Statistical hypothesis testing reflects the scientific method, adapted to the setting of research involving data analysis. In this framework, a researcher makes a precise statement about the population of interest, then aims to falsify the statement. In statistical hypothesis testing, the statement in question is the null hypothesis. If we reject the null hypothesis, we have falsified it (to some degree of confidence). According to the scientific method, falsifying a hypothesis should require an overwhelming amount of evidence against it. If the data we observe are ambiguous, or are only weakly contradictory to the null hypothesis, we do not reject the null hypothesis.

The framework of formal hypothesis testing defines two distinct types of errors. A type I error (false positive) occurs when the null hypothesis is true but is incorrectly rejected. A type II error occurs when the null hypothesis is not rejected when it actually is false. Most traditional methods for statistical inference aim to strictly control the probability of a type I error, usually at 5%. While we also wish to minimize the probability of a type II error, this is a secondary priority to controlling the type I error.

All the standard statistical testing procedures perform well at controlling type I error under ideal conditions, but most of them can break down and give misleading results in practice. That is, if we claim to be conducting a test that has a 5% false positive rate, is this the actual false positive rate? As we discussed earlier in the setting of confidence intervals, several complications that can arise in practice will result in a statistical procedure not performing as intended. In order to reduce the risk of this happening, statisticians use statistical theory and computer simulations to assess the operating characteristics of testing procedures in various challenging settings.

Normality of the data

One of the commonly-stated “assumptions” that is often raised as a caveat when presenting statistical findings is the issue of the data being normally distributed. While it is true that in some circumstances, strongly non-normal data can cause statistical tests to be misleading, in most settings, the data are not required to follow a normal distribution. In addition, there are other issues unrelated to normality that are potentially more likely to produce misleading results.

Concerns about normality primarily center on the calibration of rejection regions for a test statistic, or equivalently, the manner in which p-values are computed. As with confidence intervals, the Z-score plays the central role. To be concrete, suppose we are conducting a test comparing two population

means using independent samples (a “two sample t-test”). The corresponding sample means are x1_bar and x2_bar, and the test is based on the difference between them, which is x1_bar - x2_bar. This difference has a standard error, which we denote here by s (there are a few different ways to compute this standard error, s can refer to any of them here). The Z-score is (x1_bar - x2_bar) / s.

Under the very strong assumption that the data are normally distributed, the Z-score follows a Student t-distribution, with degrees of freedom depending on the way that the standard error was constructed (in most cases the degrees of freedom will be between m-2 and m, where m is the combined sample size of the two samples being compared). If the data are not normally distributed, we can appeal to the central limit theorem (CLT), which states that the Z-score will be approximately normally distributed as long as the sample size is not too small.

As discussed earlier in the setting of confidence intervals, there is no universal rule that states when the sample size is large enough to justify invoking the CLT. Rules of thumb between 20 and 50 are often stated. A smaller sample size is sufficient to invoke the CLT when the data approximately follow a normal distribution, and a larger sample size is needed if the data are strongly non-Gaussian.

While normality is a consideration in some settings, it is mainly relevant when the sample size is small and the data are strongly non-Gaussian. Other issues can cause statistical tests to break down in a broader range of settings, so normality of the data should be seen as one of several factors that can impact the performance of a statistical test, and is often a relatively minor one at that.

A useful approach in practice is to use the Student t-distribution to calculate rejection regions and p-values, even when the data are not expected to follow a normal distribution. Doing so is slightly conservative, in that the rejection region based on the t-distribution will be slightly smaller than the rejection region based on the normal distribution, and p-values based on the t-distribution will be slightly larger than p-values based on the normal distribution. As the sample size grows beyond around 50, there is little practical difference between using the t-distribution and the normal distribution when carrying out statistical hypothesis tests.

Clustering and data dependence

One additional issue that can adversely impact the performance of a statistical test is the presence of unknown (or unmodeled) correlations or clustering in the data. For example, if the data values are

observed in sequence (e.g. over time), with each value possibly being correlated with its neighbors, then we have “autocorrelation”, which is a form of dependence. Alternatively, we may have some form of grouping or clustering in the data. Methods for addressing these issues will be discussed in Course 3 of this specialization.

Causality

Another issue to be aware of when conducting statistical tests is that of confounding and causality. This issue is especially relevant when interpreting the results of a statistical test. Suppose, for example, that two groups of people differ significantly in terms of some trait, e.g. people with fewer dental cavities are seen to have statistically lower risk of heart disease compared to people with more cavities. It is important to note that this effect could be due to a lurking factor, or to some form of selection bias. For example, the people with fewer cavities may be less likely to be smokers. Note that this is arguably not a problem with the statistical hypothesis test itself -- it may well be true that people with fewer cavities are less likely to have heart disease (in the sense of there being a real association between these two factors). Rather, it is an issue of what substantive conclusions may be drawn, especially when people draw a causal conclusion where one is not warranted.

