

# Datasets



Statistical Reasoning  
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 2

# Outline

---

Quantitative data come as **datasets** in particular formats. Your research projects will use **cross-sectional data** in Stata format (**.dta**).

Objectives

Structure

Availability

Exploration

Practice



## Objectives

---

At the early stage of your project, you will look for **cross-sectional data**, which is a **time-consuming task**. Consider yourself done only when the dataset is open and ready for analysis in Stata.

- **Find a dataset**, based on availability and research interests.
- **Download its documentation**, and especially the codebook.
- **Convert and reshape** if necessary (Stata Guide, Section 7).
- **Explore and describe** by writing up a draft do-file.

It is strongly recommended that you have formed pairs and have a Stata dataset ready for exploration **by next week**. The results of your exploratory analysis will later appear in Assignment No. 1.

## Data structure

---

**Cross-sectional data** capture the characteristics of a **sample** of comparable **units** at a **single point** in time:

- **Units** can be individual respondents, states, organizations...
- **Observations** vary by their characteristics, *not* by unit type
- **Sampling** will vary depending on representativity requirements

**Time series** capture **repeated** observations over time of either sampled or nonsampled units:

- **Cross-sectional time series** (CSTS) capture fixed, nonsampled units at different time intervals
- **Longitudinal data** capture sampled units (called cohort or panel) at different time intervals

# Example: Industry Canada File Sharing Survey (2006)

Using individual-level ‘micro’ data on illegal downloading practices among a representative sample of the Canadian population aged 15+:

	id	prov	qregn	date	age	sex	download	q1	q2	q3	q4
1	1065	ON	Ontario	20060502	Less than 25 years old	Male	NON-DOWNLOADER	Yes	20	10	Very strong
2	1129	AB	Alberta	20060423	Less than 25 years old	Female	NON-DOWNLOADER	No	-	-	Somewhat strong
3	1152	QC	Quebec	20060519	Less than 25 years old	Female	DOWNGRADER	No	-	-	Somewhat strong
4	1166	ON	Ontario	20060429	Less than 25 years old	Male	NON-DOWNLOADER	Yes	20	4	Moderate
5	1191	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Yes	20	15	Somewhat strong
6	1214	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Don't Know/Refused	-	-	Very limited
7	1215	QC	Quebec	20060422	Less than 25 years old	Female	NON-DOWNLOADER	Yes	10	6	Very strong
8	1245	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	No	-	-	Very strong
9	1266	BC	British Columbia	20060419	25 years old or more	Female	NON-DOWNLOADER	No	-	-	Very limited
10	1315	QC	Quebec	20060430	25 years old or more	Male	NON-DOWNLOADER	No	-	-	Somewhat limited
11	1317	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Don't Know/Refused	25	5	Very strong
12	1643	ON	Ontario	20060423	25 years old or more	Female	DOWNGRADER	Don't Know/Refused	20	3	Somewhat strong

- **Observations:** rows hold data for a single sampled unit
- **Variables:** columns hold all values for a single variable
- **Missing data:** “Do Not Know / Refused to Answer”, ‘.’
- **Value formats:** numeric, **string**, **encoded** (values/labels)

## Format requirements

---

Check your dataset against the following list:

- The dataset format is **DTA** ([.dta](#)).  
*Otherwise, non-Stata format: [convert](#).*
- The data are available at **only one point in time** (e.g. year 2009).  
*Otherwise, time series: [subset](#).*
- The columns do **not hold time variables** (e.g. y1960, y1961, . . . )  
*Otherwise, 'wide' data: [reshape](#).*

If you need to reformat your dataset before analysing it, all guidelines and operations are detailed in the Stata Guide, Sections 5–8.

## Availability

---

Generally speaking, things are getting much better:

- **“Data Deluge”**: scientists, journalists and organizations share more and more data online.
- **Open formats**: converting, expanding and sharing data through different media is increasingly possible.
- **Visualization projects**: online platforms that support graphical views of data are becoming very popular.

For a few examples, check the *Guardian’s Data Blog*:

<http://www.guardian.co.uk/news/datablog>

More relevant links appear in the README file of the Datasets folder, as well as on the course blog:

<http://srqm.tumblr.com/tagged/data>

# Example: Gapminder



## Recommended sources

---

Start with recommended datasets and repositories:

- **Teaching Pack ▷ Datasets:**

- **ESS 2008:** European Social Survey
  - **WVS 2000:** World Values Survey
  - **QOG 2011:** Quality of Government

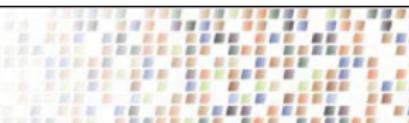
**Note:** consult with us if you are planning to use a dataset that is not from the Teaching Pack.

