Samples and descriptive statistics

Prottoy A. Akbar

Principles of Empirical Analysis (ECON-A3000)
Lecture 2

Logistics

- Do you have your name placards from last time?
- Pre-class assignment 1 was due 15 minutes ago.
    - You are allowed up to 2 skips without penalty.
    - pass/fail grade based on effort

Logistics

- Do you have your name placards from last time?
- Pre-class assignment 1 was due 15 minutes ago.
  - You are allowed up to 2 skips without penalty.
  - pass/fail grade based on effort
- In-class worksheet 1 today!
  - pick up from upfront
  - when done, you can take a photo/scan and submit on MyCourses before next class.
  - pass/fail grade based on accuracy

Today's learning objectives

- Descriptive statistics
  - ~~mean, variance and standard deviation~~
  - median and quantiles
  - density functions
  - joint distributions
  - correlation and covariance
- Sample and Population
  - representativeness
  - sampling error

Descriptive statistics

Descriptive statistics (Review)

- **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
(Stata also allows shortened format e.g.  sum earn, d)
```

- 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 (Review)

- Variance:

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

- Standard deviation:

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

- To be able to compare across variables, 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 $Q(p)$.
- 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

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

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

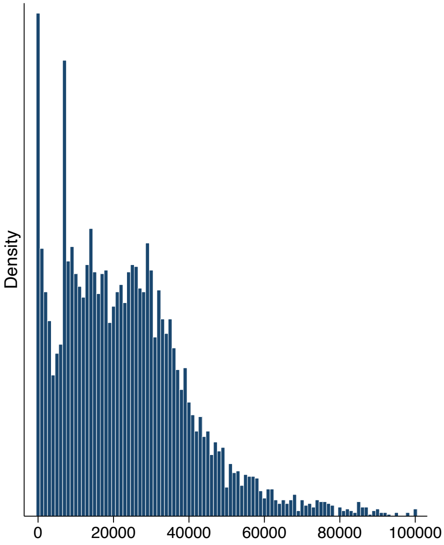
$$f_X(x) \geq 0 \quad \text{and} \quad \sum_x f_X(x) = 1$$

- 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

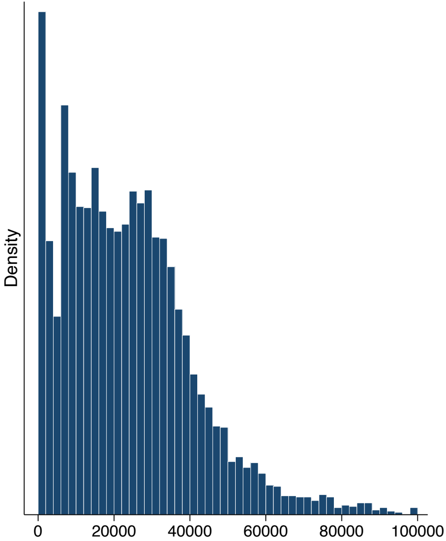
- The empirical counterpart of the density function of a discrete variable is a **histogram**.
  - the height of the bar describes the fraction of observations that take the value x



