Center, Spread, and Shape in Inference: Claims, Caveats, and Insights

Dr. Nancy Pfenning (University of Pittsburgh) AMATYC November 2008

Preliminary Activities

1. I would like to produce an interval estimate for the mean years of education (starting with first grade) of Math/Stats teachers at two-year colleges. I record years of education for each teacher, calculate the sample mean and standard deviation, and use those to set up a 95% confidence interval for the population mean: $\bar{x} \pm 2\frac{s}{\sqrt{n}}$. (The sample is large enough that the t multiplier for 95% confidence is approximately 2.) What do I claim about this interval?

   I claim to be 95% confident that the mean years of education of all Math/Stats teachers at two-year colleges is in the interval.

2. A set of 6 numbered cards represent how many phone numbers each person in a family of 6 knows by heart. I would like to produce an interval estimate for the mean number known by all 6, based on information from 4 of the 6 family members who happen to be available. Assume the population standard deviation is 3. We pick 4 cards at random and set up the interval $\bar{x} \pm 2\frac{3}{\sqrt{4}} = \bar{x} \pm 3$. What do I claim about this interval?

   I claim to be 95% confident that the unknown population mean is in that interval.

3. A set of 16 numbered cards represent the number of fillings for a group of 16 students. Suppose a dentist visiting the school has time to examine 2 students. Based on their mean number of fillings, he is to test the hypothesis that the mean for the whole group is 4 against the alternative that it is less than 4. The test is to be carried out at the 0.05 level. We'll assume the standard deviation for all 16 students is 3.

   Two cards are picked at random from those 16, and we calculate their mean $\bar{x}$. The test statistic is $\frac{\bar{x}-4}{3/\sqrt{2}}$. Once we find the test statistic is not less than -1.645, we do not reject the null hypothesis at the 0.05 level. In fact, the mean of all 16 cards is 4, so the null hypothesis is true. In the long run, how often will it be rejected?

   I claim it will be rejected 5% of the time.

Introduction

Our key inference results in introductory statistics—confidence intervals and hypothesis tests about means or proportions—are constructed from what we claim to know about the center, spread, and shape of the sampling distributions. In fact, such claims must be made cautiously, because if certain criteria concerning the background of our data are not met, the center, spread, and shape of the sampling distributions are affected. We can be specific about what aspects of our interval or test are undermined, depending on how the three key

conditions are violated. Furthermore, we can carry out activities with students to give them a concrete feel for how improper sampling design can affect three specific aspects of our inference claims. Because of limited time, we'll focus only on quantitative variables.

Claims

The key results obtained are summarized in the following claims about the **sampling distribution of sample mean**: If the population for a quantitative variable has mean μ and standard deviation σ, then (under the right circumstances) the distribution of sample mean for a sample of size n has

- **center:** mean μ
- **spread:** standard deviation $\frac{\sigma}{\sqrt{n}}$
- **shape:** approximately normal

These pave the way for us to make the following **inference** claims:

If sampled values of a quantitative variable for a sample of size n have mean $\bar{x}$, and the variable's standard deviation is known to be σ, then a 95% confidence interval for μ is $\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$. To test the hypothesis $H_0 : \mu = \mu_0$, we standardize $\bar{x}$ to $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ and decide whether to reject H_0 based on the probability of z this low, high, or different from 0, depending on the form of the alternative hypothesis.

Caveats

Students—and, admittedly, we teachers—are inclined to take this information and run with it, although we do occasionally mention the following caveats:

- The sample must be unbiased.
- Selections must be independent; the population should be at least 10 times the sample size.
- The sample size must be large enough.

However, to avoid the discomfort of assigning students problems that reach a dead-end, we tend to give them carte blanche: assume the sample is random, and large enough to guarantee normality through the Central Limit Theorem, and that our selections are independent. Often, because these assurances start to become repetitive, we let students assume that, by default, all the necessary conditions are in place.

