WHAT IS DATA?

UF UNIVERSITY of FLORIDA

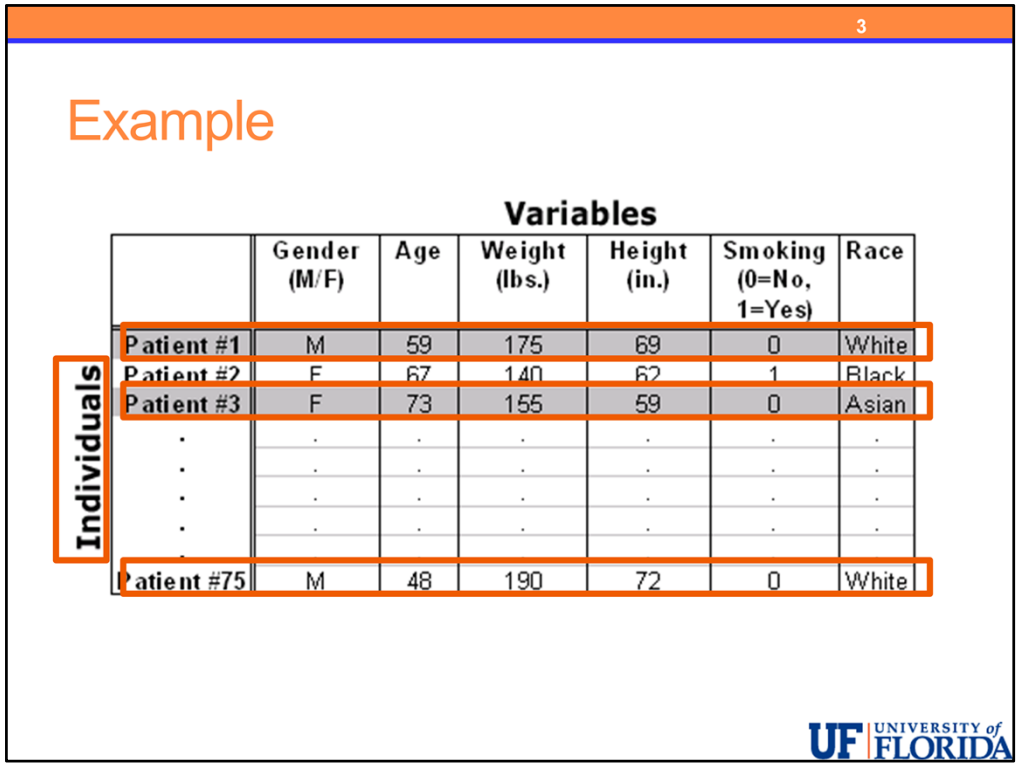
Example

Variables

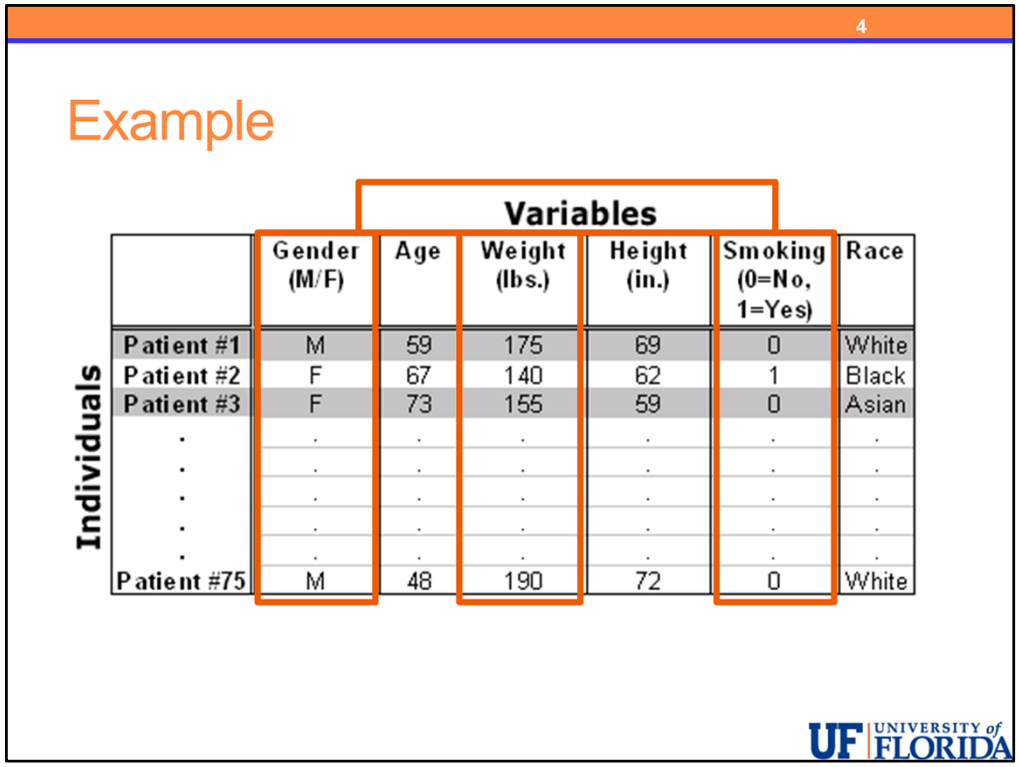
Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (0=No, 1=Yes)	Race
Patient #1	M	59	175	69	0	White
Patient #2	F	67	140	62	1	Black
Patient #3	F	73	155	59	0	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	190	72	0	White

