Chapter 1

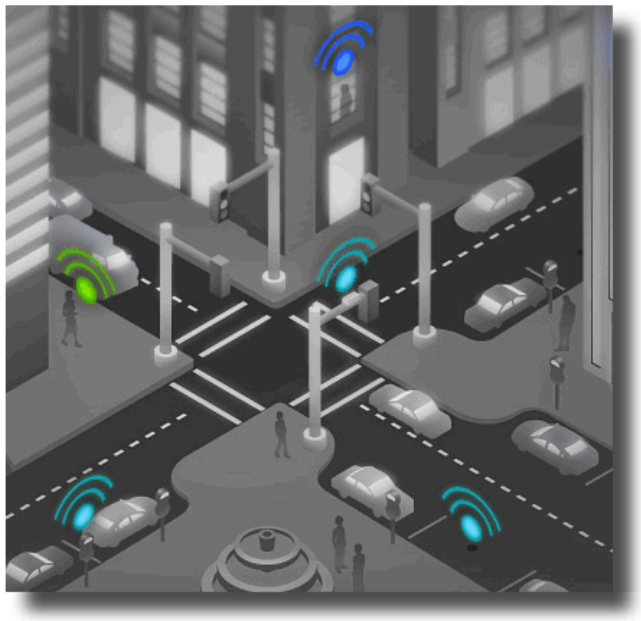
Introduction to Data

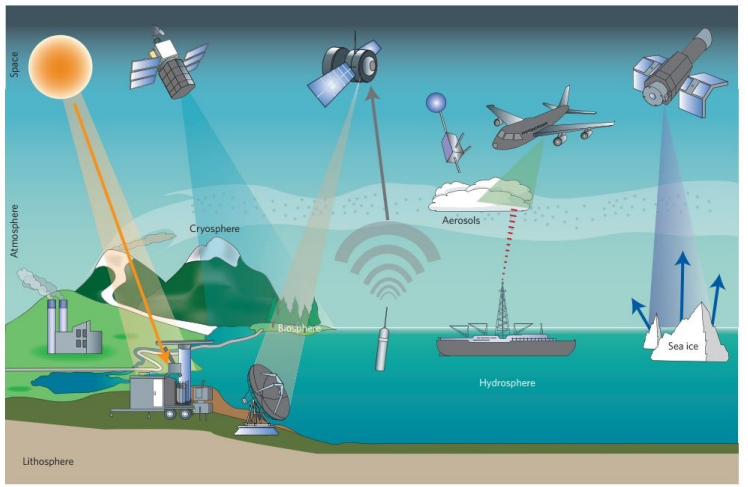
Statistics is…

1. the science of data and variability
2. the methodology of collecting, analyzing and drawing conclusions from data
3. making effective use of the data around us to make decisions about ourselves and our surroundings
4. all of the above

Data

- Data are produced by people, machines, sensors, computers, phones
- Data can be very large
- Data are everywhere!

