Confidence Intervals

Point Estimation Vs Interval Estimation

- Point estimation gives us a particular value as an estimate of the population parameter.

- Interval estimation gives us a range of values which is likely to contain the population parameter. This interval is called a confidence interval.

- This procedure also tells us how likely this interval is to contain the actual parameter. That value is called the confidence coefficient or confidence level. Usually α denotes 1−confidence level. So, confidence level=$1-\alpha$.

- A confidence interval is a random interval.

- A two-sided interval is given by two statistics (lower and upper bounds for the interval, respectively).
- An upper or lower interval is given by one statistic only (upper or lower bounds, respectively).

Interpretation for the Confidence Coefficient

- Remember that the parameter is fixed but the confidence interval is random.

- If different samples are drawn, the confidence intervals will be usually different.

- But if you follow the same rule (use the same statistics for the bounds of the interval) to construct the confidence intervals repeatedly, in the long run the proportion of the time the interval will contain the actual parameter will be approximately equal to the confidence coefficient.

Example

We want to compare the mean family income in two states. For state 1, we had a random sample of $n_1 = 100$ families with a sample mean of $\bar{x}_1 = 35000$. For state 2, we had a random sample of $n_2 = 144$ families with a sample mean of $\bar{x}_2 = 36000$. Past studies have shown that for both states $\sigma = 4000$. Estimate $\mu_1 - \mu_2$ and place a two-standard-error bound on the error of estimation. How much confidence do we have in this interval?

Confidence Intervals for Large Sample Size

If the sample size is large, how can we find a confidence interval for a parameter θ given a certain confidence coefficient?

Example

We want to compare the proportion of families earning above \$40,000 in two states. For state 1, we had a random sample of $n_1 = 100$ families with a sample proportion of $\hat{p}_1 = 0.30$ making above \$40,000. For state 2, we had a random sample of $n_2 = 144$ families with a sample proportion of $\hat{p}_2 = 0.32$.

Find a 99% confidence interval for $p_1 - p_2$.

One-sided Confidence Intervals

- One-sided confidence intervals do exist, but aren't used very often.

- Large sample one-sided confidence intervals:

  - Upper bound only: $P(\frac{\theta-\hat{\theta}}{SE_{\hat{\theta}}} \leq z_{\alpha}) = 1-\alpha$ yields an interval of the form $(-\infty, \hat{\theta} + z_{\alpha}SE_{\hat{\theta}}]$.

  - Lower bound only: $P(\frac{\theta-\hat{\theta}}{SE_{\hat{\theta}}} \geq -z_{\alpha}) = 1-\alpha$ yields an interval of the form $[\hat{\theta} - z_{\alpha}SE_{\hat{\theta}}, \infty)$.

- Can be useful when you're interested in the percentage of samples for which the parameter is less/greater than the one boundary point.

Example

We want to compare the proportion of families earning above \$40,000 in two states. For state 1, we had a random sample of $n_1 = 100$ families with a sample proportion of $\hat{p}_1 = 0.30$ making above \$40,000. For state 2, we had a random sample of $n_2 = 144$ families with a sample proportion of $\hat{p}_2 = 0.32$.

Find a lower bound for $p_1 - p_2$ in which you have 99% confidence.

Summary of Conf. Intervals (Large Sample)

Two-sided confidence interval for parameter θ is given by $\hat{\theta} \pm z_{\frac{\alpha}{2}} SE_{\hat{\theta}}$.

One-sided confidence intervals for parameter θ are given by $(-\infty, \hat{\theta} + z_{\alpha} SE_{\hat{\theta}}]$ or $[\hat{\theta} - z_{\alpha} SE_{\hat{\theta}}, \infty)$ where

- $\hat{\theta}$ is the point estimator
- $SE_{\hat{\theta}}$ is the standard error for the estimator
- Confidence coefficient: $1 - \alpha$

Confidence Intervals for Small Sample

- When sample sizes are large, we can assume normality of the sampling distribution for sample means or proportions (by CLT).

- Cases sometimes arise in which we can only obtain a small random sample.

- If we know that the population is normal or near normal, we can form confidence intervals based on the t distribution.

Small Sample Confidence Interval for μ

- Assume we have a small random sample ($n < 30$) from a population for which σ is unknown and the distribution is assumed to be near normal.
- In this case, we know that $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ has approximately a t distribution with $n-1$ degrees of freedom.
- Want a way to construct bounds such that, if we use the procedure with many different random samples, $1-\alpha$ proportion of samples will yield intervals that contain the mean.
- These bounds are given by $(\bar{x}-t_{\frac{\alpha}{2}}^{n-1}\frac{s}{\sqrt{n}}, \bar{x}+t_{\frac{\alpha}{2}}^{n-1}\frac{s}{\sqrt{n}})$.

Example: Small Sample Conf. Interval

A bottling factory fills thousands of 20oz bottles daily with soda, but not all the bottles are filled to the same level. A random sample of bottles was taken from the factory line, containing the following amounts of soda (in oz):

19.8 20.1 19.7 19.2 19.9 20.0 19.8 19.9 19.7

Assuming that the distribution of amounts of soda is approximately normal, find a 95% confidence interval for the mean amount of soda contained in the bottles. What does this say about the dependability of the factory's process?

Small Sample Confidence Interval for $\mu_1 - \mu_2$

- Assume we have small random samples from 2 populations for which the distributions are assumed to be near normal.

- As long as the variances of the populations can be assumed to be equal, the same general confidence interval strategy is effective.

- We can use the same general format for the confidence interval, which is $(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}^{n_1+n_2-2} SE_{\bar{x}_1-\bar{x}_2}$, but we need the estimator for the standard error of the difference in means.

Estimator for Standard Error of $\bar{x_1} - \bar{x_2}$

- In the past, we saw that the standard error for the difference in means was $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- Since we're assuming equal variances, this simplifies to $\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

- In most cases, we won't know σ so we need an estimator that takes the information from both samples into account.

- This estimator is $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$.

- So when we have small samples from near-normal populations with equal variances, the standard error estimate is $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Example: Matched Design

A task in a factory can be performed in 2 ways; management is interested in knowing which method takes less time. A random sample of employees is timed performing the task using method A and method B.

Method A	6.0	5.0	7.0	6.2	6.0	6.4
Method B	5.4	5.2	6.5	5.9	6.0	5.8
Difference	0.6	-0.2	0.5	0.3	0.0	0.6

Find a 90% confidence interval for the difference in the completion time. Assume that the time differences are distributed approximately normally.

Example: Conf. Interval for $\mu_1 - \mu_2$

Suppose in the previous example two different samples of 6 employees were chosen at random. Assume that the samples are independent. The employees of the first sample were timed performing the task using method A and the employees of the second sample, using method B. Use the same data as before to find a 90% confidence interval for the difference in the completion time. Assume that the SD for the time for both methods are equal and time for both methods are normal.