taking a random draw from the population and measuring their earnings and years of completed schooling. In our die rolling example, the experiment involves rolling a fair die three times and recording the sequence of outcomes.

Sometimes we may be interested particular combinations of outcomes. Define an *event* to be a subset $A \subseteq \Omega$ of the sample space. For example we might be interested in the event that a random draw from the population of white males earns less than \$50,000 per year. The probability of this event is simply

$$p(A) = \sum_{a \in A} p(a),$$

where $a = (y, x)$ and $A = \{y : y < \$50,000\}$. This corresponds to a summation across all cells in the bottom ten rows of Figure 1.1.1. Performing this summation we get 50.59; over 50 percent of the white males aged 40 in the year 2000 earned less than \$50,000 per annum. A more refined event is observing that our draw earns less than \$50,000 and has not completed high school. Now $A = \{y, x : y < \$50,000,\ x < 12\}$. Now we sum the bottom ten elements of the left-most column in the table. This yields 10.1 percent.

As a different type of example, consider the event that the sum of our three die roles

exceeds 15 (this calculation may appeal to fans of Dungeons and Dragons). To calculate this probability we need to count the number of elements in the subset of die role sequences with sums greater than or equal to 15. This set includes any sequence that includes only 5s and 6s. The per die role probability of the event "five or six" is one third. Hence the probability of the event of all three roles consist of fives or sixes is $\left(\frac{1}{3}\right)^3 = \frac{1}{27}$. Any sequence of two sixes and either a three or a four also sums to at least 15. There are six such sequences, so that observing one of them occurs with probability 6/216. Finally any sequence consisting of a 4, 5 and a 6 also sums to 15. There are again six such sequences. Hence the probability of the event that three die roles sum to at least 10 is

$$\frac{1}{27} + \frac{6}{216} + \frac{6}{216} = \frac{22}{216} \simeq 0.102$$

or about 10 percent. It is common for discrete probability calculations to involve counting arguments.

The table in Figure 1.1.1 completely characterizes the joint distribution of schooling and earnings in the population of white males aged 40 in the year 2000. What can it tell us about the relationship between these two variables? The table is a bit bewildering, but with some work we can discern an important regularity: most of the cells in the lower left-hand and upper right-hand regions of the table have close to zero probability mass attached to them. We see few individuals with low levels of schooling, but high earnings and *vice versa.* Instead most of the probability mass is concentrated on the diagonal block of cells running from the upper left-hand to the lower right-hand portions of the table. This suggests that higher levels of schooling tend to be paired with higher levels of earnings in our population. This is an association and may or may not reflect any structural economic relationships between schooling and earnings.

Figure 1.1.1 characterizes a joint distributions of two random variables. What if we are just interested in the variation of schooling or earnings alone? The **marginal probability mass** attached to the event $X = x_j$ is

$$p\left(x_k\right) = \sum_{j=1}^{J} p\left(y_j, x_k\right). \tag{1.1.3}$$

Equation (1.1.3) computes the marginal probability of the event that a random draw from our population has exactly $X = x_k$ years of schooling. By 'marginal' we mean that we are only interested in the frequency distribution of X alone, independent of any possible relationship between X and Y . To calculate the fraction of workers with exactly $X = x_k$ years of schooling in the population we sum over the $j = 1, \ldots, J$ frequencies for earnings

and schooling combinations where schooling is held fixed at $X = x_k$. Looking at our table, we simply sum up all the elements in a given column. For example, the marginal probability of the event $X = 16$ (i.e., that a random draw has 16 years of schooling, corresponding to an undergraduate college degree) is

$$p(16) = p(x_6) = \sum_{j=1}^{J} p(y_j, x_6) = 0.0051 + 0.0000 + 0.00150 + \cdots 0.0096 + 0.0091 = 0.1656.$$

About seventeen percent of while males aged 40 in the year 2000 had completed 16 years of schooling. The histogram at the bottom of the figure provides a graphical depiction of the marginal schooling distribution. The vertical axis of the histogram gives the frequency of each schooling level in the population. Note the sizable masses of probability at 12 (0.4256) and 16 (0.1656) years of completed schooling, corresponding to high school and college graduation respectively.

