MATH 243 PARTIAL LECTURE NOTES (16 APRIL 2008)

N. CHRISTOPHER PHILLIPS

We can't afford to ask every registered voter in the US which presidential candidate they intend to vote for. So we choose only *some* of them, and try to estimate the results for everybody from the results for this group.

The *sample* is the group of people we actually interview. (In other cases, it could be the objects we test, etc.)

The *population* is the group of people we actually want information about, in this case all registered voters in the US.

We will later see how big we have to choose the sample. For now, we concentrate on avoiding bias.

How *not* to choose a sample.

The Emerald online poll. (It used to be on page 2 in the *Oregon Daily Emerald* most Mondays, but now seems to have disappeared. It asked people to comment online about such questions as whether so and so should remain as basketball coach, whether the Westmoreland housing should be sold, etc.)

It only considers people who know about it and care enough to respond. They are not representative.

This is a *voluntary response sample.* See Example 8.5 and Problem 8.32 of the book for a voluntary response sample which said 60% "yes" when more reliable sampling methods give about 55%–60% "no". (Example 7.1 in the previous edition was more dramatic. It had a voluntary response sample which said 67% "no" when more reliable sampling methods give about 72% "yes".)

Date: 16 April 2008.

The following are *convenience samples*:

The first 100 people you can interview as they go into the main entrance of the Valley River Center mall.

The people in the front row of my Math 243 class, for a study of Math 243 student attitudes towards statistics.

The people in the back row of my Math 243 class, for the same study.

1000 people found on the streets of Eugene, for an opinion poll on national issues.

Voluntary response samples and convenience samples are likely to be *biased:* they favor certain outcomes over others.

The most elementary way to avoid this is with a *simple random sample:* a sample selected at random in such a way that every possible sample of that size is equally likely to be selected. For example, in a fair state lottery, the set of numbers chosen is a simple random sample of all numbers in the appropriate range.

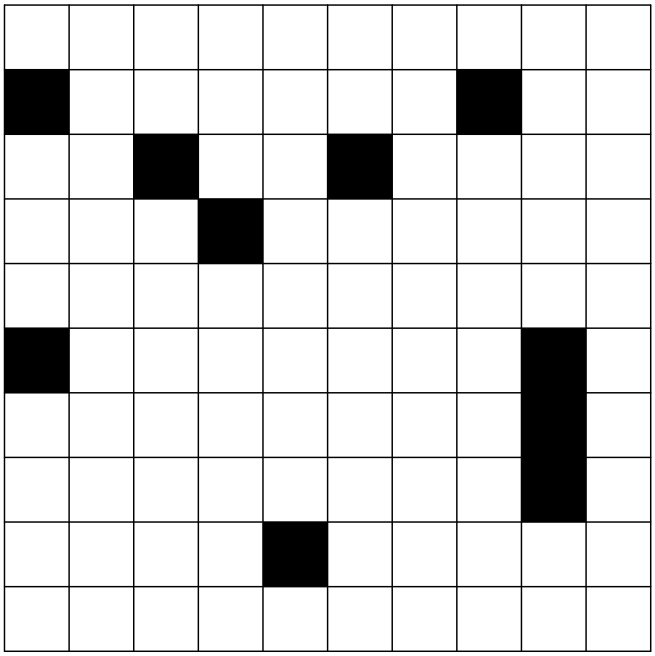
See the book about choosing a simple random sample using a table of random numbers (such as Table B in the book). One can also use a random number generator, although they produce only "pseudorandom" numbers. For most sampling purposes, this is likely to be good enough.

A particular simple random sample is not likely to be regularly distributed (although, by definition, *every* sample is a possible outcome).

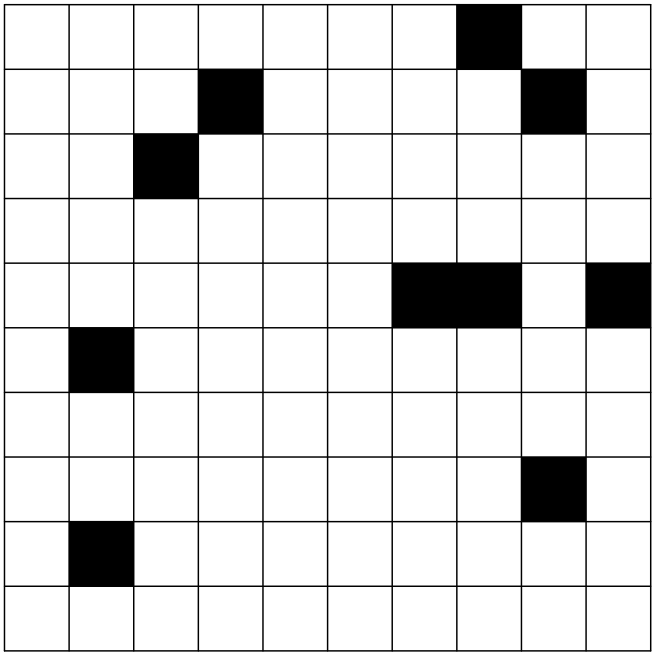
Do *not* choose every 12th widget off the assembly line as your sample. Maybe inside there is a machine that paints 12 widgets at once, and one of its nozzles is defective. For most starting places, you will see no improperly painted widgets, but for some starting places, *every* widget you see will be improperly painted.

