BIOL 422 & 680, Pedro Peres-Neto, Biology, Concordia University

A pedagogical guide for understanding the issues underlying

Multiple hypothesis testing

Should we trust the results from multiple statistical tests? The short answer is no! But why not? It turns out that to answer this question one has to have a very good understanding of the logic (not the mathematics) underlying statistical hypothesis testing. Although statistical hypothesis testing is the most used approach in research to assess the significance of an effect of interest in data (e.g., do odd- and even- day born individuals vary in their preference for chocolate?); **YET** conceptually (though not operationally), statistical *hypothesis testing is an intimate stranger*. Operationally it is not an intimate stranger as a large number of people can conduct a statistical test and arrive at a statistical conclusion (reject or do not reject the H_0) without clearly understanding the concepts involved. This text is meant to provide you the necessary conceptual basis to improve your understanding about the logic underlying hypothesis testing. The exposition might sound pretty redundant and obvious to some, but not to all and will hopefully help to build a much-needed clear understanding of the logic underlying statistical hypothesis testing.

1) IN ORDER to provide evidence for or against a null hypothesis (H_0), we need to first accept it as true. This sounds confusing at first to most. We accept it as true to be able to identify the corresponding sampling distribution for the test statistic of interest (e.g., the sampling distributions associated with the theoretical value for which the null hypothesis is true, i.e., $\mu_{odd} - \mu_{even} = 0$). The standard t distribution is such a distribution. The standard t distribution corresponds to the sampling distribution for many types of statistics under the null hypotheses (e.g., regression slopes = 0, correlations = 0), including the difference between two samples means. In this case, *the standard t distribution contains all infinite t standardized values calculated by subtracting all infinite pairs of means sampled from two populations with the same mean* (assuming H_0 is true); each pair of differences are standardized by the standard error to form the standardized t distribution.

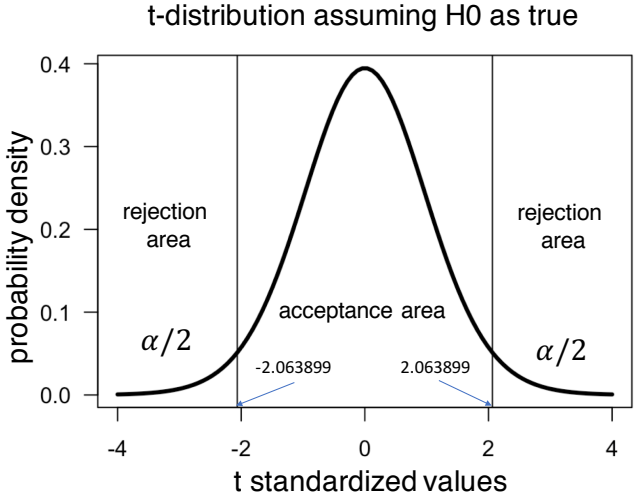
This is key: although the standardized t distribution refers to the distribution of all possible standardized t values assuming H_0 as true, standardized t values can be calculated for two sample means from the same (H_0 is true) or different (H_0 is false) populations. We never know if two samples

come or not from two populations with the same mean (that's why we make statistical guesses based on hypothesis testing). Again, all values in the standardized t distribution represent the standardized t values assuming H_0 as true (i.e., calculated as explained in the passage in italic in the paragraph above; worth reading it again if you are unsure). However, the observed standardized t value (i.e., the one for your two sample means) may or may not represent samples taken from two populations with the same mean. That is what we are trying to determine! The logic underlying statistical hypothesis testing (under a frequentist framework) is to establish whether the observed standardized t value (your value) fits well among the values of the standardized t distribution produced by assuming H_0 as true (i.e., a likely or probable standardized t value among all infinite values for the distribution) or fits badly (an unlikely or improbable value).

If the observed standardized t value (yours), for which we have no idea whether it represents the difference between two samples taken from the same or different populations, is very unlikely under a distribution for which all values were calculated by assuming H_0 as true (standardized t distribution), then we should be inclined to reject H_0; even though in the (unknown) reality we could be wrong (i.e., albeit too unlikely, the observed standardized t value did come from samples of the same population (by rejecting, you would inadvertently (without intention) had committed a mistake that you would have to live with the consequences of that mistake; this mistake is commonly called a false positive in statistics).

