The Effective Use of Statistical Tests

Some General Advice About Statistical Tests

We have now seen a number of situations were we can use statistical tests in order to help us make conclusions about our data. Some of the results of these tests have been quite important. For example we were able to show that we have strong evidence that the mean body temperature is not 37° Celsius, which was a result that held for more than 100 years. But any powerful tool must be used responsibly! In this section, we will talk about potential pitfalls that can occur when working with statistical tests of hypotheses and how you might avoid them.

- **Do not misinterpret p-values.**
  - A p-value does not tell you how likely it is that the null hypothesis is true. It does tell us how likely the observed data would be if the null hypothesis were true.
  - A p-value is a measure of the strength of the evidence, so report your p-value. It is better to state the actual number rather than saying p-value < 0.05.
- **Testing cannot correct flaws in the design of the data collection.**
  - Lack of randomization in choosing a sample, lack of control and lack of randomization in assigning treatments will lead to bias, confounding, and we will not be able to make causal conclusions.
- **Always use two-sided tests** (unless you are really sure that one direction is of no interest).
  - The p-value for the one-sided test is only half the p-value of the two-sided test, so for the one-sided test we are more likely to get a significant result.
  - Taking a look at your data and seeing if it is larger or smaller than you expect, and then deciding whether you should do a one-sided or a two-sided test is cheating.
- **Statistical significance is not the same thing as practical significance.**
  - Very small effects can be highly statistically significant when a test is based on a large sample; this does not mean it is practically important and the decision must be based on the context.
  - A large p-value does not necessarily mean that the null hypothesis is true. There may not be enough power to reject it.
- **Small p-values may occur:**

  - By chance.
  - Because of problems related to data collection.
  - Because of violations of the conditions necessary for the particular testing procedure being used.
  - Because the null hypothesis is false.
- **If multiple tests are carried out, some are likely to be significant by chance alone.**
  - If $\alpha = 0.05$, we expect significant results 5% of the time, even when the null hypothesis is true.
  - Be suspicious when you see a few significant results when many tests have been carried out, for example significant results on a few subgroups of the data.
- **Data snooping!**
  - Test results are not reliable if the statements of the hypotheses are suggested by the data.
  - Hypotheses should be specified before any data are collected.

The tests we have learned in previous sections for proportions and means require:

- independent observations,
- the sampling distribution of the estimators to be (approximately) normally distributed.

Any statistical procedure is **robust** if it is not sensitive to the violations in this necessary conditions. That is, the p-value is approximately correct even if the necessary conditions are not completely satisfied.

The question of how large of a sample size is required for a test depends on the scenario and which parameter we are trying to make inferences about.

- For tests of proportions, larger sample sizes are needed the further the true value of p is from 0.5.
- For tests of means, the $t-test$ is robust even for small sample sizes, except when there is extreme skew or outliers.

Statistical tests should always be preceded by exploratory analyses, using plots and summary statistics. They may indicate any problems with the data such as skewness or outliers, and illustrate any effect that you were seeking. If you cannot visualize something in a plot, you may wonder if it is practically important, even if it is statistically significant.

Fundamentally, testing cannot correct flaws in the data collection design. Be careful when collecting the data to avoid bias, confounding, and the inability to generalize, and other

consequences of lack of randomization.

It is easy to put data into a statistical software package and generate p-values; it is much harder to truly understand them and to assess if they really answer the question you want. But statistical testing works well if you design a study carefully to investigate focused questions, and then use tests to assess the evidence for or against your hypothesis.