Political Science 100a/200a

Fall 2001

Confidence intervals and hypothesis testing, Part I[1]

1 Estimating the uncertainty attached to a sample mean: s^2 vs. σ^2

- Recall the problem of descriptive inference: We want to use the data we collect to say something about the value of some parameter, like the percent who favor Bush prior to an election, or the number killed in civil wars since 1945.
- But either
  1. We don't observe the whole population, as in the case of survey research where we only sample a small fraction of registered voters, etc. ...
  2. Or we observe "the population," but what we are really interested in is the underlying social or economic process that generates these values, such as number or magnitude of civil wars, and we believe that there are numerous random factors that going into making such quantities. (In the simplest case of this sort, we want to say something about a repeatable process like a coin flip or a missile system's accuracy, based on a finite number of experiments or observations.)
- From the central limit theorem and some of our other results, we have drawn some powerful conclusions about *the sample mean* as an estimator for an underlying population parameter μ (e.g. the proportion who favor Bush, or the true number killed in civil wars since 1945, or the likelihood that a BMD system works on a given trial). In particular,

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

  or equivalently,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0, 1).$$

- You should understand what this says: The sample mean $\bar{x}$ has an approximately normal distribution with mean μ (it is unbiased) and variance σ^2/n.
- But now the question is, How can we apply this result in practice when we have a bunch of sample data?

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

- e.g.: Suppose we sample 25 countries and estimate life expectancy in each one. We know that $\bar{x}$, the sample mean, is an unbiased estimator for life expectancy by country around the world. But how do we estimate the uncertainty attached to this estimate? Where do we get an estimate for the standard deviation of the sample mean, $\sigma/\sqrt{n}$?
- We encounter a problem: Our theoretical result requires us to have σ^2, the variance of the *population* variable. This is of course a problem because we don't observe the population – that is why we are drawing and examining a sample in the first place!
- We have, however, a *natural candidate* to use as an estimate for σ^2. Why not just use the *variance of the sample values*?
- Consider, for example, the variance of the sample, call this σ_s^2.

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

  where x_i is the ith value in the sample and n is the size of the sample.
- Our next question should be: Is this a good estimate of σ^2, the true population variance?
- What makes something a good estimator? One criterion, as we have discussed, is *unbiasedness*. If we took many random samples of 25 countries, computed σ_s^2 for each one, and then averaged them, would the result be centered on the true population value σ^2?
- In other words, is it true that $E(\sigma_s^2) = \sigma^2$?
- The answer turns out to be No, not quite. Below, I show that

$$E(\sigma_s^2) = \tfrac{n-1}{n}\sigma^2$$

- What does this imply? σ_s^2 is a slightly biased estimator of the population variance σ^2; it is a little bit too small on average, although as the sample size n gets larger it is almost unbiased. (You could show that it "converges in probability" to the right value.)
- To get an unbiased estimator, all we have to do is multiply by $n/(n-1)$ so

$$\begin{aligned}\tfrac{n}{n-1}\sigma_s^2 &= \tfrac{n}{n-1}\tfrac{1}{n}\textstyle\sum(x_i - \bar{x})^2 \\ &= \tfrac{1}{n-1}\textstyle\sum(x_i - \bar{x})^2.\end{aligned}$$

- This quantity is what we call *the sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- This is the quantity that Stata and (I think) most other statistical packages computes when it shows you variance or standard deviation. Note that we can compute s^2 using only information available in our sample. (Recall also that s^2 is an estimate of the population variance σ^2, NOT an estimate of the variance of the sample mean $\bar{x}$. Important to keep this straight.)
- s^2 is a good estimate for the population variance σ^2 in the sense that on average (across hypothetical random samples) it will be right.
- So, this gives us a new possibility for estimating the uncertainty attached to our estimate $\bar{x}$. Why don't we just use

