Descriptive Statistics I

Matti Sarvimäki

Principles of Empirical Analysis
Lecture 2

Course outline and learning objectives

- Data and measurement
  1. introduction, data
  2. **today: descriptive statistics**
  3. more descriptive statistics
- Experimental methods
  1. causality and research designs
  2. statistical significance
  3. statistical power
  4. noncompliance
- Quasi-experimental methods
  1. observational data and quasi-experiments
  2. difference-in-difference (DiD)
  3. regression discontinuity design (RDD)

Today's learning objectives

- Descriptive statistics
  - mean, median, quantiles
  - variance and standard deviation
  - density functions
- Sample and Population
  - representativeness
  - sampling error

Descriptive statistics

Income distribution

- Aim: learning to characterize distributions
- Example: income distribution
  - the learning objectives could be fulfilled with any distribution
  - but this one is particularly central to much research and policy debate
- Outline
  - today: basics using Statistics Finland's teaching data
  - next week: making sense of some of the key papers on inequality and social mobility

Source: The Economist, 28 Nov 2019

Statistics Finland's teaching data

- We use Statistic Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee (FLEED) dataset
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here; in English here → Variable description)

Statistics Finland's teaching data

- We use Statistic Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee (FLEED) dataset
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here; in English here → Variable description)
- We will use annual earned income for this analysis
  - Statistic Finland's metadata: "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**."
  - initial source: Finland's Tax Authority

Statistics Finland's teaching data

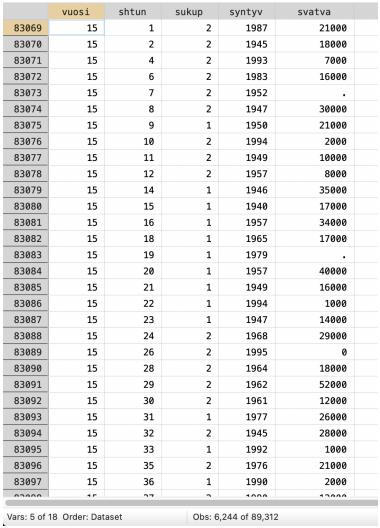
- We use Statistic Finland's teaching data for this lecture and some your exercises
  - random sample of the old Finnish Linked Employer-Employee (FLEED) dataset
    - now under the name FOLK for research purposes (taika.stat.fi)
  - lots of information about all working age residents living in Finland and their employers (data description here; in English here → Variable description)
- We will use annual earned income for this analysis
  - Statistic Finland's metadata: "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**."
  - initial source: Finland's Tax Authority
- In teaching data, all income is
  1. rounded to the nearest 1,000 euros
  2. top-coded at 100,000

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data



	vuosi	shtun	sukup	syntyv	svatva
83069	15	1	2	1987	21000
83070	15	2	2	1945	18000
83071	15	4	2	1993	7000
83072	15	6	2	1983	16000
83073	15	7	2	1952	.
83074	15	8	2	1947	30000
83075	15	9	1	1950	21000
83076	15	10	2	1994	2000
83077	15	11	2	1949	10000
83078	15	12	2	1957	8000
83079	15	14	1	1946	35000
83080	15	15	1	1940	17000
83081	15	16	1	1957	34000
83082	15	18	1	1965	17000
83083	15	19	1	1979	.
83084	15	20	1	1957	40000
83085	15	21	1	1949	16000
83086	15	22	1	1994	1000
83087	15	23	1	1947	14000
83088	15	24	2	1968	29000
83089	15	26	2	1995	0
83090	15	28	2	1964	18000
83091	15	29	2	1962	52000
83092	15	30	2	1961	12000
83093	15	31	1	1977	26000
83094	15	32	2	1945	28000
83095	15	33	1	1992	1000
83096	15	35	2	1976	21000
83097	15	36	1	1990	2000

Vars: 5 of 18 Order: Dataset Obs: 6,244 of 89,312

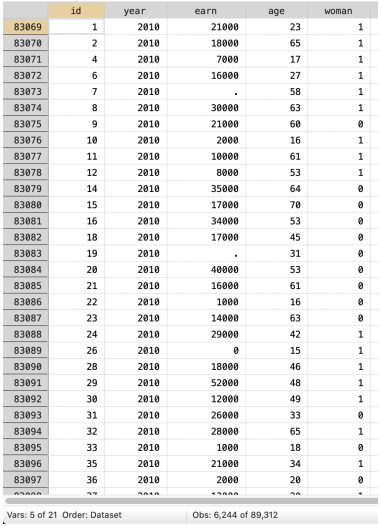
Source: FLEED teaching data

```
browse shtun vuosi sukup syntyv svatva if vuosi==15
```

