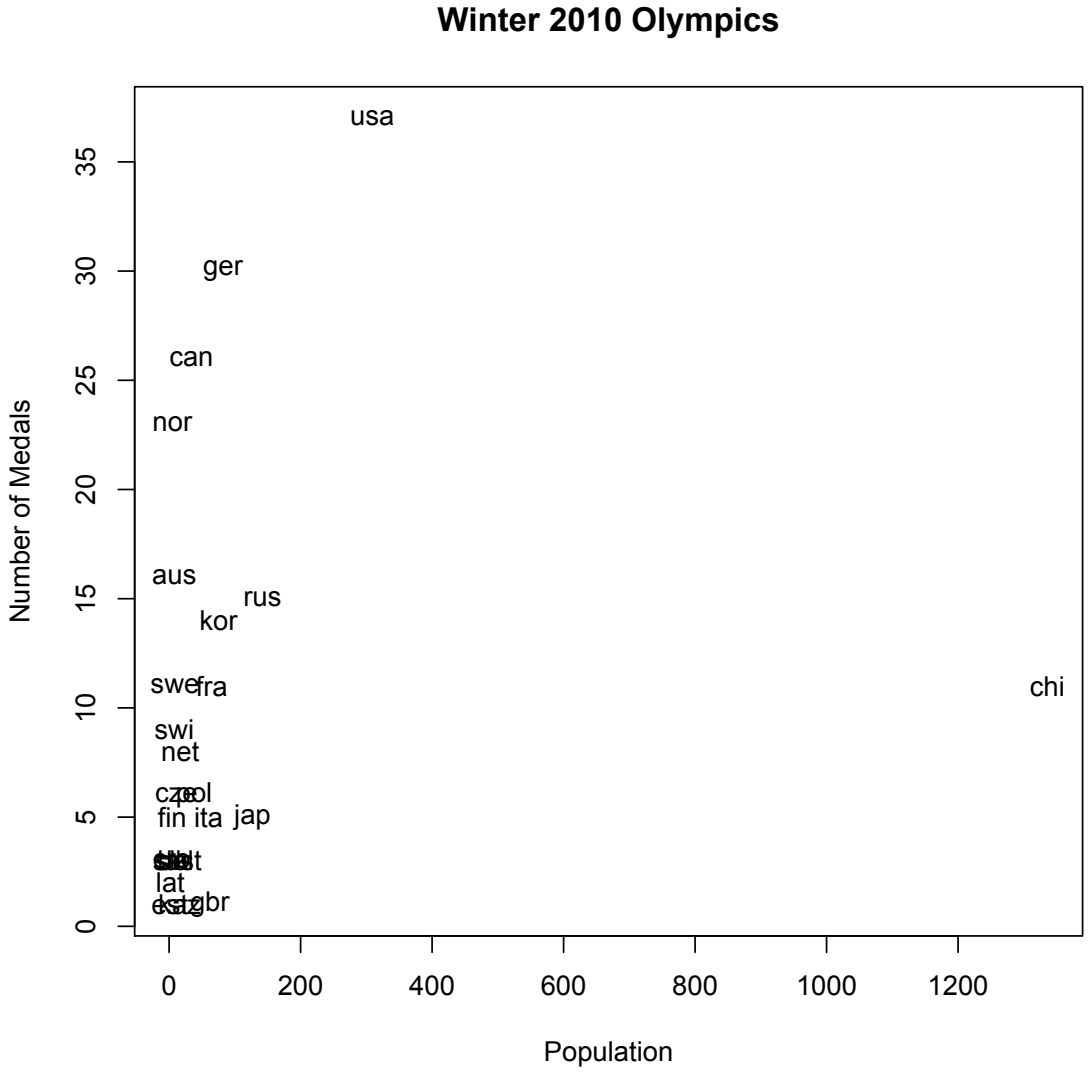
Week 7

Olympics: Does medal count depend on population size?
Sampling: Political Polls. Populations and Samples
Randomized Response Technique
Hill: Evaluating School Choice Programs
Heilig et al: Leveraging Chance in HIV Research
Assignment #7: Sample Midterm II, Due March 9, 4pm. (Is web-posted now.)