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

instead of

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

2 Application: Confidence intervals

- Suppose we do this. How do we "summarize the uncertainty" associated with our estimate? One valuable approach is to construct a *confidence interval.* (See FPP, ch21, sections 2-3).
- Intuitively, you have already been exposed to confidence intervals many times, whenever you hear a poll result reported as (for instance), "34% of Americans don't like Al Gore's beard, according to a poll with a plus or minus 3% margin of error."
- What does this really mean?
- Draw picture of normal distribution around μ with standard deviation $\sigma/\sqrt{n}$. ... Indicate a particular sample mean $\bar{x}$. Show that in about 95% of all such random samples, $\bar{x}$ will lie within 2 s.d.s on either side of the population value, μ.
- Thus, if you could draw random samples of this size many, many times, in about 95% of them, an interval of two s.d.s around the sample mean would "cover" the true value μ. (discuss weird locution ...)
- More formally,

  Def$^{\underline{n}}$: A $\beta\%$ *confidence interval* is the interval on either side of the sample mean $\bar{x}$ (symmetric around $\bar{x}$) that would take in $\beta\%$ of the area under the probability distribution for the sample mean.

  - e.g.: A 95% confidence interval around the sample mean extends $1.96\sigma/\sqrt{n}$ to either side of $\bar{x}$ Show with diagram ... Why 1.96?

– e.g.: In the above case, our estimate for a 95% confidence interval extends

$$1.96\frac{s}{\sqrt{25}}$$

on either side of $\bar{x}$ (not using the finite sample correction)

– What good is this?

– A confidence interval can be interpreted as follows: We don't know what the true population mean μ is, but if we draw 25 country samples repeatedly, many, many times, 95% of the time the 95% confidence interval would "cover" the true mean, μ.

– The confidence interval is NOT to be interpreted as follows (at least not by a frequentist): It is not true that the probability that the true mean falls within the 95% confidence interval is 95%. It either does or it doesn't.

– A confidence interval is often a fairly intuitive and helpful way of summarizing uncertainty about a parameter being estimated than just giving the standard error of the sample mean, because it gives the reader a sense of the plausible range of the error of the estimate. e.g., public opinion polls are reported with a "margin of error" which is the size of one side of a confidence interval (typically 95%, I think).

– Increasingly, in political science research, you find people reporting confidence intervals rather than "significance levels" (about which more later), because confidence intervals tell you something more substantive about the parameter of interest (i.e., the range in which the parameter likely falls).

• Example 1: A confidence interval for an estimate of life expectancy across countries ...

– Use Stata to sample 25 countries, compute mean and sd of sample, construct 95% confidence interval. Does it "cover" the true value?

• Example 2: A confidence interval for the number killed in civil wars since 1945

– Show in Stata data on numbers killed by war for 1945-99

– These estimates are *incredibly* noisy, however. You might think this would make estimating the *total* number killed by adding up these noisy estimates an even more noisy, hopeless endeavor.

– Let x_i be an estimate of number killed for the ith war, and let S be the sum of a bunch of estimates. Then we have

$$\begin{aligned} S &= x_1 + x_2 + x_3 + \ldots + x_n, \text{ and thus} \\ var(S) &= var(x_1) + var(x_2) + \ldots + var(x_n) \end{aligned}$$

- Notice that we are treating S as a random variable that is the sum of a bunch of random variables here. That is, we see just one set of estimates, but we imagine that these estimates are the product of a stochastic process that could produced quite different numbers.
- Further, what can we say about the probability distribution of the sum S?
- So we can estimate the variance (and thus the sd) of the sum if we can an estimate for the variance of the estimate for each war.
- Let's just make a guess: Suppose that the sd of each estimate is proportional to the number killed – thus, for bigger wars the uncertainty associated with the number killed is proportionately larger.
- Let's try assuming that the sd is 10% of the estimate we have and see what this gives us. **gen sdest = .1*iissdth, gen varest = sdest^2, sum varest**.
- Thus our estimate for the sd of the sum S is what?
- Ok, now construct a confidence interval ...

