ECON-C4100 - Capstone: Econometrics I

Lecture 2: Statistics recap

Otto Toivanen

Statistics recap

- In this lecture, we go through material that you have already been taught. Why bother?
  1. Repetition reinforces learning.
  2. You may have had this material taught in a way that did not resonate with you, so another try may be worthwhile.
  3. As you will find out,
     - "simple" descriptive statistics can be a powerful way of understanding and communicating what is going on.
     - a key insight is to understand how to relate the descriptive statistics to your research question.
     - estimation of the mean is a simple way of introducing the principle of **Least Squares** estimators.

Statistics recap

- What will we do? Go through
  1. univariate descriptive statistics and their use;
  2. conditional descriptive statistics and their use; and
  3. estimation of the mean and the concepts needed to accomplish this.

Learning outcomes

- By the end of the lecture, you know
  1. the reasons for, the importance of, and the steps of practical implementation of ensuring reproduceability of your analysis.
  2. the meaning of central concepts for descriptive statistics of a single variable,
  3. how to characterize the distribution of a variable, and
  4. the basics of how to link statistics to the social phenomenon of interest.

Documentation of your data, sources & analysis

- An extremely important and topical issue in science, social science and economics included, is the *replication crisis*.
- The replication crises refers to the (too) low fraction of studies that upon inspection can be replicated.
  - **Observational data:** Replication of an existing study using 1:1 the same data.
  - **Experiments:** Replication of an existing study by conducting the same experiment again.
- Scientific practice is continually improved to tackle this issue.
- As recently as in 2020, *Frances Arnold*, a Chemistry Nobel prize winner (2018), retracted her *Science* article because of problems in reproducing the results.

1. Documentation of your data, sources & analysis

- Some sources:
  - *Wikipedia.*
  - *Journal Science / economic experiments.*
  - An early investigation with bleak results: Dewald, W., Thursby, J. & Anderson, R. (1986). Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review, 76*(4), 587–603. Their goal was to replicate original studies using the original data.

Documentation of your data, sources & analysis

- Objective: document your empirical research in such a way that a knowledgeable researcher can
  1. go to your original data sources,
  2. recreate your analysis data,
  3. redo you analysis to reproduce your results, and
  4. critically evaluate the decisions you've made going from the raw data to the data used in the analysis and the choices you've made in the analysis.
- An example: in litigation involving empirical analysis, both sides are given access to each others' data and code for replication purposes.
- We strive to accustomize you to this standard right from the start.

2. Univariate descriptive statistics

- What are these?
  1. Ways to summarize the data.
  2. Ways to understand what the data "looks like".
  3. Ways to communicate to your audience "what is going on" in the data regarding the phenomenon you are studying.
- As a major side-benefit, in producing these you may spot anomalies in your data, such as
  1. missing data on some variables
  2. absurdly high values due to mistakes.
  3. logically impossible values (e.g. negative age).

Income distribution

- We use the income distribution to study descriptive statistics.
- The distribution of income has for long been of central interest to researchers and policy makers.

Income distribution

- Major reasons for economists to be interested:
  1. understand the causes of income inequality - efficient allocation of resources, or the result of various constraints?
  2. understand the effects of income inequality on resource allocation.
  3. to design policies that alleviate the inefficiencies.
  4. to design policies that reach stated distributional goals with minimal efficiency losses.

Income distribution

- In today's lecture, we study data relating to the Finnish income distribution.
- The objective is not to teach you facts about the Finnish income distribution, but to use it to reach our learning objectives.

Measuring income

- Q1: what income measure to use?
- Q2: whose income to measure?
- For the purposes of this lecture and next, we are going to use *earned income*, as defined by Statistics Finland.
- The data come from Statistic Finland: Teaching data.

Measuring income: Statistics Finland definition of Earned income

Earned income is the **sum of earned and entrepreneurial income** *received by households and income recipients during the year. The earned income concept of the income distribution statistics includes income items taxed in taxation both as* **earned and capital income**. *From the statistical year 1999 onwards, the concept of earnings has been used for earned income in the income distribution statistics. The content of the concept has not changed.*

Source: Statistic Finland: Metadata: Concepts.

FLEED

- FLEED is (was: It has been superceded by FOLK) the *Finnish Linked Employer-Employee* data.
- All residents in Finland enter the data at age 15.
- The data links rich personal & family information with information on their employer.

FLEED

- Such data allow a researcher to track an individual in multiple interesting ways and allows the study of a wide variety of questions. For example,
  1. Huttunen, K. & Kellokumpu, J. (2016). The effect of job displacement on couples' fertility decisions. *Journal of Labor Economics*, *34*(2), 403–42 study the impact of job loss on fertility, using plant closures to generate (quasi-experimental) variation.
  2. Aghion, P., Akcigit, U., Hyytinen, A. & Toivanen, O. (2018). On the returns to invention within firms: Evidence from finland. *American Economic Association, Papers & Proceedings*, *108*, 208–12 identify the coworkers of inventors of patents to study the impact of invention on coworkers.
- Statistics Finland allows us to download data on a (small) random sample of individuals in the data.

