Multiple Hypothesis Testing and False Discovery Rate
(Some materials are from Answers.com)
STATC141

Type I and Type II errors

- **Type I error**, also known as a **"false positive"**: the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. Plainly speaking, it occurs when we are observing a difference when in truth there is none (or more specifically - no statistically significant difference). So the probability of making a type I error in a test with rejection region R is $P(R \mid H_0 \text{ is true})$.
- **Type II error**, also known as a "**false negative**": the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. Plainly speaking, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is $1 - P(R \mid H_a \text{ is true})$. The power of the test can be $P(R \mid H_a \text{ is true})$.

Understanding Type I and Type II Errors

Hypothesis testing is the art of testing if variation between two sample distributions can just be explained through random chance or not. If we have to conclude that two distributions vary in a meaningful way, we must take enough precaution to see that the differences are not just through random chance. At the heart of Type I error is that we don't want to make an unwarranted hypothesis so we exercise a lot of care by minimizing the chance of its occurrence. Traditionally we try to set Type I error as .05 or .01 - as in there is only a 5 or 1 in 100 chance that the variation that we are seeing is due to chance. This is called the 'level of significance'. Again, there is no guarantee that 5 in 100 is rare enough so significance levels need to be chosen carefully. For example, a factory where a six sigma quality control system has been implemented requires that errors never add up to more than the probability of being six standard deviations away from the mean (an incredibly rare event). Type I error is generally reported as the p-value.

Statistics derives its power from random sampling. The argument is that random sampling will average out the differences between two populations and the differences between the populations seen post "treatment" could be easily traceable as a result of the treatment only. Obviously, life isn't as simple. There is little chance that one will pick random samples that result in significantly same populations. Even if they are the same populations, we can't be sure whether the results that we are seeing are just one time (or rare) events or actually significant (regularly occurring) events.

Multiple Hypothesis Testing

In Statistics, **multiple testing** refers to the potential increase in Type I error that occurs when statistical tests are used repeatedly, for example while doing multiple comparisons to test null hypotheses stating that the averages of several disjoint populations are equal to each other (homogeneous).

Intuitively, even if a particular outcome of an experiment is very unlikely to happen, the fact that the experiment is repeated multiple times will increase the probability that the outcome appears at least once. As an example, if a coin is tossed 10 times and lands 10 times on tail, it will usually be considered evidence that the coin is biased, because the probability of observing such a series is very low for a fair coin ($2^{-10} \approx 10^{-3}$). However, if the same series of ten tails in a row appears as part of 10,000 tosses with the same coin, it is more likely to be seen as a random fluctuation in the long series of tosses.

If the significance level for a given experiment is α, the experimentwise significance level will increase exponentially (significance decreases) as the number of tests increases. More precisely, assuming all tests are independent, if *n* tests are performed, the experimentwise significance level will be given by $1 - (1 - \alpha)^n \approx n\alpha$ when α is small Thus, in order to retain the same overall rate of false positives in a series of multiple tests, the standards for each test must be more stringent. Intuitively, reducing the size of the allowable error (alpha) for each comparison by the number of comparisons will result in an overall alpha which does not exceed the desired limit, and this can be mathematically proved true. For instance, to obtain the usual alpha of 0.05 with ten tests, requiring an alpha of $.005 = 0.05/10 = \alpha/n$ for each test can be demonstrated to result in an overall alpha which does not exceed 0.05. This technique is known as **Bonferroni correction**.

However, it can also be demonstrated that this technique may be conservative (depending on the correlation structure among tests), *i.e.* will in actuality result in a true alpha of significantly less than 0.05; therefore raising the rate of false negatives, failing to identify an unnecessarily high percentage of actual significant differences in the data. For this reason, there has been a great deal of attention paid to developing better techniques for multiple testing, such that the overall rate of false positives can be maintained without inflating the rate of false negatives unnecessarily.

False Discovery Rate

For large-scale multiple testing (for example, as is very common in genomics when using technologies such as DNA microarrays) one can instead control the false discovery rate (FDR), defined to be the expected proportion of false positives among all significant tests.

