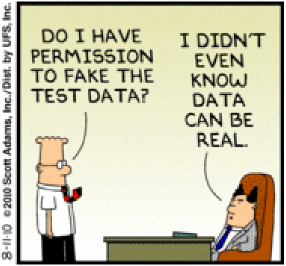
DATASETS

1. Data sources
2. Data structure
3. Data exploration
4. Practice


DO I HAVE PERMISSION TO FAKE THE TEST DATA?
I DIDN'T EVEN KNOW DATA CAN BE REAL.
© 2010 Scott Adams, Inc./Dist. by UFS, Inc.
8-11-10

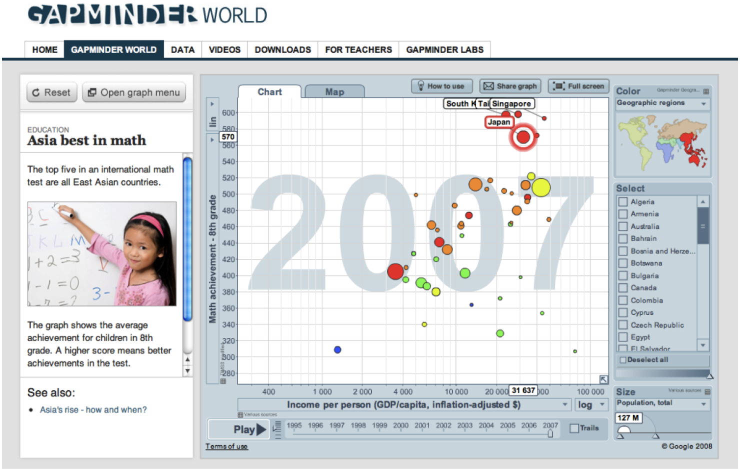
The New York Times

February 4, 2013

The Philosophy of Data

By DAVID BROOKS

If you asked me to describe the rising philosophy of the day, I'd say it is data-ism. We now have the ability to gather huge amounts of data. This ability seems to carry with it certain cultural assumptions — that everything that can be measured should be measured; that data is a transparent and reliable lens that allows us to filter out emotionalism and ideology; that data will help us do remarkable things — like foretell the future.



EDUCATION

Asia best in math

The top five in an international math test are all East Asian countries.

The graph shows the average achievement for children in 8th grade. A higher score means better achievements in the test.

See also:

- Asia's rise - how and when?

See also: “Hans Rosling: Stats that reshape your worldview” (TED, 2006) cessda Council of European Social Science Data Archives

ICPSR INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

NSD TheMacroDataGuide – An International Social Science Resource

DATA.GOV EMPOWERING PEOPLE

DATA.GOV.UK Beta Opening up Government

data.gouv.fr BETA INNOVATION TRANSPARENCE . OUVERTURE

ckan *The open source data portal software*

https://github.com/briatte/srqm/wiki/data

Data structures

Cross-sectional data

- **Comparable units** sampled over a single time period
- **Units** can be individual respondents, countries, firms, …
- **Observations** vary by their characteristics, *not* by unit type

Time series

- **Repeated observations** over time, pooled or sampled
- **Cross-sectional time series** (CSTS): fixed, nonsampled units
- **Longitudinal data**: e.g. cohorts of patients, stocks, voters, …

Sample characteristics

Survey methodology

- Target and survey population
- Sampling frame and **randomization**
- Standardized questionnaires

Issues in representativeness

- Undercoverage
- Unit nonresponse
- Postsurvey adjustments (**sampling weights**)

Documentation: codebooks

Contents

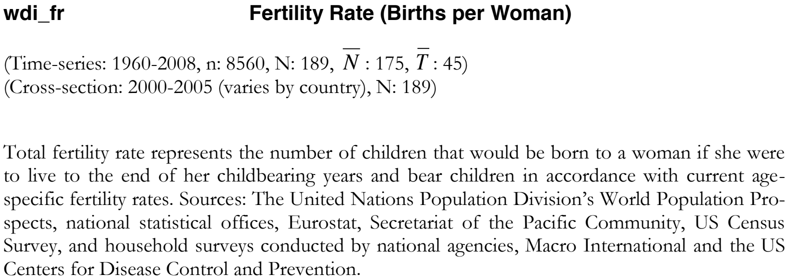
- **Definitions:** unit of analysis, questionnaire, measurements…
- **Survey design:** sampling strategy, time period, weights…
- **Referencing:** authors, affiliations, bibliographic citation

Important

Knowing the data in depth is never an option: the data is never better than your knowledge of it. “Garbage In, Garbage Out.”

