Read M&M §3.1 and M&M §3.2, but skip bit about tables of random digits (use Minitab). Read M&M §3.3 and M&M §3.4.

A little bit about randomization in experimental design. Simple random samples. Comparison between sampling and full enumeration. Sampling variability.

1. Why randomize?

Chapter 3 of Moore and McCabe introduces one of the key ideas of modern statistics, the paradoxical idea that one should inject more randomness into the process of producing data in order to be more confident about the conclusions that can be drawn. For experimental design, the idea takes the form of pre-experimental randomization for the pattern in which treatments are applied to experimental units. In sampling theory, it takes the form of a random choice for the subset of a populations that is examined in detail.

In a nutshell, the reason for adding randomness is to convert unknown or unknowable systematic differences (between experimental units or members of a population) into random quantities whose behavior is controlled by the laws of probability.

2. Randomization in experiental design

In the early 1950's a vaccine was developed against polio. To test its effectiveness, a large-scale field trial was carried out. Some school children were to receive the new vaccine, and some were to be given a placebo, with the choice of treatment for each child made by the 'toss of a fair coin'. Neither the children nor their parents nor the doctors who were to assess the effectiveness of the vaccine were told who received vaccine and who received placebo.

Why were these precautions necessary?

(i) Why not vaccinate all children in the experiment then see whether the polio rates were lower than the previous year?

(ii) Why not allow each child's parents to decide whether the child is vaccinated or not?

(iii) Why not vaccinate all children from one area of the country and leave the others unvaccinated?

(iv) Why not vaccinate all second graders in the study and leave the first and third graders unvaccinated?

(v) Why bother with the placebo?

(vi) Why not let local school officials decide who should be vaccinated and who not?

(vii) Why keep the nature of the treatment a secret from the doctors who were to judge the effectiveness of the vaccine?

I have extracted these questions and the background information from the first chapter of "Statistics" (2nd edition) by Freedman, Pisani, Purves, and Adhikari. (The book is one of the popular competitors to the Moore & McCabe text.) See Section 3.3 of M&M for answers to similar questions.

3. Sample versus complete enumeration

	counts	cum.pct
Under 1	1782	1.4
1 and 2	4439	4.8
3 and 4	4000	7.8
5	1855	9.3
6	1732	10.6
7 to 9	5118	14.5
10 and 11	3442	17.1
12 and 13	2929	19.4
14	1391	20.5
15	1403	21.5
16	1400	22.6
17	1445	23.7
18	2615	25.7
19	3651	28.5
20	3735	31.4
21	3694	34.2
22 to 24	8467	40.7
25 to 29	14345	51.7
30 to 34	12004	60.9
35 to 39	9174	67.9
40 to 44	7475	73.7
45 to 49	5643	78
50 to 54	4410	81.4
55 to 59	4062	84.5
60 and 61	1604	85.7
62 to 64	2592	87.7
65 to 69	4401	91.1
70 to 74	3963	94.1
75 to 79	3506	96.8
80 to 84	2232	98.5
85	1965	100

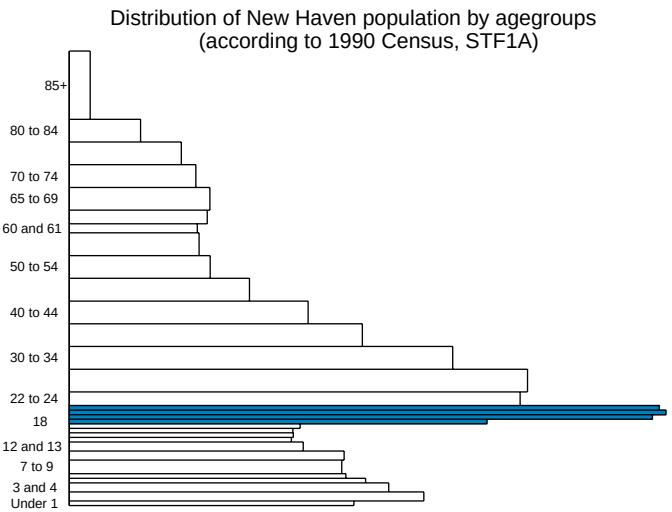
Every ten years, the Bureau of the Census tries to record information about the population of the United States. Some of the information is based on questions asked of everyone; some is based on questions asked of only a subset of persons (those in households that receive the 'long form' of the Census questionnaire).

For example, the short form asks for the age of every person in the household. The Bureau publishes tables for various geographical regions (states, counties, towns, ...) giving the breakdown of the population into 31 age groups: under 1, 1 and 2, 3 and 4, ..., 85 and over.

The table at left shows the breakdown by age for the 130474 persons counted in the City of New Haven. The counts come from "table P011 of Summary Tape File 1A", which I pulled down from the Census Bureau web site. The third column shows the cumulative percentages of the population in the age group or a younger age group. For example, 40.7% of the population of the City was 24 or younger, and 51.7% was 29 or younger. The median age must be somewhere in the the 25 to 29 range.