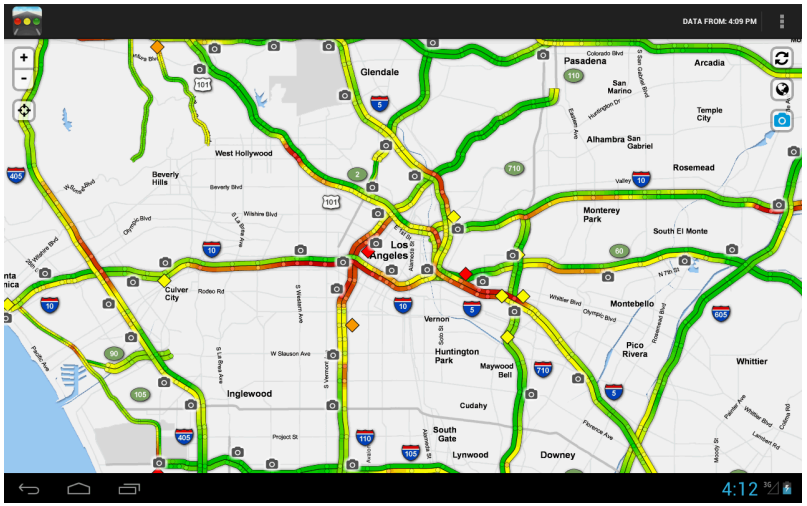


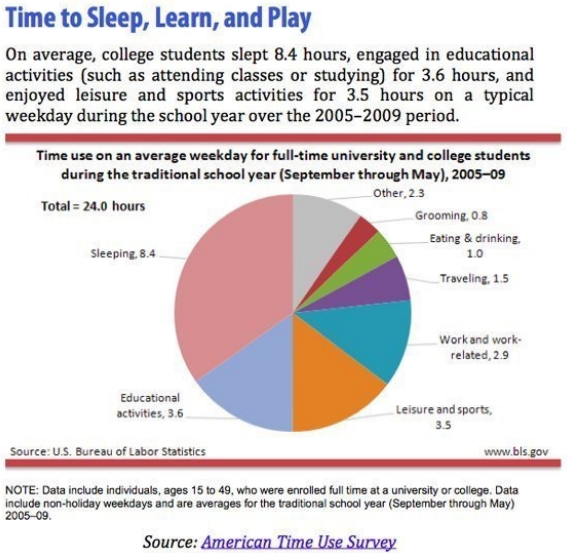


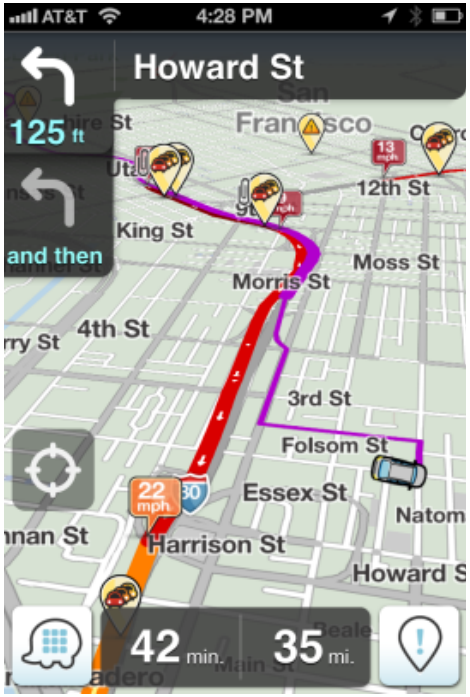
Data

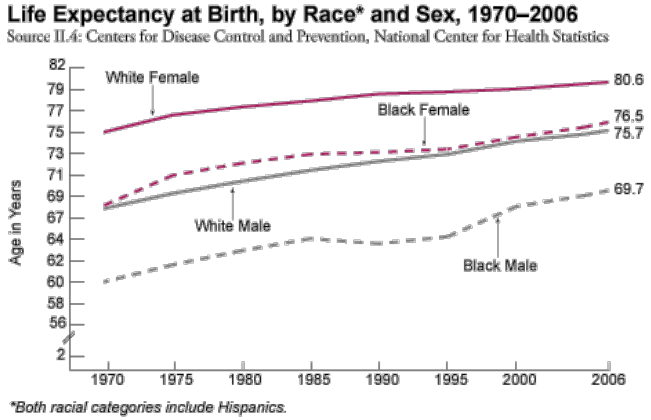
The data flows into here…

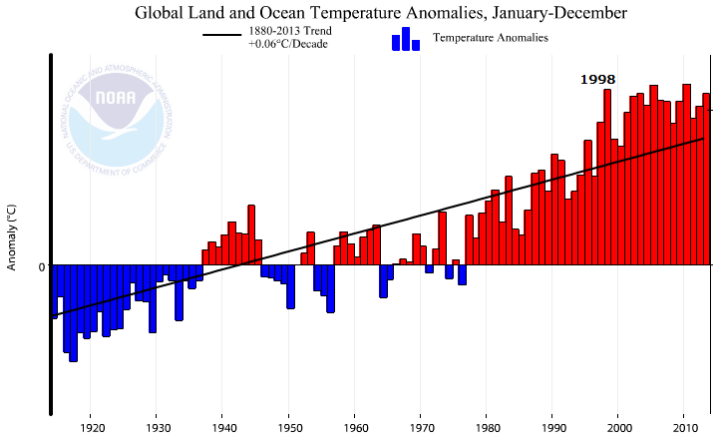
where summaries like these are made!