UF UNIVERSITY of FLORIDA

Datasets typically look something like this.



Data consists of information on INDIVIDUALS



The information on each individual is organized into variables which measure or categorize particular attributes or characteristics of each individual.

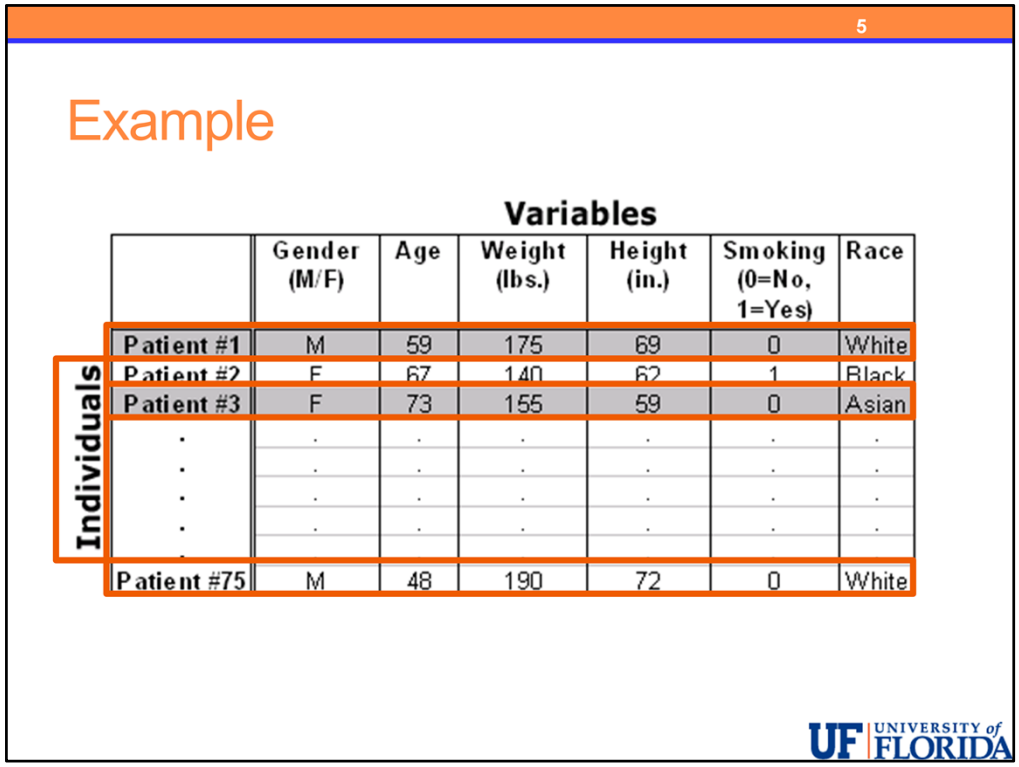
In studies on people, we often collect information on demographic variables such as age, gender, race/ethnicity, socioeconomic status, marital status, income, and so on.