False discovery rate (FDR) controls the expected proportion of incorrectly rejected null hypotheses (type I errors) in a list of rejected hypotheses. It is a less conservative comparison procedure with greater power than familywise error rate (FWER) control, at a cost of increasing the likelihood of obtaining type I errors. (Bonferroni correction controls FWER; FWER = P(the number of type I errors ≥ 1)).

The q-value is defined to be the FDR analogue of the p-value. The q-value of an individual hypothesis test is the minimum FDR at which the test may be called significant.

To estimate the q-value and FDR, we need following notations:

- m is the number of tests
- m_0 is the number of true null hypotheses
- $m - m_0$ is the number of false null hypotheses
- U is the number of true negatives (the tests not rejected when the null is true)
- V is the number of false positives (the tests rejected when the null is true)
- T is the number of false negatives (the tests not rejected when the null is not true)
- S is the number of true positives (the tests rejected when the null is not true)
- $R = V + S$ is the total number of tests rejected.
- $H_1 ... H_m$ the null hypotheses being tested (so there are m tests)
- In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S, T, U, and V are all unobservable random variables.

The false discovery rate (FDR) is given by $E(\frac{V}{V+S}) = E(\frac{V}{R})$ and one wants to keep this value below a threshold α:

The *Simes* procedure ensures that its expected value $E(\frac{V}{R})$ is less than a given α (Benjamini and Hochberg 1995). This procedure is only valid when the m tests are independent. Let $H_1 ... H_m$ be the null hypotheses and $P_1 ... P_m$ their corresponding p-values. Order these values in increasing order and denote them by $P_{(1)} ... P_{(m)}$. For a given α, find the largest k such that $P_{(k)} \le \frac{k}{m}\alpha$. Then reject all $H_{(i)}$ for $i = 1,...,k$. The q-value of $H_{(k)}$ can be estimated by $mP_{(k)}/k$. It is also the FDR if we reject all the null hypotheses with p-values $\le P_{(k)}$.

The positive discover rate (pFDR) (Storey 2002) is given by

$$E(\frac{V}{R} \mid R > 0) = P(\text{null is true} \mid \text{observed statistic in the rejection region})$$

$$= \frac{P(\text{observed statistic in the rejection region} \mid \text{null is true})P(\text{null is true})}{P(\text{observed statistic in the rejection region})}$$

If we reject the null when the p-value $\le P_{(k)}$,

- P(observed statistic in the rejection region$\mid H_{(k)}$ is true) = $P_{(k)}$,
- P(null is true) = $\frac{W(\lambda)}{(1-\lambda)m}$, where $W(\lambda) =$ the number of hypotheses with pvalues $\ge \lambda$. This equation is based on the consideration that the large p-values are most likely to come from the null, uniformly distributed p-values. So usually a relatively large λ will be used.
- P(observed statistic in the rejection region) = 1- $W(P_{(k)})/m$

The q-value of $H_{(k)}$ controlling the pFDR then can be estimated by

$$\frac{P_{(k)}W(\lambda)}{(1-\lambda)\left(m-W(P_{(k)})\right)} .$$

It is also the estimated pFDR if we reject all the null hypotheses with p-values $\leq P_{(k)}$.

Maximum Likelihood Estimation

Given a family of probability distributions parameterized by θ (which could be vector-valued), associated with either a known probability density function (continuous distribution) or a known probability mass function (discrete distribution), denoted as f_θ, we may draw a sample x_1, x_2, ..., x_n of n values from this distribution and then using f_θ we may compute the probability density associated with our observed data:

$$f_\theta(x_1,...,x_n \mid \theta) .$$

As a function of θ with x_1, ..., x_n fixed, this is the likelihood function

$$L(\theta) = f_\theta(x_1,...,x_n \mid \theta) .$$

The method of maximum likelihood estimates θ by finding the value of θ that maximizes $L(\theta)$. This is the **maximum likelihood estimator** (**MLE**) of θ.

This contrasts with seeking an unbiased estimator of θ, which may not necessarily yield the MLE but which will yield a value that (on average) will neither tend to over-estimate nor under-estimate the true value of θ.

The maximum likelihood estimator may not be unique, or indeed may not even exist.