Is there any way to determine the mean age of the population from the tabulated data? How would one handle the contribution from the open-ended 85+ range? Why does the Bureau present the data in this grouped format, and why does it break ages into these particular 31 categories?

There are other ways to look at the data. The histogram at left (or should it be called a barplot?) was derived from the tabulated data. The area of each bar in the barplot is proportional to the fraction of the population in each of the ranges. So for ranges of 5 years I had to divide the count by 5 to scale back to counts per year, and so on. For the top category, over 85, I arbitrarily spread the count over 15 years.

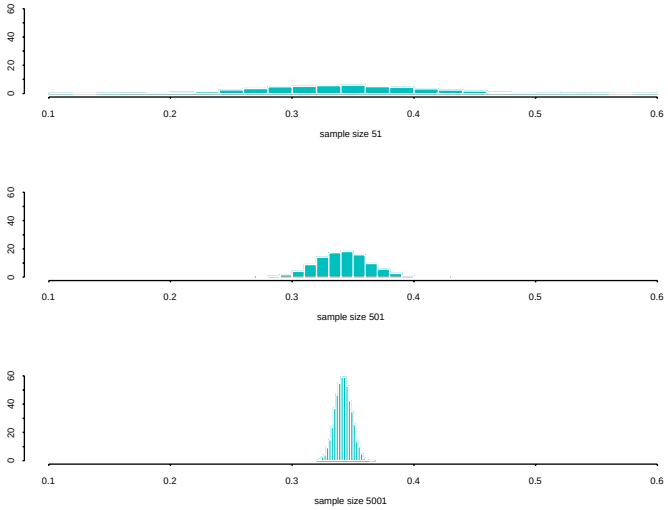


What does the spike for the ages 18 through 21 reflect? Do you think the distribution of population by age for neighboring towns would be similar?

Now suppose the Census had not been taken but that we had wanted to know what fraction of the New Haven population was 21 years of age or younger. We could have estimated the fraction by taking a simple random sample (SRS) and calculating the fraction under-21 in the sample. Why does a sample tell us about the whole population? How large a sample do we need?

In theory it is straightforward to determine what we could have learned from a sample of size n. Write out a list of all possible samples of size n. Calculate the proportion under-21 in each sample and write it on a slip of paper. Put all the slips in a very large hat. Taking a sample is then like drawing a slip at random from the hat, assuming that each slip has the same chance of being selected.

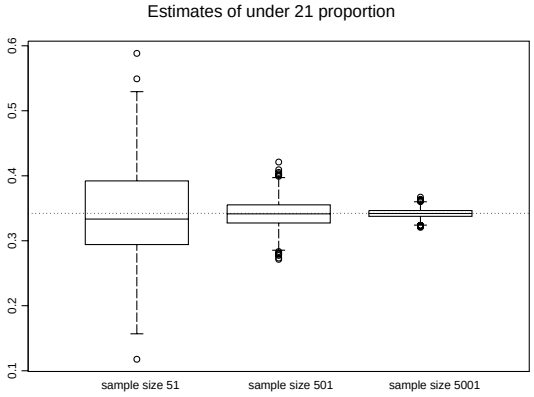
In fact there are too many different possible samples, even for quite small n, to allow a complete listing of the distribution of the under-21 proportions calculated from the samples. Instead, I used a computer to take many different samples, for three different values of n (51, 501, or 5001) from the population described by the tabulation: for each choice of n I repeatedly (4000 times) generated samples from the population, recording the proportion under-21 for each sample. (Actually I cheated slightly. My samples were not quite SRS's, but the difference is unimportant for the present discussion.) For each n I then had the results from 4000 different samples.



The three histograms show the distributions of the under-21 proportions for the 4000 samples. Probability theory assures me that the histograms are close the histograms we would get if we ground out the under-21 proportions for all possible samples of size n.

I have drawn the three histograms on the same scale to allow comparison between the distributions for different n. Notice that each histogram is roughly normal, centered at the true value (.342) shown in the tabulation. The spread decreases as the sample size increases. For the samples of size 5001, most of the under-21 proportions lie very close to the true value. By sampling less than 4% of the whole population we can, with high probability, get an estimate that lies quite close to the true under-21 proportion of the population.

The increasing concentration about the true value is perhaps more clearly seen in the following boxplots that show the results of my simulation. (The dotted horizontal line marks the true proportion of under-21 persons in the New Haven population.)



For each of my samples I also calculated the age group in which the median of the sample lay. I couldn't be more precise about the location of the median because of the way the tabulated data were grouped. Of course in a real sampling situation one would try to determine the exact age, and not just the age group, for each person in the sample. The following table gives the percentage of the sample medians that fell in each age group.