We will classify variables in two main ways in this course, by their role (which we will briefly mention now) and their type (which we will discuss next)

The variable we wish to predict, explain, or study is commonly called the response variable, the outcome variable, or the dependent variable.

Any variable we are using to predict or explain differences in the response variable is commonly called an explanatory variable, an independent variable, a predictor variable, or a covariate.

Recognizing whether the role of a variable is as the explanatory or response variable will be an important skill in this courses.



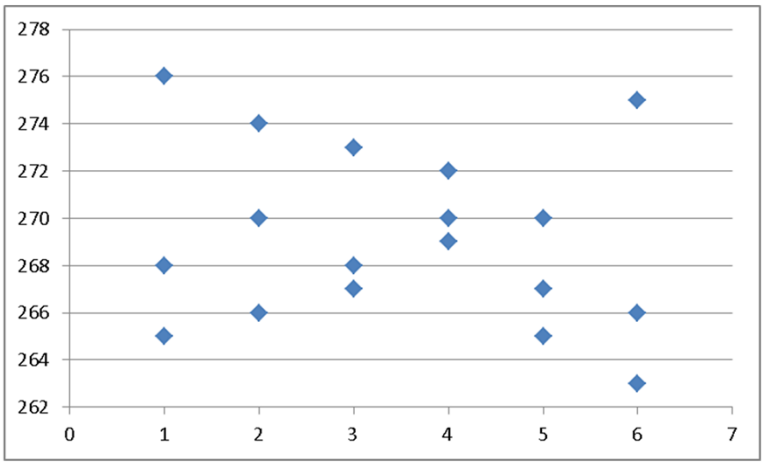
Let's return to the INDIVIDUALS in a dataset…

Some other terms for the individuals in our dataset that are commonly used are

Observations, cases, subjects, experimental units, or a description that is specific to the individuals themselves such as patients, students, batteries, hearing aids, cities, counties, countries, etc.

In our course, unless otherwise discussed and specified, an important assumption is that the individuals in our dataset are independent of each other. The best way to secure this independence is through random sampling, which we will discuss in more detail later in the course.

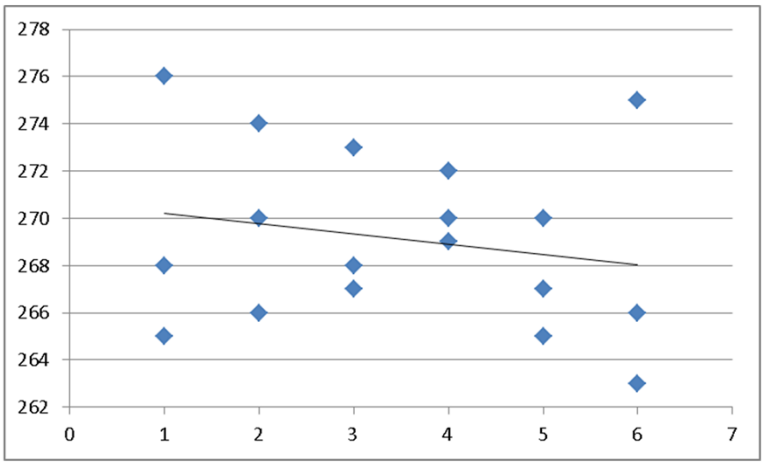
Hazards of Hidden Dependence



UF UNIVERSITY of FLORIDA

Here is a scatterplot

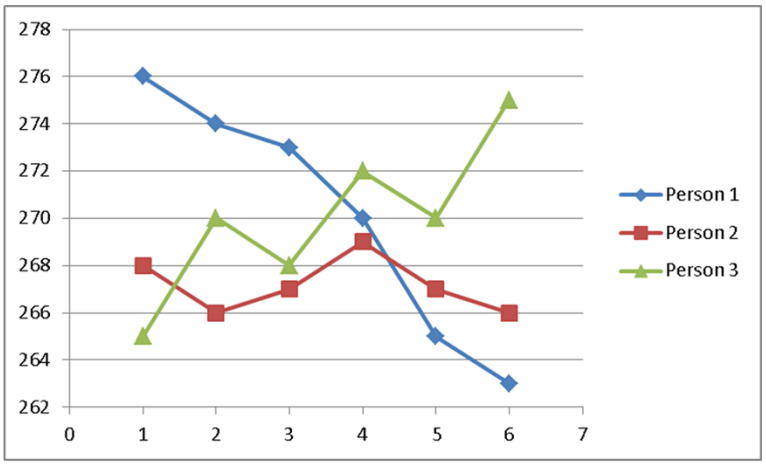
Hazards of Hidden Dependence



UF UNIVERSITY of FLORIDA

This is the overall trend in this data as it is displayed

Hazards of Hidden Dependence



UF UNIVERSITY of FLORIDA

Once we visualize the dependence in the data, in this case that each individual was measured over time, different patterns are revealed.