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data
  - second: let's clean it a little bit

```
rename shtun id
gen year=1995+vuosi
gen woman=(sukup==2)
replace woman=.  if sukup==.
gen age=year-syntyv
rename svatva earn
keep if year==2010
order id year earn age woman
```



	id	year	earn	age	woman
83069	1	2010	21000	23	1
83070	2	2010	18000	65	1
83071	4	2010	7000	17	1
83072	6	2010	16000	27	1
83073	7	2010	.	58	1
83074	8	2010	30000	63	1
83075	9	2010	21000	60	0
83076	10	2010	2000	16	1
83077	11	2010	10000	61	1
83078	12	2010	8000	53	1
83079	14	2010	35000	64	0
83080	15	2010	17000	70	0
83081	16	2010	34000	53	0
83082	18	2010	17000	45	0
83083	19	2010	.	31	0
83084	20	2010	40000	53	0
83085	21	2010	16000	61	0
83086	22	2010	1000	16	0
83087	23	2010	14000	63	0
83088	24	2010	29000	42	1
83089	26	2010	0	15	1
83090	28	2010	18000	46	1
83091	29	2010	52000	48	1
83092	30	2010	12000	49	1
83093	31	2010	26000	33	0
83094	32	2010	28000	65	1
83095	33	2010	1000	18	0
83096	35	2010	21000	34	1
83097	36	2010	2000	20	0

Vars: 5 of 21 Order: Dataset | Obs: 6,244 of 89,312

Source: FLEED teaching data

`browse id year earn age woman`

First look at the data

- Let's have a look at 2010 earned income
  - total of 6,244 individuals in the data.
  - income information for only 5,973 (inc. those with now zero income)
- How to make sense of these data?
  - first: let's look at the data
  - second: let's clean it a little bit

```
rename shtun id
gen year=1995+vuosi
gen woman=(sukup==2)
replace woman=.  if sukup==.
gen age=year-syntyv
rename svatva earn
keep if year==2010
order id year earn age woman
```

  - still: 5,973 is an awful lot of numbers...
- We need to find ways to summarize the data in an informative, but parsimonious manner

	id	year	earn	age	woman
83069	1	2010	21000	23	1
83070	2	2010	18000	65	1
83071	4	2010	7000	17	1
83072	6	2010	16000	27	1
83073	7	2010	.	58	1
83074	8	2010	30000	63	1
83075	9	2010	21000	60	0
83076	10	2010	2000	16	1
83077	11	2010	10000	61	1
83078	12	2010	8000	53	1
83079	14	2010	35000	64	0
83080	15	2010	17000	70	0
83081	16	2010	34000	53	0
83082	18	2010	17000	45	0
83083	19	2010	.	31	0
83084	20	2010	40000	53	0
83085	21	2010	16000	61	0
83086	22	2010	1000	16	0
83087	23	2010	14000	63	0
83088	24	2010	29000	42	1
83089	26	2010	0	15	1
83090	28	2010	18000	46	1
83091	29	2010	52000	48	1
83092	30	2010	12000	49	1
83093	31	2010	26000	33	0
83094	32	2010	28000	65	1
83095	33	2010	1000	18	0
83096	35	2010	21000	34	1
83097	36	2010	2000	20	0

Vars: 5 of 21 Order: Dataset | Obs: 6,244 of 89,312

Source: FLEED teaching data

`browse id year earn age woman`

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible
- Let's start with Stata's summarize command

```
summarize earn, detail
```

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Descriptive statistics

- **Descriptive statistics:** ways of summarizing information to make data understandable
  - objective: reduce the amount of numbers as much as possible while losing as little information as possible
- Let's start with Stata's summarize command

```
summarize earn, detail
```

- It gives us the key descriptive statistics:
  - sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - single number measures of variation
  - selected quantiles

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Measures of variation

- Variance:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Standard deviation:

$$SD(x) = \sqrt{Var(x)}$$

- Sometimes we normalize the standard deviation with mean. This is called the coefficient of variation. In this example:

$$CV(x) = \frac{SD(x)}{\bar{x}} = \frac{17,164}{23,297} = .74$$

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Quantiles

- Quantile $Q(p)$: value such that a fraction p of observations take at most value $Q(p)$.

