Organizing Data

We'll usually assume that a *data set* is organized according to the *units* (also called *subjects* or *cases*) that it describes.

For each unit, the values of certain *variables* are recorded. These values may be *categorical* or *numerical* (also called *quantitative*).

Examples of categorical variables:

sex:	male / female
faculty:	arts&science / engineering / medicine
drive:	front-wheel / rear-wheel / four-wheel

Examples of numerical variables:

net worth:	Any real number (dollars)
age:	Any non-negative real number (years)
percent alcohol:	A real number between 0 and 100
number of arrests:	Any non-negative integer

Some variables have associated measurement units (eg, years). A *discrete* variable takes on integer values. A *continuous* variable takes on real values, but may have a restricted range.

Example: Crash-Test Dummies

The U.S. National Transportation Safety Board crashed vehicles into a wall at 35 miles per hour, and recorded the effect on dummies in the driver's and passenger's seat.

Among the variables measured for these tests were the following:

make:	Make of the car
model:	Model of car within that make
year:	Model year for the car
head:	A measure of the extent of head injuries
chest:	A measure of chest deceleration
side:	Which side of the car the dummy was on
protection:	Protection for dummies (eg, airbag)
weight:	Weight of the vehicle in pounds

'Make', 'model', 'side', and 'protection' are categorical variables.

'Head', 'chest', and 'weight' are numerical variables, with positive real values.

What about 'year'?

Values Observed for Some of the Units

make	model	year	head	chest	side	protection	weight
Buick	Elect._Park_Ave	88	1467	54	Drvr	manual_belts	3360
Chevrolet	Camaro	91	585	39	Drvr	d_airbag	3191
Chevrolet	G-20_Beauville	87	1387	67	Drvr	manual_belts	4887
Chevrolet	Suburban	87	1477	50	Drvr	manual_belts	5619
Chevrolet	S-10_Blazer_4X4	91	1026	71	Drvr	manual_belts	3820
Daihatsu	Charade	88	768	43	Drvr	manual_belts	1820
Ford	Club_Wagon	90	2613	59	Drvr	manual_belts	5103
Ford	F-150	88	1074	56	Drvr	manual_belts	3757
Isuzu	Stylus	91	580	57	Drvr	d_airbag	2333
Mercury	Tracer	89	940	48	Drvr	manual_belts	2280
Mitsubishi	Eclipse	90	772	44	Drvr	Motorized_belts	2594
Mitsubishi	Wagon	89	805	49	Drvr	manual_belts	3441
Nissan	Pulsar_Nx	88	1134	40	Drvr	manual_belts	2480
Nissan	Stanza	90	1105	59	Drvr	Motorized_belts	2790
Nissan	Stanza	91	546	56	Drvr	Motorized_belts	2740
Plymouth	Acclaim	91	762	55	Drvr	d_airbag	2860
Pontiac	Trans_Sport	90	761	42	Drvr	manual_belts	3740
Toyota	4-Runner_4x4	90	1306	48	Drvr	manual_belts	3860
Acura	Integra_RS	90	637	42	Pass	Motorized_belts	2490
Audi	100	89	710	31	Pass	d_airbag	3100
Chevrolet	Astro	89	1838	64	Pass	manual_belts	4002
Chrysler	Le_Baron	90	2043	46	Pass	d_airbag	3000
Daihatsu	Charade	88	642	37	Pass	manual_belts	1820
Ford	Mustang	90	438	50	Pass	d_airbag	3445
Ford	Taurus	90	609	40	Pass	d_airbag	3080
Honda	Civic_Dx	88	533	38	Pass	manual_belts	2077
Isuzu	Amigo	90	744	63	Pass	manual_belts	2900
Jeep	Wrangler_YJ_4x4	87	1229	40	Pass	manual_belts	3120
Lexus	ES250	90	630	47	Pass	d_airbag	3280
Nissan	NL_Xev_Pickup	88	1242	53	Pass	manual_belts	2631

Example: Calcium and Blood Pressure

Data from an experiment regarding the effect of calcium on systolic B.P. in 21 Black men:

B.P. before	Treatment received	B.P. after
107	Calcium	100
110	Calcium	114
123	Calcium	105
129	Calcium	112
112	Calcium	115
111	Calcium	116
107	Calcium	106
112	Calcium	102
136	Calcium	125
102	Calcium	104
123	Placebo	124
109	Placebo	97
112	Placebo	113
102	Placebo	105
98	Placebo	95
114	Placebo	119
119	Placebo	114
112	Placebo	114
110	Placebo	121
117	Placebo	118
130	Placebo	133

B.P. before and after treatment are continuous numerical variables. Whether calcium or a "placebo" was give is a categorical variable.