1.2 Expectation

The expectation of X is simply the probability/frequency weighted average

$$\mathbb{E}[X] = \sum_{x \in \mathbb{X}} x p(x) = \sum_{k=1}^{K} x_k p(x_k). \tag{1.2.1}$$

Using (1.2.1) and Figure 1.1.1 we get

$$\mu_X = 0.1209 \times 11 + 0.4256 \times 12 + \cdots + 0.0142 \times 20 = 13.44.$$

The 'average' 40 year old white male in 2000 had completed 13.44 years of schooling. Notice how we used the marginal probabilities attached to each schooling level to weight the different years of schooling values when computing its expected value. For simplicity I treated all individuals with schooling less than 12 years "as if" they had completed 11 years of schooling.

Recall that for earnings we are proceeding 'as if' individuals earnings equal the midpoint of the bin in which they fall. However, to calculate average earnings it remains to assign a value to the top-most earnings bin. If we assume that high earnings levels follow a Pareto distribution (say above \$200,000), then

$$\Pr(Y \leq y | Y \geq 200,000) = 1 - \left(\frac{200,000}{y}\right)^{\alpha}$$

for some parameter α. We can calibrate this parameter by observing that

$$\Pr\left(Y > 250,000 | \, Y \geq 200,000\right) = \left(\frac{200,000}{250,000}\right)^{\alpha},$$

which, using the information in Figure 1.1.1 , gives

$$\alpha = \frac{\ln\left(\frac{2.64}{1.83+2.64}\right)}{\ln\left(\frac{200,000}{250,000}\right)} \simeq 4.$$

Now observe that, for $y \geq 250,000$, we have that

$$\Pr\left(Y \geq y | \, Y \geq 250,000\right) = \frac{\Pr\left(Y \geq y | \, Y \geq 200,000\right)}{\Pr\left(Y \geq 250,000 | \, Y \geq 200,000\right)} = \frac{\left(\frac{200,000}{y}\right)^{\alpha}}{\left(\frac{200,000}{250,000}\right)^{\alpha}} = \left(\frac{250,000}{y}\right)^{\alpha}.$$

This implies that the conditional distribution of Y above \$250,000 is also Pareto. This gives, by the properties of the Pareto distribution, an expected value of Y, conditional on the event that it exceeds \$250,000 of $\mathbb{E}\left[Y | \, Y \geq \$250,000\right] = \frac{\alpha}{\alpha-1}\$250,000 \approx \$350,000$. We will therefore proceed 'as if' all individuals earning more than \$250,000 earned exactly \$350,000. If above calculations are not entirely clear, then I suggest returning to them later in the course after we have covered more of the underlying concepts.

Using our definition of expectation we get

$$\mu_Y = 0.0559 \times \$2,500 + 0.0173 \times \$7,500 + \cdots + 0.0264 \times \$350,000 = \$65,159.$$

The expected value of earnings in our population of 40 year old White males is about \$65,000.

1.3 Inequalities

We will encounter a variety of probability inequalities in this course. Such inequalities have many uses. As an introduction consider the Markov Inequality (MI). This inequality gives us a probability for the event that a non-negative random variable Y is "far" from its expectation. Say we want to bound the event that Y exceeds some (large) value a. From our definition of the expected value we have that

$$\mathbb{E}\left[Y\right] = \sum_{y \in \mathbb{Y}} y p\left(y\right)$$

where $\mathbb{Y}$ denotes the sample space and y a (generic) element of that space. We can break the sample space into two parts $A = \{y \in \mathbb{Y},\, y \geq a\}$, and the relative complement of A in $\mathbb{Y}$.

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in A} yp(y) + \sum_{y \in \mathbb{Y} \setminus A} yp(y) \\ &\geq \sum_{y \in A} yp(y),\end{aligned}$$

where the second inequality follows if Y is non-negative (as we will assume). Next observe that for all $y \in A$ we have that $y \geq a$. This fact, along with the inequality above, gives

$$\begin{aligned}\mathbb{E}[Y] &\geq \sum_{y \in A} ap(y) \\ &= a\left[\sum_{y \in A} p(y)\right] \\ &= a \Pr(A) \\ &= a \Pr(Y \geq a).\end{aligned}$$

Re-arranging the last line we get.