Example: Quality of Government

Codebook p. 229:



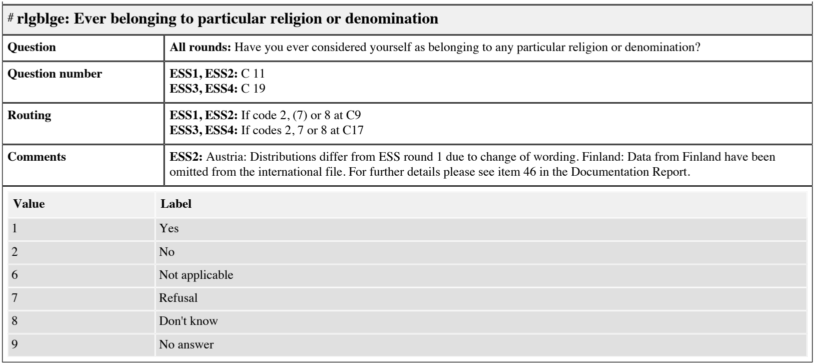
wdi_fr **Fertility Rate (Births per Woman)**

(Time-series: 1960-2008, n: 8560, N: 189, $\overline{N}$: 175, $\overline{T}$: 45)
(Cross-section: 2000-2005 (varies by country), N: 189)

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. Sources: The United Nations Population Division's World Population Prospects, national statistical offices, Eurostat, Secretariat of the Pacific Community, US Census Survey, and household surveys conducted by national agencies, Macro International and the US Centers for Disease Control and Prevention.

Example: European Social Survey

Codebook p. 191:



# rlgblge: Ever belonging to particular religion or denomination

Question	**All rounds:** Have you ever considered yourself as belonging to any particular religion or denomination?
Question number	**ESS1, ESS2:** C 11 **ESS3, ESS4:** C 19
Routing	**ESS1, ESS2:** If code 2, (7) or 8 at C9 **ESS3, ESS4:** If codes 2, 7 or 8 at C17
Comments	**ESS2:** Austria: Distributions differ from ESS round 1 due to change of wording. Finland: Data from Finland have been omitted from the international file. For further details please see item 46 in the Documentation Report.

Value	Label
1	Yes
2	No
6	Not applicable
7	Refusal
8	Don't know
9	No answer

Formatting

Requirements — Stata Guide, Sections 5–8

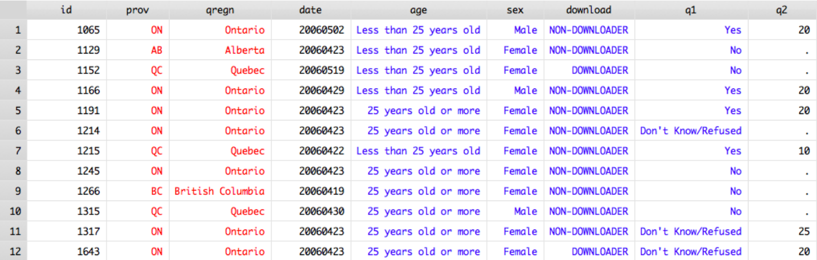
- The dataset format is **DTA** … otherwise convert
- The data is **cross-sectional** … otherwise subset
- The columns hold **only variables** … otherwise reshape

Course datasets — `SRQM/data` folder

- All files are preprocessed `.dta`
- `gss2010` and `nhis2009` hold several years
- See the `README` file for details

Example: Industry Canada File Sharing Survey, 2006

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



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

- **Layout:** one observation per row, one variable per column
- **Formats:** `numeric`, `string`, encoded (values/labels)
- **Missing data:** encoded as `.` (dot), interpreted as $+\infty$