Data

(building blocks of statistics)

- More than just numbers
- Collections of numbers, measurements, or any type of observation that someone records
- Can be quantitative (numerical) or qualitative (categorical)
- Statistics is used to analyze and interpret data

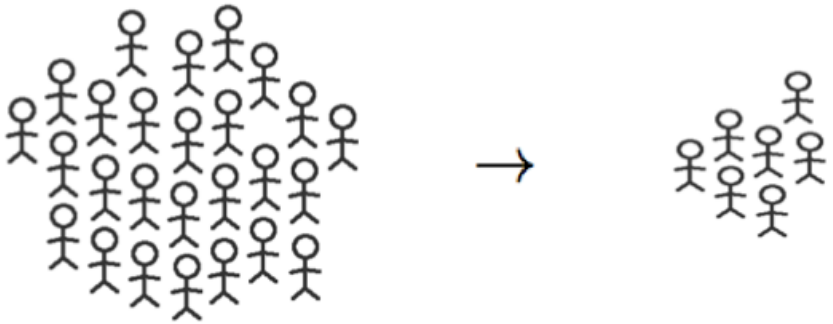
Data Collection

- Election polls
- Surveys
- Google Analytics (browser history)
- Smartphone Apps
- Sales transactions
- Hospital and school records
- Sports
- Twitter/Facebook posts
- Satellites

Populations and Samples

A POPULATION is the collection of observations of interest. This number is usually very large and nearly impossible to obtain measurements from

A SAMPLE is a *portion* of a POPULATION of interest. A sample is usually taken to measure a characteristic about a population. The size of a sample is usually denoted by *n*.



Begin by Studying Data

- Data arise from observations
- For a single observation, we might note several attributes
- These attributes are called variables

Example

#	Full Name	Pos.	Ht.	Wt.	Yr.	Hometown / High School
0	Alex Olesinski	F	6-10	200	So.	Roswell, N.M. / La Lumiere School [IN]
2	Lonzo Ball	G	6-6	190	Fr.	Chino Hills, Calif. / Chino Hills HS
3	Aaron Holiday	G	6-1	185	So.	Chatsworth, Calif. / Campbell Hall HS
10	Isaac Hamilton	G	6-5	195	Sr.	Los Angeles, Calif. / St. John Bosco HS
13	Ike Anigbogu	F/C	6-10	250	Fr.	Corona, Calif. / Centennial HS
14	Gyorgy Goloman	F	6-11	215	Jr.	Kormend, Hungary / The Sagemont School [FL]
15	Jerrold Smith	G	6-0	165	Sr.	Los Angeles, Calif. / St. Bernard's HS
20	Bryce Alford	G	6-3	185	Sr.	Albuquerque, N.M. / La Cueva HS
21	Alec Wulff	G	6-3	185	Jr.	Laguna Beach, Calif. / Laguna Beach HS
22	TJ Leaf	F	6-10	225	Fr.	El Cajon, Calif. / Foothills Christian HS
23	Prince Ali	G	6-3	190	So.	The Bronx, N.Y. / The Sagemont School [FL]
34	Ikenna Okwarabizie	C	6-9	250	Jr.	Lagos, Nigeria / East HS [IA]
40	Thomas Welsh	C	7-0	245	Jr.	Redondo Beach, Calif. / Loyola HS

- Each row is an observation
- Each column is a variable

Types of Variables

- Numerical: the values of the variable are numbers.
  - Examples: weight, height, temperature, GPA

- Categorical: the values of the variable are categories or classifications
  - Examples: eye color, year in school, class subject

Identifier Variables

- The information is unique to each individual or object in the dataset.
    - Examples: ID number, driver's license number, social security number
- These variables are useful for the researcher but not important for analysis.