- **Datasets/README ▷ Datasets:**

- **ICPSR:** <http://www.icpsr.umich.edu/>
  - **CESSDA:** <http://www.cessda.org/>

**Note:** some data sources will ask you to register online before granting you full access.

# Example: ICPSR (topic search: Health)



## Find & Analyze Data

[Find ICPSR Data](#)

[Bibliography of Data-Related Literature](#)

[Variables Database](#)

[Analyze Data Online](#)

[Thematic Collections](#)

[Restricted Data](#)

[Publication-Related Archive](#)

## Study Search Results

If you weren't looking for datasets, you can also perform your query against our [citations database](#) or database of [variables/questions](#). You may also wish to look over our [searching tips](#), or utilize our [subject thesaurus](#) to refine your search.

### Query:

584 results found

Sort by: [Title A-Z](#)

1 - 25 [>](#)

Study No.	Study Title/Investigator	Time Period
9460	<a href="#">ABC News AIDS Poll, June 1990</a> ABC News	1990
9458	<a href="#">ABC News Genetic Engineering Poll, April 1990</a> ABC News	1990
9246	<a href="#">ABC News National Cholesterol Test Poll, April 1989</a> ABC News	1989

# Example: ICPSR (variable search: Presidential)



## Q32: Presidential Promises

Name: Q32

Label Presidential Promises

Question In general, do you think presidents can keep most of the promises they make or is this something most presidents cannot do? IF CANNOT, ASK: Who do you think is mostly to blame for that, the president himself, Congress, special interest groups, or the media? [!oc 1/64][##md1=0][##md2=0][##label= Presidential Promises]

Response Categories:	Code	Label	Freq.	%
	1	Can keep	34	6.7
	2	Cannot, President	4	0.8
	3	Cannot, Congress	37	7.3
	4	Cannot, special interest groups	15	2.9
	5	Cannot, media	13	2.6
	6	Combination (Vol.)	19	3.7
	9	DK/NA	6	1.2
	Missing Data			
	.		381	74.9
	Total			
			NaN	100%

*Disclaimer:* The frequencies for this variable are not weighted. They are purely descriptive and may not be representative of the study population. Please use with caution and consult the study documentation.

Source: This variable was taken from: CBS News Monthly Poll #4, November 2000. You may:

- [View the study home page](#)

# Example: ICPSR (geography search: Russia)



## Find & Analyze Data

[Find ICPSR Data](#)

[Bibliography of Data-Related Literature](#)

[Variables Database](#)

[Analyze Data Online](#)

[Thematic Collections](#)

[Restricted Data](#)

[Publication-Related Archive](#)

## Study Search Results

If you weren't looking for datasets, you can also perform your query against our [citations database](#) or database of [variables/questions](#). You may also wish to look over our [searching tips](#), or utilize our [subject thesaurus](#) to refine your search.

Query: geography:Russia [X](#),

76 results found		Sort by:	Time Period (newest)	1 - 25
Study No.	Study Title/Investigator	Time Period		
28101	<a href="#">Voice of the People End of Year Survey, 2009</a> Gallup International Association			2009
21441	<a href="#">Voice of the People, 2007</a> Gallup International Association			2007
23562	<a href="#">ICT Diffusion and Distribution Dataset, 1990-2007</a> Howard, Philip N.; Busch, Laura; Cohen, Spencer			2007

# Example: CESSDA (keyword search: Legal Systems)



The screenshot shows the CESSDA website interface. On the left, there is a sidebar with a search bar, a British flag icon, and a language dropdown set to English. Below these are various navigation links under the heading "CESSDA Catalogue". The main content area displays a search result for the term "legal systems". The results are categorized under "Study".

**Search Term: **legal systems****

**Study**   **Section**   **Variable**

Study
Multiscopo ISTAT - Security of citizens - 1997 - 1998
Multiscopo ISTAT - Security of citizens - 2002
Confidence in Courts of Justice : District Court Survey 1999
[ICVS] International Crime Victim Survey (ICVS) : Series information
[AZK01] <sup>3</sup> Crime Victim Survey : Slovenia, 2001
Confidence in Courts of Justice : Extraordinary Appeals 1989-1998
Reading Comprehension Among Inmates at Åna County Jail, 1993
Luottamus tuomioistuimiin : käräjäoikeuskysely 1999
Luottamus tuomioistuimiin : ylimääräiset muutoksenhaut 1989 - 1998
[ICVS_97] <sup>3</sup> International Crime Victim Survey (ICVS), 1989-1997

1-10 of 35 | Next ▶

**Top Terms:**

- ↪ LAW, CRIME AND LEGAL SYSTEMS

**Broader Terms:**

- ↪ LAW, CRIME AND LEGAL SYSTEMS

## Source requirements

---

You need to be able to reference your dataset in full before use. This implies collecting information on:

- **the source**, with its full name and online address
  - e.g. *World Health Organization (WHO)*, [website]
- **the unit of analysis**, with its restrictions
  - e.g. *American adult resident population with U.S. citizenship*
- **the sampling strategy**, with a reference if needed
  - e.g. “*cf. ‘Sample Design’ in the Survey Description [source]*”
- **the total number of observations**
  - e.g.  $N = 1,524$

**Note:** all these characteristics need to appear in your research project.

## Exploring the documentation

---

**Knowing the data is not an option.** You naturally do not have to 'read' through the data itself, but **you need to read everything else.**

The **codebook** is essential to **measurement**:

- Data collection and measurement are publicly documented to allow for sceptical scrutiny of sources and method.
- The unit of continuous data, scale of ordinal data or categories of nominal data are given with their construction notes.

## *Example: Measurement*

---

### **World Development Indicators**

<http://go.worldbank.org/U0FSM7AQ40>

**wdi\_fr              Fertility rate (births per woman)**

(Time-series: 1960-2007, n: 4986, N: 187,  $\bar{N}$ : 104,  $\bar{T}$ : 27)

(Cross-section: 1999-2002 (varies by country), N: 186)

Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates. Source: World Bank staff estimates from various sources including census reports, the United Nations Population Division's World Population Prospects, national statistical offices, household surveys conducted by national agencies, and Macro International

## Exploring the identifiers

---

**Data comes with identifiers:** variables have names and labels, and their values can also carry a label.

Identifiers are essential to writing commands:

- Passing commands requires correct variable names, labels and values, as in `su rlgdgr if rlgdnm==6`.
- Excluding missing values frequently requires recoding them to the Stata `.` symbol(s) and passing the `if !mi()` selector.

## *Example: Coding*

---

# dscrgrp: Member of a group discriminated against in this country	
Question	All rounds: Would you describe yourself as being a member of a group that is discriminated against in this country?
Question number	ESS1, ESS2: C 16 ESS3, ESS4: C 24
Comments	ESS2: Slovenia: Coding error. All respondents with code 2 in C16 (DSCRGRP) have got code 1 in C17 (DSCRREF).
Value Label	
1	Yes
2	No
7	Refusal
8	Don't know
9	No answer

# Selections

---

Stata offers two ways to select observations:

- **Range selection** with `in`:

- All observations have a row number `n` between 1 and sample size `N`.  
The number is arbitrary and non statistically meaningful.
  - Range selection is principally useful to look at a few observations.  
e.g. `list in 1/10, list in -25/1`

- **Logical selection** with `if`:

- “equal to” (`==`) or “not equal to” (`!=`)
  - “greater/lesser than” (`>/<`) and “... or equal to” (`>=/<=`)
  - “and” (`&`), “or” (`|`)
  - “missing” (`mi()`) or “nonmissing” (`!mi()`)

Logical operators apply to virtually all data operations.

## Examples

---

Think of selecting observations as formulating linguistic statements:

- **Identification**, e.g. “I am not Sidney Poitier”

```
drop if name=="Sidney Poitier"
```

- **Validity**, e.g. “Raise your hand if you’re absent”

```
gen vote=1 if absent==1
```

- **Conditions**, e.g. “Twist and Shout”

```
replace beatles=1 if !mi(twist) & !mi(shout)
```

That logic allows to run commands on particular groups:

- `drop if mi(age) | age < 65` means

“drop all observations where age is missing or under 65.”

- `li country if gdp >= 5000 & !mi(gdp)`” means

“list country if GDP is above or equal to 5,000 and nonmissing.”

## Practice: Body Mass Index

---

$$\text{BMI} = \frac{\text{mass (kg)}}{(\text{height(m)})^2} = \frac{\text{mass (lb)} \times 703}{(\text{height(in)})^2}$$