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Quantiles

- Quantile $Q(p)$: value such that a fraction p of observations take at most value $Q(p)$.
- Some quantiles have names, e.g., **median**
  - $Q(.5)$: 50% of observations below this value

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Quantiles

- Quantile $Q(p)$: value such that a fraction p of observations take at most value $Q(p)$.
- Some quantiles have names, e.g., **median**
  - $Q(.5)$: 50% of observations below this value
  - Some other named quantiles
    - quartiles: $Q(.25)$, $Q(.5)$, $Q(.75)$
    - deciles: $Q(.1)$, $Q(.2)$, ..., $Q(.9)$
    - percentiles: $Q(.01)$, $Q(.02)$, ..., $Q(.99)$

earn

	Percentiles	Smallest		
1%	**0**	**0**		
5%	**1000**	**0**		
10%	**3000**	**0**	Obs	**5,973**
25%	**10000**	**0**	Sum of Wgt.	**5,973**
50%	**21000**		Mean	**23296.67**
		Largest	Std. Dev.	**17163.61**
75%	**33000**	**100000**		
90%	**45000**	**100000**	Variance	**2.95e+08**
95%	**55000**	**100000**	Skewness	**1.006775**
99%	**78000**	**100000**	Kurtosis	**4.340098**

Source: FLEED teaching data

Quantiles

- Quantile $Q(p)$: value such that a fraction p of observations take at most value $Q(p)$.
- Some quantiles have names, e.g., **median**
  - $Q(.5)$: 50% of observations below this value
  - Some other named quantiles
    - quartiles: $Q(.25)$, $Q(.5)$, $Q(.75)$
    - deciles: $Q(.1)$, $Q(.2)$, ..., $Q(.9)$
    - percentiles: $Q(.01)$, $Q(.02)$, ..., $Q(.99)$
- The width of the distribution is often characterized with percentile ratios:
  - 90/10 ratio: $Q(.9)/Q(.1) = 15$
  - 90/50 ratio: $Q(.9)/Q(.5) = 2.1$
  - 50/10 ratio: $Q(.5)/Q(.1) = 7$

earn

	Percentiles	Smallest		
1%	0	0		
5%	1000	0		
10%	3000	0	Obs	5,973
25%	10000	0	Sum of Wgt.	5,973
50%	21000		Mean	23296.67
		Largest	Std. Dev.	17163.61
75%	33000	100000		
90%	45000	100000	Variance	2.95e+08
95%	55000	100000	Skewness	1.006775
99%	78000	100000	Kurtosis	4.340098

Source: FLEED teaching data

Density functions

Density function: discrete variables

- If the distribution of the random variable X is *discrete*, it's **density function** is

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

  i.e. the **probability that the random variable, X, takes a specific value, x**

Density function: discrete variables

- If the distribution of the random variable X is *discrete*, it's **density function** is

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

i.e. the **probability that the random variable, X, takes a specific value, x**

- note that the following conditions must hold

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

Density function: discrete variables

- If the distribution of the random variable X is *discrete*, it's **density function** is

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

  i.e. the **probability that the random variable, X, takes a specific value, x**
  - note that the following conditions must hold

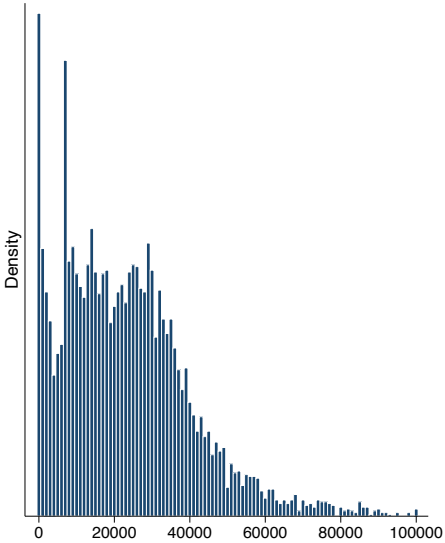
$$f_X(x) \geq 0 \text{ and } \sum_x f_X(x) = 1$$

- Thus, the probability that X takes a value within the set A is

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

Histogram

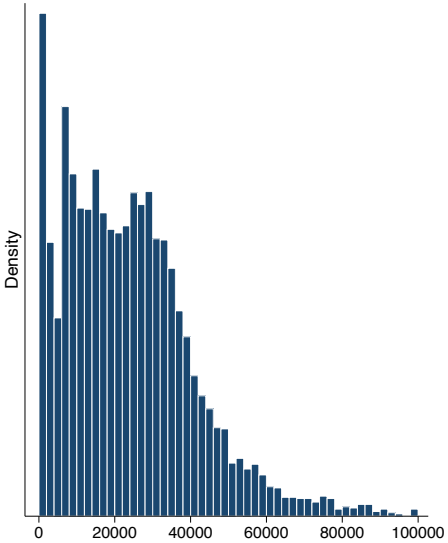
- The empirical counterpart of the density function of a discrete variable is a **histogram**.
  - the height of the bar describes the fraction of observations that take the value x