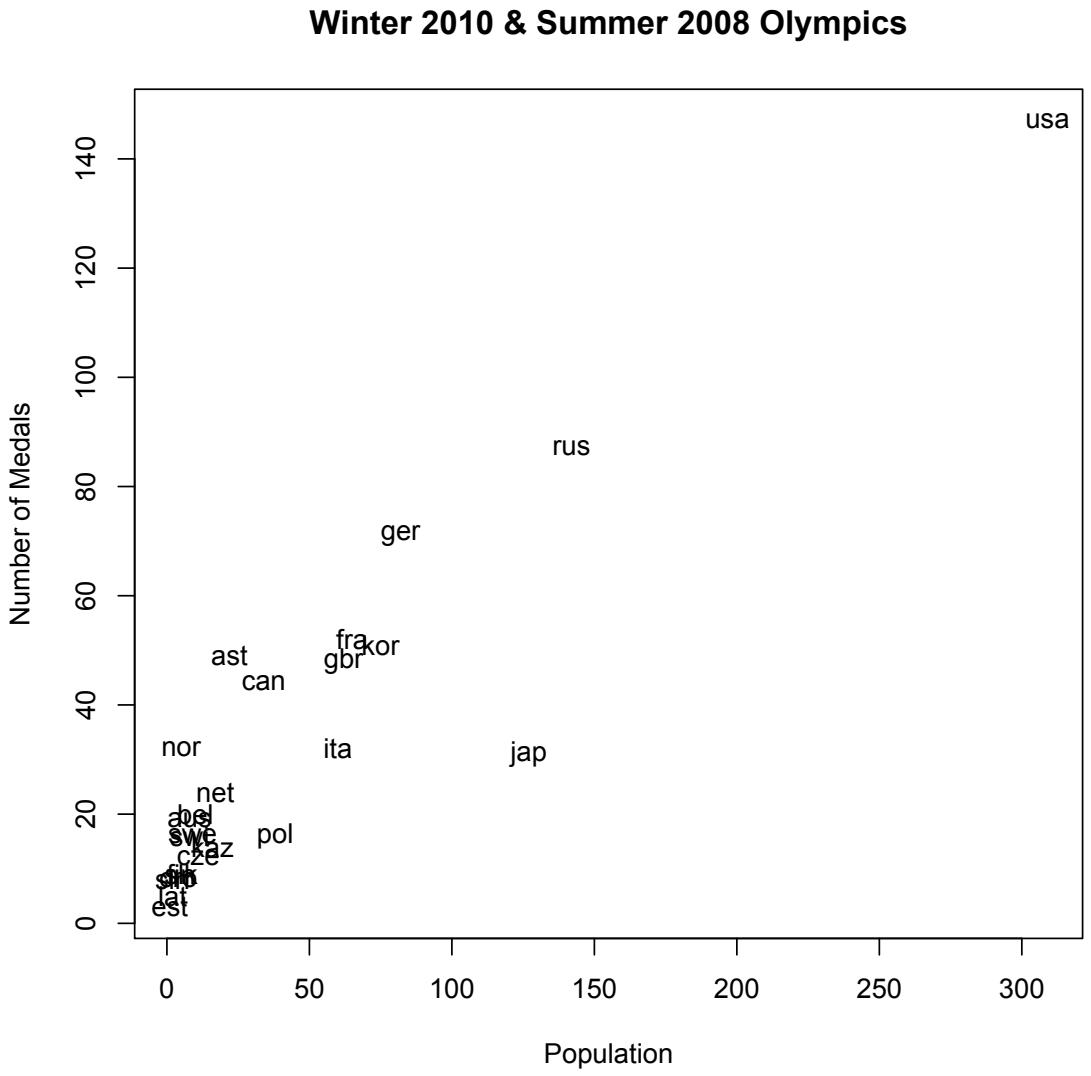
Olympics:

Q: Is it possible to compare a country's success in winning Olympic medals in a way that allows for the country's population size? (R program: medals.demo())

