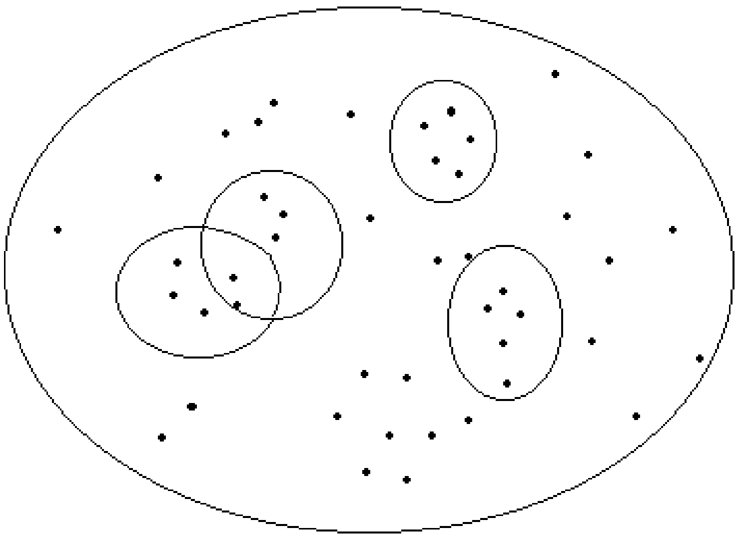
<u>Where Are We?</u>

1. **Identify the Question**
   - A. **Describe the problem**
   - B. **State the question(s)**
   - C. **Check the data format**
   - D. **Reflect on the study design**
2. **Analyze the Data**
   - A. **Identify the relevant variables**
   - B. **Determine the appropriate analysis**
   - C. **Conduct the analysis**
   - D. **Interpret the results**
   - E. **Consider whether additional analyses are necessary**
3. **Draw Conclusions from the Data**
   - A. **Re-state the question(s)**
   - B. **Answer the question(s) based on analyses.**
   - C. **Evaluate the strengths and weaknesses**

Gathering Good Data

Gathering Data: Sampling

- *Population:* The entire group of people, animals, things that we want to learn about.
- *Unit:* Any individual in the population. Also called an *observation*, or just plain *individual*.
- *Sample:* A part of the population from which we actually collect information to learn about the population.
- *Sampling Frame:* A list of units from which we choose the sample.
- *Variable:* A piece of information measured in the same way for all units in the sample (or population).



Example: Public Opinion Polls

Gallup, CBS/New York Times, USA Today, and other organizations, all conduct polls throughout the year asking people's opinions about various issues of the day. In a typical poll,

- The *population* is all US Residents ages 18 and older. Noncitizens and even illegal immigrants are included.

- A *unit* is anyone living in the US.

- A *sample* is, say, 1000 persons interviewed by telephone.

- The *sampling frame* is a list of telephone numbers from which the sample is drawn.

- The *variables* are the answers that each person gives to the questions in the poll.

Example: Medical Survey

Doctors studying the quality of care given to patients, say, at risk of heart disease, may study medical records and record how frequently doctors prescribed aspirin or beta-blockers.

- The *population* is all US patients at risk of heart disease.

- A *unit* is anyone seeing a doctor and at risk.

- A *sample* is, say, 3000 patients' medical records from various hospitals around the country.

- The *sampling frame* is the set of all medical records available to the researchers.

- The main *variable* of interest is whether or not the physician prescribes aspirin or beta-blockers. Other *variables* that might be of interest include the patient's level of risk, the size and quality of the hospital, the caseload of the doctor, etc.

Example: Internet Questionnaire

To popularize its web site and perhaps to gather information tracking new trends in public opinion, the Gallup Organization used to offer the opportunity for anyone who came to http://www.gallup.com to fill out a survey form on issues of the day. The results would be updated regularly and reported on the web site.

- The *population* was unstated. But the nature of the questions suggested Gallup was interested in all US adult citizens.

- A *unit* was anyone living in the US.

- The *sample* was whatever group of people had filled out the survey form.

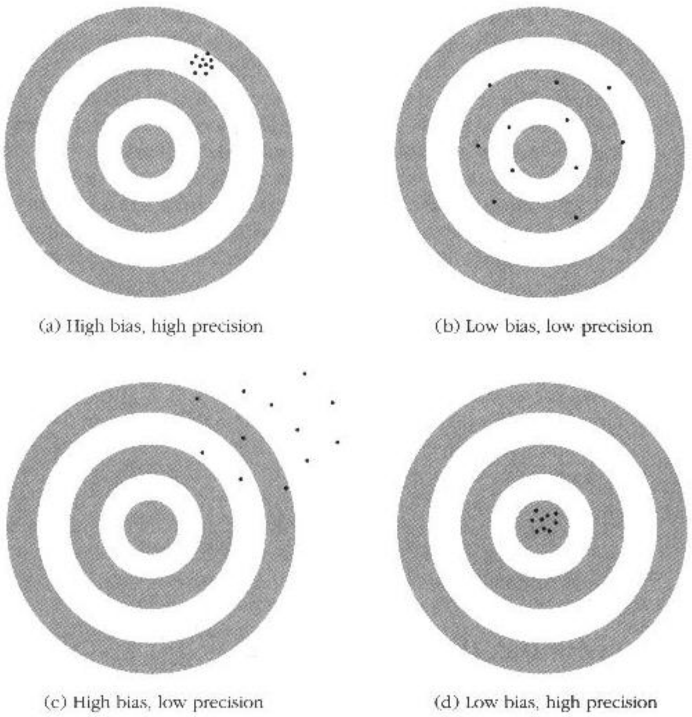
- The *sampling frame* was the set of all web users.