Example

Student ID	Major	GPA	Year	Male
20429	Mathematics	3.64	Senior	1
28503	English	3.25	Junior	1
27604	Dance	3.87	Freshman	0
39875	Statistics	3.15	Sophmore	0
60983	History	2.67	Freshman	0
19875	Biology	3.01	Senior	1
12309	Sociology	3.34	Junior	0

What are the variables? How many observations?
Categorical vs Numerical?
What questions can we ask and answer using this data?

Context is key

To understand data we have to ask the following questions:

- Who, or what, was observed?
- What variables were measured?
- How were they measured?
- What are the units of measurement?
- Who collected the data?
- How did they collect the data?
- Where were the data collected?
- Why were the data collected?
- When were the data collected?

Organizing and Reporting Categorical Data

- Frequencies (or counts) are one way to report a categorical variable
- A two way table (or frequency table) is a way to display the counts of two categorical variables

	Dog	Cat	Total
Freshmen	42	10	52
Seniors	9	39	48
Total	51	49	100

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

Two-way Tables

	Dog	Cat	Total
Freshmen	42	10	52
Seniors	9	39	48
Total	51	49	100

- How many people were seniors and owned a cat?
- What percentage of people surveyed owned a dog?
- What proportion of freshmen owned a cat?

Two-way Tables

	Dog	Cat	Total
Freshmen	42	10	52
Seniors	9	39	48
Total	51	49	100

- How many people were seniors and owned a cat? **39**
- What percentage of people surveyed owned a dog?

  51/100 = 51%
- What proportion of freshmen owned a cat? **10/52**

Two-way Tables

	Dog	Cat	Total
Freshmen	42	10	52
Seniors	9	39	48
Total	51	49	100

- Are freshmen or seniors more likely to own a dog?

  percentage of freshmen who owned a dog:

  percentage of seniors who owned a dog:

Two-way Tables

	Dog	Cat	Total
Freshmen	42	10	52
Seniors	9	39	48
Total	51	49	100

- Are freshmen or seniors more likely to own a dog?

  percentage of freshmen who owned a dog: **42/52 = 81%**

  percentage of seniors who owned a dog: **9/48 = 19%**

Two-way Tables

Clicker!

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- How many people surveyed were younger than 25 years of age?

  A. 626

  B. 37

  C. 242

  D. 2827

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- How many people surveyed were younger than 25 years of age?

  A. 626

  B. 37

  C. 242

  D. 2827

Two-way Tables

Clicker!

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- What percentage of people surveyed were between 25 and 54 years of age?

A. 25%

B. 63%

C. 72%

D. 100%

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- What percentage of people surveyed were between 25 and 54 years of age?

A. 25%

B. 63%

C. 72%

D. 100%

(626+679+481)/2827 = 1786/2827 = 63%

Two-way Tables

Clicker!

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- What proportion of divorcees were 65 or older? 25%
  - A. 48/444
  - B. 48/2827
  - C. 444/2827
  - D. 479/2827

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

- What proportion of divorcees were 65 or older? 25%

A. 48/444

B. 48/2827

C. 444/2827

D. 479/2827

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

	Never Married	Married**	Total
Younger than 35	457	411	868
Older than 35	205	1754	1959
Total	662	2165	2827

Younger than 35
total = 868
# of never married
= 457

Older than 35
total = 1959
# of never married
= 205

- Are people younger than 35 more likely to never be married than people older than 35?

