Data Analysis and Statistical Methods Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 13 (MWF) Designing the experiment: Margin of Error

Suhasini Subba Rao

Terminology: The population and sample mean

- The population mean μ is the mean of the entire population. The population can be (and often is) infinite.

- Suppose that $X_1, \ldots, X_n$ are numbers drawn from the population. The sample mean is the average of $X_1, \ldots, X_n$.

Terminology: Standard deviations and errors

- The standard deviation is a measure of variation/spread of a variable (in the population). This is typically denoted as σ. See Lecture 4.

- The standard error is a measure of variation/spread of the sample mean. The standard error of the sample mean is

$$\frac{\sigma}{\sqrt{n}}.$$

  See Lecture 12.

- Usually, σ is unknown. To get some idea of the spread, we estimate it from the sample $\{X_i\}_{i=1}^n$ using the formula

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

We call s the sample standard deviation. It is an estimator of the standard deviation σ. Usually $s \neq \sigma$. Often $s < \sigma$ (especially, when the sample size is not large).

- Since σ is usually unknown, the standard error $\sigma/\sqrt{n}$ is usually unknown. Instead we estimate it using the sample standard error is

$$\frac{s}{\sqrt{n}}.$$

Margin of Error

- According to a recent survey, Americans walk on average 18 miles a week with a margin of error of 2.5 miles.

- What does this mean in terms of confidence intervals?

- 18 miles corresponds to the (sample mean) average number of miles walked in the sample, the margin of error is the plus and minus in the confidence interval.

  In other words for a 95% confidence interval

$$\text{Margin of Error} = 1.96 \times \frac{\sigma}{\sqrt{n}}.$$

- The smaller the margin of error the more precisely we can pin point the population mean. Of course it is worth bearing in mind that we can never be sure that our confidence interval contains the mean, which is why we prescribe a level (such as a 95%) to the interval. In other words, we can never be sure that the population mean is within the prescribed margin of error of the sample mean.

- Of course, it is clear a small margin error gives a better level of precision than a large one.

- In the next few slides, we recall factors which influence the MoE.

Relationships: Sample size and MoE

- We compare the 95% confidence intervals for $n = 9$ and $n = 25$. We see

$$\text{n=9} \qquad \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{9}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{9}}\right]$$

$$\text{n=25} \qquad \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{25}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{25}}\right]$$

What are the lengths of the above intervals?

  - For n=9 the margin of error is $1.96 \times \frac{\sigma}{\sqrt{9}}$.
  - For n=25 the margin of error is $1.96 \times \frac{\sigma}{\sqrt{25}}$.

Important The margin of error does not depend on $\bar{X}$.

Example: $\bar{X} = 10.38$, $\sigma = \sqrt{33}$.

$$\text{n=9} \qquad [10.38 \pm 1.96\frac{\sqrt{33}}{\sqrt{9}}] = [6.63, 14.13] \quad \text{MoE=3.75}$$

$$\text{n=25} \qquad [10.38 \pm 1.96\frac{\sqrt{33}}{\sqrt{25}}] = [8.12, 12.63] \quad \text{MoE} = 2.255$$

- The second interval has a smaller margin of error.

- When the sample size is large the estimator "tends" to be closer to the true parameter. Thus the confidence interval will be narrower; since the margin of error is smaller.

Relationships: the standard deviation and MoE

- The variability in the sample, measured by σ will have an impact of the reliability of an estimator and its margin of error.
- Example: Suppose $\bar{X} = 10.38$, $n = 9$, but the variability in the two populations are different:

$$\sigma = 5.7 \qquad \left[10.38 \pm 1.96\frac{5.7}{\sqrt{9}}\right] = [6.63, 14.13], \quad \mathrm{MoE} = 3.75$$

$$\sigma = 10 \qquad \left[10.38 \pm 1.96\frac{10}{\sqrt{9}}\right] = [3.38, 16.91], \quad \mathrm{MoE} = 6.8$$

- The more variablility within the population (as measure by the standard deviation) the more variability in the sample mean (as measure by the standard error). Usually we cannot do anything about this.

How large an interval to use?

- You read in a newspaper that "the proportion of the public that support same-sex marriage is $55\% \pm 15\%$."

- This means a survey was done, the proportion in the survey who said they supported same-sex marriage was 55% and the confidence interval for the population proportion is $[55 - 15, 55 + 15]\% = [40, 70]\%$.

