

Sociology 541
Thursday, February 22 2001
WEEK 6

Review

A sampling distribution of sample means has three important characteristics.

1. The distribution of sample means will approximate the form of a normal distribution if the size of each of the samples is equal to or greater than 30. (Central Limit Theorem)
2. The **mean** of the distribution of sample means ($\mu_{\bar{x}}$) equals the real population mean (μ).

If the population mean is the mean of the sampling distribution and the sampling distribution approximates the normal distribution, we can start to make statements about how likely it is to observe a particular sample mean.

We know that 68% of all sample means are likely to be +/-1 standard deviation from the population mean, 95% within +/- 2 standard deviations of the population mean and so on.

3. We can also make statements about the dispersion or variability in a sampling distribution, just as we do with a sample or population distribution. The measure of dispersion for a **sampling distribution** is called the **standard error** and is directly related to the standard deviation of a population.

Formula to calculate the standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The **standard error** is the number that results from repeatedly taking samples of size n from the population, finding the mean for each set of n observations and calculating the standard deviation of the \bar{x} values.

The spread of the sampling distribution depends on two factors.

1. Spread of the population distribution
2. Sample size n

Example

Suppose a population has a mean of 100 and a standard deviation of 15 and we draw a random sample of N=400 cases. What is the sampling error (standard error) of the sampling distribution of means for all samples of size 400 that can be drawn from the population?

In reality, we don't know the standard deviation of the population. Instead, we use the sample standard deviation. However, we make an adjustment to the denominator, using n-1 rather than n, since the standard deviation of the sample is likely to be an underestimate of the population standard deviation, especially when n is small.

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n-1}}$$

In summary:

When we conduct research, usually all we have is information on the distribution of the variable in the sample. Using n and s from the sample, we can calculate the standard error of the sampling distribution, telling us the amount of variation our measure has in the population. Once we know this, we can assess how good an estimate our sample mean is of the population mean.

Using Sample Data to Estimate Population Parameters

Point Estimation

A **point estimate** of a population parameter is a single value based on sample data that is used to estimate the population parameter.

A sample statistic (e.g. the sample mean) is a point estimate of the population mean.

We characterize these statistics in a couple of different ways:

1. A statistic with mean value equal to the value of the population parameter being estimated is said to be an **unbiased** point estimate of the population parameter. A statistic that is not unbiased is said to be **biased**.
2. A statistic with as small a standard error as possible is said to be **efficient**.

Confidence Intervals

Interval Estimation

Since the point estimate of a population parameter rarely equals the population parameter, it would be more useful to come up with a range of plausible values for our population parameter based on the information contained in the sample at our disposal.

A **confidence interval (interval estimate)** for a population parameter is an interval or range of plausible values for the population parameter. It is constructed so that, with a chosen degree of confidence, the value of the population parameter will be captured inside the interval.

The probability that the confidence interval contains the parameter is called the **confidence coefficient**. This is a chosen number close to 1, such as 0.95 or 0.99.

Error probability: The probability that a confidence interval does not contain the parameter.

Error Probability = 1 – the confidence coefficient.

Notation

Error probability is denoted by α .
(1 - α) is the confidence coefficient.

A Large-Sample Confidence Interval for a Population Mean

The Central Limit Theorem states that for large random samples, the sampling distribution of \bar{x} is approximately normal.

Once the sample is selected, if the sample mean does fall within 1.96 standard error units of the population mean, then the interval from $\bar{x} - 1.96 \sigma_{\bar{x}}$ to $\bar{x} + 1.96 \sigma_{\bar{x}}$ contains the population mean with probability of 0.95.

The interval $\bar{x} \pm 1.96 \sigma_{\bar{x}}$ is an interval estimate for the population mean with a confidence coefficient of 0.95. It's also known as a 95% confidence interval.

Lower confidence limit (LCL): $\bar{x} - 1.96 \sigma_{\bar{x}}$

Upper confidence limit (UCL): $\bar{x} + 1.96 \sigma_{\bar{x}}$

Example

GSS survey questions have asked respondents how many female partners they have had sex with since their 18th birthday. In 1994, of the 1055 respondents who responded with a number higher than 0, the distribution was highly skewed to the right with a sample mean of 10.2 and standard deviation of 10.1. Construct a 95% confidence interval around the population mean.

Controlling the Confidence Coefficient and Error Probability

Sometimes, we may want a greater degree of confidence in our interval. For example, we might construct a 99% confidence interval for the population mean.

We know that 99% of observations from a normal distribution occur within 2.58 standard deviations of the mean, so 0.99 of sample means fall within 2.58 standard errors of the population mean.

Construct a 99% confidence interval for the population mean μ for example above:

We can express these confidence intervals more generally:

When n is large, a $(1 - \alpha)100\%$ **confidence interval for the population mean, μ** , is

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)$$

where z_{α} is the upper $(\frac{\alpha}{2})100\%$ point of the standard normal distribution. An abbreviated formula for the interval is

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

where $+$ gives the upper limit and $-$ gives the lower limit of the interval.

When obtaining a confidence interval for the population mean when the sample size is large, if σ is unknown, we replace it by the sample standard deviation s in the formula for the confidence interval giving

$$\bar{x} \pm z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

This formula is more useful in practice.

For 95% and 99% confidence intervals, z equals 1.96 and 2.58, respectively.

What is the value of z for a 98% confidence interval?

80 % confidence interval?