3 Another problem with our approach: s^2 is a random variable

- Let's return to the problem of estimating a confidence interval for a population mean based on a sample (e.g., life expectancy across countries).
- You may be wondering, Isn't there something circular about using the sample to estimate the variance of the (unobserved) population (the "box," in FPP)? What about the uncertainty attached to s^2 as an estimate of σ^2? Shouldn't this be factored in somehow?
- Yes, it should.
- Look again at what we did: we substituted the sample variance s^2 for the true population variance σ^2 in our theoretical result, using

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

instead of

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

- But this was kind of devious. Note that because it is based on a sample, and would be a little different for each different sample we could draw, the sample variance s^2 is *a random variable.*
- If so, then what gives us the license to stick it into $N(\mu, \cdot)$ and still believe that what we have is distributed normally?

- Given that s^2 is a random variable, wouldn't you expect that this would *add* to our uncertainty about the sample mean, $\bar{x}$?
- In fact it does, and to be really correct we need to account for this additional uncertainty introduced by the fact that we only have an estimate of σ^2, not the true population value.
- It turns out that in a large sample (big n), this additional uncertainty doesn't really matter. The central limit theorem will ensure that despite the added uncertainty introduced by s^2, the distribution of $\bar{x}$ will become approximately normal with variance s^2/n.
- But what about in a small sample? Here the fact that we only have an estimate of σ^2 that is likely to have error attached to it becomes important. It is possible (though difficult) to establish the following result:

  Th$^{\underline{\mathbf{m}}}$: If a random variable X has Normal distribution, and we draw a sample from X with n observations, then the sample mean $\bar{x}$ has a *t distribution with $n-1$ degrees of freedom.*
- A t distribution looks very much like a normal distribution except that it has fatter tails. More weight is put on values relatively far from the mean. So using the t distribution is bit more conservative about how precise an estimate of the sample mean you are getting.
- However, as sample size n gets larger, a t distribution with $n-1$ degrees of freedom converges quickly to a normal distribution, so with a large sample (as low as 25 if the underlying variable is not highly asymmetric) using the t distribution is essentially equivalent to using a normal distribution. (Show with Stata ...)
- In reading political science articles using regression and related methods, you will constantly encounter authors talking about "t statistics." This is what they are referring to. When you are testing hypothesis in a regression model – e.g., that after controlling for per capita income, subSaharan Africa states have significantly lower life expectancies on average – the *test statistic* spit out by a standard regression is a t-statistic. Illustrate with Stata ...

Question: Construct a 95% confidence interval for our estimate of the mean life expectancy around the globe by country, using the more conservative (and appropriate, for a 25 country sample) t distribution.

- Stata has the built in function **invt(df, p)**, where df is the number of "degrees of freedom," which is the sample size minus one, and p is the probability you want. I.e., typing **invt(24, .95)** will give the distance from zero you have to go on either side to

get 95% of the area under a t distribution with 24 degrees of freedom. Draw ... this gives 2.06. So you need to go 2.06 standard units on either side to get a 95% confidence interval.

- For our problem, a standard unit is a standard deviation of the sample mean, $s/\sqrt{n} = ?/5 = ?$. So a 95% confidence interval is: ....

4 Hypothesis testing: the core logic

- These results concerning the probability distribution of the sample mean allow us to test hypotheses about unobserved parameters of social science interest, such as the *population* mean of the proportion of likely Gore voters, average life expectancy by country for the whole world, or the "true" probability that two democracies will fight each other.
- You have already seen examples of the logic at work. The most basic example is the coin toss experiment.

Question: You have a coin you suspect may be biased in favor of heads. How can you decide?

- You try the experiment of tossing it 10 times, and you find that it comes up heads 8 times.
- You formulate your *null hypothesis* that the coin is fair, and ask: What is the probability that I would see 8 or more heads in ten tosses if this coin were in fact fair? Formally, let

$$\begin{aligned} H_0 &= \text{coin is fair,} \\ H_1 &= \text{coin is biased in favor of heads} \end{aligned}$$

- We will ask what is $P(8 \text{ or more heads} | H_0)$. If very small, we will "reject the null."
- This is a question we can answer without the central limit theorem, just by using probability theory.
- The probability of 8 heads in 10 tosses of a fair coin is $B(8; 10, .5) = \binom{10}{8}/2^{10} = .044$. Likewise, we can calculate $B(9; 10, .5)$ and $B(10; 10, .5)$, add them together to get the probability of getting 8 or more heads in 10 tosses if the coin were in fact fair: approximately .055.
- .055 is an example of a *p-value*. You might see it reported as $p = .055$.

