Political Science 100a/200a

Fall 2001

The central limit theorem and the law of large numbers[1]

1 Estimating a population statistic from a sample statistic

Question 1: You want to predict the outcome of an upcoming presidential election, but you lack the time and money to interview every single possible voter about his or her vote intentions. So you decide to interview of a random sample of registered (possible) voters. Can you use the results of this sample to generalize about the population, and how accurate is this method likely to be?

Question 2: Suppose you want to know the average and standard deviation of life expectancy across countries in a particular year, *but the only way to get the data is a painful and difficult process of consulting records country-by-country.* You do not have time and research money to do this for all countries in a given year. What should or can you do?

- Suppose you decide instead to draw a random sample of 25 country names, and to collect life expectancy data for these. Can you use this sample to estimate the average and standard deviation of life expectancy for the population?
- Since we have the actual data on life expectancy for (almost) the full set of countries in 1994(?), we are in position to evaluate how accurate and reliable this procedure would be.
- So let's try it: **use lifeexp, sample 16, sum, graph lifeexp**. Let's *pretend* that this is the sample of 25 countries that you construct.
  - Note that the distribution of the sample *values* looks sort of like the distribution of the population values. The idea is to use this sample to make inferences about the population (or random process) from which it is drawn.
  - So use **sum** to get basic descriptive statistics $\bar{x}$ and sd on the sample and then compare to population values. How do they compare?
- In reality, we *only ever see the sample we are working with*, so we can't know how far our estimates are from the truth.

[1]Notes by James D. Fearon, Dept. of Political Science, Stanford University, November 2001.

- But we can perform an experiment with Stata to see how "good" is the sample mean, $\bar{x}$, *in general* as an estimate for the population mean μ.
- We can write a program in Stata that will
    1. Draw a random sample of 25 countries and compute the mean of life expectancy for the countries in this sample.
    2. Record this result in a new data file as an observation (a case).
    3. Repeat this procedure as many times as we like, thus building up a data set that consists of (say) the mean life expectancies for each of the 25 country samples we drew.
    4. We can then look at the distribution of these estimates based on 25-country samples, to see how well they match up with the true mean in the population. The spread (or standard deviation) of this distribution will give us an indication of the likelihood that an estimate based on a 25 country sample is way off. Ok?
- **NB**: Of course, *in reality*, you will only ever have a *sample*, and not the population or direct access to the social or political process you are trying to draw an inference about. Here we are doing an *experiment* where we imagine that we have the population and then see how good are estimates of the population are based on smaller samples. Clear?
- Show **xmean.do**, discuss ... run **simul xmean , reps(.)** for 10, 50, 100, 1000, 10000 replications.
- Compare the distribution of our 25-country sample means to the distribution of **lifeexp** in the original sample. What do you notice?
- Several features of this experiment turn out (and we will show) to be entirely general – they would occur (almost) no matter what the distribution of the population variable of interest. These features are:
    1. The distribution of the sample mean is *centered on* the true mean in the population. This is called *unbiasedness*, and it is one natural criterion for a sample statistic to be a good estimate of a population statistic.

       Def$^{\underline{n}}$: A sample statistic b is an *unbiased estimator* for a population parameter β if $E(b) = \beta$.
    2. The distribution of the sample mean is approximately normal. This is immensely useful because it allows us to easily figure how much uncertainty attaches to the particular estimate we get – we make use of the standard normal curve.

       The approximate normality of the sample mean is explained by the central limit theorem, which says that the sum of a large number of of random variables has an approximately normal distribution *almost no matter what is the distribution of the component random variables.* (That is, as you add more and more random variables

together, the distribution of their sum more and more closely approximates a normal distribution.)

3. The standard deviation of the sample mean is smaller than the standard deviation of the population variable. In fact, we will show that the standard deviation of the sample mean is $\sigma/\sqrt{n}$, where n is now the size of the sample being drawn and σ is the sd of the population. (Show that this works for cases above, where $n = 25$.)

   So, as we increase the sample size, our estimate of the population mean becomes *much* more precise, and this doesn't depend in any way on the size of the population (discuss relevance to polling ...).