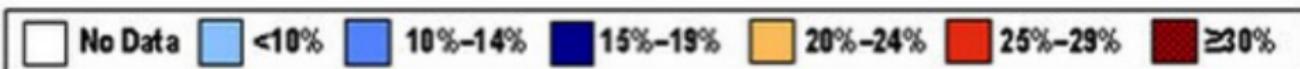
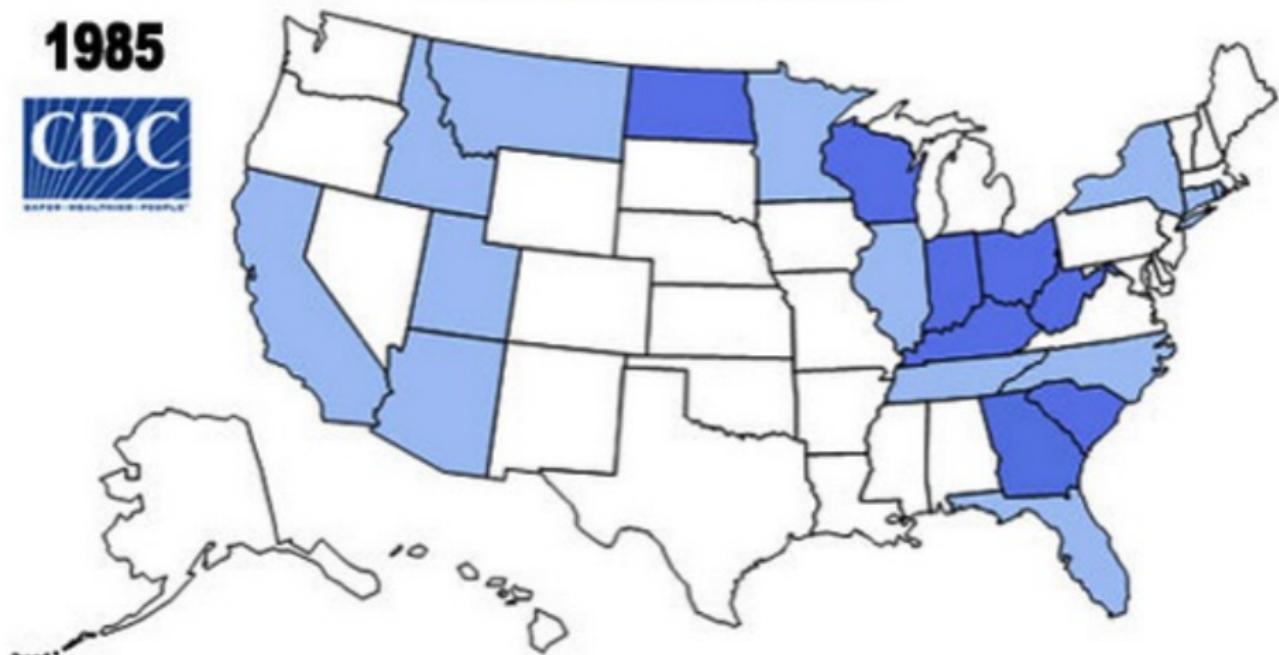
- For **normal weight** adults,  $18.5 < \text{BMI} < 25$ .
- For **overweight** adults,  $25 \leq \text{BMI} < 30$ .
- For **obese** adults,  $\text{BMI} \geq 30$ .
  
- National Health Interview Survey (NHIS)
- Sample: U.S. adult population, 1997–2009



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

[<previous](#) [next>](#) play stop

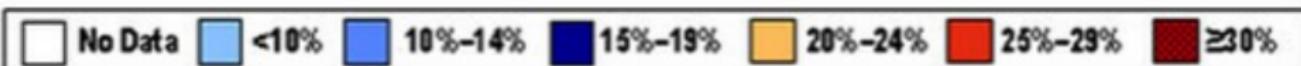
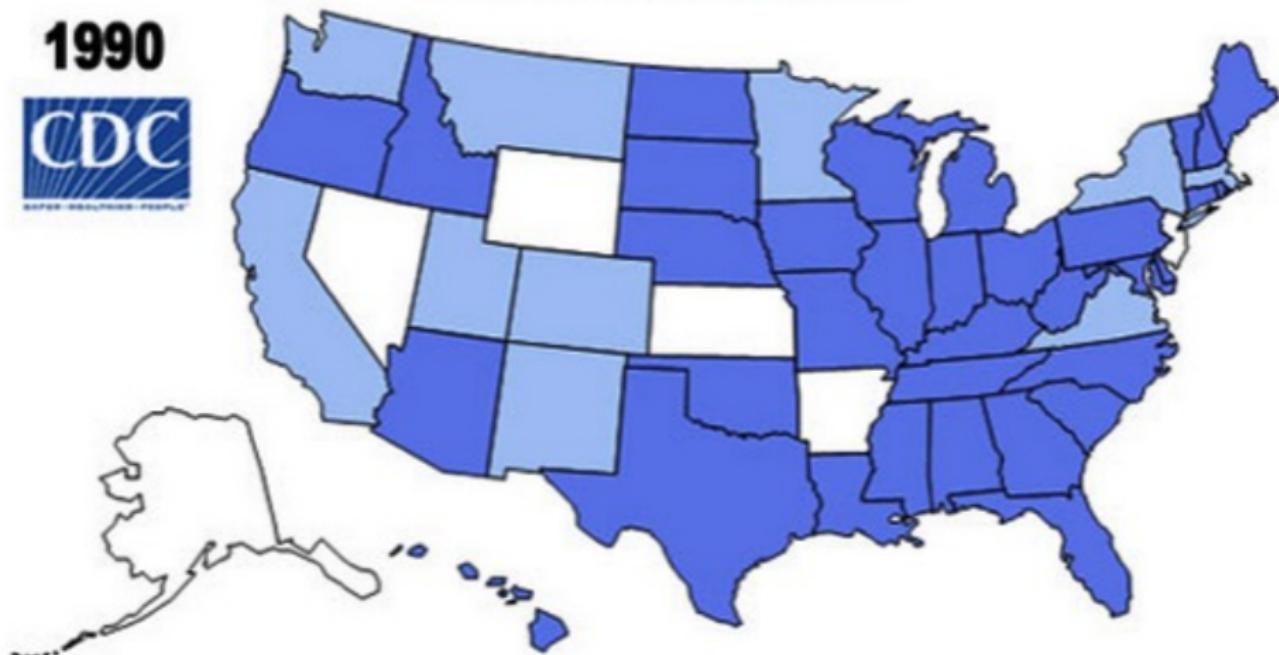
1985