Definition 1.3.1. (Markov Inequality) If Y is a non-negative random variable and $a > 0$, then

$$\Pr(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}.$$

A particularly useful special case of the Markov Inequality is Chebyshev's Inequality, which we will using to prove a Law of Large Numbers later in the course. Define the event $A = \{y \in \mathbb{Y},\, |y - \mu_Y| \geq a\}$ and observe that

$$\begin{aligned}\mathbb{E}\left[(Y - \mu_Y)^2\right] =& \sum_{y \in A} (Y - \mu_Y)^2 p(y) + \sum_{y \in \mathbb{Y} \setminus A} (Y - \mu_Y)^2 p(y) \\ \geq& \sum_{y \in A} (Y - \mu_Y)^2 p(y) \\ \geq& \sum_{y \in A} a^2 p(y) \\ =& a^2 \Pr(A)\end{aligned}$$

where the first inequality follows from non-negativity of $(Y - \mu_Y)^2$ and the second from the definition of A. Re-arranging we get:

Definition 1.3.2. (CHEBYSHEV'S INEQUALITY) For any random variable $Y \in \mathbb{Y} \subseteq \mathbb{R}^1$ and $a > 0$ any positive real number

$$\frac{\mathbb{V}(Y)}{a^2} \geq \Pr\left(|Y - \mu_Y| \geq a\right).$$

1.4 Conditional distributions

We can use our joint and marginal probability mass functions to compute the conditional probability of certain events. For example what is the conditional probability that worker earns $\$50,000$ per annum given that he has exactly 12 years of schooling? The **conditional probability mass function** is

$$p\left(y|\, x\right) = \begin{cases} \frac{p(x,y)}{p(x)} & \text{if } x \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases}. \tag{1.4.1}$$

Equation (1.4.1) gives the population fraction of individuals earning exactly $Y = y_k$ dollars per year conditional on them also having exactly $X = x_j$ years of schooling. The marginal frequency of workers with 12 years of schooling is 0.4256, the joint frequency of workers with 12 years of schooling and annual earnings of $[\$50,000, \$55,000)$ is 0.032. Therefore the conditional frequency of earning $[\$50,000, \$55,000)$ per year – where the conditioning is on schooling being equal to exactly 12 years – is $0.032/0.4256 = 0.0752$. Among those men with exactly 12 years of schooling, roughly 7.5 percent earn between $\$50,000$ and $\$55,000$ per year.

The conditional probability mass function for earnings given each of our 10 possible schooling levels is reported in Figure 1.4.1. An inspection of this table clarifies the relationship between earnings and schooling in our population. Consider the earnings distribution among high school dropouts (reported in column one of the table). Inspecting this column we see that the distribution of earnings for this subpopulation is concentrated at lower earnings levels. Over 83 percent of this group earns no more than $\$50,000$ a year. In contrast only about than 10 percent of workers with 20 years of schooling have earnings levels in this region. Overall, inspection of the conditional probability mass function reinforces the perception that higher earnings tend to be associated with higher levels of schooling.

The 'tends to' part of this statement is important. For each of the schooling levels the distribution of earnings attaches positive probability to nearly every possible earnings level. That is, among those with a common level of schooling, there will always be some very poor individuals and some very rich individuals. However, distributions which condition on high

levels of schooling tend to place *greater* probability mass on high earnings realizations.

1.5 Conditional expectation function (CEF)

We can use conditional probabilities as weights in order to calculate the expected value of earnings given a specific level of schooling. The **conditional expected value** of Y given $X = x_j$ is

$$\mathbb{E}\left[Y \middle| X = x_j\right] = \mu_{Y|X}\left(x_j\right) = \sum_{k=1}^{K} p\left(y_k \middle| x_j\right) y_k. \tag{1.5.1}$$

Using the conditional probabilities in our second table we get an average earnings level for drop outs of

$$\mu_{Y|X}\left(11\right) = 0.1387 \times \$2,500 + 0.0336 \times \$7,500 + \cdots + 0.000 \times \$350,000 = \$30,489.$$

The average 40 year White high school dropout earned approximately $\$30,000$ per year in 2000. The conditional expected value is simply the average earnings level within the subpopulation for which the conditioning criterion is true. From inspection of the conditional pmf we know that while some dropouts earn more, and others less, *on average* they earn $\$30,000$ per year.