Multiplicity

A third pitfall that arises with statistical hypothesis testing, as well as with other forms of statistical inference such as confidence intervals, is that of “multiplicity”, which we will discuss next. Note that a variety of terms have been used to describe this issue, including “data dredging”, “multiple testing”, and “p hacking” (in reference to “p values”).

Most statistical inference procedures, including both confidence intervals and hypothesis testing, are based on an idealized research design in which a single analysis with narrow scope is conducted using a data set. In practice, data analysis often involves data exploration coupled with formal inference. It is now widely accepted that doing this heedlessly can lead to misleading results, and in particular often leads to statistical evidence for research findings being overstated. This is a large topic; here we comment on a few aspects of it, starting with two examples of how multiplicity in data analysis can cause problems.

- Suppose that a researcher habitually conducts a two-sample t-test comparing group means, then reports a confidence interval for the difference in population means only if the null hypothesis of the t-test is rejected. On one hand, this researcher is exhibiting good judgment, as it is often recommended to report an effect size along with the results of any hypothesis test (the point estimate of the population mean difference is an effect size in this setting). However, the confidence intervals selected in this way will have lower than the nominal coverage probability. This is an example of a broader issue sometimes called “selective inference”.
- Suppose that there are natural ways to divide the population into subgroups. For example, imagine that a researcher is interested in whether people who sleep less than 7 hours per night on average during one month have greater gain of body weight in the following year. The researcher’s initial plan may have been to consider the general population, but perhaps the investigator then decides to carry out analyses separately in women and in men, in older people and in younger people, in smokers and in non-smokers, etc. Although it is possible that hypothesized effect may actually be much stronger in some of these subgroups than in others, repeatedly testing the same data on different subgroups will be likely to give rise to falsely positive evidence for an association. For example, if the researcher conducts 3 independent tests on the same data (e.g. on different subgroups of the data), with each test having a 5% false positive probability, then the probability that at least one of the tests will yield a false positive is over 14%.

In recent years, researchers have identified more and more ways that multiple-testing can arise in practice, corrupting statistical findings. Almost always, multiple testing leads to overstatements of the confidence in findings, or to erroneous findings being reported. Fortunately, there are many approaches to remedying this issue. The easiest of these to apply is the Bonferroni correction, which essentially involves multiplying all p-values by the number of tests that were performed. For example, if we conduct 5 hypothesis tests, and one of them yields a p-value of 0.02, then we should adjust this p-value to 0.1 (= 0.02 * 5). Thus, this test which would have been deemed to be “statistically significant” if conducted in isolation will be not be seen as such following the Bonferroni adjustment.

Power

A final consideration we will discuss here is statistical power. Power is often defined in a narrow sense as the probability of rejecting the null hypothesis when the null hypothesis is false. Loosely speaking, this is the probability of not making a type II error. More broadly, power can refer to any aspect of the study design or data analysis that would make it more likely for meaningful results to be attained. There is a branch of statistics focusing on formal “power analysis” that aims to develop concrete and quantitative ways to assess the statistical power in a given setting. We will not delve into these methods here, and instead will discuss at a higher level how low power intersects with and exacerbates some of the other complicating considerations discussed above.

The type I error rate is controlled by the researcher (say at 5%), but this only represents the risk of drawing a false conclusion from a single test. In recent years, the notion of the “false discovery rate” (FDR) has been advanced to understand how often the conclusions drawn in a research process involving multiple formal inferential procedures are mistaken. Focusing on hypothesis tests, we can consider a situation where, for example, five tests are to be conducted. If two of the underlying null hypotheses are false, but the power to reject them is low (say it is only 20%), then around 25% of all the rejected null hypotheses were incorrectly rejected. This shows how the FDR is different than the type I error rate, which remains controlled here at 5%.

People sometimes incorrectly believe that controlling the type I error rate at 5% means that there is only a 5% chance that any reported finding is wrong. As illustrated here, the probability that a reported finding is wrong can be much higher than the type I error rate. This “error inflation” is primarily driven by two factors -- a researcher who pursues hypotheses that are unlikely to be correct, and a researcher who carries out studies with low statistical power will both have higher FDR in their work overall. The latter issue is in principle addressable by encouraging researchers to pursue fewer, but higher quality studies (i.e. to pursue fewer studies with larger sample sizes rather than many studies with small sample sizes). The former issue is harder to address, but it reflects the fact that in some fields, especially difficult areas of science such as genetics and neuroscience, there is a poor fundamental understanding of the systems under study, which may lead to people speculating and pursuing hypotheses with weak theoretical grounding.