Source: FLEED teaching data
`hist earn, disc`

Histogram

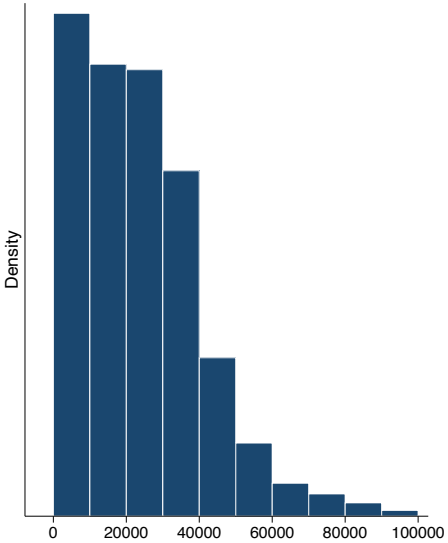
- The empirical counterpart of the density function of a discrete variable is a **histogram**.
  - the height of the bar describes the fraction of observations that take the value x
- More generally: we can divide the observations into **bins** and draw a histogram of them
  - each observation is allocated 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.



Source: FLEED teaching data
`hist earn, bin(50)`

Histogram

- The empirical counterpart of the density function of a discrete variable is a **histogram**.
  - the height of the bar describes the fraction of observations that take the value x
- More generally: we can divide the observations into **bins** and draw a histogram of them
  - 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.
- Changing the number of bins may allow us to see the same data differently



Source: FLEED teaching data

```
hist earn, bin(10)
```

Density function: continuous variables

- If the distribution of the random variable X is *continuous*, the probability that X takes a value within the set A is

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

  - note that continuous variable can take infinite values and thus the likelihood that X takes a specific value is zero, i.e. $\mathbb{P}(X = x) = \int_x^x f_X(x)dx = 0$.

Density function: continuous variables

- If the distribution of the random variable X is *continuous*, the probability that X takes a value within the set A is

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

  - note that continuous variable can take infinite values and thus the likelihood that X takes a specific value is zero, i.e. $\mathbb{P}(X = x) = \int_x^x f_X(x)dx = 0$.

- An interpretation of the density function for a continuous stochastic variable is as the probability wrt. to small variation, $h > 0$, the following holds:

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

  where $(X = x \pm h/2)$ means the event $x - h/2 \leq X \leq x + h/2$.

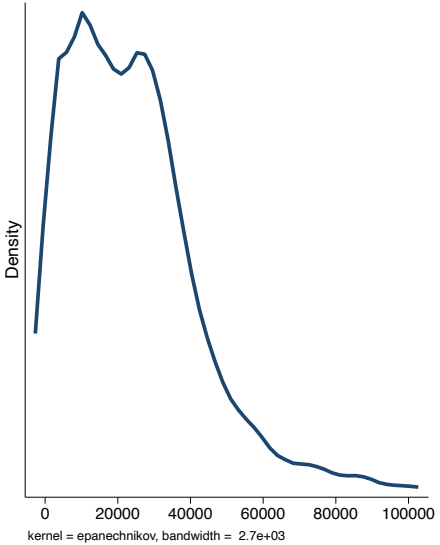
- This is the basis for the definition of a **kernel**.

Kernel density estimator

- A kernel density estimator is essentially a local (weighted) average for each value x:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

  - **bandwidth** (h): how much data around x is used
  - **kernel function** (K_h): how do we weight observations within the bandwidth, i.e, do observations further away from x get lower weight?
- By default, Stata chooses an "optimal" bandwidth



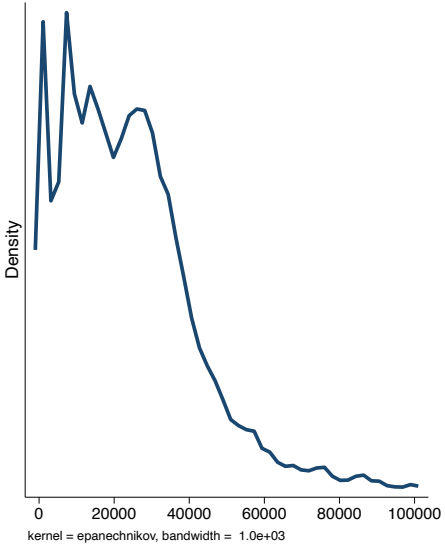
Source: FLEED teaching data
`kdensity earn`