- The *variables* were the answers to the questions on the survey form.

What is **Good Data***?*

- *Bias* is a systematic tendency for samples to deviate from the pattern of observations in the population.
- *Precision* consistency from one observation to the next in a sample.

A good sample has *low bias* and *high precision*.



(a) High bias, high precision

(b) Low bias, low precision

(c) High bias, low precision

(d) Low bias, high precision

Biased Sampling Methods

There are two common ways to generate bias in samples:

- *Wrong sampling frame,* not identical to the population we want to learn about.
  - Often caused by "samples of convenience".
- *Selection effect,* the fact that the person is responding and the nature of the response are related somehow.
  - usually caused by "voluntary response samples."

Simple Random Samples (SRS's)

Guaranteed (by math!):

- To be *unbiased* for their sampling frame.
- To allow calculation of the *precision* (via "square root law").

> A *simple random sample (SRS)* of size n from a population, chosen in such a way that every possible set of size n is equally likely to be chosen.

One way to make a simple random sample "by hand" is as follows:

1. *Label.* Number all the units in the sampling frame from 0 to the max, but make sure each unit has the same number of digits
   - if there are less than 100 units, number them 00, 01, ...
   - if there are less than 1000 units, number them 000, 001, ...

   and so forth.
2. *Table.*
   - Start anywhere in a table of random digits and circle successive groups (pairs or triples or whatever) of numbers.
   - Take units whose numbers are drawn in this way for your sample (no repeats!).
   - If you circle a number that is not on any of your units (say, you circle 37 even though there are only 30 units), just ignore it and go on to the next group.

Example: SRS

The chess club has 11 members. University rules require a club executive committee of three members. To be fair, the club will select its executive committee as an SRS. The club members are:

Able	Ernst	Ibis
Baker	Forrest	Jackson
Caldwell	Gibson	Kellogg
Darling	Hanson	

TABLE A. Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244

Making Good Comparisons

- An *Observational study* observes (measures variables for) inviduals but does not attempt to influence the responses.
  - *What is the pattern in the population?*
  - E.g.: An opinion poll or sample survey.
- An *experiment* deliberately imposes some treatment on individuals in the study in order to observe their responses to treatment.
  - *Does the treatment cause a change in the response?*
  - E.g.: A randomized controlled drug trial.

Some vocabulary:

- *Units*, also called *subjects*.
- *Response variable* a variable we want to learn about by manipulating the treatment
- *Explanatory variable* A variable that explains or causes changes in the response variable(s)
- *Treatement* Any unique combination of explanatory variables.

Example: Computer Education

To compare computer software that teaches reading with a standard reading curriculum, an educator tests the reading ability of a group of 60 fourth graders, then divides them into two classes of 30 students each. One class uses the computer, the other studies the standard curriculum. After a year, she retests the students and compares the average increases in reading ability in the two classes.

- What are the *explanatory* and *response* variables?

- Was this an *experiment* or an *observational* study?

- Will the results help us determine which method *causes* better learning (or are there *lurking variables*)?

Example: Treating breast cancer

The most common treatment for breast cancer was once mastectomy (removal of the breast). It is now usual to remove the tumor and nearby lymph nodes, followed by radiation. To study whether these treatments differ in their effectiveness, a medical team examines the records of 25 large hospitals and compares the survival times after surgery of all women who have had either treatment.

- What are the *explanatory* and *response* variables?

- Was this an *experiment* or an *observational* study?

- Will the results help us determine which method *causes* longer survival (or are there *lurking variables*)?

What is a **Good Comparison***?*

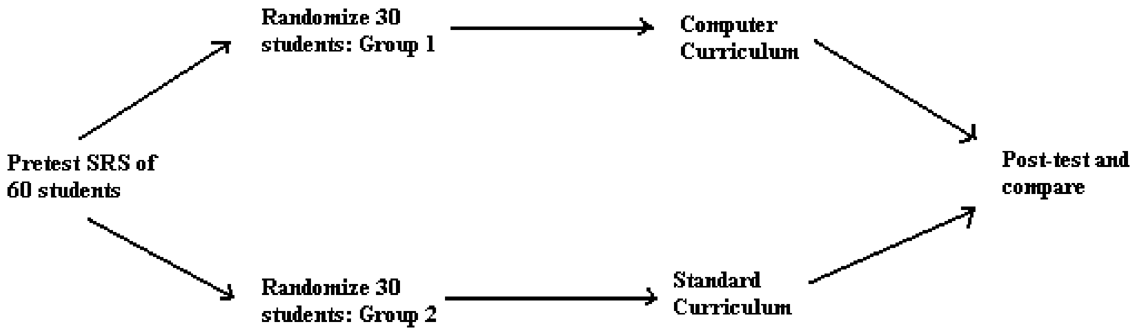
- Low bias
- High precision
- No counfounding

> – A **lurking variable** is an important explanatory variable that you forgot to include in the study.
>
> – Two variables (explanatory or lurking) are **confounded** when their effects on a response variable cannot be distinguished from each other.

- *In observational studies* we can seldom rule out lurking variables and other confounds, *so it is unusual to be able to make a causal claim.*
- *In a well-designed experiment* all lurking variables and confounds will be ruled out, so that *causal explanations are possible.*

Randomized Comparative Experiments

Also known as randomized comparative trials (RCT's), these are the simplest experiments that allow an unambiguous causal explanation.



- SRS (or something similar) to make sure that results reflect the population of interest.
- Random assignment to produce groups that are similar in all respects before treatment is applied.
- Apply comparative treatments at the same time and under the same circumstances so any influences other than the treatment differences act equally on the two groups.

Therefore, differences due to the treatments alone, and generalize to the population.