The term "unlikely" is subjective! As such, we need to establish what unlikely represents by setting a significance level (α). If there is a chance smaller than alpha (say 5% or 0.05) to find a value in the standardized t distribution that is as extreme or extreme in contrast to the observed (your standardized t value), then we state "unlikely" to have occurred *given the null hypothesis*. The term "given the null hypothesis" means "given that the null hypothesis is true".

2) Sampling distributions for statistical testing are "built" assuming H_0 as true - the case of the standardized t distribution.



The distribution presented here is the standardized t distribution for 24 degrees of freedom given that say 26 students filled the survey (2 degrees of freedom are lost because the standard deviation of both samples are used). The two vertical lines correspond to the two-standardized t-values (-2.063899

and 2.063899) that separate the acceptance and rejection regions based on a significance level (α) equal to 0.05 (two-sided or two-tailed test). It is vital to insist that all the values in this distribution were achieved by assuming H_0 as true (i.e., variation in t standardized values are just due to pure sampling variation between two sample means coming from populations with the same mean). The standardized t distribution varies from -∞ to ∞ (in the graph above, the distribution was truncated between -4 and 4). Therefore, any possible (including yours) observed standardized t value lies within this distribution.

3) THE P-VALUE is the number of values in the standardized t distribution (i.e., assuming H_0 as true) that are (equal or smaller) or (equal or greater; i.e., two-sided test) than the observed t-statistic (yours) divided by the total number of samples used to build the standardized t-distribution. This is a "pedagogical simplification" as this number is infinite, so we use the areas under the distribution curve to calculate this probability. As such, the P-value is the probability of finding the observed or more extreme results when the null hypothesis (H_0) of a study question is true (i.e., no true difference in preferences between populations of odd and even birthdays for a given preference (e.g., Chocolate). So, imagine that your observed P-value is tiny (P = 0.00001); this denotes that the chance of sampling two sample means that led to an observed standardized t value (yours) as extreme or more extreme in the standardized t distribution is 0.001%. So, values as extreme or more extreme than your observed standardized t value is pretty rare among the standardized t values that were used to build a distribution made by all standardized t values for which the H_0 is true (i.e., the standardized t distribution).

4) STATISTICAL SIGNIFICANCE is reached when the observed standardized t value (based on a study) is very unlikely (according to a particular chosen significance level α) among the values in the standardized t distribution (i.e., for which H_0 is true). Decisions involve:

- **Do not reject H_0** if the P-value (for the observed difference) is greater than alpha. In this case, obviously, the observed t standardized value for the observed difference between the two means (odd - even) for any given preference (e.g., Chocolate) is either greater than -2.063899 or smaller than 2.063899 (i.e., they lie in the acceptance area). If we do not reject H0, then we say that there is a lack of statistical significance or lack of support against H_0.

- **Reject H_0** if the P-value (for the observed difference) is smaller than alpha. In this case, obviously, the observed t standardized value for the observed difference between the two means (odd - even) for any given preference (e.g., Chocolate) is either smaller than -2.063899 or greater than 2.063899 (i.e., they lie in the acceptance area). If we reject the H_0, then we say that there is statistical significance or support against H_0.

5) RATIONALE: a statistical decision can be right or wrong though we don't know whether this decision is truly (in reality) right or wrong. By reading carefully the text so far, you should be able to understand that the significance level (α) is simply the probability of committing a Type I error (i.e., reject H_0 when in reality H_0 is true). This is because all the values in the standardized t distribution were calculated on the basis of samples from two populations with the same mean. So, even the values that are smaller than -2.063899 or greater than 2.063899, though unlikely, they are possible values in the standardized t distribution (for which H_0 is true). So, by stating that we reject H_0 whenever the observed standardized t statistic (yours) is significant when its associated probability (P-value) is small than α, we are basically accepting a chance of 5% (0.05) to be wrong when stating that we should reject H_0. Again, this error is often referred to false positives.

Now, if the chance of committing a Type I error (i.e., false positive) is 0.05 for one single observed standardized t statistic, what is the expected error for when multiple standardized t statistics are being contrasted against the same standardized t distribution? Because the standardized t distribution varies from $-\infty$ to ∞ and as such includes all possible infinite standardized t values; by the way, they also include the standardized t values contrasting differences between two samples of populations with different means). Regardless if you increase the number of tests, this will certainly increase the chances of finding values that are large, thus falling within the rejection area. This is key to understand the initial question and answer - **Should we trust the results from multiple statistical tests? I hope you understand the reason why the answer is no!**