FLEED

- Examples of variables that FLEED contains:
  - education (level and field)
  - # children aged ≤ 7 and ≤ 18.
  - employment months.
  - region
  - and many more...

FLEED - defining our analysis sample

- Lets have a look at FLEED earned income.
- I choose *year* = 2010 (year 15 in the data).
- Whose income are we studying?
- We have a total of 6 244 individuals in the data.
- However, we have income information for only 5 973.
- 48.8% of these are male; on average, they were born in 1968.
- Let's take these 5 973 individuals as our analysis sample, and today concentrate on just earned income.

FLEED - income variable

- Note: In our data, the reported annual income is
  1. rounded to the nearest 1000 euros and
  2. top-coded at 100 000.
- Our job: to make sense of these income data.
- How to proceed? 5 973 is an awful lot of numbers...
- Clearly cannot just skim them (and even if we could, that would not work with the actual data with circa 2.5M individuals in any given year).

What is the objective of (univariate) descriptive statistics?

- In some sense, there is an overflow of information: so many observations that we cannot see the forest for the trees.
- *Descriptive Statistics:* ways of summarizing information in the data in order to make it visible and understandable.
- *Objective*: to reduce the number of figures we need to digest while losing as little information as possible.

A first look at income

- How would you characterize the income of individuals in our data?
- Mean? Median? ... ?
- Income in euros? Something else, if so, what?

Variable	Mean	Median
Income	23 296	21 000

- Why are the mean and the median different / what does the difference imply?

Measures of variation

- How to further characterize the *variation* of income?
- *Standard deviation*.

$$sd = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} = \sqrt{\mathbb{E}[x^2] - (\mathbb{E}[x])^2} \tag{1}$$

- *Coefficient of variation*.

$$CV = sd/mean \tag{2}$$

More univariate descriptive statistics

Variable	Mean	Median	*sd*	*CV*
Income	23 296	21 000	17 163	0.74

- What do we learn from the standard deviation and coefficient of variation regarding these data?

Characterizing the distribution

- While *sd* and *CV* help us understand the variation in income in our data, they are not very informative of the (asymmetries in the) *distribution* of income.
- Two functions, the *density* and *cumulative density* functions, and various empirical measures based on them, help us visualize the distribution.

Density functions - discrete variable

- If the distribution of the stochastic variable X is *discrete*, we can write the probability that X takes a value within the set A as

$$\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x) \tag{3}$$

where $f_X(x)$ is the density function of X.

- We can write the density function of a discrete stochastic variable as

$$f_X(x) = \mathbb{P}(X = x), \tag{4}$$

- with the following conditions holding:

$$f_X(x) \geq 0 \text{ and } \sum_x f_X(x) = 1. \tag{5}$$

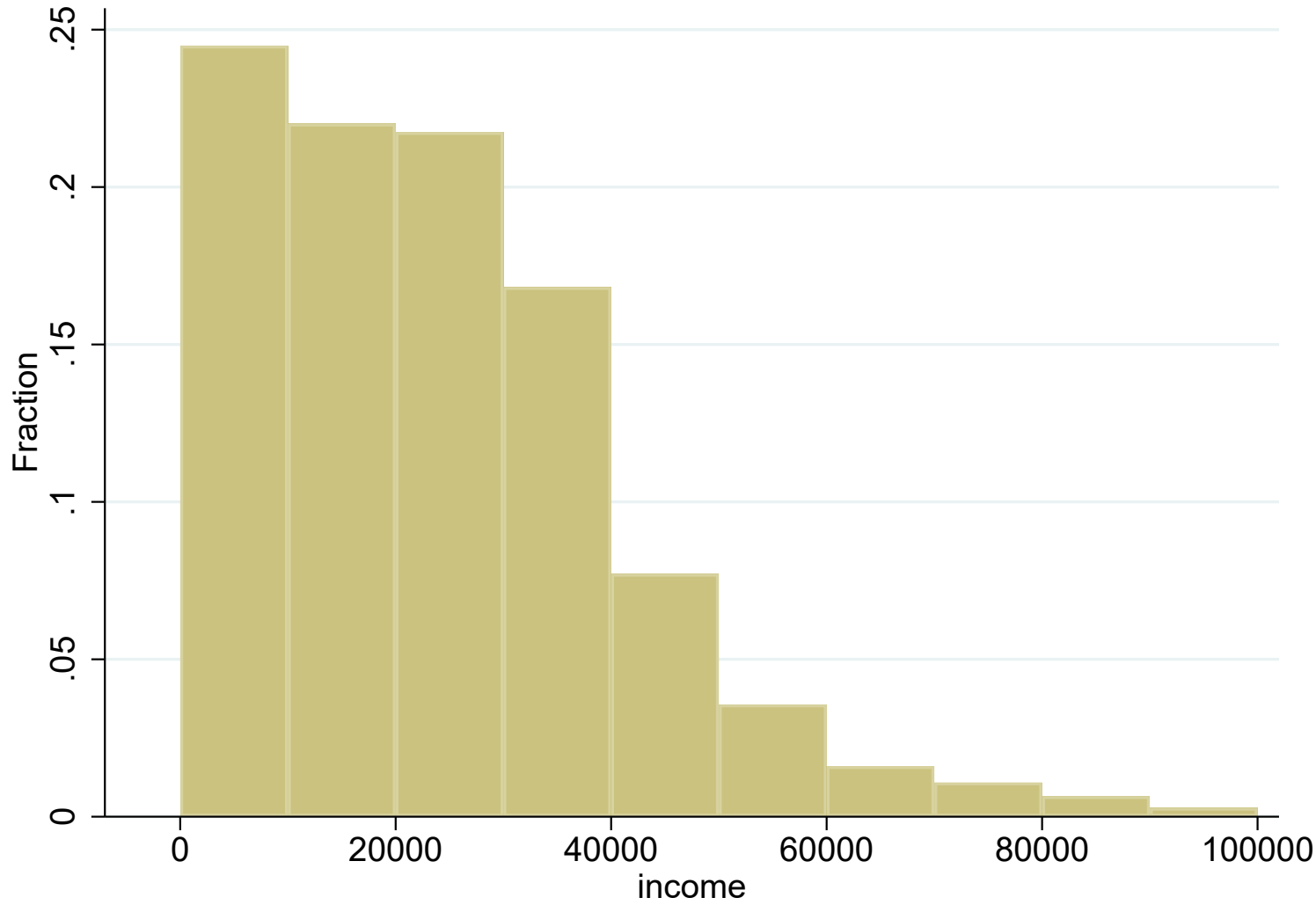
Histogram

- The empirical counterpart of the density function of a discrete variable is a *histogram*.
- More generally useable, a histogram allows us to plot the distribution of a variable by dividing the observations into bins.
- Each observation is allocated to a single bin, and all observations are allocated to some bin.
- The width of the bin describes the values that observations within the bin can take.
- The height of the bin describes the number (fraction) of observations in the bin.

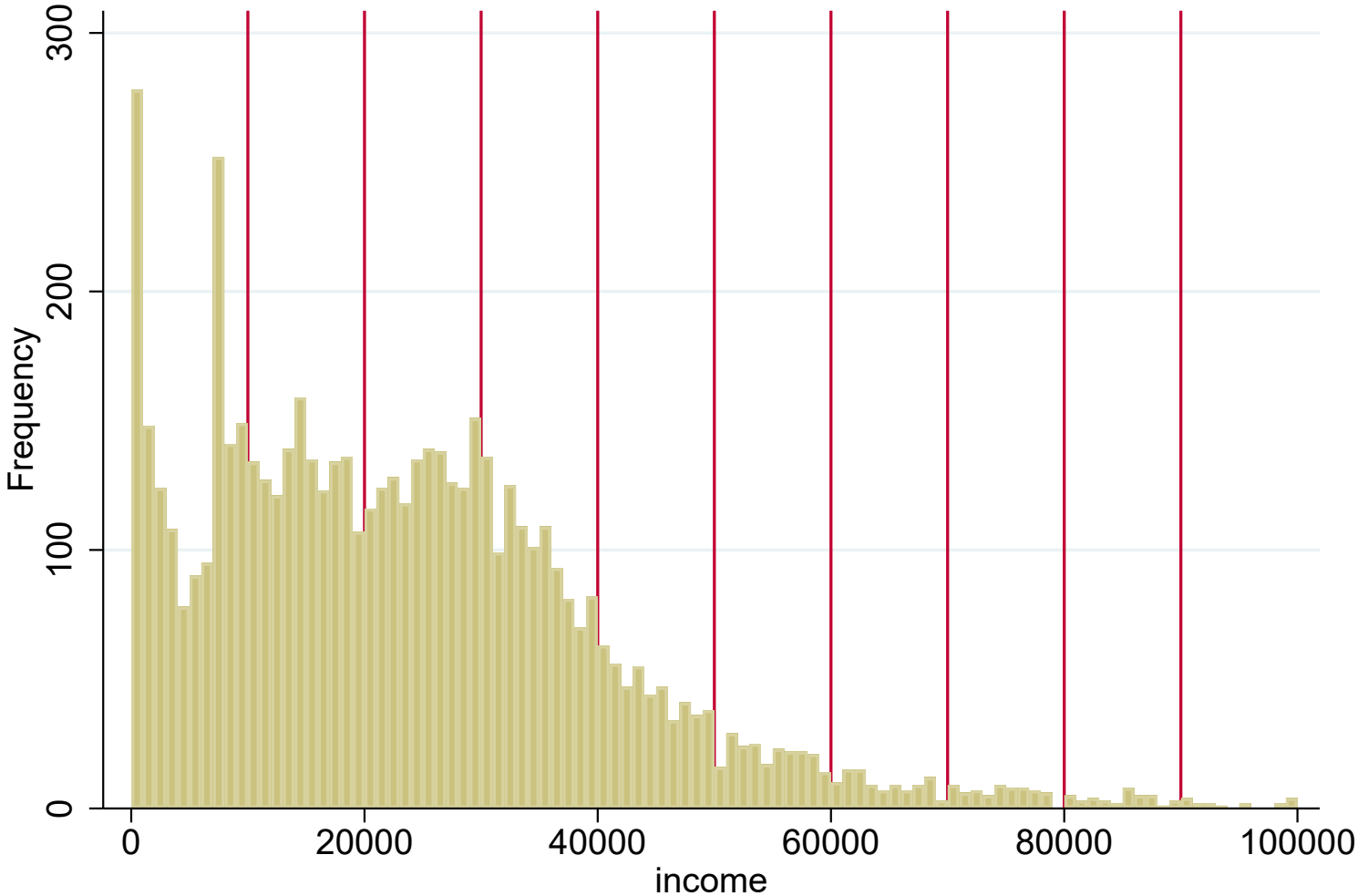
Histogram - Income distribution in Finland 2010