Thus, in the interest of guaranteeing them situations where they can plug into a handy formula with impunity, we deprive them of two valuable lessons:

1. There are direct consequences that link each caveat with one of the claims about center, spread, and shape, respectively. Thus, there are meaningful connections between what is learned about producing and summarizing data, and about sampling distributions, and what is applied to perform inference. These connections are threads that tie together all the most important material in an introductory statistics course.

2. In many real-life situations, the conditions needed to justify the claims are not met.

Unjustified claims about center

In general, bias can arise in two ways, each of which can undermine our claim that the mean of the distribution of sample means equals the population mean.

1. Bias arising from the **sampling design**: the sample does not truly represent the larger population of interest. Do our teachers truly represent the entire population of math/stat teachers at two-year colleges?

2. Bias arising from the **study design**: the sampled values are assessed improperly, resulting in biased summaries. Note that this type of bias can arise even if the sampled individuals are truly representative of the larger population. Is asking teachers to report their years of education out loud a good way to obtain completely accurate data values?

Are we really 95% confident that the mean years of education for the entire population of math/stat teachers at two-year colleges falls in the interval we produced earlier?

Unjustified claims about spread

When we present students with the concept of random variables, we often mention rules for means and variances: the mean of the sum of two random variables equals the sum of the means, and the variance of the sum equals the sum of the variances. But the latter *only holds if the two variables are independent*. If there is dependence, then $\sigma^2_{X_1+X_2}$ can be substantially greater or less than $\sigma^2_{X_1} + \sigma^2_{X_2}$.

The additivity of variances lets us assert that the variance of $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is

$$\frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

and that the standard deviation of $\bar{X}$ is the square root of this, $\frac{\sigma}{\sqrt{n}}$.

However, these claims break down if the X_i are *dependent*, a circumstance that can arise in real-life sampling situations, including when we sample without replacement from a population that is *not* sufficiently large relative to the sample size.

Thus, if we fail to satisfy the requirement that the population be at least 10 times the sample size, the claimed standard deviation $\frac{\sigma}{\sqrt{n}}$ in our confidence interval formula is incorrect.

Are we really 95% confident that the interval $\bar{x} \pm 2\frac{3}{\sqrt{4}} = \bar{x} \pm 3$ based on a sample of 4 cards contains the mean for all 6 cards?

Unjustified claims about shape

Thanks to the Central Limit Theorem, we can say that the distribution of sample mean for random samples of size n is approximately normal if the underlying distribution itself is approximately normal, or if the sample size is large enough to offset non-normality in the underlying distribution.

If the sample is too small and the shape of the sampling distribution is not normal, then the multiplier 2 in our confidence interval is incorrect, because it is based on a normal probability distribution. Likewise, our P-value in a hypothesis test would be incorrect.

Earlier, we picked 2 cards at random from 16 cards whose mean was 4 and standard deviation was 3, and tested the true null hypothesis $H_0 : \mu = 4$ against the alternative $H_a : \mu < 4$ at the 0.05 level. Can we really claim that H_0 will be rejected 5% of the time in the long run?

Insights through activities

1. A biased sample mean: Our activity today had to do with teachers' years of education. You can survey students about another variable that may produce biased responses: ask them to report out loud how many classes they missed the week before. Note that bias can occur here on two fronts. First, the students who are missing class that day are not represented in the survey! Second, students may not feel comfortable telling their instructor how many classes they missed. In both cases, the bias leads to our sample mean providing an underestimate of the true population mean number of classes missed.

   Alternatively, you can ask students to pick a number "at random" (off the top of their heads) from 1 to 20. We use the sample mean number picked, $\bar{x}$, and the standard deviation of the numbers 1 to 20 (5.77) to construct a 95% confidence interval for the "unknown" population mean, 10.5. Presumably we are 95% confident that the population mean falls in this interval. However, such selections—which are actually haphazard, not random—tend to be biased towards higher numbers. Our sample mean is a biased estimator for μ, and our confidence interval is wrong.