Source: FLEED teaching data
`hist earn, disc`

Histogram

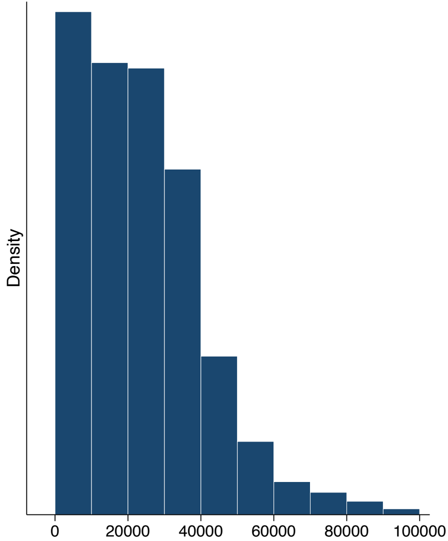
- The empirical counterpart of the density function of a discrete variable is a **histogram**.
    - the height of the bar describes the fraction of observations that take the value x
- More generally: we can divide the observations into **bins** and draw a histogram of them
    - each observation is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

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

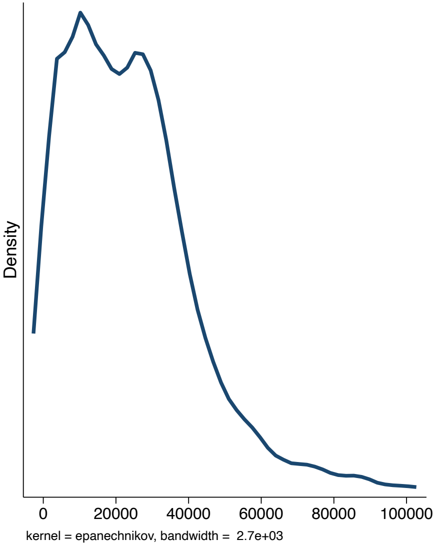
  where $(X = x \pm h/2)$ means the event $x - h/2 \leq X \leq x + h/2$.
- This is the basis for the definition of a **kernel**.

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



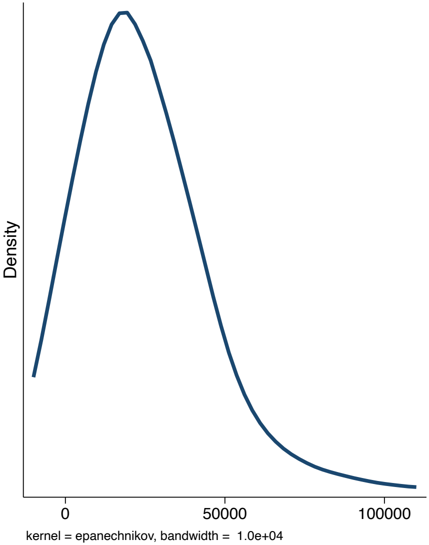
Source: FLEED teaching data
`kdensity earn`

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
  - larger bandwidth disregards more data



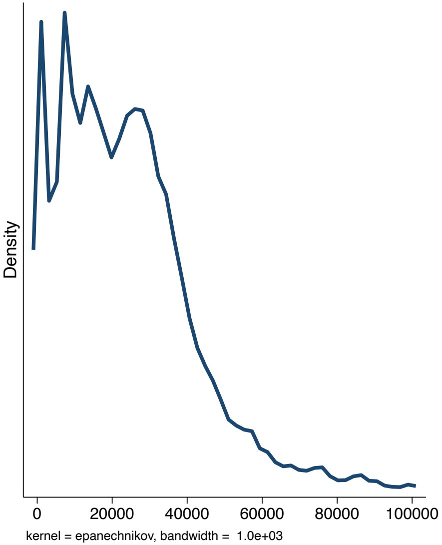
Source: FLEED teaching data
`kdensity earn, bw(10000)`

Kernel density estimator

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

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

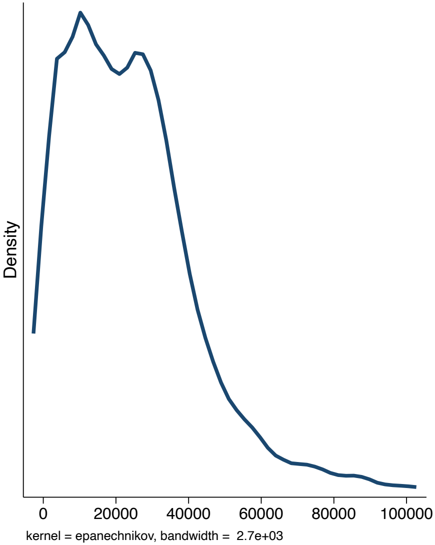
  - **bandwidth** (h): how much data around x is used
  - **kernel function** (K_h): how do we weight observations within the bandwidth, i.e, do observations further away from x get lower weight?
- By default, Stata chooses an "optimal" bandwidth
  - larger bandwidth disregards more data
  - smaller bandwidth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



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

Kernel density estimator

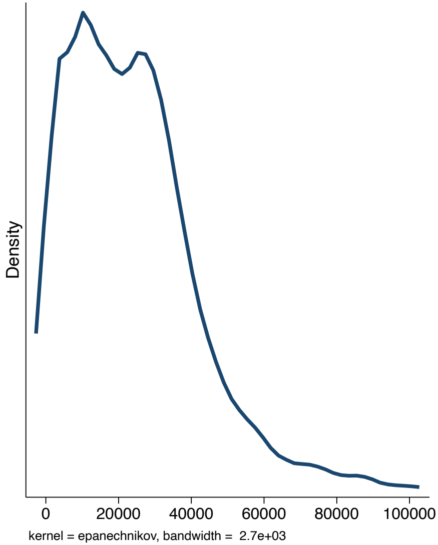
What fraction of the sample has incomes above 40,000?



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

Kernel density estimator

What fraction of the sample has incomes above 40,000?

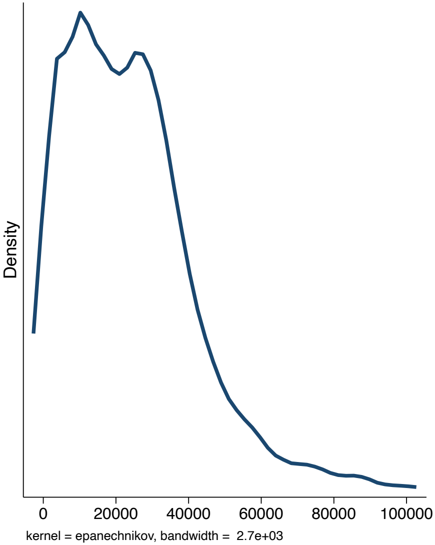