- Let's first distribute the individuals in our data into 10 equally wide bins.
  - $\rightarrow$ width of bin 10 000 euros.
- Let's then distribute the individuals in our data into 100 equally wide bins.
  - $\rightarrow$ width of bin 1 000 euros.

Histogram - Income distribution in Finland 2010



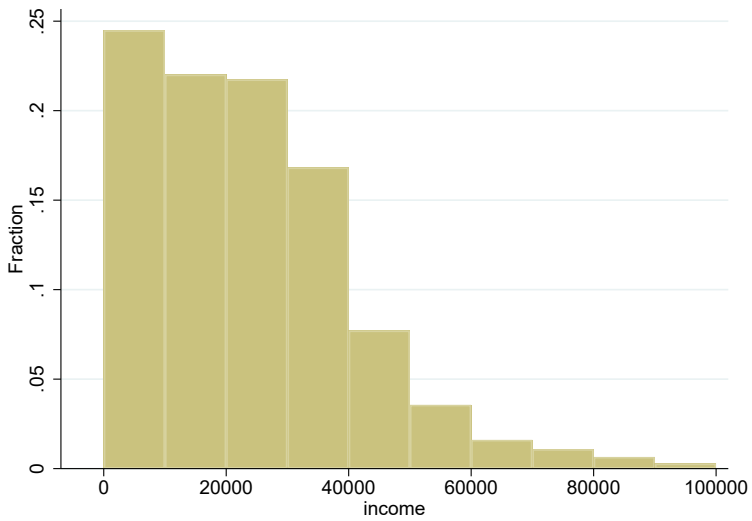
Histogram - Income distribution in Finland 2010

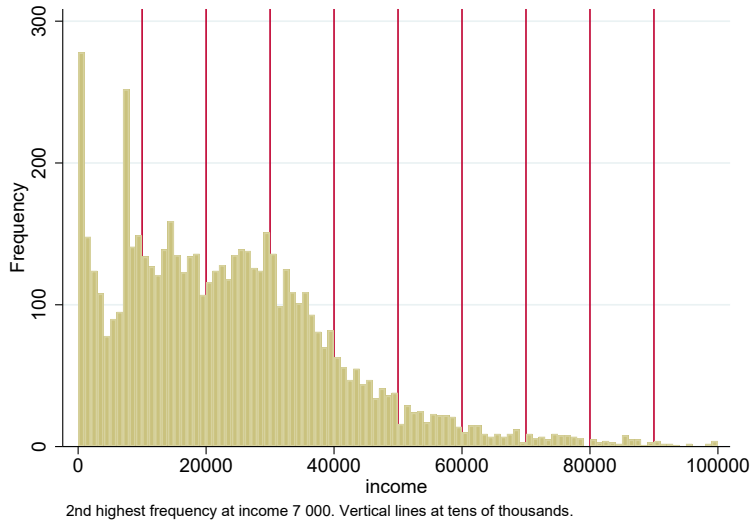


2nd highest frequency at income 7 000. Vertical lines at tens of thousands.