2 Explaining the simulation results

- Suppose we draw a random sample of values $x_1, x_2, x_3, \ldots, x_n$ from some population (e.g., life expectancies). Recall that the sample mean is defined as

$$\bar{x} = \sum_{i=1}^{n} x_i = \frac{1}{n}\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{x_1}{n} + \frac{x_2}{n} + \ldots + \frac{x_n}{n}.$$

- The experiments above suggested how we can think of the sample mean $\bar{x}$ *as a random variable*. How? (This is very important to understand.) In any given social science research problem, we only see ONE sample mean, but we can think about it as though it were a draw from a distribution of possible values, i.e., a random variable. Examples:

  1. In the life expectancy example above, there were 157 some countries with data on life expectancy in the population. How many possible 25-country samples can be drawn from this set? $\binom{157}{25}$.

     So let our sample space S be *the set of all possible 25-country samples from the population of 157.* With random sampling all draws are equally likely.

     Each different draw implies a sample mean $\bar{x}(s)$, where $s \in S$ is a particular sample. Thus, *the sample mean is a random variable*, a function that assigns a number to each point in a sample space. Clear? (this is important). Keep in mind that we only observe one sample, the realized value for our particular draw.

  2. There are some z million registered voters (let's suppose). We draw a random sample of 2500 and ask them about their vote intentions. There are $\binom{z}{2500}$ possible samples (a huge number), and these comprise our sample space. The sample mean of (say) "intention to vote for Gore" (1 or 0) is a thus random variable, $\bar{x}(s)$.

  3. You want to estimate the probability that an oddly shaped die will turn up an ace (a 1). You roll it 100 times and use the average number of aces (the sample mean) as an estimate. Note that here the population and the sample space from which

the sample is drawn is infinite. But no matter: Here the sample mean is a random variable in the sense of being the result of well-defined chance process.

4. You want to estimate the probability that a 40-year-old smoker has cancer, and to do so you draw a random sample of 40-year-old smokers. What is the "population"? All current 40-yr-old smokers? All current and future?

- If we can think of the sample mean $\bar{x}$ as a random variable, we can also think about its expectation or expected value. What is $E(\bar{x})$?

$$E(\bar{x}) = E(\frac{x_1 + x_2 + \ldots + x_n}{n}).$$

- But each of these particular values, x_1, x_2, ..., is a draw from the population, i.e., a sample of size 1. So we can treat each particular value as a particular realization of the random variable that is the value of one draw from the population. Call this random variable X_i for the ith draw. Next, what is the expectation of a sum of random variables multiplied by a constant?

$$E(\bar{x}) = \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)).$$

- But what is $E(X_i)$? What is the expected value of a single draw from the population? Just the expectation of the random variable X, or $E(X) = \mu$, the population mean. Thus,

$$E(\bar{x}) = \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

- This explains one of the simulation results above: the expectation of the sample mean is the population mean, or in other words, the distribution of the sample mean is centered on the true, underlying mean of the population. In other words still: *the sample mean is an unbiased estimate of the population mean.*

- Notice that this argument did not depend in any way on the particular distribution of values in the population (the random variable X).

- So thinking about the sample mean $\bar{x}$ as a random variable we have established that $E(\bar{x}) = \mu$. We can also ask about the *variance* and standard deviation of this random variable:

$$\begin{aligned} var(\bar{x}) &= var(\tfrac{X_1+X_2+\ldots+X_n}{n}) \\ &= var(\tfrac{X_1}{n} + \tfrac{X_2}{n} + \cdots + \tfrac{X_n}{n}) \\ &= \tfrac{1}{n^2}var(X_1) + \tfrac{1}{n^2}var(X_2) + \ldots + \tfrac{1}{n^2}var(X_n) \end{aligned}$$

- Be sure you understand what happened in the last step. Two properties of the variance of a random variable were used here. What were they? Why could we "take variance through" in this manner?
- What is $var(X_i)$? Each individual value is a draw from the population "box", so it has variance equal to the population variance σ^2. Thus ...

$$\begin{aligned} var(\bar{x}) &= \tfrac{1}{n^2}var(X_1) + \tfrac{1}{n^2}var(X_2) + \ldots + \tfrac{1}{n^2}var(X_n) \\ &= \tfrac{1}{n^2}\sigma^2 + \tfrac{1}{n^2}\sigma^2 + \ldots + \tfrac{1}{n^2}\sigma^2 \\ &= \tfrac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \tfrac{1}{n^2}n\sigma^2 = \tfrac{\sigma^2}{n}. \end{aligned}$$

