Multiple testing in the world of business – when and how?

Nicolle Clements

Department of Decision and System Sciences,
Saint Joseph's University,
5600 City Avenue, Philadelphia, PA, USA
Email: nclement@sju.edu

Abstract: Multiplicity of data and compounding errors is often overlooked in data analysis for applied business scenarios. Statistical theory around multiple testing provides a framework for describing appropriate error rates and offers methods to control them in order to protect against wrong conclusions. However, these multiple testing procedures are often misunderstood and underutilised in applied business problems. In this article, existing multiple testing methodologies are reviewed and summarised. Specific numeric examples are shown to illustrate the techniques and demonstrate the statistical power of each. Finally, three cases are given of business-related situations when multiple testing can be overlooked in data analysis.

Keywords: multiple testing; familywise error rate; FWER; false discovery rate; FDR; type I error; business analytics.

Reference to this paper should be made as follows: Clements, N. (2019) 'Multiple testing in the world of business – when and how?', *Int. J. Business and Data Analytics*, Vol. 1, No. 1, pp.16–29.

Biographical notes: Nicolle Clements is a PhD Statistician and an Assistant Professor in the Department of Decision System Sciences at Saint Joseph's University, where she also serves as the Academic Coordinator of MSBIA Program. She holds a Doctorate in Statistics from the Temple University's Fox School of Business, Master of Science in Statistics from Virginia Polytechnic Institute and State University (Virginia Tech) and Bachelor of Science in Mathematics from Millersville University. Her PhD research was in the area of high dimensional multiple testing procedures and she currently conducts applied work in spatial and environmental applications of analytics, multiplicative time-series modelling and statistical analysis of substance abuse treatments. Much of her research is focused on adapting standard statistical models to be used in non-traditional applications. In addition to her research, she frequently teaches courses at the undergraduate, graduate and executive level on topics such as: statistics, data mining and R statistical programming.

1 Introduction

Testing a single hypothesis typically involves making a choice between two complementary statements about a population parameter, referred to as the null and alternative hypothesis. Multiple testing refers to simultaneous testing of several hypotheses within a data analysis. This scenario is rather common and often overlooked,

in many business applications, which brings the research question to this article. Some examples include:

1 Fitting a multiple linear regression model to identify which coefficients are statistically different from zero.

2 Screening for changes across multiple locations in a geographic region.

3 Evaluating an experimental design with respect to multiple outcomes and trying to decide which outcomes in the experiment yield significant effects.

The decision to accept or reject the null hypothesis is based on the information gathered from a sample, which is a subset of the population. Since information from the entire population is often infeasible, it is possible that the sample data can lead to erroneous decisions regarding the null and alternative. Specifically, type I and type II errors are concerning and unintended consequences. A type I error occurs when a true null hypothesis is incorrectly rejected (also known as a 'false positive'), while a type II error occurs by incorrectly failing to reject a false null hypothesis (also known as a 'false negative'). Often, the goal of data analysts is to develop testing procedures that minimise the probability of making these errors without sacrificing the power to detect false nulls. However, it is impossible to minimise these errors simultaneously, since reducing one type of error will inflate the other error. Controlling type I error has traditionally been the focus of statistical testing methods (Bretz et al., 2010).

Recently, instead of single hypothesis testing, researchers are being confronted with testing hypotheses together, where n can be very large. One cannot employ the same testing procedures used for a single hypothesis because the probability of making an error compounds as n gets large. This is called the multiplicity effect. Multiple testing procedures are necessarily different than single testing procedures because they take into account the number of hypotheses being tested to ensure simultaneous control of error. When testing several hypotheses together, called a family of hypotheses, an appropriate compound error measure must first be defined. Then, a procedure is developed that allows one to control this error rate at a desired level, called α.