Kernel density estimator

- A kernel density estimator is essentially a local (weighted) average for each value x:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

  - **bandwidth** (h): how much data around x is used
  - **kernel function** (K_h): how do we weight observations within the bandwidth, i.e, do observations further away from x get lower weight?
- By default, Stata chooses an "optimal" bandwidth
  - smaller bandwidth creates more noice



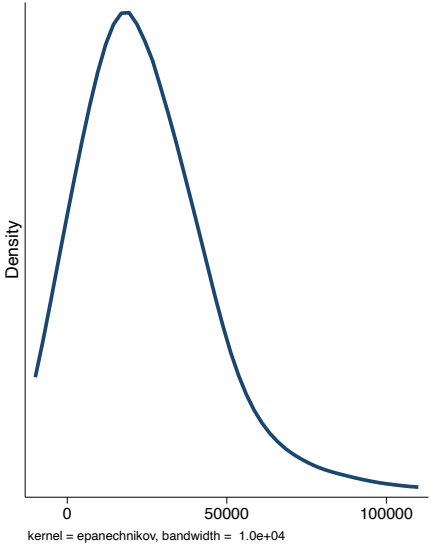
Source: FLEED teaching data
`kdensity earn, bw(1000)`

Kernel density estimator

- A kernel density estimator is essentially a local (weighted) average for each value x:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

  - **bandwidth** (h): how much data around x is used
  - **kernel function** (K_h): how do we weight observations within the bandwidth, i.e, do observations further away from x get lower weight?
- By default, Stata chooses an "optimal" bandwidth
  - smaller bandwidth creates more noice
  - larger bandwidth disregards more data



Source: FLEED teaching data
`kdensity earn, bw(10000)`

Cumulative density function

- **Cumulative density function** (CDF) for a continuous variable is defined as:

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

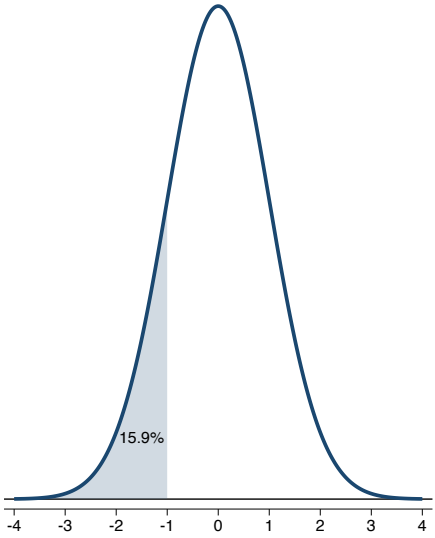
- It answers the question: **what fraction of the observations have values of x below t?**

Cumulative density function

- **Cumulative density function** (CDF) for a continuous variable is defined as:

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

- It answers the question: **what fraction of the observations have values of x below t?**
  - e.g. for standardized normal distribution: $F_X(-1) = 0.159$



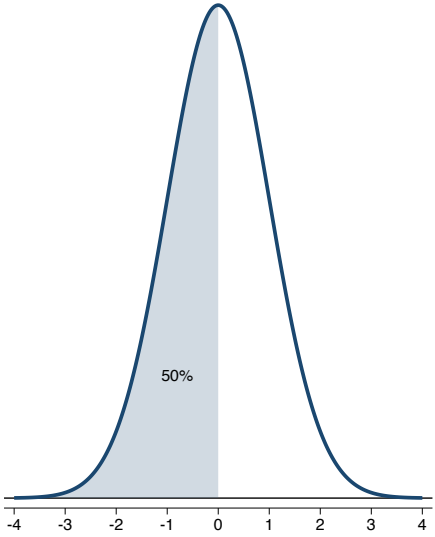
Density function of the standard normal distribution

Cumulative density function

- **Cumulative density function** (CDF) for a continuous variable is defined as:

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

- It answers the question: **what fraction of the observations have values of x below t?**
  - e.g. for standardized normal distribution:
    $F_X(-1) = 0.159$
    $F_X(0) = 0.5$



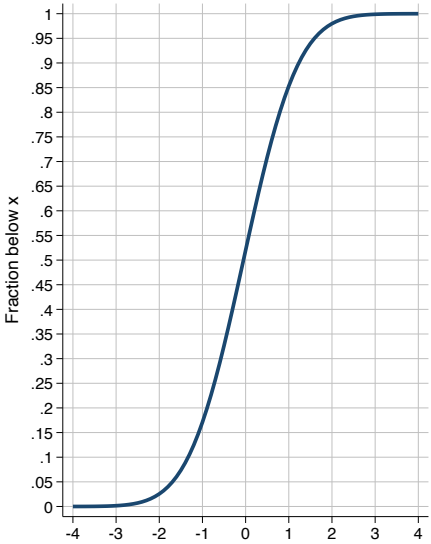
Density function of the standard normal distribution

