

Confidence intervals and hypothesis tests are both forms of formal statistical inference. There is a strong connection between them. When working with a single parameter, tests and intervals are essentially equivalent, but this equivalence is lost when there is more than one parameter.

If we have a confidence interval for a single target parameter, say the population mean (μ), then we can convert this interval into a hypothesis test by declaring that the null hypothesis $\mu=c$ is rejected if c is not inside the confidence interval. Here c is a specific numerical value that must be pre-specified. The type I error rate of the resulting test is 1 minus the coverage probability of the confidence interval. Thus, a 95% coverage probability interval is equivalent in this sense to a hypothesis test with 5% type I error rate (i.e. $\alpha = 0.05$). There are variations on this equivalence for one-sided tests, which are equivalent to one-sided intervals, and for two-sided tests, which are equivalent to two-sided intervals.

Going the other way, if we have a hypothesis testing procedure for a parameter μ , then we can conduct a test of the null hypothesis $\mu=c$ for all possible values of c , at a fixed type I error rate. The set of all values of c for which the null hypothesis is not rejected is a confidence interval. If the tests are conducted with type I error rate α (e.g. $\alpha=0.05$), then the coverage probability of the resulting interval is $1 - \alpha$ (e.g. 0.95). This process is often called “inverting a hypothesis test to construct a confidence set”. Note that in practice it is not actually possible to conduct the test for all possible values of c , because there are infinitely many such values. There are numerical algorithms to address this computational issue, but we will not cover them here.

More subtle questions arise when we are working with two or more parameters. For example, suppose we have samples of female and male office workers, and are comparing their incomes. We can construct two 95% confidence intervals -- one for the women and one for the men. It is natural to think that if these two intervals do not overlap, then we can reject the two-sample test that the population means are different, and conversely, if the two intervals do overlap, then we do not reject the hypothesis that the population means are different. But in fact, only one of these assertions is true. When the confidence intervals do not overlap, then the two-sample test will always reject its null hypothesis. But the converse is actually not true. There are situations where the confidence intervals overlap, but where the hypothesis test will still reject its null hypothesis.

In general, it is safer to conduct a hypothesis test explicitly if that is the type of result that will be reported. But if for some reason it is not possible to conduct the hypothesis test, but the intervals are available (e.g. they appear in a paper but the primary data are not accessible), then it is important to

remember that the test results cannot always be inferred from knowing only whether the confidence intervals overlap.