Def^{n}: (a bit loose; cf. FPP) The *p-value* of a hypothesis test is the probability that you would see this data or worse (for the null hypothesis) if the null hypothesis were true.

- So it is fairly unlikely that this data would be produced by a fair coin. But still, in a bit more than one in twenty such experiments we would see this many heads or worse.
- What next? Social scientists (and scientists much more generally) have developed conventions here, such as: We reject the null hypothesis in favor of the alternative hypothesis if the p value is less than .05, or .01, or .10, etc.
- In fact, the standard convention in political science is that you might report a relationship that has a p value of .10 or less, and below that how small p is is taken as a measure of how decisively the data reject the null.
- Notice that it is entirely possible that you could *wrongly reject the null hypothesis*. e.g., if you saw 9 heads in 10 tosses, you would conclude that the p value was .01, so you would surely "reject the null" in favor of your alternative the coin is biased in favor of heads.
- But one in 100 (hypothetical) times this will be a mistake. You will be committing what is called a *Type I* error – wrongly rejecting the null hypothesis when the null is correct.
- What's the alternative? If we tighten the standard for rejecting the null – e.g., we will "accept the null" for p values of less than, say, .001 – then we will be increasing the odds of making a *Type II error*, which means wrongly accepting the null when the null hypothesis is false. There is a tradeoff between Type I and Type II errors. The convention resolves this conservatively, making us reluctant to conclude againt the (random) null hypothesis unless the data are pretty highly unlikely to observed if the null is true.

5 Applying the logic: Using the normal approximation with Bernoulli trials

Question: What if you collected more data? Suppose you flip the coin 100 times and get 62 heads.

- Again, our test is to ask what is the probability we observe this many heads or more if the null hypothesis were true.

- We could ask Stata to figure this out precisely. **set obs 101**, **range x 0 100**, **gen bin = comb(100,x)/2^{100}**, **graph bin x ,s(.)**, **egen p = sum(x) if x > 61**. Or much more simply in Stata 7: **di Binomial(100,62,.5)**.
- But there is another way that is more useful in general because *we can apply it to problems where we don't have such a well-defined underlying stochastic process that generates the data* (e.g., life expectancy across countries): Use the normal approximation.
- By the central limit theorem, the sum of the number of heads in 100 tosses of a fair coin has an approximately normal distribution. What is $E(X)$ and $var(X)$, if X is the number of heads that come up? show that $sd(X) = \sqrt{n}\sigma$, which FPP call "square root law," so here $sd(X) = \sqrt{100}\sqrt{p(1-p)} = \sqrt{100}\sqrt{.5 * .5} = 5$.
- So the number of heads in 100 flips has an approx. normal distribution with mean 50 and s.d. of 5. What is the probability of observing 62 or more heads? Ask how many standard units 62 is away from 50 and use the normal table: Approximately
$$z = \frac{62 - 50}{5} = 2.4,$$
using **di 1 - normprob(2.4)** in Stata gives .008, which is pretty close to .0104 we calculated directly from the binomial distribution. (In Stata 7, the cumulative normal function has been changed to **norm(z)**.)
- We can do slightly better by asking what is the probability of observing 61.5 or more heads (draw diagram). This is what FPP discuss as a continuity correction. Then $z = (61.5 - 50)/5 = 2.3$, so $p = .0107$ which is very close to the true value indeed.
- There is a useful and important thing to note about the expression for z above. z is a *test statistic*, and has the general form
$$z = \frac{\text{observed value } - \text{value expected}|H_0}{\text{est. standard error}}$$
- Explain. Be sure to read and understand FPP, chapter 26 on this.