Cumulative density function

- **Cumulative density function** (CDF) for a continuous variable is defined as:

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

- It answers the question: **what fraction of the observations have values of x below t?**
  - e.g. for standardized normal distribution:
    $F_X(-1) = 0.159$
    $F_X(0) = 0.5$
- Plot all of these points to draw the entire CDF

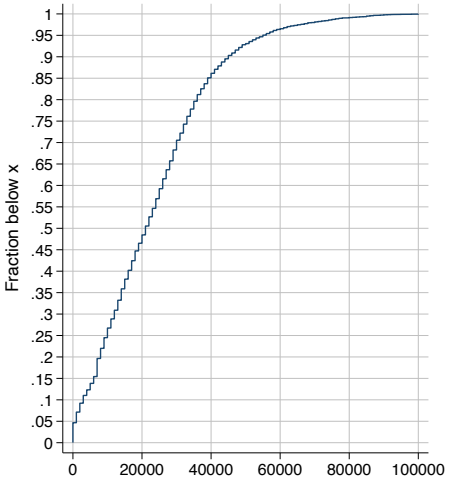


CDF of the standard normal distribution

CDF of Income in 2010

- Let's return to the teaching data and calculate the fraction of individuals in our analysis sample who earn at most x euros, $x = 1000, 2000, ...$

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
...	...	...	...	...
Total	5973	1.00	5973	1.00



Source: FLEED teaching data
`distplot earn`

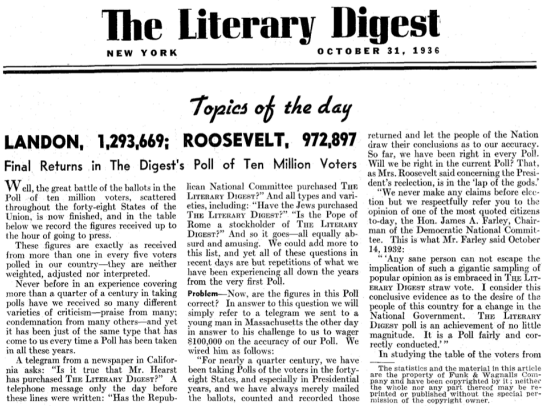
Population and sample

Population and sample

- Population
  - the entire group that you want to draw conclusions about (N units)
- Sample
  - specific group we select out of the population and collect data from (n units)
  - aim is to make an **inference** of the population
    - infer: deduce or conclude (information) from evidence and reasoning rather than from explicit statements
- Things to worry about
  - sampling bias: sample does not represent population
  - sampling error: exceptional observations sampled by chance

Sampling bias: 1936 US Presidential Election Polls

- In 1936, Literary Digest sent 10 million "straw" ballots asking who people were planning to vote in the upcoming election
  - 2.4 million were returned: 57% to Landon and 43% to Roosevelt



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased The Literary Digest?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased The Literary Digest?" "Is the Pope of Rome a stockholder of The Literary Digest?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of voters from

Source: Sidetrade Tech Hub.

Sampling bias: 1936 US Presidential Election Polls

- In 1936, Literary Digest sent 10 million "straw" ballots asking who people were planning to vote in the upcoming election
  - 2.4 million were returned: 57% to Alf Landon and 43% to Roosevelt
  - Roosevelt won the elections 62% to 37%
- George Gallup also conduced a poll
  - sample size just 50,000
  - prediction: 56% to Roosevelt
- Discuss: why was Gallup's data better?
  - compare to the 2020 polling

"All the News That's Fit to Print."

The New York Times.

FINAL EXTRA

NEW YORK, WEDNESDAY, NOVEMBER 4, 1936.

ROOSEVELT SWEEPS THE NATION; HIS ELECTORAL VOTE EXCEEDS 500; LEHMAN WINS; CHARTER ADOPTED

FEW HOUSE SHIFTS — Democrats May Add to Vast Majorities in Both Chambers — THREE SENATORS TRAIL — Barbour, Hastings and Metcalf Appear to Have Lost Seats.

Landon Congratulates President, Who Replies

UNION PARTY VOTE FAR BELOW BOASTS — Coughlin Group Appears to Have Exercised Little Influence on the Electorate.