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

[<previous](#) [next>](#) play stop

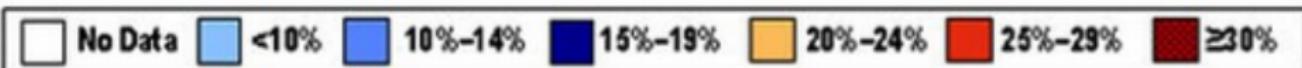
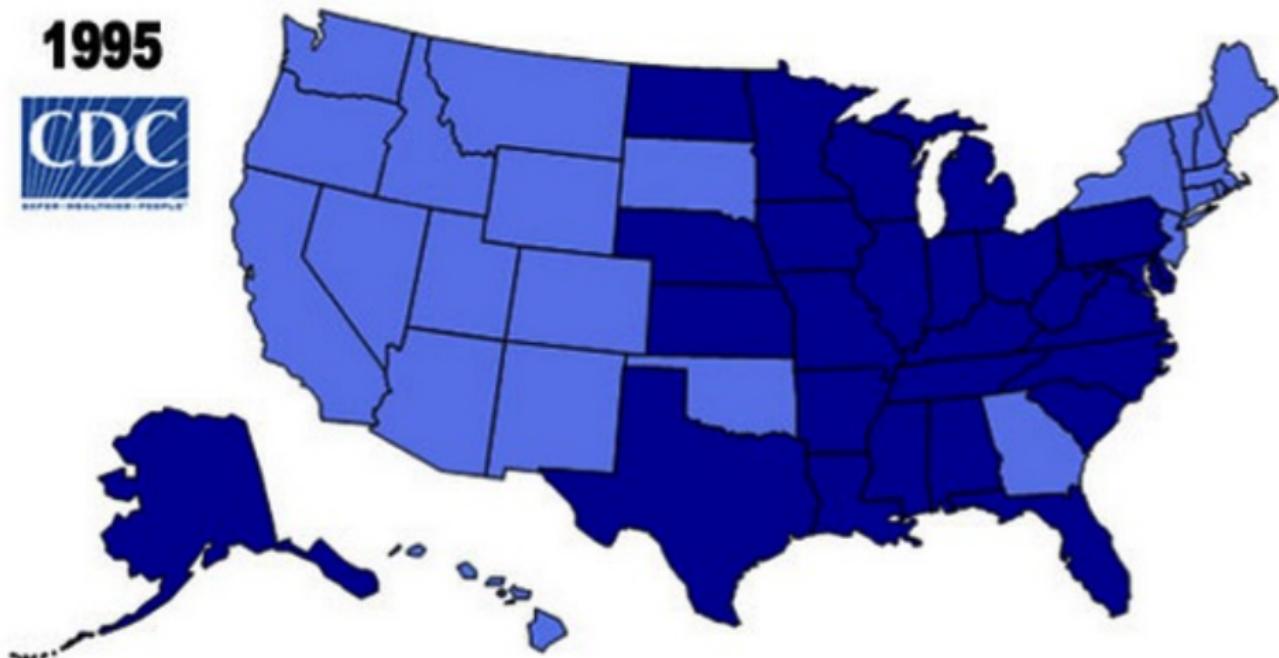
1990