Notice change in y-axis compared to previous figure.

Histogram - Income distribution in Finland 2010





- What, if any, differences between the histograms?

Quantiles

- Besides histograms, *quantiles* can be used to display the distribution.
- Quantile $Q(p)$ is defined as the value such that a fraction p of observations take at most value $Q(p)$.
- Median is an example of a quantile: What value of $X = Q(50)$ is such that exactly half of the observations lie below and the other half above X?.
- Some quantiles have names: median, tercile, quartile, quintile, percentile.

Quantiles

- Various percentiles of the income distribution in our analysis sample.

Percentile	p1	p5	p10	p25	p50	p75	p90	p95	p99
Income	0	1 000	3 000	10 000	21 000	33 000	45 000	55 000	78 000

- What do we learn?
- An often used measure of inequality is the p9010 - ratio, i.e., $p90/p10$.
- Ask yourself: what is the p9010 ratio in our data? What does it mean?

Density functions - continuous variable

- If the distribution of the stochastic variable X is *continuous*, we can write the probability that X takes a value within the set A as

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx \tag{6}$$

- For a continuous variable, the following holds:

$$\mathbb{P}(X = x) = \int_x^x f_X(x)dx = 0.$$

Density functions - continuous variable

- Therefore, an interpretation of the density function for a continuous stochastic variable is as the probability wrt. to small varition, namely that for small $h > 0$, the following holds:

$$f_X(x) \approx \frac{\mathbb{P}(X = x \pm h/2)}{h}, \tag{7}$$

- where $(X = x \pm h/2)$ means the event $x - h/2 \leq X \leq x + h/2$.
- The density function of a continuous variable satisfies

$$f_X(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} f_X(x) = 1. \tag{8}$$

- Equation (7) is the basis for the definition of a *kernel*.

Kernel function

- Define a kernel function, normalized such that

$$\int_{-\infty}^{\infty} K(u)du = 1$$

- and symmetric

$$K(u) = K(-u).$$

1. Uniform kernel: $K(u) = 1/2$, support $|u| \leq 1$.
2. Triangular kernel: $K(u) = (1 - |u|)$, support $|u| \leq 1$.
3. Epanechnikov kernel (parabolic):$K(u) = \frac{3}{4}(1 - u^2)$, support $|u| \leq 1$.

- For all these kernels, $K(u) = 0$ outside the support.