- So (very important), we conclude that the variance of the sample mean $\bar{x}$ is σ^2/n, and the standard deviation of the sample mean is $\sigma/\sqrt{n}$.
- illustrate that this is approximately correct with simulation results from Stata ...
- So what happens to the standard deviation (dispersion) of the sample mean as the size of the sample (n) gets larger?
- Why is this happening? What is the intuition? Friedman has a nice way of thinking about it in terms of the *sum* of draws from box model (i.e., random variable), instead of the *mean* (average) of the draws:
  - The expectation of the sum is $E(X_1 + X_2 + \ldots + X_n) = n\mu$.
  - The variance of the sum of draws is $var(X_1 + X_2 + \cdots + X_n) = var(X_1) + var(X_2) + \ldots + var(X_n) = n\sigma^2$. So the *variance* of the sum of the draws increases linearly with n.
  - But standard deviation is the square root of variance, so the standard deviation of the sum of the draws (which Friedman calls the "standard error") is $\sqrt{n}\sigma$.
  - Thus the standard error of the sum of the draws increases only with the *square root* of the number of draws.
  - And accordingly, the standard deviation of the average of the sum of the draws is $\sqrt{n}\sigma/n = \sigma/\sqrt{n}$, as shown above.
  - e.g.: flip a coin 100 times and count the number of heads. The expected number is 50, and the standard error is $\sqrt{100}\sqrt{p(1-p)} = \sqrt{100}\sqrt{.5 * .5} = 5$.
    But the expected *percentage* of heads is thus $50/100 = 1/2$, and the standard error as a percentage of the number of flips is $5/100 = .05$.
    We could have gotten the last number directly from $\sigma/\sqrt{n} = \sqrt{.5 * .5}/\sqrt{100} = .5/10 = .05$. Ok?

  - Problem-solving note: Sometimes a question/problem requires you to think about the sum of a bunch of random draws, and sometimes about the mean. Note the difference.

- This last example is of a form that is worth developing in the more general case: *Many* problems take the form of estimating a percentage or proportion.

  - e.g.: Public opinion polling on vote intentions try to estimate the proportion of voters who will vote for Gore, Bush, Nader and Buchanan.

- Schematically, we are trying to estimate the mean and sd of the proportion of 1's in a box with an unknown number of 1's and 0's in it. From the above general results, we have for a sample of size n,

$$E(\bar{x}) = \mu$$

$$var(\bar{x}) = \sigma^2/n = \frac{p(1-p)}{n}$$

$$sd(\bar{x}) = \sqrt{\frac{p(1-p)}{n}}.$$

- e.g.: Suppose that the truth is that 48% of those who will vote will vote for Gore, and 52% for someone else. What is the standard error of the estimate of this population proportion if we have a random sample of 100 voters? 1000? 2500? 10,000?

n	s.e.		
100	$\sqrt{.48(.52)/100}$	=	.05
1000	$\sqrt{.48(.52)/1000}$	=	.016
2500	$\sqrt{.48(.52)/2500}$	=	.01
10,000	$\sqrt{.48(.52)/2500}$	=	.005

- Be sure to read the FPP chapters on the Gallup poll and the unemployment statistics to get a sense of the various obstacles that make the reality of doing this much trickier (though the theory and core ideas behind it are in the above).

- Very important point to understand: Notice that in the above analysis the accuracy of the estimates (as indicated by the standard error of the sample proportion above) depends on the sample size, *but not the size of the population.* Why is this?

- See FPP chapter 20, part 4 for a very nice exposition. Intuition says that if we sample 2500 voters in New Mexico and 2500 in Texas, our results in New Mexico should be much more accurate because we sampled 1 in 500 voters there and only 1 in about 5,000 in Texas. But this is an intuition that is just wrong.

- When the size of the sample is small relative to the size of the population (usually the case in survey research), what matters is the size of the sample. Compare to estimating the amount of salt dissolved in a container of water, by sampling a cubic centimeter worth – obviously it won't matter whether the cubic centimeter is drawn from a cup or a jug, if the solution is well mixed.
- When the size of the sample is large relative to the size of the population, we do need a *correction factor* to get a good estimate of the standard error:

$$sd(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

- The reason is that in sampling without replacement from the population, each draw leaves the composition of the box slightly different, and thus changes the probability of getting a '1' on the next draw. These differences are trivial when there are many tickets in the box and the sample is small. (Note what happens above when N is large relative to n.) Illustrate with Stata results above ...