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

[<previous](#) [next>](#) play stop

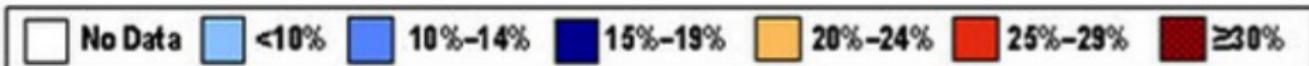
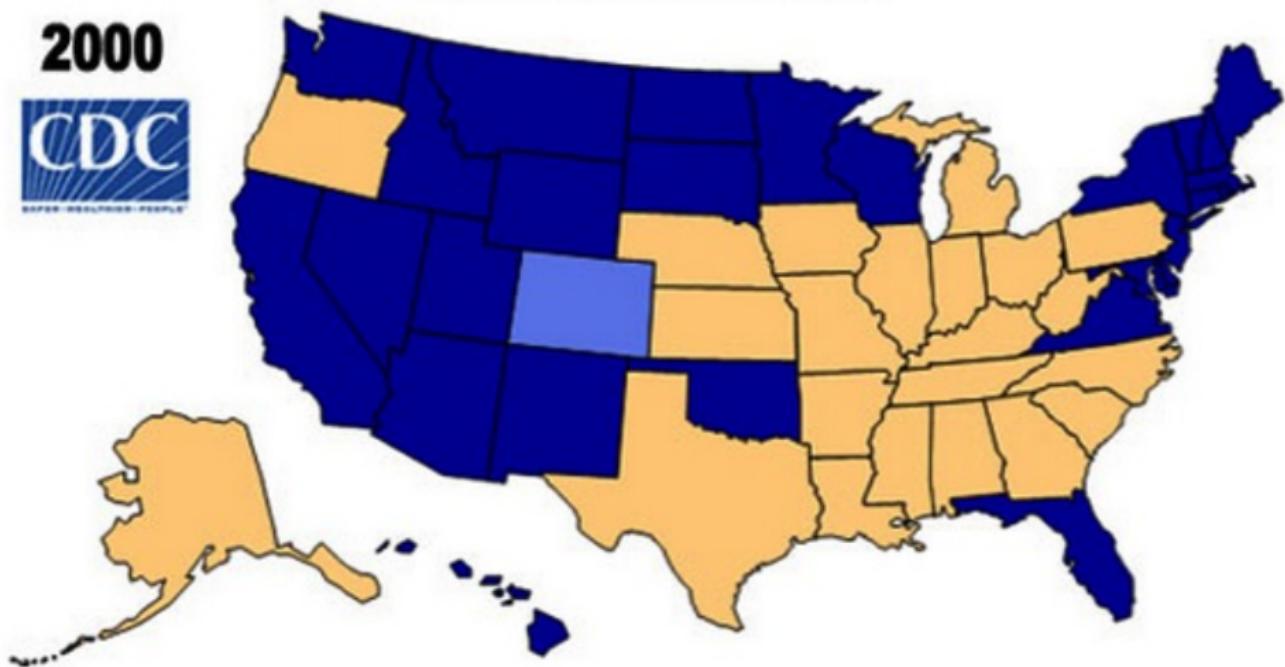
1995



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

[<previous](#) [next>](#) play stop

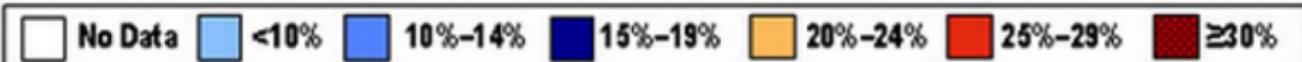
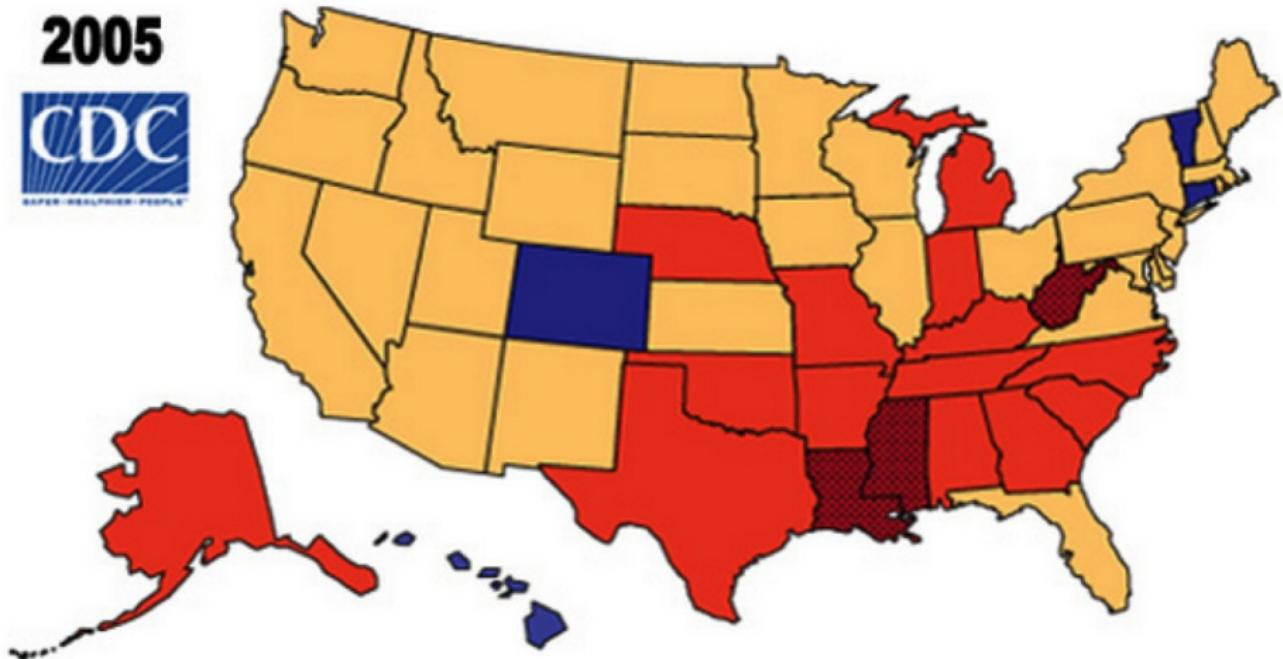
2000