Data exploration

Open and describe

```
* Load a dataset; -clear- wipes previous data in memory.
use data/ess2008, clear
* Describe all variables (names and variable labels).
describe
* Describe a few variables (command shorthand: -d-).
d gndr agea edu* trst*
```

Search for variables

```
* Look for keywords in variable names and labels.
lookfor democ health
* The -lookfor_all- package searches across datasets.
lookfor_all democ health, dir(data)
```

Selection by range

Browsing and counting

```
* All observations have a row number _n from 1 to _N.
di _N
* List first ten observations.
li in 1/10
* View entire dataset; NEVER modify the data by hand.
browse
* Establish sample size ‘N’.
count
```

Reminder

Use `help` when you need more details on a command.

Selection by conditions

Logical operators

```
* Count French and German respondents aged 18 to 24.
count if (agea >= 18 & agea < 24) & ///
  (cntry == "FR" | cntry == "DE")
* Keep observations for a selection of countries.
keep if inlist(cntry, "FR", "DE", "IT", "SP")
```

Missing values

```
* Delete data if missing age, sex or marital status.
drop if mi(agea, gndr, maritala)
* Create a married ‘dummy’ variable when applicable.
gen married = (maritala == 1) if !mi(maritala)
```

Data preparation

1. Weight and subset

```
* Set up survey weights.
svyset [pw=dweight]
* Keep only some observations.
keep if cntry == ”FR”
```

2. Subset and count

```
* Study the missing values for a set of variables.
misstable pat gndr agea edulvla trstep
* Drop observations with missing data.
drop if mi(gndr, agea, edulvla, trstep)
* Get final sample size.
count
```

Practice: NHIS dataset

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

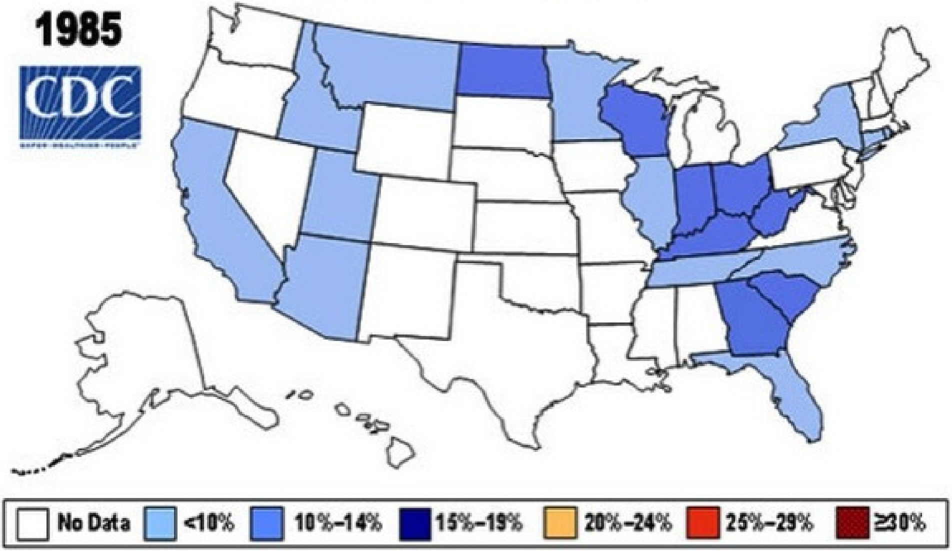
- For **normal weight** adults, $18.5 < \text{BMI} < 25$.
- For **overweight** adults, $25 \leq \text{BMI} < 30$.
- For **obese** adults, $\text{BMI} \geq 30$.