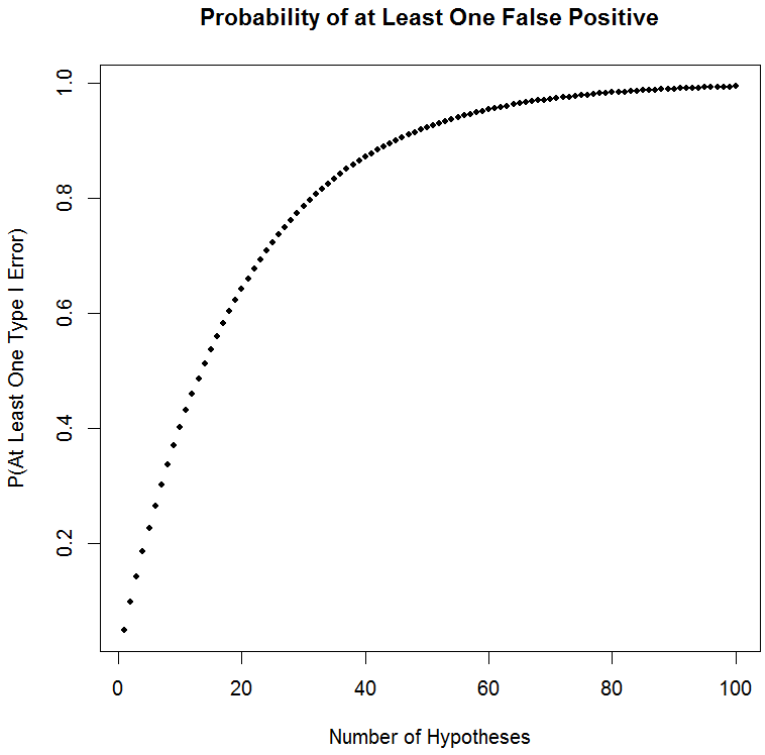
If the multiplicity of tests is not taken into account, then the probability that some of the true null hypotheses are rejected by pure chance may be undeservedly large. For illustration, consider the case of $N = 100$ hypotheses being simultaneously tested, all of them being true, with the size and level of each test exactly equal to α. For $\alpha = 0.05$, five true hypotheses are expected to be rejected erroneously. Further, if all tests are mutually independent, then the probability that at least one true null hypothesis will be falsely rejected is $1 - 0.95^{100} = 0.994$!. Figure 1 illustrates the increasing likelihood of making a type I error as the number of hypotheses, n increases.

Of course, this problem does not exist if there is a priori focus is on a particular hypothesis. In this case, the decision can still be based on the corresponding marginal p-value. The problem of multiplicity only surfaces if the list of p-values is searched for significant results, a posteriori. Unfortunately, the latter is much more common. In such a case, multiple testing procedures should be used to guard against committing one or more type I errors.

The objective of this article is to give an overview of multiple testing concepts, define several measures of error and corresponding testing procedures and give examples of situations in which multiple testing problems in business research may occur. The current

business literature does not provide a comprehensive overview in one document, so this article will fill that gap. The remainder of the paper is structured as follows. In Section 2, some notation is defined that will be used throughout the paper. Afterwards, Section 3 describes two types of type I error controlling rates are explained along with multiple corresponding error-controlling procedures. Then, a guide as when to worry about multiple testing adjustments is provided in Section 4. Finally, examples are given where multiple testing should be used in business scenarios in Section 5 and wrapped up with some concluding remarks in Section 6.

Figure 1 Assuming independence, the probability of making at least one type I error increases significantly as the number of hypotheses tested gets large



2 Notations

Suppose X is data generated from an unknown probability distribution P. Consider the problem of simultaneously testing n hypotheses from this data. The set of null hypotheses are written as H_1, H_2, …, H_n and the corresponding alternative hypotheses are $\bar{H}_1, \bar{H}_2, \ldots, \bar{H}_n$. Also, the analogous test statistics, critical values and p-values are T_1, T_2, …, T_n, c_1, c_2, …, c_n and p_1, p_2, …, p_n, respectively. Many multiple testing procedures use ordered p-values, which are denoted by subscripts in parentheses: $p_{(1)}$, $p_{(2)}$, …, $p_{(n)}$. The i^{th} ordered p-value, $p_{(i)}$, corresponds to the null hypothesis $H_{(i)}$, which is not necessarily the same as H_i.

Table 1 gives the various outcomes when testing hypotheses simultaneously, where H_{i0}: $\theta_i = \theta_{i0}$ is the null hypothesis and H_{i1}: $\theta_i \neq \theta_{i0}$ is the two-sided alternative, for $i = 1, 2, \ldots, n$. Of these quantities in Table 1, only n, A and r are known after applying a particular multiple testing procedure. The number of type I errors, V and the number of type II errors, T, are unknown but desirably small. Most multiple testing procedures focus on controlling V in some capacity.

Table 1 Multiple testing outcomes from testing n hypotheses

		Decision		*Total*
		Fail to reject null	*Reject null*	
Truth	*Null true* ($\theta = \theta_0$)	U (correct decisions)	V (type I errors)	n_0
	Alternative true ($\theta \neq \theta_0$)	T (type II errors)	S (correct decisions)	n_1
	Total	A	R	n

Multiple testing procedures can be categorised into single-step tests or multi-step tests, also called stepwise procedures. Single step testing procedures define one critical value for which all p-values are then compared. In other words, suppose the common critical value is c, then a single-step procedure rejects all hypotheses H_i if the corresponding $p_i \leq c$. Multi-step or stepwise, procedures compare each p-value to a different threshold. Consider the set of ordered p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ and the corresponding set of critical values $c_1, c_2, \ldots, c_n$. A step-down procedure rejects $H_{(i)}$ for which $p_i \leq p_k$, where $k = \max\{i: p_j \leq c_j, \forall j < i\}$. In other words, if the largest p-value $p_{(n)} < c_{(n)}$, reject $H_{(n)}$ and continue for $i = n - 1, n - 2, \ldots, 3, 2, 1$ comparing $p_{(i)}$ with c_i, rejecting $H_{(i)}$ if $p_{(i)} < c_i$. Continue until, for the first time, $p_{(i)} > c_i$. If $p_{(n)} > c_n$, reject no hypotheses. If $p_{(i)} < c_i$, for all values of i, reject all the hypotheses. A step-up procedure rejects $H_{(i)}$ corresponding to $p_{(i)} \leq p_{(k^*)}$ where $k^* = \max\{i: p_{(i)} \leq c_i\}$.