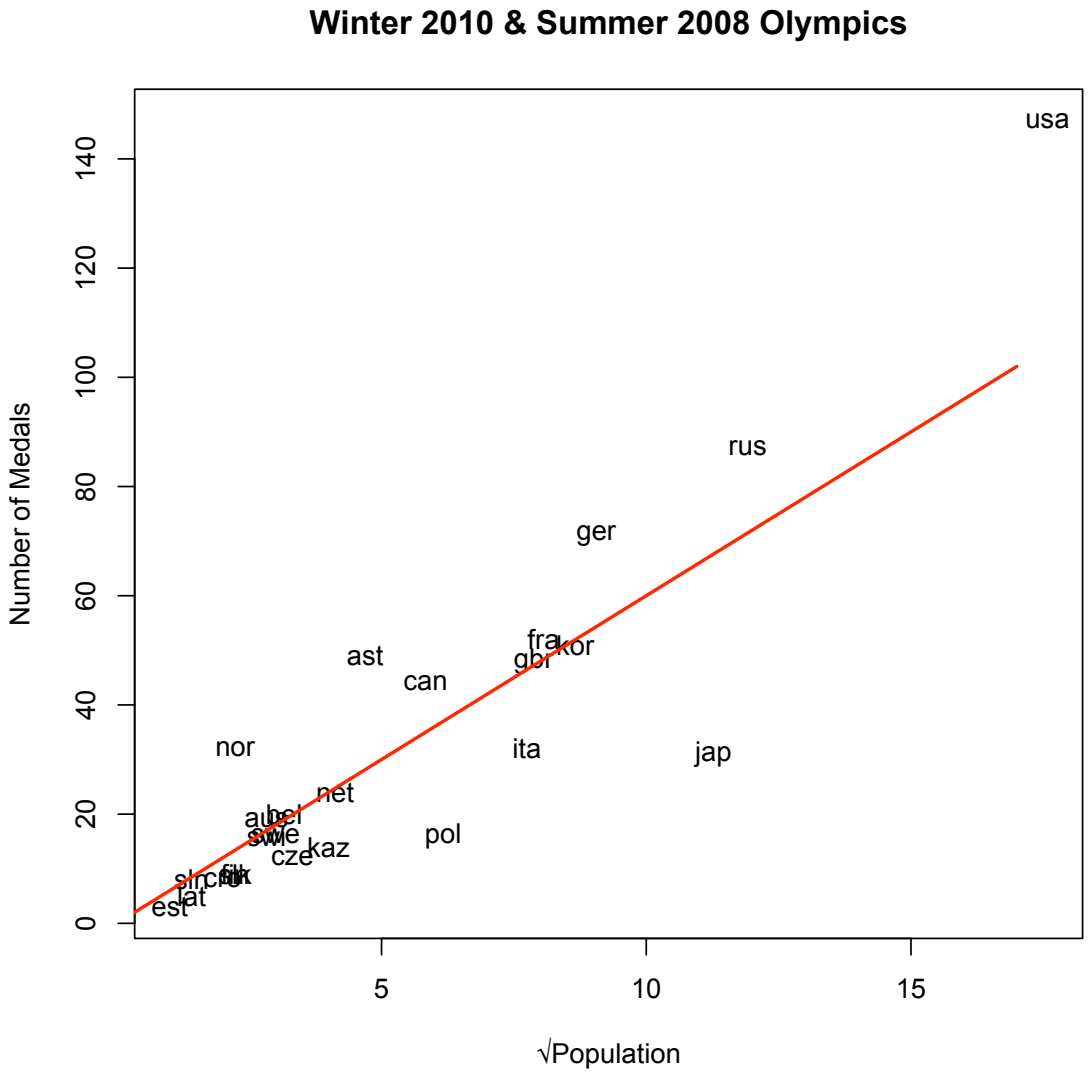
	country	medals.s	pop	gold	medals.sw
1	usa	37	309	9	147
2	ger	30	82	10	71
3	nor	23	5	9	32
4	can	26	34	14	44
5	rus	15	142	3	87
6	kor	14	75	6	51
7	aus	16	8	4	19
8	fra	11	65	2	52
9	swi	9	8	6	16
10	chi	11	1336	5	111
11	swe	11	9	5	16
12	net	8	17	4	24
13	cze	6	11	2	12
14	pol	6	38	1	16
15	ita	5	60	1	32
16	ast	3	22	2	49
17	slk	3	5	1	9
18	jap	5	127	0	31
19	lat	2	2	0	5
20	bel	3	10	1	20
21	cro	3	4	0	8
22	sln	3	2	0	8
23	gbr	1	62	1	48
24	est	1	1	0	3
25	fin	5	5	0	9
26	kaz	1	16	0	14