Properties of the Confidence Interval for a Mean

1. Interpretation: If we repeatedly select random samples of size n and each time constructed a 95% confidence interval, then in the long run about 95% of the intervals would contain the population mean.
2. The greater the confidence level, the wider the confidence interval.
3. The larger the sample, the better we can estimate μ .

Example

Suppose instead that the information in example above were based on sample of size $n=4220$ (four times the actual sample size).

Large Sample Confidence Interval for a Population Proportion

Formulae to calculate descriptive statistics for a proportion.

Mean:

$$\bar{x} = p$$

Standard Deviation for a Large Sample:

$$s = \sqrt{p(1-p)}$$

Re-calculate the mean response and standard deviation for the three years in the question below using these formulae.

The General Social Survey asks whether respondents favor or oppose capital punishment for murder. Following are the responses for 1972, 1980 and 1987:

	1972	1980	1987
Yes = 1	852	982	1012
No = 0	632	390	354

Confidence Intervals for Population Proportions

Notation

π : the parameter representing a population proportion (analogous to μ for a population mean).

Point estimate of population proportion π is the sample proportion (denoted by p).

Population proportion π is $p = (\text{number of outcomes of interest in the sample})/(\text{total number sampled})$.

The **standard error** for a proportion for a large sample is

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

When n is large, a $(1 - \alpha)100\%$ **approximate confidence interval for the population proportion, π** , is

$$\left(p - z_{\frac{\alpha}{2}} \times \sqrt{\frac{p(1-p)}{n}}, p + z_{\frac{\alpha}{2}} \times \sqrt{\frac{p(1-p)}{n}} \right).$$

Example

1994 GSS question: Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason.

1934 respondents with 895 responding yes and 1039 responding no.

What is the estimated proportion in the population responding yes with 95% confidence intervals?

When asked whether abortion should be available if the woman becomes pregnant as a result of rape, 1616 said yes and 318 said no. (N=1934)

Calculate new mean and 95% confidence intervals:

Choosing the Sample Size

Sample Size for Estimating Population Mean

When estimating the population mean for a large sample, the margin of error (E) or width of the interval (i) is $\pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$. For a given level of confidence, say $(1 - \alpha)100\%$, and width of confidence interval i , the sample size required is

$$n \geq \left(\frac{z_{\frac{\alpha}{2}} \times \sigma}{i} \right)^2.$$

If σ is not known, we replace it with s . The smallest integer value for n satisfying this relationship is chosen. In practice, σ is not known.

Example

An estimate is needed of the mean acreage of farms in Canada. The estimate must be correct to within 25 acres with a confidence level of 0.95. A preliminary study suggests that 200 acres is a reasonable guess for the standard deviation of farm size. How large a sample of farms is required?

Suppose now that a sample is selected of the size believed to be needed to estimate the mean correct to within 25 acres with a confidence level of 0.95. Suppose, however, that the sample has a standard deviation of 400 acres, rather than 200. Then, how close can we expect the sample mean to be to the true mean?

Sample Size for Estimating Proportions

Confidence intervals for a proportion, like the one for a mean, apply for large samples. When a proportion is between about 0.30 and 0.70, the usual sample size criterion for a mean (30) works fine. When proportions are less than 0.3 or greater than 0.7, the distribution is skewed and requires a larger sample size to achieve normality. In that case, you should have at least ten observations both in the category of interest and not in it.

When we are estimating a population proportion π , the sample size required for a given level of confidence, say a $(1 - \alpha)100\%$, and width of interval i , is

$$n \geq \left(\frac{z_{\frac{\alpha}{2}}}{i} \right)^2 \times p(1 - p).$$

If information about the expected proportion (p) from a previous sample is not available, then the worst case scenario would be used, which happens when $\pi = 0.5$. The worst case scenario uses the value of p which gives the largest value of $p(1 - p)$.

Example

A group of social scientists want to estimate the proportion of school children in Boston who are living with only one parent. They decide they want a sample size that, with a confidence level of 0.95, the error will not exceed 0.04.

What if you are now informed that, based on past studies, the proportion of school children in Boston who were living with only one parent was no more than 0.25. What is an adequate sample size?

PROBLEMS

1. A certain university claims its recent graduates earn an average annual income of \$20,000. To determine the legitimacy of this claim, we take a random sample of 100 alumni who had graduated within the last 2 years. In the process, we get a sample mean of only \$18,500 and a sample standard deviation of \$7000. How probable is it that we would get a sample mean of \$18,500 or less if the true population mean is actually \$20,000?
2. Revisiting the problem regarding the sample size for farms to estimate acreage. Suppose now that you only need the estimate of mean acreage to be correct with a confidence level of 0.75. What sample size is required now?
3. Out of an election day sample of 400 individuals who voted in the gubernatorial election, 160 voted for Jones and 240 for Smith.
 - a. Assuming this is a random sample of all voters, construct a 99% confidence interval for the proportion of votes that Jones will receive. Do you think that Jones will lose the election? Why?
 - b. Suppose, instead, that the sample size had been 40, of whom 16 voted for Jones. Again, find the 99% confidence interval and, if possible, predict the winner. How does the result compare to (a)?
4. In the 1994 GSS, responses of 1964 subjects to the question, "On the average day, about how many hours do you personally watch television?" had a mean of 2.8. The standard error of the mean was 0.05.
 - a. Calculate a 98% confidence interval for the mean daily time spent watching television. Interpret.
 - b. Find the standard deviation of the time spent watching television. Do you think that the distribution is bell-shaped? Why or why not?