Example: Election Poll

CBS conducted a series of polls on voter preference for the 1988 U.S. Presidential election. Here are the data on a few of the people interviewed:

candidate preference	sex	ethnicity	age	education	state	days before election
Bush	male	other	45-64	hs-grad	GA	9
Dukakis	male	other	30-44	some-college	IA	9
Bush	female	other	18-29	some-college	SC	9
Bush	male	other	30-44	college-grad	MD	9
Bush	male	other	30-44	some-college	CA	8
Bush	male	other	18-29	not-hs-grad	OH	8
Bush	female	other	30-44	some-college	MS	8
Dukakis	male	other	45-64	hs-grad	MO	8
No-answer	male	other	30-44	hs-grad	MI	7
Dukakis	female	other	65+	some-college	MI	7
Dukakis	male	other	18-29	college-grad	PA	7
Bush	female	other	45-64	college-grad	KY	6
Dukakis	female	other	45-64	hs-grad	SC	4
Bush	male	other	65+	some-college	OR	4
Bush	female	other	30-44	college-grad	IN	4
No-answer	female	black	45-64	not-hs-grad	VA	4
Bush	male	other	65+	hs-grad	NJ	4
Bush	female	other	65+	some-college	CO	3
Dukakis	female	other	18-29	hs-grad	KY	3
Dukakis	male	black	18-29	hs-grad	MD	3
Bush	female	other	18-29	college-grad	WA	2
Dukakis	male	other	18-29	hs-grad	CA	2

Example: Survival on the Titanic

The following is known (with some errors) for the 2201 passengers and crew on the *Titanic*:

```
class:     1st, 2nd, or 3rd class passenger, or crew
age:       child or adult
sex:       female or male
survived:  Whether or not they survived
```

Here is the data on a few of the people:

class	age	sex	survived
1st	adult	male	yes
1st	adult	female	yes
2nd	adult	male	yes
2nd	adult	male	no
2nd	adult	female	yes
2nd	child	male	yes
3rd	adult	male	yes
3rd	adult	male	no
3rd	child	male	no
crew	adult	male	yes
crew	adult	male	yes
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no
crew	adult	male	no

Reference: The article by Robert Dawson in the on-line *Journal of Statistics Education*, vol. 3, no. 3, at `http://www.amstat.org/publications/jse/`

Where Do the Units and the Values of Variables Come From?

The *population* is the set of all units that we are interested in.

The *sample* is the set of units for which we have measurements. In a *census*, the sample is the entire population.

In a *survey* or in an *observational study*, we observe the values of the variables.

In an *experiment*, we control some of the variables, and observe others.

Surveys

When the population of interest is finite, we might be able to do a census, but looking only at a smaller sample is likely to be cheaper.

And often we are interested in an infinite (or indefinite size) population — eg, all cars of a given make and model that will ever be made.

A *simple random sample* of a given size is obtained by a procedure that has an equal chance of picking any sample of that size.

Using a simple random sample eliminates *bias* — a systematic tendency to get things wrong.

Some more complicated sampling schemes, such as *stratified* and *cluster* sampling, are also unbiased, and may be cheaper (for a given level of accuracy).

Sampling Bias

In practice, eliminating bias completely is very difficult.

First, the *sampled population* may differ from the *target population*:

> For an election poll meant to predict the winner, the target population is *people who will vote*.
>
> The sampled population won't be this. It might be *people who say they will vote* or *people who are eligible to vote*.

Second, it may be impossible to obtain the true values of some variables for units in the sample:

> Some people contacted in an election poll may refuse to say who they plan to vote for, or they may not tell the truth.

The fraction who refuse to answer may be different for people who plan to vote for different candidates — leading to bias.

Sampling Variability

Even if we have eliminated bias, the answer we get from a sample that isn't the whole population will probably not be quite right.

Random sampling variability — who ended up by chance in the sample — will affect the results.

A basic principle of statistics:

> **Sampling variability can be reduced by using a larger sample.**

A sample of around 1000 is typical for election polls, and gives answers accurate to about 3%.

In contrast, it makes very little difference what *fraction of the population* is in the sample, unless the sample approaches the entire population (in which case results are more accurate than they would be for the same sample size with a bigger population).

Census or Sample?

Consider the data on people on the *Titanic*. Is this a census, or a sample from a larger population?

The answer depends on the target population, and that depends on the question you're trying to answer.

Some questions this data might help answer:

- In England at the time, were boys more likely to be taken on ocean voyages than girls?
- Were there so many children on the *Titanic* that attending to them interfered with rapidly putting people on lifeboats?
- Did the crew of the *Titanic* discriminate in favour of the upper classes when putting people on lifeboats?