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

	Never Married	Married**	Total
Younger than 35	457	411	868
Older than 35	205	1754	1959
Total	662	2165	2827

Younger than 35 who are never married = 457/868 = 53%

Older than 35 who are never married = 205/1959 = 10%

- Are people younger than 35 more likely to never be married than people older than 35? **Younger**

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

	Never Married	Married**	Total
Younger than 35	457	411	868
Older than 35	205	1754	1959
Total	662	2165	2827

Never married total = 662
# of younger than 35 = 457

Married** total = 2165
# of younger than 35 = 411

- Are people who have never been married more likely to be younger than 35 than those who have been married**?

Two-way Tables

		Marital Status					
		Married	Widowed	Divorced	Separated	Never married	Total Count
Age Category	Less than 25	37	1	5	5	194	242
	25 to 34	271	13	63	16	263	626
	35 to 44	379	11	129	44	116	679
	45 to 54	275	18	123	13	52	481
	55 to 64	186	31	76	7	20	320
	65 or older	197	209	48	8	17	479
	Total Count	1345	283	444	93	662	2827

	Never Married	Married**	Total
Younger than 35	457	411	868
Older than 35	205	1754	1959
Total	662	2165	2827

Never married who are younger than 35 = 457/662 = 69%

Married** who are younger than 35 = 411/2165 = 19%

- Are people who have never been married more likely to be younger than 35 than those who have been married**? **Never married**

Causality

- Establishing causality means to show than an outcome is effected by some treatment.
- **Treatment group:** individuals who receive the treatment of interest in an experiment
- **Control group:** individuals who do NOT receive treatment
- Examples:
    - Clinical trial for a new medicine to lower blood pressure
    - Experiment for a new weight loss exercise plan
    - A new training method to see if an athlete can improve run time

Association is NOT Causation

- Unless the individuals of the study are identical in every way except for the treatment, we cannot conclude causation which means that the treatment caused the outcome.
- If a certain type of outcome occurs more frequently in one group, we can conclude that the treatment and outcome are associated.
- A confounding variable is a characteristic other than the treatment that causes both outcomes.
- Example: People with grey hair are observed to have more wrinkles. Does this mean that grey hair causes wrinkles?

  Grey hair is associated with wrinkles, but old age causes both grey hair and wrinkles so grey hair isn't the cause of wrinkles.

Observational Study

- In an observational study, researchers do not assign choices; they simply observe them.
- No treatment is applied to any individual or subject.
- Observational studies are valuable for discovering trends and possible associations.
- It is NOT possible for observational studies to demonstrate a causal relationship.
- Examples:
    - Recording the number of passes a basketball player makes during a game.
    - Counting the number of people wearing a UCLA shirt on campus.
    - Counting the number of people who have brown hair in our class.

Controlled Experiments

- In an experiment, the researcher/experimenter deliberately manipulates the treatment variable and assigns the subjects to those treatments, usually at random.
- There must be at least one treatment variable to manipulate and at least one outcome variable to measure
- The outcome variable is observed and compared for the different groups of subjects who have been treated differently
- It IS possible to show a causal relationship with an experiment.
- Examples:
    - New forms of advertising to see if sales increase from the previous quarter
    - New lighting in classrooms to see if less students fall asleep

Principals of Experimental Design

- Large sample size: This ensures that the study captures the full range of variability amongst the population and allows small differences to be noticed.
- Controlled and randomized: Random assignment of subjects to treatment or control groups to minimize bias.
- Double-blind: Neither subjects nor researchers know who is in which group.
- Placebo (if appropriate): This format controls for possible differences between groups that occur simply because subjects think their treatment is effective.

Bias and Random Assignment

- Bias is the tendency to overestimate or underestimate a population parameter due to a measurement process.

  Examples:

  - Polling only conservatives to estimate who will win an election.
  - Surveying people at the Wooden Center to estimate the average time a student spends working out a week.
  - A researcher putting the heaviest people in the same group for a diet study.

- Random assignment helps minimize bias.

  Examples:

  - Use a computer or random number generator that randomly assigns the people being studied into the control and treatment groups.
  - Randomly pull number out of a bag to assign individuals or subjects to groups.