2. An incorrect standard deviation: We sample 4 cards without replacement from a population with mean 5.5, standard deviation approximately 3. In the first case, we will sample from a population of 40 cards (10 times the sample size. In the second case, we sample 4 from 6 (the population is only 50% larger than the sample). In the third case, the population is the same size as the sample. To give our activity context, we can say the values represent how many phone numbers people have memorized.

   (a) Sample 4 cards from 40: an ordinary deck minus all the face cards, so it consists of 4 each of the numbers 1 through 10.

   (b) Sample 4 cards from 6: the numbers 1, 3, 5, 6, 8, 10.

   (c) Sample 4 cards from 4: the numbers 2, 3, 8, 9.

In each case, we sample 4 cards without replacement at least 20 times from the given population, recording the sample mean value each time. Assuming selections are independent, the distribution of sample mean selection has a mean of 5.5 and a standard deviation of $\frac{3}{\sqrt{4}} = 1.5$. We calculate the mean and standard deviation of each group of sample means. In all three cases, the mean of sample means will in fact be approximately 5.5. In the first case, the standard deviation should conform well to the claimed value, 1.5. In the second case, the standard deviation will be less. In the third case, it is of course zero!

If we sample 4 individuals without replacement from a population of 6 individuals whose standard deviation is 3, and use their sample mean to set up a confidence interval for the population mean, the interval will be wider than it should be (margin of error $2\frac{3}{\sqrt{4}} = 3$ instead of about $2(.8) = 1.6$). We would claim that we are only 95% confident that the population mean falls in that interval, but because these sample means don't stray as far from 5.5, our probability of capturing 5.5 in our confidence interval will be considerably higher than 95%. In the extreme, if we sample 4 from 4 cards, our "95%" confidence interval actually has a 100% success rate, because our sample mean is always 5.5 and there is no margin of error at all.

As far as a hypothesis test is concerned, we would calculate z to be $\frac{\bar{x}-5.5}{3/\sqrt{4}}$, whereas the actual denominator for 4 cards sampled without replacement from 6 is less than $3/\sqrt{4}$. Our calculated z is less extreme than it should be, making us vulnerable to Type II Errors, failing to reject a false null hypothesis.

3. A non-normal sampling distribution: A right-skewed distribution of numbers is created from a deck of cards: use all 4 of each of the numbers 1 and 2, and only 1 each of the numbers 3 through 10. (This distribution models the number of fillings in a group of people where most have just one or two, but a few have more.) The population mean is 4 and the standard deviation is 3. Each student picks 2 cards at random from these 16 cards and reports his or her sample mean. Each uses the sample mean to test $H_0 : \mu = 4$ vs. $H_a : \mu < 4$ at the 0.05 level. The long-run probability of rejecting, theoretically, is 0.05. However, the lowest possible sample mean is 1, which standardizes to $\frac{1-4}{3/\sqrt{2}} = -1.41$. A sample mean low enough to produce a P-value of 0.05 or less must fall below (to the left of) -1.645, but this is impossible. We can take hundreds such samples, and never manage to reject H_0 at the 5% level.

   In general, our P-value based on z probabilities is incorrect if we are standardizing a sample mean or sample proportion for a small sample taken from a skewed population.

Summarizing claims and caveats in confidence intervals

A 95% confidence interval for μ is constructed as

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}$$

- If there is bias arising from a poor sampling or study design, then the correct interval is not centered at $\bar{x}$.
- If there is dependence in our sampled individuals (such as when we sample without replacement from a relatively small population), then the standard deviation $\frac{\sigma}{\sqrt{n}}$ is incorrect.
- If the sample size is not large enough to offset non-normality in the shape of the population, then the multiplier 2 is incorrect.

[Note 1: the same problems arise if we standardize with s and base our confidence interval on a t distribution. Note 2: Another source of sampling dependence would be if the individuals are aware of each other's responses, and tend to adjust their own responses accordingly. Yet another is the use of a two-sample procedure for paired data. In particular, the claimed formula for the standard deviation of the differnce between sample means is undermined if the variables X_1 and X_2 are dependent via a paired design.]