Next: The first five results I got from choosing simple random samples of size 10 from a 10×10 square (say, of students in a classroom with 10 rows of 10 seats each). Selected locations are black. The choices were made using a random number generator on my computer, but you can do the same using Table B in the book.

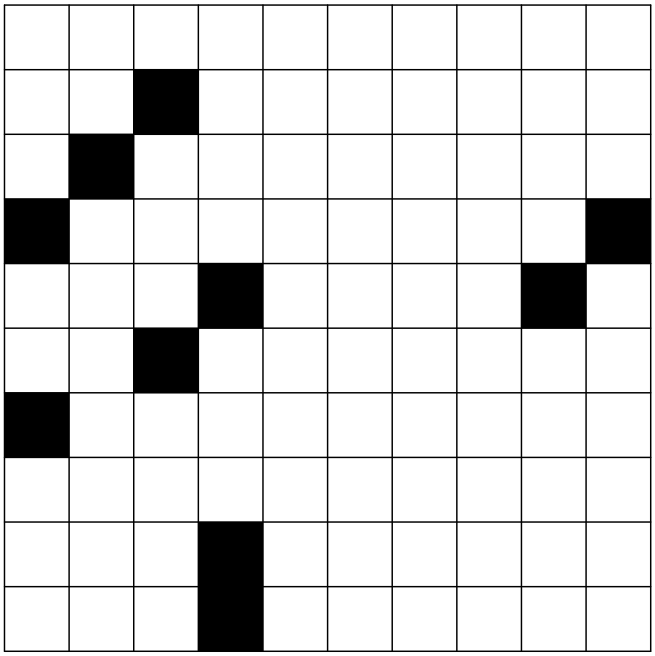
Sample 1:



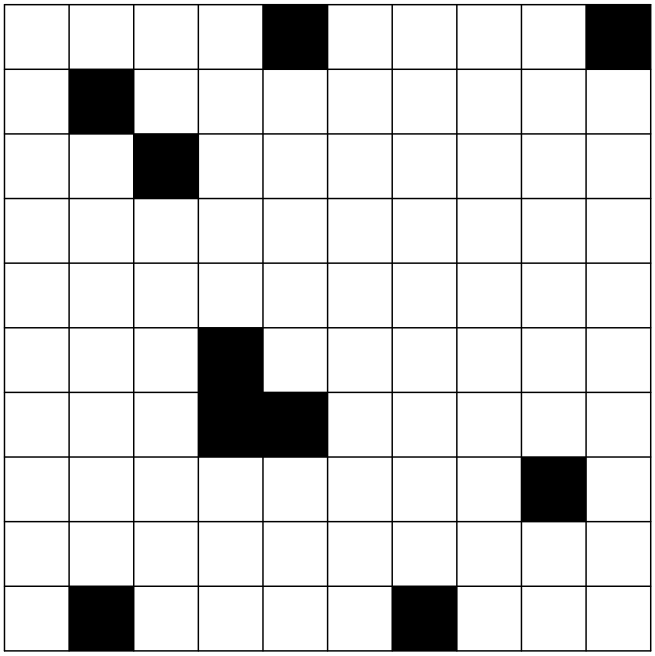
Sample 2:



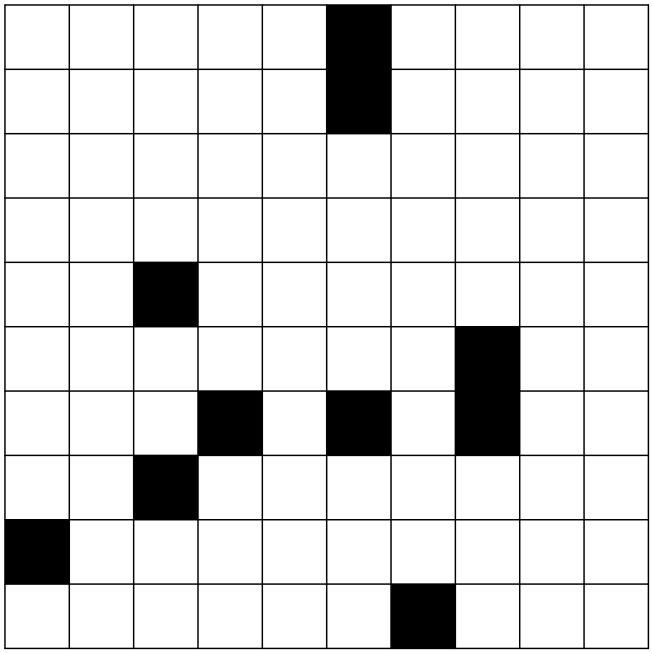
Sample 3:



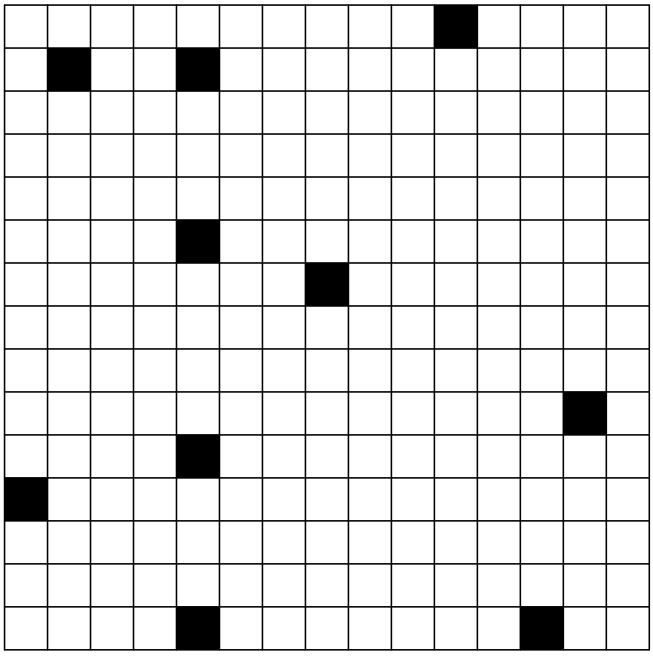
Sample 4:



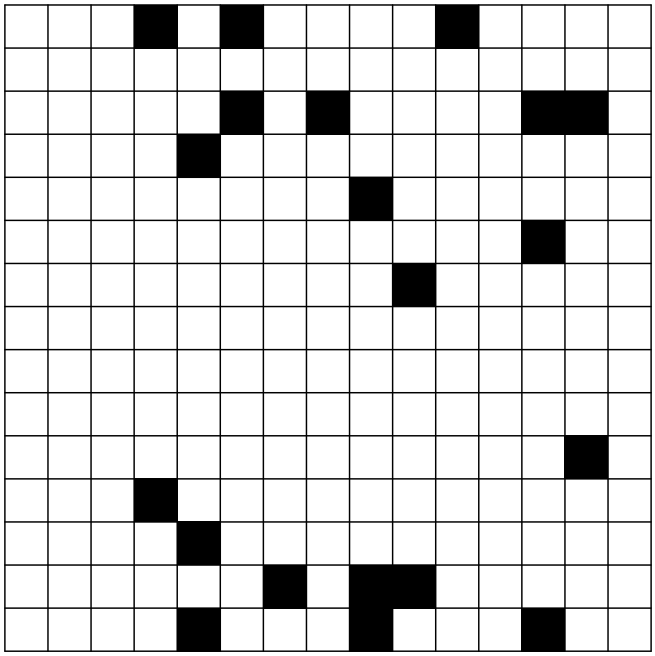
Sample 5:



A simple random sample of size 10 from a 15×15 grid:



A simple random sample of size 20 from a 15×15 grid:



Hazards of opinion polls and questionnaires; some apply more generally to many kinds of observational studies.

Hazard: Confusing questions. If people don't understand the question, the results will be useless.

Hazard: Biased questions. This can be either deliberate or accidental.

Here is a blatant example. Senator Snort's opponent's campaign conducts a survey using the question, "In view of Senator Snort's recent conviction for drunken driving, would you vote to reelect him?" Then it issues a press release claiming Senator Snort's campaign is in trouble, or sends a letter to potential campaign contributors using the result as evidence that Senator Snort is vulnerable.

The book contains other examples of biased questions, such as Examples 8.10 and 8.11.