3 Error rates and corresponding procedures

3.1 Familywise error rate

One of the most commonly used measures of overall type I error is called the familywise error rate (FWER). The FWER is the probability of making one or more type I errors. In other words, out of n simultaneously tested hypotheses, where V is the number of type I errors made out of n decisions (recall: V is an unknown quantity), then $FWER = Prob\{V > 0\}$. In the case of multiple hypothesis testing, the FWER should be controlled at a desired overall level, called α.

The Bonferroni procedure is one of the most frequently used adjustments utilised by researchers dealing with multiplicity because it is easy to use and it can be used in any dependence structure. However, it can be extremely conservative by controlling type I error at a value much less than the specified level. The Bonferroni correction is a single step procedure, meaning all hypotheses are compared to the same threshold for every hypothesis tested. In terms of hypothesis testing, the i^{th} null hypothesis is rejected if the corresponding p-value is less than α/n where α is the desired overall significance level and n is the number of hypotheses being simultaneously tested (Holland and Copenhaver,

1987; Hochberg and Tamhane, 1987). For an example of the critical values in this procedure, see column 3 in Table 2. Using the Bonferroni procedure will control the FWER under any data's dependence structure at level $n_0\alpha/n$.

Table 2 An illustration

Rank	*p-value*	*Critical values/reject H_{i0}?*			
		Bonferroni	*Šidák*	*Holm*	*Hochberg*
1	0.0024	0.005/yes	0.005116/yes	0.005/yes	0.005/yes
2	0.0057	0.005/no	0.005116/no	0.00556/no	0.005556/yes
3	0.0061	0.005/no	0.005116/no	0.00625/no	0.00625/yes
4	0.0391	0.005/no	0.005116/no	0.007143/no	0.007143/no
5	0.0488	0.005/no	0.005116/no	0.008333/no	0.008333/no
6	0.0630	0.005/no	0.005116/no	0.01/no	0.01/no
7	0.1294	0.005/no	0.005116/no	0.0125/no	0.0125/no
8	0.3613	0.005/no	0.005116/no	0.016667/no	0.016667/no
9	0.4689	0.005/no	0.005116/no	0.025/no	0.025/no
10	0.6725	0.005/no	0.005116/no	0.05/no	0.05/no

Note: Using of four FWER procedures comparing their critical values to simulated *p*-values and the decision made.

Šidák's (1967) procedure is another common multiple testing adjustment to control the FWER, but only when the tests are independent or positively dependent. The notion of positive dependence is satisfied by a number of multivariate distributions, including multivariate normal test statistics with positive correlations, absolute values of studentised independent normals and multivariate t and F (Benjamini and Yekutieli, 2001). Šidák's (1967) procedure is more powerful than Bonferroni's but the gain is small, so often the procedure is overlooked. Like the Bonferroni correction, Šidák's (1967) correction is also a single step procedure. Šidák's (1967) correction factor says to reject the i^{th} null hypothesis if the corresponding p-value is less than $1 - (1 - \alpha)^{1/n}$. It is clear to see that the Šidák (1967) correction gives a stronger bound than the Bonferroni correction because $\alpha/n \leq 1-(1-\alpha)^{1/n}$, for any $n \geq 1$. For an example of the implementation of this procedure, see column 4 in Table 2.

However, the Šidák (1967) correction requires the additional condition of independence or positive dependence among the tests or interval estimates, whereas the Bonferroni correction has no assumption on the dependence structure. Previously, because the Šidák (1967) correction required the user to calculate fractional powers (i.e., the n^{th} root), the computationally simpler Bonferroni correction was often the preferred adjustment factor. Now, since computing fractional powers is trivial, preference of the Bonferroni method is due in part to tradition or unfamiliarity with the Šidák (1967) method. Unfortunately, Šidák's (1967) method offers minimal gain, in terms of statistical power, for conventional significance levels (α between .01 and .10).

The Holm's (1979) procedure is a step-down procedure which controls the FWER under any dependence structure. The critical values used in hypothesis testing are

$c_i = \frac{\alpha}{n-1+1}$ for i = 1, 2, …, n. This procedure is implemented by computing and ordering the p-values $p_{(1)} \le p_{(2)} \le \ldots \le p_{(n)}$. For i = 1, 2, …, n, if $p_{(i)} \ge \frac{\alpha}{n-1+1}$, then accept $H_{(i)}$, $H_{(i+1)}$, …, $H_{(n)}$; otherwise, reject $H_{(i)}$ and increment to $i + 1$, while $i < n$. For an example of the critical values used in this procedure, see column 5 in Table 2.