BIG CHARTER VOTE — 8-Hour System for Firemen Also Wins Easily — VOTING CHANGE APPROVED — Brunner Is Victor Over Morris by Large Plurality. — ROOSEVELT SWEEP HERE

Smith Plans Comment On the Election Today

DEMOCRATS RETAIN STATE SENATE LEAD — They Are Assured of 30 Seats of the 51, One More Than Their Previous Number.

LEHMAN VOTE CUT — Bleakley Gets a Surprising Total in the City — SWEEP HELPS GOVERNOR — Roosevelt Strong in Industrial Cities—Gets Big Up-State Poll.

FRANKLIN D. ROOSEVELT

DEMOCRATS SWEEP ALL PENNSYLVANIA

JERSEY'S 16 VOTES SAFE FOR NEW DEAL

POLL SETS RECORD — Roosevelt Electoral Vote of 519 Seen as a Minimum — NO SWING TO THE BOLTERS — 'Jeffersonian Democrats' Fail to Cause Rift as Expected. — NEIGHBORS HAIL PRESIDENT

Source: New York Times, 4 Nov 1936.

Sampling error

- Random sampling removes bias
  - each object in the population has the same probability of being selected into the sample

Sampling error

- Random sampling removes bias
  - each object in the population has the same probability of being selected into the sample
- ... but **sampling error** remains
  - difference between a sample statistic and population parameter arising by chance

Sampling error

- Example: **Population mean** of income among 15–64 year olds people living in Finland in 2010 ($N \approx 3.5M$)

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i = €26,144$$

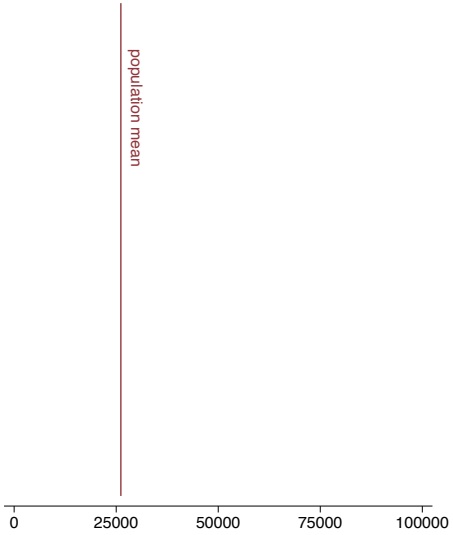
- Suppose we take a random sample of n people from the full-population data and calculate a **sample average**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- *Question*: What is the relationship between μ_x, $\bar{x}$, and n?

Sampling error

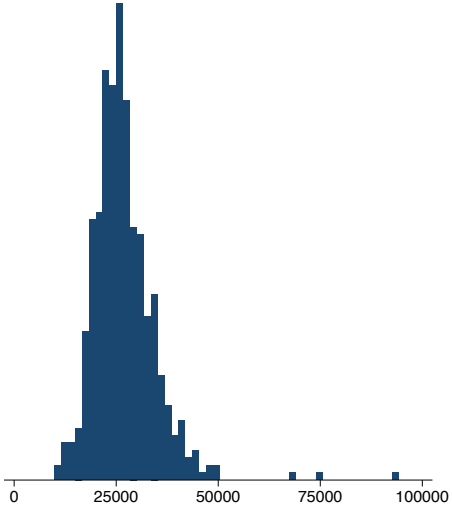
- Let's take many random samples from the full-population using different sample sizes
  -



Source: Statistics Finland's population level research data

Sampling error

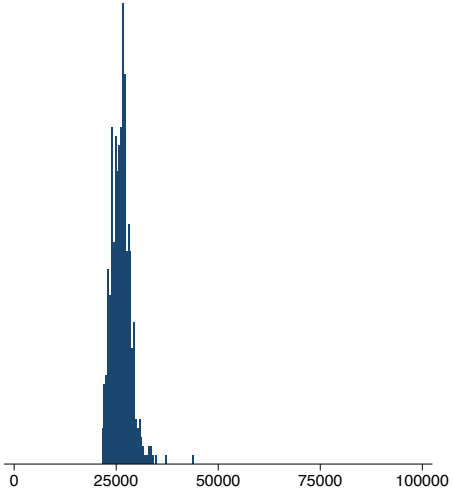
- Let's take many random samples from the full-population using different sample sizes
  - $n = 10$