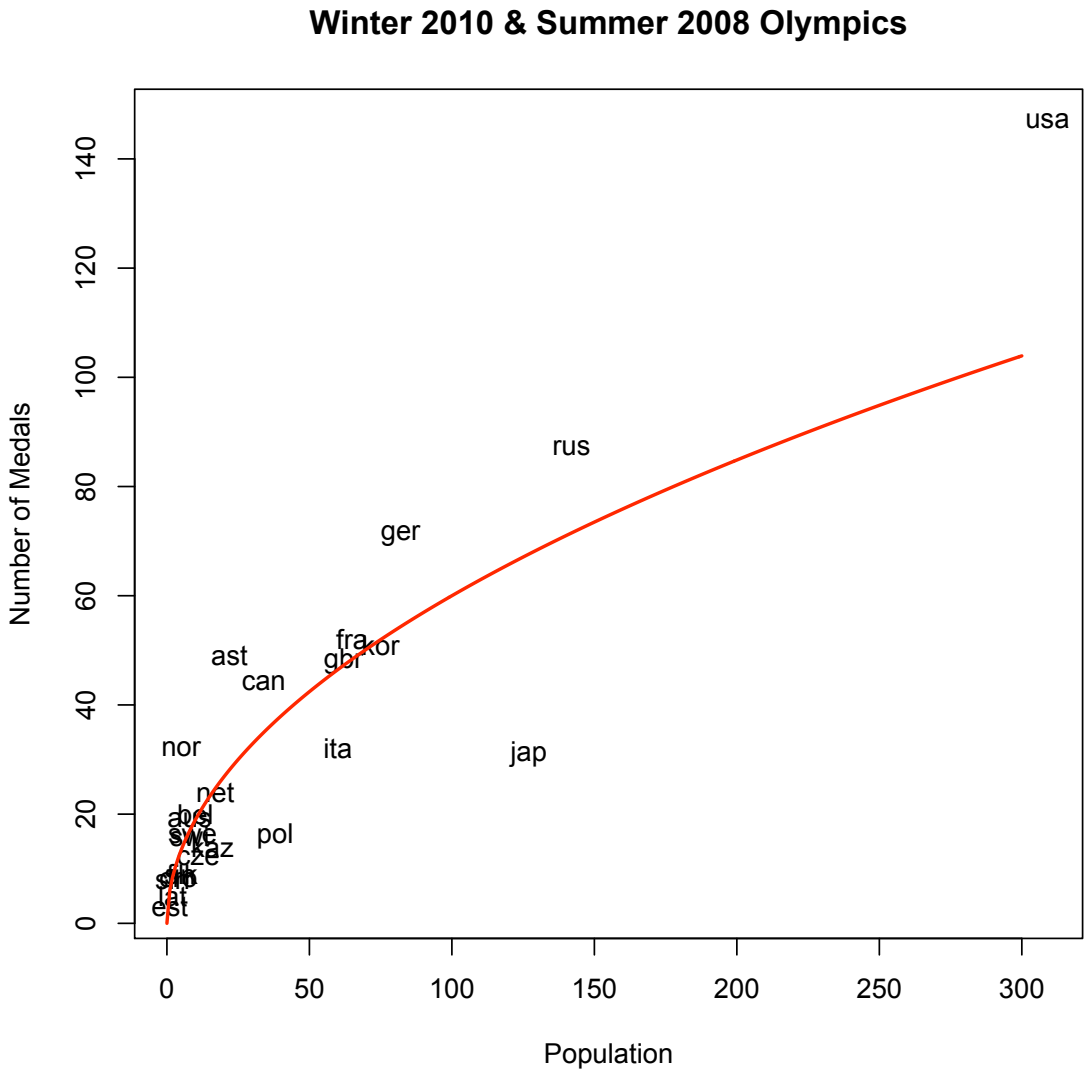
But China is so populous compared to the other countries in the winter Olympics that the relationship of medals to population is squashed. We put aside China for this further investigation of the relationship. The result is shown below.



The relationship is still a bit obscured by too many small countries in the lower left corner of the graph. Let's stretch it out by taking the square root of population. …



Now we have a clearer relationship. But is it legal to use √Population? A simpler (but equivalent) way to show the relationship is to allow the relationship to be curved. The identical fit can then be shown without the square root trick.