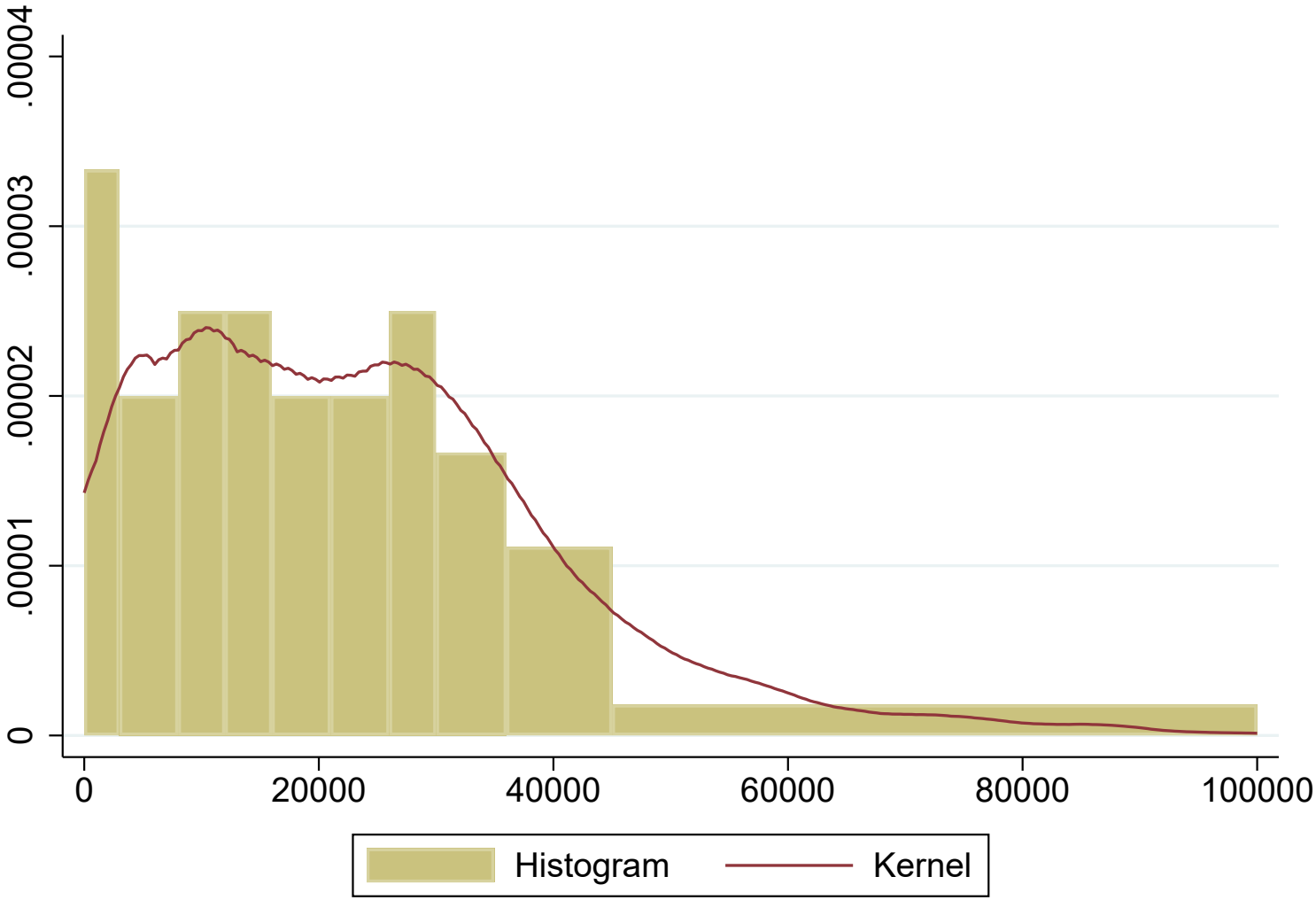
Kernel function

- A kernel density estimator is given by:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(x - x_i)/h). \tag{9}$$

- The parameter h is the *bandwidth* of the kernel.
- The choice of h is governed by a tradeoff between bias and variance.

Kernel density plot - Income distribution in Finland 2010



Note: Histogram shows income bins with 1/10th of the sample each

Cumulative density function

- The density function (= empirical measures based on it) is useful when we want to answer questions such as:
  - How many (what fraction of) people in Finland in 2010 earned between 9 and 12 000 euros (a year)?
- We might also want to answer questions such as:
  - How many people (what fraction of) people in Finland in 2010 earned at most 12 000 euros (a year)?
- The tool to answer such questions is the *cumulative density function* (cdf) which for a continuous variable is defined as:

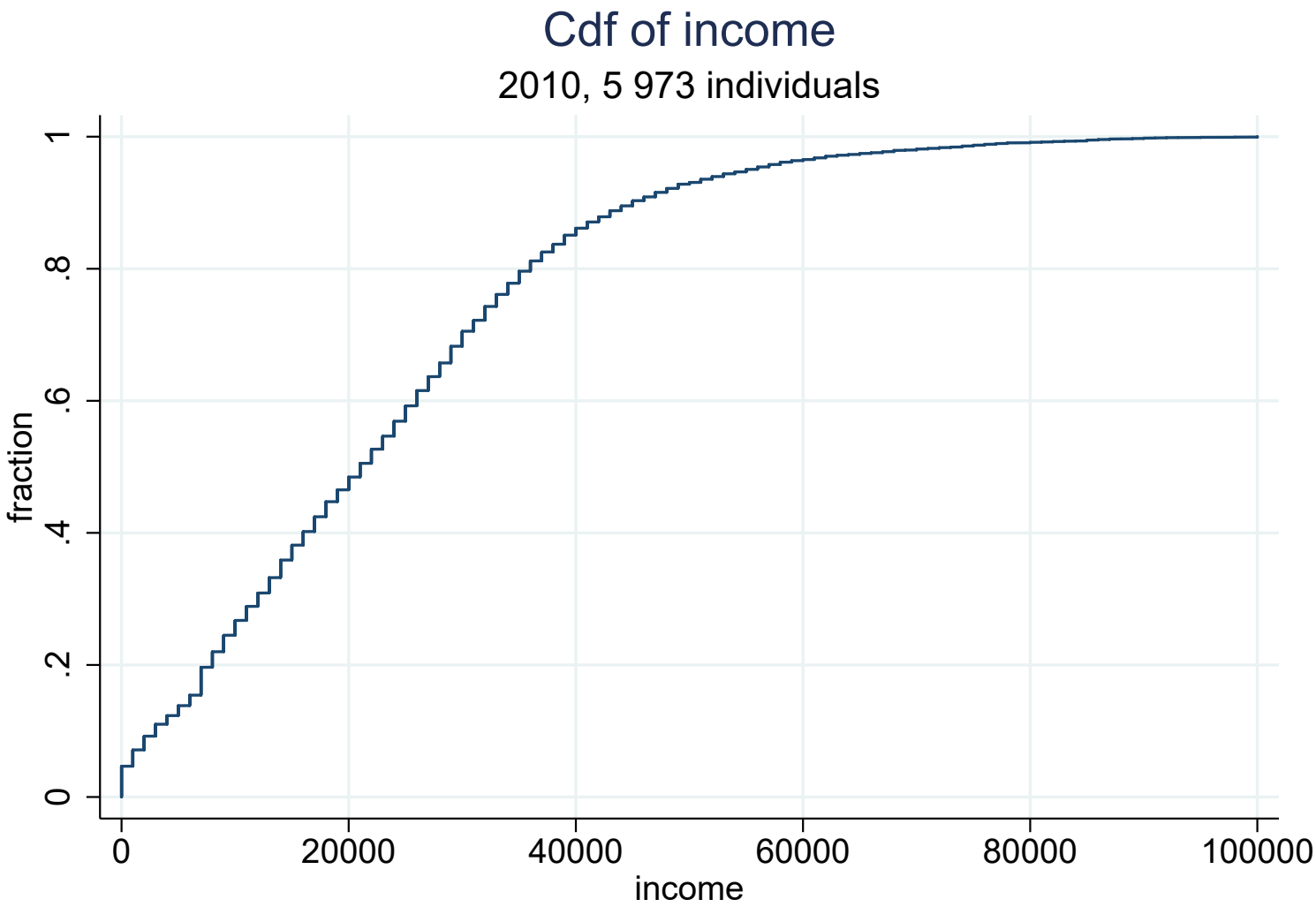
$$F_X(t) = \int_{-\infty}^{t} f_X(s)ds \tag{10}$$