Blinding

- Blinding helps prevent bias from being introduced into a study by ensuring that the participants (and sometimes the researchers) do not know who is assigned to which study group.
- Who can influence the outcome of an experiment?
  - If the researcher knows a participant is in a certain group they might interact with them depending on the group they are in.
  - If the participant knows which treatment they are receiving they might behave differently than they would if they knew nothing about their treatment.
- A study is double-blind if neither side knows who is in either group.

Placebos

- A placebo is a “fake” treatment that looks just like the treatment being tested.
- Sometimes merely applying some form of treatment is enough to induce an improvement.
- This is one of the best methods to blind a subject.

Example

A new women's antiperspirant has been developed. We're interested to determine if it actually impedes perspiration.

- How would we do this?
    - Compare how much women perspire when using the new antiperspirant to how much women perspire when using nothing

We have a group of women willing to try the new product and allow us to measure the amount they perspire.

- Is this a controlled experiment or observational study?
- Do we have a treatment group and control group?

Example

A new women's antiperspirant has been developed. We're interested to determine if it actually impedes perspiration.

- How would we do this?
    - Compare how much women perspire when using the new antiperspirant to how much women perspire when using nothing

We have a group of women willing to try the new product and allow us to measure the amount they perspire.

- Is this a controlled experiment or observational study?

  It's an experiment if we choose which group gets the treatment. Otherwise, it's an observational study.

- Do we have a treatment group and control group?

  Only if it's an experiment, treatment group: new perspirant and control group: wears nothing.

Example

Records of patients who have had broken ankles are examined to see whether those who had physical therapy achieved more ankle mobility than those who did not.

- Is this a controlled experiment or observational study?
- Is there a treatment group and control group?

Example

Records of patients who have had broken ankles are examined to see whether those who had physical therapy achieved more ankle mobility than those who did not.

- Is this a controlled experiment or observational study?

  Observational study

- Is there a treatment group and control group? **No**

Example

A researcher was interested in the effect of exercise on memory. She randomly assigned half of a group of students to run up a stairway three times and the other half to rest for an equivalent amount of time. Each student was then asked to memorize a series of random digits. She compared the numbers of digits remembered for the two groups.

- Is this a controlled experiment or observational study?
- Is there a treatment group and control group?

Example

A researcher was interested in the effect of exercise on memory. She randomly assigned half of a group of students to run up a stairway three times and the other half to rest for an equivalent amount of time. Each student was then asked to memorize a series of random digits. She compared the numbers of digits remembered for the two groups.

- Is this a controlled experiment or observational study?

  Controlled experiment

- Is there a treatment group and control group?

  Treatment group: running group

  Control group: resting group

Example

Clicker!

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Of the women who took TAC, what percent survived?

  A. 426/521 = 82%

  B. 473/539 = 88%

  C. 473/1060 = 45%

  D. 426/1060 = 40%

Example

Clicker!

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Of the women who took FAC, what percent survived?

  A. 426/521 = 82%

  B. 473/539 = 88%

  C. 473/1060 = 45%

  D. 426/1060 = 40%

Example

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Of the women who took TAC, what percent survived? **473/539 = 88%**
- Of the women who took FAC, what percent survived? **426/521 = 82%**

Example

Clicker!

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Is this a controlled experiment or observational study?

  A. Controlled experiment

  B. Observational study

Example

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Is this a controlled experiment or observational study?
  Controlled experiment

Example

Clicker!

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Can we conclude causation?

  A. Yes

  B. No

Example

Two drugs were tested to see whether they helped women who had breast cancer without lymph node involvement. The drugs are called TAC and FAC. About half of the 1060 women with breast cancer without lymph node involvement were randomly assigned to TAC and the other half were assigned to FAC. After 77 months, 473 out of 539 of the women assigned to TAC were alive, and 426 out of 521 women assigned to FAC were alive.

- Can we conclude causation?

  Since this is a controlled experiment with random assignment we can conclude causation or cause and effect. The random assignment balances out other variables, so the only difference is the treatment which must be causing the effect.