Hochberg's (1988) procedure is a step-up method that is based on the same critical values as the Holm's (1979) procedure, but is generally more powerful. Hochberg's (1988) method maintains control of FWER under independence or positive dependence of the p-values, proved by Hochberg (1988), Sarkar (1998) and Sarkar and Chang (1997). This testing procedure rejects all $p_{(i)} \le p_{(k)}$, where $k = \max\left\{i : p_{(i)} < \frac{\alpha}{n-i+1}\right\}$. For an example of how to implement this procedure, see column 6 in Table 2. Notice the gain of two additional rejections using Hochberg's (1988) step-up method compared to Holm's (1979) step-down method using the exact same critical values.

Given in Table 2 are ten independently simulated p-values, which were sorted and ranked. Based on these p-values, four FWER procedures are implemented: Bonferroni, Šidák (1967), Holm (1979) and Hochberg (1988). Table 2 lists the corresponding critical values (c_i) and the binary decision to accept or reject the null hypothesis, at significance level $\alpha = 0.05$.

3.2 False discovery rate

The false discovery rate (FDR), proposed by Benjamini and Hochberg (1995) is the second most common measure of type I errors. The FDR is the expected proportion of type I errors among all the rejected null hypotheses. If there are no rejected hypotheses, the FDR is defined to be zero. In terms of Table 1, $FDR = E\left[V/\max(R, 1)\right]$. Comparatively, the FDR is less conservative than the FWER, meaning FWER control ensures FDR control. However, a multiple testing procedure with FDR control will not necessarily maintain control of the FWER. The FDR is a widely accepted and utilised notion of type I errors in large-scale multiple testing investigations (Nichols, 2007).

The Benjamini and Hochberg (1995) method, known as the BH method for short, was proposed at the same time that they introduced the notion of the FDR error metric. This is a step-up method defined by using the ordered p-values for all n hypotheses: $p_{(1)}$, $p_{(2)}$, …, $p_{(n)}$. The method works by letting $k = \max\left\{i : p_{(i)} < i\alpha/n\right\}$ and rejecting all hypotheses whose p-values are less than or equal to $p_{(k)}$. This procedure will control the FDR at level $n_0\alpha/n$, under the assumption of independence of the null p-values. Benjamini and Yekutieli (2001) later showed that the BH method would control the FDR if the p-values have positive dependence (see also Sarkar, 2002). Since $n_0\alpha/n < \alpha$, the FDR is conservatively controlled. However, n_0, the number of true null hypotheses, is generally unknown.

The Benjamini and Yekutieli (2001) method, abbreviated as the BY method, was developed to control the FDR under any arbitrary dependence assumption among the test statistics. The BY method adjusts the BH method's k with $k^* = \max\left\{i : p_{(i)} < i\alpha / c_n n\right\}$ where $c_n = \sum_{i=1}^{n} i^{-1} \approx \ln(n)$ when n is large. Then, reject all $p_{(i)} \le p_{(k^*)}$. Although this method gives the user freedom of use under any dependence structure, the downfall is that it can be quite conservative in many situations. The work of Sarkar (2008) extended this result by proposing a different stepwise method to control FDR under arbitrary dependence using critical values $c_i = \frac{i(i+1)\alpha}{2n^2}$. Blanchard and Roquain (2009) also proposed a step-up test to be used in arbitrary dependence by using the critical values $c_i = \frac{i(i+1)(2i+1)}{3n(n+1)}$.

In Table 3, a toy example is given with sample p-values and the critical values for Benjamini and Hochberg's (1995) procedure and Benjamini and Yekutieli's (2001) procedure. In this example, $\alpha = 0.05$ and the p-values are assumed to be independent so that both procedures ensure control of the FDR. Notice in the BH method, the largest p-value, $p_{(n)}$, is always compared to the overall level of significance, α. Also, take note of how much smaller the critical values are for the BY procedure compared to BH, due to the adjustment of $c_n = 2.9289$ in the denominator.

Table 3 An illustration of two FDR procedures comparing their critical values to p-values and the decision made

Rank	*p-value*	*Critical values/reject H_{i0}?*	
		Benjamini-Hochberg	*Benjamini-Yekutieli*
1	0.0004	0.005/yes	0.0017/yes
2	0.0038	0.01/yes	0.0034/yes
3	0.0047	0.015/yes	0.0051/yes
4	0.0191	0.02/yes	0.0068/no
5	0.0218	0.025/yes	0.0085/no
6	0.0430	0.03/no	0.0102/no
7	0.0691	0.035/no	0.0119/no
8	0.1849	0.04/no	0.0137/no
9	0.2004	0.045/no	0.0154/no
10	0.3602	0.05/no	0.0171/no