	20	21	22 to 24	25 to 29	30 to 34	35 to 39	40 to 44
ss=51	0.3	1	8.7	51.4	32.4	5.7	0.5
ss=501	0	0	0	76.1	23.9	0	0
ss=5001	0	0	0	99.5	0.5	0	0

You should not place much faith in any value after the decimal point—the 4000 replications were not enough to estimate the percentages for the population too accurately. I have the left the decimal fractions to give an impression of how the sampling distribution for the median is spread out. Notice the increasing concentration in the correct category as the sample size increases.

4. How good is a sample?

In general we do not have data for the entire population to compare our sample values against. Indeed the whole point of taking a sample is to avoid the task of enumerating the whole population. How then can we know whether the sample is "representative" of the whole population? How can we tell whether our sample gives values close to the population values?

The answer is that we cannot know whether a particular sample is "representative" of the population. However, if the sampling is carried out correctly, we can appeal to probability theory for assurance that most samples will give values close to the population values.

As you will see in the next few lectures, it is possible to make probabilistic assertions about the behavior of samples even when we do not know the exact composition of the population being sampled.

Sampling is easier than complete enumeration. Sometimes a well designed sample can even do better than an attempt at complete enumeration, which would stretch resources too thinly. The decennial Census is a case in point. It is well nigh impossible to count everyone in the country by conventional means. It is generally accepted that some groups in the population suffer more heavily from undercount than others.

Many experts feel that sampling provides a better way to get at the 'hard-to-count' parts of the US population than the traditional method of repeated attempts to track down persons who do not return the Census questionnaire. Unfortunately, the basic issue has become lost in the political wrangling over attempts to reform the Census.

5. The controversy over Census 2000

Article I, section 2 of the US Constitution mandates that the population of each US state be determined every ten years:

> [3] Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number ...

The Fourteenth Amendment to the Constitution modified the apportionment of seats in the House of Representatives among the several States, requiring that it be "according to their respective numbers, counting the whole number of persons in each State, excluding Indians not taxed."

Current practice is controlled by Title 13 of the United States Code. In particular, Section 195 states:

> Use of sampling
>
> Except for the determination of population for purposes of apportionment of Representatives in Congress among the several States, the Secretary [of Commerce] shall, if he considers it feasible, authorize the use of the statistical method known as "sampling" in carrying out the provisions of this title [13 USCS §§1 et seq.].

In a 1997 report to Congress†, the Bureau of the Census commented on some shortcomings of previous decennial censuses:

> The national census count became more accurate with each consecutive census from 1940 to 1980. Although it surpassed all previous censuses in terms of design, execution and resources used, the 1990 census took a large step backwards in terms of accuracy. While the 1980 census had fallen 2.8 million people below an accurate count, the census count in 1990 fell 4.7 million people short, missing 1.8 percent of the population, according to demographic analysis estimates. Moreover, the undercount in 1990 was not spread evenly across the nation; children and minorities were disproportionately undercounted.

The Bureau was not alone in its concerns:

> In the wake of the 1990 census, there was a consensus among the Census Bureau, professional statisticians, and Congress that significant changes were required for the upcoming 2000 census; the Census Bureau could not continue to employ the methods it had been using. In 1991, bipartisan legislation passed unanimously by Congress and signed by President Bush directed the National Academy of Sciences (the Academy) to study "the means by which the Government could achieve the most accurate population count possible."

In response to the Academy report (and the recommendations from other study groups), the Bureau proposed extensive changes for Census 2000, including the use of "statistical sampling to account for those who cannot otherwise be accounted for." The Bureau further cited the Academy Panel on 'Census Requirements in the Year 2000 and Beyond', which concluded that,

> [i]t is fruitless to continue trying to count every last person with traditional census methods of physical enumeration. Simply providing additional funds to enable the Census Bureau to carry out the 2000 census using traditional methods, as it has in previous censuses, will not lead to improved coverage or data quality.

The apparent consensus of opinion disintegrated when it became clear that the proposed improvements might lead to a redistribution of House seats between states. The House of Representatives challenged the proposal in Federal court. The court (1998 U.S. Dist. LEXIS 13133) very recently decided that the proposal violates Title 13, and therefore ordered that "defendants are permanently enjoined from using any form of statistical sampling, including their program for nonresponse follow-up and Integrated Coverage Measurement, to determine the population for purposes of congressional apportionment." In short, unless the Supreme Court overturns the decision, the Census 2000 will be carried out by a method that is doomed to miss appreciable chunks of minority populations.

Read "The American Census: a Social History" by Margo Anderson (Yale University Press, 1988) to see how the current battle fits right in with the long history of wrangling over how to count the population.

---

† The Plan for Census 2000. Originally Issued July 1997. Revised and Reissued August 1997. All my quotes are taken from the Executive Summary in that report.