- This is an extremely large interval, it is so wide, that it is uninformative about the majority opinion of the public.

- The reason it is too wide is that the sample size is too small.

- Conclusion: This experiment was not designed well.

- Typically, before data is calculated, we need to decide how large a sample to collect.

- This is usually done by deciding how much "above and below" the estimator is acceptable. For example, an interval of the type [55-3,55+3]% = [52,58]% tells us that the majority appear to support same-sex marriage.

- The 3% is is the margin of error. Given a margin of error we can then determine the sample size to collect.

Choosing the sample size for estimating μ

- In an ideal world we would have a large sample size. A large sample size gives a small standard error, which in turn yields a small margin of error.

- However, obtaining very large samples can be impossible for many different reasons.

- How can one determine the number of observations to be included in a sample? How to choose the sample size n?

  Answer: Usually we have a margin of error in mind. We can accept the reliability of a estimator up to a certain margin or error. Once we know what margin of error is acceptable we can then choose the sample size.

Formula for choosing the sample size

- To choose the sample size according to the margin of error, we need to know (or guess apriori) the standard deviation σ (if we don't know what it is, then we err on the cautious side and use a value that seems reasonable but large).

- We recall that in the confidence interval:

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

  the margin of error is $MoE = 1.96\frac{\sigma}{\sqrt{n}}$.

- Therefore, if we want to choose the sample size such that the margin of error for a given E we need to solve for

$$MoE = 1.96\frac{\sigma}{\sqrt{n}} \quad \text{solving for } n \text{ gives} \quad n = \left(\frac{1.96\sigma}{E}\right)^2.$$

- Example: Suppose we guess that $\sigma = \sqrt{3}$ and we want the margin of error $MoE = 0.25$. The confidence interval is

$$\left[\bar{X} - 1.96 \times \frac{\sqrt{3}}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sqrt{3}}{\sqrt{n}}\right]$$

and we solve $1.96 \times \frac{\sqrt{3}}{\sqrt{n}} = 0.25$. This gives

$$n = (\frac{2 \times 1.96\sqrt{3}}{0.5})^2 = 184.4$$

- Of course, a larger value of n will give a smaller margin of error, so we round up and use $n = 185$. In other words, for this experiment we need to choose a sample size of at least 185 to be sure that the margin error is at most 0.25.

The general formula

- Changing the confidence level simply changes 1.96 to another number.

- If the objective is to use a 99% CI, we first look up 0.5% in the z-tables, $z_{0.5\%} = 2.57$. Then we solve $MoE = 2.57\frac{\sigma}{\sqrt{n}}$

$$n = \left(\frac{2.57 \times \sigma}{MoE}\right)^2.$$

- In general, for the $(100 - \alpha)\%$ CI use the formula $MoE = |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}$ and solve it to give

$$n = \left(\frac{|z_{\alpha/2}| \times \sigma}{MoE}\right)^2.$$

Example 4: Heights

- Researchers want to estimate the mean height of students at a university (in meters) with a margin of error of 0.04 (using a 95% CI level). The sample standard deviation from a small sample taken previously is 0.113. How many students must they sample to achieve their specifications?

- Solution 4:
  - Since the true population standard deviation is unknown, they use the sample standard deviation in the calculation.
  - Use the formula with $MoE = 0.04$ gives the solution
  $$n = \frac{(1.96)^2(0.113)^2}{(0.04)^2} = 30.65.$$
  - They must sample 31 people such that a 95% confidence interval has length $2 \times 0.04 = 0.08$.

Example 5 (more practice): Caffine content

- The caffine content is coffee is being analysed and it is known that standard deviation of a randomly selected coffee is 7.1mg.

  Suppose 100 cups of coffee are analysed, and the total weight of caffine in all the cups is $\sum_{i=1}^{100} X_i = 11000$mg, construct a 95% CI for the mean caffine content.

  Construct an 80% CI for the mean caffine.

  Find the minimum number of coffees which must be analysed for the 80% CI to have MoE 0.45mg?

- Solution 5:

- The total weight of caffine for the 100 cups is 11000 mg. Therefore the sample average of caffine per cup is $\bar{x} = 11000/100 = 110$.

- Calculating the 95% CI (use the formala or calculate yourself): $z_{0.05/2} = z_{0.025} = 1.96$, $n = 100$, $\sigma = 7.1$ and $\bar{X} = 110$.