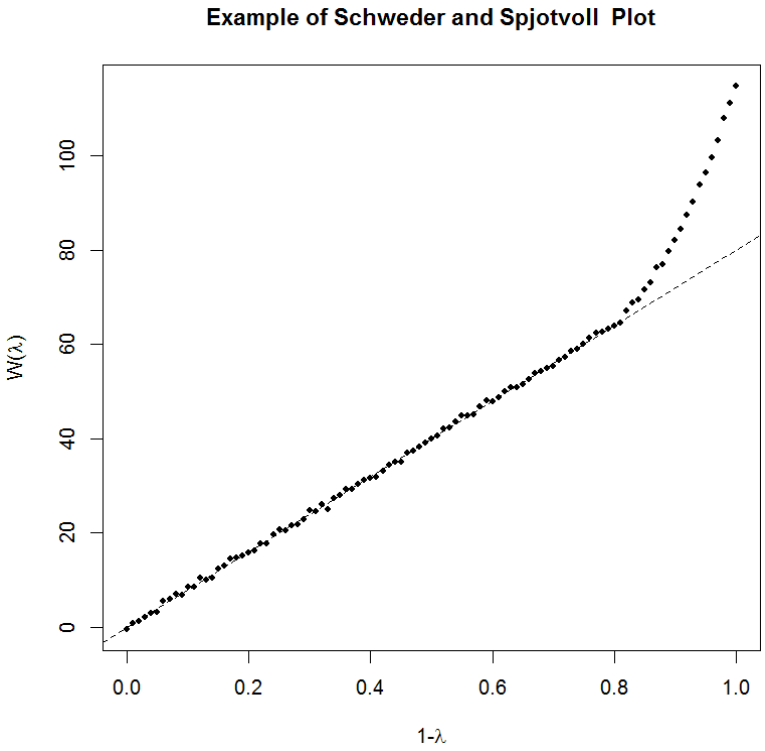
3.3 Adaptive methods

Unfortunately, many of the procedures described in the previous two sections will control the respective error rate at a level of $n_0\alpha / n$, which is less than the desired level α. The number of true null hypotheses, n_0, is typically unknown and conservatively influences the control of the FWER and FDR. However for a given procedure, a suitable estimate of n_0 obtained from the data can potentially improve the original methods. This has been the

rationale behind developing adaptive procedures through an estimation of n_0. A few popular adaptive methods are described next.

Schweder and Spjotvoll (1982) were the first to propose a simple informal graphical procedure to estimate the number of true null hypotheses, n_0. If $W(\lambda)$ denotes the number of p-values greater than λ, then $E[W(\lambda)] = n_0(1 - \lambda)$ since the p-value should be large for a true null hypothesis. Thus, a plot of $W(\lambda)$ against $(1 - \lambda)$ should indicate a straight line with slope $n0$ for large λ. The plot will tend to show a linear behaviour for larger p-values, the slope of which gives an estimate of the number of true nulls. The points deviating from the straight line will most likely correspond to false nulls. So, by looking at the plot, one can get a visual estimate of the number of true null hypotheses. In Figure 2, an illustration of a Schweder and Spjotvoll (1982) plot is given.

Figure 2 Illustration of a Schweder and Spjotvoll (1982) plot used to estimate the number of true nulls by the slope of the line



Note: In this example, the slope is approximately 80, so $n_0 \approx 80$ could be used in adaptive methods.

The adaptive BH method of Storey et al. (2004) is among the most often used of the adaptive methods. Their estimate is $\hat{n}_0 = \min\left\{\frac{W_n(\lambda)+1}{1-\lambda}, n\right\}$ for any fixed tuning parameter $\lambda \in (0, 1)$ and $W_n(\lambda) = \sum_{i=1}^{n} I\left(p_i \geq \lambda\right)$ where $I(p_i \geq \lambda)$ is the indicator function which takes a value of 1 when $p_i \geq \lambda$ and a value of 0 when $pi < \lambda$. So, $W_n(\lambda)$ is a count of

the number of p-values greater than or equal to λ. In a sense, it is a count of the number of 'large' p-values that likely come from the null hypothesis.

Other estimators of n_0 exist, such as the one proposed in Benjamini et al. (2006). But regardless of the estimator used, adaptive methods work by using the estimate, $\hat{n}_0$, in place of n_0 in the critical value calculations. For example, the adaptive Bonferroni method rejects p-values that are less than $\alpha/\hat{n}_0$ instead of α/n, which yields at least as many, if not more, rejected null hypotheses (i.e., higher power).

As another example, the adaptive BH method that corresponds to an estimate of n_0 rejects all hypotheses whose p-values are less than or equal to $p_{(\hat{k})}$ where $\hat{k} = \max\left\{i : p_{(i)} < \frac{i\alpha}{\hat{n}_0}\right\}$. This adaptive BH method based on Storey, Taylor and Siegmund will control the FDR under independence of the p-values (Benjamini et al., 2006; Storey et al., 2004), as well as under certain form of asymptotic weak dependence.

4 When should analysts worry about multiplicity?

The answer is, it depends! If the goal of data analysis is to control the type I error rate for individual tests, an adjustment for multiplicity is unnecessary. However, if the goal is to ensure simultaneous control over the family of hypotheses, thus control the FWER or the FDR, a multiplicity adjustment is essential. Regrettably, no single answer exists to when it is appropriate to control which error rate. Different data analysts may have diverse but still rational opinions.