Question: Test the null hypothesis that global life expectancy by country is 65 years against the alternative hypothesis that it is less than 60 years.

- Compute
$$z = \frac{\bar{x} - 65}{s/\sqrt{n}}$$
which is the number of standard units $\bar{x}$ is away from 65, and then calculate $norm(z)$. (Discuss one tailed test ...)
- Note that Stata has a command that will do t tests automatically: **ttest varname = #**, for example. Illustrate ... (does not use finite sample correction)

6 Proof that s^2 is an unbiased estimator for σ^2

- Why is σ_s^2 a biased estimate of σ^2? First an intuition, then a proof.
  - First, notice that one of the components that goes into the estimate σ_s^2 is the sample mean itself, since

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j)^2$$

    which can be rewritten inside the sum as

$$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i(1 - \frac{1}{n}) - \frac{1}{n}\sum_{j\neq i}x_j)^2$$

  - Explain rewrite. Notice that this is almost as if we had a new random variable $x_i(1-1/n)$. This variable has to have lower variance than σ^2 because of the $1-1/n$ term which is effectively reducing everything towards the mean.
  - Intuitively, what is happening is that by using the sample mean in constructing σ_s^2, we introduce an influence that *lowers* σ_s^2 relative to we are trying to estimate, σ^2.
  - Imagine that in our 25 country sample, we happen to get a country with very low life expectancy. Note that this pulls our sample mean down, towards the very low life expectancy number. In effect, this is reducing the size of our estimate σ_s^2 relative to what it would be if we were using μ instead of $\bar{x}$ when calculating σ_s^2.
  - Now, a proof:
    1. First, note that

$$\begin{aligned} E(\sigma_s^2) &= E(\tfrac{1}{n}\textstyle\sum(x_i - \bar{x})^2) \\ &= E(\tfrac{1}{n}(\textstyle\sum x_i^2) - \bar{x}^2) \\ &= \left(\tfrac{1}{n}E(\textstyle\sum x_i^2)\right) - E(\bar{x}^2) \\ &= E(x_i^2) - E(\bar{x}^2) \end{aligned}$$

       using a fact about rewriting the expression for variance you've seen a couple of times, and then properties of expectations.
    2. Next, note that

$$\begin{aligned} E((x_i - \mu)^2) &= \sigma^2 \text{ by definition, so} \\ E(x_i^2 - 2x_i\mu + \mu^2) &= \sigma^2 \\ E(x_i^2) - \mu^2 &= \sigma^2, \text{ so} \\ E(x_i^2) &= \sigma^2 + \mu^2. \end{aligned}$$

3. In just the same way we can use the result that

$$E((\bar{x}-\mu)^2) \;=\; \tfrac{\sigma^2}{n}$$

to show that

$$E(\bar{x}^2) \;=\; \mu^2 + \tfrac{\sigma^2}{n}$$

4. Now we have expressions for $E(x_i^2)$ and $E(\bar{x}^2)$, so we can return to the result in step 2 above, getting

$$\begin{aligned} E(\sigma_s^2) &= E(x_i^2) - E(\bar{x}^2) \\ &= \sigma^2 + \mu^2 - \mu^2 - \tfrac{\sigma^2}{n} \\ &= \sigma^2\left(1 - \tfrac{1}{n}\right) \\ &= \sigma^2\left(\tfrac{n-1}{n}\right) \end{aligned}$$

5. This shows that σ_s^2 is a biased estimate of σ^2. It is a little too small on average.
6. An unbiased estimate is:

$$\begin{aligned} s^2 \;\equiv\; \tfrac{n}{n-1}\sigma_s^2 &= \tfrac{n}{n-1}\tfrac{1}{n}\textstyle\sum(x_i-\bar{x})^2 \\ s^2 \;\equiv\; & \qquad\quad \tfrac{1}{n-1}\textstyle\sum(x_i-\bar{x})^2. \end{aligned}$$

- That takes care of the question about why σ_s^2 is biased.