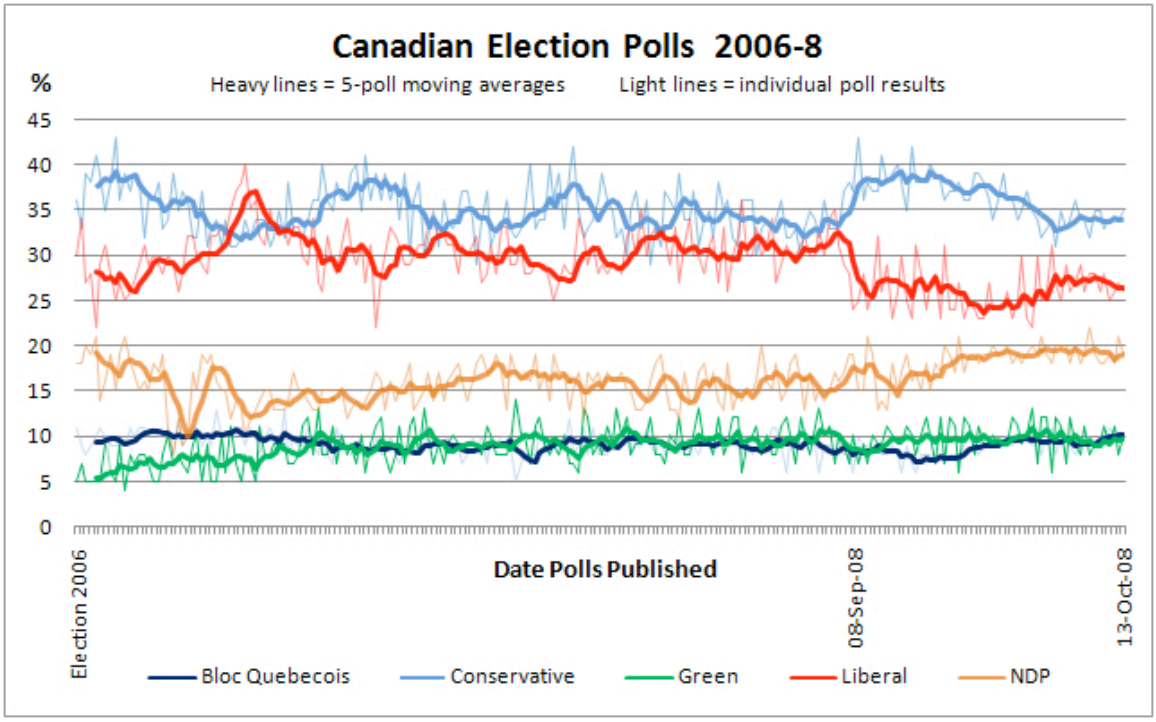


What use is this graph? Clearly there are important differences that we know about, such as per capita GNP, that would allow some countries to fund a more medal-winning team. But at least we can see the effect of population base and adjust for that in comparing countries. The residuals in the above graph are a measure of the extent that a country's medal count is not explained by population. And for countries like Norway, Canada, and Slovenia, and others, this is a useful thing!

Sampling:

Political Polls

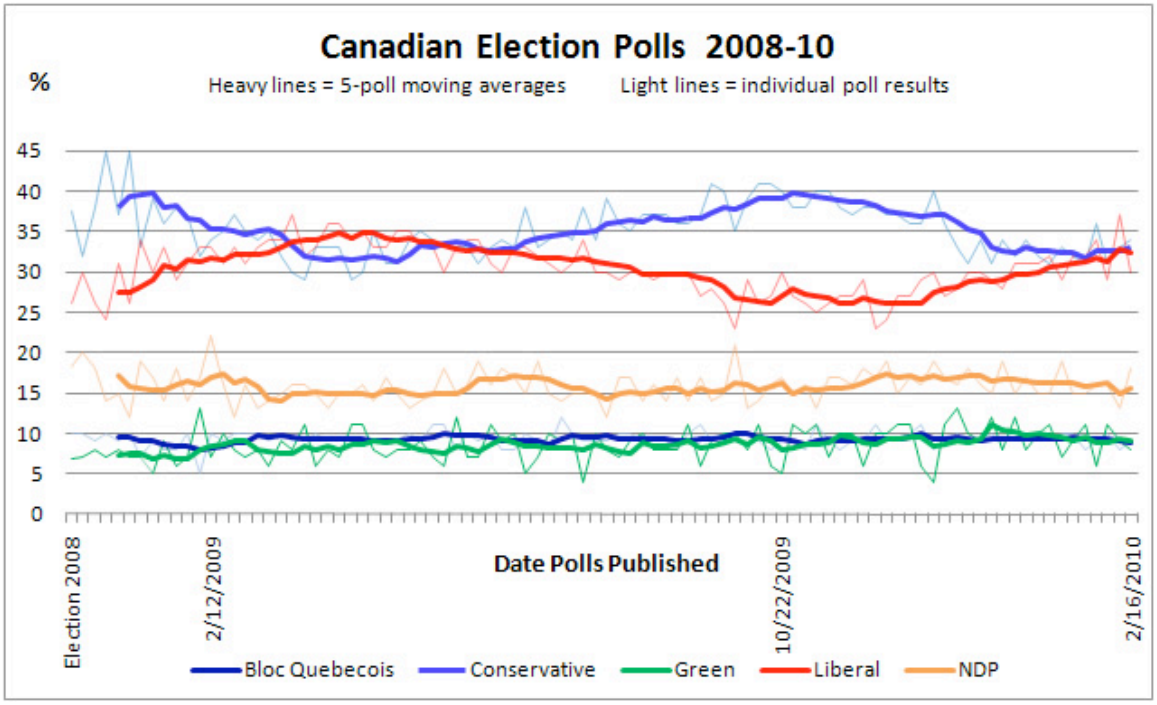
Just prior to the Fall 2008 election, the following graph was published on the cover of the magazine Liaison:



The graph was produced by Andrew Heard of the Political Science Dept at SFU.

The heavy lines are smoothes of the finer lines, using the moving average technique. The finer lines join up points from surveys done by various organizations in Canada. We know the result was a minority government of the Conservative party, which would be no surprise to someone looking at this chart before the election. (Canada's first-past-the-post election result system tends to amplify small differences in the popular vote, but not enough in this instance to give the Conservatives a majority.)

Here is an updated version of the above graph, also supplied by SFUs Andrew Heard:



Here is the question most important for our purposes:
Q: How can an opinion poll based on a few thousand responses be a reliable indication of the aggregate opinions of millions?

National political polls tend to sample about 2,000 people while the population of Canada is about 35,000,000. If these sample surveys produce reliable information, they are quite useful and even amazing, right?

Here is the key formula that explains this amazing result:

$$SD_{sample.mean} = \frac{SD_{population}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Here N is the population size, and n is the sample size.

Except for the last square root, this formula should look familiar. Remember that averages (=means) vary less that the things that are averaged, and the reduction in variability is a factor $\sqrt{n}$. We need to explain the sudden addition of that last term $\sqrt{\frac{N-n}{N-1}}$.

When we talk about sampling from a normal probability distribution, or from any probability distribution, we usually implicitly assume that the removal of an item from the population (drawn into the sample) does not change the make up of the population – just as if the population were infinite, or, if we were sampling *with*

replacement. But in polls from real populations (as opposed to abstract ones defined by a probability distribution) we might need to consider that we are really sampling *without replacement.*

Q: If we choose a sample without replacement from a real (and finite) population, and we decide to sample the whole population, how variable would the sample mean be in such a process?

Common sense answers this question, but it is comforting that the above formula also provides the same answer!

Now, in political polls like the ones mentioned in the intro above, the sample size n is much smaller than the population size N. In this case, what happens to the factor $\sqrt{\frac{N-n}{N-1}}$? Clearly, it will be very close to 1, and multiplying it by the more familiar part of the formula will not change anything. (rewriting $\sqrt{\frac{N-n}{N-1}}$ as $\sqrt{1-\frac{n-1}{N-1}}$ makes this even more clear). A rule of thumb is, when n/N is less than 1/20, ignore the factor. The factor is commonly called "the finite population correction factor" but perhaps a more descriptive name for it would be " the correction factor for sampling without replacement". The two names are equivalent – as a bit of reflection reveals.

One important conclusion from the discussion is that, in political polling, one can ignore the factor $\sqrt{\frac{N-n}{N-1}}$, and so one is left with the usual square root relationship between the population SD and the SD of the sample mean.

Q: Political polls usually end up with results expressed in term of proportions of percentages, whereas means are usually computed for quantitative data (like income, imports or GNP). How does the above discussion apply to percentages?

Remember that a proportion is an average of 0-1 data and any proportion can be considered to be based on 0-1 data (proportion of 1s). So the above discussion applies.