Source: FLEED teaching data
`kdensity earn, bw(1000)`

Kernel density estimator

What fraction of the sample has incomes above 40,000?

Your responses



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

Cumulative density function

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

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

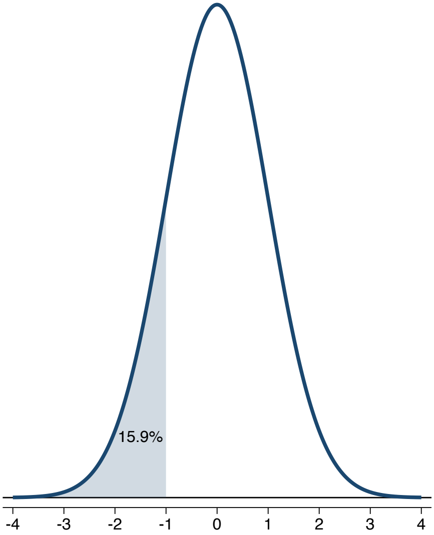
- It answers the question: **what fraction of the observations have values of** x **below** t**?**

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: $F_X(-1) = 0.159$



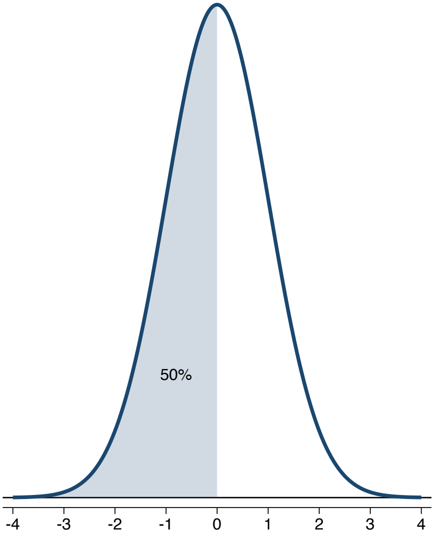
Density function of the standard normal distribution

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:
    $F_X(-1) = 0.159$
    $F_X(0) = 0.5$



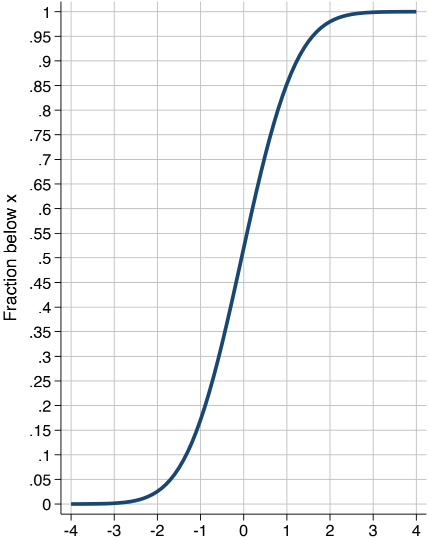
Density function of the standard normal distribution

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

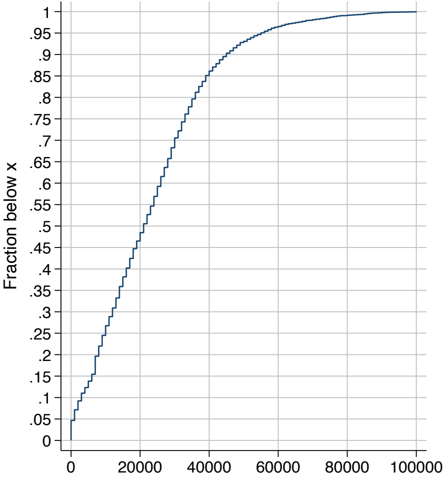


CDF of the standard normal distribution

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, $x = 1000, 2000, \ldots$

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`

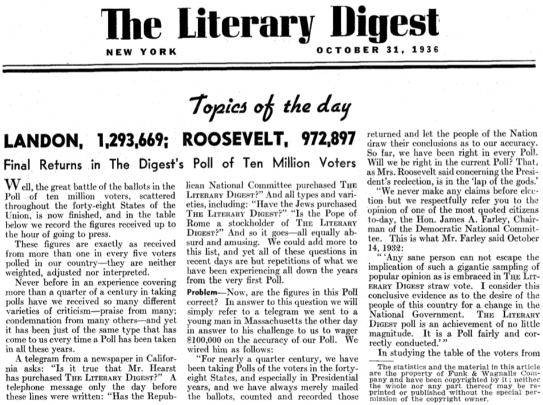
Population and sample

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 for 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 the 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 for in the upcoming election
  - 2.4 million were returned: 57% to Landon and 43% to Roosevelt
  - Roosevelt won the elections 62% to 37%
- George Gallup also conducted 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

Landon Congratulates President, Who Replies

BIG CHARTER VOTE

8-Hour System for Firemen Also Wins Easily

Smith Plans Comment On the Election Today

LEHMAN VOTE CUT

Bleakley Gets a Surprising Total in the City

POLL SETS RECORD

Roosevelt Electoral Vote of 519 Seen as a Minimum

DEMOCRATS RETAIN STATE SENATE LEAD

UNION PARTY VOTE FAR BELOW BOASTS

FRANKLIN D. ROOSEVELT