If we analyzed this data ignoring the information about which individual corresponds to each measurement at each time point we would be ignoring this dependence, which is usually not the most appropriate analysis!

Hazards of Hidden Dependence

Person	Time	Weight
1	1	276
1	2	274
1	3	273
1	4	270
1	5	265
1	6	263
2	1	268
2	2	266
2	3	267
2	4	269
2	5	267
2	6	266
3	1	265
3	2	270
3	3	268
3	4	272
3	5	270
3	6	275

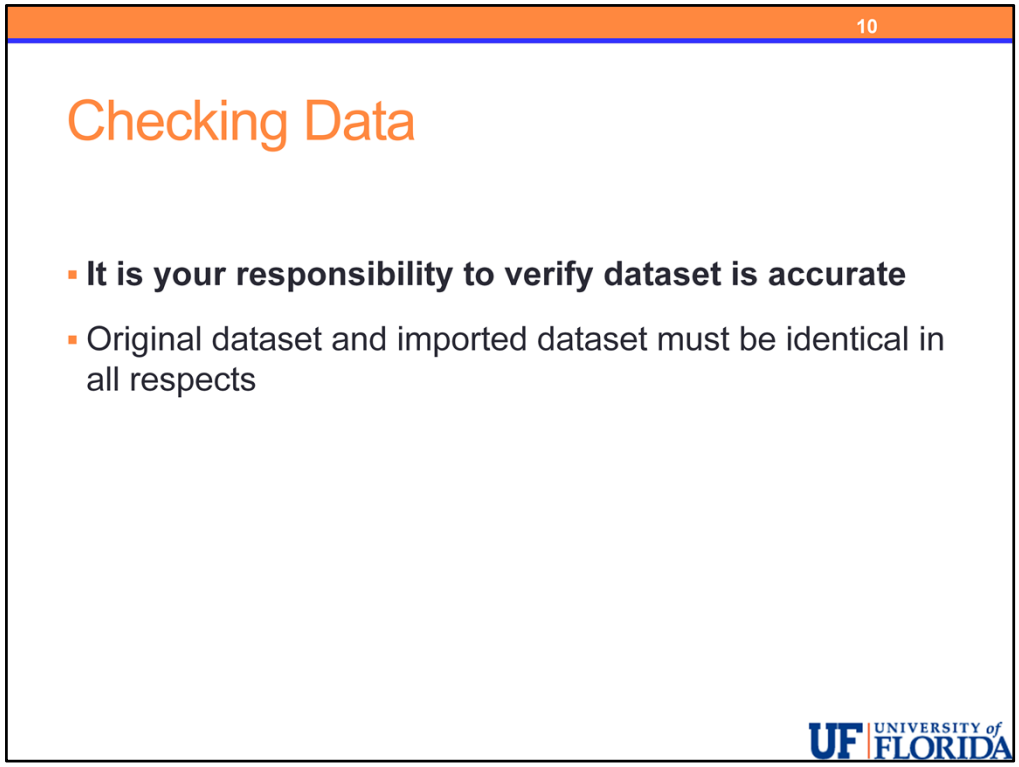
6	263
5	265
1	265
2	266
6	266
3	267
5	267
1	268
3	268
4	269
4	270
2	270
5	270
4	272
3	273
2	274
6	275
1	276

UF UNIVERSITY of FLORIDA

These two datasets are the same!

The one on the left clearly defines the dependence in the data

The one on the right does not and in fact, without access to the original, we would find it impossible to sort out the dependence in the observations in the data



Throughout the class, it is your responsibility to check your data. When you import data into the statistical software package or use data you are given in a particular format, be certain to check the data carefully.

If you do not and there is a problem which you do not notice and repair, you may be required to complete the assignment again using the correct data.

WHAT IS DATA?