One other technical item needs to be re-introduced here. A population of 0s and 1s has an SD that can be computed from a short-cut formula which depends only on the proportion of 1s, p, in the population.

$$SD_{0-1\,population} = \sqrt{p(1-p)}$$

and so the SD of a sample mean (in sampling with replacement, as is most usual) is

$$SD_{sample.proportion.\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Note that $\hat{p}$ is an estimate based on the proportion of 1s in the sample, but p is the actual (and usually unknown) proportion of 1s in the population. When we want $SD_{sample.proportion.\hat{p}}$ and do not know the true population p we usually plug in our estimate $\hat{p}$ so we estimate the $SD_{sample.proportion.\hat{p}}$ by $\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$.

So here is an example: Suppose a population of 10,000,000 people is sampled with a sample of size 1000. Since 1000/10,000,000 is less than 1/20 we ignore the correction factor. In our sample of 1000 we find 423 1s and 577 0s. So our estimate of p is $\hat{p}$=423/1000 = .423 and the accuracy of this estimate is indicated by the

$$SD_{sample.proportion.\hat{p}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = \frac{\sqrt{.423(.577)}}{\sqrt{1000}} = .016$$

We can report this in various equivalent ways:

1) proportion of 1s in the population is estimated as .423 with SD = .016
2) percentage of 1s in the population is estimated as 42.3% with SD 1.6%
3) percentage of 1s in the population is estimated as in (39.1%, 45.5%) 19 times out of 20.
4) A 95% Confidence Interval for the population percentage of 1's is (39.1%,45.5%)

Q: When we plug in $\hat{p}$ for p in the SD formula for $\hat{p}$, we change an exact formula to an approximate formula. Is there a more conservative procedure – what is the largest possible value of $\sqrt{p(1-p)}$?

The largest possible value of $\sqrt{p(1-p)}$ is ½, the value when p=1/2 so a conservative estimate of $SD_{sample.proportion.\hat{p}} = \frac{1/2}{\sqrt{n}}$.

Q: Using the conservative formula, how big does a sample have to be (in a big population) to estimate the sample proportion to within 3 percentage points?

We want the width of the 95% Confidence Interval for a proportion to be .03 percentage points, so 2 times $\frac{1/2}{\sqrt{n}}$ must be .03. Thus $\sqrt{n}$ must be $\frac{1}{.03}$ and n must be at least 1111.

So?

Political polls are usually based on a few thousand people. If you say "why more than 1111?", the answer is that real-world political polls are not simple random samples but more complicated designs, and to get the same information, these more complicated samples have to be bigger than a simple random sample. But the advantage of random sampling is still an important part of these more complex designs.

Q: Did we assume Normality in the above discussion? (Yes. The factor "2" for a 95% Confidence Interval is based on the Normal distribution, and the justification for it is the Central Limit theorem, which says that averages (and proportions) have approx N distributions, with the approx getting better and better as the sample size increases.)

Q: When a survey respondent is asked their political preference, does the respondent worry about the confidentiality of the information, and will the respondent answer truthfully without regard to confidentiality?

A: Y usually, N not always.

Randomized Response Surveys

How to get reliable confidential information without anonymity!

Every one in the class uses a coin to toss, resulting in H or T. Each student should keep the outcome of the toss private.

According to the outcome answer either question 1 or 2 by raising your hand for an answer "Yes" when asked by the prof.

Q1. (to be answered if your toss was a head). Is your coin a head?
Q2. (to be answered if your toss was a tail). Are you a regular user of marijuana (usually once per week or more)?

To raise your hand in this situation, you need to be assured that your peers will not know if you got a H or T, and so a yes could mean that you did get a head **or** it could mean that you are answering Yes to the marijuana question. But nobody except you will know which question you are asking.

I did this in 2002 with the STAT 100 class of about 100 students. Then I had 60 Yes responses (60 hands up). Since about half of these were Q1 answerers, only about 10 of the other 50 were answering Yes to Q2. So the percentage of regular marijuana users was estimated to be 20% (10 out of 50).

Was the confidentiality of response really protected? Consider the 60 answering Yes to whichever question they answered. There are 5 yesses to Q1 for each 1 yes to

Q2. So a person answering Yes had a probability of 5/6 or 83% to be answering the innocuous Q1.