DEMOCRATS SWEEP ALL PENNSYLVANIA

JERSEY'S 16 VOTES SAFE FOR NEW DEAL

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 ($N \approx 3.5M$)

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

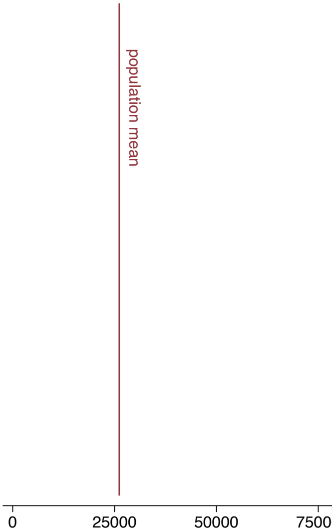
- Suppose we take a random sample of n people from the full-population data and calculate a **sample average**

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

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

Sampling error

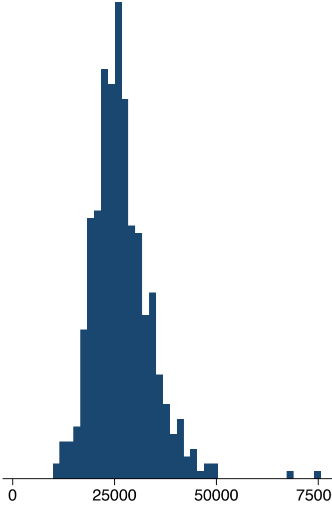
- Let's take many random samples from the full-population using different sample sizes
    -



Source: Statistics Finland's population level r

Sampling error

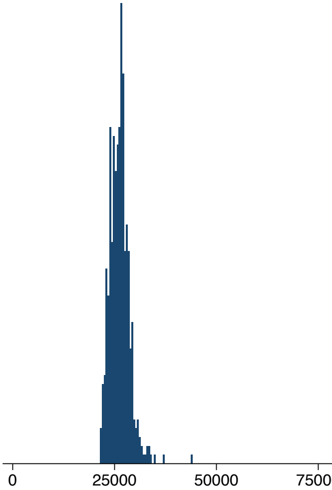
- Let's take many random samples from the full-population using different sample sizes
  - $n = 10$



Source: Statistics Finland's population level r

Sampling error

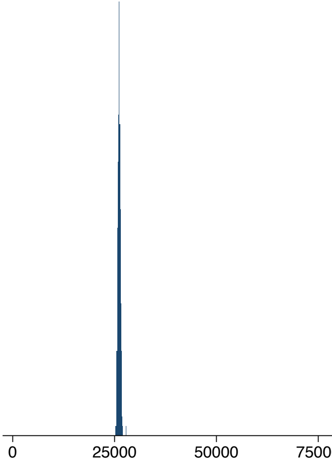
- Let's take many random samples from the full-population using different sample sizes
  - $n = 100$



Source: Statistics Finland's population level r

Sampling error

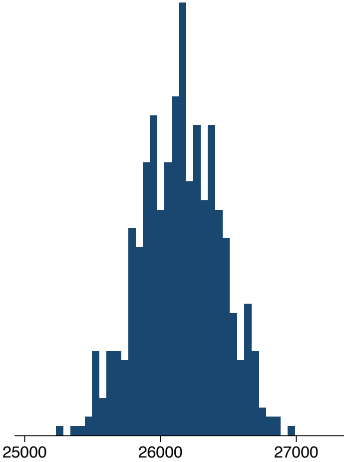
- Let's take many random samples from the full-population using different sample sizes
  - $n = 5,973$



Source: Statistics Finland's population level r

Sampling error

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

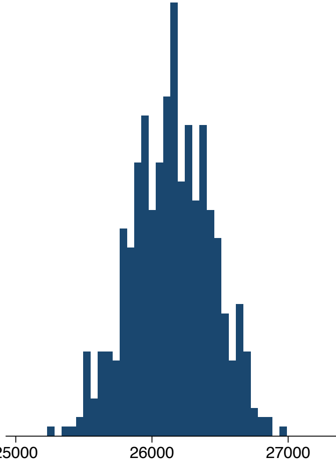


Source: Statistics Finland's population level r

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 r

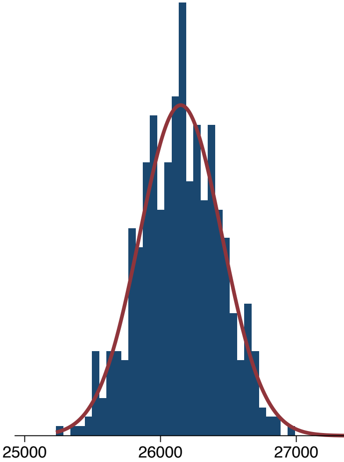
Sampling error

- 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 r

Joint distributions

Cross tabulation

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)

edul	woman 0	1	Total
Less/unknown	1,128	894	2,022
Secodary	1,430	1,313	2,743
Bachelor	439	651	1,090
Master	181	185	366
Lis./PhD	17	6	23
Total	3,195	3,049	6,244

Source: FLEED teaching data
`tabulate edul woman`

Joint distribution