- The CI is:

$$\left[110 - 1.96 \times \frac{7.1}{\sqrt{100}}, 110 + 1.96 \times \frac{7.1}{\sqrt{100}}\right] = [108.6, 111.4]\,.$$

- To construct an 80% CI only one thing has to change, that is we only have to replace the 1.96 above with another number. To find this value go to the normal table an look inside it for 0.1, you should see -1.28. Replace 1.96 with 1.28 to give

$$\left[110 - 1.28 \times \frac{7.1}{\sqrt{100}}, 110 + 1.28 \times \frac{7.1}{\sqrt{100}}\right] = [109.1, 110.9].$$

- If we want the MoE to have length 0.45, then the interval

$$\left[\bar{X} - 1.28 \times \frac{7.1}{\sqrt{n}}, \bar{X} + 1.28 \times \frac{7.1}{\sqrt{n}}\right]$$

must have length 1. This means that

$$MoE = 0.5 = 1.28 \times \frac{7.1}{\sqrt{n}}.$$

Solve this (or use the formula) to give

$$n = \left(\frac{1.28 \times 7.1}{0.45}\right)^2 = 400.$$

- Hence we need to sample at least 400 cups to obtain a margin of error which is 0.45 (half of what existed previously).

Example 6 (when standard deviation is unknown)

How large a sample size do we require such that the margin of error for a 95% confidence interval for the mean of human heights is maximum 0.25 inch. The standard deviation is unknown, but it is believed that σ lies somewhere between 2-5 inches.

- Why this question matters: In general the standard deviation will be unknown. But we can 'guess' limits on how large or small it is based on own expertize.

Solution 6

The more variable the data the larger the confidence interval. Therefore, when given a range of standard deviations and our aim is that the margin of error should be *no larger* than 0.25 (i.e. 0.25 or **less**), then we need to use the largest standard deviation in the given range in the calculation

- In other words

$$n = \left(\frac{1.96 \times \sigma_{\text{largest}}}{0.25}\right)^2 = \left(\frac{1.96 \times 5}{0.25}\right)^2 = 1537.$$

- For any other $\sigma < 5$, using $n = 1537$ will lead to a Margin of Error which is *less* than 0.25. To see why, recall

$$MoE = 1.96 \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{\sigma}{\left(\frac{1.96\times 5}{0.25}\right)} = 0.25 \times \frac{\sigma}{5}.$$

Thus we see that if the true $\sigma < 5$, then the MoE will be less than 0.25, since $\sigma/5 < 1$.

- If we use $\sigma = 2$ in the margin of error calculation, then $n = 246$. However, if the true $\sigma > 2$ using $n = 246$ will lead to a Margin of Error which is *larger* than 0.25.

Margin of Error calculations using software

- There are various software tools on the web that will do margin of error calculations. For example,

  https://www.emathhelp.net/calculators/probability-statistics/margin-of-error-calculator/

- Here is one by survey monkey

  https://www.surveymonkey.com/mp/margin-of-error-calculator/. This calculator is specifically designed for calculating the MoE of proportions; where the standard deviation need not be specified. Here is another one

  http://www.raosoft.com/samplesize.html, which which can give smaller sample sizes if a proportion is specified. We cover this later on in the course.

- The calculations done in class assume that the population size is infinite (or that the sample is a SRS, that is the same person is sampled again).

However, when surveying certain populations the population size will be finite.

- Therefore, some calculators will also ask for the population size. Using the finite population size they make what is called a finite sample correction. You can read more about it here:

https://en.wikipedia.org/wiki/Margin_of_error#Effect_of_population_size.

Example 7

A confidence interval for the length of parrots is [4,10] inches. It is based on a sample size n. By what factor should the sample size increase such that the margin of error reduces to 1?

Solution 7

- The original margin of error is 3. Thus $1.96 \times \sigma/\sqrt{n} = 3$.
- We want to increase the sample size such that it decreases to 1.

$$1.96 \times \frac{\sigma}{\sqrt{\text{Factor} \times n}} = 1$$

$$\Rightarrow \quad 1.96 \times \frac{\sigma}{\sqrt{\text{Factor}} \times \sqrt{n}} = \frac{3}{\sqrt{\text{Factor}}} = 1.$$

  Solving for this we see that we need to increase the sample size by factor 9 in order to decrease the margin of error by a factor 3.

- Conclusion An extremely large increase in sample size has to be made for a moderate reduction in margin of error.