Source: Statistics Finland's population level research data

Sampling error

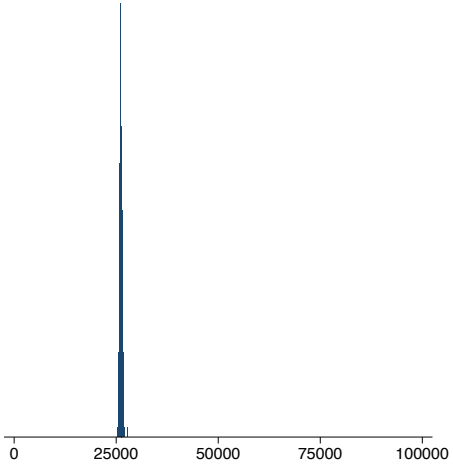
- Let's take many random samples from the full-population using different sample sizes
  - $n = 100$



Source: Statistics Finland's population level research data

Sampling error

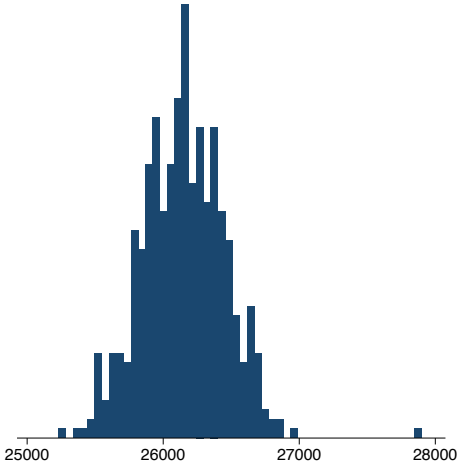
- Let's take many random samples from the full-population using different sample sizes
  - $n = 5,973$



Source: Statistics Finland's population level research data

Sampling error

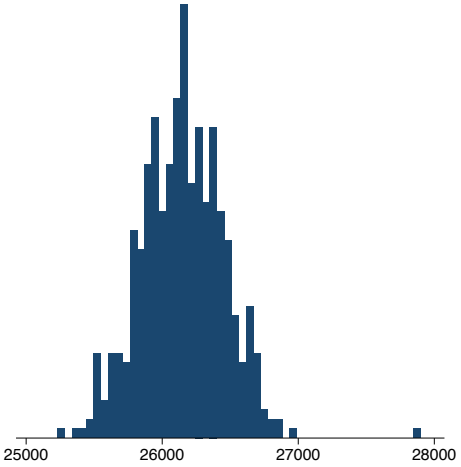
- Let's take many random samples from the full-population using different sample sizes
  - $n = 5,973$ (zooming in)



Source: Statistics Finland's population level research data

Sampling error

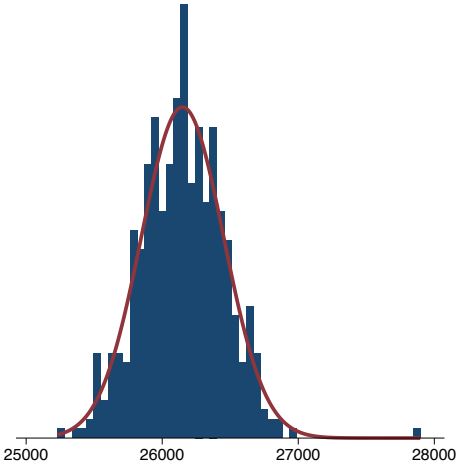
- Let's take many random samples from the full-population using different sample sizes
  - $n = 5,973$ (zooming in)
- Take-aways
  1. the larger the sample size, the closer the sample averages tend to be to the population mean
  2. sample averages are distributed relatively symmetrically around population mean



Source: Statistics Finland's population level research data

Sampling error

- Let's take many random samples from the full-population using different sample sizes
  - $n = 5,973$ (zooming in)
- Take-aways
  1. the larger the sample size, the closer the sample averages tend to be to the population mean
  2. sample averages are distributed relatively symmetrically around population mean
- These properties are also know as
  1. The Law of Large Numbers
  2. The Central Limit Theorem
- They are deep results at the heart of statistics
  - properly discussed in MS-A0503 and/or 2nd year econometrics; here, we just build intuition



Source: Statistics Finland's population level research data

Summary

- Today's lecture was mainly about learning the basic concepts
  1. Concepts you need to know understand well
     - mean, variance, standard deviation, coefficient of variation
     - quantiles, median, percentiles, deciles
     - density function, CDF
  2. Things to worry when using samples
     - representativeness
     - sampling error

Summary

- Today's lecture was mainly about learning the basic concepts
  1. Concepts you need to know understand well
     - mean, variance, standard deviation, coefficient of variation
     - quantiles, median, percentiles, deciles
     - density function, CDF
  2. Things to worry when using samples
     - representativeness
     - sampling error
- Next time, we will
  1. learn a few more key concepts
  2. make sense of recent research on inequality