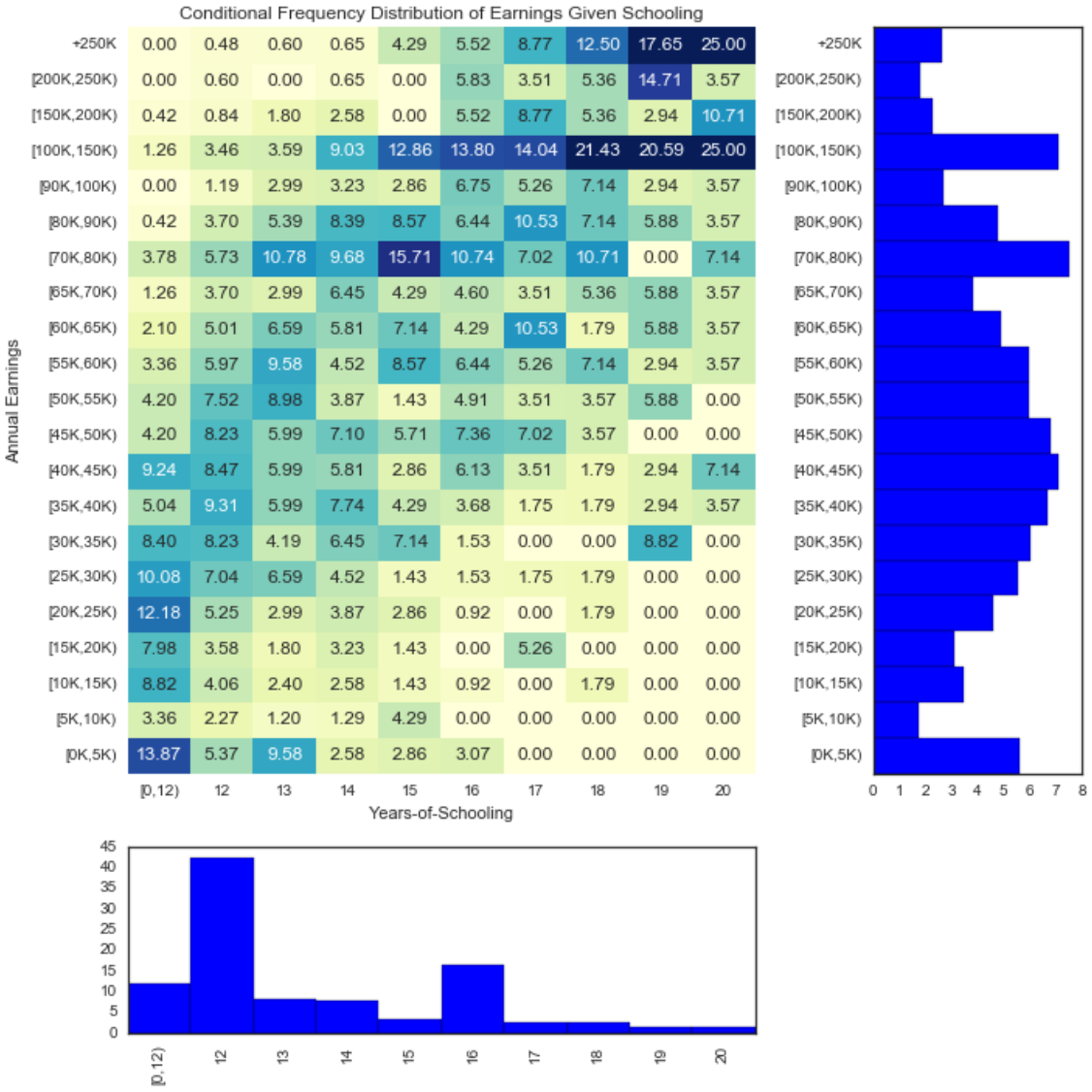
By varying the value of x_j in (1.5.1) we trace out the **conditional expectation function** (CEF) of Y given $X = x$. The CEF provides a conceptually simple summary of the relationship between $Y$and X in the population. Figure 1.5.1 computes the conditional mean of earnings for each of the $J = 10$ possible schooling levels. As is apparent from the figure average earnings are clearly increasing in years of completed schooling.