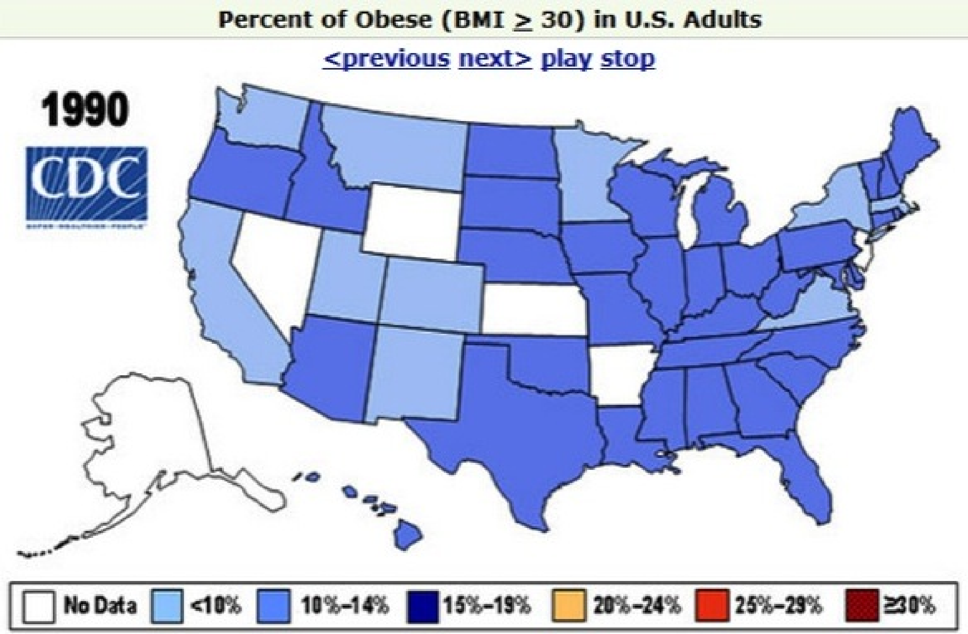
Data:

- National Health Interview Survey (NHIS)
- Sample: U.S. adult population, 2009

Percent of Obese (BMI ≥ 30) in U.S. Adults

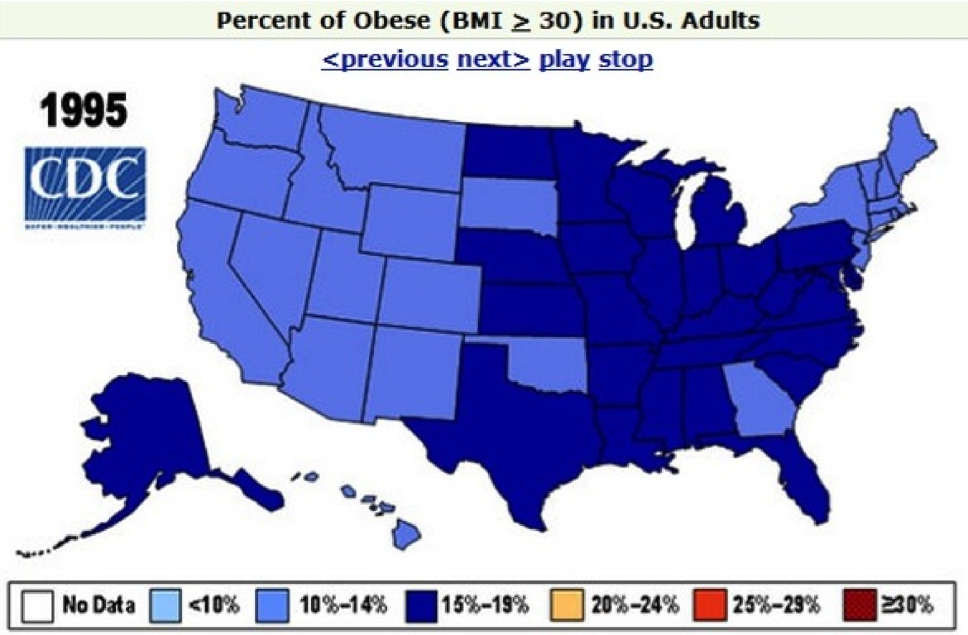
<previous next> play stop



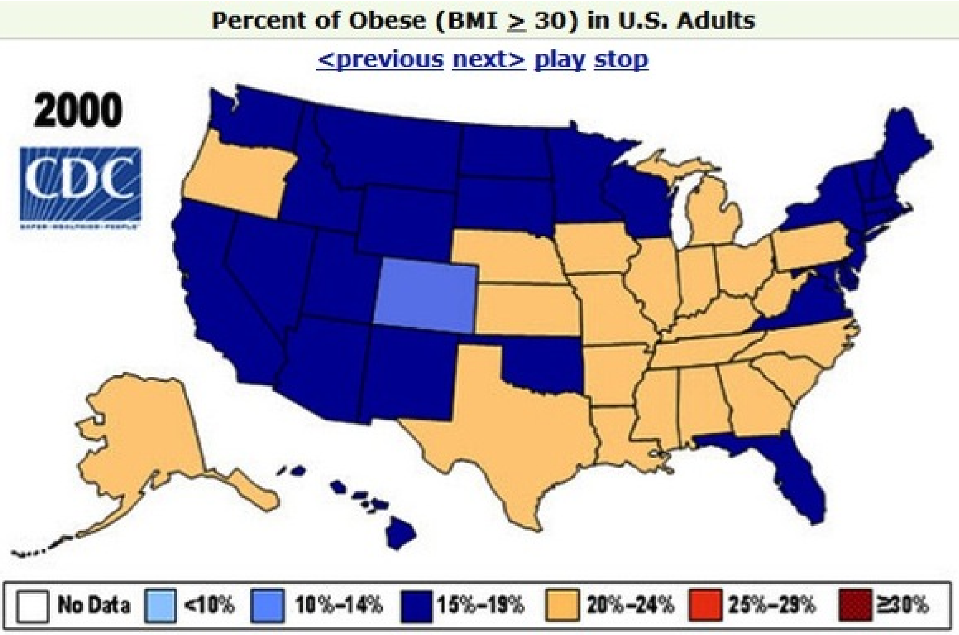


Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop

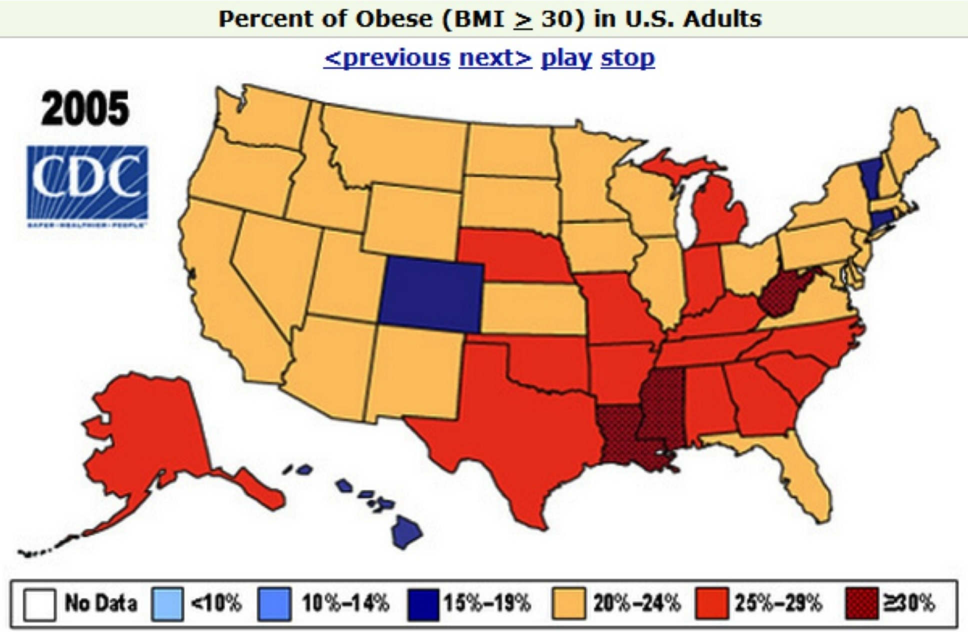


Percent of Obese (BMI ≥ 30) in U.S. Adults

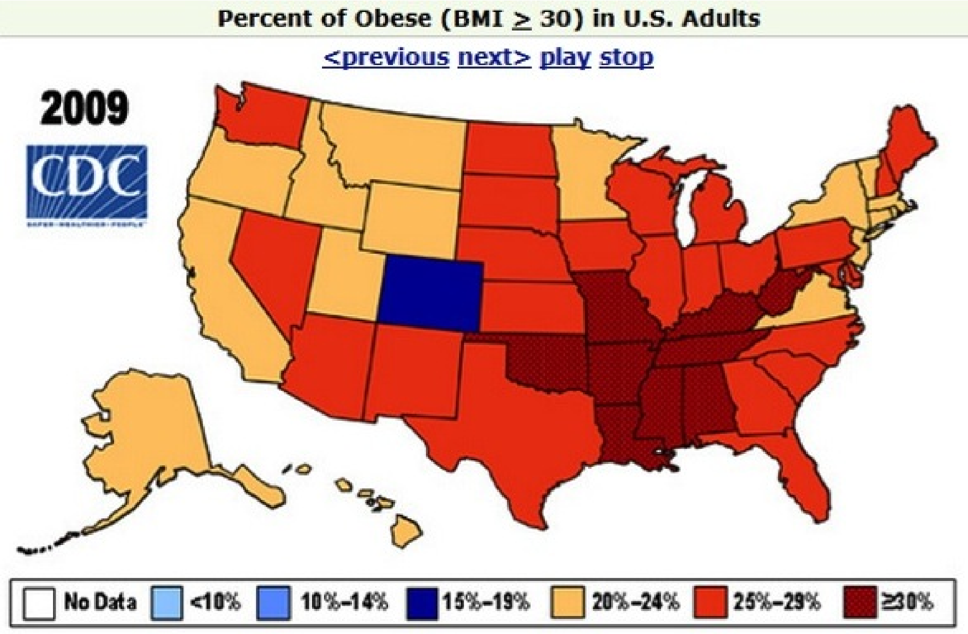


Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop

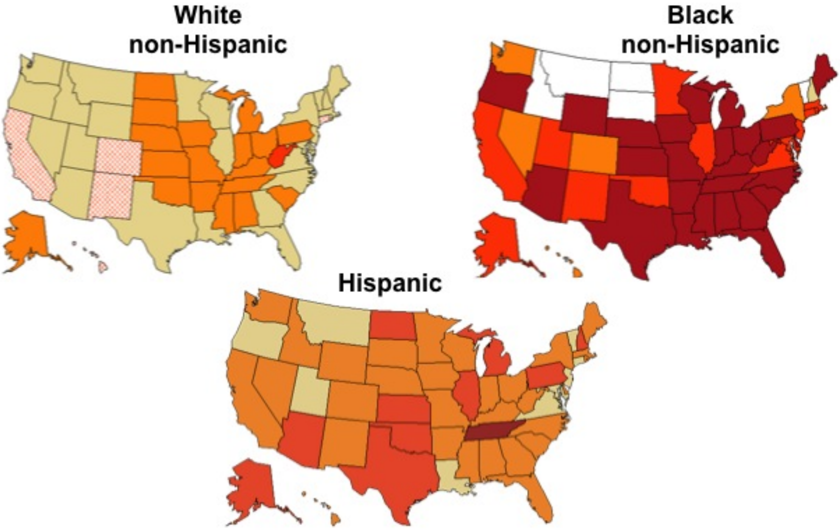


Percent of Obese (BMI ≥ 30) in U.S. Adults



Percent of Obese (BMI ≥ 30) in U.S. Adults

Another dimension of the issue



White non-Hispanic

Black non-Hispanic

Hispanic

Practice session

Class

```
* Get the do-file for this week.
srqm fetch week2.do
* Open to read and replicate.
doedit code/week2
```

Coursework

- Finish the do-file and read all comments at home.
- Read from the codebooks in the `data/` folder.
- Start writing some draft code to describe a dataset.

Exercises

Ex 2.1. European Social Survey 2008

1. Load the data.
2. Find all variables on discrimination.
3. How many countries are there in the dataset?

Ex. 2.2. Quality of Government 2011

1. Load the data.
2. Find all variables on corruption.
3. Which one has the most observations?