In addition to choosing which error rate should be under control, the analyst must next evaluate the dependency structure of the data. If the hypotheses can be assumed independent, any of the methods mentioned above will work for the respective error rate (e.g., Holm or Hochberg for FWER or Benjamini-Hochberg for FDR). However, if the hypotheses have a dependency between them, the analyst must use one of the conservative methods to control the type I errors (Bonferroni for FWER or Benjamini-Yekutieli for FDR).

In essence, the data analyst needs to make sequential decisions about the objective of the analysis. First, the analysis must determine if there is a need to make conclusions about multiple hypotheses, parameters or objectives. If yes, the analyst must then decide the level of control he/she wants to have over type I errors (namely, control the FWER or FDR). Finally, the analyst needs to estimate the dependence structure (independent, positively dependent or other) of the data before selecting the multiple testing procedures to use for adjustments in the analysis.

5 Multiple testing in practice

The need to simultaneous test of several hypotheses can easily be overlooked in many business applications. In this section, examples are given where multiple testing should be used in business scenarios.

5.1 Example 1: multiple testing in regression models

Suppose an econometrician is interested in trying to find relevant predictors of demand for a service. There are two outcome/dependent variables describing the demand (indicator for use of service yes/no and the frequency of occasions). Suppose ten predictor/independent variables could theoretically explain the demand, such as age, gender, education, income, price, native language, socio-economic status, etc. Running two separate multiple regressions will yield 20 coefficients estimations and their *p*-values. With enough independent variables in the regression models, the econometrician would sooner or later find at least one variable with a statistically significant correlation between the dependent and independent variables. These significant correlations may or may not be spurious! If the regression model is the only one under consideration and the econometrician is not interested in performing model selection, then multiple testing adjustments should be applied when drawing conclusions about the coefficients (Mumdrom et al., 2006).

Expanding on the use of multiple testing in regression models, in some situations a joint test of a composite hypothesis regarding regression coefficients should be used to draw conclusions. Dmitrienko et al. (2009) reviewed multiple comparison procedures used in pharmaceutical statistical regression models by focusing on different drug development applications. Farcomeni (2008) and Veazie (2006) provided more examples from the literature that are helpful in understanding when to and when not to, combine hypotheses. Some of their examples are summarised in the next two paragraphs.

Veazie (2006) pointed to two articles in the *Journal of Health Economics* that both used regression models on a dependent variable by fitting a second order polynomial. Also known as quadratic regression, polynomial models are a common practice to capture nonlinear relationships. In these articles, the null hypothesis for each coefficient of the polynomial was rejected according to its individual *p*-value. It was concluded that the explanatory variable had a quadratic relationship with the response variable. Veazie (2006) suggests that the authors rejected the joint hypothesis that both coefficients were simultaneously zero. However, this is different from a researcher testing second-order nonlinearity (as opposed to testing the parabolic shape). In this case, an individual test of the coefficient on the second-order term (i.e., the coefficient on the squared variable) is appropriate because the value of the first order term is meaningless in the judgement of nonlinearity.

As a different situation in regression models, Veazie (2006) points to a recent article in *Medical Care*. The article's methodology categorised a count variable into three size-groups and used a set of dummy variables to represent the two largest groups and the smallest group was referred to as the baseline category. The results explained that based on the individual significance of the two dummy variables, the hypothesis that both coefficients were zero was rejected. The article concluded that the dependent variables were related to having larger counts based on the underlying concept. In this conclusion, they collapsed two categories into a single statement about being larger on the underlying variable. Yet, if the authors meant that both categories are larger than the reference group, then it is a test of both coefficients being simultaneously zero that is relevant. A joint test is warranted here to avoid type I errors.

5.2 Example 2: multiple testing in spatial screening for significant locations in a geographic region

Consider a data analysis project with an objective of identifying trends present in time series data across multiple geographic locations. This type of time series data is common in many applications, such as a business's sales data at various retail stores in a region, US census data in each state collected every ten years or even hourly weather observation taken from different weather stations.

As a specific example, Clements et al. (2014) studied vegetation monitoring in East Africa based on the normalised difference vegetation index (NDVI) series from satellite remote sensing data that was collected between 1982 and 2006 over 8 kilometre grid points. The NDVI is a simple graphical indicator that can be used to analyse remote sensing measurements to assess whether region being observed contains live green vegetation or not. Trend changes in vegetation can give valuable information to decision makers about effective land use and development which is fundamental in planning agricultural endeavours. In particular, decision makers want to gain knowledge of current vegetation trends and use them to make accurate predictions. Vegetation trends are also closely related to sustainability issues, such as management of conservation areas and wildlife habitats, precipitation and drought monitoring, improving land usage for livestock and finding optimum agriculture seeding and harvest dates for crops. For this reason, there are many decision makers, agencies and organisations that are invested in the study of land use and land cover trends, linking them to climate change and the socioeconomic consequences of these changes.