While the conditional expectation function is straitforward to visualize and interpret, this simplicity comes at a cost. Relative to the conditional pmf reported in our second table, Figure 1.5.1 contains less information about the variability of earnings conditional on any given schooling level. One way to summarize this variability is by the **conditional variance** of earnings given schooling.

$$\mathbb{V}\left(Y \middle| X = x_j\right) = \sigma^2_{Y|X}\left(x_j\right) = \sum_{k=1}^{K} \left(y_k - \mu_{Y|X}\left(x_j\right)\right)^2 p\left(y_k \middle| x_j\right). \tag{1.5.2}$$

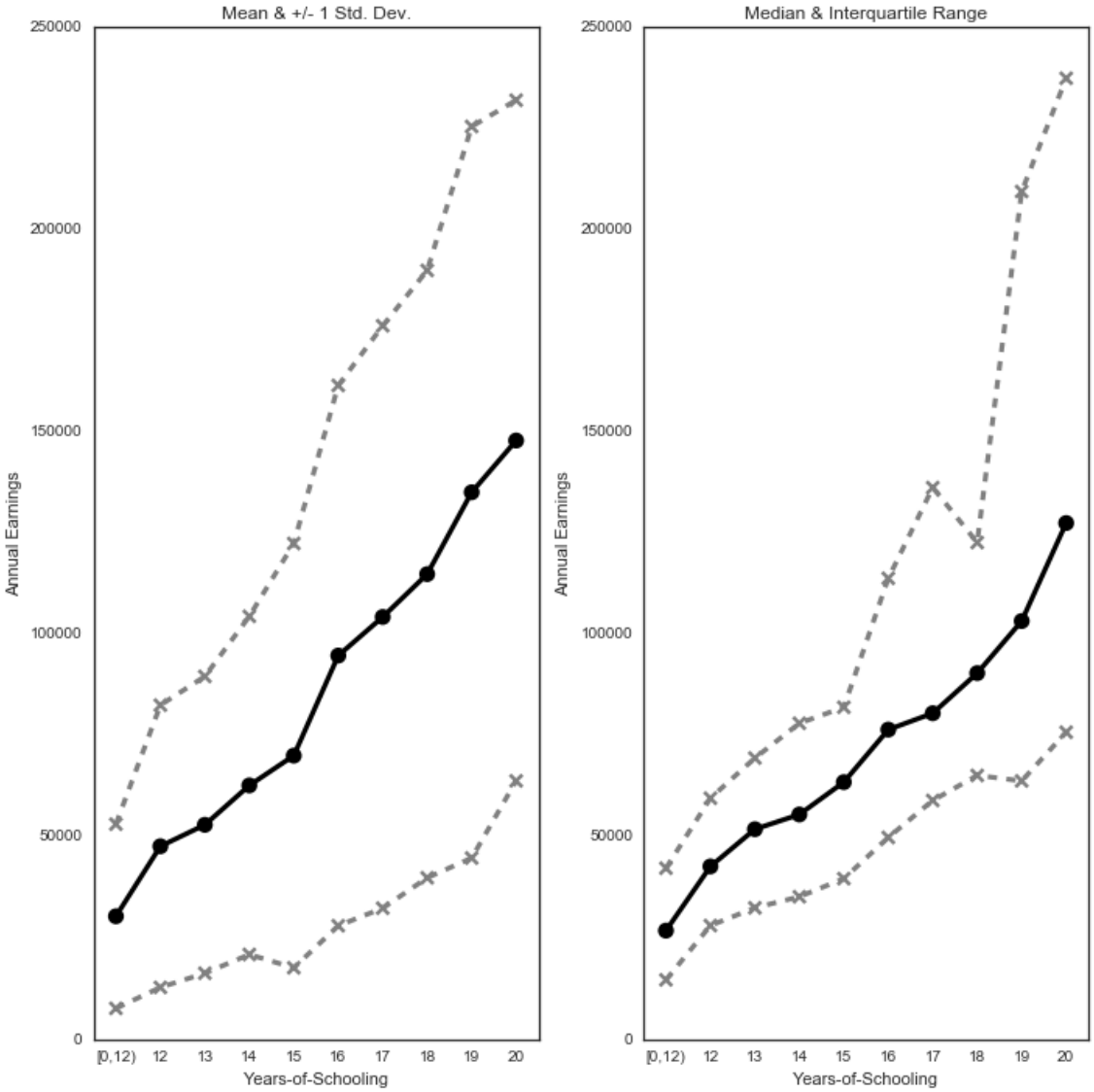
Equation (1.5.2) corresponds to the (conditional) average squared deviation of earnings about its (conditional) mean. If (1.5.2) is large then 'lots' of probability mass is attached to realizations of Y that are far from its average. The square root of variance corresponds to