3 Interlude: the law of large numbers

- Recall that the frequentist "definition" of probability went like this: The probability of an event is the *relative frequency* that the event would occur if the experiment or trial were run many, many times, over and over again. We are now in a position to show that this "definition" at least has a certain internal, mathematical coherence.

Th$^{\underline{\mathbf{m}}}$: (The Law of Large Numbers.) Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with $E(X_i) = \mu$ and $var(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any number $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- Explain in words: As the number of draws from the box gets very large, the probability that average of the draws (the sample mean) is farther than ϵ away from the average of the numbers in the box (the true mean) gets very, very small.
- Illustrate with Stata: toss a coin 1000 times, compute the running percentage of heads, and plot the running percentage against the number of tosses. **set obs 1000**, **gen n = _n**, **gen x = uniform() > .5**, **gen heads = sum(x)**, **gen pctheads = heads/n** **graph pctheads n , s(.) yline(.5) yla(0,.5,1) ti(pct heads in n tosses)** .
- Where is this coming from? Can already kind of see it from the result that the standard deviation of the sample mean $\bar{x}$ is $\sigma/\sqrt{n}$, which approaches zero as n gets large. But this

doesn't say anything directly yet about the probability of getting a result far from the sample mean, since we haven't said anything yet about the distribution of the sample mean.

- In fact we don't even need to: We can prove the theorem using Chebyshev's inequality:
  1. Chebyshev's inequality said that for any X, at least $1 - 1/k^2$ percent of the observations lie within k s.d.'s of the mean, $\bar{X}$.
  2. This implies that that for any X, *at most* $1/k^2$ percent of the observations lie *outside* of k s.d.'s from the mean, $\bar{X}$.
  3. Restated in random variable terms, this means that
  $$P(|X - \mu| > k\sigma) \leq 1/k^2.$$
  e.g., for *any* random variable, the probability of drawing a value at least two s.d.s from the mean is at most 1/4. (3 s.d.s, at most 1/9, etc.)
  4. Look back at the Law of Large Numbers: We need an ϵ in the $P(..)$ part. So let's try letting $\epsilon = k\sigma$. This implies that $k = \epsilon/\sigma$. So we can write the inequality in (3) as
  $$P(|X - \mu| > \epsilon) \leq 1/(\epsilon/\sigma)^2 = \sigma^2/\epsilon^2.$$
  5. SO, if we are considering the sample mean $\bar{X}_n$, we have
  $$P(|\bar{X}_n - \mu| > \epsilon) \leq \sigma^2_{\bar{X}_n}/\epsilon^2 = \frac{\sigma^2}{n\epsilon^2}.$$
  6. And what happens to this as n gets very large? Goes to zero. Review that this proves the claim.
- It is NOT crucial that you understand the logic or math of this proof. It is important that you understand the idea or intuition of the law of large numbers, which FPP calls "the law of averages."
  - One way to put is this: As one's sample gets large, the probability that the sample mean is very different from the mean of the population gets very small.
  - Another way to put is this: If an event occurs with probability p, then the proportion of times the event occurs in n trials will approach p with a very high degree of confidence as n gets very large.
- One further concept illustrated in the law of large numbers: Consider some statistic that can be computed for a sample of size n, not necessarily the sample mean. Call it Z_n. Let α be the parameter value that Z_n is trying to estimate (like the population mean re the sample mean). Then methods types say that

  Def$^{\underline{n}}$: *Z_n converges in probability* to α if it is true that $P(|Z_n - \alpha| < \epsilon)$ approaches zero as n gets very large, for any $\epsilon > 0$.

- e.g.: Suppose you want to know what the percentage of nonwhites is in the U.S. congressional district that has the largest such percentage. (Assume this data isn't readily available). Suppose you draw a sample of Congressional districts and use the maximum value for the sample as your estimator ($Z_n = \max\{x_1, x_2, \ldots, x_n\}$).
  - Is this estimator biased or unbiased? (i.e., will it be systematically different from the true value)
  - Does it converge in probability to the true value as the sample size grows?
- If and when you get to more advanced statistical procedures such as regression analysis with a dichotomous dependent variable (e.g., war or not, democracy or nondemocracy, etc.), you will encounter many estimators that are not unbiased like the sample mean, but which do have the weaker nice property of converging in probability to the "right" value.