Q: Is Randomized Response Technique just a parlour game or a serious confidentiality strategy?

The government and many other large companies have huge data basis that they may wish to make available to legitimate users, but they do not want to compromise the confidentiality of the information by providing enough info about individuals that some individuals are identified by their data even if names and addresses are removed. One strategy that is used is that the data provided is messed up with random errors so that individual files are modified but the aggregate of the files has the statistical information wanted by the user. So variation so the technique are actually used.

Sampling Studies

The basic idea in sampling is that it is a lot easier and cheaper to base information about a population on a sample from the population, rather than measure the entire population. And, if the sample is a random sample, or at least using a design based on randomization, then the sample itself provides information about the accuracy of the sample as a representative of the population. As we have seen with political polls, when the population is huge, sampling is a particularly efficient way of getting population information.

However, in more complex situations, sampling studies can be a huge challenge, and are only attempted when the goal is very worthwhile. The two readings described below describe three sampling studies that sought important information, and in spite of the expertise in the design of the studies, many practical problems were encountered. The discussion of the goals, and the efforts to overcome the obstacles, is instructive for anyone trying to judge the validity of findings from such studies.

Hill: Evaluating School Choice Programs (pp 69-87)

A question many people would like to know the answer to is: Are private schools better than public schools? But this is too big a question. One that might have an answer is "Are private schools better at teaching the three Rs than public schools"? The Hill article describes a serious effort to find an answer to this question.

The article first discusses why the studies that have been done so far do not answer the question satisfactorily – mainly because they are observational studies. Often experiments with human subjects are unethical and cannot get approval. But the article describes a situation in which a "natural" experiment is possible meaning that it is possible to randomly assign the proposed "cause" (private school vs public school) to the experimental units (students accepted into the scholarship program)

so that the other possible influences would be balanced between the two comparison groups (that is, the group of students assigned to the private school and the group of students assigned to the public school.)

But there were practical problems that made the comparison a bit ambiguous even though the design of the study was set up as an "experiment". There were *missing data*: although the students were to be tested after three years in the assigned school system, not all of them showed up for the testing. Another problem was *non-compliance*: not all students who were assigned to one type of school accepted that assignment. Finally, there were the subjective choices made my the evaluators when certain variables were defined for the analysis. The article discusses the ways in which the investigators tried to lessen the impact of these deviations from a purely objective experiment, and they no doubt were able to lessen some of the potential problems, but as usual the practical problems in studies of human subjects always leave the final inference to be to some extent subjective.

In so far as the problems can be ignored, the final result was that the public school system did seem to help the children from low-income families, but only the ones from African-American families. This conclusion left lots of room for further studies!

Heilig et al "Leveraging Chance in HIV Research" (pp 227-242)

In the Hill study on school types, the main problem was to get the human subjects to do what they were assigned to do. But the list of experimental units was readily available. In this HIV study the main problem was to get a list of the study subjects, namely the young "men who have sex with other men" in one survey and "children born to mothers with HIV" in the other survey.

In the Young Men's Survey (YMS), a list of meeting venues for homosexual men was constructed and a random subset of these visited for information collection. All eligible subjects at the selected venues were asked to provide information for the survey. The times of the visits were also randomized. These randomization strategies allowed for scale-up of the information to the entire population of interest. For participants agreeing to participate, blood testing was done to determine the presence of HIV. Table 1 provides some results (p 233). Using the odds ratio, it was determined that black men in this study population had a significantly higher risk of HIV than white men in the same population.

[Note on terminology:

Incidence: number of **new** cases per year per 1000 population
Prevalence: number of **existing** cases per year per 1000 population]

In the HIVNET Clinical Trial in Uganda, the aim was to compare the effectiveness of two drugs, ZDV and NVP, in prevention of HIV transmission from mother to infant.

The design of a classical clinical trial includes the following elements: *double blind* treatments, including the use of a *placebo,* and *randomization* of treatments to cases. (You need to understand these terms and why they are important.) However, the Uganda trial did not use the double-blind strategy. The Thailand and US/France trials did use the double blind strategy.

The results of the three studies are summarized in Table 3 on page 240. The three trials have comparable results: The NVP drug is superior to the ZDV in this context.

On page 239, mention is made of a method of "survival analysis" for "censored" data. We will explore this technique in Week 9 of the course.

Reminder: Be sure to access Assignment #7 from the course web page.