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)
- Alternatively, cross tabulation cells may report the share of observations with value (y, x)

	woman		
edul	0	1	
Less/unknown	18.07	14.32	
Secodary	22.90	21.03	
Bachelor	7.03	10.43	
Master	2.90	2.96	
Lis./PhD	0.27	0.10	
			100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Joint distribution

- A simple, yet efficient way to display (small) data of two variables is **cross tabulation**
  1. the no. rows = no. values that Y can take
  2. the no. columns = no. values that X can take
  3. the cells report no. observations with value (y, x)
- Alternatively, cross tabulation cells may report the share of observations with value (y, x)
- This is the empirical counterpart of the **joint density function**

$$f_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$$

i.e., the probability that random variable X takes the value x *and* that random value Y takes the value y

	woman		
edul	0	1	
Less/unknown	18.07	14.32	
Secodary	22.90	21.03	
Bachelor	7.03	10.43	
Master	2.90	2.96	
Lis./PhD	0.27	0.10	
			100.00

Source: FLEED teaching data
`tabulate edul woman, cell nofreq`

Summary

- Today's lecture was mainly about learning the basic concepts
    1. Concepts you need to know and understand well
        - density function, CDF
        - joint distributions
    2. Things to worry when using samples
        - representativeness
        - sampling error

- Next lecture: Conditional descriptive statistics
    - and apply them to make sense of recent research on inequality

Upcoming assignments

- **In-class worksheet 1** due on MyCourses before next lecture
  - You may upload a photo/scan (preferred)
  - If not, can turn in paper copy in-person beginning of next lecture.
- **Pre-class assignment 2** due 15 minutes before lecture
- **Homework 1** due Jan 17
  - Now you have all the conceptual tools to get started
  - Attend Exercise session 1 tomorrow for practical tools (e.g. Stata)