4 Return to problem of estimating a population statistic from a sample statistic

- We saw in the simulations that the distribution of the sample mean looked approximately normal. This is not an accident.
- In the first part of the 18th century, de Moivre showed that the distribution of the number of heads in n tosses of a coin that lands heads with probability p is approximately Normal as n gets large, with mean equal to np and variance $np(1-p)$.
- Subsequently it was realized that this was a reflection of a much more general truth, which is depicted in a variety of versions of *central limit theorems*. Here is a loose version of a very general one:

  Th$^{\underline{\mathbf{m}}}$: Suppose that random variables $X_1, X_2, \ldots, X_n$ are independent and that some technical conditions (on the third central moment) hold. Then the sum of these random variables has a distribution that is approximately normal with the approximation getting very good as n gets large.
- Here is a more particular version, concerning the case of the sum of a sequence of random variables that all have the same probability distribution.

  Th$^{\underline{\mathbf{m}}}$: Let X_i be random variable with mean μ and variance σ^2. Let $Z_n = X_1 + X_2 + \ldots + X_n$ be the random variable formed by adding together a sequence of n independent draws of X_i. Then as n grows large,

  $$Z_n \overset{a}{\sim} N(n\mu, n\sigma^2)$$

  ($\overset{a}{\sim}$ reads "has a distribution that is approximately ..." .)

- Notice that this theorem says nothing about how the random variables that are the components of the sum are distributed themselves – this doesn't matter, which is kind of amazing.
  - It is true, though, that how rapidly the convergence to Normality takes place depends on the shape of the distributions of the components of the sum. The more normal looking they are, the more rapid the convergence. How rapid is it?
  - Illustrate with sum of 12 uniform distributions in Stata: **set obs 1000**, **for num 1/12: gen xX = uniform()**, **gen z = 0**, **for num 1/12: gen z = z + xX**, **sum**, **graph z ,bin(25) normal**.
  - 12 isn't very many, a uniform distribution isn't very normal, but look at that.
  - By contrast, do same but with **for num 1/12: gen xX = uniform() > .9**.
  - The general point is this: How rapid the convergence is depends on how Normal looking are the distribution of the variables that are components of the sum. The uniform distribution has fatter tails than the normal distribution, but is symmetric. The skewed Bernoulli distributions are both asymmetric and not distributed with much mass at the center.

- How does the central limit theorem bear on the distribution of the sample mean? Recall that the sample mean is based on *a sum of random variables* (i.e., draws from the population "box"). This is why the sample mean has a distribution that is *approximately* normal.

- This implies our main result:

  Th$^{\underline{\mathbf{m}}}$: If the random variables $(x_1, x_2, x_3, \ldots, x_n)$ form a random sample of size n from a given distribution with mean μ and variance $\sigma^2 < \infty$, then as the size of the sample grows large, the sample mean

  $$\bar{x} \overset{a}{\sim} N(\mu, \sigma^2/n).$$

- In other words, the sample mean is an unbiased estimator for the population mean, and has an approximately normal distribution with variance σ^2/n.

- Recall (from properties of Normal distributions) that the sum Z of two normally distributed random variables X and Y has a normal distribution also. This implies further that:

  Th$^{\underline{\mathbf{m}}}$: If a random variable X has a normal distribution, then the distribution of the sample mean $\bar{x}$ for *any* sized sample from X is *exactly* normal.

  Can you see why?

- The central limit theorem helps explain why so many quantities in nature and even in social science have an approximately normal distribution: In brief, *any variable whose*

value is produced by the sum of a fairly large number of other factors or influences should have an approximately normal distribution.

  - e.g.: length of tree leaves, heights and weights of people (see FPP), measurement errors (for repeated scientific measurements) of all sorts, golfers' 18 scores, total weekend receipts at a restaurant or other store, total targets hit by US bombers in Afghanistan by week, etc...

- In short, when you see a variable in social science that has a roughly normal distribution, it is good and potentially fruitful idea to ask if you can imagine it being produced as the sum of number of other factors or components.