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

[<previous](#) [next>](#) play stop

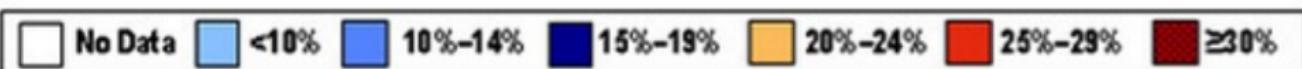
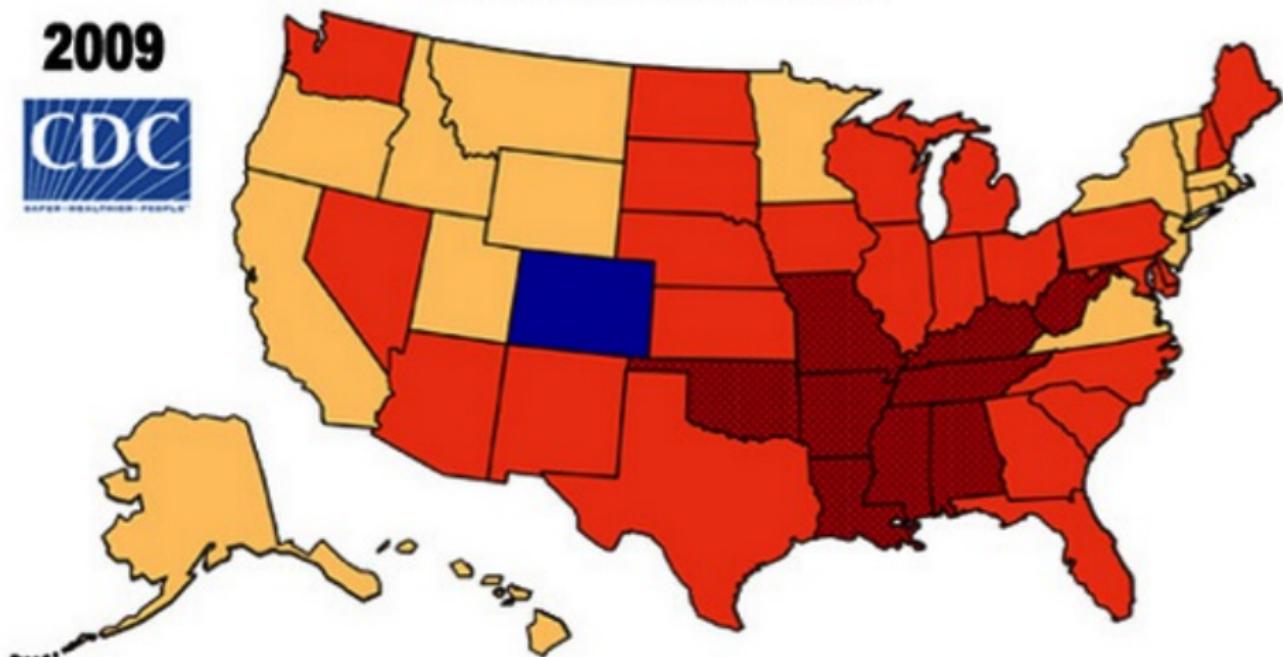
**2005**