Here is a quite subtle example. The question:

Imagine you are purchasing rice in your local grocery store. You can choose between two types of rice. One is regular long-grain white rice that ***has not been*** genetically engineered. This ***non-genetically engineered*** rice does ***not*** contain vitamin A. The other rice option is *Golden Rice. Golden Rice* **has been** genetically engineered **to contain** vitamin A. One serving of *Golden Rice* will satisfy 30% of your daily requirement for vitamin A, as outlined by the FDA. Now, imagine that you are in a grocery store and the price of a 1 lb. bag of regular long-grain white long grain is \$0.65. Would you purchase a 1 lb. bag of long grain *Golden Rice* if it cost \$0.75?

(Versions of this question with different prices were also used.)

It may not look like it, but it turns out that this question is biased.

From a news article describing the study this question was used in:

Participants in marketing studies tend to misrepresent consumer behavior in ... hypothetical situations .... This misrepresentation is especially widespread in surveys that question willingness–to–pay, an attribute that measures how much consumers would pay for a given product ....

"Any economist will tell you that people have incentives to misrepresent their preferences in a survey," Lusk said. .... Economists call this tendency for survey participants to misrepresent their true opinions hypothetical bias.

In this study, Lusk assessed whether applying a technique called cheap talk could reduce consumer hypothetical bias when responding to survey questions about willingness–to–pay.

News article at:
`http://www.sciencedaily.com/releases/2003/10/031028061950.htm`.
Journal article:
Jayson L. Lusk, *Effects of Cheap Talk on Consumer Willingness–to–Pay for Golden Rice*, American Journal of Agricultural Economics **85**(2003), 840–856.

(More links will be on the course website.)

Hazard: Nonresponse.

Example: Telephone surveys. People who are away from home (on vacation?), or who hang up on telephone pollsters, may very well have different opinions than those who are home and are willing to respond.

In particular, what is likely to happen to a telephone survey about the National Do Not Call List?

Nonresponse (missing or unavailable data) can occur in other kinds of observational studies as well. For example, a remote weather instrument might malfunction occasionally, or break and not be fixed for a while.

Following links from some National Weather Service pages (such as `http://www.wrh.noaa.gov/pqr/snowdata.php`) gets links to weather data such as snow depth at various sites in the Cascades.

At `http://www.nwac.us/~nwac/products/OSOTIM` (Mt Hood/ Timberline Upper), "Wind sensors not heating and may rime".

At `http://www.nwac.us/~nwac/products/OSOCRA` (Crater Lake), "System unavailable". (Sometimes at other sites a measuring instrument will fail and not be repaired immediately.)

At those that record snow depth, such as `http://www.nwac.us/~nwac/products/OSOTML` (Mt Hood/Timberline Base), sometimes the snow depth numbers will be obviously ridiculous.

Hazard: Undercoverage.

Example: Telephone polls miss people who don't have telephones, or with unlisted numbers. (These days, I presume they miss people whose only phone is a cell phone.) A sample which automatically omits these people is biased. How badly biased depends on how many such people there are, and how much they differ from the general population on the issues the pollster wants to know about.

Previously (and this may still be true now), polls for presidential elections excluded Alaska and Hawaii. (Perhaps this was because phone calls to those places were more expensive?)

The reported margins of error in polls do *not* take account of either nonresponse or undercoverage.

Found at `http://www.hostedsurvey.com/article-satisfaction.html`:

"Let me give you an example of how fanatical good survey research people are. At the Survey Research Institute at University of Michigan, these are the types of stories they told about collecting data. (Stories are a good way of conveying cultural values.) One ongoing survey followed up with people at regular intervals, asking similar questions to assess changes over time. At one point, one respondent had been put in jail. The only visitors allowed were family members. So the woman surveyor told the jail she was his mother and went in and did the interview. In another case, a house address came up in a canyon in Los Angeles. When the surveyor went to the address, the house wasn't there! It had been in a mudslide and was gone. The surveyor was told to rent a jeep and go look for the house. When they [sic] found it, there were people in it and the designated one was interviewed."

Hazard: Response bias.

Examples of biases covered by this term:

- People may lie if asked about illegal behavior, unethical behavior, or about subjects which are normally not discussed in public.
- The tone of voice of the interviewer may (unconciously) communicate approval or disapproval of some of the alternatives in a multiple choice survey question. (Do your attitudes toward genetically modified foods affect your intonation if you read aloud the question about Golden Rice?)
- On certain subjects, men and women might be more willing to respond to an interviewer of the same sex (or, I suppose, to an interviewer of the opposite sex).