To test for significant trend in each location, Clements et al. (2014) apply the monotonic trend test proposed by Brillinger (1989) for a time series consisting of a signal and stationary autocorrelated errors. The NDVI annual averages were used as the observed time series. This test examines the null hypothesis that the series has a signal, that is, constant in time against the alternative hypothesis that the signal is monotonically increasing or decreasing in time. Thus, p-values generated for each site (8 km × 8 km grid of land), provides evidence of vegetation change occurring over the years – the smaller the p-value, the higher is the evidence of a significant vegetation change. For each site, a decision must be made regarding the significance of vegetation change that might have occurred over the years at that site and, if vegetation change is found significant, determine the direction in which this change has taken place. This must be done simultaneously for all sites (≈ 50,000) in the East African region in a multiple testing framework designed to ensure a control over a meaningful combined measure of statistical type I errors. It is important to provide an upper bound on type I errors (false detections), since there is large risk associated with falsely declaring an area to have significant vegetation changes.

5.3 Example 3: multiple testing in design of experiments

The rise of the design of experiments within business organisations or what is sometimes referred to as field experiments, has the potential to transform organisational decision-making. Carefully designed experiments also can provide new insight into many areas of business, such as product design, human resources or public policy. Companies that invest in randomised design of experiments have a lot to learn from their data. Yet, if the experiment is run incorrectly or not analyses properly, the organisation cannot receive

the advantages of this scientific process. Goeman and Solari (2014) explain how this applies to multiplicity in genomics. The article describes the exploratory nature of genomics experiments, in which researchers look at the selection of genes before or after testing and at the role of validation experiments. In this experimental setting, multiplicity must be accounted for in gene selection. The following paragraphs describe a more explicit example of the misuse of multiple testing in a design of experiment.

A common mistake in designing an experiment is the lack of definition of which hypotheses are of interest for one experimental study. For example, consider an experiment in which three new treatments (*X*, *Y*, *Z*) are compared with a standard treatment, called the control (*C*). One could consider all six pairwise contrasts (*X* vs. *C*, *X* vs. *Y*, *X* vs. *Z*, *Y* vs. *C*, *Y* vs. *Z* and *Z* vs. *C*) as one experiment or family of comparisons. However, often the main goal, or primary hypothesis, is to compare the new treatments with the control (*X* vs. *C*, *Y* vs. *C* and *Z* vs. *C*). Secondary to comparing new treatments with the control is to compare the new treatments with each other (*X* vs. *Y*, *X* vs. *Z* and *Y* vs. *Z*). These three would constitute as secondary analysis and is not the main goal of the experiment. In this situation, it may be appropriate to perform separate multiplicity adjustments in each tier of the experiment.

In general, it is reasonable to suggest that the FWER should be under control when the results of a well-defined family of multiple tests should be summarised in one conclusion for the whole experiment, using methods such as Bonferroni, Šidák (1967), Holm (1979) and Hochberg (1988). For example, if each new treatment is significantly different from the control, the conclusion that all three treatments differ from the standard treatment should be based upon an adequate control of the FWER and not the marginal *p*-values. Otherwise, the type I error of the final conclusion is not necessarily controlled, which means that the aim of the design of experiment is not reached.

6 Conclusions

It is not common practice among applied business researchers to use multiple testing procedures in analysis applications, such as variable selection in regression analysis. Rather, it is much more common to see each specific testing conducted at the nominal level ($\alpha = 0.05$). However, other fields, particularly the medical field, are beginning to put multiple testing practices to use (Farcomeni, 2008; Streiner and Norman, 2011; Goeman and Solari, 2014). In this article, it is explained that using unadjusted testing procedures where multiplicity exists, the associated overall type I error rate may be inflated. Specific numeric examples are shown to illustrate the techniques and demonstrate the statistical power of each.

For example, in regression analysis, the type I error rate could be magnified by as much as two to six times the nominal level, depending upon the number of predictors in the model relative to the number of predictors that have a non-zero relationship with the response. Consequently, one or more variables may be identified as ‘significant’ predictors of the response that are not actually needed in the model. In other words, the amount of variance in the response explained by the model’s variables could be negligible (Mumdrom et al., 2006).

There is an argument for an alternative approach to significance testing for analysis of data that should be mentioned. Bayesian statistics differ from the testing procedures discussed in this article because they focus on minimising what is called the 'Bayes risk under additive loss', rather than controlling type I error rates. In principal, control of type I error is not required to make valid inferences from a Bayesian perspective. Some conceptual and practical difficulties involved with the control of type I error can be avoided by using Bayesian methods, especially in the case of multiplicity. However, this article concentrated on classical statistical methodology based upon significance testing. The article assumes that significance tests are going to be used for data analysis. Under this assumption, this article summarised some available procedures to adjust for multiple testing. Since Bayes methods do not provide adjustments of *p*-values, as they do not give *p*-values at all, they are not discussed in this article.