# Percent of Obese (BMI $\geq 30$ ) in U.S. Adults

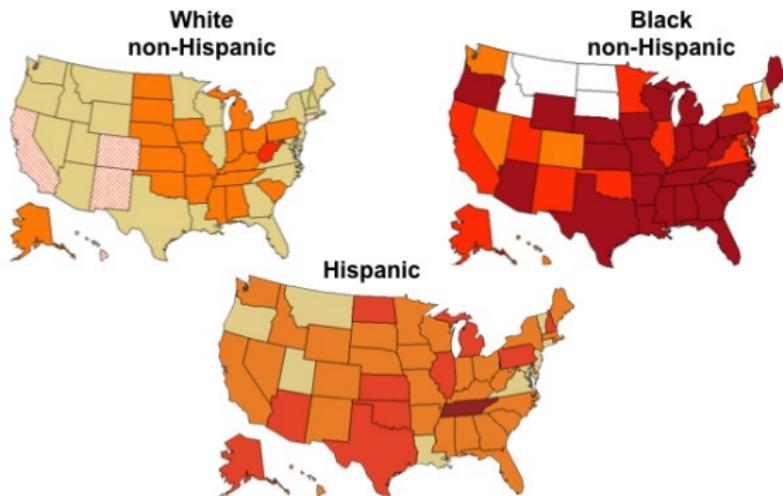
[<previous](#) [next>](#) play stop

2009



## Another dimension of the issue

---



State-specific Prevalence of Obesity (BMI  $\geq 30$ ) Among U.S. Adults, by Race/Ethnicity, 2006–2008. Source: CDC.

## Prologue

---

Set up Stata if you have not done so yet:

```
set mem 500m // if Stata 11-
set more off // add , perm on
```

Now select your main SRQM folder as the **working directory** by adapting this command to your system and personal folder hierarchy:

```
cd "/Users/fr/Documents/SRQM/"
```

Finally, log the session and open the NHIS dataset:

```
log using "Replication/week2.log", replace
use "Datasets/nhis2009.dta", clear
```

## Data exploration

---

Our first step verifies whether the survey is cross-sectional. If we find that the data spans over several years, we will suppress observations for all but one year of data.

\* List all variables in the dataset.

`describe`

\* Check whether the survey is cross-sectional.

`tab year`

\* Delete all observations except for one survey year.

`drop if year != 2009`

\* Locate variables of interest.

`lookfor height weight`

\* List their values for the first ten observations.

`list height weight in 1/10`

## Variable transformation

---

Our next step is to compute the Body Mass Index for each observation in the dataset (i.e. for each respondent to the survey) from their height and weight by using the `height` and `weight` variables, and the formula for BMI.

```
* Create the Body Mass Index from height and weight.  
gen bmi = weight*703/(height^2)  
* Add a description label to the variable.  
label variable bmi "Body Mass Index"  
describe bmi  
* List a few values.  
list bmi in 50/60  
list bmi in -10/1
```

## Summary statistics

---

We now turn to analysing the newly created `bmi` variable, using the `summarize` command (shorthand `su`) to obtain its mean, min and max values, as well as standard deviation, which we will cover later on.

```
su bmi
```

\* Add the ‘detail’ option for precise statistics.

```
su bmi, detail
```

\* Create a histogram for the distribution of BMI.

```
histogram bmi, normal name(bmi, replace)
```

The **histogram** describes the **distribution** of the variable in the sample, i.e. the distribution of different values of BMI among the respondents to the survey. The `freq` option specifies to use percentages; the `normal` option overlays a normal distribution to the histogram bars; and the `name` option temporarily saves the graph.

## Independent variables

---

Body Mass Index is our **dependent** variable, i.e. the one that we want to explain. We have reason to believe that some **independent** variables like gender, health status and race could be influencing BMI.

`lookfor sex health race`

\* Summarize BMI for each value of ‘sex’.

`bysort sex: su bmi height weight`

\* Read the frequencies for the ‘health’ variable.

`fre health`

\* Summarize BMI for each value of ‘health’.

`bysort health: su bmi weight`

\* Graph the mean BMI of each ethnic group.

`gr dot bmi, over(raceb) ytitle("Average Body Mass Index") name(bmi_race, replace)`

## Note: Logical operators

---

```
drop if year != 2009
```

This command deletes all observations for which the variable `year` is **different** (`!=`) from 2009. An equivalent command would be:

```
keep if year==2009
```

This command keeps only observations for which the `year` variable is **equal** (`==`) to 2009. Notice that the “equal to” operator in Stata is a double equal sign (`==`).

```
su bmi weight if age >= 18 & age < 25
```

This command reads as: “run the `summarize` command on the `bmi` variable for observations with an `age` value **greater than or equal to 18 and (&) lesser than** 25.” We will learn logical operators shortly.

## Note: Graph options

---

```
gr dot bmi, over(raceb) ytitle("Average Body Mass  
Index") name(bmi_race, replace)
```

- `over(raceb)` creates a line and a dot at the mean value of BMI for each category of the `raceb` variable.
- `ytitle("Average... Index")` provides a legible title for the axis on which BMI appears.
- `name(bmi_race, replace)` names the graph `bmi_race` and keeps it in memory; it will `replace` any previous graph with that name.

You might object:

*“So many commands! So many options! This is madness!”*

But no...

**THIS  
IS  
STATA**

A scene from the movie 300 featuring Leonidas (Gerard Butler) shouting in front of his troops. A large black 'X' is drawn across the left side of the frame, covering the first two lines of text.