Cumulative density function

- Let's study the number / fraction of individuals in our analysis sample that earn at most x euros, $x = 1000, 2000, ..., 12000$.

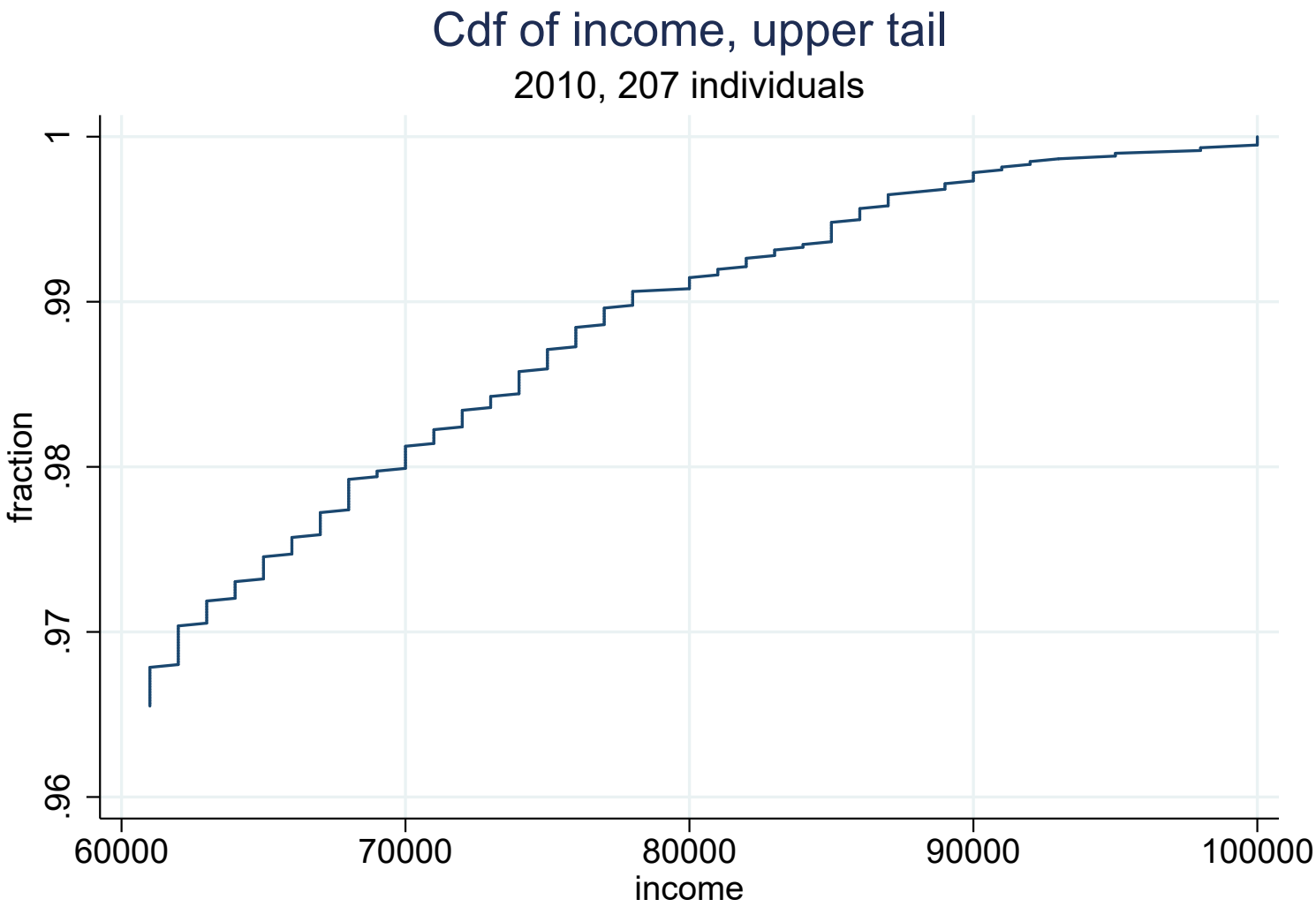
income	#	pdf	cumul	cdf
0	278	0.05	278	0.05
1 000	148	0.02	426	0.07
2 000	124	0.02	550	0.09
3 000	108	0.02	658	0.11
4 000	78	0.01	736	0.12
5 000	90	0.02	826	0.14
6 000	95	0.02	921	0.15
7 000	252	0.04	1173	0.20
8 000	141	0.02	1314	0.22
9 000	149	0.02	1463	0.24
10 000	134	0.02	1597	0.27
11 000	127	0.02	1724	0.29
12 000	121	0.02	1845	0.31
...	...	...	...	...
100 000	5973	1.00	5973	1.00