In summary, to ensure valid statistical inference in the case of multiplicity, methods to adjust for multiple testing are necessary. Adjustments should be used in all confirmatory studies where a clearly defined family of tests exists and one final conclusion and decision will be drawn. In such cases, the maximum type I error rate under any family of null hypotheses should be under control at the desired level α. The simple, but commonly used, Bonferroni procedure is often not appropriate due to low power. However, there are a number of more powerful procedures available for in various multiplicity situations, as summarised and explained in this article. These methods deserve wider knowledge and application in business research than what is currently provided in the literature.

References

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society*, Vol. 57, No. 1, pp.289–300.

Benjamini, Y. and Yekutieli, D. (2001) 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics*, Vol. 29, No. 4, pp.1165–1188.

Benjamini, Y., Krieger, K. and Yekutieli, D. (2006) 'Adaptive linear step-up procedures that control the false discovery rate', *Biometrika*, Vol. 93, No. 3, pp.491–507.

Blanchard, G. and Roquain, E. (2009) 'Adaptive FDR control under independence and dependence', *Journal of Machine Learning Research*, Vol. 10, No. 1, pp.2837–2871.

Bretz, F., Hothorn, T. and Westfall, P. (2010) *Multiple Comparisons Using R*, Chapman and Hall/CRC, New York.

Brillinger, D.R. (1989) 'Consistent detection of a monotonic trend superposed on a stationary time series', *Biometrika*, Vol. 76, No. 1, pp.23–30, MR0991419.

Clements, N., Sarkar, S., Zhao, Z. and Kim, D. (2014) 'Applying multiple testing procedures detect changes in East African vegetation', *Annals of Applied Statistics*, Vol. 8, No. 1, pp.286–308.

Dmitrienko, A., Tamhane, A.C. and Bretz, F. (2009) *Multiple Testing Problems in Pharmaceutical Statistics*, Taylor and Francis, Boca Raton.

Farcomeni, A. (2008) 'A review of modern multiple testing, with particular attention to the false discovery proportion', *Statistical Methods in Medical Research*, Vol. 17, No. 4, pp.347–388

Goeman, J. and Solari, A. (2014) 'Multiple hypothesis testing in genomics', *Statistics in Medicine*. Vol. 33, No. 11, pp.1946–1978.

Hochberg, Y. (1988) 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, Vol. 75, No. 4, pp.800–802.

Hochberg, Y. and Tamhane, A. (1987) *Multiple Comparison Procedures*, Wiley, New York, New York.

Holland, B.S. and Copenhaver, M.D. (1987) 'An improved sequentially rejective Bonferroni test procedure', *Biometrics*, Vol. 43, No. 2, pp.417–423.

Holm, S. (1979) 'A simple sequential rejective multiple test procedure', *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp.65–70.

Mumdrom, D.J., Perrett, J.J., Schaffer, J., Piccone, A. and Rooseboom, M. (2006) 'Bonnferroni adjustments in tests for regression coefficients', *Multiple Linear Regression Viewpoints*, Vol. 32, No. 1, pp.1–6.

Nichols, T. (2007) *False Discovery Rate*, Ample Translations [online] http://www-personal.umich.edu/~nichols/FDR/ (accessed 10 July 2014).

Sarkar, S.K. (1998) 'Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture', *Annals of Statistics*, Vol. 26, No. 2, pp.494–504.

Sarkar, S.K. (2002) 'Some results on false discovery rate in stepwise multiple testing procedures', *Annals of Statistics*, Vol. 30, No. 1, pp.239–257.

Sarkar, S.K. (2008) 'Generalizing Simes' test and Hochberg's step-up procedure', *The Annals of Statistics*, Vol. 36, No. 1, pp.337–363.

Sarkar, S.K. and Chang, C.K. (1997) 'The Simes method for multiple hypothesis testing with positively dependent test statistics', *Journal of the American Statistical Association*, Vol. 92, No. 440, pp.1601–1608.

Schweder, T. and Spjotvoll, E. (1982) 'Plots of p-values to evaluate many tests simultaneously', *Biometrika*, Vol. 69, No. 3, pp.493–502.

Šidák, Z.K. (1967) 'Rectangular confidence regions for the means of multivariate normal distributions', *Journal of the American Statistical Association*, Vol. 62, No. 318, pp.626–633.

Storey, J., Taylor, J. and Siegmund, D. (2004) 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach', *Journal of the Royal Statistical Society, B*, Vol. 66, No. 1, pp.187–205.

Streiner, D. and Norman, G. (2011) 'Correction for multiple testing: is there a resolution?', *Chest Journal*, Vol. 140, No. 1, pp.16–18.

Veazie, P.J. (2006) 'When to combine hypotheses and adjust for multiple tests', *Health Services Research*, Vol. 41, No. 3, Part 1, pp.804–818 [online] http://doi.org/10.1111/j.1475-6773.2006.00512.x.