Figure 1.4.1: Conditional frequency distribution of annual earnings at age 40 given schooling among white males



SOURCE: National Longitudinal Survey of Youth (1979) and author's calculations. Figure rendered using Python (iPython Notebook `Ec240a_Python_Notebook_1.ipynb`)
NOTES: See notes to Figure 1.1.1.

Figure 1.5.1: Expected and Median Annual Earnings by Years of Schooling



SOURCE: National Longitudinal Survey of Youth (1979) and author's calculations. Figure rendered using Python (iPython Notebook `Ec240a_Python_Notebook_1.ipynb`)
NOTES: See notes to Figure 1.1.1

the standard deviation. The dashed lines in Figure 1.5.1 corresponds to the CEF ± one conditional standard deviation.

1.6 Conditional quantile function (CQF)

An alternative way of summarizing the conditional relationship between earnings and schooling involves **quantiles**. Let $F_{Y|X}(y|X=x_j)$ denote the cumulative distribution function of Y given $X=x_j$:

$$F_{Y|X}(y|X=x_j)=\Pr(Y\leq y|X=x_j)=\sum_{k=1}^{K}\mathbf{1}(y_k\leq y)\,p(y_k|x_j),$$

where $\mathbf{1}(y_k\leq y)$ is an indicator function taking a value of 1 is $y_k\leq y$ and zero otherwise. For any $\tau\in(0,1)$ we call

$$Q_{Y|X}(\tau|X=x_j)=F^{-1}_{Y|X}(\tau|X=x_j)=\inf\left\{y:F_{Y|X}(y|X=x_j)\geq\tau\right\}$$

the τ^{th} conditional quantile of Y given $X=x_j$. By varying x_j we trace out the τ^{th} **conditional quantile function** (CQF). $Q_{Y|X}(\tau|X=x_j)$ equals the minimal earnings level for which *at least* $\tau\times100$ percent of the subpopulation homogenous in $X=x_j$ earns less. If Y were a continuously-valued random variable with a strictly increasing distribution function, then $Q_{Y|X}(\tau|X=x_j)$ would correspond to the earnings level for which exactly $\tau\times100$ percent of the population earn less. In the present example, where Y is (artificially) discretely-valued the interpretation is a bit 'coarser' (in that $Q_{Y|X}(\tau|X=x_j)$ may be constant in τ over some regions of the unit interval).

If we set $\tau=1/2$ we get the **conditional median function** (CMF) . The median earnings level in the subpopulation of dropouts (inspecting column one of our second table) is just $\$26,948$ per year. The right-hand panel of Figure 1.5.1 plots $Q_{Y|X}(1/2|X=x_j)$ for each of the 10 possible schooling levels. For most values of schooling the conditional median of earnings is less than the corresponding conditional mean. This is a frequent feature of earnings distributions, reflecting the presence of right-tail inequality.