Cdf of the income distribution in Finland 2010



Ask yourself: Based on the figure, what fraction of people in our data earn at most 20 000 / 60 000 euros?

Cdf of the income distribution in Finland 2010, upper tail



Ask yourself: Based on the figure, what fraction of people in our data earn at most 60 000 / 80 000 euros?

Back to objectives of descriptive statistics

- The objective of descriptive statistics is to reveal features of your data that speak to your research question.
- While it is necessary to display means and sd's, those rarely in themselves satisfy this objective.
- Therefore always ask yourself:
  - If I was the reader / in the audience, what would I like to learn about the data given the research question posed by the researcher?
- Example: A topical question regarding the income distribution is
  - How is the "pie" shared? Watch video by Lucas Chancel.
  - Note: in so doing we are actually already entering the world of *conditional* descriptive statistics.

Share of wage returns to invention by worker type

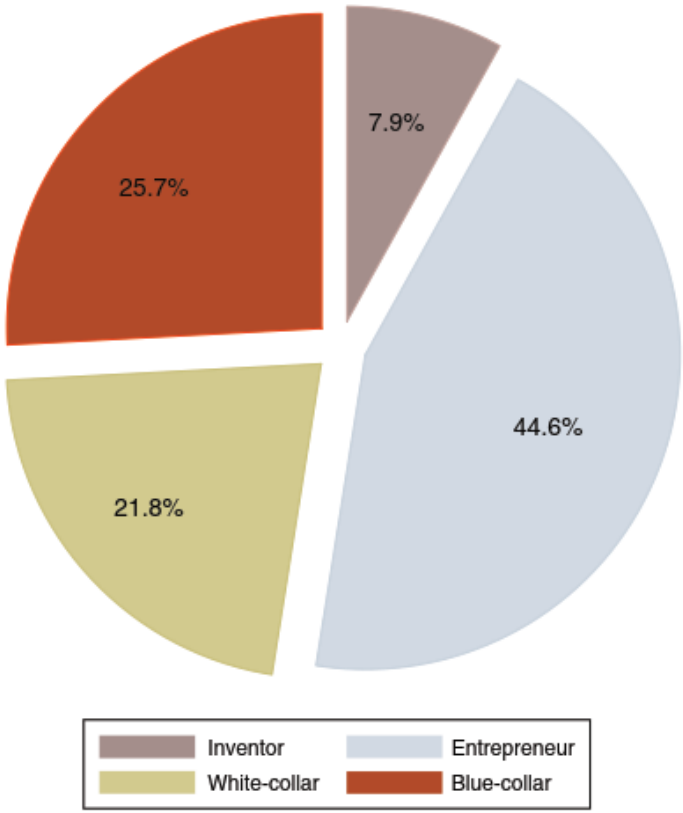


FIGURE 2. RETURNS DISTRIBUTION

Aghion, P., Akcigit, U., Hyytinen, A. & Toivanen, O. (2018). On the returns to invention within firms: Evidence from finland. *American Economic Association, Papers & Proceedings*, *108*, 208–12.